A Distributional View of High Dimensional Optimization

Inauguraldissertation
zur Erlangung des akademischen Grades
eines Doktors der Naturwissenschaften
der Universität Mannheim

vorgelegt von

Felix Benning
aus Nürtingen

Mannheim, 2025

Dekan:	Prof. Dr. Claus Hertling, Universität Mannheim
Referent:	Prof. Dr. Leif Döring, Universität Mannheim
Korreferent:	Prof. Dr. Simon Weissmann, Universität Mannheim
Korreferent:	Prof. Dr. Steffen Dereich, Universität Münster

Tag der mündlichen Prüfung: 18. Juli 2025

To Veronika

Contents

Acknowledgements

Writing a dissertation in mathematics is an exercise for the privileged. I was fortunate to be born into a stable family, where my parents – especially my mother – always had the energy to discuss any rule they set. I was lucky to have had Hans Schmiga as my mathematics teacher in school, who bothered to introduce any topic with a derivation for those who were interested. And had the tutor of the mathematics preparatory course for my economics bachelor not responded to my question 'What does matrix multiplication actually *mean*?'[1] with 'I don't know, ask a mathematician', I might never have taken the maths minor that allowed me to gradually shift towards mathematics.

I was also fortunate that my parents had the means to ensure that I could focus on my studies. And by chance, when I began my studies in Mannheim, my supervisor Leif Döring had just started as a junior professor. He is incredibly gifted at teaching the fundamentals, and I was lucky to be part of one of the few cohorts he taught Analysis 1 and 2 to. His ability to motivate abstract concepts and teach how to structure formally correct proofs left a lasting impression on me.

Beyond these pivotal events that led to this attempt at a PhD, there were so many people that helped me through its challenges it is impossible to recount them all. Special thanks goes to my friend Julie Naegelen, who supported me through the ups and downs and always found the time to be a brutal proofreader. And I cannot put into words how grateful I am to my girlfriend Veronika for putting up with my absentminded, sleepless self.

To my family, friends, collegues – everyone that helped me get here, on purpose or by accident – thank you!

[1] I now know: A matrix is a basis representation of a linear map; matrix multiplication is defined to represent the composition of linear maps.

Preface

> *"He asks this shepherd's boy: 'How many seconds in eternity?' The shepherd's boy says, 'there is this mountain of pure diamond. It takes an hour to climb it and an hour to go around it, and every hundred years a little bird comes and sharpens its beak on the diamond mountain. And when the entire mountain is chiseled away, the first second of eternity will have passed.' You may think that's a hell of a long time. Personally, I think that's a hell of a bird."*
>
> — The 12th Doctor, "Heaven Sent"

In many areas of modern science and engineering, complex models are no longer built manually but instead shaped through optimization. As this *machine learning* paradigm has grown in prominence, so too has the importance of general-purpose optimization tools.

Each time a model is modified – whether by changing its architecture, its loss function, or its input representation – the underlying optimization problem shifts. In such a setting, designing a custom optimization method for every iteration is not only impractical, but counterproductive. Consequently, there is little appetite for crafting bespoke optimization algorithms tailored to each new objective. Instead, practitioners rely on black-box optimization.

Black-box optimization provides a clean separation of concerns between the user and the underlying optimization process. In this framework, the user only needs to implement an oracle that can evaluate the objective function and its derivatives up to a certain order. The black-box optimizer takes care of the rest: selecting the next query points, exploring the objective landscape, and iterating toward an optimal solution.

> A k-th order black box optimizers requires a k-th order oracle which provides all derivatives $f(x), \nabla f(x), \dots$ up to order k at an arbitrary point x chosen by the optimization algorithm.

For black box optimization to work, certain assumptions about the objective function f are necessary. It is fairly intuitive that an optimizer cannot optimize a discontinuous function with an infinite domain, so *continuity* is the most basic assumption. However, as we will show in Chapter 1, continuity assumptions alone lead to a curse of dimensionality: the number of oracle evaluations required to achieve a meaningful optimization result grows exponentially with the dimension of the problem. In practical machine learning settings, where models can have millions of parameters, this exponential increase in complexity makes optimization seem impossible. In classical (convex) optimization theory this problem is solved by forcing information into the gradient – a lower bound on the gradient norm using the distance of the function value to the global optimum.[2] Unfortunately, this assumption also implies that all critical points are global optima. This is unrealistic for machine learning, where exponentially more saddle points occur than local minima.[3] Classical optimization theory therefore suggests that non-convex, global optimization is utterly hopeless in high dimension.

[2] This is known as the PL inequality. See Equation (1.2).

[3] Dauphin et al., 2014.

In practice, however, optimization in machine learning seems to work – even in high-dimensional spaces where classical optimization theory suggests failure. This observation implies that there must be a theory that accounts for the *realistic* behavior of objective functions in machine learning, rather than relying on the worst-case scenario that is often used in classical theory.

In Chapter 1, we will examine how the worst-case objective functions that lead to the curse of dimensionality are adversarially chosen: every oracle evaluation returns zero for the function value, gradient and higher oder derivatives. This deprives the optimizer of useful directional information, which would be provided by the gradient for any other value. In contrast, a randomly selected objective function is unlikely to exhibit this pathological behavior.

It is therefore natural to study the **optimization of random functions** to explain the realistic behavior of optimizers. This approach is known as 'Bayesian optimization' or 'Kriging' in the literature. After the first chapter on Worst-case optimization (Chapter 1), the thesis therefore structured as follows.

Outline of Part I (Distributional optimization). In Chapter 2 we properly introduce this Bayesian approach to optimization and Chapter 3 addresses its measure theoretical problems. The perhaps most significant challenge of this approach lies in the selection of a distribution over functions. A pragmatic solution to this problem are uniformity assumptions, specifically isotropy, which postulates that no coordinate system for the random function is inherently favoured. This assumption and other approaches to uniformity are introduced in Chapter 4. Having established the ingredients of random function optimization, we turn to the challenge of optimizing *high-dimensional* random functions for which classical Bayesian optimization methods are ill suited. In Chapter 5 we introduce a Bayesian method, we call 'Random Function Descent', that is scalable to high dimension. Under the isotropy assumption (Chapter 4), this method happens to coincide with gradient descent using a specific step size schedule – bridging the gap between classical worst case optimization approaches and Bayesian optimization. Finally, in Chapter 6 we analyze the behavior of classical gradient span optimization algorithms. We show that asymptotically, as the dimension increases, the optimization progress converges to a deterministic sequence. This implies a tight concentration around the average case performance of optimizers in high dimension.

Outline of Part II (Supervised machine learning). Having motivated the random objective function assumption as an answer to the hopelessness of worst-case optimization in Part I, we replace this argument of desperation with a more axiomatic approach in in Chapter 7. In this approach, specific to supervised machine learning, we will derive random objective functions from the assumption of exchangeable data. Then we outline how this results in an elegant interpretation of the most commonly used loss functions in machine learning as maximum a posteriori estimators. In Section 7.4 we will present evidence against stationarity of the objective function, giving more importance to non-stationary isotropy as characterized in Section 4.4.

Finally, in Chapter 8 we analyze the losslandscape induced by shallow neural networks on random regression problems. Under extremely general unversality assumptions about the target function,[4] we find that on the domain of efficient network parameters[5] the objective function is a Morse function. This implies positive curvature in all directions at local minima and thereby isolated local minima with strongly convex neighborhoods.

[4] all continuous functions lie in the support of the distribution over target functions

[5] the parameters whose response functions cannot be reproduced by smaller networks

Attribution Chapters based on papers I contributed to are often *verbatim* copies of these papers with minor modifications.

Chapter 1: Summarizes well known results on worst-case optimization theory. The generalization to the modulus of continuity is non-standard but unlikely to be new.

Chapter 2: Summarizes well known results about Bayesian optimization.[6]

[6] e.g. Garnett, 2023.

Chapter 3: Verbatim copy of Benning (2025) except for its introduction, which was moved to the previous chapter.

Chapter 4: Section 4.2 is based on Section F in Benning and Döring (2024b), Section 4.3 and 4.4 are based on Benning and Schölpple (2025). Section 4.5 and 4.6 are based on Benning and Döring (2024a)

Chapter 5: Based on Benning and Döring (2024b).
Chapter 6: Based on Benning and Döring (2024a).
Chapter 7: To the best of our knowledge the framework of exchangable data for supervised learning is new, however the motivation of the MSE and crossentropy via the maximum likelihood is well known.[7] The random linear model was already introduced in Benning and Döring (2024b) except for the analysis of the distribution of the noise.
Chapter 8: Verbatim copy of Benning and Dereich (2025) with shortened introduction.

[7] e.g. Goodfellow et al., 2016.

Bibliography

Benning, Felix (2025). *Measure Theory of Conditionally Independent Random Function Evaluation*. arXiv: 2504.08513. Pre-published.

Benning, Felix and Steffen Dereich (2025). *In Almost All Shallow Analytic Neural Network Optimization Landscapes, Efficient Minimizers Have Strongly Convex Neighborhoods*. arXiv: 2504.08867. Pre-published.

Benning, Felix and Leif Döring (2024a). *Gradient Span Algorithms Make Predictable Progress in High Dimension*. arXiv: 2410.09973. Pre-published.

— (2024b). "Random Function Descent". In: *Advances in Neural Information Processing Systems*. NeurIPS. Vol. 37. Vancouver, Canada: Curran Associates, Inc., pp. 111248–111298. arXiv: 2305.01377. URL: https://proceedings.neurips.cc/paper_files/paper/2024/hash/c980ad0fb46f0cfb0faabcd42b30a67a-Abstract-Conference.html (visited on 2025-04-13).

Benning, Felix and Max David Schölpple (2025). *Schoenberg Characterization of Continuous Non-Stationary Isotropic Positive Definite Kernels*. arXiv: 2506.22048 [math]. URL: http://arxiv.org/abs/2506.22048 (visited on 2025-07-09). Pre-published.

Dauphin, Yann N et al. (2014). "Identifying and Attacking the Saddle Point Problem in High-Dimensional Non-Convex Optimization". In: *Advances in Neural Information Processing Systems*. Vol. 27. Montréal, Canada: Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper/2014/hash/17e23e50bedc63b4095e3d8204ce063b-Abstract.html (visited on 2022-06-10).

Garnett, Roman (2023). *Bayesian Optimization*. 1st edition. Cambridge, United Kingdom ; New York, NY: Cambridge University Press. 358 pp. ISBN: 978-1-108-42578-0. URL: https://bayesoptbook.com/.

Goodfellow, Ian, Yoshua Bengio, and Aaron Courville (2016). *Deep Learning*. MIT Press. 801 pp. ISBN: 978-0-262-33737-3. Google Books: omivDQAAQBAJ.

Notation

$[a, b)$	closed-open interval $[a, b) = \{x \in \mathbb{R} : a \leq x < b\}$
$[i\!:\!j)$	closed-open discrete interval $[i\!:\!j) = [i, j) \cap \mathbb{Z}$ with integers $\mathbb{Z}$
$S((x_i)_{i\in I})$	the forward shift on sequences, i.e. $S((x_i)_{i\in I}) = (x_{i+1})_{i\in I}$
$\langle x, y\rangle$, $x \cdot y$	the inner product between x and y
$\|x\|$	a norm, typically induced by the inner product $\|x\|^2 = \langle x, x\rangle$
$\mathbb{S}^{d-1}$	the unit sphere in $\mathbb{R}^d$ given by $\mathbb{S}^{d-1} = \{x \in \mathbb{R}^d : \|x\| = 1\}$
$\mathbb{N}$, $\mathbb{N}_0$	the natural numbers starting at 1 and 0 respectively
ℓ^2	the hilbert space of sequences $\ell^2 = \{(x_i)_{i\in\mathbb{N}} : \sum_{i=1}^{\infty} x_i^2 < \infty\}$
$E(d)$	the group of (Euclidean) isometries, i.e. $\|\phi(x) - \phi(y)\| = \|x - y\|$ for any $\phi \in E(d)$
$O(d)$	the orthogonal group of linear isometries $O(d) = \{\phi \in E(d) : \phi \text{ linear}\}$
$T(d)$	the group of translations, i.e. $\phi(x) = x + v$ for some v
$\overline{z}$	the conjugate $\overline{z} = x - iy$ of the complex number $z = x + iy \in \mathbb{C}$
$\mathbb{I}$	the identity (matrix)
$\mathbf{1}$	the function $x \mapsto 1$ or vector $(1, \ldots, 1)$ in finite dimensional space
$\mathbf{1}_A$	the indicator function of the set A, i.e. $\mathbf{1}_A(x)$ is 1 if $x \in A$ and 0 else
δ_x	the dirac distribution $\delta_x(A) = \mathbf{1}_A(x)$
$\overline{A}$	the closure of the set A
$\omega_f(\delta)$	the modulus of continuity of the function f, see Definition 1.1.3
$C(\mathbb{X}, \mathbb{Y})$	the set of continuous functions with domain $\mathbb{X}$ and co-domain $\mathbb{Y}$
$(\Omega, \mathcal{A}, \mathbb{P})$	the underlying probability space Ω, σ-algebra of measurable sets $\mathcal{A}$ and probability measure $\mathbb{P}$
$\mathcal{B}(E)$	the Borel sigma algebra induced by the open subsets of E, assuming the topology is obvious
$\mathbb{E}$	the expectation, i.e. $\mathbb{E}[X] = \int X(\omega)\mathbb{P}(d\omega)$
a.s.	almost surely
iid	independent, identically distributed
$\mathcal{N}(\mu, \sigma^2)$	the normal distribution with mean $\mu \in \mathbb{R}$ and variance $\sigma^2 \geq 0$
$\mathcal{N}(\mu, \Sigma)$	the multivariate normal distribution with mean $\mu \in \mathbb{R}^d$ and covariance matrix $\Sigma \in \mathbb{R}^{d\times d}$
$\mathcal{N}(\mu, \mathcal{C})$	the distribution of a Gaussian random function with mean function μ and covariance function $\mathcal{C}$
$\mathcal{U}(a, b)$	the uniform distribution on the interval (a, b)
$\mu_{\mathbf{f}}$	the mean function $\mu_{\mathbf{f}}(x) = \mathbb{E}[\mathbf{f}(x)]$ of the random function $\mathbf{f}$
$\mathcal{C}_{\mathbf{f}}$	the covariance function $\mathcal{C}_{\mathbf{f}}(x, y) = \mathrm{Cov}(\mathbf{f}(x), \mathbf{f}(y))$ of the random function $\mathbf{f}$

Chapter 1

Worst case black box optimization

As outlined in the preface, this chapter is concerned with the hopelessness of worst case global optimization in high dimension. We start with zero-order black box optimizers in Section 1.1 and proceed to explain why differentiability of the objective function and higher order oracles change little about this fact in Section 1.2. In Section 1.2 we further outline how classical optimization theory solves this problem by forcing information into the gradient, at the cost of assuming all critical points to be global minima in the process. Since the assumption that 'all critical points are global minima' is completely implausible for many applications such as machine learning[1] we will motivate a distributional view of optimization, i.e. optimization on random functions in Chapter 2.

Before we do so – to cover classical optimization theory 'once and for all' and to hint at a possible connection to random function optimization theory – there are two sections sprinkled into this chapter marked with a star*. In these we present classical optimization results in terms of the modulus of continuity, which is more general than the common Lipschitz assumptions.[2] These result may become useful for the optimization of random functions since there are entropy bounds on their modulus of continuity.[3]

[1] e.g. Dauphin et al., 2014.

[2] While this generality is not commonly found in standard textbooks, we do not believe that these modulus of continuity results are new.

[3] e.g. Adler and Taylor, 2007.

1.1 High dimensional global optimization is hopeless

Without any assumptions about the objective function to optimize, it would be necessary to try every possible input. This becomes impossible as soon as the function has an infinite domain. A natural assumption is that the evaluation of an objective function f at x is informative of $f(x')$ for x' in a neighborhood of x. Intuitively this is the assumption of continuity, which classical optimization theory always assumes in some form or another.[4]

A very intuitive continuity assumption is L-Lipschitz continuity, which limits the rate of change of a function f to $L \geq 0$, i.e.

$$|f(x) - f(y)| \leq L\|x - y\|.$$

A differentiable function f is in fact L-Lipschitz if and only if its derivative is bounded by L.

The constant L acts as a conversion factor of the distance between x and y and their function values. Ensuring that all function values $f(y)$ of points y selected from a δ-ball around x are within an ϵ tolerance of $f(x)$ requires $\delta \leq \frac{\epsilon}{L}$ which leads to

$$|f(x) - f(y)| \leq L\|x - y\| \leq L\delta \leq \epsilon.$$

[4] In a probabilistic framework with a random objective function $\mathbf{f}$ it is possible to capture the information provided by $\mathbf{f}(x)$ for $\mathbf{f}(x')$ with a covariance, or more generally a conditional distribution. This allows far more types of dependence than a simple "similar if close-by" relation that the continuity assumption represents.

Sufficient samples to find ϵ-optimum To find the global optimum to ϵ-precision it is thus sufficient to cover the space with $\delta = \frac{\epsilon}{L}$-balls, sample from the center of every ball and choose the largest (or respectively smallest function value). The number of samples required is thus upper bounded by the δ-covering number.

Definition 1.1.1 (Covering number)**.** The *δ-covering number* of a space $\mathbb{X}$ is given by the smallest number of closed δ-balls that cover $\mathbb{X}$ and is denoted by $N_{\mathbb{X}}(\delta)$.

Considering a square $[0,1]^d$ it is intuitive that the covering number should behave asymptotically like δ^{-d}. It is easiest to see this using the sup-norm as the balls are then simply cubes with side length δ. More generally the number $d \geq 0$ which asymptotically satisfies

$$N_{\mathbb{X}}(\delta) \sim C\delta^{-d}$$

for some constant C and $\delta \to 0$ is called the 'box-counting dimension'[5] and the box-counting dimension of an m-dimensional submanifold of $\mathbb{R}^n$ is equal to m.[6]

[5] e.g. Falconer, 2003.

[6] Falconer, 2003, Sec. 3.2.

Necessary samples to find ϵ-optimum The inequalities used above can be made tight in the class of Lipschitz functions by setting the slope of a bump function equal to L (cf. Figure 1.1). Specifically, the bump function

$$\varphi_\delta(x) = L(\|x\| - \delta)\mathbf{1}_{\|x\| \leq \delta}$$

is an L-Lipschitz function that is zero outside the δ-ball at the origin. To show that an algorithm cannot pick fewer than $N_{\mathbb{X}}(\delta)$ evaluations we essentially place the optimum using the bump function g in the last remaining δ-ball (cf. Figure 1.2).[7] To assume without loss of generality that the algorithm picks one of its evaluation points as the decision point we gift it one additional point and therefore only prove that there exists no algorithm which can pick *fewer* than $N_{\mathbb{X}}(\delta) - 1$ evaluation points. We then consider what happens to the algorithm when all evaluations are zero. This results in a sequence of evaluation points (including the final decision) $x_1, \dots x_{N_{\mathbb{X}}(\delta)-1}$ which we can interpret as centers of closed δ-balls. These balls cannot cover $\mathbb{X}$ as this would be a contradiction to the definition of $N_{\mathbb{X}}(\delta)$. Thus there exists a point x_* which is of at least $\delta^+ > \delta$ distance to every evaluation point x_i. Then the evaluation collected by the algorithm are consistent with the bump function $g_{\delta^+}(\cdot - x_*)$, which is zero at distance δ^+ away from x_* and thus at the x_i. But the minimum of φ_δ^+ is at $-\epsilon^+$ with

$$\epsilon^+ := L\delta^+ > L\delta = \epsilon$$

and the error of the algorithm on g_{δ^+} is therefore greater than ϵ.

Put together we have proven the following result

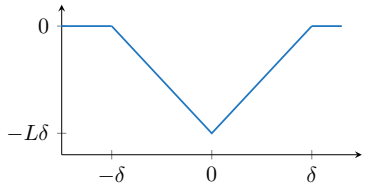


FIGURE 1.1: The L-Lipschitz bump function φ_δ.

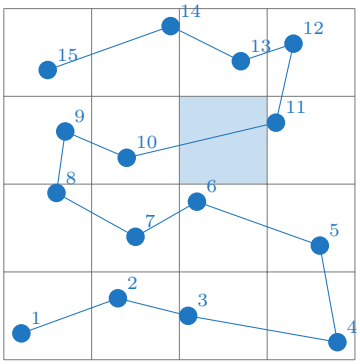


FIGURE 1.2: Sampling a 4×4 grid with $15 < 4 \cdot 4$ points always leaves a tile untouched.

[7] In this picture we motivate the packing number (smallest number of disjoint balls, i.e. tiling), whereas in the proof we use the covering number again and draw balls around the points. This is because it is harder to visualize that there must be an uncovered point left-over than an empty field. While the packing number has the same asymptotic behavior (Falconer, 2003) tight results require the covering number.

Proposition 1.1.2. *The optimal evaluation complexity* $\text{EvalComp}^*(\epsilon)$ *to guarantee an ϵ-optimal function value in the space of L-Lipschitz functions with a zero order black box algorithm on the domain $\mathbb{X}$ is bounded by*

$$N_{\mathbb{X}}(\tfrac{\epsilon}{L}) - 1 \leq \text{EvalComp}^*(\epsilon) \leq N_{\mathbb{X}}(\tfrac{\epsilon}{L}).$$

In particular, we have

$$\text{EvalComp}^*(\epsilon) \sim C(\tfrac{\epsilon}{L})^{-d}$$

for some constant $C > 0$ and box-counting dimension d of $\mathbb{X}$. Moreover this complexity is achieved by a naive 'grid search' algorithm, which covers the space by $N_{\mathbb{X}}(\frac{\epsilon}{L})$ balls, choosing the centers as its evaluation points and selecting the point with the largest (or respectively smallest) function value.

Not only does this imply that the optimal algorithm is essentially a dumb type of grid search, but since the 'Box-counting dimension' d of $\mathbb{X}$ coincides with the intuitive dimension on Manifolds and in particular $\mathbb{R}^d$ this implies exponential complexity in the ambient dimension. From this perspective black box optimization is essentially doomed in high dimension d.

1.1.1 Modulus of continuity: Grid search*

Not every continuous function is Lipschitz continuous and one may be curious, whether or not it is possible to generalize Prop. 1.1.4. Indeed, with the modulus of continuity it is possible to quantify continuity more generally.

Definition 1.1.3 (Modulus of continuity)**.** The *modulus of continuity of* f is defined by

$$\omega_f(\delta) := \sup\{|f(x) - f(y)| : x, y \in \mathbb{X},\ d(x, y) \leq \delta\},$$

where d is a metric on the domain $\mathbb{X}$ of f.

Functions which can be moduli of continuity can be characterized as follows.[8].

[8] e.g. Efimov (2025), see Section 1.A for a proof.

Proposition 1.1.4 (Characterization of moduli of continuity)**.** *For a function* $\omega : [0, \infty) \to [0, \infty)$ *the following are equivalent:*

(i) There exists a uniformly continuous function $f : \mathbb{X} \to \mathbb{R}$*, where* $\mathbb{X}$ *is a convex subset of a normed vectorspace, such that* ω *is the modulus of continuity of* f*, i.e.* $\omega = \omega_f$

(ii)
- *a)* $\omega(0) = 0$*,*
- *b)* ω *is non-decreasing,*
- *c)* ω *is continuous in* 0*,*
- *d)* ω *is subadditive, i.e.* $\omega(\eta + \delta) \leq \omega(\eta) + \omega(\delta)$

If these statements are true, ω *is continuous, its modulus of continuity is itself, i.e. for all* $\eta, \delta \geq 0$

$$|\omega(\eta) - \omega(\delta)| \leq \omega(|\eta - \delta|),$$

and we call ω *"a modulus of continuity".*

With moduli of continuity characterized it is possible to define more general continuity classes.

Definition 1.1.5 (Continuity class)**.** Functions which are dominated by some modulus of continuity ω, i.e.

$$\mathcal{F}_\omega := \{f \colon \mathbb{X} \to \mathbb{R} : \omega_f(\delta) \leq \omega(\delta)\}$$

are the continuity class associated to ω.

Of these classes, Lipschitz continuos functions are simply an example.

Example 1.1.6. $\mathcal{F}_\omega$ with $\omega(\delta) = L\delta$ is exactly the space of L-Lipschitz continuous functions, since for Lipschitz functions we have

$$\omega_f(\delta) \equiv \sup_{d(x,y)\leq\delta} |f(x) - f(y)| \leq \sup_{d(x,y)\leq\delta} Ld(x, y) = L\delta.$$

Recall that the conversion from ϵ to δ was previously given by $\delta = \frac{\epsilon}{L}$. More generally the conversion from function value scale to domain scale is given by $\delta = \omega^{-1}(\epsilon)$, which necessitates the introduction of the pseudo inverse

$$\omega^{-1}(\epsilon) := \sup\{\delta \geq 0 : \omega(\delta) \leq \epsilon\}$$

to cover the non-bijective case.

For completeness sake let us properly define "grid-search".

Definition 1.1.7 (Generalized (ϵ, ω)-grid search)**.** Let $\delta = \omega^{-1}(\epsilon)$ and let $x_i \in \mathbb{X}$ be selected such that the balls $\mathrm{B}_\delta(x_1), \ldots, \mathrm{B}_\delta(x_{N_\mathbb{X}(\delta)})$ cover $\mathbb{X}$ (which is possible by the definition of the covering number $N_\mathbb{X}(\delta)$). Then evaluate $f : \mathbb{X} \to \mathbb{R}$ at every x_i and return

$$\hat{x} = \operatorname*{arg\,min}_{i=1,\ldots,N(\delta)} f(x_i).$$

With grid search properly defined it is possible to generalize Prop. 1.1.2.

Theorem 1.1.8 (Grid search has optimal worst-case complexity)**.** *Generalized (ϵ, ω)-grid search is guaranteed to find an ϵ-minimal function value in the class $\mathcal{F}_\omega$ using $N_{\mathbb{X}}(\omega^{-1}(\epsilon))$ function evaluations. The optimal worst-case evaluation complexity* $\text{EvalComp}^*(\epsilon)$ *is furthermore bounded by*

$$N_{\mathbb{X}}(\omega^{-1}(\epsilon)) - 1 \leq \text{EvalComp}^*(\epsilon) \leq N_{\mathbb{X}}(\omega^{-1}(\epsilon))$$

For the box-counting dimension d of the domain $\mathbb{X}$ there thus exists a constant C such that

$$\text{EvalComp}^*(\epsilon) \sim C[\omega^{-1}(\epsilon)]^d.$$

Proof. Let $\delta = \omega^{-1}(\epsilon)$ and $x_1, \ldots, x_{N_{\mathbb{X}}(\delta)}$ be the points chosen by grid search such that that their δ neighborhoods are a cover $\mathbb{X}$. For any $x \in \mathbb{X}$ there exists x_j such that their distance is smaller than $\delta = \omega^{-1}(\epsilon)$ which implies

$$f(x_j) - f(x) \leq |f(x_j) - f(x)| \leq \omega_f(\delta) \leq \omega(\delta) \leq \epsilon$$

Then we have for the point $\hat{x}$ selected by grid search

$$f(\hat{x}) = \arg\min_i f(x_i) \leq f(x_j) \leq f(x) + \epsilon.$$

As x was arbitrary we have ϵ-optimality, i.e.

$$f(\hat{x}) \leq \inf_x f(x) + \epsilon.$$

Optimality: We now want to show that there exists no algorithm which requires fewer than $N(\omega^{-1}(\epsilon)) - 1$ oracle evaluations. Assume that some algorithm would start with x_1 and choose x_i if $f(x_j) = 0$ for all $j < i$ and would terminate after n steps with $n < N(\omega^{-1}(\epsilon)) - 1$. It then chooses point $\hat{x}$. We are now going to gift this algorithm one last evaluation (setting $x_{n+1} = \hat{x}$) and can thus assume that the algorithm picks

$$\hat{x} = \arg\min_{i=1,\ldots,n+1} f(x_i)$$

without loss of generality. Furthermore $n + 1 < N(\omega^{-1}(\epsilon))$ so so the $\mathrm{B}_\delta(x_i)$ with $\delta = \omega^{-1}(\epsilon)$ do not cover $\mathbb{X}$. Therefore there exists $x^* \in \mathbb{X}$ with

$$\delta^+ := \min_{i=1,\ldots,n+1} d(x^*, x_i) > \delta,$$

Since $\delta = \omega^{-1}(\epsilon)$ we have by definition of ω^{-1} and $\delta^+ \geq \delta$ that $\omega(\delta^+) > \epsilon$. We now define

$$f(x) := \min\big\{\omega(d(x, x^*)) - \omega(\delta^+), 0\big\}$$

Since $d(x_i, x^*) \geq \delta^+$ we have by the non-decreasing property of ω that

$$\omega(d(x_i, x^*)) \geq \omega(\delta^+)$$

and thus $f(x_i) = 0$. So the algorithm will select these x_i but we have

$$f(x_*) = -\omega(\delta^+) < -\epsilon.$$

This implies

$$\arg\min_i f(x_i) = 0 > \inf_x f(x) + \epsilon.$$

This would be a contradiction to ϵ optimality if $f \in \mathcal{F}_\omega$. To show $f \in \mathcal{F}_\omega$, consider that $x \mapsto \min\{x, 0\}$ is 1-Lipschitz resulting in

$$|f(x) - f(y)| \leq |\omega(d(x, x^*)) - \omega(d(y, x^*))| \overset{(*)}{\leq} \omega(d(x, y)).$$

Where $(*)$ is the reverse triangle inequality for $\omega \circ d$, which is a semi-metric because it is positive, symmetric and satisfies the triangle inequality since ω is increasing and sub-additive as a modulus of continuity (Prop. 1.1.4), so we have

$$\omega(d(x, y)) \overset{\text{increasing}}{\leq} \omega(d(x, z) + d(z, y)) \overset{\text{subadditive}}{\leq} \omega(d(x, z)) + \omega(d(z, y)).$$

Noting that x and y were arbitrary, we thus conclude

$$\omega_f(\delta) = \sup_{d(x,y) \leq \delta} |f(x) - f(y)| \leq \sup_{d(x,y) \leq \delta} \omega(d(x, y)) \leq \omega(\delta). \qquad \square$$

1.2 Differentiable objective functions

Assume the objective function f is differentiable with continuous gradients to ensure they are useful. It is again instructive to start with the L-Lipschitz continuity condition for gradients, i.e.

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|. \tag{1.1}$$

The objective function f is then also called 'L-smooth'. The important question is whether this restricted function class allows for speed-ups in optimization. In particular, does the use of gradient evaluations, i.e. first order optimization, help?

For intuition consider the L-smooth bump function (cf. Figure 1.3)[9]

$$\varphi_\delta(x) = \begin{cases} \frac{L}{4}[2\|x\|^2 - \delta^2] & \|x\| \leq \frac{\delta}{2} \\ -\frac{L}{2}(\delta - \|x\|)^2 & \frac{\delta}{2} \leq \|x\| \leq \delta \\ 0 & \|x\| \geq \delta. \end{cases}$$

While it may be no longer possible to reach the height $L\delta$, it is possible to reach height ϵ within a larger δ-ball, specifically $\delta = 2\sqrt{\epsilon/L}$. With the same covering argument as before it can be shown that any algorithm which evaluates less than $N_{\mathbb{X}}(\delta)$ places leaves a δ^+-ball open where the φ_δ bump function can be placed. This implies that at every evaluation location not only the function value but also its gradients are zero. Gradient evaluations therefore do not have any benefit at all. By the same covering argument one can obtain

$$N_{\mathbb{X}}(2\sqrt{\epsilon/L}) - 1 \leq \text{EvalComp}^*(\epsilon) \leq N_{\mathbb{X}}(2\sqrt{\epsilon/L})$$

and for some constant $C \geq 0$ and the box-counting dimension d this implies

$$\text{EvalComp}^*(\epsilon) \sim C(\epsilon/L)^{d/2}.$$

While L-smoothness halves the dimension it therefore does not fix the issue of exponential complexity in the dimension. Furthermore the gradient evaluations are seemingly useless as zero-order grid search is still optimal. More generally, global optimization has worst-case sample complexity of the order $(\epsilon/L)^{-d/r}$ to get within ϵ-distance of the minimum of a function in $C^r([0,1]^d)$ with r-th derivatives bounded by L.[10]

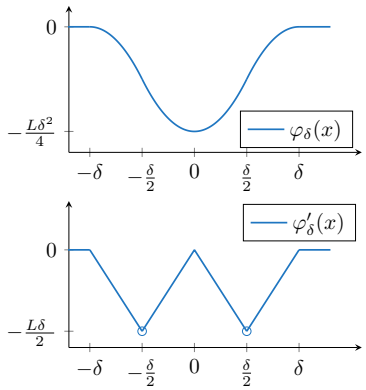


FIGURE 1.3: L-smooth bump function φ_δ

[9] The proof of L-smoothness is more annoying than one might think based on the pictures (details in Lemma 1.A.1).

Why are gradient evaluations seemingly useless?

Gradients are useful because they point towards a reasonable search direction, *assuming they are non-zero*. In high dimension directional information is invaluable since there are many directions to choose from. But the adversarially chosen bump function does not grant this information since gradients are zero and thus directionless. Similarly, if you had lower and higher function evaluations then, due to continuity, the vicinities of lower function values are more likely to contain the global minimum. This could inform search decisions. But the adversarial bump function withholds any information by being completely flat.

However, if information is available an algorithm ought to use it and grid search decidedly does not. Its optimality therefore hinges on the fact that in the worst case there is no information to use. Classical optimization theory therefore forces information upon the function by assuming the gradient cannot be zero outside the global minimum. One such example is the Polyak-Łojasiewicz (PL) inequality,[11] which demands

$$\mu(f(x) - \min_x f(x)) \leq \tfrac{1}{2}\|\nabla f(x)\|^2 \tag{1.2}$$

for some $\mu > 0$. And this assumption is sufficient to obtain exponential convergence of the objective value to the global optimum with the additional assumption of L-smoothness.[11] Of course this assumption also implies that all critical points are global minima, as does any weaker assumption that forces the gradient to be non-zero outside of a global minimum. This implies that it is sufficient to find *any* critical point, which was not the case in the adversarial example where every evaluation point was critical.

[10] Wasilkowski, 1992.

[11] e.g. Karimi et al., 2016.

Remark 1.2.1 (Convexity). The assumption of strong convexity, i.e.

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \tfrac{\mu}{2}\|y - x\|^2,$$

implies a lower bound[12] of the gradient in terms of the distance to the unique minimizer x_* (that exists by convexity)

$$\mu\|x_* - x\| \leq \|\nabla f(x)\|. \tag{1.3}$$

For L-smooth functions (1.1) this is stronger than the PL-inequality, since[13]

$$f(x) - f(x_*) \overset{L\text{-smooth}}{\leq} \tfrac{L}{2}\|x - x_*\|^2 \overset{(1.3)}{\leq} \tfrac{L}{2\mu}\|\nabla f(x)\|^2.$$

As minima have locally convex neighborhoods, not much generality is lost to assume convexity at this point, since the convex neighborhoods determine asymptotic behavior. And convex function theory is fully developed with tight lower bounds and convergence algorithms that achieve these lower bounds.[14]

[12] Due to $f(x) - f(x_*) \geq 0$, reordering strong convexity with $y = x_*$ and the Cauchy-Schwarz inequality imply

$$\begin{aligned}\tfrac{\mu}{2}\|x_* - x\|^2 &\leq \langle f(x), x_* - x\rangle \\ &\leq \|\nabla f(x)\|\|x_* - x\|.\end{aligned}$$

Division by $\|x_* - x\|$ yields the claim (this proof is from Karimi et al., 2016).

[13] The first inequality follows from Lemma 1.2.6 for L-smooth f and $\nabla f(x_*) = 0$.

[14] e.g. Nesterov, 2018; Garrigos and Gower, 2023; Goh, 2017; Bottou et al., 2018.

1.2.1 Modulus of continuity: Descent lemma and convergence*

In this section we want to briefly mention that the "Decent Lemma" that drives most convergence proofs in classical optimization theory can also be generalized to the modulus of continuity. Accordingly, define the modulus of continuity of the gradient to be

$$\omega_{\nabla f}(\delta) := \sup\{\|\nabla f(x) - \nabla f(y)\| : \|x - y\| \leq \delta\}.$$

Recall that L-smoothness is defined by L-Lipschitz gradients, i.e. the special case

$$\omega_{\nabla f}(\delta) = L\delta. \tag{1.4}$$

With the help of this modulus of continuity it is possible to generalize the 'Descent Lemma'.[15]

[15] e.g. Garrigos and Gower, 2023, Lemma 2.28.

Lemma 1.2.2 (General Descent Lemma)**.** *Steps in the direction of the gradient result in*

$$f\left(x - \eta\tfrac{\nabla f(x)}{\|\nabla f(x)\|}\right) \leq f(x) - \int_0^\eta \|\nabla f(x)\| - \omega_{\nabla f}(t)dt,$$

where the minimizing step size is given by $\eta^* = \omega_{\nabla f}^{-1}(\|\nabla f(x)\|)$ *and if* $\omega_{\nabla f}$ *is differentiable*

$$f\left(x - \eta^*\tfrac{\nabla f(x)}{\|\nabla f(x)\|}\right) \leq f(x) - \underbrace{\left(\int_0^1 \frac{1-t}{\omega'_{\nabla f}(t\|\nabla f(x)\|)}dt\right)}_{\overset{L\text{-smooth}}{=}\frac{1}{2L}}\|\nabla f(x)\|^2.$$

Remark 1.2.3 (Step size vs. learning rate). In the case of L-smooth objective functions (1.4), the optimal step size is given by

$$\eta^* = \omega_{\nabla f}^{-1}(\|\nabla f(x)\|) = \tfrac{\|\nabla f(x)\|}{L}.$$

The resulting optimization step is thus characterized by

$$x_{n+1} = x_n - \underbrace{\eta^*}_{\text{'stepsize'}}\tfrac{\nabla f(x_n)}{\|\nabla f(x_n)\|} = x_n - \underbrace{h^*}_{\text{'learning rate'}}\nabla f(x_n)$$

with 'learning rate' $h^* = \frac{1}{L}$. Classical optimization theory typically considers these learning rates associated to the un-normalized gradient, while step sizes appear to be more natural in the general theory using moduli of continuity. Step sizes are also more natural in the theory of optimization on random functions as will become apparent in the chapter on Random Function Descent (Chapter 5)

We explain the significance of the Descent Lemma by proving that convergence proofs can be easily obtained by its use. We provide two examples, one proves convergence and the other convergence rates.

Theorem 1.2.4 (Convergence)**.** *Assume f is L-smooth, $\inf_x f(x) > -\infty$ and $x_{n+1} = x_n - \frac{1}{L}\nabla f(x_n)$. Then*

$$\lim_{n\to\infty} \|\nabla f(x_n)\| = 0.$$

Proof. By Lemma 1.2.2 we have

$$f(x_{n+1}) - f(x_n) \leq -\tfrac{1}{2L}\|\nabla f(x_n)\|^2$$

and therefore

$$\sum_{k=0}^{n} \|\nabla f(x_k)\|^2 \leq 2L \sum_{k=0}^{n} (f(x_k) - f(x_{k+1})) \overset{\text{telescope}}{\leq} 2L\big(f(x_0) - \inf_x f(x)\big).$$

This bound is finite by assumption and holds uniformly over n, thus the squared gradient norms are infinitely summable and thereby converge to 0. □

It is clear that if the gradients converge to zero and the only place where the gradient may be zero is in global minima, then the function values should also converge to the minimum. The PL-inequality (1.2) ensures this is the case and indeed, the following result provides convergence rates for the case of L-smooth functions which satisfy the PL-inequality (1.2).

Theorem 1.2.5 (Convergence rates)**.** *Let f be L-smooth, $f_* := \inf_x f(x) > -\infty$, f satisfies the PL-inequality (1.2) and*

$$x_{n+1} = x_n - \tfrac{1}{L}\nabla f(x_n),$$

then we have exponential convergence[16]

$$f(x_n) - f_* \leq \big(1 - \tfrac{\mu}{L}\big)^n \big(f(x_0) - f_*\big).$$

[16] typically referred to as 'linear convergence' in the optimization literature due to the linear reduction in every step.

Proof. Since $\frac{\|\nabla f(x_n)\|}{L}$ is the minimizing step size of the upper bound we have by the Descent Lemma (Lemma 1.2.2)

$$\begin{aligned} f(x_{n+1}) - f_* &\overset{\text{Lem. 1.2.2}}{\leq} f(x_n) - f_* - \tfrac{1}{2L}\|\nabla f(x_n)\|^2 \\ &\overset{(1.2)}{\leq} (1 - \tfrac{\mu}{L})(f(x_n) - f_*). \end{aligned}$$

Here we used the the PL-inequality (1.2), i.e. $\mu(f(x_n) - f_*) \leq \frac{1}{2}\|\nabla f(x)\|$, in the last line. Induction over n finishes the proof. □

Clearly, weaker assumptions about continuity weaken the first inequality provided by the Descent Lemma (Lemma 1.2.2). Similarly, lower bounds on the gradient that are weaker than the PL-inequality will weaken the second inequality. In essence convergence can be proven as long as the combination of these assumptions results in a sufficient reduction factor.

In the remainder of this section we prove the Descent Lemma. The main vehicle of the proof of Lemma 1.2.2 is the following bound based on the Taylor approximation.

Lemma 1.2.6 (Taylor approximation with trust bound)**.** *Let f be a continuously differentiable function whose gradient has modulus of continuity $\omega_{\nabla f}$. Then*

$$f(y) \leq f(x) + \langle \nabla f(x), y - x\rangle + \underbrace{\int_0^{\|y-x\|} \omega_{\nabla f}(t)dt}_{\overset{L\text{-smooth}}{=} \frac{L}{2}\|y-x\|^2}.$$

Proof. By the fundamental theorem of calculus (FTC), the Cauchy-Schwarz inequality (CS) and $\mathbf{d} = y - x$ we have

$$\begin{aligned}
f(y) &= f(x) + (f(y) - f(x)) \\
&\overset{\text{FTC}}{=} f(x) + \int_0^1 \big\langle \nabla f\big(x + \lambda \mathbf{d}\big), \mathbf{d} \big\rangle d\lambda \\
&= f(x) + \langle \nabla f(x), y - x \rangle + \int_0^1 \underbrace{\big\langle \nabla f\big(x + \lambda \mathbf{d}\big) - \nabla f(x), \mathbf{d} \big\rangle}_{\overset{\text{CS}}{\leq} \|\nabla f(x+\lambda \mathbf{d}) - \nabla f(x)\| \|\mathbf{d}\|} d\lambda \\
&\leq f(x) + \langle \nabla f(x), y - x \rangle + \int_0^1 \omega_{\nabla f}(\lambda \|\mathbf{d}\|) \|\mathbf{d}\| d\lambda \\
&= f(x) + \langle \nabla f(x), y - x \rangle + \int_0^{\|\mathbf{d}\|} \omega_{\nabla f}(t) dt. \qquad \square
\end{aligned}$$

Proof of Descent Lemma (1.2.2). Simply plug $y = x - \eta \frac{\nabla f(x)}{\|\nabla f(x)\|}$ into Lemma 1.2.6 to obtain[17]

$$\begin{aligned}
f(y) &\leq f(x) + \underbrace{\left\langle \nabla f(x), -\eta \frac{\nabla f(x)}{\|\nabla f(x)\|} \right\rangle}_{= \eta \|\nabla f(x)\|} + \int_0^\eta \omega_{\nabla f}(t) dt \\
&= f(x) - \int_0^\eta \|\nabla f(x)\| - \omega_{\nabla f}(t) dt. \qquad (1.5)
\end{aligned}$$

What remains is to find the optimal step size. Since $\omega_{\nabla f}$ is monotonously increasing, the integrand (1.5) is positive until

$$\eta^* = \omega_{\nabla f}^{-1}(\|\nabla f(x)\|)$$

and negative afterwards. This is therefore the minimizing step size. Plugging this result into (1.5) we obtain

$$\begin{aligned}
f\left(x - \eta^* \tfrac{\nabla f(x)}{\|\nabla f(x)\|}\right) &\leq f(x) - \int_0^{\eta^*} \|\nabla f(x)\| - \omega_{\nabla f}(t) dt \\
&= f(x) - \int_0^1 \frac{1 - t}{\omega'_{\nabla f}(t \|\nabla f(x)\|)} ds \|\nabla f(x)\|^2.
\end{aligned}$$

with the substitution $s = \frac{\omega_{\nabla f}(t)}{\|\nabla f(x)\|}$. $\square$

[17] Note that the gradient direction minimizes the bound of Lemma 1.2.6. This is because minimization of the bound

$$\min_{\mathbf{d}: \|\mathbf{d}\| = \eta} f(x) + \langle \nabla f(x), \mathbf{d} \rangle + B(\|\mathbf{d}\|),$$

with $B(\eta) = \int_0^\eta \omega_{\nabla f}(t) dt$, only depends on the direction $\mathbf{d} = y - x$ in the inner product, which forces the gradient direction. The proof is a straightforward application of the Cauchy-Schwarz inequality (see Corollary 1.A.3).

1.A Technical Appendix

Lemma 1.A.1 (L-smooth bump function)**.** *For any $\delta, L > 0$ we have that the function g is L-smooth, where g is defined by*

$$g : \begin{cases} \mathbb{R}^d \to \mathbb{R} \\ x \mapsto h(\|x\|) \end{cases} \qquad h(r) : \begin{cases} [0, \infty) \to \mathbb{R} \\ r \mapsto \begin{cases} \frac{L}{4}[2r^2 - \delta^2] & r \leq \frac{\delta}{2} \\ -\frac{L}{2}(\delta - r)^2 & \frac{\delta}{2} \leq r \leq \delta \\ 0 & r \geq \delta. \end{cases} \end{cases}$$

Proof. It is straightforward to confirm (cf. Figure 1.3)

$$\left|\tfrac{h'(r)}{r}\right| \leq L \quad \text{and} \quad |h''(r)| \leq L.$$

for all $r \notin \{0, \frac{\delta}{2}, \delta\}$. Furthermore, the Hessian of g is (for all $\|x\| \notin \{0, \frac{\delta}{2}, \delta\}$) given by

$$\nabla^2 g(x) = h''(\|x\|) \tfrac{x x^T}{\|x\|^2} + \tfrac{h'(\|x\|)}{\|x\|} \left(\mathbb{I} - \tfrac{x x^T}{\|x\|^2}\right)$$

We can decompose any vector v of unit length ($\|v\| = 1$) into the orthogonal components

$$v = \alpha \tfrac{x}{\|x\|} + v^{\perp},$$

which implies

$$|v^T \nabla^2 g(x) v| = \left|h''(\|x\|)\alpha^2 + \tfrac{h'(\|x\|)}{\|x\|}(1-\alpha^2)\right| \leq L.$$

Thus for $\|x\| \notin \{0, \frac{\delta}{2}, \delta\}$ the operator norm of $\nabla^2 g(x)$ is bounded by L. Since $p(t) := \nabla g(x + t(y-x))$ is continuous and *piecewise* differentiable except for a finite number of points, we can apply the fundamental theorem of calculus to obtain

$$\begin{aligned}\|\nabla g(y) - \nabla g(y)\| = \|p(1) - p(0)\| &\leq \int_0^1 \underbrace{\|\nabla^2 g(x + t(y-x))\|}_{\leq L} \|y - x\| dt \\ &\leq L\|y - x\|.\end{aligned}$$

With the understanding that the integral would formally have to be broken into pieces. □

We finally provide a proof for the well known characterization of the modulus of continuity.[18]

[18] e.g. Efimov, 2025.

Proposition 1.1.4 (Characterization of moduli of continuity)**.** *For a function $\omega : [0, \infty) \to [0, \infty)$ the following are equivalent:*

(i) There exists a uniformly continuous function $f : \mathbb{X} \to \mathbb{R}$, where $\mathbb{X}$ is a convex subset of a normed vectorspace, such that ω is the modulus of continuity of f, i.e. $\omega = \omega_f$

(ii)
- *a) $\omega(0) = 0$,*
- *b) ω is non-decreasing,*
- *c) ω is continuous in 0,*
- *d) ω is subadditive, i.e. $\omega(\eta + \delta) \leq \omega(\eta) + \omega(\delta)$*

If these statements are true, ω is continuous, its modulus of continuity is itself, i.e. for all $\eta, \delta \geq 0$

$$|\omega(\eta) - \omega(\delta)| \leq \omega(|\eta - \delta|),$$

and we call ω "a modulus of continuity".

Proof. **(i) ⇒ (ii):** $\omega(0) = 0$ is follows immediately from the definition like the fact that it is non-decreasing. Continuity in 0 is necessitates by the fact that f is uniformly continuous. We prove sub-additivity in two steps:

Step 1: First we prove that whenever $\|x - y\| \leq \eta + \delta$, there exists $z \in \mathbb{X}$ such that $\|x - z\| \leq \eta$ and $\|z - y\| \leq \delta$.

For this we define $\lambda = \frac{\delta}{\eta + \delta}$ and select $z = \lambda x + (1 - \lambda)y$ (which we are allowed to do by convexity). Then

$$\|x - z\| = \|(1-\lambda)(x-y)\| = (1-\lambda)\|x - y\| = \frac{\eta}{\eta + \delta}\|x - y\| \leq \eta.$$

And similarly for the other equation.

Step 2: Now we simply consider the definition

$$\begin{aligned}\omega_f(\eta+\delta) &= \sup\{|f(x)-f(y)| : x,y\in\mathbb{X}, \|x-y\|<\eta+\delta\}\\ &\leq \sup\Big\{|f(x)-f(z)|+|f(z)-f(y)| :\\ &\qquad\qquad x,y\in\mathbb{X}, \|x-y\|<\eta+\delta, z=\tfrac{\delta}{\eta+\delta}x+\tfrac{\eta}{\eta+\delta}y\Big\}\\ &\leq \sup\Big\{|f(x)-f(z)| : x,z\in\mathbb{X}, \|x-z\|\leq\eta\Big\}\\ &\quad+\sup\Big\{|f(z)-f(y)| : y,z\in\mathbb{X}, \|y-z\|\leq\delta\Big\}\\ &= \omega_f(\eta)+\omega_f(\delta).\end{aligned}$$

Where we use that giving more choice only increases the supremum.

(ii) ⇒ (i): By proving the modulus of continuity has itself as its modulus of continuity, we prove existence and the additional claim. By monotonicity and sub-additivity of the modulus of continuity for $\eta,\delta\geq 0$

$$\omega(\eta) \overset{\text{monotonicity}}{\leq} \omega(|\eta-\delta|+|\delta|) \overset{\text{sub-additivity}}{\leq} \omega(|\eta-\delta|)+\omega(\delta)$$

So by symmetry of δ and η

$$|\omega(\eta)-\omega(\delta)|\leq\omega(|\eta-\delta|). \qquad \square$$

1.A.1 Constrained linear optimization

Let U be a vectorspace. We define the projection of a vector w onto U by

$$P_U(w) := \operatorname*{arg\,min}_{v\in U}\|v-w\|^2$$

Lemma 1.A.2 (Constrained maximiziation of scalar products)**.** *For a linear subspace $U\subseteq\mathbb{R}^d$, we have*

$$\max_{\substack{v\in U\\ \|v\|=\lambda}}\langle v,w\rangle = \lambda\|P_U(w)\| \tag{1.6}$$

$$\operatorname*{arg\,max}_{\substack{v\in U\\ \|v\|=\lambda}}\langle v,w\rangle = \lambda\frac{P_U(w)}{\|P_U(w)\|} \tag{1.7}$$

Before we get to the proof let us note that this immediately results in the following corollary about minimization.

Corollary 1.A.3 (Constrained minimization of scalar products)**.**

$$\min_{\substack{v\in U\\ \|v\|=\lambda}}\langle v,w\rangle = -\lambda\|P_U(w)\| \tag{1.8}$$

$$\operatorname*{arg\,min}_{\substack{v\in U\\ \|v\|=\lambda}}\langle v,w\rangle = -\lambda\frac{P_U(w)}{\|P_U(w)\|} \tag{1.9}$$

Proof of Corollary 1.A.3. The trick is to move one '−' outside from $w=-(-w)$

$$\min_{\substack{v\in U\\ \|v\|=\lambda}}\langle v,w\rangle = -\max_{\substack{v\in U\\ \|v\|=\lambda}}\langle v,-w\rangle = -\lambda\|P_U(w)\|$$

where we have used in the last equation that the projection is linear (we can move the minus sign out) and the norm removes the inner minus sign. The arg min argument is similar. $\square$

Proof of Lemma 1.A.2. **Step 1**: We claim that

$$v^* = \lambda \frac{P_U(w)}{\|P_U(w)\|}$$

results in the value $\langle v^*, w\rangle = \lambda\|P_U(w)\|$.

For this we consider

$$\begin{aligned} P_U(w) &= \arg\min_{v\in U} \underbrace{\|v-w\|^2}_{=\|v\|^2-2\langle v,w\rangle+\|w\|^2} \\ &= \arg\min_{v\in U} \underbrace{\|v\|^2 - 2\langle v, w\rangle}_{=:f(v)} \end{aligned} \tag{1.10}$$

we know that $t \mapsto f(t\langle w\rangle_U)$ is minimized at $t = 1$ by the definition of $\langle w\rangle_U$. The first order condition implies

$$0 \overset{!}{=} \frac{d}{dt} = 2t\|P_U(w)\|^2 - 2\langle P_U(w), w\rangle$$

and thus

$$1 = t^* = \frac{\langle P_U(w), w\rangle}{\|P_U(w)\|^2}$$

Multiplying both sides by $\lambda\|P_U(w)\|$ finishes this step

$$\lambda\|P_U(w)\| = \Big\langle \underbrace{\lambda\frac{P_U(w)}{\|P_U(w)\|}}_{=v^*}, w\Big\rangle. \tag{1.11}$$

Step 2: By (1.11), we know that we can achieve the value we claim to be the maximum (and know the location v^* to do so). So if we prove that we can not exceed this value, then it is a maximum and v^* is the arg max. This would finish the proof. What remains to be shown is therefore

$$\langle v, w\rangle \le \lambda\|P_U(w)\| \qquad \forall v \in U : \|v\| = \lambda.$$

Let $v \in U$ with $\|v\| = \lambda$. Then for any $\mu \in \mathbb{R}$ we can plug μv into f from (1.10) to get

$$\begin{aligned} \mu^2\lambda^2 - 2\mu\langle v, w\rangle = f(\mu v) &\overset{(1.10)}{\ge} f(P_U w) = \|P_U(w)\|^2 - 2\langle P_U w, w\rangle \\ &= -\langle P_U w, w\rangle \end{aligned}$$

where the last equation follows from (1.11) with $\lambda = \|P_U w\|$. Reordering we get for all μ

$$\langle P_U w, w\rangle + \mu^2\lambda^2 \ge 2\mu\langle v, w\rangle$$

We now select $\mu = \frac{\|P_U w\|}{\lambda} > 0$ and divide both sides by μ to get

$$2\langle v, w\rangle \le \underbrace{\Big\langle \underbrace{\frac{P_U(w)}{\mu}}_{=v^*}, w\Big\rangle}_{\overset{(1.11)}{=}\lambda\|P_U w\|} + \lambda\|P_U(w)\| = 2\lambda\|P_U w\|$$

Dividing both sides by 2 yields the claim. □

References

Adler, Robert J. and Jonathan E. Taylor (2007). *Random Fields and Geometry*. Springer Monographs in Mathematics. New York, NY: Springer New York. ISBN: 978-0-387-48112-8. DOI: 10.1007/978-0-387-48116-6.

Bottou, Léon, Frank E. Curtis, and Jorge Nocedal (2018). "Optimization Methods for Large-Scale Machine Learning". arXiv: 1606.04838 [cs, math, stat]. URL: http://arxiv.org/abs/1606.04838 (visited on 2021-06-08).

Dauphin, Yann N et al. (2014). "Identifying and Attacking the Saddle Point Problem in High-Dimensional Non-Convex Optimization". In: *Advances in Neural Information Processing Systems*. Vol. 27. Montréal, Canada: Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper/2014/hash/17e23e50bedc63b4095e3d8204ce063b-Abstract.html (visited on 2022-06-10).

Efimov, A.V. (2025). *Continuity, Modulus Of*. In: *Encyclopedia of Mathematics*. ISBN: 1-4020-0609-8. URL: http://encyclopediaofmath.org/index.php?title=Continuity,_modulus_of&oldid=51338 (visited on 2025-02-20).

Falconer, Kenneth (2003). *Fractal Geometry: Mathematical Foundations and Applications*. 2nd ed. Wiley. ISBN: 978-0-470-84861-6. DOI: 10.1002/0470013850.

Garrigos, Guillaume and Robert M. Gower (2023). *Handbook of Convergence Theorems for (Stochastic) Gradient Methods*. DOI: 10.48550/arXiv.2301.11235. Pre-published.

Goh, Gabriel (2017). "Why Momentum Really Works". In: *Distill*. DOI: 10.23915/distill.00006.

Karimi, Hamed, Julie Nutini, and Mark Schmidt (2016). "Linear Convergence of Gradient and Proximal-Gradient Methods Under the Polyak-Łojasiewicz Condition". In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Lecture Notes in Computer Science. Springer, Cham, pp. 795–811. ISBN: 978-3-319-46128-1. DOI: 10.1007/978-3-319-46128-1_50. arXiv: 1608.04636.

Nesterov, Yurii Evgen'evič (2018). *Lectures on Convex Optimization*. Second edition. Springer Optimization and Its Applications; Volume 137. Cham: Springer. ISBN: 978-3-319-91578-4. DOI: 10.1007/978-3-319-91578-4.

Wasilkowski, G. W. (1992). "On Average Complexity of Global Optimization Problems". In: *Mathematical Programming* 57.1, pp. 313–324. ISSN: 1436-4646. DOI: 10.1007/BF01581086.

Part I

Distributional optimization

Chapter 2

Distributional optimization

While it is nice to have worst-case guarantees (when the problem exhibits convexity, the PL-inequality or similar) the optimization problems in machine learning are decidedly not of this type. In fact they exhibit exponentially more saddle points than minima[1] and in particular not all critical points are global minima.

[1] Dauphin et al., 2014.

Despite this, some optimizers from convex function theory such as gradient descent find themselves successfully optimizing machine learning objectives. Other convex optimizers are less successful. The geometric alternative to momentum suggested by Bubeck et al. (2015) for example is virtually unknown in machine learning despite an equivalent worst case complexity to Nesterov's momentum on convex functions. Explaining these successes and failures requires a different mathematical framework that can tackle global optimization questions. And the "just find a local minimum, it is not going to be that bad, typically" heuristic screams for an average case analysis or more generally a distributional analysis.

A distributional analysis has lead to algorithmic improvements in many prominent cases:

- **Sorting algorithms.** Quicksort[2] is successful in practice with an average performance of $\mathcal{O}(n \log(n))$ despite its worst case performance of $\mathcal{O}(n^2)$ and the existence of Merge sort with worst case performance $\mathcal{O}(n \log(n))$.
- **Data structures.** E.g. Hash-tables or Binary-search trees, which sometimes have much better performance on average than in the worst case.
- **Lossless compression algorithms.** Such algorithms have to be injections, so they can not map into a smaller set (a set of numbers with fewer bits) than the origin (pigeonhole principle). But the average case is bounded by entropy instead.[3]
- **Linear programming.** The simplex algorithm has polynomial time complexity on average despite its exponential time complexity in the worst case.[4]
- "Random by design" algorithms such as **Monte Carlo integration**.

[2] Hoare, 1962.

[3] Shannon, 1948, Theorem 9, "Shannon's source coding theorem".

[4] e.g. Borgwardt, 1986; Smale, 1983; Spielman and Teng, 2004; Todd, 1991.

Moreover, the empirical performance of many algorithms concentrates around the average case performance. This suggests that an analysis of the distribution of performance should often lead to tight concentration results. But to analyze the distribution of performance or even just the average performance of an algorithm it is necessary to either

1. randomize the algorithm itself, or
2. make assumptions about the distribution of inputs given to the algorithm.

2.1 Randomized Optimizer

The most salient randomized algorithm for global optimization may be **simulated annealing**.[5] If we can generate samples from the softmax (a.k.a. Gibbs or

[5] e.g. Van Laarhoven and Aarts, 1987; Salamon et al., 2002.

Boltzmann) distribution

$$X \sim \mu_T(dx) = \frac{\exp(\frac{1}{T}f(x))}{\int_{\mathbb{X}} \exp(\frac{1}{T}f(y))dy}dx$$

then as the temperature T cools[6] ($T \to 0$), the distribution concentrates around the global maximizers of f in $\mathbb{X}$. The difficulty of this method lies in sampling from this distribution. A key insight for the case of differentiable f is that *Langevin Diffusion*,[7] i.e. the Îto SDE given by gradient flow and disturbed by Brownian noise $B(t)$

$$dX(t) = \underbrace{\nabla f(X(t))dt}_{\text{gradient flow}} + \underbrace{\sqrt{2T}dB(t)}_{\text{noise}}, \qquad \text{(Langevin Diffusion)}$$

converges to μ_T as $t \to \infty$. The idea is then, to reduce the temperature T slowly enough over time t such that $X(t)$ remains distributed according to μ_T. For example, under suitable assumptions about f, the temperature schedule $T(t) := \frac{c}{\log(t)}$ for a sufficiently large $c > 0$ can accomplish this task and $X(t)$ concentrates on the global maximum.[8] To turn this into an optimization algorithm it is necessary to show that these properties can be retained for a discretized version that can be used in practice.[9]

Remark 2.1.1 (Stochastic gradient descent). Langevin dynamics suggest that noise is sometimes desirable – especially at the beginning of training. This may explain the success of Stochastic gradient descent.

Remark 2.1.2 (Randomization via preprocessing). The randomization of an algorithm also often takes the form of a pre-processing step of the input. For example: Random shuffling as the first step can guarantee that the input of a sorting algorithm is uniform. With the help of these pre-processing steps the analysis can often be reduced to assumptions about the distribution of inputs to the algorithm.

[6] the state "anneals", inspired by the concept of annealing in metallurgy where metals are heated above a certain temperature before a slow gradual cooling process.

[7] Langevin Diffusion and its discretized version are grouped under the umbrella term *Langevin Dynamics*. Sampling from μ_T using this procedure is also known as *Langevin Monte Carlo* (Chen et al., 2024).

[8] Chiang et al., 1987.

[9] further information in e.g. Raginsky et al., 2017; Xu et al., 2018; Zou et al., 2021; Chen et al., 2024.

2.2 Random objective functions

The input to an optimization algorithm is the objective function. A distributional assumption over the input is thus a distribution over objective functions. The optimization of such a *random function* is a fundamental problem that has independently emerged across multiple research domains, each developing its own terminology for very similar methods.

In **geostatistics**, random functions are typically referred to as *random fields*, which are used to model spatial distributions, such as ore deposits at various locations.[10] In this domain, interpolating the underlying function based on limited sample points is known as *Kriging*[11] and used to guide subsequent evaluations – such as where to drill the next pilot hole.

In **Bayesian optimization** (BO),[12] concerned with general black-box function optimization, it is standard to assume a Gaussian prior and refer to random functions as *Gaussian processes*. In BO the subsequent evaluation point x_{n+1} is selected based on the posterior distribution of the random function $\mathbf{f}$ conditional on the previous evaluations $\mathbf{f}(x_0), \dots \mathbf{f}(x_n)$, which coincides with Kriging in the Gaussian case. In recent years, BO has gained prominence in machine learning, particularly for hyperparameter tuning, where evaluating the function (e.g. training a model) is costly.

In the field of **compressed sensing** the goal is to reconstruct a signal x from noisy observations $y = Ax + \varsigma$ with sensing matrix A and noise ς. This task is generally achieved by minimizing a regression objective of the form $\mathbf{f}(x) = \|Ax - y\|^2 + R(x)$ with regularization R. In the analysis of *Approximate Message Passing* algorithms, the sensing matrix A is assumed to be random.[13] This turns the regression objective $\mathbf{f}$ into a random function.

Finally, in **statistical physics**, random functions appear in the study of spin glasses,[14] where they are often referred to as *Hamiltonians* but more recently

[10] Krige, 1951; Matheron, 1963; Stein, 1999.

[11] Kriging utilizes the 'Best linear unbiased estimator' (BLUE), which coincides with the conditional expectation in the Gaussian case and is otherwise the minimum variance estimator in the space of unbiased linear estimators. The BLUE is effectively used if a Gaussian prior is assumed despite the underlying prior not being Gaussian.

[12] Kushner, 1964; Jones et al., 1998; P. I. Frazier, 2018; Garnett, 2023.

[13] Donoho et al., 2009; Donoho et al., 2010; Bayati and Montanari, 2011.

[14] e.g. Montanari, 2019; Subag, 2021; Huang and Sellke, 2022.

[15] Auffinger et al., 2023.

also as random functions.[15] The optima of these energy landscapes have been extensively studied because they correspond to stable physical states. This research on high-dimensional random functions has also lead to insights about the loss landscapes found in machine learning.[16]

[16] e.g. Dauphin et al., 2014; Choromanska et al., 2015.

Definition 2.2.1 (Random function)**.** A *Random function* is a collection of random variables $\mathbf{f} = (\mathbf{f}(x))_{x\in\mathbb{X}}$ indexed by the domain $\mathbb{X}$ of the random function $\mathbf{f}$. Since random variables are defined as functions from a probability space $(\Omega, \mathcal{A}, \mathbb{P})$ random functions technically have two inputs: Elementary events $\omega \in \Omega$ and the 'actual' input $x \in \mathbb{X}$. The notation

$$\mathbf{f}\colon \mathbb{X} \to \mathbb{R}$$

is therefore abuse of notation and stands for $\mathbf{f}\colon \Omega \times \mathbb{X} \to \mathbb{R}, (\omega, x) \mapsto \mathbf{f}_\omega(x)$. To emphasize the difference between random functions $\mathbf{f}$ and normal functions f we denote random functions in bold by convention.

Bayesian optimization[17] For the optimization of the random function $\mathbf{f}$ Baysian optimizers select evaluation locations X_{n+1} in $\mathbb{X}$ based on the previously seen evaluations $\mathcal{F}_n^X := \sigma(\mathbf{f}(X_0), \dots, \mathbf{f}(X_n))$. That is X_{n+1} is selected $\mathcal{F}_n^X$ measurably using the distribution of $\mathbf{f}(x)$ conditioned on $\mathcal{F}_n^X$. The following definition generalizes this setting to allow for the inclusion of derivatives of $\mathbf{f}$, accommodates for the possibility of noise on the evaluations and permits randomized algorithms (Section 2.2.1).

[17] e.g. Kushner, 1964; J. B. Mockus and L. J. Mockus, 1991; P. I. Frazier, 2018; Garnett, 2023.

Definition 2.2.2 (Feasible optimization algorithm)**.** Let $\mathbf{f} : \mathbb{X} \to \mathbb{R}$ be a continuous random objective function, i.e. $\mathbf{f}$ is a random element in $C(\mathbb{X}, \mathbb{Y})$. Let $\varsigma = (\varsigma_n)_{n\in\mathbb{N}_0}$ be a sequence of (location dependent) noise in $C(\mathbb{X}, \mathbb{Y})$, where ς_n is the noise on the n-th evaluation of $\mathbf{f}$. For some random element $\mathbf{M}$ such that $\mathbf{f}$ is measurable with respect to $\mathbf{M}$ (e.g. $\mathbf{M} = \mathbf{f}$), we assume the noise sequence to be iid conditional on $\mathbf{M}$ and each ς_n centered conditional on $\mathbf{M}$.[18]

If $\mathbf{f}$ and $(\varsigma_n)_{n\in\mathbb{N}_0}$ are in $C^k(\mathbb{X}, \mathbb{Y})$, the sequence of random evaluation locations $X = (X_n)_{n\in\mathbb{N}_0}$ in $\mathbb{X}$ is generated by a *feasible k-th order optimization algorithm*, if X_{n+1} is independent from $(\mathbf{f}, \varsigma)$ conditional on[19]

$$\mathcal{F}_n^X := \sigma(\mathcal{F}, Y_0, \dots, Y_n), \qquad n \in \{-1, 0, 1, \dots\} \tag{2.1}$$

where $\mathcal{F}$ is the initial information and $Y_n = \mathbf{Y}_n(X_n)$ is the n-th observation with

$$\mathbf{Y}_n(x) := \mathrm{J}^k\mathbf{f}(x) + \mathrm{J}^k\varsigma_n(x),$$

where $\mathrm{J}^k f(x)$ is the jet of differentials of f in x up to k-th order[20].

[18] This formulation is both very general and motivated by stochastic losses in the supervised machine learning setting. (Section 7.2)

[19] as introduced by (Kallenberg, 2002) we write $\xi \perp\!\!\!\perp_{\mathcal{G}} \eta$, if ξ is independent from η conditional on $\mathcal{G}$. If ξ is a random variable in a Polish space and $\sigma(W) = \mathcal{G}$, there exists a measurable function h and $U \sim \mathcal{U}(0,1)$ independent of W, η such that

$$\xi = h(W, U)$$

(Kallenberg, 2002, Prop. 6.13). Conditional independence therefore generalizes the case where $\xi = h(W)$, i.e. where ξ is measurable with respect to $\mathcal{G}$. In our case this allows for the randomized selection of X_{n+1} based on a distribution that is $\mathcal{F}_n^X$ measurable.

[20] e.g. for $k = 1$ the jet is of the form $\mathrm{J}^k\mathbf{f}(x) = (\mathbf{f}(x), \nabla\mathbf{f}(x))$

Bayesian optimization algorithms select the next evaluation point based on this conditional distribution with the help of *acquisition functions*. In the following we present prominent examples with maximization as the objective.

Example 2.2.3 (Acquisition functions)**.**

- *Probability of improvement* (PI) chooses

$$X_{n+1} = \arg\max_{x\in\mathbb{X}} \mathbb{P}\Big(\mathbf{f}(x) > \tau_n \mid \mathcal{F}_n^X\Big)$$

for some τ_n adapted to the filtration $\mathcal{F}_n^X$. In the noiseless case where the values $(\mathbf{f}(X_k))_{k<n}$ are contained in $\mathcal{F}_n^X$ the canonical selection is essentially the running maximum

$$\tau_n := \max_{k=0,\dots,n-1} \mathbf{f}(X_k) + \epsilon \tag{2.2}$$

where $\epsilon \geq 0$ is a minimal improvement.

- *Expected improvement* (EI) chooses

$$X_{n+1} = \arg\max_{x\in\mathbb{X}} \mathbb{E}\Big[\big(\mathbf{f}(x) - \tau_n\big)_+ \mid \mathcal{F}_n^X\Big]$$

with $(a)_+ = \max\{a, 0\}$ and τ_n adapted to $\mathcal{F}_n^X$. The canonical choice for τ_n in the noiseless setting is again given by (2.2).

- *Upper confidence bound* (UCB)[21] maximization selects

$$X_{n+1} = \arg\max_{x\in\mathbb{X}} \mu_n(x) + \sqrt{\beta_n}\sigma_n(x)$$

where $\mu_n(x) = \mathbb{E}[\mathbf{f}(x) \mid \mathcal{F}_n^X]$ represents the conditional expectation, $\sigma_n(x) = \mathrm{Var}(\mathbf{f}(x) \mid \mathcal{F}_n^X)$ the conditional variance and β_n is a parameter that weighs the value of exploration at time step n.

[21] The name of this acquisition function clashes with the presentation of Bayesian optimization with minimization as the objective.

With $\epsilon = 0$ in (2.2) the PI acquisition function highly values safe, infinitesimal improvements. To encourage exploration, non-zero ϵ have to be selected in practice to get reasonable convergence. However EI can work 'hyperparameter free'[22] with $\epsilon = 0$ as it inherently values larger improvements.

[22] ignoring the selection of the distributional model.

Remark 2.2.4 (Convergence proof). For the zero order case (without gradients) there are convergence proofs for the UCB algorithm[23] and for the EI algorithm.[24] Even if the prior distribution of $\mathbf{f}$ is not known and has to be estimated some convergence results are available.[25]

[23] Srinivas et al., 2010.

[24] Vazquez and Bect, 2010.

[25] Berkenkamp et al., 2019.

2.2.1 Randomized Bayesian optimizer

Before we take a closer look at Bayesian optimization algorithms let us quickly note that it is also possible to combine the assumption about the random objective function with randomized algorithms.

Thomson sampling[26] randomly samples the subsequent evaluation point X_{n+1} according to the posterior distribution of the arg max of the random objective $\mathbf{f}$

[26] Thompson, 1933.

$$X_{n+1} \sim \mathbb{P}\big(\arg\max_x \mathbf{f}(x) \in \cdot \mid \mathcal{F}_n^X\big)$$

However, due to the complex nature of this posterior distribution,[27] Thomson sampling is considered impractical for implementation.[28]

[27] except for special priors on the objective or small domains $\mathbb{X}$

[28] Garnett, 2023, Sec. 8.7.

2.2.2 Optimal Bayesian algorithms

Intuitively, the best performance an optimizer can achieve with zero further evaluations and available information $\mathcal{F}$ is given by the value

$$V_0\{\mathcal{F}\} := \max_{x\in\mathbb{X}} \mathbb{E}[\mathbf{f}(x) \mid \mathcal{F}].$$

It is achieved with the greedy choice of

$$Q_0\{\mathcal{F}\} := \arg\max_{x\in\mathbb{X}} \mathbb{E}[\mathbf{f}(x) \mid \mathcal{F}].$$

To consider the case of n evaluations we want to proceed by backwards induction. After a single evaluation, one has $n-1$ evaluations left and those evaluations are affected by the shifted noise sequence $S(\varsigma) = (\varsigma_{n+1})_{n\in\mathbb{N}_0}$.

Thus, if $V_{n-1}^{S(\varsigma)}\{\mathcal{F}\}$ is the best performance that is achievable with information $\mathcal{F}$ and $n-1$ further evaluations, then with n evaluations available, the best one can do with information $\mathcal{F}$ is

$$V_n^\varsigma\{\mathcal{F}\} := \max_{x\in\mathbb{X}} \mathbb{E}\big[V_{n-1}^{S(\varsigma)}\{\mathcal{F}, \mathbf{Y}_0(x)\} \mid \mathcal{F}\big] \qquad (n\text{-step lookahead value})$$

$$Q_n^\varsigma\{\mathcal{F}\} := \arg\max_{x\in\mathbb{X}} \mathbb{E}\big[V_{n-1}^{S(\varsigma)}\{\mathcal{F}, \mathbf{Y}_0(x)\} \mid \mathcal{F}\big]. \qquad (n\text{-step lookahead choice})$$

Clearly, with $V_0^\varsigma := V_0$ and $Q_0^\varsigma := Q_0$ this results in well defined V_n^ς and Q_n^ς.

Remark 2.2.5 (Knowledge Gradient). The one step lookahead, which selects

$$X_n := Q_1^{S^n(\varsigma)}\{\mathcal{F}_{n-1}^X\} = \arg\max_{x\in\mathbb{X}} \mathbb{E}\Big[\max_y \mathbb{E}\big[\mathbf{f}(y) \mid \mathcal{F}_{n-1}^X, \mathbf{Y}_n(x)\big] \,\Big|\, \mathcal{F}_{n-1}^X\Big]$$

is known as the *knowledge gradient*[29] (KG) acquisition function in Bayesian optimization.

[29] see e.g. P. Frazier et al. (2009), therein a convergence proof can also be found.

The following theorem shows that backwards induction using the Q functions results in the algorithm with optimal average case performance after N evaluations.

Theorem 2.2.6 (N-evaluation optimal algorithm)**.** *In the setting of feasible optimization algorithms (Definition 2.2.2), the algorithm that chooses*

$$X_n := Q_{N-n}^{S^n(\varsigma)}\{\mathcal{F}_{n-1}^X\} \qquad \text{for} \quad n \in \{0,\dots,N\}, \tag{2.3}$$

and $X_n = X_N$ for $n > N$, is a feasible k-th order optimization algorithm that achieves the best average performance in the set of all feasible algorithms after N evaluations of k-th order information. That is, if $\tilde{X}_0,\dots,\tilde{X}_N$ is generated by a feasible k-th order algorithm, then almost surely

$$\mathbb{E}[\mathbf{f}(X_N) \mid \mathcal{F}] = V_N^\varsigma\{\mathcal{F}\} \geq \mathbb{E}[\mathbf{f}(\tilde{X}_N) \mid \mathcal{F}],$$

where we recall that $\mathcal{F} = \mathcal{F}_{-1}$ is the initial information.

Remark 2.2.7 (Randomization does not provide an advantage). Observe that X_n is selected as a function of $\mathcal{F}_{n-1}^X$ and randomization as in Thomson sampling 2.2.1 is unnecessary to obtain optimal performance.

Proof. **Step 1:** More than the first equality, we will prove for all $n \in \{0,\dots,N\}$

$$\mathbb{E}[\mathbf{f}(X_N) \mid \mathcal{F}] = \mathbb{E}[V_{N-n}^{S^n(\varsigma)}\{\mathcal{F}_{n-1}^X\} \mid \mathcal{F}]$$

by backwards induction. Since $V_N^\varsigma\{\mathcal{F}_{-1}^X\}$ is measurable with respect to $\mathcal{F}_{-1}^X = \mathcal{F}$ by definition, the case $n = 0$ implies the first part of the claim

$$\mathbb{E}[\mathbf{f}(X_N) \mid \mathcal{F}] = \mathbb{E}[V_N^{S^0(\varsigma)}\{\mathcal{F}_{-1}^X\} \mid \mathcal{F}] = V_N^\varsigma(\mathcal{F}).$$

The induction start with $n = N$ is given by

$$\begin{aligned}
\mathbb{E}\big[V_0^{S^N(\varsigma)}\{\mathcal{F}_{N-1}^X\} \mid \mathcal{F}\big] &= \mathbb{E}\big[\max_{x\in\mathbb{X}} \mathbb{E}[\mathbf{f}(x) \mid \mathcal{F}_{N-1}^X] \mid \mathcal{F}\big] \\
&= \mathbb{E}\Big[\mathbb{E}\big[\mathbf{f}(Q_0\{\mathcal{F}_{N-1}^X\}) \mid \mathcal{F}_{N-1}^X\big] \mid \mathcal{F}\Big] \\
&= \mathbb{E}\big[\mathbb{E}[\mathbf{f}(X_N) \mid \mathcal{F}_{N-1}^X] \mid \mathcal{F}\big] \\
&\overset{\text{tower}}{=} \mathbb{E}[\mathbf{f}(X_n) \mid \mathcal{F}].
\end{aligned}$$

For the induction step $(n+1) \to n$, recall that by definition of X_n and Q_n^ς we have

$$\begin{aligned}
X_n &= Q_{N-n}^{S^n(\varsigma)}\{\mathcal{F}_{n-1}^X\} \\
&= \arg\max_{x\in\mathbb{X}} \mathbb{E}\big[V_{N-n-1}^{S^{n+1}(\varsigma)}\{\mathcal{F}_{n-1}^X, Y_0(x)\} \mid \mathcal{F}_{n-1}^X\big].
\end{aligned}$$

And thus by definition of V_n^ς and $\mathcal{F}_n^X$ and Example 3.1.4

$$\begin{aligned}
V_{N-n}^{S^n(\varsigma)}\{\mathcal{F}_{n-1}^X\} &= \mathbb{E}\big[V_{N-n-1}^{S^{n+1}(\varsigma)}\{\mathcal{F}_{n-1}^X, Y_0(X_n)\} \mid \mathcal{F}_{n-1}^X\big]. \\
&= \mathbb{E}\big[V_{N-(n+1)}^{S^{n+1}(\varsigma)}\{\mathcal{F}_{(n+1)-1}^X\} \mid \mathcal{F}_{n-1}^X\big].
\end{aligned}$$

Since we have the claim for $n+1$ we thus obtain for n by induction

$$\mathbb{E}[V_{N-n}^{S^n(\varsigma)}\{\mathcal{F}_{n-1}^X\} \mid \mathcal{F}] \overset{\text{tower}}{=} \mathbb{E}\big[V_{N-(n+1)}^{S^{n+1}(\varsigma)}\{\mathcal{F}_{(n+1)-1}^X\} \mid \mathcal{F}\big] \overset{\text{ind.}}{=} \mathbb{E}[\mathbf{f}(X_n) \mid \mathcal{F}].$$

Step 2: To avoid notational clutter we drop the tilde and show for any initial information $\mathcal{F}$ and any $X = (X_n)_{n\in\mathbb{N}_0}$ generated by a feasible k-th order optimization algorithm by induction over n

$$\mathbb{E}[\mathbf{f}(X_n) \mid \mathcal{F}] \leq V_n^{S^m(\varsigma)}\{\mathcal{F}\}.$$

for all $m \in \mathbb{N}_0$. The selection $n = N$ and $m = 0$ then yields the claim.

Observe that X_0 is independent from $\mathbf{f}, \varsigma$ conditional on $\mathcal{F}$ and therefore there exists $U \sim \mathcal{U}(0,1)$ independent from $(\mathbf{f}, \varsigma, \mathcal{F})$ such that

$$X_0 = h(W, U) \tag{2.4}$$

for some random element W that generates $\mathcal{F}$[30] and a measurable function h. For the induction start with $n = 0$ this implies for all $m \in \mathbb{N}_0$

[30] the identity map from the underlying probability space $(\Omega, \mathcal{A})$ into $(\Omega, \mathcal{F})$ does the job.

$$\begin{aligned}
\mathbb{E}[\mathbf{f}(X_0) \mid \mathcal{F}] &\overset{\text{tower}}{=} \mathbb{E}\big[\mathbb{E}[\mathbf{f}(X_0) \mid \mathbf{f}, W] \mid \mathcal{F}\big] \\
&\overset{(2.4)}{=} \mathbb{E}\Big[\int_0^1 \mathbf{f}(h(W,u))du \mid \mathcal{F}\Big] \\
&= \int_0^1 \underbrace{\mathbb{E}\big[\mathbf{f}(h(W,u)) \mid \mathcal{F}\big]}_{\leq \max_x \mathbb{E}[\mathbf{f}(x)\mid\mathcal{F}]} du \leq V_0\{\mathcal{F}\} = V_0^{S^m(\varsigma)}\{\mathcal{F}\},
\end{aligned}$$

where we use that $h(W,u)$ is an $\mathcal{F}$ measurable random variable and Example 3.1.4.

Let us now consider the induction step $n \to n+1$. Since the shifted process $S(X) := (X_{n+1})_{n\in\mathbb{N}_0}$ retains the conditionally independence property with respect to the shifted filtration $S(\mathcal{F}^X) := (\mathcal{F}^X_{n+1})_{n\in\{-1,0,\dots\}}$ we can apply the induction assumption for n and $\mathcal{F} = \mathcal{F}_0^X$ to obtain for all $m \in \mathbb{N}_0$

$$\begin{aligned}
\mathbb{E}[\mathbf{f}(X_{n+1}) \mid \mathcal{F}_0^X] = \mathbb{E}\big[\mathbf{f}(S(X)_n) \mid S(\mathcal{F}^X)_{-1}\big] \\
&\overset{\text{ind.}}{\leq} V_n^{S^{m+1}(\varsigma)}\{S(\mathcal{F}^X)_{-1}\} \\
&= V_n^{S^{m+1}(\varsigma)}\{\mathcal{F}_0^X\}
\end{aligned}$$

We therefore have

$$\begin{aligned}
\mathbb{E}[\mathbf{f}(X_{n+1}) \mid \mathcal{F}] &\overset{\text{tower}}{=} \mathbb{E}\big[\mathbb{E}[\mathbf{f}(X_{n+1}) \mid \mathcal{F}_0^X] \mid \mathcal{F}\big] \\
&\overset{\text{monot.}}{\leq} \mathbb{E}\big[V_n^{S^{m+1}(\varsigma)}\{\mathcal{F}_0^X\} \mid \mathcal{F}\big] \\
&= \mathbb{E}\big[V_n^{S^{m+1}(\varsigma)}\{\mathcal{F}, \mathbf{Y}_0(X_0)\} \mid \mathcal{F}\big] \\
&\overset{\text{tower}}{=} \mathbb{E}\Big[\mathbb{E}\big[V_n^{S^{m+1}(\varsigma)}\{\mathcal{F}, \mathbf{Y}_0(X_0)\} \mid \mathbf{f}, \varsigma, W\big] \mid \mathcal{F}\Big] \\
&\overset{(2.4)}{=} \int_0^1 \mathbb{E}\Big[V_n^{S^{m+1}(\varsigma)}\{\mathcal{F}, \mathbf{Y}_0(h(W,u))\} \mid \mathcal{F}\Big] du \\
&\overset{(*)}{\leq} \max_x \mathbb{E}[V_n^{S^{m+1}(\varsigma)}\{\mathcal{F}, \mathbf{Y}_0(x)\} \mid \mathcal{F}] \\
&= V_{n+1}^{S^m(\varsigma)}(\mathcal{F}).
\end{aligned}$$

For $(*)$ we use that $h(W,u)$ is $\mathcal{F}$ measurable and Example 3.1.4. □

Remark 2.2.8 (Utility). One may assign a utility u to reaching a certain function value $\mathbf{f}$. But in this case maximizing $\mathbf{f}(X_n)$ turns into maximizing $u(\mathbf{f}(X_n))$ and one can just consider $\tilde{\mathbf{f}} = u \circ \mathbf{f}$ instead.

Remark 2.2.9 (Regret). The performance of the intermediate optimization steps $\mathbf{f}(X_n)$ with $n \leq N$ is completely disregarded for the optimality of Theorem 2.2.6. Only the final performance matters. This allows algorithms to prioritize information gain and therefore exploration up until timestep N. One could instead consider a weighted average of $\mathbf{f}(X_n)$ as a performance target. To avoid the requirement that

the weights $\gamma = (\gamma_n)_{n\in\mathbb{N}_0}$ have to be be summable one can instead consider the weighted regret

$$R_\gamma := \sum_{n=0}^{\infty} \gamma_n \big(\max_{y\in\mathbb{X}} \mathbf{f}(y) - \mathbf{f}(X_n)\big)$$

which is equivalent but allows $\gamma_n = 1$ for all n as long as $\mathbf{f}(X_n)$ converges quickly enough against the global maximum.[31]

The iteration we considered translates to

$$V_\gamma^\varsigma\{\mathcal{F}\} := \max_{x\in\mathbb{X}} \mathbb{E}\Big[\underbrace{\gamma_0\big(\mathbf{f}(x) - \max_{y\in\mathbb{X}} \mathbf{f}(y)\big)}_{\substack{\text{immeditate}\\ \text{value}}} + \underbrace{V_{S(\gamma)}^{S(\varsigma)}\{\mathcal{F}, \mathbf{Y}_0(x)\}}_{\text{future value}} \mid \mathcal{F}\Big]$$

using the shift operator $S(\gamma) := (\gamma_{n+1})_{n\in\mathbb{N}_0}$. If the sequence of γ is eventually zero, i.e. $S^k(\gamma) = \mathbf{0}$ for some k, then termination with $V_{\mathbf{0}}\{\mathcal{F}\} := 0$ ensures V_γ is well defined.[32] In the general case one has to define the value function V_γ^ς as the attained expected negative regret of the algorithm that maximizes the expected negative regret and then prove this recursion as a theorem. This is similar to value functions in reinforcement learning.

[31] Theorem 2.2.6 treats optimality for the case $\gamma_n = \mathbf{1}_{N=n}$ where the stated algorithm achieves minimal expected regret.

[32] For $\gamma_n = \mathbf{1}_{N=n}$ it follows $V_\gamma\{\mathcal{F}\} = V_N\{\mathcal{F}\} - \mathbb{E}\big[\max_{y\in\mathbb{X}} \mathbf{f}(y) \mid \mathcal{F}\big]$.

2.2.3 Practical Bayesian optimization

Section 2.2.2 demonstrates that optimal average performance algorithms are relatively easy to obtain in an abstract form. The problem is that abstract maximizers of acquisition functions are not necessarily computable. To make the acquisition functions calculable it is typically assumed that $\mathbf{f}$ is Gaussian.

More generally, assume $\mathbf{f}$ and the random variables generating the filtration $\mathcal{F}_n^X$ to be jointly Gaussian.[33] The posterior distribution of $\mathbf{f}(x)$ given $\mathcal{F}_n^X$ is then Gaussian by Theorem 2.A.1, i.e.

$$\mathbf{f}(x) \mid \mathcal{F}_n^X \sim \mathcal{N}(\mu_n(x), \sigma_n^2(x)),$$

where we denote by $\mu_n(x) := \mathbb{E}[\mathbf{f}(x) \mid \mathcal{F}_n^X]$ the conditional mean and by $\sigma_n^2(x) := \mathrm{Var}(\mathbf{f}(x) \mid \mathcal{F}_n^X)$ the conditional variance.

Recall the UCB acquisition function in Example 2.2.3 was already defined using only this conditional mean and variance. In the following theorem we present results that show that PI and EI can also be expressed using $\mu_n(x)$ and $\sigma_n^2(x)$ assuming a Gaussian distribution.

[33] Treating the, typically random, evaluation points X_n as deterministic (cf. Chapter 3, in particular Corollary 3.1.8)

Theorem 2.2.10 (More explicit acquisition functions)**.** *Let $\mathbf{f}$ and be jointly Gaussian. Then*

1. *Probability of improvement (PI)*

$$\arg\max_{x\in\mathbb{X}} \mathbb{P}\big(\mathbf{f}(x) > \tau_n \mid \mathcal{F}_n^X\big) = \arg\max_{x\in\mathbb{X}} \underbrace{\frac{\mu_n(x) - \tau_n}{\sigma_n(x)}}_{=:\pi_n(x)}$$

   with τ_n adapted to the filtration $\mathcal{F}_n^X$.

2. *Expected improvement (EI)*

$$\mathbb{E}\Big[\big(\mathbf{f}(x) - \tau_n\big)_+ \mid \mathcal{F}_n^X\Big] = \sigma_n(x)\Big(\varphi\big(\pi_n(x)\big) + \pi_n(x)\Phi\big(\pi_n(x)\big)\Big),$$

   where τ_n and π_n are as for PI, and φ is the density and Φ the cumulative distribution function of the standard normal distribution.

Proof (e.g. Garnett, 2023, chap. 8). Since the posterior distribution of $\mathbf{f}(x)$ given $\mathcal{F}_n^X$ is Gaussian with mean μ_n and variance σ_n^2 we have that

$$Z_n(x) := \frac{\mathbf{f}(x) - \mu_n(x)}{\sigma_n(x)}$$

is standard normal conditional on $\mathcal{F}_n^X$. Reordering the inequality in the PI acquisition function therefore implies it is equal to

$$\mathbb{P}\big(\mathbf{f}(x) > \tau_n \mid \mathcal{F}_n^X\big) = \Phi\Big(\frac{\mu_n(x) - \tau_n}{\sigma_n(x)}\Big) = \Phi(\pi_n(x)).$$

Since the cumulative distribution function Φ of the standard normal distribution is monotonously increasing, maximization of this term is equivalent to maximizing its input $\pi_n(x)$.

Similarly transforming $\mathbf{f}(x)$ and τ_n in the EI acquisition we obtain

$$\mathbb{E}[(\mathbf{f}(x) - \tau_n)_+ \mid \mathcal{F}_n^X] = \sigma_n(x)\mathbb{E}\big[(Z_n(x) - \pi_n(x))_+ \mid \mathcal{F}_n^X\big]$$

since $\sigma_n(x)$ is non-negative and $\mathcal{F}_n^X$ measurable. Since $\pi_n(x)$ is $\mathcal{F}_n^X$ measurable and $Z_n(x)$ standard normal conditional on $\mathcal{F}_n^X$ it is thus sufficient to calculate for $Y \sim \mathcal{N}(0,1)$ and $x \in \mathbb{R}$

$$\begin{aligned}\mathbb{E}[(Y+x)_+] &= \frac{1}{\sqrt{2\pi}}\int_{-x}^{\infty}(y+x)\exp(-\tfrac{y^2}{2})dy \\ &= \int_{-x}^{\infty}\frac{d}{dy}\Big(-\tfrac{1}{\sqrt{2\pi}}\exp(-\tfrac{y^2}{2})\Big)dy + x(1-\Phi(-x)) \\ &= \varphi(x) + x\Phi(x). \qquad \square\end{aligned}$$

While the knowledge gradient acquisition function (Remark 2.2.5) may be evaluated in $\mathcal{O}(n\log(n))$ time[34] for domains $\mathbb{X}$ of size $n = \#\mathbb{X}$ it becomes intractable for infinite domains in all but a few special cases.[35]

Observe that the acquisition functions of Theorem 2.2.10 contain an arg max that may be difficult to obtain depending on the nature of the posteriors μ_n and σ_n. Roughly speaking: If the posterior mean μ_n is as difficult to optimize as the function $\mathbf{f}$ then one has not gained anything with this approach. This is the motivation for the 'Random Function Descent' acquisition function (Chapter 5).

[34] P. Frazier et al., 2009.

[35] Garnett, 2023, Sec. 8.3.

Higher evaluation costs justify more sophisticated acquisition functions If the evaluation of $\mathbf{f}$ is significantly more expensive than evaluations of μ_n and σ_n, then the conversion of the problem from $\arg\max_x \mathbf{f}(x)$ to the surrogate maximization problem $\arg\max_x F_n(\mu_n(x), \sigma_n(x))$, for some function F_n induced by the acquisition function, can be worth it. An example is 'hyperparameter tuning' in machine learning, where the evaluation of one 'hyperparameter' is the final error after training the entire model using this hyperparameter. Compared to the training of a machine learning model the evaluation of most posteriors is very cheap.

2.A Conditional Gaussian distribution

In Section 2.2.3 we claimed that the posterior distribution of $\mathbf{f}(x)$ given $\mathcal{F}_n^X$ is Gaussian. For this we assumed that the noisy evaluations $Y_0, \dots, Y_n$ generating the filtration[36] $\mathcal{F}_n^X$ are joint Gaussian together with $\mathbf{f}$, treating the evaluation locations as deterministic.[37] This implies $(\mathbf{f}(x), Y_0, \dots, Y_n)$ is a Gaussian vector for any x and to calculate the posterior distribution of $\mathbf{f}(x)$ given $Y_0, \dots, Y_n$ we need to be able to obtain the posterior distribution of multivariate Gaussian random vectors.

Theorem 2.A.1 presents the well known explicit conditional distribution of multivariate Gaussian random vectors.[38]

[36] Here we omit the initial information $\mathcal{F}$ which may be added as another Gaussian random variable Y_{-1} such that all variables are joint Gaussian.

[37] see Chapter 3 especially Example 3.1.7 and Corollary 3.1.8

[38] e.g. Eaton, 2007, Prop. 3.13.

Theorem 2.A.1 (Conditional Gaussian distribution)**.** *Let $X \sim \mathcal{N}(\mu, \Sigma)$ be a multivariate Gaussian vector where the mean is a block vector and the covariance matrix a block matrix of the form*

$$\mu = \begin{bmatrix}\mu_1 \\ \mu_2\end{bmatrix} \quad \text{and} \quad \Sigma = \begin{bmatrix}\Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22}\end{bmatrix},$$

then the conditional distribution of X_2 given X_1 is

$$X_2 \mid X_1 \sim \mathcal{N}(\mu_{2|1}, \Sigma_{2|1}),$$

with conditional mean and variance

$$\begin{aligned}\mu_{2|1} &:= \mu_2 + \Sigma_{21}\Sigma_{11}^{-1}(X_1 - \mu_1)\\ \Sigma_{2|1} &:= \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}.\end{aligned}$$

where Σ_{11}^{-1} may be a generalized inverse.[39]

[39] cf. Eaton, 2007.

Proof. For the general statement we point to the standard literature.[40] For this proof we will assume for simplicity that Σ is invertible.

[40] e.g. Eaton, 2007.

Let $\bar{X} := X - \mu$ be the centered version of X. Then there exists a unique lower triangular matrix L such that $\Sigma = LL^T$ (i.e. the Cholesky Decomposition). This results in the following representation[41]

$$X - \mu =: \begin{bmatrix}\bar{X}_1\\ \bar{X}_2\end{bmatrix} = \begin{bmatrix}L_{11} & 0\\ L_{21} & L_{22}\end{bmatrix}\begin{bmatrix}Y_1\\ Y_2\end{bmatrix} = LY$$

with independent standard normal Y_i, i.e. $Y \sim \mathcal{N}(0, \mathbb{I})$. L_{11} is invertible since Σ_{11} and therefore the map from Y_1 to X_1 is bijective. Conditioning on X_1 is therefore equivalent to conditioning on Y_1. But we have

$$X_2 = \mu_2 + \bar{X}_2 = \underbrace{\mu_2 + L_{21}Y_1}_{\text{conditional expectation}} + \underbrace{L_{22}Y_2}_{\text{conditional distribution}} \tag{2.5}$$

[41] It is straightforward to check that $Y := L^{-1}(X - \mu)$ is a multivariate Gaussian vector of centered, uncorrelated entries with variance 1. They are therefore iid standard normal since uncorrelated multivariate Gaussian random variables are independent. Sometimes this representation is even taken as the definition of a multivariate normal distribution. This is the only real place we use the invertibility of Σ in its entirety, the invertibility of Σ_{11}^{-1} is used later on because we do not want to get into the business of generalized inverses here.

So it follows that

$$X_2 \mid X_1 \sim \mathcal{N}(\mu_{2|1}, \Sigma_{2|1})$$

with

$$\begin{aligned}\mu_{2|1} &:= \mathbb{E}[X_2 \mid X_1] &&= \mu_2 + L_{21}Y_1\\ \Sigma_{2|1} &:= \operatorname{Cov}[X_2 \mid X_1] = \mathbb{E}\Big[\big(X_2 - \mathbb{E}[X_2 \mid X_1]\big)^2 \mid X_1\Big] &&= L_{22}L_{22}^T.\end{aligned}$$

What is left to do, is to find a representation for the L_{ij} using the block matrices of Σ. For this we observe

$$\begin{bmatrix}\Sigma_{11} & \Sigma_{12}\\ \Sigma_{21} & \Sigma_{22}\end{bmatrix} = \Sigma = LL^T = \begin{bmatrix}L_{11}L_{11}^T & L_{11}L_{21}^T\\ L_{21}L_{11}^T & L_{22}L_{22}^T + L_{21}L_{21}^T\end{bmatrix}. \tag{2.6}$$

Using $Y_1 = L_{11}^{-1}\bar{X}_1$ and the insertion of an identity matrix this implies

$$L_{21}Y_1 = L_{21}(L_{11}^T L_{11}^{-T})(L_{11}^{-1}\bar{X}_1) \overset{(2.6)}{=} \Sigma_{21}\Sigma_{11}^{-1}(X_1 - \mu_1).$$

resulting in the desired conditional expectation $\mu_{2|1}$. The conditional variance follows alike

$$\begin{aligned}\Sigma_{2|1} = L_{22}L_{22}^T &\overset{(2.6)}{=} \Sigma_{22} - L_{21}L_{21}^T\\ &= \Sigma_{22} - \underbrace{L_{21}(L_{11}^T}_{\overset{(2.6)}{=}\Sigma_{21}}\underbrace{L_{11}^{-T})(L_{11}^{-1}}_{\overset{(2.6)}{=}\Sigma_{11}^{-1}}\underbrace{L_{11})L_{21}^T}_{\overset{(2.6)}{=}\Sigma_{12}}.\end{aligned}$$

□

Remark 2.A.2 (Decomposition). $X_2 \mid X_1 \sim \mathcal{N}(\mu_{2|1}, \Sigma_{2|1})$ always implies that there exists there exists $Y_2 \sim \mathcal{N}(0, \mathbb{I})$ independent of X_1 such that the following equality is true in distribution

$$X_2 = \mu_{2|1} + \sqrt{\Sigma_{2|1}}Y_2, \tag{2.7}$$

where $\sqrt{\Sigma}$ denotes the cholesky decomposition of Σ. If the covariance matrix is moreover invertible, then Y_2 can be determined constructively as in the proof of Theorem 2.A.1 such that equation holds always and not just in distribution (cf. (2.5)). This might be true in the non-invertible case as well but would require deeper insights about singular matrices.

2.B Sample regularity of random functions

In the previous section we made the assumption that $(\mathbf{f}, Y_0, \dots, Y_n)$ was joint Gaussian, treating the evaluation locations as deterministic.[42] Since the observations are of the form

$$Y_i = \mathbf{Y}_i(X_i) = \mathrm{J}^k\mathbf{f}(X_i) + \mathrm{J}^k\varsigma(X_i),$$

we need $\mathrm{J}^k\mathbf{f}$ and $\mathrm{J}^k\varsigma$ to be joint Gaussian random functions. For the case $k = 0$ it is sufficient if $\mathbf{f}$ and ς are joint Gaussian random functions. But for $k > 0$ the question is what requirements $\mathbf{f}$ and ς need to satisfy such that $\mathrm{J}^k\mathbf{f}$ and $\mathrm{J}^k\varsigma$ are joint Gaussian. But the question is not really whether or not it is joint Gaussian, since derivatives are Gaussian as limits of Gaussian random variables

$$D_v\mathbf{f}(x) = \lim_{t\to 0} \tfrac{1}{t}(\mathbf{f}(x + tv) - \mathbf{f}(x)). \tag{2.8}$$

The real question is, do derivatives exist? Specifically, since Gaussian random functions are understood in terms of mean and covariance – what requirements do mean and covariance function need to satisfy to ensure the samples $\mathbf{f}$ are regular enough for $\mathbf{f}$ to be k times continuously differentiable.

Clearly, if the mean is in $C^k(\mathbb{X}, \mathbb{Y})$, then we may add it to a centred Gaussian random function in $C^k(\mathbb{X}, \mathbb{Y})$ to obtain a Gaussian random function in $C^k(\mathbb{X}, \mathbb{Y})$ with the desired mean and covariance. We can thus assume centered Gaussian random functions without loss of generality.

Gaussian processes, like most infinite dimensional random objects, are typically constructed as a collection of random variables $\mathbf{f} = (\mathbf{f}(x))_{x\in\mathbb{X}}$ using Kolmogorov's extension theorem.[43] For such a collection of random variables $\mathbf{f}$ one first needs to prove there exists a version[44] that lies in $C^k(\mathbb{X}, \mathbb{Y})$.

The requirements on the covariance kernel for sample path regularity are well known.[45] To understand these requirements consider a centered Gaussian random function $\mathbf{f}$. If the derivative $\partial_{x_i}\mathbf{f}(x)$ exists, then by (2.8) and and an argument why the expectation and limit may be exchanged[46]

$$\mathbb{E}[\partial_{x_i}\mathbf{f}(x)] = 0$$

and

$$\mathrm{Cov}(\partial_{x_i}\mathbf{f}(x), \partial_{y_j}\mathbf{f}(y)) = \mathbb{E}[\partial_{x_i}\mathbf{f}(x)\partial_{y_j}\mathbf{f}(y)] = \partial_{x_i}\partial_{y_j}\mathcal{C}_{\mathbf{f}}(x, y). \tag{2.9}$$

So clearly, for $\partial_{x_i}\mathbf{f}(x)$ to exist, $\partial_{x_i}\partial_{y_j}\mathcal{C}_{\mathbf{f}}(x, y)$ must exist. And in fact the existence of $\partial_{x_i}\partial_{y_i}\mathcal{C}_{\mathbf{f}}(x, y)$ is equivalent to the existence of $\partial_{x_i}\mathbf{f}(x)$ as an L^2 limit of (2.8).[47] In this case derivative and expectation may be freely exchanged, i.e. we also have for the mixed derivatives

$$\mathrm{Cov}(\partial_{x_i}\mathbf{f}(x), \mathbf{f}(y)) = \partial_{x_i}\mathcal{C}_{\mathbf{f}}(x, y). \tag{2.10}$$

However for $\partial_{x_i}\mathbf{f}(x)$ to exist as a continuous random function, slightly more regularity on the covariance function is necessary. It turns out that Hölder continuity of $\partial_{x_i}\partial_{y_i}\mathcal{C}_{\mathbf{f}}(x, y)$ in both entries leads to Hölder continuity of the sample function $\partial_{x_i}\mathbf{f}(x)$.[48]

In summary, if $\mathbf{f}$ is a Gaussian random function in $C(\mathbb{X}, \mathbb{Y})$, then $\mathrm{J}^k\mathbf{f}$ is a Gaussian random function and the covariance function can be calculated using the rules in (2.10) and (2.9).

[42] see Chapter 3.

[43] e.g. Klenke, 2014, Thm. 14.36.

[44] if $\mathbf{f}$ is in $C^k(\mathbb{X}, \mathbb{Y})$ up to a probability zero set N, it can be set to zero on N resulting in the version $\tilde{\mathbf{f}}_\omega(x) = \mathbf{f}_\omega(x)\mathbf{1}_{N^\complement}(\omega)$, which is in $C^k(\mathbb{X}, \mathbb{Y})$ for all elementary events $\omega \in \Omega$ such that $\tilde{\mathbf{f}}$ becomes a random variable in $C^k(\mathbb{X}, \mathbb{Y})$, assuming the Borel σ-algebra on $C^k(\mathbb{X}, \mathbb{Y})$ is compatible with the product σ-algebra, see Remark 3.4.3

[45] see e.g. Costa et al. (2024), Scheurer (2009, chap. 5), Gihman and Skorokhod (1974, Ch. IV, §3, Def. 3 and following), Talagrand (1987) and references therein.

[46] Scheurer, 2009, Sec. 5.3.2.

[47] Scheurer, 2009, Thm. 5.3.10.

[48] Costa et al., 2024, Thm. 7.

References

Auffinger, Antonio, Andrea Montanari, and Eliran Subag (2023). "Optimization of Random High-Dimensional Functions: Structure and Algorithms". In: *Spin Glass Theory and Far Beyond*. 5 Toh Tuck Link, Singapore: WORLD SCIENTIFIC, pp. 609–633. ISBN: 9789811273919. DOI: 10.1142/9789811273926_0029. arXiv: 2206.10217 [math.PR].

Bayati, Mohsen and Andrea Montanari (2011). "The Dynamics of Message Passing on Dense Graphs, with Applications to Compressed Sensing". In: *IEEE Transactions on Information Theory* 57.2, pp. 764–785. ISSN: 1557-9654. DOI: 10.1109/TIT.2010.2094817.

Berkenkamp, Felix, Angela P. Schoellig, and Andreas Krause (2019). "No-Regret Bayesian Optimization with Unknown Hyperparameters". In: *Journal of Machine Learning Research* 20.50, pp. 1–24. ISSN: 1533-7928. URL: http://jmlr.org/papers/v20/18-213.html (visited on 2024-08-15).

Borgwardt, Karl Heinz (1986). *The Simplex Method: A Probabilistic Analysis*. Softcover reprint of the original 1st ed. 1987 edition. Berlin Heidelberg: Springer. 282 pp. ISBN: 978-3-540-17096-9.

Bubeck, Sébastien, Yin Tat Lee, and Mohit Singh (2015). "A Geometric Alternative to Nesterov's Accelerated Gradient Descent". arXiv: 1506.08187 [cs, math]. URL: http://arxiv.org/abs/1506.08187 (visited on 2021-07-07).

Chen, August Y., Ayush Sekhari, and Karthik Sridharan (2024). *Langevin Dynamics: A Unified Perspective on Optimization via Lyapunov Potentials*. DOI: 10.48550/arXiv.2407.04264. arXiv: 2407.04264 [cs]. Pre-published.

Chiang, Tzuu-Shuh, Chii-Ruey Hwang, and Shuenn Jyi Sheu (1987). "Diffusion for Global Optimization in $\mathbb{R}^n$". In: *SIAM Journal on Control and Optimization* 25.3, pp. 737–753. ISSN: 0363-0129. DOI: 10.1137/0325042.

Choromanska, Anna et al. (2015). "The Loss Surfaces of Multilayer Networks". In: *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*. Artificial Intelligence and Statistics. San Diego, CA, USA: PMLR, pp. 192–204. arXiv: 1412.0233 [cs]. URL: https://proceedings.mlr.press/v38/choromanska15.html (visited on 2023-01-23).

Costa, Nathaël Da et al. (2024). *Sample Path Regularity of Gaussian Processes from the Covariance Kernel*. DOI: 10.48550/arXiv.2312.14886. arXiv: 2312.14886 [cs]. Pre-published.

Dauphin, Yann N et al. (2014). "Identifying and Attacking the Saddle Point Problem in High-Dimensional Non-Convex Optimization". In: *Advances in Neural Information Processing Systems*. Vol. 27. Montréal, Canada: Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper/2014/hash/17e23e50bedc63b4095e3d8204ce063b-Abstract.html (visited on 2022-06-10).

Donoho, David L., Arian Maleki, and Andrea Montanari (2009). "Message-Passing Algorithms for Compressed Sensing". In: *Proceedings of the National Academy of Sciences* 106.45, pp. 18914–18919. DOI: 10.1073/pnas.0909892106.

— (2010). "Message Passing Algorithms for Compressed Sensing: I. Motivation and Construction". In: *2010 IEEE Information Theory Workshop on Information Theory (ITW 2010, Cairo)*. 2010 IEEE Information Theory Workshop on Information Theory (ITW 2010, Cairo), pp. 1–5. DOI: 10.1109/ITWKSPS.2010.5503193.

Eaton, Morris L. (2007). *Multivariate Statistics: A Vector Space Approach*. Ed. by R. A. Vitale. Vol. 53. Lecture Notes-Monograph Series. Beachwood, Ohio, USA: Institute of Mathematical Statistics. 519 pp. ISBN: 978-0-940600-69-0. DOI: 10.1214/lnms/1196285102.

Frazier, Peter, Warren Powell, and Savas Dayanik (2009). "The Knowledge-Gradient Policy for Correlated Normal Beliefs". In: *INFORMS Journal on Computing* 21.4, pp. 599–613. ISSN: 1091-9856. DOI: 10.1287/ijoc.1080.0314.

Frazier, Peter I. (2018). "Bayesian Optimization". In: *Recent Advances in Optimization and Modeling of Contemporary Problems*. INFORMS TutORials in Operations Research. Phoenix, Arizona, USA: INFORMS, pp. 255–278. ISBN: 978-0-9906153-2-3. DOI: 10.1287/educ.2018.0188. arXiv: 1807.02811 [cs, math, stat].

Garnett, Roman (2023). *Bayesian Optimization*. 1st edition. Cambridge, United Kingdom ; New York, NY: Cambridge University Press. 358 pp. ISBN: 978-1-108-42578-0. URL: https://bayesoptbook.com/.

Gihman, Iosif Il'ich and Anatoliĭ Vladimirovich Skorokhod (1974). *The Theory of Stochastic Processes I*. Classics in Mathematics. Berlin, Heidelberg: Springer.

ISBN: 978-3-540-20284-4 978-3-642-61943-4. DOI: 10.1007/978-3-642-61943-4.

Hoare, C. A. R. (1962). "Quicksort". In: *The Computer Journal* 5.1, pp. 10–16. ISSN: 0010-4620. DOI: 10.1093/comjnl/5.1.10.

Huang, Brice and Mark Sellke (2022). "Tight Lipschitz Hardness for Optimizing Mean Field Spin Glasses". In: *2022 IEEE 63rd Annual Symposium on Foundations of Computer Science (FOCS)*. 2022 IEEE 63rd Annual Symposium on Foundations of Computer Science (FOCS), pp. 312–322. DOI: 10.1109/FOCS54457.2022.00037.

Jones, Donald R., Matthias Schonlau, and William J. Welch (1998). "Efficient Global Optimization of Expensive Black-Box Functions". In: *Journal of Global Optimization* 13.4, pp. 455–492. ISSN: 1573-2916. DOI: 10.1023/A:1008306431147.

Kallenberg, Olav (2002). *Foundations of Modern Probability*. Red. by J. Gani, C. C. Heyde, and T. G. Kurtz. Probability and Its Applications. New York, NY: Springer. ISBN: 978-1-4419-2949-5 978-1-4757-4015-8. DOI: 10.1007/978-1-4757-4015-8.

Klenke, Achim (2014). *Probability Theory: A Comprehensive Course*. Universitext. London: Springer. ISBN: 978-1-4471-5360-3 978-1-4471-5361-0. DOI: 10.1007/978-1-4471-5361-0.

Krige, Daniel G. (1951). "A Statistical Approach to Some Basic Mine Valuation Problems on the Witwatersrand". In: *Journal of the Southern African Institute of Mining and Metallurgy* 52.6, pp. 119–139. URL: https://journals.co.za/doi/abs/10.10520/AJA0038223X_4792 (visited on 2025-03-25).

Kushner, H. J. (1964). "A New Method of Locating the Maximum Point of an Arbitrary Multipeak Curve in the Presence of Noise". In: *Journal of Basic Engineering* 86.1, pp. 97–106. ISSN: 0021-9223. DOI: 10.1115/1.3653121.

Matheron, Georges (1963). "Principles of Geostatistics". In: *Economic Geology* 58.8, pp. 1246–1266. ISSN: 0361-0128. DOI: 10.2113/gsecongeo.58.8.1246.

Mockus, J. B. and L. J. Mockus (1991). "Bayesian Approach to Global Optimization and Application to Multiobjective and Constrained Problems". In: *Journal of Optimization Theory and Applications* 70.1, pp. 157–172. ISSN: 1573-2878. DOI: 10.1007/BF00940509.

Montanari, Andrea (2019). "Optimization of the Sherrington-Kirkpatrick Hamiltonian". In: *2019 IEEE 60th Annual Symposium on Foundations of Computer Science (FOCS)*. 2019 IEEE 60th Annual Symposium on Foundations of Computer Science (FOCS), pp. 1417–1433. DOI: 10.1109/FOCS.2019.00087.

Raginsky, Maxim, Alexander Rakhlin, and Matus Telgarsky (2017). "Non-Convex Learning via Stochastic Gradient Langevin Dynamics: A Nonasymptotic Analysis". In: *Proceedings of the 2017 Conference on Learning Theory*. Conference on Learning Theory. PMLR, pp. 1674–1703. URL: https://proceedings.mlr.press/v65/raginsky17a.html (visited on 2025-02-21).

Salamon, Peter, Paolo Sibani, and Richard Frost (2002). *Facts, Conjectures, and Improvements for Simulated Annealing*. Mathematical Modeling and Computation. Society for Industrial and Applied Mathematics. 144 pp. ISBN: 978-0-89871-508-8. DOI: 10.1137/1.9780898718300.

Scheurer, Michael (2009). "A Comparison of Models and Methods for Spatial Interpolation in Statistics and Numerical Analysis". Göttingen. 142 pp. DOI: 10.53846/goediss-2461.

Shannon, C. E. (1948). "A Mathematical Theory of Communication". In: *The Bell System Technical Journal* 27.3, pp. 379–423. ISSN: 0005-8580. DOI: 10.1002/j.1538-7305.1948.tb01338.x.

Smale, Steve (1983). "On the Average Number of Steps of the Simplex Method of Linear Programming". In: *Mathematical Programming* 27.3, pp. 241–262. ISSN: 1436-4646. DOI: 10.1007/BF02591902.

Spielman, Daniel A. and Shang-Hua Teng (2004). "Smoothed Analysis of Algorithms: Why the Simplex Algorithm Usually Takes Polynomial Time". In: *Journal of the ACM* 51.3, pp. 385–463. ISSN: 0004-5411. DOI: 10.1145/990308.990310.

Srinivas, Niranjan et al. (2010). "Gaussian Process Optimization in the Bandit Setting: No Regret and Experimental Design". In: *Proceedings of the 27th International Conference on Machine Learning*. ICML. Haifa, Israel. arXiv: 0912.3995.

Stein, Michael L. (1999). *Interpolation of Spatial Data*. Springer Series in Statistics. New York, NY: Springer. ISBN: 978-1-4612-7166-6 978-1-4612-1494-6. DOI: 10.1007/978-1-4612-1494-6.

Subag, Eliran (2021). "Following the Ground States of Full-RSB Spherical Spin Glasses". In: *Communications on Pure and Applied Mathematics* 74.5, pp. 1021–1044. ISSN: 1097-0312. DOI: 10.1002/cpa.21922.

Talagrand, Michel (1987). "Regularity of Gaussian Processes". In: *Acta Mathematica* 159 (none), pp. 99–149. ISSN: 0001-5962, 1871-2509. DOI: 10.1007/BF02392556.

Thompson, William R. (1933). "On the Likelihood That One Unknown Probability Exceeds Another in View of the Evidence of Two Samples". In: *Biometrika* 25.3–4, pp. 285–294. URL: https://academic.oup.com/biomet/article-pdf/25/3--4/285/513725/25--3--4--285.pdf (visited on 2025-03-05).

Todd, Michael J. (1991). "Probabilistic Models for Linear Programming". In: *Mathematics of Operations Research* 16.4, pp. 671–693. ISSN: 0364-765X. DOI: 10.1287/moor.16.4.671.

Van Laarhoven, Peter J. M. and Emile H. L. Aarts (1987). *Simulated Annealing: Theory and Applications*. Dordrecht: Springer Netherlands. ISBN: 978-90-481-8438-5 978-94-015-7744-1. DOI: 10.1007/978-94-015-7744-1.

Vazquez, Emmanuel and Julien Bect (2010). "Convergence Properties of the Expected Improvement Algorithm with Fixed Mean and Covariance Functions". In: *Journal of Statistical Planning and Inference* 140.11, pp. 3088–3095. ISSN: 0378-3758. DOI: 10.1016/j.jspi.2010.04.018.

Xu, Pan et al. (2018). "Global Convergence of Langevin Dynamics Based Algorithms for Nonconvex Optimization". In: *Advances in Neural Information Processing Systems*. Vol. 31. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper/2018/hash/9c19a2aa1d84e04b0bd4bc888792bd1e-Abstract.html (visited on 2025-02-21).

Zou, Difan, Pan Xu, and Quanquan Gu (2021). "Faster Convergence of Stochastic Gradient Langevin Dynamics for Non-Log-Concave Sampling". In: *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence*. Uncertainty in Artificial Intelligence. PMLR, pp. 1152–1162. URL: https://proceedings.mlr.press/v161/zou21a.html (visited on 2025-02-21).

Chapter 3

Measure theory for random function optimization

In this chapter we highlight and address measure theoretical challenges that must be overcome for a rigorous mathematical treatment of random function optimization. In particular, we show how non-rigorous but intuitive claims about the conditional distributions of the objective function $\mathbf{f}$ can be made rigorous. Before that though, we need to address measurability itself.

Measurability of random evaluations. The evaluation $\mathbf{f}(X)$ of a random function $\mathbf{f} = (\mathbf{f}(x))_{x\in\mathbb{X}}$ at a random location X is a complicated measure theoretical object. Ex ante not even the **measurability of** $\mathbf{f}(X)$, i.e. its existence as a random variable, is certain. In the study of stopping times this problem is sometimes defined away by the requirement that the stochastic process $\mathbf{f} = (\mathbf{f}(s))_{s\in\mathbb{R}_+}$ is *progressive*.[1] This leaves the user with the burden to confirm that their process is in fact progressive. However, the main reason we do not take this approach is that it is tailored for stochastic processes with one dimensional input and stopping times.

[1] e.g. Kallenberg, 2002, Lem. 7.5.

We prefer the assumption that the evaluation function $e(f, x) := f(x)$ is *measurable*, which immediately implies that $\mathbf{f}(X) = e(\mathbf{f}, X)$ is a random variable for any random variable X. This assumption holds with great generality: If $\mathbf{f}$ is a *continuous* random function with locally compact, separable, metrizable domain $\mathbb{X}$ and polish co-domain $\mathbb{Y}$, then the evaluation function e is continuous and thereby measurable (cf. Theorem 3.4.1). This accounts for almost all continuous applications, in particular the case $\mathbb{X} \subseteq \mathbb{R}^d$ and $\mathbb{Y} = \mathbb{R}^n$.

Conditional distributions. During the optimization of a random function $\mathbf{f} = (\mathbf{f}(x))_{x\in\mathbb{X}}$ one typically selects evaluation locations X_{n+1} in $\mathbb{X}$ based on the previously seen evaluations $\mathcal{F}_n := \sigma(\mathbf{f}(X_0), \dots, \mathbf{f}(X_n))$. Since X_{n+1} is thereby measurable with respect to $\mathcal{F}_n$, the sequence $(X_n)_{n\in\mathbb{N}}$ is called *previsible*[2] with respect to the filtration $(\mathcal{F}_n)_{n\in\mathbb{N}}$. Our main result is a formalization of the intuitive notion that previsible evaluation locations may be treated as if they are deterministic during the calculation of the conditional distribution

[2] see also Definition 3.1.5

$$\mathbb{P}(\mathbf{f}(X_n) \in \cdot \mid \mathbf{f}(X_0), \dots, \mathbf{f}(X_{n-1})).$$

In the case of Gaussian random functions $\mathbf{f}$ for example, $(\mathbf{f}(x_0), \dots, \mathbf{f}(x_n))$ is a multivariate Gaussian vector with well known conditional distribution $\mathbf{f}(x_n)$ given $(\mathbf{f}(x_0), \dots, \mathbf{f}(x_{n-1}))$ when the evaluation locations are deterministic. But $\mathbf{f}(X)$ is not necessarily Gaussian if X is random[3] and the calculation of conditional distributions becomes much more difficult. Treating previsible inputs as deterministic ensures the calculation is feasible but it lacks theoretical foundation.

[3] consider $X = \arg\min_{x\in K} \mathbf{f}(x)$ for some compact set $K \subseteq \mathbb{X}$.

We formalize the hope, that previsible evaluation locations may be treated as deterministic as follows. Let $(\kappa_{x_{[0:n]}})_{x_{[0:n]}\in\mathbb{X}^{n+1}}$ be a collection of regular conditional distributions for $\mathbf{f}(x_n)$ given $\mathbf{f}(x_{[0:n)}) = (\mathbf{f}(x_0), \dots, \mathbf{f}(x_{n-1}))$[4] indexed by the

[4] e.g. Klenke, 2014, Def. 8.28.

evaluation locations $x_{[0:n]} = (x_0, \dots, x_n)$, where we use the following notation for discrete intervals

$$[i{:}j] := [i, j] \cap \mathbb{Z}, \qquad [i{:}j) := [i, j) \cap \mathbb{Z}, \qquad \text{etc.} \qquad \text{(discrete intervals)}$$

For all locations $x_{[0:n]}$ we thus have for all measurable sets A

$$\mathbb{P}\big(\mathbf{f}(x_n) \in A \mid \mathbf{f}(x_{[0:n)})\big) \overset{\text{a.s.}}{=} \kappa_{x_{[0:n]}}\big(\mathbf{f}(x_{[0:n)}); A\big).$$

Recall that this collection is easy to come by in the Gaussian case. For previsible locations $X_{[0:n]}$ the *hope* is therefore that for all measurable sets A

$$\mathbb{P}\big(\mathbf{f}(X_n) \in A \mid \mathbf{f}(X_{[0:n)})\big) \overset{\text{a.s.}}{=} \kappa_{X_{[0:n]}}\big(\mathbf{f}(X_{[0:n)}); A\big). \tag{3.1}$$

In practice, this hope is often naively treated as self-evident.[5] But while the collection of probability kernels $(\kappa_{x_{[0:n]}})_{x_{[0:n]} \in \mathbb{X}^{n+1}}$ may be treated as a function in $x_{[0:n]}$, there is no guarantee this function is even measurable. This means that the term on the right in (3.1) might not even be a well defined random variable, let alone satisfy the equation (cf. Example 3.1.2).

[5] e.g. Srinivas et al., 2012, Lemma 5.1, p. 3258.

Outline In Section 3.1 we introduce the concepts needed to formalize the treatment of previsible random variables as deterministic and state our main results for continuous random functions. In Section 3.2 we prove the building blocks for our main results, which are then used in Section 3.3 to state and prove our main result with the additional generalization to conditionally independent evaluation locations and noisy evaluations. Section 3.4 is concerned with the topological foundations that ensure the evaluation function $e(f, x) = f(x)$ is measurable on the space of continuous functions; and the limitations of this approach.

3.1 Main results

To ensure that we may plug random variables into the index of a collection of regular conditional distributions, we introduce the concept of a 'joint' probability kernel.

Definition 3.1.1 (Joint probability kernels)**.** A probability kernel κ is a *joint probability kernel* for the collection $(\kappa_x)_{x \in I}$ of probability kernels with index set I, if for all $x \in I$

$$\kappa_x(\omega; A) = \kappa(\omega, x; A) \qquad \forall \omega, A.$$

We call κ a *joint conditional distribution*, if the collection of probability kernels $(\kappa_x)_{x \in I}$ is a collection of regular conditional distributions.

Due to the measurability of a probability kernel for fixed sets A, the existence of a joint conditional distribution ensures that the object in (3.1) is a well defined random variable.

In the following, we provide sufficient conditions for a joint regular conditional distribution to be *consistent* – a term we use informally to describe the setting in which random evaluation locations can be treated as if they were deterministic, as conjectured in (3.1).

The meaning of 'consistent' is often clear from context and should therefore help with the interpretation of our results. Nevertheless there are slightly different requirements (e.g. measurable/previsible/conditionally independent) on the random evaluation locations in every case, so we accompany any occurrence of the term with a clarification of its precise meaning.

Example 3.1.2 (A joint conditional distribution that is not consistent)**.** Consider a standard uniform random variable $U \sim \mathcal{U}(0, 1)$, an independent standard normal random variable $Y \sim \mathcal{N}(0, 1)$ and define $\mathbf{f}(x) = Y$ for all $x \in \mathbb{X} := [0, 1]$. As a constant function $\mathbf{f}$ is clearly a continuous Gaussian random function. Let

$$\kappa(u, x; B) := \mathbb{P}_Y(B) \qquad \tilde{\kappa}(u, x; B) := \mathbb{P}_Y(B)\mathbf{1}_{U \neq x} + \delta_0(B)\mathbf{1}_{U = x}.$$

Then clearly for all $x \in \mathbb{X}$

$$\kappa(U, x; B) \overset{\text{a.s.}}{=} \mathbb{P}(\mathbf{f}(x) \in B \mid U) \overset{\text{a.s.}}{=} \tilde{\kappa}(U, x; B)$$

and thereby both κ and $\tilde{\kappa}$ are joint regular probability kernels for $\mathbf{f}(x)$ conditioned on U. But $\tilde{\kappa}$ is not consistent, because for most measurable sets B

$$\mathbb{P}(\mathbf{f}(U) \in B \mid U) = \mathbb{P}_Y(B) \neq \delta_0(B) = \tilde{\kappa}(U, U; B),$$

even though U is clearly measurable with respect to U.

Observe that for any fixed x, the kernels in the example above coincide almost surely. That is, they coincide up to a null set N_x, specifically $N_x = \{U = x\}$. But the union of these null sets over all possible values of x is not a null set. Joint null set can ensure that the consistency of one kernel implies the other. A sufficient criterion for such a joint null set is a notion of continuity.

The following result implies that a continuous, joint conditional distribution is consistent and such a conditional distribution exists.

Theorem 3.1.3 (Consistency for dependent evaluations $\mathbf{f}(x)$)**.** *Let $\mathcal{F}$ be a sub σ-algebra of the underlying probability space $(\Omega, \mathcal{A}, \mathbb{P})$ and $\mathbf{f}$ a random variable in the space of continuous function $C(\mathbb{X}, \mathbb{Y})$, with locally compact, separable metrizable domain $\mathbb{X}$ and Polish co-domain $\mathbb{Y}$. Then there exists a consistent joint conditional distribution κ for $\mathbf{f}(x)$ given $\mathcal{F}$, which means*

$$\mathbb{P}(\mathbf{f}(X) \in B \mid \mathcal{F})(\omega) = \kappa(\omega, X(\omega); B)$$

for all $\mathcal{F}$-measurable X.

Furthermore $x \mapsto \kappa(\omega, x; \cdot)$ is continuous with respect to the weak topology on the space of measures for all ω. Let $\tilde{\kappa}$ be another joint conditional distribution for $\mathbf{f}(x)$ given $\mathcal{F}$, which is continuous in this sense. Then there exists a joint null set N such that for all $\omega \in N^{\complement}$, all $x \in \mathbb{X}$ and all borel sets $B \in \mathcal{B}(\mathbb{Y})$

$$\kappa(\omega, x; B) = \tilde{\kappa}(\omega, x; B).$$

In particular, $\tilde{\kappa}$ is also a consistent joint conditional distribution.

This Theorem will be proven as a special case of Theorem 3.2.1. The following example demonstrates its use in the optimization of random functions.

Example 3.1.4 (Conditional minimization)**.** Assume the setting of Theorem 3.1.3 and that $\mathbf{f}(x)$ is integrable. Using the consistent joint conditional distribution κ there exists a measurable function $H(\omega, x) := \int y \, \kappa(\omega, x; dy)$ such that by disintegration[6] we have for $\mathbb{P}$-almost all ω and all $x \in \mathbb{X}$

$$\mathbb{E}[\mathbf{f}(x) \mid \mathcal{F}](\omega) = H(\omega, x) \quad \text{and} \quad \mathbb{E}[\mathbf{f}(X) \mid \mathcal{F}](\omega) = H(\omega, X(\omega))$$

for any $\mathcal{F}$-measurable X. That is, we can treat $\mathcal{F}$-measurable random variables X in $\mathbb{X}$ as if they were deterministic inputs to $\mathbb{E}[\mathbf{f}(x) \mid \mathcal{F}]$. This implies for any such X

$$\inf_{x \in \mathbb{X}} \mathbb{E}[\mathbf{f}(x) \mid \mathcal{F}] \leq \mathbb{E}[\mathbf{f}(X) \mid \mathcal{F}]. \tag{3.2}$$

And if $X_* := \arg\min_{x \in \mathbb{X}} \mathbb{E}[\mathbf{f}(x) \mid \mathcal{F}]$ is a $\mathcal{F}$-measurable random variable,[7] we have

$$\inf_{x \in \mathbb{X}} \mathbb{E}[\mathbf{f}(x) \mid \mathcal{F}] = \inf_{\substack{X \ \mathcal{F}\text{-meas.} \\ \text{r.v. in } \mathbb{X}}} \mathbb{E}[\mathbf{f}(X) \mid \mathcal{F}] = \mathbb{E}[\mathbf{f}(X_*) \mid \mathcal{F}]. \tag{3.3}$$

So far, the random function evaluation $\mathbf{f}(x)$ only occurred as a dependent variable. In the following we analyze the case where $\mathbf{f}(x)$ is conditioned on. While consistency was an issue when $\mathbf{f}(x)$ is a dependent variable, it turns out that every joint conditional distribution is consistent when $\mathbf{f}(x)$ is conditioned on.

Definition 3.1.5 (Previsible)**.** The *previsible setting* is

[6] e.g. Kallenberg, 2002, Thm. 6.4.

[7] This is a non-trivial problem by itself (see e.g. Aliprantis and Border, 2006, Thm. 18.19).

[8] There always exists such a random element W since the identity map from the measurable space $(\Omega, \mathcal{A})$ into $(\Omega, \mathcal{F})$ is measurable and clearly generates $\mathcal{F}$.

- an underlying probability space $(\Omega, \mathcal{A}, \mathbb{P})$,
- a sub-σ-algebra $\mathcal{F}$ (the 'initial information') with W a random element such that $\mathcal{F} = \sigma(W)$,[8]
- $\mathbf{f}$ a random function in the space of continuous functions $C(\mathbb{X}, \mathbb{Y})$, where $\mathbb{X}$ is a locally compact, separable metrizable space and $\mathbb{Y}$ a polish space.

A sequence $X = (X_n)_{n\in\mathbb{N}_0}$ of random evaluation locations in $\mathbb{X}$ is called *previsible*, if X_{n+1} is measurable with respect to

$$\mathcal{F}_n^X := \sigma(\mathcal{F}, \mathbf{f}(X_0), \dots, \mathbf{f}(X_n)) \qquad \text{for } n \geq -1.$$

Theorem 3.1.6 (Previsible sampling)**.** *Assume the previsible setting (Def. 3.1.5).*

(i) Let Z be a random variable in a standard Borel space $(E, \mathcal{B}(E))$ and κ a joint conditional distribution for Z given $\mathcal{F}, \mathbf{f}(x_0), \dots, \mathbf{f}(x_n)$, i.e.

$$\mathbb{P}(Z \in A \mid \mathcal{F}, \mathbf{f}(x_0), \dots, \mathbf{f}(x_n)) \overset{a.s.}{=} \kappa(W, \mathbf{f}(x_0), \dots, \mathbf{f}(x_n), x_{[0:n]}; B)$$

for all $x_{[0:n]} \in \mathbb{X}^{n+1}$ and $A \in \mathcal{B}(E)$. Then κ is consistent, i.e. for all previsible sequences $(X_k)_{k\in\mathbb{N}_0}$ and all $A \in \mathcal{B}(E)$

$$\mathbb{P}(Z \in A \mid \mathcal{F}_n^X) \overset{a.s.}{=} \kappa(W, \mathbf{f}(X_0), \dots, \mathbf{f}(X_n), X_{[0:n]}; A).$$

(ii) Let κ be a joint conditional distribution for $\mathbf{f}(x_n)$ given $\mathcal{F}, \mathbf{f}(x_{[0:n)})$ such that

$$x_n \mapsto \kappa(y_{[0:n)}, x_{[0:n]}; \,\cdot\,)$$

is continuous with respect to the weak topology on the space of measures for all $x_{[0:n)} \in \mathbb{X}^n$ and $y_{[0:n)} \in \mathbb{X}^n$.

Then κ is consistent, i.e. for all previsible $(X_k)_{k\in\mathbb{N}_0}$ and $B \in \mathcal{B}(\mathbb{Y})$

$$\mathbb{P}(\mathbf{f}(X_n) \in B \mid \mathcal{F}_{n-1}^X) \overset{a.s.}{=} \kappa(W, \mathbf{f}(X_{[0:n)}), X_{[0:n]}; B).$$

Theorem 3.1.6 is proven as a special case of Theorem 3.3.2. There we allow X_{n+1} to be random, conditional on $\mathcal{F}_n$ and only require it to be independent from $\mathbf{f}$ conditional on $\mathcal{F}_n$. This result also covers noisy evaluations of $\mathbf{f}$.

We want to highlight that continuity of the kernel is only required for the case where function evaluations are dependent variables. While consistency is therefore never an issue, the existence of such a joint conditional distribution is uncertain in general. However in the Gaussian case, the joint conditional distribution is known explicitly.

Example 3.1.7 (Gaussian case)**.** Let $\mathbf{f} = (\mathbf{f}(x))_{x\in\mathbb{X}}$ be a Gaussian random function with mean and covariance functions

$$\mu_0(x) = \mathbb{E}[\mathbf{f}(x)] \qquad \text{and} \qquad \mathcal{C}_{\mathbf{f}}(x, y) = \operatorname{Cov}(\mathbf{f}(x), \mathbf{f}(y)).$$

The conditional distribution of $\mathbf{f}(x_n)$ given $\mathbf{f}(x_{[0:n)})$ is the conditional distribution of a multivariate Gaussian random vector $\mathbf{f}(x_{[0:n]})$, which is well known to be $\mathcal{N}\big(\mu_n(x_{[0:n]}, \mathbf{f}(x_{[0:n)})), \Sigma_n(x_{[0:n]})\big)$,[9] with

[9] e.g. Theorem 2.A.1 or Eaton (e.g. 2007, Prop. 3.13)

$$\mu_n(x_{[0:n]}, y_{[0:n)}) := \mu_0(x_n) + \sum_{i,j=0}^{n-1} \mathcal{C}_{\mathbf{f}}(x_n, x_i)\big[\Sigma_0(x_{[0:n)})^{-1}\big]_{ij}\big(y_j - \mu_0(x_j)\big)$$

$$\Sigma_n(x_{[0:n]}) := \Sigma_0(x_n) - \sum_{i,j=0}^{n-1} \mathcal{C}_{\mathbf{f}}(x_n, x_i)\big[\Sigma_0(x_{[0:n)})^{-1}\big]_{ij}\mathcal{C}_{\mathbf{f}}(x_j, x_n),$$

where $\Sigma_0(x_{[0:n)})_{ij} := \operatorname{Cov}(\mathbf{f}(x_i), \mathbf{f}(x_j))$. This induces the joint conditional distribution

$$\kappa(y_{[0:n)}, x_{[0:n]}; B) \propto \int_B \exp\Big(-\tfrac{1}{2}(t-\mu_n)^T\Sigma_n(x_{[0:n]})^{-1}(t-\mu_n)\Big)dt \tag{3.4}$$

[10] e.g. Costa et al., 2024; Talagrand, 1987, Thm. 3.

with $\mu_n := \mu_n(x_{[0:n]}, y_{[0:n)})$. Now if $\mathbf{f}$ is a *continuous* Gaussian random function, Theorem 3.1.6 (i) is immediately applicable. For the applicability of (ii) observe that a continuous Gaussian random function must have continuous mean μ_0 and covariance $\mathcal{C}_\mathbf{f}$.[10] And the continuity of μ_0 and $\mathcal{C}_\mathbf{f}$ is sufficient for μ_n and Σ_n to be continuous in x_n. This implies the characteristic function of the joint conditional distribution

$$\hat{\kappa}(y_{[0:n)}, x_{[0:n]}; t) = \exp\Big(it^T \mu_n(x_{[0:n]}, y_{[0:n)}) - \tfrac{1}{2} t^T \Sigma_n(x_{[0:n]}) t\Big)$$

is continuous in x_n, which implies continuity of κ in the weak topology by Lévy's continuity theorem.[11]

[11] e.g. Kallenberg, 2002, Thm. 5.3.

Corollary 3.1.8 (Gaussian previsible sampling)**.** *For a continuous Gaussian random function* $\mathbf{f}$*,* (3.4) *is a consistent joint probability kernel for* $\mathbf{f}$ *in the sense of Theorem 3.1.6.*

3.2 Previsible sampling

In this section we establish the building blocks for our main results. As the analysis of multiple function evaluations will ultimately proceed via induction, we restrict our attention here to the case of a single function evaluation. This section thereby lays the groundwork for the most general form of our results, presented in Section 3.3.

Throughout this section any σ-algebra is implicitly assumed to be a sub σ-algebra of the underlying probability space $(\Omega, \mathcal{A}, \mathbb{P})$. $\mathbb{X}$ always denotes a locally compact, separable metrizable space and $\mathbb{Y}$ a polish space.

Theorem 3.2.1 (Consistency for dependent $\mathbf{f}(x)$)**.** *Let* $\mathcal{F}$ *be a* σ*-algebra,* Z *a random variable in the standard borel space* $(E, \mathcal{B}(E))$ *and* $\mathbf{f}$ *a random variable in* $C(\mathbb{X}, \mathbb{Y})$*. Then there exists a consistent joint conditional distribution* κ *for* $Z, \mathbf{f}(x)$ *given* $\mathcal{F}$*. That is*

$$\mathbb{P}(Z, \mathbf{f}(X) \in B \mid \mathcal{F})(\omega) = \kappa(\omega, X(\omega); B)$$

for all $\mathcal{F}$*-measurable* X*.*

Furthermore $x \mapsto \kappa(\omega, x; \cdot)$ *is continuous with respect to the weak topology on the space of measures for all* $\omega \in \Omega$*. If* $\tilde{\kappa}$ *is another joint conditional distribution for* $Z, \mathbf{f}(x)$ *given* $\mathcal{F}$*, that is continuous in this sense, then there exists a joint null set* N *such that for all* $\omega \in N^\complement$*, all* $x \in \mathbb{X}$ *and all borel sets* B

$$\kappa(\omega, x; B) = \tilde{\kappa}(\omega, x; B).$$

In particular, $\tilde{\kappa}$ *is also a consistent joint conditional distribution.*

Remark 3.2.2 (Existence of consistent joint conditional distribution). Note that for the existence of a consistent joint conditional distribution we only require a regular conditional distribution for $Z, \mathbf{f}$ given $\mathcal{F}$ to exist and measurability of the evaluation map e. This part of the result can therefore be made to hold with greater generality.

Proof. Observe that $E \times C(\mathbb{X}, \mathbb{Y})$ is a standard borel space since $C(\mathbb{X}, \mathbb{Y})$ is Polish (Theorem 3.4.1). There therefore exists a regular conditional probability distribution $\kappa_{Z,\mathbf{f}|\mathcal{F}}$.[12] Using this probability kernel, we define the kernel

[12] e.g. Kallenberg, 2002, Thm. 6.3.

$$\kappa(\omega, x; B) := \int \mathbf{1}_B(z, e(f, x)) \kappa_{Z,\mathbf{f}|\mathcal{F}}(\omega; dz \otimes df)$$

which is a measure in $B \in \mathcal{B}(E) \otimes \mathcal{B}(\mathbb{Y})$ by linearity of the integral, so we only need to prove it is measurable in $(\omega, x) \in \Omega \times \mathbb{X}$ to prove it is a probability kernel. This follows from measurability of the evaluation function e (Theorem 3.4.1) and the application of Lemma 14.20 by Klenke (2014) to the probability kernel $\tilde{\kappa}(\omega, x; A) :=$

[13] e.g. Kallenberg, 2002, Thm 6.4.

$\kappa_{Z,\mathbf{f}|\mathcal{F}}(\omega; A)$ in the equation above. By 'disintegration'[13] this probability kernel is moreover a regular conditional version of $\mathbb{P}(Z, \mathbf{f}(X) \in B \mid \mathcal{F})$ for all $\mathcal{F}$-measurable X, i.e. for all $B \in \mathcal{B}(E) \otimes \mathcal{B}(\mathbb{Y})$ and for $\mathbb{P}$-almost all ω

$$\mathbb{P}\big(Z, \mathbf{f}(X) \in B \mid \mathcal{F}\big)(\omega) \overset{\text{disint.}}{=} \int \mathbf{1}_B(z, e(f, X(\omega)))\kappa_{Z,\mathbf{f}|\mathcal{F}}(\omega; dz \otimes df) \overset{\text{def.}}{=} \kappa(\omega, X(\omega); B).$$

The kernel is thereby consistent (and a joint kernel since the constant map $X \equiv y$ is $\mathcal{F}$ measurable).

For continuity observe that we have $\lim_{x\to y}(z, f(x)) = (z, f(y))$ for any $f \in C(\mathbb{X}, \mathbb{Y})$. For open U this implies

$$\liminf_{x\to y} \mathbf{1}_U(z, f(x)) \geq \mathbf{1}_U(z, f(y)),$$

because if $(z, f(y)) \in U$, then eventually $(z, f(x))$ in U due to openness of U. An application of Fatou's lemma[14] yields for all open U

[14] e.g. Klenke, 2014, Thm. 4.21.

$$\liminf_{x\to y} \kappa(\omega, x; U) \geq \int \liminf_{x\to y} \mathbf{1}_U(z, f(x))\kappa_{Z,\mathbf{f}|\mathcal{F}}(\omega; dz \otimes df) \geq \kappa(\omega, y; U).$$

And we can conclude weak convergency by the Portemanteau theorem[15] since $E \times \mathbb{Y}$ is metrizable.

[15] Klenke, 2014, Thm. 13.16.

Let $\tilde{\kappa}$ be another continuous joint probability kernel. Since $E \times \mathbb{Y}$ is second countable, there is a countable base $\{U_n\}_{n\in\mathbb{N}}$ of its topology, which generates the Borel σ-algebra $\mathcal{B}(E) \otimes \mathcal{B}(\mathbb{Y})$. And since $\mathbb{X}$ is separable, it has a countable dense subset Q. There must therefore exist a zero set N such that

$$\kappa(\omega, q; U_n) = \tilde{\kappa}(\omega, q; U_n), \qquad \forall \omega \in N^{\complement},\ n \in \mathbb{N},\ q \in Q,$$

because both kernels are regular conditional version of $\mathbb{P}(Z, \mathbf{f}(q) \in U_n; \mathcal{G})$ and the union over $\mathbb{N} \times Q$ is a countable union. Since $\{U_n\}_{n\in\mathbb{N}}$ generates the σ-algebra, we deduce for all $\omega \in N^{\complement}$ and all $q \in Q$ that $\kappa(\omega, q; \cdot) = \tilde{\kappa}(\omega, q; \cdot)$. As Q is dense in $\mathbb{X}$ we have by continuity of the joint kernels for all $\omega \in N^{\complement}$ and all $x \in \mathbb{X}$

$$\kappa(\omega, x; \cdot) = \tilde{\kappa}(\omega, x; \cdot). \qquad \square$$

In Example 3.1.2 we showed a joint probability distribution to exist, which is not consistent. In this example, the random function was a dependent variable. In Theorem 3.2.1 we gave a sufficient condition for a unique continuous and consistent conditional distributions to exist in this case where the function value is the dependent variable. It is now time to consider the case where functions evaluated at random points are conditional variables. The following result shows that we never have to worry about the consistency of probability kernels where the function value is a conditional variable, if a consistent joint probability kernel exists for function values as dependent variables.

Proposition 3.2.3 (Consistency shuffle)**.** *Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space and let $\xi_1, (\xi_2^y)_{y\in D}, \xi_3$ be random variables in the measurable spaces $(E_i, \mathcal{E}_i)$ with measurable domain $(D, \mathcal{D})$.*

> *If there exists a consistent joint probability kernel $\kappa_{3,2|1}$ for ξ_3, ξ_2^y given ξ_1, then **any** joint probability kernel $\kappa_{3|2,1}$ for ξ_3 given ξ_2^y, ξ_1 is consistent.*

Formally, assume there exists a consistent probability kernel $\kappa_{3,2|1}$ for ξ_3, ξ_2^y given ξ_1, i.e. for all $A \in \mathcal{E}_3 \otimes \mathcal{E}_2$ and all measurable functions $g\colon E_1 \to D$

$$\mathbb{P}\big(\xi_3, \xi_2^{g(\xi_1)} \in A \mid \xi_1\big) \overset{a.s.}{=} \kappa_{3,2|1}(\xi_1, g(\xi_1); A), \tag{3.5}$$

where we assume $\xi_2^{g(\xi_1)}$ *is a random variable, i.e. measurable. Then if there exists a joint conditional probability kernel* $\kappa_{3|1,2}$ *for* ξ_3 *given* ξ_1, ξ_2^y *such that for all* $y \in D$ *and* $A_3 \in \mathcal{E}_3$

$$\mathbb{P}(\xi_3 \in A_3 \mid \xi_1, \xi_2^y) \overset{a.s.}{=} \kappa_{3|2,1}(\xi_1, \xi_2^y, y; A_3), \tag{3.6}$$

then $\kappa_{3|2,1}$ *is consistent, i.e. we have for all* $A_3 \in \mathcal{E}_3$ *and measurable* $g\colon E_1 \to D$

$$\mathbb{P}(\xi_3 \in A_3 \mid \xi_1, \xi_2^{g(\xi_1)}) \overset{a.s.}{=} \kappa_{3|2,1}\big(\xi_1, \xi_2^{g(\xi_1)}, g(\xi_1); A_3\big).$$

Remark 3.2.4 (Possible generalization). Note that we keep g fixed throughout the proof. So if consistency of $\kappa_{3,2|1}$ only holds for a specific g, then we also obtain consistency of $\kappa_{3|2,1}$ only for this specific function g. For consistency of $\kappa_{3|2,1}$ it is therefore sufficient to find a $\kappa_{3,2|1}^g$ that is only consistent w.r.t. g for each g.

Proof. Let $g\colon E_1 \to D$ be a measurable function. By definition of the conditional expectation we need to show for all $A_3 \in \mathcal{E}_3$ and all $A_{1,2} \in \mathcal{E}_1 \otimes \mathcal{E}_2$

$$\mathbb{E}\Big[\mathbf{1}_{A_{1,2}}(\xi_1, \xi_2^{(g(\xi_1))})\kappa_{3|2,1}\big(\xi_1, \xi_2^{(g(\xi_1))}, g(\xi_1); A_3\big)\Big] = \mathbb{E}\big[\mathbf{1}_{A_{1,2}}(\xi_1, \xi_2^{(g(\xi_1))})\mathbf{1}_{A_3}(\xi_3)\big]$$

Without loss of generality we may only consider $A_{1,2} = A_1 \times A_2 \in \mathcal{E}_1 \times \mathcal{E}_2$ since the product sigma algebra $\mathcal{E}_1 \otimes \mathcal{E}_2$ is generated by these rectangles. Since

$$\kappa_{2|1}^g(x_1; A_2) := \kappa_{3,2|1}(x_1, g(x_1); E_3 \times A_2)$$

is a regular conditional version of $\mathbb{P}(\xi_2^{(g(\xi_1))} \in \cdot \mid \xi_1)$ by assumption (3.5) we may apply disintegration[16] to the measurable function

$$\varphi(x_1, x_2) \mapsto \mathbf{1}_{A_2}(x_2)\kappa_{3|2,1}(x_1, x_2, g(x_1); A_3)$$

to obtain

$$\begin{aligned}\mathbb{E}\big[\varphi\big(\xi_1, \xi_2^{g(\xi_1)}\big) \mid \xi_1\big] &\overset{\text{a.s.}}{=} \int \varphi(\xi_1, x_2)\kappa_{2|1}^g(\xi_1; dx_2) \\ &\overset{\text{def.}}{=} \int \varphi(\xi_1, x_2)\kappa_{3,2|1}(\xi_1, g(\xi_1); E_3 \times dx_2).\end{aligned} \tag{3.7}$$

We thereby have

$$\begin{aligned}&\mathbb{E}\Big[\mathbf{1}_{A_1}(\xi_1)\mathbf{1}_{A_2}(\xi_2^{g(\xi_1)})\kappa_{3|2,1}\big(\xi_1, \xi_2^{g(\xi_1)}, g(\xi_1); A_3\big)\Big] \\ &= \mathbb{E}\Big[\mathbf{1}_{A_1}(\xi_1)\varphi\big(\xi_1, \xi_2^{g(\xi_1)}\big)\Big] \\ &\overset{(3.7)}{=} \mathbb{E}\Big[\mathbf{1}_{A_1}(\xi_1)\int \varphi(\xi_1, x_2)\kappa_{3,2|1}(\xi_1, g(\xi_1); E_3 \times dx_2)\Big] \\ &\overset{\text{Lemma 3.2.5}}{=} \mathbb{E}\Big[\mathbf{1}_{A_1}(\xi_1)\kappa_{3,2|1}(\xi_1, g(\xi_1); A_3 \times A_2)\Big] \\ &\overset{(3.5)}{=} \mathbb{E}\Big[\mathbf{1}_{A_1}(\xi_1)\mathbf{1}_{A_3 \times A_2}(\xi_3, \xi_2^{g(\xi_1)})\Big] \\ &= \mathbb{E}\Big[\mathbf{1}_{A_1}(\xi_1)\mathbf{1}_{A_2}(\xi_2^{g(\xi_1)})\mathbf{1}_{A_3}(\xi_3)\Big].\end{aligned}$$

The crucial step is the application of Lemma 3.2.5, which provides an integral representation of a regular conditional distribution of $\xi_3, \xi_2^y \mid \xi_1$ that *couples* the two conditional kernels.

Lemma 3.2.5. *For all* $A_2 \in \mathcal{E}_2$, $A_3 \in \mathcal{E}_3$, *all* $y \in \mathbb{X}$ *and* $\mathbb{P}_{\xi_1}$*-almost all* x_1

$$\begin{aligned}\kappa_{3,2|1}(x_1, y; A_3 \times A_2) &= \int \varphi(x_1, x_2)\kappa_{3,2|1}(x_1, y; E_3 \times dx_2) \\ &= \int \mathbf{1}_{A_2}(x_2)\kappa_{3|2,1}(x_1, x_2, y; A_3)\kappa_{3,2|1}(x_1, y; E_3 \times dx_2).\end{aligned} \tag{3.8}$$

[16] e.g. Kallenberg, 2002, Thm. 6.4.

[17] e.g. Klenke, 2014, chap. 8.

In the remainder of the proof we will show this Lemma. To this end pick any $A_1 \in \mathcal{E}_1$. Then by definition of the conditional expectation[17]

$$\begin{aligned}
&\mathbb{E}[\mathbf{1}_{A_1}(\xi_1)\kappa_{3,2|1}(x_1, y; A_3 \times A_2)] \\
&\overset{(3.5)}{=} \mathbb{E}[\mathbf{1}_{A_1}(\xi_1)\mathbf{1}_{A_2}(\xi_2^y)\mathbf{1}_{A_3}(\xi_3)] \\
&\overset{(3.6)}{=} \mathbb{E}[\mathbf{1}_{A_1}(\xi_1)\mathbf{1}_{A_2}(\xi_2^y)\kappa_{3|2,1}(\xi_1, \xi_2^y; A_3)] \\
&\overset{(*)}{=} \mathbb{E}\left[\mathbf{1}_{A_1}(\xi_1)\int \mathbf{1}_{A_2}(x_2)\kappa_{3|2,1}(\xi_1, x_2, y; A_3)\kappa_{3,2|1}(\xi_1, y; E_3 \times dx_2)\right]
\end{aligned}$$

Note that the constant function $g \equiv y$ is always measurable for the application of (3.5). The last step $(*)$ is implied by disintegration[18]

[18] e.g. Kallenberg, 2002, Thm. 6.4.

$$\mathbb{E}[f(\xi_1, \xi_2^y) \mid \xi_1] \overset{\text{a.s.}}{=} \int f(\xi_1, x_2)\kappa_{2|1}^y(\xi_1; dx_2)$$

of the measurable function $f(x_1, x_2) := \mathbf{1}_{A_2}(x_2)\kappa_{3|2,1}(x_1, x_2; A_3)$ using the probability kernel

$$\kappa_{2|1}^y(x_1; A_2) := \kappa_{3,2|1}(x_1, y; E_3 \times A_2)$$

which is a regular conditional version of $\mathbb{P}(\xi_2^y \in \cdot \mid \xi_1)$ by assumption (3.5), since the constant function $g \equiv y$ is measurable. □

Corollary 3.2.6 (Automatic consistency)**.** *Let Z be a random variable in a standard borel space $(E, \mathcal{B}(E))$, $\mathbf{f}$ a random variable in $C(\mathbb{X}, \mathbb{Y})$. Let W be a random element in an arbitrary measurable space $(\Omega, \mathcal{F})$. If there exists a joint conditional distribution κ for Z given $W, \mathbf{f}(x)$ then κ is automatically consistent. That is, for all $B \in \mathcal{B}(E)$ all $\sigma(W)$ measurable X*

$$\mathbb{P}(Z \in B \mid W, \mathbf{f}(X)) \overset{a.s.}{=} \kappa(W, \mathbf{f}(X), X; B),$$

Proof. Since $X = g(W)$ for some measurable function g, Proposition 3.2.3 with $(\xi_3, \xi_2^y, \xi_1) = (Z, \mathbf{f}(y), W)$ yields the claim, since a consistent joint probability kernel for ξ_3, ξ_2^y given ξ_1 exists by Theorem 3.2.1. □

3.3 Conditionally independent sampling

In this section we will first introduce generalizations to of our main result (Theorem 3.1.6) and then proceed to prove this more general result (Theorem 3.3.2).

Conditional independence Sometimes X_{n+1} is not previsible itself, but sampled from a previsible distribution. That is, a distribution constructed from previously seen evaluations (e.g. Thompson sampling[19]). In this case, X_{n+1} is not previsible, but independent from $\mathbf{f}$ conditional on $\mathcal{F}_n$ denoted by $X_{n+1} \perp\!\!\!\perp_{\mathcal{F}_n} \mathbf{f}$.[20] Note that conditional independence is almost always equivalent to $X_{n+1} = h(\xi, U)$ for a measurable function h, a random (previsible) element ξ that generates $\mathcal{F}_n$ and a standard uniform random variable U independent from $(\mathbf{f}, \mathcal{F}_n)$.[21]

[19] Thompson, 1933.

[20] as introduced in Kallenberg (2002, p. 109).

[21] Kallenberg, 2002, Prop. 6.13.

Noisy evaluations In many optimization applications only noisy evaluations of the random objective function $\mathbf{f}$ at x may be obtained. We associate the noise ς_n to the n-th evaluation x_n, such that the function $\mathbf{f}_n = \mathbf{f} + \varsigma_n$ returns the n-th observation $Y_n = \mathbf{f}_n(x_n)$. While the noise may simply be independent, identically distributed constants, observe that this framework allows for much more general location-dependent noise. The only requirement is that ς_n is a continuous function, such that $\mathbf{f}_n$ is a continuous function.

Definition 3.3.1 (Conditionally independent evolution)**.** The *general conditional independence setting* is given by

- an underlying probability space $(\Omega, \mathcal{A}, \mathbb{P})$,

- a sub-σ-algebra $\mathcal{F}$ (the 'initial information') with W a random element such that $\mathcal{F} = \sigma(W)$,
- A sequence $(\mathbf{f}_n)_{n\in\mathbb{N}_0}$ of continuous random functions in $C(\mathbb{X}, \mathbb{Y})$, where $\mathbb{X}$ is a locally compact, separable metrizable space and $\mathbb{Y}$ a polish space.
- a random variable Z in a standard borel space $(E, \mathcal{B}(E))$ (representing an additional quantity of interest).

A sequence $X = (X_n)_{n\in\mathbb{N}_0}$ of random evaluation locations in $\mathbb{X}$ is called a *conditionally independent evolution*, if $X_{n+1} \perp\!\!\!\perp_{\mathcal{F}_n^X} (Z, (\mathbf{f}_n)_{n\in\mathbb{N}_0})$ for the filtration

$$\mathcal{F}_n^X := \sigma(\mathcal{F}, \mathbf{f}_0(X_0), \dots, \mathbf{f}_n(X_n), X_{[0:n]}) \qquad \text{for } n \geq -1.$$

Theorem 3.3.2 (Conditionally independent sampling)**.** *Assume the general conditional independence setting (Definition 3.3.1).*

(i) Let κ be a joint conditional distribution for Z given $\mathcal{F}, \mathbf{f}_0(x_0), \dots, \mathbf{f}_n(x_n)$, i.e. for all $x_{[0:n]} \in \mathbb{X}^{n+1}$ and $A \in \mathcal{B}(E)$

$$\mathbb{P}\big(Z \in A \mid \mathcal{F}, \mathbf{f}_0(x_0), \dots, \mathbf{f}_n(x_n)\big) \overset{a.s.}{=} \kappa\big(W, \mathbf{f}_0(x_0), \dots, \mathbf{f}_n(x_n), x_{[0:n]}; B\big).$$

Then for all conditionally independent evolutions $(X_k)_{k\in\mathbb{N}_0}$ and $B \in \mathcal{B}(E)$

$$\mathbb{P}\big(Z \in A \mid \mathcal{F}_n^X\big) \overset{a.s.}{=} \kappa\big(W, \mathbf{f}_0(X_0), \dots, \mathbf{f}_n(X_n), X_{[0:n]}; B\big).$$

(ii) Let κ be a joint conditional distribution for $Z, \mathbf{f}_n(x_n) \mid \mathcal{F}, (\mathbf{f}_k(x_k))_{k\in[0:n)}$ such that

$$x_n \mapsto \kappa(y_{[0:n)}, x_{[0:n]}; \,\cdot\,)$$

is continuous in the weak topology for all $x_{[0:n)} \in \mathbb{X}^n$ and all $y_{[0:n)} \in \mathbb{X}^n$.

Then for all conditionally independent evolutions $(X_k)_{k\in\mathbb{N}_0}$ and $B \in \mathcal{B}(\mathbb{Y})$

$$\mathbb{P}\big(Z, \mathbf{f}_n(X_n) \in B \mid \mathcal{F}_{n-1}^X, X_n\big) \overset{a.s.}{=} \kappa\big(W, (\mathbf{f}_k(X_k))_{k\in[0:n)}, X_{[0:n]}; B\big).$$

Remark 3.3.3 (Gaussian case). If $(\mathbf{f}_n)_{n\in\mathbb{N}_0}$ is a sequence of continuous, joint Gaussian random functions, then the same procedure as in Example 3.1.7 leads to a consistent joint conditional distribution in the sense of Theorem 3.3.2.

The proof of (i) will follow from repeated applications of the following lemma. (ii) will follow from (i) and an application of Theorem 3.2.1.

Lemma 3.3.4 (Consistency allows conditional independence)**.** *Let Z be a random variable in a standard borel space $(E, \mathcal{B}(E))$, $\mathbf{f}$ a continuous random function in $C(\mathbb{X}, \mathbb{Y})$. Let W be a random element in an arbitrary measurable space $(\Omega, \mathcal{F})$. If there exists a joint conditional distribution κ for Z given $W, \mathbf{f}(x)$ then for any random variable $X \perp\!\!\!\perp_W (Z, \mathbf{f})$ in $\mathbb{X}$*

$$\mathbb{P}(Z \in B \mid W, X, \mathbf{f}(X)) = \kappa(W, \mathbf{f}(X), X; B).$$

Proof. Observe that X is clearly measurable with respect to $W^+ := (W, X)$. Our proof strategy therefore relies on constructing a joint conditional distribution for Z given $(W^+, \mathbf{f}(x))$ using κ and apply Corollary 3.2.6.

Since X independent from $(Z, \mathbf{f})$ conditional on W there exists a standard uniform $U \sim \mathcal{U}(0, 1)$ independent from $(W, Z, \mathbf{f})$ such that $X = h(W, U)$ for some measurable function h.[22] Since U is independent from $W, Z, \mathbf{f}$ we have by[23]

$$\mathbb{P}(Z \in B \mid W, U, \mathbf{f}(x)) = \mathbb{P}(Z \in B \mid W, \mathbf{f}(x)) = \kappa(W, \mathbf{f}(x), x; B)$$

for all $x \in \mathbb{X}$. Since $\sigma(W, X, \mathbf{f}(x)) \subseteq \sigma(W, U, \mathbf{f}(x))$ and $(W, \mathbf{f}(x))$ is measurable with respect to $\sigma(W, X, \mathbf{f}(x))$ we therefore have

$$\mathbb{P}(Z \in B \mid \underbrace{W, X}_{=W^+}, \mathbf{f}(x)) = \kappa(W, \mathbf{f}(x), x; B) =: \kappa^+(\underbrace{W, X}_{=W^+}, \mathbf{f}(x), x; B),$$

where κ^+ is defined as constant in the second input. An application of Corollary 3.2.6 to κ^+ yields the claim. □

[22] Kallenberg, 2002, Prop. 6.13.

[23] Kallenberg, 2002, Prop. 6.6.

Proof of Theorem 3.3.2. We will prove (i) by induction over $k \in \{0, \dots, n+1\}$. For any conditionally independent evolution $(X_n)_{n\in\mathbb{N}_0}$, the induction claim is

$$\begin{aligned}\mathbb{P}\big(Z \in A \mid W, (\mathbf{f}_i(X_i))_{i\in[0:k)}, (\mathbf{f}_i(x_i))_{i\in[k:n]}, X_{[0:k)}\big) \\ = \kappa\big(W, (\mathbf{f}_i(X_i))_{i\in[0:k)}, (\mathbf{f}_i(x_i))_{i\in[k:n]}, X_{[0:k)}, x_{[k:n]}; A\big).\end{aligned} \tag{3.9}$$

The induction start with $k = 0$ is given by assumption and $k = n+1$ is the claim, so we only need to show the induction step $k \to k+1$. For this purpose we want to define $\tilde{W} = \big(W, (\mathbf{f}_i(X_i))_{i\in[0:k)}, (\mathbf{f}_i(x_i))_{i\in(k:n]}, X_{[0:k)}\big)$ and the kernel

$$\tilde{\kappa}_{x_{(k:n]}}\big(\tilde{W}, \mathbf{f}_k(x_k), x_k; A\big) := \kappa\big(W, (\mathbf{f}_i(X_i))_{i\in[0:k)}, (\mathbf{f}_i(x_i))_{i\in[k:n]}, X_{[0:k)}, x_{[k:n]}; A\big),$$

which is formally defined for any fixed $x_{(k:n]}$ by mapping the elements of $\tilde{W}$ into the right position. By induction (3.9) we thereby have

$$\mathbb{P}(Z \in A \mid \tilde{W}, \mathbf{f}(x_k)) = \tilde{\kappa}_{x_{(k:n]}}(\tilde{W}, \mathbf{f}(x_k), x_k; A).$$

We can thereby finish the induction using Lemma 3.3.4 if we can prove X_k is independent from $(Z, \mathbf{f})$ conditional on $\tilde{W}$. For this we will use the characterization of conditional independence in Proposition 6.13 by Kallenberg (2002).

Since X_k is independent from $(Z, \mathbf{f})$ conditionally on $\mathcal{F}_{k-1}$ there exists, by this Proposition, a uniform random variable $U \sim \mathcal{U}(0,1)$ independent from $(Z, \mathbf{f}, \mathcal{F}_{k-1})$ such that $X_k = h(\xi, U)$ for some measurable function h and a random element ξ that generates $\mathcal{F}_{k-1}$. Due to $\mathcal{F}_{k-1} \subseteq \sigma(\tilde{W})$ the element ξ is a measurable function of $\tilde{W}$ and therefore $X_k = \tilde{h}(\tilde{W}, U)$ for some measurable function $\tilde{h}$. Since U is independent from $(Z, \mathbf{f}, \mathcal{F}_{k-1})$, it is independent from $\tilde{W}$ as $\sigma(\tilde{W}) \subseteq \sigma(\mathbf{f}, \mathcal{F}_{k-1})$. Using Prop. 6.13 from Kallenberg (2002) again, X_k is thereby independent from $(Z, \mathbf{f})$ conditional on $\tilde{W}$.

What remains is the proof of (ii). Let x_n be fixed and define $\tilde{Z} = (Z, \mathbf{f}_n(x_n))$. Since κ is a joint conditional distribution for $Z, \mathbf{f}_n(x_n)$ given $\mathcal{F}, (\mathbf{f}_k(x_k))_{k\in[0:n)}$ the kernel

$$\kappa_{x_n}(y_{[0:n)}, x_{[0:n)}; B) := \kappa(y_{[0:n)}, x_{[0:n]}; B)$$

clearly satisfies the requirements of (i) and thereby

$$\mathbb{P}\big(Z, \mathbf{f}(x_n) \in B \mid \mathcal{F}^X_{n-1}\big) = \kappa\big(\underbrace{W, (\mathbf{f}_k(X_k))_{k\in[0:n)}, X_{[0:n)}}_{=:\tilde{W}}, x_n; B\big) =: \tilde{\kappa}(\tilde{W}, x_n; B).$$

Since $\tilde{W}$ generates $\mathcal{F}^X_{n-1}$ we are almost in the setting of Theorem 3.2.1, as we have continuity in x_n. However, since X_n is not previsible we have to repeat the same trick we used in the proof of Lemma 3.3.4. Namely, X_n is measurable with respect to $W^+ := (\tilde{W}, X)$ and we will have to construct a joint conditional distribution for $Z, \mathbf{f}(x_n)$ given W^+.

Since X_n is independent from $Z, \mathbf{f}$ conditional on $\tilde{W}$, there exists a standard uniform $U \sim \mathcal{U}(0,1)$ independent from $(\tilde{W}, Z, \mathbf{f})$ such that $X_n = h(\tilde{W}, U)$ for some measurable function h.[24] Since U is independent from $(\tilde{W}, Z, \mathbf{f})$, we have by[25]

$$\mathbb{P}(Z, \mathbf{f}(x_n) \in B \mid \tilde{W}, U) \overset{\text{a.s.}}{=} \mathbb{P}(Z, \mathbf{f}(x_n) \in B \mid \tilde{W}) \overset{\text{a.s.}}{=} \tilde{\kappa}(\tilde{W}, x_n; B)$$

for all $x_n \in \mathbb{X}$. Since $\sigma(\tilde{W}, X_n) \subseteq \sigma(\tilde{W}, U)$ and $\tilde{W}$ is measurable with respect to $\sigma(\tilde{W}, X)$ we therefore have

$$\mathbb{P}(Z, \mathbf{f}(x_n) \in B \mid \underbrace{\tilde{W}, X_n}_{=W^+}) \overset{\text{a.s.}}{=} \tilde{\kappa}(\tilde{W}, x; B) =: \kappa^+(\underbrace{\tilde{W}, X_n}_{=W^+}, x_n; B),$$

where κ^+ is defined as constant in the second input. Clearly, by definition of κ^+ via κ, κ^+ is continuous in x_n and as a continuous joint conditional distribution it is consistent by Theorem 3.2.1. This finally implies the claim

$$\begin{aligned}\mathbb{P}(Z, \mathbf{f}(X_n) \in B \mid \underbrace{\mathcal{F}_{n-1}, X_n}_{=W^+}) &\overset{\text{a.s.}}{=} \kappa^+(W^+, X_n; B) \\ &= \kappa(W, (\mathbf{f}_k(X_k))_{k\in[0:n)}, X_{[0:n)}, X_n; B).\end{aligned}$$ □

[24] Kallenberg, 2002, Prop. 6.13.

[25] Kallenberg, 2002, Prop. 6.6.

3.4 Topological foundation

In this section we show the evaluation function to be continuous and therefore measurable for continuous random functions. For compact $\mathbb{X}$ this result can be collected from various sources.[26] But we could not find a reference for the result in this generality, so we provide a proof.

Theorem 3.4.1 (Continuous functions)**.** *Let $\mathbb{X}$ be a locally compact, separable and metrizable space*[27]*, $\mathbb{Y}$ a polish space and $C(\mathbb{X}, \mathbb{Y})$ the space of continuous functions equipped with the* compact-open[28] *topology. Then*

(i) *the evaluation function*

$$e\colon \begin{cases} C(\mathbb{X},\mathbb{Y}) \times \mathbb{X} \to \mathbb{Y} \\ (f,x) \mapsto f(x) \end{cases}$$

is continuous and therefore measurable.

(ii) *$C(\mathbb{X}, \mathbb{Y})$ is a **polish space**, whose topology is generated by the metric*

$$d(f,g) := \sum_{n=1}^{\infty} 2^{-n} \frac{d_n(f,g)}{1+d_n(f,g)} \quad \text{with} \quad d_n(f,g) := \sup_{x \in K_n} d_{\mathbb{Y}}(f(x),g(x))$$

for any metric $d_{\mathbb{Y}}$ that generates the topology of $\mathbb{Y}$ and any compact exhaustion[29] *$(K_n)_{n\in\mathbb{N}}$ of $\mathbb{X}$, that always exists because $\mathbb{X}$ is hemicompact!*

(iii) *The Borel σ-algebra of $C(\mathbb{X}, \mathbb{Y})$ is equal to the restriction of the product sigma algebra of $\mathbb{Y}^{\mathbb{X}}$ to $C(\mathbb{X}, \mathbb{Y})$, i.e. $\mathcal{B}(C(\mathbb{X}, \mathbb{Y})) = \mathcal{B}(\mathbb{Y})^{\otimes\mathbb{X}}|_{C(\mathbb{X},\mathbb{Y})}$.*

Remark 3.4.2 (Topology of pointwise convergence). The topology of point-wise convergence ensures that all projection mappings $\pi_x(f) = f(x)$ are continuous. It coincides with the product topology.[30] Thm. 3.4.1 (iii) ensures that the Borel-σ-algebra generated by the topology of point-wise convergence coincides with the Borel σ-algebra generated by the compact-open topology.

Remark 3.4.3 (Construction). The main tool for the construction of probability measures, Kolmogorov's extension theorem,[31] allows for the construction of random measures on product spaces. This is only compatible with the product topology, i.e. the topology of point-wise convergence. But the evaluation map is generally not continuous with respect to this topology.[32] (iii) ensures that this does not pose a problem as long as $\mathbb{X}$ and $\mathbb{Y}$ satisfy the requirements of Theorem 3.4.1 and the constructed random process has a continuous version[33].

Remark 3.4.4 (Limitations). While the compact-open topology can be defined for general topological spaces, the continuity of the evaluation map crucially depends on $\mathbb{X}$ being locally compact.[34] For $\mathbb{X}$ and $\mathbb{Y}$ polish spaces, this implies $C(\mathbb{X}, \mathbb{Y})$ is generally only well behaved if $\mathbb{X}$ is locally compact.

Remark 3.4.5 (Discontinuous case). Without continuity it is already difficult to obtain a random function $\mathbf{f}$ that is almost surely measurable and can be evaluated point-wise. The construction of Lévy processes in càdlàg[35] space only works on ordered domains such as $\mathbb{R}$, where 'right-continuous' has meaning. Typically, discontinuous random functions are therefore only constructed as generalized functions in the sense of distributions[36] that cannot be evaluated point-wise.[37] In particular, we cannot hope to evaluate generalized random functions at random locations.

Proof. Since $\mathbb{X}$ is locally compact, (i) follows from Proposition 2.6.11 and Theorem 3.4.3. by Engelking (1989).

For (ii) let us begin to show that $\mathbb{X}$ is **hemicompact/exhaustible by compact sets**. Since the space $\mathbb{X}$ is locally compact, pick a compact neighborhood for every point. The interiors of these compact neighborhoods obviously cover $\mathbb{X}$. Since every regular, second countable space[27] is Lindelöf,[38] we can pick a countable subcover,

[26] e.g. Engelking (1989, Thm. 4.2.17) and Kechris (1995, Thm. 4.19)

[27] technically, we do not need $\mathbb{X}$ to be metrizable but only regular and second countable, which is equivalent to separability in metrizable spaces (Engelking, 1989, Cor. 4.1.16). With this definition it is more obvious that a locally compact polish space satisfies the requirements, but we will assume the more general setting in the proof.

[28] For $K \subseteq \mathbb{X}$ compact and $U \subseteq \mathbb{Y}$ open, the sets

$$M(K,U) := \{f \in C(\mathbb{X},\mathbb{Y}) : f(K) \subseteq U\}$$

form a sub-base of the *compact-open* topology (e.g. Engelking, 1989, Sec. 3.4). I.e. the compact-open topology it is the smallest topology such that all $M(K,U)$ are open. Recall that the set of finite intersections of a sub-base form a base of the topology and elements from the topology can be expressed as unions of base elements.

[29] The set $\mathbb{X}$ is *hemicompact* if it can be *exhausted by the compact sets* $(K_n)_{n\in\mathbb{N}}$, which means that the compact set K_n is contained in the interior of K_{n+1} for any n and

$$\mathbb{X} = \bigcup_{n\in\mathbb{N}} K_n.$$

[30] Engelking, 1989, Prop. 2.6.3.

[31] e.g. Klenke, 2014, Sec. 14.3.

[32] Engelking, 1989, Prop. 2.6.11.

[33] cf. Talagrand (1987, Thm. 3), Costa et al. (2024) and references therein

[34] Engelking, 1989, Thm. 3.4.3 and comments below.

[35] french: continue à droite, limite à gauche, "right-continuous with left-limits"

[36] The set of distributions is defined as the topological dual to a set of test functions. In particular, distributions are *continuous* linear functionals acting on the test functions. Thereby one may hope that Theorem 3.4.1 is applicable, but the set of test functions is typically not locally compact (cf. Remark 3.4.4).

[37] e.g. Schäffler, 2018.

[38] Engelking, 1989, Thm. 3.8.1.

such that the interiors of the sequence $(C_i)_{i\in\mathbb{N}}$ of compact sets cover the domain $\mathbb{X}$. We inductively define a compact exhaustion $(K_n)_{n\in\mathbb{N}}$ with $K_1 := C_1$. Observe that the set K_n is covered by the interiors $(\operatorname{int} C_i)_{i\in\mathbb{N}}$. Since K_n is compact, we can choose a finite sub-cover $(\operatorname{int} C_i)_{i\in I}$ and define $K_{n+1} := \bigcup_{i\in I} C_i \cup C_{n+1}$. Then by definition K_n is contained in the interior of the compact set K_{n+1} and due to $C_n \subseteq K_n$ this sequence also covers the space $\mathbb{X}$ and is thereby a compact exhaustion.

It is straightforward to check that the metric defined in (ii) is a metric, so we will only prove this **metric induces the compact-open** topology.

(I) **The compact-open topology is a subset of the metric topology.** We need to show that the sets $M(K,U)$ are open with respect to the metric. This requires for any $f \in M(K,U)$ an $\epsilon > 0$ such that the epsilon ball $B_\epsilon(f)$ is contained in $M(K,U)$.

We start by constructing a finite cover of $f(K)$. For any $x \in K$ there exists $\delta_x > 0$ with $B_{2\delta_x}(f(x)) \subseteq U$ for balls induced by the metric $d_{\mathbb{Y}}$ as U is open. Since K is compact, $f(K) \subseteq U$ is a compact set covered by the balls $B_{\delta_x}(f(x))$. This yields a finite subcover $B_{\delta_1}(f(x_1)), \dots, B_{\delta_m}(f(x_m))$ of $f(K)$.

Using this cover we will prove the following criterion: Any $g \in C(\mathbb{X},\mathbb{Y})$ is in $M(K,U)$ if

$$\sup_{x\in K} d_{\mathbb{Y}}(f(x),g(x)) < \delta := \min\{\delta_1,\dots,\delta_m\}. \tag{3.10}$$

For this criterion note that for any $x \in K$ there exists $i \in \{1,\dots,m\}$ such that $f(x) \in B_{\delta_i}(f(x_i))$. This implies

$$d(g(x),f(x_i)) \le d(g(x),f(x)) + d(f(x),f(x_i)) \le 2\delta_i,$$

which implies $g(K) \subseteq \bigcup_{i=1}^m B_{2\delta_i}(f(x_i)) \subseteq U$ and therefore $g \in M(K,U)$.

Consequently, if there exists $\epsilon > 0$ such that $g \in B_\epsilon(f)$ implies criterion (3.10), then we have $B_\epsilon(f) \subseteq M(K,U)$ which finishes the proof. And this is what we will show. Since K is compact and the interiors of K_n cover the space, there exists a finite sub-cover $K \subseteq \bigcup_{i\in I} K_i$ and therefore some $m = \max I$ such that K is in the interior of K_m. By definition of d_m it is thus clearly sufficient to ensure $d_m(f,g) < \delta$. And since $\varphi(x) = \frac{x}{1+x}$ is a strict monotonous function $\epsilon := 2^{-m}\varphi(\delta)$ does the job, since $2^{-m}\varphi(d_m(f,g)) \le d(f,g) \le \epsilon$ implies $d_m(f,g) \le \delta$.

(II) **The metric topology is a subset of the compact-open topology.** Since the balls $B_\epsilon(f)$ form a base of the metric topology it is sufficient to prove them open in the compact-open topology. If for any $g \in B_\epsilon(f)$ there exists a compact $C_1,\dots,C_m \subseteq \mathbb{X}$ and open $U_1,\dots,U_m \subseteq \mathbb{Y}$ such that $g \in \bigcap_{j=1}^m M(C_j,U_j) \subseteq B_\epsilon(f)$, then the ball is open since these finite intersections are open sets in the compact-open topology and their union over g remains open. But since there exists $r > 0$ such that $B_r(g) \subseteq B_\epsilon(f)$, it is sufficient to prove for any $r > 0$ that there exist compact C_j and open U_j such that

$$g \in V := \textstyle\bigcap_{j=1}^m M(C_j,U_j) \subseteq B_r(g). \tag{3.11}$$

For this purpose pick K_N from the compact exhaustion with sufficiently large N such that $2^{-N} < \frac{r}{2}$. Pick a finite cover $O_{x_1},\dots O_{x_m}$ of K_N from the cover $\{O_x\}_{x\in K_N}$ with $O_x := g^{-1}(B_{r/5}(g(x)))$ and define the sets

$$C_j := \overline{O_{x_j}} \cap K_N \qquad U_j := B_{r/4}(f(x_j)).$$

Clearly the U_j are open and the C_j are compact and we will now prove they satisfy (3.11). Observe that $g \in V$ since for all j

$$g(C_j) \subseteq g(\overline{O_{x_j}}) \subseteq \overline{B_{r/5}(f(x_j))} \subseteq U_j.$$

Pick any other $h \in V$. Then for all $x \in K_N$ there exists i such that $x \in O_{x_i} \subseteq C_i$ and by definition of V this implies $h(x) \in U_i$ and also $g(x) \in U_i$ and thereby

$d_{\mathbb{Y}}(h(x), g(x)) \leq r/2$. This uniform bound implies $d_N(h, g) \leq r/2$ and therefore

$$d(g,h) \leq \Big(\sum_{n=1}^{N} 2^{-n} d_n(g,h)\Big) + \Big(\sum_{n=N+1}^{\infty} 2^{-n}\Big) \leq d_N(g,h) + 2^{-N} < r,$$

since $d_n(f, g) \leq d_N(f, g)$ for $n \leq N$. Thus $h \in B_r(g)$ which proves (3.11).

As $C(\mathbb{X}, \mathbb{Y})$ is clearly metrizable, what is left to prove are its separability and completeness. Separability could be proven directly similarly to the proof of Theorem 4.19 in Kechris (1995) but for the sake of brevity this result follows from Theorem 3.4.16 and Theorem 4.1.15 (vii) by Engelking (1989) and the fact that $\mathbb{X}$ and $\mathbb{Y}$ are second countable. Completeness follows from the fact that any Cauchy sequence f_n induces a Cauchy sequence $f_n(x)$ for any x by definition of the metric. And by completeness of $\mathbb{Y}$ there must exist a limiting value $f(x)$ for any x. The continuity of f follows from the uniform convergence on compact sets, since every compact set is contained in some K_n from the compact exhaustion (cf. last paragraph in (I)).

What is left to prove is (iii). Since the projections are continuous with respect to the compact open topology, they are measurable with respect to the Borel-σ-algebra. The product sigma algebra, which is the smallest sigma algebra to ensure all projections are measuralbe, restricted to the continuous functions is therefore a subset of the Borel σ-algebra. To prove the opposite inclusion, we need to show that the open sets are contained in the product σ-algebra. Since the space is second countable[39] and every open set thereby a countable union of its base, it is sufficient to check that the open ball $B_\epsilon(f_0)$ for $\epsilon > 0$ and $f_0 \in C(\mathbb{X}, \mathbb{Y})$ is in the product sigma algebra restricted to $C(\mathbb{X}, \mathbb{Y})$. But since $B_\epsilon(f_0) = H^{-1}([0, \epsilon))$ with $H(f) := d(f, f_0)$, it is sufficient to prove H is $\sigma(\pi_x : x \in \mathbb{X})$-$\mathcal{B}(\mathbb{R})$-measurable, where π_x are the projections. H is measurable if $H_n(f) = d_n(f, f_0)$ is measurable, as a limit, sum, etc.[40] of measurable functions. But since $\mathbb{X}$ is separable,[41] i.e. has a countable dense subset Q, we have by continuity of f and f_0

$$d_n(f, f_0) = \sup_{x \in K_n} d_{\mathbb{Y}}(f(x), f_0(x)) = \sup_{x \in K_n \cap Q} d_{\mathbb{Y}}(\pi_x(f), \pi_x(f_0)).$$

Since $d_{\mathbb{Y}}$ is continuous and thereby measurable,[42] H_n is measurable as a countable supremum of measurable functions.[43] □

[39] Engelking, 1989, Cor. 4.1.16.

[40] Klenke, 2014, Thm. 1.88-1.92.

[41] Engelking, 1989, Cor. 1.3.8.

[42] Klenke, 2014, Thm. 1.88.

[43] Klenke, 2014, Thm. 1.92.

References

Aliprantis, Charalambos D. and Kim C. Border (2006). *Infinite Dimensional Analysis*. 3rd ed. Berlin/Heidelberg: Springer-Verlag. ISBN: 978-3-540-29586-0. DOI: 10.1007/3-540-29587-9.

Costa, Nathaël Da et al. (2024). *Sample Path Regularity of Gaussian Processes from the Covariance Kernel*. DOI: 10.48550/arXiv.2312.14886. arXiv: 2312.14886 [cs]. Pre-published.

Eaton, Morris L. (2007). *Multivariate Statistics: A Vector Space Approach*. Ed. by R. A. Vitale. Vol. 53. Lecture Notes-Monograph Series. Beachwood, Ohio, USA: Institute of Mathematical Statistics. 519 pp. ISBN: 978-0-940600-69-0. DOI: 10.1214/lnms/1196285102.

Engelking, Ryszard (1989). *General Topology: Revised and Completed Edition*. Berlin: Heldermann Verlag. 540 pp. ISBN: 978-3-88538-006-1.

Kallenberg, Olav (2002). *Foundations of Modern Probability*. Red. by J. Gani, C. C. Heyde, and T. G. Kurtz. Probability and Its Applications. New York, NY: Springer. ISBN: 978-1-4419-2949-5 978-1-4757-4015-8. DOI: 10.1007/978-1-4757-4015-8.

Kechris, Alexander S. (1995). *Classical Descriptive Set Theory*. Vol. 156. Graduate Texts in Mathematics. New York, NY: Springer. ISBN: 978-1-4612-8692-9 978-1-4612-4190-4. DOI: 10.1007/978-1-4612-4190-4.

Klenke, Achim (2014). *Probability Theory: A Comprehensive Course*. Universitext. London: Springer. ISBN: 978-1-4471-5360-3 978-1-4471-5361-0. DOI: 10.1007/978-1-4471-5361-0.

Schäffler, Stefan (2018). *Generalized Stochastic Processes*. Compact Textbooks in Mathematics. Cham: Springer International Publishing. ISBN: 978-3-319-78767-1 978-3-319-78768-8. DOI: 10.1007/978-3-319-78768-8.

Srinivas, Niranjan et al. (2012). "Information-Theoretic Regret Bounds for Gaussian Process Optimization in the Bandit Setting". In: *IEEE Transactions on Information Theory* 58.5, pp. 3250–3265. ISSN: 1557-9654. DOI: 10.1109/TIT.2011.2182033.

Talagrand, Michel (1987). "Regularity of Gaussian Processes". In: *Acta Mathematica* 159 (none), pp. 99–149. ISSN: 0001-5962, 1871-2509. DOI: 10.1007/BF02392556.

Thompson, William R. (1933). "On the Likelihood That One Unknown Probability Exceeds Another in View of the Evidence of Two Samples". In: *Biometrika* 25.3–4, pp. 285–294. URL: https://academic.oup.com/biomet/article-pdf/25/3--4/285/513725/25--3--4--285.pdf (visited on 2025-03-05).

Chapter 4

Distributions over functions

To perform Bayesian optimization it is necessary to choose a prior distribution over objective functions or at least a family of prior distributions.

A minimal assumption is that every continuous function lies in the support of the objective function distribution. A slightly weaker form of such a *universality*[1] assumption will be used in Chapter 8.

But for most applications, such a minimal assumption leaves a family of distributions that is way too big to make useful statements. In this section we therefore introduce uniformity assumptions about the distribution of the objective function as an effective way to shrink the set of prior distributions to a manageable level.

A uniformity assumption can be justified in two ways: First, if we do not have any information about the distribution it seems like a reasonable zero-knowledge prior to weigh every possible input equally. And second, it may sometimes be possible to enforce this uniformity, for example via a preprocessing step (cf. Remark 2.1.2).

But what *is* a uniform prior in the case of functions? We will present two different approaches: The first will motivate the Gaussian assumption by infinite divisibility (Section 4.1). The second motivates the stationary, isotropy assumption with general input invariance (Section 4.2). In other words, we will motivate *stationary, isotropic Gaussian random functions* as a zero-knowledge prior. Unfortunately this assumptions does not result in a single distribution but in a family of distributions. In Section 4.3 we thus present characterizations of various input-invariant families. The remaining sections are dedicated to the proof of the characterization of the non-stationary covariance kernels. These have not been characterized so far and are motivated by the fact that objective functions in machine learning are typically not stationary but may be isotropic cf. Section 7.4.

[1] Universality is typically defined as 'The Reproducing kernel hilbertspace (RKHS) lies dense in the continuous functions' (e.g. Micchelli et al., 2006; Carmeli et al., 2010), but since the RKHS generally lies dense in the support of a random function (Bogachev, 1998, Thm. 3.6.1) this definition implies this more intuitive concept.

4.1 Infinite divisibility

A naive approach to build a uniform prior over functions is to sample every value $N(x)$ of a function N independent and identically distributed. Since optimization requires some continuity one may want to smooth this white noise, perhaps with a Gaussian blur. Convolution with a mollifier such as the Gaussian blur is integration and integration requires measurability; but unfortunately, the naive white noise is almost surely *not* measurable. A more fruitful approach is to construct white noise as a generalized function with measurability in mind. Accordingly, consider what properties the integrals

$$W(A) = \int \mathbf{1}_A(x) N(x) dx$$

should have if they were well defined. We can then take these properties as definition for the measure W, which is a generalized function.[2] Since the values $N(x)$ are iid we would expect:

(a) $W(A)$ and $W(B)$ to be independent for disjoint A and B (by independence),

[2] Observe that if we define W, then the noise at a certain point $N(x)$ is heuristically equal to the limit $\lim_{\epsilon \to 0} \frac{W(\mathrm{B}_\epsilon(x))}{|\mathrm{B}_\epsilon(x)|}$, where $\mathrm{B}_\epsilon(x)$ is an ϵ ball around x and $|\mathrm{B}_\epsilon(x)|$ its Lebesgue measure (volume). Of course this limit does not actually exist, since N is discontinuous. But this conveys the intuition why a measure can generalize a function (see 'generalized function' or 'Schwarz distribution').

(b) $W(A \cup B) = W(A) + W(B)$ for disjoint A and B (by linearity of the integral),

(c) $W(A)$ to be equal in distribution to $W(B)$ if the volume of A is equal to that of B (since the $N(x)$ are identically distributed).

Observe that any set A of non-zero volume can be subdivided into n equally sized disjoint parts, which are then iid distributed by (a) and (c). Hence, the summability in (b) implies the distribution of $W(A)$ is infinitely divisible. The Lévy-Khintchine representation Theorem[3] then characterizes the possible distributions of $W(A)$.

Gaussian white noise,[4] characterized by $W(A) \sim \mathcal{N}(0, |A|)$ in place of (c), may still be an arbitrary selection from the general set of Lévy white noise[5] (see also 'Lévy random field'); but – while the Gaussian assumption was a forgone conclusion for practical reasons (cf. Section 2.2.3) – the fact that it is picked from a reduced, plausible set of infinitely divisible distributions may alleviate some concern.

[3] e.g. Ken-Iti, 1999, Theorem 8.1.

[4] e.g. Adler and Taylor, 2007, Section 1.4.3 and Section 5.3.

[5] e.g. Fageot and Humeau, 2021.

Smoothing noise Before we move on, let us quickly touch on two prominent smoothing approaches of white noise to obtain distributions over continuous functions.

Example 4.1.1 (Brownian motion & Lévy process)**.** Let W be Gaussian white noise on $[0, \infty)$, then the integrated white noise $B(t) := W([0, t))$ is the Brownian motion.[6] For general Lévy noise W defined on $[0, \infty)$, $W([0, t))$ is a Lévy process which the the Lévy-Itô decomposition breaks up into, drift, Brownian motion and a jump process.[7] In particular the assumption of continuity eliminates all distributions except for the Gaussian one.

[6] see e.g. Adler and Taylor (2007, Sec. 1.4.3) including a generalization to 'Brownian sheets' on $[0, \infty)^d$.

[7] e.g. Applebaum, 2009, Theorem 2.4.16.

Example 4.1.2 (Convolution with mollifier/Blur)**.** Let W be Lévy white noise and assume $\mathbb{E}[W(A)] = 0$ and $\mathbb{E}[W(A)^2] = |A|$, where $|A|$ is the Lebesgue measure of A. Then integrals and thereby convolutions with mollifiers are well defined[8]

$$\mathbf{f}(t) = k * W(t) = \int k(t - x) W(dx).$$

[8] Adler and Taylor, 2007, Sec. 5.2.

The resulting random function $\mathbf{f}$ is Gaussian if W is Gaussian noise. $\mathbf{f}$ is further always centered with covariance[9]

$$\mathbb{E}[\mathbf{f}(t)\mathbf{f}(s)] = \int k(t - x)k(s - x)dx = \int k((t - s) - y)k(y)dy =: C(t - s).$$

[9] Adler and Taylor, 2007, Lemma 5.3.1.

If we choose the Gaussian blur $k(x) = \exp(-\frac{\|x\|^2}{s^2})$ as the mollifier, then the covariance is given by

$$\begin{aligned}
\mathbb{E}[\mathbf{f}(t)\mathbf{f}(s)] &= \int \exp\Big(-\frac{\|t - x\|^2 + \|s - x\|^2}{s^2}\Big) dx \\
&\overset{y = x - \frac{s+t}{2}}{=} \int \exp\Big(-\frac{\|\frac{t-s}{2} - y\|^2 + \|\frac{s-t}{2} - y\|^2}{s^2}\Big) dy \\
&= \int \exp\Big(-\frac{2\|\frac{t-s}{2}\|^2 - 2\langle (t - s) + (s - t), y\rangle + 2\|y\|^2}{s^2}\Big) dy \\
&= e^{-\frac{\|t-s\|^2}{2s^2}} \int e^{-\frac{\|y\|^2}{2(s^2/4)}} dy \\
&= \sigma^2 \exp\Big(-\frac{\|t - s\|^2}{2s^2}\Big),
\end{aligned}$$

for an appropriate choice of σ^2 that may be changed by scaling of the mollifier k. This covariance, known as the 'Squared exponential', 'Radial Basis Function' or 'Gaussian' kernel, plays a crucial role in the characterization of stationary isotropic covariance kernels (cf. Section 4.3).

Remark 4.1.3 (Representation). The 'spectral representation theorem'[10] gives a characterizing result in the opposite direction: Any continuous stationary complex random function $\mathbf{f}$ may be obtained by application of the mollifier $k(t, x) = e^{i\langle t, x\rangle}$ to complex ν-noise, where ν is a intensity measure replacing the Lebesgue measure in the definition of white noise.

[10] Adler and Taylor, 2007, Theorem 5.4.2.

4.2 Input invariance

A different approach to get something akin to a uniform distribution on the space of measures, is to consider axiomatic invariance properties. On subsets of $\mathbb{R}^d$ with non-zero volume, the uniform distribution is defined as the restriction of the Lebesgue measure to these sets. The Lebesgue measure in turn is characterized by its translation invariance which encapsulates the notion that any two points are alike.[11] Translations in function space is however the wrong operations to demand invariance from:

1. There does not exist a non-zero translation invariant measure in infinite dimensional banach spaces.[12]
2. Translation in function space is an extremely noisy operation which completely changes the landscape of the function, because translation entails the addition of an arbitrarily rough function.

Invariance with respect to transformations of the input to the function are much more natural however. If these transformations are bijections, the minimizer found in the transformed space can be translated back into a minimizer of the original function and these transformations are thus also amendable to preprocessing.[13]

Definition 4.2.1 (Distributional Φ-invariance)**.** We say a random function $\mathbf{f} = (\mathbf{f}(x))_{x\in\mathbb{X}}$ is (distributionally) Φ-invariant to a set of transformations $\phi\colon \mathbb{X} \to \mathbb{X}$, typically a group, if[14]

$$\mathbb{P}_{\mathbf{f}\circ\phi} = \mathbb{P}_{\mathbf{f}} \qquad \forall \phi \in \Phi.$$

There are a few prominent examples for $\mathbb{X} = \mathbb{R}^d$:

- If Φ is the set of *translations* $T(d)$, we call $\mathbf{f}$ **stationary**.
- If Φ is the set of *linear isometries*, i.e. the orthogonal group $O(d)$, we call $\mathbf{f}$ (non-stationary) **isotropic**.
- If Φ is the set of (Euclidean) *isometries* $E(d)$, we call $\mathbf{f}$ **stationary isotropic**.

We further say a random function $\mathbf{f}$ is n-weakly Φ-invariant, if for all $\phi \in \Phi$, all $k \leq n$ and all x_i

$$\mathbb{E}[\mathbf{f}(\phi(x_1)) \cdot \cdots \cdot \mathbf{f}(\phi(x_k))] = \mathbb{E}[\mathbf{f}(x_1) \cdot \cdots \cdot \mathbf{f}(x_k)].$$

Since second moments fully determine Gaussian distributions, 2-weakly invariance is special, because it is equivalent to full input invariance in the Gaussian case. So an omitted n equals 2.

Remark 4.2.2 (Almost sure invariance). We defined a distributional invariance. The function $\mathbf{f}$ itself is typically not input invariant to Φ. Almost sure Φ-invariance, i.e. $\mathbf{f}(\phi(x)) = \mathbf{f}(x)$ almost surely, results in n-weak invariance of the form

$$\mathbb{E}[\mathbf{f}(\phi_1(x_1)) \cdot \cdots \cdot \mathbf{f}(\phi_2(x_k))] = \mathbb{E}[\mathbf{f}(x_1) \cdot \cdots \cdot \mathbf{f}(x_k)].$$

for all selections of $\phi_i \in \Phi$ and $x_i \in \mathbb{X}$ and $k \leq n$. Since the distributional invariance we introduced only allows for the selection of a single transformation ϕ in its weak form, some authors refer to distributional invariance as 'simultaneous invariance' or 'total invariance'.[15]

Different notions of weak invariance have simple characterizations in terms of the mean and covariance functions.[16] We present these in Theorem 4.2.3 of which the stationary isotropic and stationary case are already well known.

Theorem 4.2.3 (Functional characterization of weak input invariance)**.** *The random function $\mathbf{f}\colon \mathbb{R}^d \to \mathbb{R}$ is*

1. *weakly stationary, if and only if there exists $\mu \in \mathbb{R}$ and covariance function $C : \mathbb{R}^d \to \mathbb{R}$ such that for all x, y*

$$\mu_{\mathbf{f}}(x) = \mu, \qquad \mathcal{C}_{\mathbf{f}}(x, y) = C(x - y),$$

[11] Translation invariance circumvents the issue that all continuous measures assign zero measure to points. It is therefore not sufficient to require that the probability of any two points are alike to define a uniform distribution in $\mathbb{R}^d$.

[12] Gelfand and Vilenkin, 1964, Sec. 5.3, Thm. 4.

[13] Preprocessing akin to random shuffling requires a uniform distribution over the set of transformations. This exists for the orthogonal group $O(d)$ but not the group of translations $T(d)$ and foreshadows that the stationarity assumption may be problematic, see Section 7.4.

[14] Since random functions technically have two inputs (cf. Definition 2.2.1) we should clarify that the transformation ϕ only applies to x, i.e. $\mathbf{f} \circ \phi(x) = \mathbf{f}(\phi(x))$.

[15] e.g. Haasdonk and Burkhardt, 2007; Brown et al., 2025.

[16] The mean and covariance function of a random function $\mathbf{f}$ are given by $\mu_{\mathbf{f}}(x) := \mathbb{E}[\mathbf{f}(x)]$ and $\mathcal{C}_{\mathbf{f}}(x, y) := \mathrm{Cov}(\mathbf{f}(x), \mathbf{f}(y))$ respectively

2. *weakly non-stationary isotropic, if and only if there exists a mean function* $\mu\colon \mathbb{R}_{\geq 0} \to \mathbb{R}$ *and covariance function* $\kappa\colon D \to \mathbb{R}$ *with* $D = \{\lambda \in \mathbb{R}^2_{\geq 0} \times \mathbb{R} : |\lambda_3| \leq 2\sqrt{\lambda_1 \lambda_2}\} \subseteq \mathbb{R}^3$ *such that for all* x, y

$$\mu_{\mathbf{f}}(x) = \mu\big(\tfrac{\|x\|^2}{2}\big)$$
$$\mathcal{C}_{\mathbf{f}}(x,y) = \kappa\big(\tfrac{\|x\|^2}{2}, \tfrac{\|y\|^2}{2}, \langle x, y\rangle\big),$$

3. *weakly stationary isotropic, if and only if there exists* $\mu \in \mathbb{R}$ *and a function* $C : \mathbb{R}_{\geq 0} \to \mathbb{R}$ *such that for all* x, y

$$\mu_{\mathbf{f}}(x) = \mu, \qquad \mathcal{C}_{\mathbf{f}}(x,y) = C\big(\tfrac{\|x-y\|^2}{2}\big).$$

Proof. The proof essentially follow as a corollary from a characterization of isometries (Proposition 4.A.2). The form of the covariance is proven in the non-stationary isotropic case as part of Theorem 4.4.1.[17] We leave the other characterizations as an exercise. □

[17] There we prove the covariance is of the form

$$\mathcal{C}_{\mathbf{f}}(x,y) = \kappa(\|x\|, \|y\|, \langle x, y\rangle).$$

Through $\tilde{\kappa}(t,s,\gamma) := \kappa(\sqrt{2t}, \sqrt{2s}, \gamma)$ this is clearly equivalent. However, this parametrization is much better suited for derivatives (cf. Section 4.5).

Remark 4.2.4. While straightforward to prove, these characterizations are profound: They provide information about the covariance of $\mathbf{f}(x)$ and $\mathbf{f}(y)$ *within* the same function $\mathbf{f}$ from a uniformity assumption *over* possible realizations of $\mathbf{f}$.

Since weak invariance is equivalent to invariance in the Gaussian case, the conditions on the mean and covariance above fully characterize the invariant Gaussian distributions.

4.3 Characterization of covariance kernels

While any value $\mu \in \mathbb{R}$ in the stationary case, or function $\mu\colon \mathbb{R}_{\geq 0} \to \mathbb{R}$ in the isotropic case, is admissible as a mean function,[18] not every kernel function κ results in a valid covariance kernel. Specifically, since the variance of random variables is necessarily positive, the covariance function $\mathcal{C}_{\mathbf{f}}$ of a random function $\mathbf{f}$ necessarily has to satisfy

$$0 \leq \operatorname{Var}\Big(\sum_{i=1}^{n} a_i \mathbf{f}(x)\Big) = \sum_{i,j=1}^{m} a_i \overline{a_j} \mathcal{C}_{\mathbf{f}}(x_i, x_j)$$

for all evaluation locations x_i and (complex valued) scalars a_i. The covariance $\mathcal{C}_{\mathbf{f}}$ is thus necessarily 'positive definite'. A function $K\colon \mathbb{X} \times \mathbb{X} \to \mathbb{C}$ is called a *positive definite kernel*, if for all $m \in \mathbb{N}$, distinct locations $x_1, \ldots, x_m \in \mathbb{X}$ and $c = (c_1, \ldots, c_m) \in \mathbb{C}^m$

$$\sum_{i,j=1}^{m} c_i \overline{c_j} K(x_i, x_j) \geq 0. \tag{4.1}$$

[18] simply take a centered random function and add this mean

Note that any positive definite kernel is also hermitian[19] and it is well known that for any positive definite kernel K a centered Gaussian random function $\mathbf{f}$ exists with $\mathcal{C}_{\mathbf{f}} = K$.[20] The positive definite kernels thereby coincide with the covariance kernels. If the inequality in (4.1) is strict whenever $c \neq 0$, we call the kernel *strict positive definite*.[21] To make sense of the vast space of positive definite kernels, characterizations have been identified for classes of kernels invariant to certain input transformations. Corresponding to the notion of distributional invariance for random functions in Definition 4.2.1 we define invariance for kernels in Definition 4.3.1 and we summarize a list of such characterizations in Table 4.1.

[19] e.g. Schölkopf and Smola, 2002, Problem 2.13.

[20] The proof of this fact is essentially an application of Kolmogorov's extension theorem (e.g. Klenke, 2014, Thm. 14.36), since positive definite functions induce consistent finite dimensional multivariate Gaussian distributions.

[21] this implies that no linear combination of finitely many random function values is deterministic and is a weak (finite) form of 'universality' (Micchelli et al., 2006).

Definition 4.3.1 (Φ-invariance, kernels)**.** Let Φ be a set of transformations $\phi\colon \mathbb{X} \to \mathbb{X}$, typically a group. A kernel K is called Φ-invariant if

$$K(\phi(x), \phi(y)) = K(x,y) \qquad \forall \phi \in \Phi, x, y \in \mathbb{X},$$

and we denote the set of Φ-invariant kernels by $\mathcal{P}_\Phi := \mathcal{P}_\Phi(\mathbb{X})$. And for $\mathbb{X} = \mathbb{R}^d$ the kernel K is called

TABLE 4.1: Characterizations of continuous pos. def. kernels $K\colon \mathbb{X} \times \mathbb{X} \to \mathbb{C}$. Here $\Omega_d(t) := \mathbb{E}[e^{itX}]$ for $X \sim \mathcal{U}(\mathbb{S}^{d-1})$, $\hat{C}_n^{(\lambda)}$ are the normalized Gegenbauer polynomial of degree n and index $\lambda = \frac{d-2}{2}$, $\mathbb{S}^\infty := \{x \in \ell^2 : \|x\| = 1\}$ and $\mathbb{R}^\infty := \ell^2$.

$\mathbb{X}$	Φ	$K(x,y)$	Characterization	
$\mathbb{R}^d$	$T(d)$	$\kappa(x-y)$	$\kappa(u) = \int e^{i\langle v,u\rangle} d\mu(v)$	μ finite measure
$\mathbb{R}^d$	$E(d)$	$\kappa(\|x-y\|)$	$\kappa(r) = \int_0^\infty \Omega_d(rs)\mu(ds)$	μ finite measure
$\mathbb{R}^\infty$	$E(\infty)$	$\kappa(\|x-y\|)$	$\kappa(r) = \int_0^\infty \exp(-s^2\frac{r^2}{2})\mu(ds)$	μ finite measure
$\mathbb{S}^{d-1}$	$O(d)$	$\kappa(x\cdot y)$	$\kappa(t) = \sum_{n=0}^\infty \gamma_n \hat{C}_n^{(\lambda)}(t)$	$\gamma_n \geq 0$
$\mathbb{S}^\infty$	$O(\infty)$	$\kappa(x\cdot y)$	$\kappa(t) = \sum_{n=0}^\infty \gamma_n t^n$	$\gamma_n \geq 0$
$X \times \mathbb{S}^{d-1}$	$\mathbb{I} \otimes O(d)$	$\kappa(x_X, y_X, x_\mathbb{S}\cdot y_\mathbb{S})$	$\kappa(r,s,t) = \sum_{n=0}^\infty \alpha_n(r,s)\hat{C}_n^{(\lambda)}(t)$	α_n pos. definite
$X \times \mathbb{S}^\infty$	$\mathbb{I} \otimes O(\infty)$	$\kappa(x_X, y_X, x_\mathbb{S}\cdot y_\mathbb{S})$	$\kappa(r,s,t) = \sum_{n=0}^\infty \alpha_n(r,s)t^n$	α_n pos. definite
$\mathbb{R}^d \times \mathbb{S}^{d-1}$	$T(d) \otimes O(d)$	$\kappa(x_\mathbb{R} - y_\mathbb{R}, x_\mathbb{S}\cdot y_\mathbb{S})$	$\kappa(r,s,t) = \sum_{n=0}^\infty \alpha_n(r-s)\hat{C}_n^{(\lambda)}(t)$	α_n pos. definite
$\mathbb{R}^d \times \mathbb{S}^\infty$	$T(d) \otimes O(\infty)$	$\kappa(x_\mathbb{R} - y_\mathbb{R}, x_\mathbb{S}\cdot y_\mathbb{S})$	$\kappa(r,s,t) = \sum_{n=0}^\infty \alpha_n(r-s)t^n$	α_n pos. definite
$\mathbb{R}^d$	$O(d)$	$\kappa(\|x\|, \|y\|, x\cdot y)$	$\kappa(r,s,t) = \sum_{n=0}^\infty \alpha_n(r,s)\hat{C}_n^{(\lambda)}(\frac{t}{rs})$	α_n pos. definite & zero at origin
$\mathbb{R}^\infty$	$O(\infty)$	$\kappa(\|x\|, \|y\|, x\cdot y)$	$\kappa(r,s,t) = \sum_{n=0}^\infty \alpha_n(r,s)\left(\frac{t}{rs}\right)^n$	α_n pos. definite & zero at origin

- *stationary*, if Φ is the group of translations $T(d)$
- *isotropic*, if Φ is the orthogonal group $O(d)$ of linear isometries[22]
- *stationary isotropic*, if Φ is the group of (Euclidean) isometries $E(d)$.

Since any kernel invariant to a set Φ of transformations is also invariant to any subset of transformations, we have $\mathcal{P}_\Psi \supseteq \mathcal{P}_\Phi$ whenever $\Psi \subseteq \Phi$. Consequently, the set of stationary isotropic kernels $\mathcal{P}_{E(d)}$ is a subset of $\mathcal{P}_{T(d)}$ and $\mathcal{P}_{O(d)}$. While the stationary kernels $\mathcal{P}_{T(d)}$ have long been characterized,[23] translation invariance is typically assumed before invariance with respect to the orthogonal group is even considered.[24] This leaves the non-stationary isotropic positive definite kernels without characterization, which is surprising: Virtually all kernel families used in practice are isotropic, up to 'geometric anisotropies'[25] that reduce to isotropy after a change of coordinate system. In contrast, the generalization of $\mathcal{P}_{E(d)}$ to $\mathcal{P}_{O(d)}$ includes many practical non-stationary kernel families, such as the dot-product kernels[26] the kernels arising from infinite width neural networks, which are all non-stationary isotropic (cf. Example 4.4.7) and the covariance kernel of an objective function induced by a random linear model (Section 7.4).

Our characterization of non-stationary isotropic kernels, based on the work of Guella and Menegatto (2019), thereby fills a fundamental gap, capturing virtually all kernel families used in practice.

We further provide necessary and sufficient conditions for isotropic kernels to be strictly positive definite (Section 4.4.1) and provide examples in 4.4.2.

4.4 Characterization of isotropic kernels on $\mathbb{R}^d$

To prove Theorem 4.4.1, we decompose isotropic kernels K in two scalar and one spherical component. The spherical component translates to the Gegenbauer polynomials,[27] which were already used by Schoenberg (1942) to characterize the $O(d)$-invariant kernels on $\mathbb{S}^{d-1}$. We denote the Gegenbauer polynomial of index $\lambda = \frac{d-2}{2}$ by $C_n^{(\lambda)}$. For the subsequent characterization we use the *normalized Gegenbauer polynomials* defined by[28]

$$\hat{C}_n^{(\lambda)}(t) := \begin{cases} C_n^{(\lambda)}(t)/C_n^{(\lambda)}(1) & \lambda < \infty \\ t^n & \lambda = \infty. \end{cases}$$

The limiting case follows from Lemma 1 by Schoenberg (1942), which shows for any $t \in [-1,1]$ that $\hat{C}_n^{(\lambda)}(t)$ converges uniformly in n to t^n as $\lambda \to \infty$. In our setting, the index $\lambda = \frac{d-2}{2}$ is determined by the dimension d and the high dimensional case can therefore be well approximated by the case $d = \infty$ which corresponds to the

Our contribution (cf. Thm. 4.4.1) is summarized in the last two lines of Table 4.1. The first three results can be found in the book by Sasvári (2013). The very first result goes back to Bochner (1933) and can be generalized to groups. Results 2-5 go back to Schoenberg (1938) and Schoenberg (1942). In his honor results 6-9 have been called Schoenberg characterizations (Guella and Menegatto, 2019; Estrade et al., 2019; Berg and Porcu, 2017).

[22] a linear isometry U is norm preserving, and thus for all x

$$\|x\|^2 = \|Ux\|^2 = x^T U^T U x$$

This forces $UU^T = \mathbb{I}$ such that U is orthogonal. Conversely, any orthogonal matrix U induces a linear isometry.

[23] Bochner, 1933.

[24] e.g. Rasmussen and C. K. Williams, 2006, Sec. 4.2.

[25] Stein, 1999, p. 50.

[26] e.g. Rasmussen and C. K. Williams, 2006, Sec. 4.2.2.

[27] see e.g. Reimer, 2003.

[28] note that $|C_n^{(\lambda)}(t)| \leq C_n^{(\lambda)}(1)$ and $0 < C_n^{(\lambda)}(1) < \infty$ (e.g. Reimer, 2003, Eq. (2.16), (2.13)).

sequence space $\mathbb{R}^\infty := \ell^2 = \{(x_i)_{i\in\mathbb{N}} : \|x\|^2 < \infty\}$. The limiting case $\hat{C}_n^{(\infty)}(t) = t^n$ is not only an approximation of the finite dimensional cases, it also corresponds to kernels that are valid in all dimensions (Lemma 4.4.3).

Theorem 4.4.1 (Characterization of positive definite isotropic kernels)**.** *Let $d \in \{2, 3, \dots, \infty\}$ and $\lambda = \frac{d-2}{2}$. For a continuous kernel $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{C}$, the following assertions are equivalent:*

(i) The kernel is positive definite and isotropic, i.e.

$$K(x, y) = K(Ux, Uy)$$

for every linear isometry U.

(ii) The kernel is positive definite and has representation

$$K(x, y) = \kappa\big(\|x\|, \|y\|, \langle x, y\rangle\big)$$

for a continuous $\kappa : M \to \mathbb{C}$ with $M = \{(r, s, \gamma) \in [0, \infty)^2 \times \mathbb{R} : \gamma \le rs\}$.

(iii) With the convention $\hat{C}_n^{(\lambda)}\big(\langle \frac{x}{\|x\|}, \frac{y}{\|y\|}\rangle\big) := 1$ if $x = 0$ or $y = 0$, the kernel has representation

$$K(x, y) = \sum_{n=0}^{\infty} \alpha_n^{(d)}\big(\|x\|, \|y\|\big)\hat{C}_n^{(\lambda)}\big(\langle \tfrac{x}{\|x\|}, \tfrac{y}{\|y\|}\rangle\big),$$

where $\alpha_n^{(d)} : [0, \infty)^2 \to \mathbb{C}$ are continuous, positive definite kernels. These kernels are summable, i.e. $\sum_{l=0}^{\infty} \alpha_n^{(d)}(r, r) < \infty$ for any $r \in [0, \infty)$, and zero at the origin, i.e.

$$\alpha_n^{(d)}(r, 0) = \alpha_n^{(d)}(0, r) = 0 \qquad \forall r \in [0, \infty),\ n > 0. \tag{4.2}$$

Moreover for $d < \infty$ the kernels $\alpha_n^{(d)}$ can be represented as

$$\alpha_n^{(d)}(r, s) = Z \int_{-1}^{1} \kappa(r, s, rs\rho)\hat{C}_n^{(\lambda)}(\rho)(1 - \rho^2)^{\frac{d-3}{2}} d\rho, \tag{4.3}$$

where the normalization constant Z is defined by

$$Z = Z(n, d) = \Big[\int_{-1}^{1} [\hat{C}_n^{(\lambda)}(\rho)]^2(1 - \rho^2)^{\frac{d-3}{2}} d\rho\Big]^{-1}.$$

Remark 4.4.2 (Embedding). Embedding the finite dimensional spaces $\mathbb{R}^d$ into the space of sequences $\ell^2 = \{(x_i)_{i\in\mathbb{N}} : \|x\|^2 < \infty\}$, we naturally obtain

$$\mathbb{R}^1 \subseteq \mathbb{R}^2 \subseteq \cdots \subseteq \mathbb{R}^\infty := \ell^2 \ .$$

Since any positive definite kernel remains positive definite on a subspace it follows that

$$\mathcal{P}_{O(1)}(\mathbb{R}) \supseteq \mathcal{P}_{O(2)}(\mathbb{R}^2) \supseteq \cdots \supseteq \mathcal{P}_{O(\infty)}(\mathbb{R}^\infty). \tag{4.4}$$

While the representation $K(x, y) = \kappa(\|x\|, \|y\|, \langle x, y\rangle)$ can be used to lift any isotropic kernel into ℓ^2, the lifted kernel is not necessarily positive definite.

In light of the remark above, one may be interested in the set of isotropic kernels that are positive definite in all finite dimensional spaces $\mathbb{R}^d$, that is on $\bigcup_{d\in\mathbb{N}} \mathbb{R}^d$, which corresponds to the space of eventually zero sequences. But these kernels coincide with the positive definite kernels on the entire sequence space ℓ^2 as shown in the following lemma.

Lemma 4.4.3 (Valid in all dimensions)**.** *The set of isotropic positive definite kernels that are valid in all dimensions coincides with the set of isotropic positive definite kernels on ℓ^2, i.e.*

$$\bigcap_{d\in\mathbb{N}} \mathcal{P}_{O(d)} = \mathcal{P}_{O(\infty)}(\ell^2).$$

Proof. '$\supseteq$' follows from (4.4). For '$\subseteq$' we need to prove that the representation κ induces a positive definite kernel on ℓ^2. For this consider vectors $x_1, \dots, x_n \in \ell_2$ and map their n dimensional subspace via a linear isometry $U \in O(\infty)$ into $\mathbb{R}^n$. By positive definiteness in $\mathcal{P}_{O(n)}(\mathbb{R}^n)$ Equation (4.1) follows. □

In particular the isotropic positive definite kernels on ℓ^2 coincide with the isotropic positive definite kernels on the space of eventually zero sequences.

The core idea of the proof of Theorem 4.4.1 is to translate results about positive definite kernels on $(0,\infty)\times\mathbb{S}^{d-1}$, independently discoverd by Guella and Menegatto (2019) and Estrade et al. (2019), to a characterization of isotropic kernels on $\mathbb{R}^d\backslash\{0\}$, and continuously extend these results to $\mathbb{R}^d$. For your convenience we restate Theorem 2.3 by Guella and Menegatto (2019) as Theorem 4.A.1 in our notation.

Lemma 4.4.4 (Continuous extension)**.** *Let $K : \mathbb{X}\times\mathbb{X} \to \mathbb{C}$ be a continuous kernel on some topological space $\mathbb{X}$ and let $A \subset \mathbb{X}$ be a set such that the closure of $\mathbb{X}\setminus A$ is $\mathbb{X}$. If K is positive definite on $\mathbb{X}\setminus A$, then K is positive definite on $\mathbb{X}$.*

Proof. Let $x_0, \dots, x_m \in \mathbb{X}$ and $c_0, \dots, c_m \in \mathbb{C}$ and choose for $i = 1, \dots, m$ a sequence $x_i^{(n)} \in \mathbb{X}\backslash A$ such that $x_i^{(n)} \overset{n\to\infty}{\longrightarrow} x_i$. By continuity and positive definiteness of K we have

$$\sum_{i,j=0}^{m} c_i\overline{c_j}K(x_i,x_j) = \lim_{n\to\infty}\sum_{i,j=0}^{m} c_i\overline{c_j}K(x_i^{(n)},x_j^{(n)}) \geq 0. \qquad \square$$

Proof of Theorem 4.4.1. **(i) ⇔ (ii):** "⇐": This follows directly from the fact that U is a linear isometry that preserves norms and inner products.
"⇒": Let $M_{\geq} := \{(r,s,\gamma) \in M : r \geq s\}$. Using two orthonormal vectors e_1, e_2, which exist due to $d \geq 2$, we define the *continuous* maps $\varphi, \psi : M_{\geq} \to \mathbb{R}^d$ by

$$\varphi(r,s,\gamma) := re_1, \qquad \psi(r,s,\gamma) := \begin{cases} \frac{\gamma}{r}e_1 + e_2\sqrt{s^2 - \frac{\gamma^2}{r^2}} & r > 0, \\ 0 & r = 0. \end{cases}$$

By the construction of $M_{\geq}$ all $(r,s,\gamma) \in M_{\geq}$ fulfill $r \geq s$ and $rs \geq |\gamma|$ and hence ψ is continuous away from $(0,0,0)$. For the continuity of ψ in $(0,0,0)$ consider

$$\|\psi(r,s,\gamma)\| \leq \tfrac{rs}{r} + s = 2s \to 0 \quad \text{for} \quad (r,s,\gamma) \to (0,0,0).$$

The goal of φ and ψ is to create two vectors of specified length and angle. Indeed, for $\theta = (r,s,\gamma)$ we have

$$\|\varphi(\theta)\| = r, \quad \|\psi(\theta)\| = s, \quad \langle\varphi(\theta),\psi(\theta)\rangle = \gamma. \tag{4.5}$$

The continuity requirement prevented the definition of φ and ψ on the entirety of M. To deal with $r < s$ we define the swap function $\tau(r,s,\gamma) := \big(s,r,\gamma\big)$. With its help we finally define $\kappa\colon M \to \mathbb{C}$ for $\theta = (r,s,\gamma)$ as

$$\kappa(r,s,\gamma) := \begin{cases} K\big(\varphi(\theta),\psi(\theta)\big) & \text{if } r \geq s \iff \theta \in M_{\geq} \\ K\big(\psi(\tau(\theta)),\varphi(\tau(\theta))\big) & \text{if } r \leq s. \end{cases} \tag{4.6}$$

Clearly, κ is continuous if it is well-defined in $r = s$ since φ, ψ and τ are continuous. For the case $r = s$, Proposition 4.A.2 and Equation (4.5) yield a linear isometry $U \in O(d)$ such that $U\varphi(\theta) = \psi(\tau(\theta))$ and $U\psi(\theta) = \varphi(\tau(\theta))$ and thus

$$K\big(\varphi(\theta),\psi(\theta)\big) \overset{\text{isotropy}}{=} K\big(U\varphi(\theta),U\psi(\theta)\big) = K\big(\psi(\tau(\theta)),\varphi(\tau(\theta))\big).$$

It remains to show that $\kappa(\|x\|,\|y\|,\langle x,y\rangle) = K(x,y)$ holds for all $x,y \in \mathbb{R}^d$. Since the other case works analogous we assume w.l.o.g. $\|x\| \geq \|y\|$ and denote by $\theta := (\|x\|,\|y\|,\langle x,y\rangle)$ the input to κ. Then by Prop. 4.A.2 and (4.5) there again exists $U \in O(d)$ such that $Ux = \varphi(\theta)$ and $Uy = \psi(\theta)$ and we conclude

$$K(x,y) = K(Ux,Uy) = K(\varphi(\theta),\psi(\theta)) = \kappa(\theta).$$

(i) ⇒ (iii): Over the homeomorphism $T\colon (0,\infty)\times\mathbb{S}^{d-1}\to\mathbb{R}^d\setminus\{0\}$ with $T(r,v)=rv$ we define the continuous kernel $K_{\mathbb{S}}\colon((0,\infty)\times\mathbb{S}^{d-1})^2\to\mathbb{C}$ by

$$K_{\mathbb{S}}((r,v),(s,w)) := K(rv,sw) \overset{(ii)}{=} \kappa\big(r,s,rs\langle v,w\rangle\big).$$

With $f_{\mathbb{S}}(r,s,\rho):=\kappa(r,s,rs\rho)$ it is clear that Theorem 4.A.1 is applicable and we obtain by its representation of $f_{\mathbb{S}}$ for any $x,y\in\mathbb{R}^d\setminus\{0\}$

$$K(x,y)=f_{\mathbb{S}}\Big(\|x\|,\|y\|,\big\langle\tfrac{x}{\|x\|},\tfrac{y}{\|y\|}\big\rangle\Big)=\sum_{n=0}^{\infty}\alpha_n^{(d)}(\|x\|,\|y\|)\hat{C}_n^{(\lambda)}\Big(\big\langle\tfrac{x}{\|x\|},\tfrac{y}{\|y\|}\big\rangle\Big),$$

where $\alpha_n^{(d)}$ are positive definite kernels on $(0,\infty)$ that are summable on the diagonal. What remains to be shown is that these kernels are continuous and can be extended to $[0,\infty)$, that this extension indeed resembles K and that (4.2) and (4.3) hold. To this end we investigate the cases $d<\infty$ and $d=\infty$ separately.

1. **Case** $d<\infty$: For all $n\in\mathbb{N}$ and all $r,s\in(0,\infty)$ we have by Theorem 4.A.1

$$\alpha_n^{(d)}(r,s)=Z\int_{-1}^{1}\overbrace{\kappa(r,s,rs\rho)}^{=f_{\mathbb{S}}(r,s,\rho)}\hat{C}_n^{(\lambda)}(\rho)(1-\rho^2)^{\frac{d-3}{2}}d\rho.$$

If we extend this representation to $[0,\infty)$, we obtain the representation (4.3) by definition. Since κ is continuous by (ii), dominated convergence using $|\hat{C}_n^{(\lambda)}(\rho)(1-\rho^2)^{\frac{d-3}{2}}|\le 1$[29] implies we can extend $\alpha_n^{(d)}$ to a continuous function on $[0,\infty)^2$. This continuous extension remains positive definite by Lemma 4.4.4. To see that this extension results in a representation of K, observe that, by the orthogonality of the Gegenbauer polynomials with respect to the weight function $w(\rho)=(1-\rho^2)^{\lambda-\frac12}$,[30] $\hat{C}_0^{(\lambda)}\equiv 1$ and the definition of Z as a normalization factor, we have for $r\ge 0$

$$\alpha_n^{(d)}(r,0)=Z\int_{-1}^{1}\kappa(r,0,0)\hat{C}_0^{(\lambda)}(\rho)\hat{C}_n^{(\lambda)}(\rho)(1-\rho^2)^{\frac{d-3}{2}}d\rho=\kappa(r,0,0)\delta_{n0},$$

where δ_{n0} denotes the Kronecker delta. This implies (4.2) and we also obtain

$$K(x,0)=\kappa(\|x\|,0,0)=\sum_{n=0}^{\infty}\underbrace{\alpha_n^{(d)}(\|x\|,\|0\|)}_{=\kappa(\|x\|,0,0)\delta_{n0}}\underbrace{\hat{C}_n^{(\lambda)}\big(\langle\tfrac{x}{\|x\|},\tfrac{0}{\|0\|}\rangle\big)}_{\overset{\text{def.}}{=}1}.$$

By an analogous argument the representation also holds for $K(0,x)$.

2. **Case** $d=\infty$**:** First note that the $\alpha_n^{(d)}$ are *continuous* kernels on $(0,\infty)$ by Guella and Menegatto (2019, Example 2.4). For the extension we select two orthonormal vectors e_1,e_2 and observe

$$\begin{aligned}K(re_1,re_1)-K(re_1,re_2)&=\sum_{n=0}^{\infty}\alpha_n^{(d)}(r,r)\langle e_1,e_1\rangle^n-\sum_{n=0}^{\infty}\alpha_n^{(d)}(r,r)\langle e_1,e_2\rangle^n\\&=\sum_{n=1}^{\infty}\alpha_n^{(d)}(r,r).\end{aligned}$$

By continuity of K we obtain

$$0=\lim_{r\to0}K(re_1,re_1)-K(re_1,re_2)=\lim_{r\to0}\sum_{n=1}^{\infty}\alpha_n^{(d)}(r,r).$$

Since $\alpha_n^{(d)}(r,r)\ge 0$ holds for all $n\in\mathbb{N}$ since the $\alpha_n^{(d)}$ are kernels, this implies $\alpha_n^{(d)}(r,r)\to 0$ for all $n\ge1$. Assume $r=0$ or $s=0$ and let $(r_l,s_l)\overset{n\to\infty}{\longrightarrow}(r,s)$. Then the Cauchy-Schwarz inequality yields

$$|\alpha_n^{(d)}(r_l,s_l)-\underbrace{\alpha_n^{(d)}(r,s)}_{=0}|\overset{\text{C.S.}}{\le}\sqrt{\alpha^{(d)}(r_l,r_l)\alpha_n^{(d)}(s_l,s_l)}\overset{l\to\infty}{\longrightarrow}0.$$

[29] e.g. Reimer, 2003, Eq. (2.16).

[30] Reimer, 2003, Theorem 2.3.

Hence, we can continuously extend $\alpha_n^{(d)}$ by zero as claimed in (4.2). Lemma 4.4.4 implies for $n > 0$ that this continuous extension of positive definite kernels $\alpha_n^{(d)}$ is positive definite. We continue with the extension in the case $n = 0$. For this consider

$$K(re_1, se_2) = \sum_{n=0}^{\infty} \alpha_n^{(d)}(r,s)\langle e_1, e_2\rangle^n = \alpha_0^{(d)}(r,s).$$

This yields the continuous extension $\alpha_0^{(d)}(r,s) := K(re_1, se_2)$. Lemma 4.4.4 shows that this extension remains positive definite. What is left to prove is that our continuous extension of the $\alpha_n^{(d)}$ kernels does in fact lead to a representation of K. To this end, let $x, y \in \mathbb{R}^d$ and w.l.o.g. assume $x = 0$ to obtain

$$K(x,y) = K(0, \|y\|e_2) = \alpha_0^{(d)}(0, \|y\|) = \sum_{n=0}^{\infty} \alpha_n^{(d)}(0, \|y\|)\langle \tfrac{0}{\|0\|}, y\rangle^n.$$

(iii) ⇒ (i): Isotropy is obvious and it remains to show positive definiteness. We split the series representation into two parts

$$K(x,y) = \underbrace{\alpha_0^{(d)}(\|x\|, \|y\|)}_{K_0(x,y)} + \underbrace{\sum_{n=1}^{\infty} \alpha_n^{(d)}(\|x\|, \|y\|)\hat{C}_n^{(\lambda)}\big(\langle \tfrac{x}{\|x\|}, \tfrac{y}{\|y\|}\rangle\big)}_{=K_{>0}(x,y)}, \tag{4.7}$$

where we use $\hat{C}_0^{(\lambda)} \equiv 1$. Let $m \in \mathbb{N}, x_1, \dots, x_m \in \mathbb{R}^d, c_1, \dots, c_m \in \mathbb{C}$ and let without loss of generality $x_i = 0$ if and only if $i = 1$. Since $\alpha_n^{(d)}(0,r) = 0$ for $n \geq 1$ we obtain

$$\sum_{i,j=1}^{m} c_i\overline{c_j}K(x_i,x_j) = \sum_{i,j=1}^{m} c_i\overline{c_j}K_0(x_i,x_j) + \sum_{i,j=2}^{m} c_i\overline{c_j}K_{>0}(x_i,x_j).$$

Since $\alpha_0^{(d)}$ is a kernel on $[0,\infty)$ we get that the first summand is non-negative. The second summand is non-negative, since it only contains $x_i \in \mathbb{R}^d \setminus \{0\}$ and the restriction $K_{>0}|_{(\mathbb{R}^d\setminus\{0\})^2}$ is positive definite using the homeomorphism $T\colon (0,\infty) \times \mathbb{S}^{d-1} \to \mathbb{R}^d \setminus \{0\}$, $(r,v) \mapsto rv$ and Theorem 4.A.1 characterizing kernels on $(0,\infty) \times \mathbb{S}^{d-1}$. □

4.4.1 Strict positive definiteness

In this section we characterize the requirements on a continuous, isotropic positive definite kernel to be *strictly* positive definite.[31] We will again utilize the connection between $(0,\infty) \times \mathbb{S}^{d-1}$ and $\mathbb{R}^d \setminus \{0\}$ and base our results on these by Guella and Menegatto (2019). Since they had to single out the case $d = 2$ case, we have to do so too.[32]

Theorem 4.4.5 (Characterization of strictly positive definite isotropic kernels)**.** *Let $d \in \{3, \dots, \infty\}$. For a continuous positive definite isotropic kernel $K\colon \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{C}$ with representation as in Theorem 4.4.1, the following assertions are equivalent:*

(i) The kernel K is strictly positive definite;

(ii) $\alpha_0^{(d)}(0,0) > 0$, and for any $m \in \mathbb{N}$, distinct $r_1, \dots, r_m \in (0,\infty)$ and $c \in \mathbb{C}^m \setminus \{0\}$ the set

$$\left\{ n \in \mathbb{N}_0 : c^T \left[\alpha_n^{(d)}(r_i, r_j)\right]_{i,j=1}^{m} \overline{c} > 0 \right\}$$

contains infinitely many even and infinitely many odd integers.

(iii) $\alpha_0^{(d)}(0,0) > 0$, and for each $\gamma \geq 0$ the even and odd kernels

$$\alpha_\gamma^e(r,s) := \sum_{2n \geq \gamma} \alpha_{2n}^{(d)}(r,s) \quad \text{and} \quad \alpha_\gamma^o(r,s) := \sum_{2n+1 \geq \gamma} \alpha_{2n+1}^{(d)}(r,s)$$

are both strictly positive definite on $(0,\infty)$.

[31] For readers interested in *universal* kernels we want to point out their characterization on the sphere by Michelli et al. (2006, Theorem 10) that may also be helpful via the connection to $(0,\infty) \times \mathbb{S}^{d-1}$ for the characterization of universal kernels in $\mathcal{P}_{O(d)}$.

[32] Warning: We consider $\mathbb{S}^{d-1}$ whereas they considered $\mathbb{S}^d$ and therefore singled out $d = 1$.

Furthermore, the following statement is sufficient for strict positive definiteness:

(iv) $\alpha_0^{(d)}(0,0) > 0$, *and* $\alpha_n^{(d)}$ *is strictly positive definite on* $(0,\infty)$ *for infinitely many even and odd* $n \in \mathbb{N}_0$.

If $\alpha_0^{(d)}(0,0) = 0$ *but all other requirements in (ii), (iii), (iv) are satisfied, then* K *is strictly positive definite on* $\mathbb{R}^d \setminus \{0\}$. *(ii) and (iii) without* $\alpha_0^{(d)}(0,0) > 0$ *are also necessary for strict positive definiteness on* $\mathbb{R}^d \setminus \{0\}$.

Proof. Recall the decomposition (4.7), i.e.

$$K(x,y) = \underbrace{\alpha_0^{(d)}(\|x\|,\|y\|)}_{=:K_0(x,y)} + \underbrace{\sum_{n=1}^{\infty} \alpha_n^{(d)}(\|x\|,\|y\|)\hat{C}_n^{(\lambda)}(\langle \tfrac{x}{\|x\|}, \tfrac{y}{\|y\|}\rangle)}_{=:K_{>0}(x,y)},$$

with positive definite kernels K_0 and $K_{>0}$. Let $m \in \mathbb{N}$, $c \in (\mathbb{C} \setminus \{0\})^m$ and let $x_1, \dots, x_m \in \mathbb{R}^d$ be distinct locations where $x_1 = 0$ w.l.o.g., then we have

$$\sum_{i,j=1}^{m} c_i \overline{c_j} K(x_i,x_j) = \sum_{i,j=1}^{m} c_i \overline{c_j} K_0(x_i,x_j) + \sum_{i,j=2}^{m} c_i \overline{c_j} K_{>0}(x_i,x_j) \tag{4.8}$$

using that $K_{>0}(x,0) = K_{>0}(0,x) = 0$ for all $x \in \mathbb{R}^d$ due to (4.2) in Theorem 4.4.1.

"(ii), (iii), (iv) $\Rightarrow$ (i)": For strict positive definiteness of K we need to show (4.8) is non-zero. Since both summands are non-negative it is sufficient if either one is non-zero. With the usual homeomorphism the conditions in (ii), (iii) and (iv) imply the conditions in Theorem 3.5, 3.7 and 3.9 of Guella and Menegatto (2019) on $(0,\infty) \times \mathbb{S}^{d-1}$. Since these conditions are satisfied for $K_{>0}$ if they are satisfied for K we thereby see that $K_{>0}$ is strictly positive definite on $\mathbb{R}^d \setminus \{0\}$. If $m > 1$ the second term in (4.8) is therefore non-zero. If $m = 1$ then the first term is

$$\sum_{i,j=1}^{m} c_i \overline{c_j} K_0(x_i,x_j) = |c_1|^2 \alpha_0^{(d)}(0,0) > 0.$$

"(i) $\Rightarrow$ (ii), (iii)": If K is strictly positive definite on $\mathbb{R}^d$, then $\alpha_0^{(d)}(0,0) > 0$ by the selection of $m = 1$, $c = 1$, $x_1 = 0$ in (4.7). Mapping strict positive definiteness of K on $\mathbb{R}^d \setminus \{0\}$ into strict positive definiteness on $(0,\infty) \times \mathbb{S}^{d-1}$, the assertions (ii) and (iii) follow from Theorem 3.5 and 3.7 of Guella and Menegatto (2019).

The final remark about strict positive definiteness on $\mathbb{R}^d \setminus \{0\}$ follows similarly from the homeomorphism to $(0,\infty) \times \mathbb{S}^{d-1}$ and the results of Guella and Menegatto (2019). □

By the same arguments Theorem 3.6, 3.8 and 3.9 of Guella and Menegatto (2019) for the case $d = 2$ yield to the following theorem.

Theorem 4.4.6 (Strict positive definiteness for $d = 2$)**.** *Let* $K\colon \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{C}$ *be a continuous positive definite isotropic kernel with decomposition as in Theorem 4.4.1. Then* necessary *conditions for strict positive definiteness include*

(i) $\alpha_0^{(2)}(0,0) > 0$, *and for any* $m \in \mathbb{N}$, $c \in \mathbb{C}^m \setminus \{0\}$ *and* $r_1, \dots, r_m \in (0,\infty)$ *the set*

$$\left\{ n \in \mathbb{Z} : c^T \left[\alpha_{|n|}^{(2)}(r_i,r_j) \right]_{i,j=1}^{m} \overline{c} > 0 \right\}$$

intersects every full arithmetic progression in $\mathbb{Z}$.

(ii) $\alpha_0^{(2)}(0,0) > 0$, *and for each full arithmetic progression* S *in* $\mathbb{Z}$ *the kernel*

$$(r,s) \mapsto \sum_{n \in S} \alpha_{|n|}^{(2)}(r,s) \qquad r,s \in (0,\infty).$$

is strictly positive definite.

and a ***sufficient*** *condition for strict positive definiteness of* K *is*

(iii) $\alpha_0^{(2)}(0,0) > 0$, *and the set* $\{n \in \mathbb{Z} : \alpha_{|n|}^{(2)}$ *is strictly positive definite*$\}$ *intersects every full arithmetic progression in* $\mathbb{Z}$.

4.4.2 Applications and examples

In applications such as support vector machines and Bayesian modelling one has to choose a family of covariance kernels and fit its parameters to the data. The family of non-stationary isotropic covariance kernels represents a joint generalization of stationary isotropic covariance kernels and 'dot product' kernels, which have so far been treated as separate cases.[33]

[33] e.g. Roos et al., 2021; Ament and Gomes, 2022.

Popular stationary isotropic covariance kernels include the squared exponential[34]

$$K(x,y) = \sigma^2 \exp\!\big(-\tfrac{\|x-y\|^2}{2s^2}\big) = \sum_{n=0}^{\infty} \underbrace{\frac{\sigma^2 \|x\|^n \|y\|^n}{n!\, s^{2n}} e^{-\frac{\|x\|^2}{2s^2}} e^{-\frac{\|y\|^2}{2s^2}}}_{=\alpha_n^{(\infty)}(\|x\|,\|y\|)} \langle \tfrac{x}{\|x\|}, \tfrac{y}{\|y\|} \rangle^n,$$

[34] cf. Gaussian smoothing in Example 4.1.2 and Schoenbergs characterization of stationary isotropic kernels in ℓ^2 in Table 4.1.

as well as the Matérn and rational quadratic kernels. Dot product kernels include the linear kernel

$$K(x,y) = \sigma_b^2 + \sigma_w^2 \langle x, y \rangle$$

used for Bayesian linear regression as well as the polynomial kernels which are powers of the linear kernel. The linear kernel is also the covariance function of the mean squared error applied to a random linear model (Section 7.4). The union of stationary isotropic kernels and dot product kernels captures essentially all kernel families used in practice[35] – up to 'geometric anisotropies'.[36]

[35] e.g. Rasmussen and C. K. Williams, 2006, Sec. 4.2.

[36] Stein, 1999, p. 50.

Strictly non-stationary isotropic covariance kernels have also naturally come up as "Neural network kernels", i.e. kernels that arise from taking the layer widths of neural networks to infinity and initializing them randomly. This approach was pioneered by C. Williams (1996) where the limit of a one layer neural network with the 'error function' as activation function resulted in the kernel

$$K(x,y) = \arcsin\Big(\frac{x \cdot y}{\sqrt{(1+\|x\|^2)(1+\|y\|^2)}}\Big)$$

Another such example is the arc-cosine kernel[37] of the form

[37] Cho and Saul, 2009.

$$K(x,y) = \frac{1}{\pi} \|x\|^n \|y\|^n J_n(\cos^{-1}(x \cdot y))$$

for certain J_n that arise when the activation function $\phi(x) = \max\{0, x\}^n$ is used instead. This includes the popular ReLU activation function with $n = 1$.

Example 4.4.7. Multilayer neural networks with activation function ϕ are (non-stationary) isotropic, centered Gaussian random functions at initialization.

Proof. Lee et al. (2018) show that multilayer neural networks are centered Gaussian processes at initialization and that the kernel of the $l+1$-th layer is recursively given by

$$K_{l+1}(x,y) = \sigma_b^2 + \sigma_w^2 F_\phi(K_l(x,y), K_l(x,x), K_l(y,y))$$

for some function F_ϕ that depends on the activation function ϕ with $K_0(x,y) = \sigma_b^2 + \sigma_w^2(x \cdot y)$. With this formula one can show recursively that the kernels are of the form $K_l(x,y) = \kappa_l(\|x\|, \|y\|, x \cdot y)$ and therefore non-stationary isotropic. In particular the output layer and thus the entire network is an isotropic Gaussian random function. □

4.5 Covariance of derivatives

In Section 2.B we explained that the covariance of derivatives of $\mathbf{f}$ is given by the derivatives of covariances

$$\mathrm{Cov}(\partial_{x_i} \mathbf{f}(x), \partial_{y_j} D_w \mathbf{f}(y)) = \partial_{x_i} \partial_{y_j} \mathcal{C}_{\mathbf{f}}(x,y)$$

In this section we want to calculate these covariances explicitly for the case of stationary and non-stationary isotropic random functions. To calculate the

covariances of more general directional derivatives we introduce some notation. For functions with one input, $D_v f(x)$ unmistakeably denotes the directional derivative of f in the direction of v. But for functions with multiple inputs $D_v f(x,y)$ is unclear. To better mark the place the directional derivative is applied to, we introduce the notation

$$f(D_v x, y) := \tfrac{d}{dt} f(x + tv, y).$$

such that

$$\operatorname{Cov}(D_v \mathbf{f}(x), D_w \mathbf{f}(y)) = \mathcal{C}_{\mathbf{f}}(D_v x, D_w y). \tag{4.9}$$

First, we calculate the covariance of derivatives for stationary isotropic random functions $\mathbf{f}$. Recall these have covariance functions of the form $\mathcal{C}_{\mathbf{f}}(x,y) = C\big(\frac{\|x-y\|^2}{2}\big)$ by Theorem 4.2.3.

Lemma 4.5.1 (Covariance of derivatives under stationary isotropy)**.** *Let $\mathbf{f} \sim \mathcal{N}(\mu, C)$ be an isotropic random function and let $\Delta = x - y$, then*

Cov	$\mathbf{f}(y)$	$D_w \mathbf{f}(y)$
$\mathbf{f}(x)$	$C\big(\frac{\|\Delta\|^2}{2}\big)$	$-C'\big(\frac{\|\Delta\|^2}{2}\big)\langle \Delta, w\rangle$
$D_v \mathbf{f}(x)$	$C'\big(\frac{\|\Delta\|^2}{2}\big)\langle \Delta, v\rangle$	$-\Big[C''\big(\frac{\|\Delta\|^2}{2}\big)\langle \Delta, w\rangle\langle \Delta, v\rangle + C'\big(\frac{\|\Delta\|^2}{2}\big)\langle w, v\rangle\Big]$

Proof. Straightforward application of chain and product rules applied to (4.9). □

Second we will consider the non-stationary isotropic kernels. Recall that these are of the form $\mathcal{C}_{\mathbf{f}}(x,y) = \kappa\big(\frac{\|x\|^2}{2}, \frac{\|y\|^2}{2}, \langle x, y\rangle\big)$ by Theorem 4.2.3. Since the (non-stationary) isotropic kernel κ have multiple inputs, we further introduce the notation

$$\kappa_i(\lambda_1, \lambda_2, \lambda_3) := \tfrac{d}{d\lambda_i}\kappa(\lambda_1, \lambda_2, \lambda_3) \qquad i \in \{1,2,3\}.$$

Lemma 4.5.2 (Covariance of derivatives under (non-stationary) isotropy)**.** *Let $\mathbf{f} \sim \mathcal{N}(\mu, \kappa)$ be (non-stationary) isotropic (Definition 4.2.1). Then the expectation of the directional derivative is given by*

$$\mathbb{E}[D_v \mathbf{f}(x)] = \mu'\big(\tfrac{\|x\|^2}{2}\big)\langle x, v\rangle, \tag{4.10}$$

and, using the abbreviation $\kappa := \kappa\big(\frac{\|x\|^2}{2}, \frac{\|y\|^2}{2}, \langle x, y\rangle\big)$ to omit inputs, the covariance of directional derivatives is given by

$$\operatorname{Cov}(D_v \mathbf{f}(x), \mathbf{f}(y)) = \frac{1}{d}\Big[\kappa_1 \langle x, v\rangle + \kappa_3 \langle y, v\rangle\Big] \tag{4.11}$$

$$\operatorname{Cov}(D_v \mathbf{f}(x), D_w \mathbf{f}(y)) = \frac{1}{d}\Big[\underbrace{\kappa_{12}\langle x, v\rangle\langle y, w\rangle + \kappa_{13}\langle x, v\rangle\langle x, w\rangle + \kappa_{32}\langle y, v\rangle\langle y, w\rangle + \kappa_{33}\langle y, v\rangle\langle y, w\rangle}_{(I)} + \underbrace{\kappa_3\langle v, w\rangle}_{(II)}\Big]. \tag{4.12}$$

In particular, if the directions v, w are orthogonal to the points x, y then (I) of (4.11) and (4.12) are zero. (II) in turn is zero, if the directions v, w are orthogonal.

Proof. Straightforward application of chain and product rules applied to (4.9). □

4.6 Strict positive definite derivatives

While covariance matrices are always positive definite, they are not always strict positive definite (i.e. invertible). Recall, that a random function $\mathbf{f}$ and its covariance function $\mathcal{C}_{\mathbf{f}}$ are *strict* positive definite if the matrix $(\mathcal{C}_{\mathbf{f}}(x_k, x_l))_{k,l=1,\dots,n}$ is strict positive definite for any finite $x_1, \dots, x_n$ (cf. Section 4.3). It is already known that, whenever $\mathbf{f}$ is stationary isotropic and non-constant, $\mathbf{f}$ is strict positive definite if the dimension satisfies $d \geq 2$.[38] But since we require the jet $\mathrm{J}^1\mathbf{f} = (\mathbf{f}, \nabla\mathbf{f})$ to be

[38] Sasvári, 2013, Theorem 3.1.5.

strict positive definite in Chapter 6, we prove a generalization of Theorem 3.1.6 in Sasvári (2013). This generalization will be sufficient to prove that all stationary isotropic covariance functions which are valid in all dimensions (cf. Lemma 4.4.3) are covered as we will see in Corollary 4.6.3. This result will be used to remove the strict positive definiteness assumption in Corollary 6.1.7.

If $\mathbf{g}$ is a random function with multivariate output, then $\mathcal{C}_{\mathbf{g}}(x_k, x_j)$ is already a matrix for fixed k, j and the collection over k, j is really a tensor. To avoid introducing this machinery and explaining positive definiteness for tensors, we will take the following equivalent statement for positive definiteness of random functions as definition.

Definition 4.6.1 (Strict positive definite random function)**.** The covariance $\mathcal{C}_{\mathbf{g}}$ of a random function $\mathbf{g} : \mathbb{R}^d \to \mathbb{R}^m$ and the random function itself is called *strict positive definite* if for all $w_k \in \mathbb{R}^m$ and distinct $x_1, \dots, x_n \in \mathbb{R}^d$ the equality

$$0 = \operatorname{Var}\Big[\sum_{k=1}^{n} w_k^T \mathbf{g}(x_k)\Big] \overset{(4.13)}{=} \sum_{k,l=1}^{n} w_k^T \mathcal{C}_{\mathbf{g}}(x_k, x_l) w_l$$

implies $w_k = 0$ for all k.

Note, that the second equality marked with (4.13) is always true and only represents an equivalent formulation, because after centering $\mathbf{g}$ (without loss of generality, using $\tilde{\mathbf{g}} = \mathbf{g} - \mathbb{E}[\mathbf{g}]$), we have

$$\sum_{k,l=1}^{n} w_k^T \mathcal{C}_{\mathbf{g}}(x_k, x_l) w_l = \mathbb{E}\Big[\sum_{k,l=1}^{n} w_k^T \mathbf{g}(x_k)\mathbf{g}(x_l)^T w_l\Big] = \operatorname{Var}\Big[\sum_{k=1}^{n} w_k^T \mathbf{g}(x_k)\Big]. \quad (4.13)$$

Theorem 4.6.2 (Strict positive definite derivatives)**.** *Let $\mathbf{f} : \mathbb{R}^d \to \mathbb{R}$ be a stationary random function (cf. Definition 4.2.1 and Theorem 4.2.3) with continuous covariance such that the support of its spectral measure contains a non-empty open set. Then up to any order k up to which $\mathbf{f}$ is almost surely differentiable, the jet*

$$\mathrm{J}^k \mathbf{f} = (\mathbf{f}, \nabla \mathbf{f}, (\partial_{ij} \mathbf{f})_{i \le j}, \dots, (\partial_{i_1 \cdots i_k} \mathbf{f})_{i_1 \le \cdots \le i_k})$$

is strict positive definite.

Theorem 4.6.2 is not more general than Sasvári (2013, Theorem 3.1.6) in the sense that conditions are weakend, but it is more general in its implication. That is, $\mathbf{g}$ is strict positive definite and not just $\mathbf{f}$.

The following corollary shows that this fully covers the stationary isotropic covariance functions valid in all dimensions (cf. Lemma 4.4.3). We use this result about stationary isotropic random functions in Corollary 6.1.7 to omit the assumption that $(\mathbf{f}, \nabla \mathbf{f})$ has to be strict positive definite, which is needed for our general result (Theorem 6.1.10).

Corollary 4.6.3. *Assume that C is a continuous stationary isotropic covariance kernel valid in all dimensions (i.e. defined on ℓ^2 by Lemma 4.4.3). Let $\mathbf{f} \sim \mathcal{N}(\mu, C)$ be a Gaussian random function, which is not almost surely constant. Then the jet*

$$\mathrm{J}^k \mathbf{f} = (\mathbf{f}, \nabla \mathbf{f}, (\partial_{ij} \mathbf{f})_{i \le j}, \dots, (\partial_{i_1 \cdots i_k} \mathbf{f})_{i_1 \le \cdots \le i_k})$$

is strict positive definite for any $k \in \mathbb{N}$.

4.6.1 Proofs

Proof of Theorem 4.6.2. For any n finite and distinct $x_1, \dots, x_n \in \mathbb{R}^d$ we need that

$$0 = \operatorname{Var}\Big[\sum_{j=1}^{n} w_j^T \mathrm{J}^k \mathbf{f}(x_j)\Big]$$

implies $w_i = 0$ for all $i = 1, \dots, n$. Without loss of generality we assume $\mathbf{f}$ (and thus $\mathrm{J}^k\mathbf{f}$) to be centered. We can then rewrite this as

$$\mathrm{Var}\Big[\sum_{j=1}^{n} w_j^T \mathrm{J}^k\mathbf{f}(x_j)\Big] = \mathrm{Var}\Big[\sum_{j=1}^{n}\sum_{l=0}^{k}\sum_{i_1\le\dots\le i_l} w_j^{(l,i_1,\dots,i_l)}\partial_{i_1\dots i_l}\mathbf{f}(x_j)\Big],$$

with appropriate indexing of the w_j. Note that l ranges over the order of differentiation contained in $\mathbf{g}$ before we sum over all the partial derivatives in the inner sum. For $k = 1$ this is for example

$$\mathrm{Var}\Big[\sum_{j=1}^{n} w_j^{(0)}\mathbf{f}(x_j) + \sum_{i=1}^{d} w_j^{(1,i)}\partial_i\mathbf{f}(x_j)\Big] = \mathrm{Var}\Big[\sum_{j=1}^{n} w_j^{(0)}\mathbf{f}(x_j) + D_{w_j^{(1,\cdot)}}\mathbf{f}(x_j)\Big] = 0.$$

We now consider the linear differential operator

$$T := \sum_{j=1}^{n}\sum_{l=0}^{k}\sum_{i_1\le\dots\le i_l} (-1)^l w_j^{(l,i_1,\dots,i_l)}\partial_{i_1\dots i_l}\delta_{x_j},$$

where $\delta_x(f) = f(x)$ is the dirac delta function and $\partial_i\delta_x(f) = -\partial_i f(x)$ is its derivative in the sense of distributions. Recall this means for test functions ϕ, that we have

$$\langle\partial_i\delta_x, \phi\rangle = -\langle\delta_x, \partial_i\phi\rangle = -\partial_i\phi(x).$$

The higher order derivatives in the sense of distributions are similarly defined. Using this operator we now obtain more succinctly

$$0 = \mathrm{Var}\Big[\sum_{j=1}^{n}\sum_{l=0}^{k}\sum_{i_1\le\dots\le i_l} w_j^{(l,i_1,\dots,i_l)}\partial_{i_1\dots i_l}\mathbf{f}(x_j)\Big] = \mathrm{Var}[T\mathbf{f}].$$

While $T\mathbf{f}$ is well defined, we now want to move this operator outside. For this we want to use the bilinearity of the covariance. But the covariance has two inputs and $T\mathcal{C}_\mathbf{f}$ is then not well defined because the differential operator T might be applied to the first or second input of $\mathcal{C}_\mathbf{f}$. To avoid this issue, we define with some abuse of notation $T^t f(t) := Tf$ such that we can write $T^t\mathcal{C}_\mathbf{f}(t,s)$ when we mean $T\mathcal{C}_\mathbf{f}(\cdot,s)$. Note that T is a linear combination of basis elements $\partial_{i_1\dots i_l}\delta_x$. So by bilinearity of the covariance it is sufficient to check, that we can move these basis elements out of the covariance, i.e.

$$\begin{aligned}\mathrm{Cov}(\partial_{i_1\dots i_l}\delta_x\mathbf{f}, \partial_{j_1\dots j_{l'}}\delta_y\mathbf{f}) &= \mathrm{Cov}(\partial_{i_1\dots i_l}\mathbf{f}(x), \partial_{j_1\dots j_{l'}}\mathbf{f}(y))\\ &= (\partial_{i_1\dots i_l}\delta_x)^t\ (\partial_{j_1\dots j_{l'}}\delta_y)^s\mathcal{C}_\mathbf{f}(t,s).\end{aligned}$$

Since T is a linear combinations of these, we get by the bilinearity of the covariance

$$0 = \mathrm{Var}[T\mathbf{f}] = T^xT^y\mathcal{C}_\mathbf{f}(x,y).$$

In the remainder of the proof we will essentially show, that this variance can be represented as an integral over the absolute value of the fourier transform of T with respect to the spectral measure of $\mathcal{C}_\mathbf{f}$. This forces the fourier transform of T and therefore T to be zero.

Using the spectral representation of $\mathcal{C}_\mathbf{f}$[39] given by

$$\mathcal{C}_\mathbf{f}(x,y) = \int e^{i\langle x-y,t\rangle}\sigma(dt),$$

[39] e.g. Sasvári, 2013, Theorem 1.7.4.

we move the operator into the integral (by moving sums and derivatives into the integral)

$$0 = T^xT^y\mathcal{C}_\mathbf{f}(x,y) = \int T^x e^{i\langle x,t\rangle}T^y e^{-i\langle y,t\rangle}\sigma(dt) = \int |Te^{i\langle\cdot,t\rangle}|\sigma(dt),$$

where we use $T\bar{f} = \overline{Tf}$ for the last equation, which follows from

- $\delta_x(\overline{f}) = \overline{\delta_x(f)}$
- $D_v\delta_x(\overline{f}) = D_v\overline{f(x)} = \overline{D_v f(x)} = \overline{D_v\delta_x(f)}$ because from $|z| = |\overline{z}|$ follows

$$\lim_{t\to 0} \frac{|\overline{f(x+tv)} - \overline{f(x)} - \overline{D_v f(x)}t|}{t} = 0,$$

- induction for higher order derivatives.

So we have that $P(t) := Te^{i\langle\cdot,t\rangle}$ must be zero σ-almost everywhere. Since

$$P(t) = \sum_{j=1}^{n}\sum_{l=0}^{k}\sum_{i_1\leq\cdots\leq i_l} w_j^{(l,i_1,\ldots,i_l)}\partial_{i_1\ldots i_l}e^{i\langle x_j,t\rangle}$$

is continuous in t, it can only be zero σ-almost everywhere if it is zero on the support of σ. Since the support of the spectral measure σ contains an open subset by assumption of the theorem, P must be zero on this open subset. But then P must be zero everywhere as an analytic function. As $P(t)$ is the fourier transform of T in the sense of distributions, i.e.

$$P(t) = \int \Big(\sum_{j=1}^{n}\sum_{l=0}^{k}\sum_{i_1\leq\cdots\leq i_l} w_j^{(l,i_1,\ldots,i_l)}\partial_{i_1\ldots i_l}\delta_{x_j}\Big)(x)e^{i\langle x,t\rangle}dx = \mathcal{F}[T](t),$$

this requires T to be zero by linearity and invertibility of the Fourier transform. But since the $\partial_{i_1\ldots i_l}\delta_{x_j}$ are linear independent[40] for distinct x_j the only way T can be zero is, if all the w_j are zero. This finishes the proof. □

The general requirement, the existence of a non-empty open sets in the support of the spectral measure, is satisfied for the stationary isotropic random functions valid in all dimensions (Corollary 4.6.3). To prove this result we require the following lemma. It proves that stationary isotropic random function in ℓ^2 is either almost surely constant or the 'Schoenberg measure' has positive measure on $(0,\infty)$. Where the Schoenberg measure ν refers to the measure in Schoenberg's characterization of stationary isotropic covariance kernels on ℓ^2 given by[41]

$$\mathcal{C}_{\mathbf{f}}(x,y) = C(\tfrac{\|x-y\|^2}{2}) = \int_{[0,\infty)} \exp(-t^2\tfrac{\|x-y\|^2}{2})\nu(dt) \tag{4.14}$$

Lemma 4.6.4 (Constant random functions)**.** *Let C be a stationary isotropic covariance kernel valid in all dimensions and let $\mathbf{f}\sim\mathcal{N}(\mu,C)$ be a continuous stationary isotropic random function.*

If the Schoenberg measure ν of C has no mass on $(0,\infty)$, i.e. $\nu((0,\infty)) = 0$, then $\mathbf{f}$ is almost surely constant.

Proof. By Schoenberg's characterization of stationary isotropic positive definite kernels in ℓ^2 (4.14) we have

$$C(r) = \int_{[0,\infty)} \exp(-t^2 r)\nu(dt) \overset{\nu((0,\infty)=0)}{=} \nu(\{0\}).$$

This implies a constant covariance

$$\operatorname{Cov}(\mathbf{f}(x),\mathbf{f}(y)) = \nu(\{0\}).$$

Assuming $\mathbf{f}$ to be centered (without loss of generality) by switching to $\tilde{\mathbf{f}}_d = \mathbf{f} - \mu$, this implies

$$\mathbb{E}[(\mathbf{f}(x)-\mathbf{f}(y))^2] = \mathbb{E}[\mathbf{f}(x)^2] - 2\mathbb{E}[\mathbf{f}(x)\mathbf{f}(y)] + \mathbb{E}[\mathbf{f}(y)^2] = 0.$$

Thus $\mathbf{f}(x) = \mathbf{f}(y)$ almost surely for all x,y. Via the union over a dense countable subset of $\mathbb{R}^d$ and a.s. continuity of $\mathbf{f}$ we get

$$\mathbb{P}(\mathbf{f}\text{ a.s. constant}) = \mathbb{P}\Big(\mathbf{f}(x) = \mathbf{f}(y) \quad \forall x,y\in\mathbb{R}^d\Big) = 1.$$ □

[40] To see the linear independence of $\partial_{i_1\ldots i_l}\delta_{x_j}$, consider that the finite x_k are distinct, so they have a minimal distance. Rescale a bump function with zero slope at the top, e.g.

$$\phi(x) = \begin{cases} \exp\Big(-\frac{1}{1-\|x\|^2}\Big) & \|x\| < 1 \\ 0 & \|x\| \geq 1 \end{cases}$$

such that it is centered on some x_j and it is zero at all other x_k (and zero at all derivatives). This implies

$$0 = \langle T,\phi\rangle = w_j^{(0)}\phi(x_j)$$

and thus $w_j^{(0)} = 0$. Then construct similar test functions to ensure the other prefactors have to be zero, by placing a non-zero slope at x_j while ensuring it is zero at all other x_k.

[41] cf. Section 4.3, Table 4.1

Using this lemma, we can prove that all stationary isotropic covariance kernels on ℓ^2 have strictly positive definite derivatives.

Corollary 4.6.3. *Assume that C is a continuous stationary isotropic covariance kernel valid in all dimensions (i.e. defined on ℓ^2 by Lemma 4.4.3). Let $\mathbf{f} \sim \mathcal{N}(\mu, C)$ be a Gaussian random function, which is not almost surely constant. Then the jet*

$$\mathrm{J}^k\mathbf{f} = (\mathbf{f}, \nabla\mathbf{f}, (\partial_{ij}\mathbf{f})_{i\le j}, \dots, (\partial_{i_1\cdots i_k}\mathbf{f})_{i_1\le\cdots\le i_k})$$

is strict positive definite for any $k \in \mathbb{N}$.

Proof of Corollary 4.6.3. Since $\psi(x) = e^{-\frac{\|x\|^2 t^2}{2}}$ is the characteristic function of $Y_t \sim \mathcal{N}(0, t\mathbb{I}_{d\times d})$, Schoenbergs characterization (4.14) of the covariance implies up to scaling

$$\begin{aligned}
\mathcal{C}_{\mathbf{f}}(x,\tilde{x}) = C\big(\tfrac{\|x-\tilde{x}\|^2}{2}\big) &= \int_{[0,\infty)} e^{-\frac{t^2\|x-\tilde{x}\|^2}{2}}\nu(dt) \\
&= \int_{[0,\infty)} \mathbb{E}[e^{i\langle x-\tilde{x}, Y_t\rangle}]\nu(dt) \\
&= \int e^{i\langle x-\tilde{x},y\rangle} \underbrace{\int_{(0,\infty)} e^{-\frac{\|y\|^2}{2t}}\nu(dt)}_{=:\varphi_\sigma(y)}\, dy + \nu(\{0\}) \\
&= \int e^{i\langle x-\tilde{x},y\rangle}\sigma(dy)
\end{aligned}$$

with spectral measure

$$\sigma(A) := \int_A \varphi_\sigma(y)dy + \tfrac{1}{d}\nu\{0\}\delta_0(A).$$

Since $\nu((0,\infty)) \neq 0$ because $\mathbf{f}$ is not almost surely constant (Corollary 4.6.4), its density φ_σ is continuous and positive $\varphi_\sigma(y) > 0$ for all $y \in \mathbb{R}^d$ and thus $\mathrm{supp}(\sigma) = \mathbb{R}^d$. In particular Theorem 4.6.2 is applicable and finishes the proof. □

References

Adler, Robert J. and Jonathan E. Taylor (2007). *Random Fields and Geometry*. Springer Monographs in Mathematics. New York, NY: Springer New York. ISBN: 978-0-387-48112-8. DOI: 10.1007/978-0-387-48116-6.

Ament, Sebastian and Carla Gomes (2022). "Scalable First-Order Bayesian Optimization via Structured Automatic Differentiation". In: *Proceedings of the 39th International Conference on Machine Learning*. ICML. Baltimore, Maryland, USA. arXiv: 2206.08366v1 [cs.LG].

Applebaum, David (2009). *Lévy Processes and Stochastic Calculus*. 2nd ed. Cambridge Studies in Advanced Mathematics. Cambridge: Cambridge University Press. ISBN: 978-0-521-73865-1. DOI: 10.1017/CBO9780511809781.

Berg, Christian and Emilio Porcu (2017). "From Schoenberg Coefficients to Schoenberg Functions". In: *Constructive Approximation* 45.2, pp. 217–241. ISSN: 1432-0940. DOI: 10.1007/s00365-016-9323-9.

Bochner, S. (1933). "Monotone Funktionen, Stieltjessche Integrale und harmonische Analyse". In: *Mathematische Annalen* 108.1, pp. 378–410. ISSN: 1432-1807. DOI: 10.1007/BF01452844.

Bogachev, Vladimir I. (1998). *Gaussian Measures*. Vol. 62. Mathematical Surveys and Monographs. Providence, RI: American Mathematical Society. 433 pp. ISBN: 0-8218-1054-5.

Brown, Theodore, Alexandru Cioba, and Ilija Bogunovic (2025). "Sample-Efficient Bayesian Optimisation Using Known Invariances". In: *Advances in Neural Information Processing Systems* 37, pp. 47931–47965. URL: https://proceedings.neurips.cc/paper_files/paper/2024/hash/55aeba84b402008d3ed10440d906b4e1-Abstract-Conference.html (visited on 2025-02-24).

Carmeli, C. et al. (2010). "Vector Valued Reproducing Kernel Hilbert Spaces and Universality". In: *Analysis and Applications* 08.01, pp. 19–61. ISSN: 0219-5305. DOI: 10.1142/S0219530510001503.

Cho, Youngmin and Lawrence Saul (2009). "Kernel Methods for Deep Learning". In: *Advances in Neural Information Processing Systems*. Vol. 22. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper/2009/hash/5751ec3e9a4feab575962e78e006250d-Abstract.html (visited on 2023-04-03).

Estrade, Anne, Alessandra Fariñas, and Emilio Porcu (2019). "Covariance Functions on Spheres Cross Time: Beyond Spatial Isotropy and Temporal Stationarity". In: *Statistics & Probability Letters* 151, pp. 1–7. ISSN: 0167-7152. DOI: 10.1016/j.spl.2019.03.011.

Fageot, Julien and Thomas Humeau (2021). *The Domain of Definition of the Lévy White Noise*. DOI: 10.48550/arXiv.1708.02500. arXiv: 1708.02500 [math]. Pre-published.

Gelfand, I. M. and N. Ya Vilenkin (1964). *Generalized Functions, Volume 4: Applications of Harmonic Analysis*. Trans. by Amiel Feinstein. AMS Chelsea Publishing. Providence, Rhode Island: American Mathematical Society. 384 pp. ISBN: 978-1-4704-2662-0. URL: https://lccn.loc.gov/2015040021 (visited on 2023-12-04).

Guella, J. C. and V. A. Menegatto (2019). "Schoenberg's Theorem for Positive Definite Functions on Products: A Unifying Framework". In: *Journal of Fourier Analysis and Applications* 25.4, pp. 1424–1446. ISSN: 1531-5851. DOI: 10.1007/s00041-018-9631-5.

Haasdonk, Bernard and Hans Burkhardt (2007). "Invariant Kernel Functions for Pattern Analysis and Machine Learning". In: *Machine Learning* 68.1, pp. 35–61. ISSN: 1573-0565. DOI: 10.1007/s10994-007-5009-7.

Ken-Iti, Sato (1999). *Lévy Processes and Infinitely Divisible Distributions*. Vol. 68. Cambridge university press. URL: https://books.google.com/books?hl=en&lr=&id=CwT5BNG0-owC&oi=fnd&pg=PR9&dq=L%C3%A9vy+processes+and+infinitely+divisible+distributions.+Cambridge+University+Press.+pp.+31%E2%80%9368.&ots=45lIP_EY8N&sig=D3cHPrRJJneEvmyRzWF6GsqoHng (visited on 2025-03-07).

Klenke, Achim (2014). *Probability Theory: A Comprehensive Course*. Universitext. London: Springer. ISBN: 978-1-4471-5360-3 978-1-4471-5361-0. DOI: 10.1007/978-1-4471-5361-0.

Lee, Jaehoon et al. (2018). "Deep Neural Networks as Gaussian Processes". In: International Conference on Learning Representations. Vancouver, Canada. arXiv: 1711.00165 [stat.ML]. URL: https://openreview.net/forum?id=B1EA-M-0Z (visited on 2023-04-03).

Micchelli, Charles A., Yuesheng Xu, and Haizhang Zhang (2006). "Universal Kernels". In: *Journal of Machine Learning Research* 7.12. URL: http://www.jmlr.org/papers/volume7/micchelli06a/micchelli06a.pdf (visited on 2025-03-21).

Rasmussen, Carl Edward and Christopher K.I. Williams (2006). *Gaussian Processes for Machine Learning*. 2nd ed. Adaptive Computation and Machine Learning 3. Cambridge, Massachusetts: MIT Press. 248 pp. ISBN: 0-262-18253-X. URL: http://gaussianprocess.org/gpml/chapters/RW.pdf (visited on 2023-04-06).

Reimer, Manfred (2003). "Gegenbauer Polynomials". In: *Multivariate Polynomial Approximation*. Basel: Birkhäuser, pp. 19–38. ISBN: 978-3-0348-8095-4. DOI: 10.1007/978-3-0348-8095-4_2.

Roos, Filip de, Alexandra Gessner, and Philipp Hennig (2021). "High-Dimensional Gaussian Process Inference with Derivatives". In: *Proceedings of the 38th International Conference on Machine Learning*. International Conference on Machine Learning. PMLR, pp. 2535–2545. URL: https://proceedings.mlr.press/v139/de-roos21a.html (visited on 2023-05-15).

Sasvári, Zoltán (2013). *Multivariate Characteristic and Correlation Functions*. De Gruyter Studies in Mathematics 50. Berlin/Boston: Walter de Gruyter. 376 pp. ISBN: 978-3-11-022399-6.

Schoenberg, Isaac J. (1938). "Metric Spaces and Positive Definite Functions". In: *Transactions of the American Mathematical Society* 44.3, pp. 522–536. URL: https://community.ams.org/journals/tran/1938-044-03/S0002-9947-1938-1501980-0/S0002-9947-1938-1501980-0.pdf (visited on 2024-04-18).

— (1942). "Positive Definite Functions on Spheres". In: *Duke Mathematical Journal* 9.1, pp. 96–108. ISSN: 0012-7094, 1547-7398. DOI: 10.1215/S0012-7094-42-00908-6.

Schölkopf, Bernhard and Alexander J. Smola (2002). *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. 1st ed. Cambridge, Mass.: MIT Press. 644 pp. ISBN: 978-0-262-53657-8.

Stein, Michael L. (1999). *Interpolation of Spatial Data*. Springer Series in Statistics. New York, NY: Springer. ISBN: 978-1-4612-7166-6 978-1-4612-1494-6. DOI: 10.1007/978-1-4612-1494-6.

Williams, Christopher (1996). "Computing with Infinite Networks". In: *Advances in Neural Information Processing Systems*. Vol. 9. MIT Press. URL: https://proceedings.neurips.cc/paper/1996/hash/ae5e3ce40e0404a45ecacaaf05e5f735-Abstract.html (visited on 2022-12-01).

4.A Appendix

Since we are interested in $\mathbb{R}^d$ we consider $\mathbb{S}^{d-1}$, whereas the source we build our characterization of isotropic kernels on[42] considers $\mathbb{S}^d$. We also work with the normalized Gegenbauer polynomials while they stated their result with the unnormalized ones. For your convenience we therefore restate their main theorem in our setting and notation.

[42] Guella and Menegatto, 2019.

Theorem 4.A.1 ($\mathbb{X} \times \mathbb{S}^{d-1}$-characterization, Guella and Menegatto, 2019)**.** *Let $d \in \{2, 3, \dots, \infty\}$ and define $\lambda := \frac{d-2}{2}$. For a kernel $K : (\mathbb{X} \times \mathbb{S}^{d-1})^2 \to \mathbb{C}$ which is isotropic on the sphere, i.e. of the form*

$$K((r, v), (s, w)) = f_{\mathbb{S}}(r, s, \langle v, w \rangle),$$

the following are equivalent

(i) The kernel K is positive definite and for all $r, s \in \mathbb{X}$ the function $f_{\mathbb{S}}(r, s, \rho)$ is continuous in $[-1, 1]$.

(ii) The kernel has series representation of the form

$$f_{\mathbb{S}}(r, s, \rho) = \sum_{n=0}^{\infty} \alpha_n^{(d)}(r, s) \hat{C}_n^{(\lambda)}(\rho), \qquad \rho \in [-1, 1],$$

where $\alpha_n^{(d)} : \mathbb{X}^2 \to \mathbb{C}$ are positive definite kernels for all $n \in \mathbb{N}_0$ that are summable, i.e. $\sum_{n=0}^{\infty} \alpha_n^{(d)}(r, r) < \infty$ for all $r \in \mathbb{X}$.

If the kernel satisfies one of these equivalent conditions and $d < \infty$ the kernels $\alpha_n^{(d)}$ moreover have the representation

$$\alpha_n^{(d)}(r, s) = Z \int_{-1}^{1} f_{\mathbb{S}}(r, s, \rho) \hat{C}_n^{(\lambda)}(\rho) (1 - \rho^2)^{(d-3)/2} d\rho$$

with normalizing constant

$$Z = Z(n, d) = \left[\int_{-1}^{1} [\hat{C}_n^{(\lambda)}(\rho)]^2 (1 - \rho)^{(d-3)/2} d\rho \right]^{-1}.$$

Proof. This is essentially Theorem 2.3 by Guella and Menegatto (2019) except that we translate their $a_n^d(r, s)$ into $\alpha_n^{(d)}(r, s) := a_n^{(d)}(r, s) C_n^{(\lambda)}(1)$ for the use with *normalized* ultraspherical polynomials. □

A key component for the characterization of invariant kernels and distributionally invariant random functions is a characterization of isometries that is most likely well known.

Proposition 4.A.2 (Characterizing isometries)**.** *Let $\mathcal{X}$ be a vector space over $\mathbb{K} \in \{\mathbb{R}, \mathbb{C}\}$ and $x_i, y_i \in \mathcal{X}$ for $i \in 1, \dots, n$, then the following pairs of statements are equivalent*

1. a) *there exists a* ***translation***[43] *φ with $\varphi(x_i) = y_i$ for all i.*
   b) *$x_i - x_j = y_i - y_j$ for all i, j*

If $\mathcal{X}$ is furthermore a hilbertspace, then

2. a) *there exists a* ***linear isometry***[44] *ϕ with $\phi(x_i) = y_i$ for all i*
   b) *$\|x_i - x_j\| = \|y_i - y_j\|$ and $\|x_i\| = \|y_i\|$ for all i, j*
   c) *$\langle x_i, y_i \rangle$ for all i, j*
3. a) *there exists an* ***isometry*** *ϕ with $\phi(x_i) = y_i$ for all i.*
   b) *$\|x_i - x_j\| = \|y_i - y_j\|$ for all i, j*
   c) *there exists a linear isometry ϕ and $v \in \mathcal{X}$ such that $y_i = \phi(x_i) + v$ for all i*

[43] A translation is a map on $\mathcal{X}$ of the form $x \mapsto x + v$ for some $v \in \mathcal{X}$

[44] An isometry $\phi \colon \mathcal{X} \to \mathcal{X}$ is a map that retains distances, i.e. for all $x, y \in \mathcal{X}$

$$\|\phi(x) - \phi(y)\| = \|x - y\|.$$

For linear maps the isometry property is equivalent to norm preservation, i.e. $\|\phi(x)\| = \|x\|$ for all $x \in \mathcal{X}$.

Proof. (1b) $\Rightarrow$ (1a)**:** we define $\varphi(x) := x + (y_0 - x_0)$, which implies

$$\varphi(x_i) = x_i - x_0 + y_0 = (y_i - y_0) + y_0 = y_i.$$

(1a) $\Rightarrow$ (1b)**:** Let $\varphi(x) = x + c$ for some c. Then we immediately have

$$y_i - y_j = \varphi(x_i) - \varphi(x_j) = x_i + c - (x_j + c) = x_i - x_j.$$

(2a) $\Rightarrow$ (2b)**:** Isometries preserve distances by definition. This implies $\|x_i - x_j\| = \|y_i - y_j\|$. And linear functions map 0 to 0, so we have

$$\|x_i\| = \|x_i - 0\| = \|\phi(x_i) - \phi(0)\| = \|y_i\|.$$

(2b) $\Rightarrow$ (2c)**:** The polarization formula $\langle x, y \rangle = \frac{1}{2}(\|x\|^2 + \|y\|^2 - \|x - y\|^2)$ yields the claim.

(2c) $\Rightarrow$ (2a)**:** With the application of the Gram-Schmidt orthonormalization procedure to both x_i and y_i we obtain

$$\begin{aligned} U_m &:= \operatorname{span}\{u_1, \dots, u_{k_m}\} = \operatorname{span}\{x_1, \dots, x_m\} \\ V_m &:= \operatorname{span}\{v_1, \dots, v_{\tilde{k}_m}\} = \operatorname{span}\{y_1, \dots, y_m\} \end{aligned}$$

for orthonormal u_i and v_i where we skip x_m (or y_m respectively) if it is already contained in $U_{k_{m-1}}$ (or $V_{k_{m-1}}$ resp.) resulting in $k_m = k_{m-1}$ (or $\tilde{k}_m = \tilde{k}_{m-1}$) respectively. We will prove inductively over $m \leq n$ that $k_m = \tilde{k}_m$ and for all $i \leq n$ and $j \leq k_m$

$$\langle x_i, u_j \rangle = \langle y_i, v_j \rangle. \tag{4.15}$$

The induction start with $m = 0$ is trivial. For the induction step observe that x_m being contained in $U_{k_{m-1}}$ is equivalent to

$$\langle x_m, x_m \rangle = \Big\langle x_m, \sum_{j=1}^{k_{m-1}} \langle x_m, u_j \rangle u_j \Big\rangle = \sum_{j=1}^{k_{m-1}} \langle x_m, u_j \rangle^2$$

And since all the inner products are the same by induction (4.15), $x_m \in U_{k_{m-1}}$ is satisfied if and only if $y_m \in V_{k_{m-1}} = V_{\tilde{k}_{m-1}}$. We therefore have $k_m = \tilde{k}_m$.

In the case where k_m increments we have by the induction assumption (4.15) and (2c)

$$\langle x_i, u_{k_m}\rangle = \Big\langle x_i, x_m - \sum_{j=1}^{k_{m-1}} \langle x_m, u_j\rangle u_j \Big\rangle$$
$$\overset{(4.15),\ (2c)}{=} \langle y_i, y_m\rangle - \sum_{j=1}^{k_{m-1}} \langle y_i, v_j\rangle\langle y_m, v_j\rangle = \langle y_i, v_{k_m}\rangle$$

for all x_i with $i \leq n$. This finishes our induction.

Since $W = \operatorname{span}(U_n \cup V_n)$ has finite dimension $d \leq 2n$, we can extend the orthonormal basis $u_1, \ldots, u_{k_n}$ of $U_n = \operatorname{span}\{x_1, \ldots, x_n\}$ to an orthonormal basis $u_1, \ldots, u_d$ of W and similarly the orthonormal basis $v_1, \ldots, v_{k_n}$ of $V_n = \operatorname{span}\{y_1, \ldots, y_n\}$ to an orthonormal basis $v_1, \ldots, v_d$ of W.

We finally define the linear map $\phi\colon \mathcal{X} \to \mathcal{X}$, which acts as the identity on $W^\perp$ and maps the basis elements u_j to v_j. This is an isometry since for all $x \in \mathcal{X}$ there exists a basis representation $x = \sum_{i=1}^n \lambda_i u_i + w$ with $\lambda_i \in \mathbb{K}^n$ and $w \in W$, which implies

$$\|\phi(x)\|^2 = \Big\|\sum_{i=1}^n \lambda_i \phi(v_i) + w\Big\|^2 = \|\lambda\|^2 + \|w\|^2 = \|x\|^2.$$

This isometry does the job, because

$$\phi(x_m) = \phi\Big(\sum_{j=1}^n \langle x_m, u_j\rangle u_j\Big) = \sum_{j=1}^n \langle x_m, u_j\rangle \phi(u_j) \overset{(4.15)}{=} \sum_{j=1}^n \langle y_m, v_j\rangle v_j = y_m.$$

(3a) $\Rightarrow$ (3b) This is precisely the distance preserving property of isometries.

(3b) $\Rightarrow$ (3c)**:** We define $\tilde{x}_i = x_i - x_0$ and similarly for y. In particular, $\tilde{x}_1 = \tilde{y}_1 = 0$. Since $\tilde{x}_i$ and $\tilde{y}_i$ satisfy the requirements of (2b), there exists a linear isometry matrix ϕ with $\phi(\tilde{x}_i) = \tilde{y}_i$. Define $v = y_1 - \phi(x_1)$ to obtain

$$\phi(x_i) + v = \phi(\tilde{x}_i) + y_1 = \tilde{y}_i + y_1 = y_i$$

(3c) $\Rightarrow$ (3a)**:** It is straightforward to show $\varphi(x) = \phi(x) + v$ is an isometry. $\square$

Chapter 5

Random Function Descent

In Section 2.2.3 we have shown that the most popular acquisition functions in Bayesian optimization can be reduced to functions of the posterior mean and covariance in the Gaussian case. While this certainly makes these algorithms more explicit, it has still left us with the task of computing the conditional mean and standard deviation and the optimization of the acquisition function. In Section 2.A we have learned how such a conditional Gaussian distribution can be computed. Recall that to compute the conditional distribution $Z_0 \mid Z_1, \dots Z_N$ of the multivariate Gaussian random vector $(Z_0, \dots, Z_N)$ the covariance matrix of $Z_1, \dots, Z_N$ has to be inverted. Further recall that gradients of a Gaussian random function $f : \mathbb{R}^d \to \mathbb{R}$ contain d Gaussian entries. In the case of first order Bayesian optimization the evaluation filtration $\mathcal{F}_n$ (2.1) therefore contains $\mathcal{O}(nd)$ Gaussian random variables. Inverting their covariance matrix thus incurs a cost of $\mathcal{O}(n^3d^3)$ in general, which renders this approach infeasible in high-dimension.

In contrast to these computationally complex Bayesian optimization algorithms classical algorithms, such as gradient descent, are very explicit and cheap to compute with $\mathcal{O}(nd)$ complexity. Taking inspiration from the classical approach we introduce the acquisition function 'Random Function Descent' in this chapter. It turns out that the minimizer of this acquisition function can be explicitly computed and is in fact equal to gradient descent with a specific step size. While the step size generally has to be tuned in the classical setting this Bayesian view provides a theoretical foundation for the choice of step size.

5.1 The random function descent algorithm

Classically, gradient descent is motivated as the minimizer of the trust bound based on the first Taylor approximation and a bound on the error (Lemma 1.2.6). In the case of L-smoothness it is of the form

$$f(y) \leq \underbrace{f(x) + \langle \nabla f(x), y - x \rangle}_{=:T[f(y) \mid f(x), \nabla f(x)]} + \tfrac{L}{2} \|y - x\|.$$

While the function f may be hard to minimize, this trust bound is a quadratic function and thereby easy to minimize with minimizer $x - \frac{1}{L}\nabla f(x)$ (cf. Lemma 1.2.2).

The eccentric notation for the Taylor approximation,[1] $T[f(y) \mid f(x), \nabla f(x)]$, is meant to highlight the connection to the stochastic Taylor approximation we define below.

[1] which is a linear approximation of f in x

Definition 5.1.1 (Stochastic Taylor approximation)**.** We define the first order stochastic Taylor approximation of a random (cost) function $\mathbf{f}$ around x by the conditional expectation

$$\mathbb{E}[\mathbf{f}(y) \mid \mathbf{f}(x), \nabla \mathbf{f}(x)].$$

This is the best L^2 approximation[2] of $\mathbf{f}(y)$ provided first order knowledge of $\mathbf{f}$ around x.

[2] Klenke, 2014, Cor. 8.17.

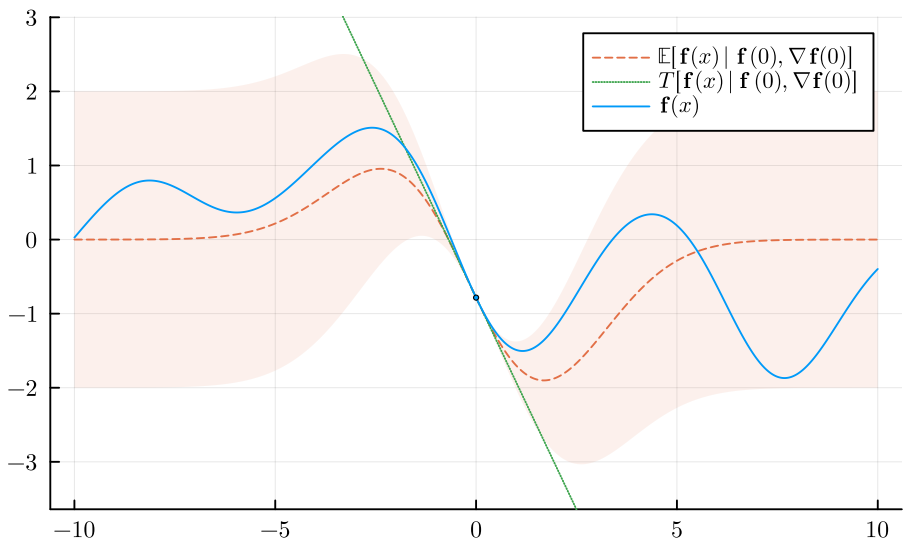


FIGURE 5.1: The stochastic Taylor approximation naturally contains a trust bound in contrast to the classical one. Here $\mathbf{f}$ is a Gaussian random function (with covariance as in Equation (5.11), with length scale $s = 2$ and variance $\sigma^2 = 1$). The ribbon represents two conditional standard deviations around the conditional expectation.

We call this the 'stochastic Taylor approximation' because this approximation only makes use of derivatives in a single point. While the standard Taylor approximation is a polynomial approximation, the 'stochastic Taylor approximation' is the best approximation in an L^2 sense. Since the stochastic Taylor approximation is already mean-reverting by itself, it naturally incorporates covariance-based trust (cf. Figure 5.1) so we omit a trust bound. A variance based trust may be added (cf. Section 5.E.2) but to obtain step sizes that are optimal on *average* we start without.[3,4]

While L-smoothness-based trust *guarantees* that the gradient still points in the direction we are going (for learning rates smaller $1/L$), covariance based trust tells us whether the derivative is still negative *on average*. Minimizing the stochastic Taylor approximation is therefore optimized for the average case. Since convergence proofs for gradient descent typically rely on an improvement *guarantee*[5], proving convergence is significantly harder in the average case and we answer this question only partially in Corollary 5.3.3.

Definition 5.1.2 (Random Function Descent – RFD)**.** Select x_{n+1} as the minimizer[6] of the first order stochastic Taylor approximation

$$x_{n+1} := \arg\min_x \mathbb{E}[\mathbf{f}(x) \mid \mathbf{f}(x_n), \nabla\mathbf{f}(x_n)].$$

Properties of RFD Before we make RFD more explicit in Section 5.2, we discuss some properties which are easier to see in the abstract Bayesian form.

First, observe that RFD is greedy and forgetful in the same way gradient descent is greedy and forgetful when derived as the minimizer of the regularized first Taylor approximation, or the Newton method as the minimizer of the second Taylor approximation. This is because the Taylor approximation only uses derivatives from the last point x_n (forgetful), and we minimize this approximation (greedy). Since momentum methods retain some information about past gradients, they are not as forgetful. We therefore expect a similar improvement could be made for RFD in the future.

Second, it is well known that classical gradient descent with exogenous step sizes (and most other first order methods) lack the scale invariance property of the Newton method.[7] Scale invariance means that scaling the input parameters x or the cost itself (e.g. by switching from the mean squared error to the sum squared error) does not change the points selected by the optimization method.

Advantage 5.1.3 (Scale invariance and equivariance)**.** *RFD is invariant to additive shifts and positive scaling of the cost* $\mathbf{f}$*. RFD is also equivariant*[8] *with respect*

[3] As we will see in Chapter 6, the average case is highly representative of the actual behavior in high-dimension.

[4] Observe that the UCB algorithm from Example 2.2.3 (for maximization!) does the opposite of a trust bound. Instead of a penalty for uncertainty, it encourages exploration with a variance-based bonus.

[5] the 'Descent Lemma' (Lemma 1.2.2)

[6] we ignore throughout the main body that arg min could be set-valued and that the x_n would be random variables (cf. Section 5.D.1 for a formal approach).

[7] e.g. Hansson, 2023; Deuflhard and Heindl, 1979.

[8] a function ϕ is *invariant* with respect to a group G, if $\phi(\rho(g)x) = \phi(x)$, where $\rho(g)$ is a representation of the element $g \in G$. ϕ is *equivariant* with respect to the group G, if $\phi(\rho(g)x) = \rho(g)\phi(x)$ for all $g \in G$.

Model		RFD step size η^* for $\mathbf{f}(x) \leq \mu$		A-RFD
		General case $\left(\text{with } \Theta = \frac{\|\nabla \mathbf{f}(x)\|}{\mu - \mathbf{f}(x)}\right)$	$\mathbf{f}(x) = \mu$	$\Theta \to 0$
Matérn	ν			
	$3/2$	$\frac{s}{\sqrt{3}} \frac{1}{\left(1 + \frac{\sqrt{3}}{s\Theta}\right)}$	$\approx 0.58s$	$\frac{1}{3}s^2\Theta$
	$5/2$	$\frac{s}{\sqrt{5}} \frac{(1-\zeta)+\sqrt{4+(1+\zeta)^2}}{2(1+\zeta)}$ with $\zeta := \frac{\sqrt{5}}{3s\Theta}$.	$\approx 0.72s$	$\frac{3}{5}s^2\Theta$
Squared-exponential	∞	$\frac{s^2}{\sqrt{\left(\frac{\mu - \mathbf{f}(x)}{2}\right)^2 + s^2\|\nabla \mathbf{f}(x)\|^2} + \frac{\mu - \mathbf{f}(x)}{2}} \|\nabla \mathbf{f}(x)\|$	s	$s^2\Theta$
Rational quadratic	β	$s\sqrt{\beta}\,\underset{\eta}{\text{Root}}\left\{-1 + \frac{\sqrt{\beta}}{s\Theta}\eta + (1+\beta)\eta^2 + \frac{\sqrt{\beta}}{s\Theta}\eta^3\right\}$	$s\sqrt{\frac{\beta}{1+\beta}}$	$s^2\Theta$

TABLE 5.1: RFD step size (cf. Figure 5.2 and Eq. (5.11), (5.13), (5.14) for the formal definitions of the models). In particular, s is the length scale in all covariance models.

to transformations of the parameter input of $\mathbf{f}$ *by differentiable bijections whose Jacobian is invertible everywhere (e.g. invertible linear maps).*

While equivariance with respect to bijections of inputs is much stronger than the affine equivariance offered by the Newton method, non-linear bijections will typically break the isotropy assumption (Chapter 4) about the random function $\mathbf{f}$ which makes RFD explicit. This equivariance should therefore be viewed as an opportunity to look for the bijection of inputs which ensures isotropy (e.g. a whitening transformation). The discussion of geometric anisotropy in Section 5.E.1 is conducive to build an understanding of this.

5.2 Relation to gradient descent

While we were able to define RFD abstractly without any assumptions on the distribution $\mathbb{P}_{\mathbf{f}}$ of the random cost $\mathbf{f}$, an explicit calculation of this acquisition function requires distributional assumptions. As we motivated in Section 2.2.3 and Chapter 4, a sensible assumption are stationary isotropic Gaussian random functions. This assumption allows for an explicit version of the stochastic Taylor approximation which then immediately leads to an explicit version of RFD.

Lemma 5.2.1 (Explicit first order stochastic Taylor approximation)**.** *For* $\mathbf{f} \sim \mathcal{N}(\mu, C)$*, the first order stochastic Taylor approximation is given by*

$$\mathbb{E}[\mathbf{f}(x - \mathbf{d}) \mid \mathbf{f}(x), \nabla \mathbf{f}(x)] = \mu + \frac{C\left(\frac{\|\mathbf{d}\|^2}{2}\right)}{C(0)}(\mathbf{f}(x) - \mu) - \frac{C'\left(\frac{\|\mathbf{d}\|^2}{2}\right)}{C'(0)}\langle \mathbf{d}, \nabla \mathbf{f}(x)\rangle.$$

The explicit version of RFD follows by fixing the step size $\eta = \|\mathbf{d}\|$ and optimizing over the direction first.

Theorem 5.2.2 (Explicit RFD)**.** *Let* $\mathbf{f} \sim \mathcal{N}(\mu, C)$*, then RFD coincides with gradient descent*

$$x_{n+1} = x_n - \eta_n^* \frac{\nabla \mathbf{f}(x_n)}{\|\nabla \mathbf{f}(x_n)\|},$$

where the RFD step sizes are given by

$$\eta_n^* := \underset{\eta \in \mathbb{R}}{\arg\min} \frac{C\left(\frac{\eta^2}{2}\right)}{C(0)}(\mathbf{f}(x_n) - \mu) - \eta \frac{C'\left(\frac{\eta^2}{2}\right)}{C'(0)}\|\nabla \mathbf{f}(x_n)\|. \tag{5.1}$$

While the descent direction is a universal property for all isotropic Gaussian random functions, it follows from (5.1) that the step sizes depend much more on the specific covariance structure. In particular it depends on the decay rate of the covariance acting as the trust bound.

Remark 5.2.3 (Scalable complexity). While Bayesian optimization typically has computational complexity $\mathcal{O}(n^3 d^3)$ in number of steps n and dimensions d,[9] RFD under the isotropy assumption has the same computational complexity as gradient descent (i.e. $\mathcal{O}(nd)$).

[9] Wu et al., 2017; Roos et al., 2021.

Remark 5.2.4 (Step until the given information is no longer informative). While L-smoothness-based trust prescribes step sizes that *guarantee* the slope to point downwards over the entire step, RFD prescribes steps which are exactly large enough that the gradient is no longer correlated to the previously observed evaluation. This is because the first order condition demands

$$0 \overset{!}{=} \nabla \mathbb{E}[\mathbf{f}(x) \mid \mathbf{f}(x_n), \nabla \mathbf{f}(x_n)] = \mathbb{E}[\nabla \mathbf{f}(x) \mid \mathbf{f}(x_n), \nabla \mathbf{f}(x_n)].$$

And for measurable functions $\phi : \mathbb{R}^{d+1} \to \mathbb{R}$ such that $\Phi = \phi(\mathbf{f}(x_n), \nabla \mathbf{f}(x_n))$ is sufficiently integrable, Φ is then uncorrelated from $\partial_i \mathbf{f}(x)$ by the first order condition

$$\operatorname{Cov}(\partial_i \mathbf{f}(x), \Phi) = \mathbb{E}\Big[\underbrace{\mathbb{E}[\partial_i \mathbf{f}(x) \mid \mathbf{f}(x_n), \nabla \mathbf{f}(x_n)]}_{=0}(\Phi - \mathbb{E}[\Phi])\Big] = 0.$$

5.3 The RFD step size schedule

While classical L-smooth theory leads to 'learning rates', RFD suggests 'step sizes' applied to normalized gradients[10] representing the actual length of the step size. In the following we thus make the distinction

$$x_{n+1} = x_n - \underbrace{h_n}_{\text{'learning rate'}} \nabla \mathbf{f}(x_n) = x_n - \underbrace{\eta_n}_{\text{'step size'}} \frac{\nabla \mathbf{f}(x_n)}{\|\nabla \mathbf{f}(x_n)\|}.$$

[10] which we also encountered in the worst case theory with the modulus of continuity (Remark 1.2.3)

To get a better feel for the step sizes suggested by RFD, it is enlightening to divide (5.1) by $\mu - \mathbf{f}(x_n)$ which results in the equivalent minimization problem

$$\eta^* := \eta^*(\Theta) := \arg\min_{\eta} q_\Theta(\eta) \qquad \text{for} \qquad q_\Theta(\eta) := -\frac{C\left(\frac{\eta^2}{2}\right)}{C(0)} - \eta \frac{C'\left(\frac{\eta^2}{2}\right)}{C'(0)}\Theta, \quad (5.2)$$

which is only parametrized by the "gradient cost quotient"

$$\Theta_n = \frac{\|\nabla \mathbf{f}(x_n)\|}{\mu - \mathbf{f}(x_n)},$$

i.e. $\eta_n^* = \eta^*(\Theta_n)$. This minimization problem can be solved explicitly for the most common[11] differentiable isotropic covariance models, see Table 5.1, Figure 5.2 and Appendix 5.C for details.

[11] Rasmussen and Williams, 2006, ch. 4.

Figure 5.2 can be interpreted as follows: At the start of optimization, the cost should be roughly equal to the average cost $\mu \approx \mathbf{f}(x)$, so the gradient cost quotient Θ is infinite and the step sizes are therefore given by $\eta^*(\infty)$ (also listed in its own column in Table 5.1). As we start minimizing, the difference $\mu - \mathbf{f}(x)$ becomes

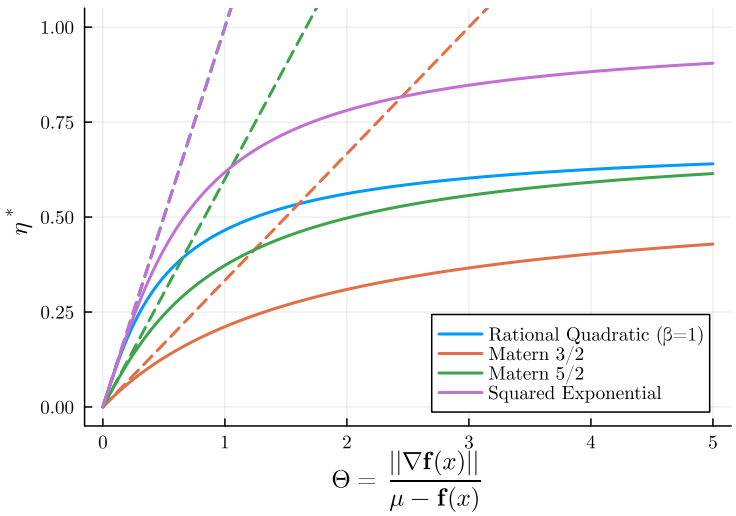


FIGURE 5.2: RFD step sizes as a function of $\Theta = \frac{\|\nabla \mathbf{f}(x)\|}{\mu - \mathbf{f}(x)}$ assuming scale $s = 1$ (cf. Table 5.1). A-RFD (Definition 5.3.1) is plotted as dashed lines. A-RFD of the rational quadratic coincides with A-RFD of the squared exponential covariance.

positive. Towards the end of minimization this difference no longer changes as the cost no longer decreases. I.e. towards the end the gradient cost quotient Θ is roughly linear in the gradient $\|\nabla \mathbf{f}(x)\|$. The derivative $\frac{d}{d\Theta}\eta^*(0)$ of $\eta^*(\Theta)$ at zero then effectively results in a constant asymptotic learning rate.

5.3.1 Asymptotic learning rate

To explain the claim above, note that the gradient cost quotient Θ converges to zero towards the end of optimization, because the gradient norm converges to zero. A first order Taylor expansion of η^* would therefore imply

$$\eta^*(\Theta) \approx \eta^*(0) + \tfrac{d}{d\Theta}\eta^*(0)\Theta = \underbrace{\tfrac{\frac{d}{d\Theta}\eta^*(0)}{\mu - \mathbf{f}(x)}}_{\text{asymptotic learning rate}} \|\nabla \mathbf{f}(x)\|$$

assuming $\eta^*(0) = 0$ and differentiability of η^*, which is a reasonable educated guess based on the the examples in Figure 5.2. But since the RFD step sizes η^* are abstractly defined as an arg min, it is necessary to formalize this intuition for general covariance models. First, we define asymptotic step sizes as an object towards which we can prove convergence. Then we prove convergence, proving they are well defined. In addition, we obtain a more explicit formula for the asymptotic learning rate.

Definition 5.3.1 (A-RFD)**.** We define the step sizes of "asymptotic RFD" (A-RFD) to be the minimizer of the second order Taylor approximation $T_2 q_\Theta$ of q_Θ around zero

$$\hat{\eta}(\Theta) := \arg\min_\eta T_2 q_\Theta(\eta) = \tfrac{C(0)}{-C'(0)}\Theta = \underbrace{\tfrac{C(0)}{C'(0)(\mathbf{f}(x) - \mu)}}_{\text{asymptotic learning rate}} \|\nabla \mathbf{f}(x)\|.$$

In the following we prove that these are truly asymptotically equal to the step sizes η^* of RFD.

Proposition 5.3.2 (A-RFD is well defined)**.** *Let $\mathbf{f} \sim \mathcal{N}(\mu, C)$ and assume there exists $\eta_0 > 0$ such that the correlation for larger distances $\eta \geq \eta_0$ are bounded smaller than 1, i.e. $\frac{C(\eta^2/2)}{C(0)} < \rho \in (0, 1)$. Then the step sizes of RFD are asymptotically equal to the step sizes of A-RFD, i.e.*

$$\hat{\eta}(\Theta) \sim \eta^*(\Theta) \quad \textit{as} \quad \Theta \to 0.$$

Note that the assumption is essentially always satisfied, since the Cauchy-Schwarz inequality implies

$$C\big(\tfrac{\|x - \tilde{x}\|^2}{2}\big) = \operatorname{Cov}(\mathbf{f}(x), \mathbf{f}(\tilde{x})) \leq \sqrt{\operatorname{Var}(\mathbf{f}(x)) \operatorname{Var}(\mathbf{f}(\tilde{x}))} = C(0),$$

where equality requires the random variables to be almost surely equal.[12] If the random function is not periodic or constant, this will generally be strict. In the proof, this requirement is only needed to ensure that η^* is not very large. The smallest local minimum of q_Θ is always close to $\hat{\eta}$ even without this assumption (which ensures it is a global minimum).

[12] Klenke, 2014.

Figure 5.2 illustrates that $\eta^* \to 0$ should imply $\Theta \to 0$, resulting in a weak convergence guarantee.

Corollary 5.3.3. *Assume $\eta^* \to 0$ implies $\Theta \to 0$, the cost $\mathbf{f}$ is bounded, has continuous gradients and RFD converges to some point x_∞. Then x_∞ is a critical point and the RFD step sizes η^* are asymptotically equal to $\hat{\eta}$.*

For the squared exponential covariance model we formally prove that η^* is strictly monotonously increasing in Θ and thus $\eta^* \to 0$ implies $\Theta \to 0$ (Prop. 5.C.3). The 'bounded' and 'continuous gradients' assumptions are almost surely satisfied for all sufficiently smooth covariance functions,[13] where three times differentiable is more than enough smoothness.

[13] cf. Adler and Taylor, 2007.

5.3.2 RFD step sizes explain common step size heuristics

Asymptotically, RFD suggests constant learning rates, similar to the classical L-smooth setting. We thus define these asymptotic learning rates (as the limit of the learning rates h_n of iteration n) to be

$$h_\infty := \frac{C(0)}{C'(0)(\mathbf{f}(x_\infty) - \mu)}, \tag{5.3}$$

where $\mathbf{f}(x_\infty)$ is the cost we reach in the limit. If we used these asymptotic learning rates from the start, step sizes would become too large for large gradients, as RFD step sizes exhibit a plateau (cf. Figure 5.2). To emulate the behavior of RFD with a piecewise linear function, we could introduce a cutoff whenever our step size exceeds the initial step size $\eta^*(\infty)$, i.e.

$$x_{n+1} = x_n - \min\left\{h_\infty, \frac{\eta^*(\infty)}{\|\nabla \mathbf{f}(x_n)\|}\right\}\nabla \mathbf{f}(x_n). \qquad \text{(gradient clipping)}$$

At this point we have rediscovered 'gradient clipping'.[14] As the rational quadratic covariance has the same asymptotic learning rate h_∞ for every β, its parameter β controls the step size bound $\eta^*(\infty)$ of gradient clipping (cf. Table 5.1, Figure 5.2).

[14] Pascanu et al., 2013.

Pascanu et al. (2013) motivated gradient clipping with the geometric interpretation of movement towards a 'wall' placed behind the minimum. This suggests that clipping should happen towards the end of training. This stands in contrast to a more recent step size heuristic, "(linear) warmup",[15] which suggests smaller learning rates at the start (i.e. $h_0 = \frac{\eta^*(\infty)}{\|\nabla \mathbf{f}(x_0)\|}$) and gradual ramp-up to the asymptotic learning rate h_∞. In other words, gradients are not clipped due to some wall next to the minimum, but because the step sizes would be too large at the start otherwise. Goyal et al. (2018) further observe that 'constant warmup' (i.e. a step change of learning rates akin to gradient clipping) performs worse than gradual warmup. Since RFD step sizes suggest this gradual increase, we argue that they may have discovered RFD step sizes empirically (also cf. Figure 5.3).

[15] Goyal et al., 2018.

5.4 Mini-batch loss and covariance estimation

In practice we typically do not have access to evaluations of the cost $\mathbf{f}$ but rather access to noisy evaluations $Y_n = \mathrm{J}^k \mathbf{Y}_n(X_n)$ with $\mathbf{Y}_n(x) = \mathbf{f}(x) + \varsigma_n(x)$, where we typically assume the noise ς_n is centered and iid conditional on $\mathbf{f}$.[16] But at the cost of multiple evaluations the variance of the noise can be reduced. Specifically, let b_n be the number of evaluations (the 'mini-batch size') used to produce the observation $\mathbf{Y}_n$, then

[16] More precisely iid conditional on some $\mathbf{M}$ such hat $\mathbf{f}$ is measurable w.r.t. $\mathbf{M}$, see Definition 2.2.2

$$\mathbf{Y}_n(x) = \frac{1}{b_n}\sum_{i=1}^{b_n} \mathbf{f}(x) + \varsigma_n^{(i)}(x) = \mathbf{f}(x) + \underbrace{\frac{1}{b_n}\sum_{i=1}^{b_n} \varsigma_n^{(i)}(x)}_{=:\varsigma_n(x)}, \tag{5.4}$$

where we assume the noise $\varsigma_n^{(i)}$ is iid and centered conditional on $\mathbf{f}$. This implies

$$\mathrm{Var}(\mathbf{Y}_n(x)) = \mathrm{Var}(\mathbf{f}(x)) + \tfrac{1}{b_n}\mathrm{Var}(\varsigma_n^{(1)}(x)) \overset{\text{isotropy}}{=} C(0) + \tfrac{1}{b_n} C_\varsigma(0), \tag{5.5}$$

where we assume $\mathbf{f} \sim \mathcal{N}(\mu, C)$ and $\varsigma_n^{(i)} \sim \mathcal{N}(0, C_\varsigma)$ in the last equation for simplicity.[17] We therefore have control over the variance of ς_n. This will help us estimate both C and C_ς from evaluations.

[17] Although this step did not yet require the distributional Gaussian assumption beyond the mean and variance.

5.4.1 Variance estimation

Recall that the asymptotic learning rate h_∞ in Equation (5.3) only depends on $C(0)$ and $C'(0)$. So if we estimate these values, we are certain to get the right RFD step sizes asymptotically without knowing the entire covariance kernel C.

Equation (5.5) reveals that for $Z_n := (\mathbf{Y}_n(x) - \mu)^2$ we have

$$\mathbb{E}[Z_n] = \beta_0 + \tfrac{1}{b_n}\beta_1 \qquad \text{i.e.} \qquad Z_b = \beta_0 + \tfrac{1}{b_n}\beta_1 + \text{noise}$$

with bias $\beta_0 = C(0)$ and slope $\beta_1 = C_\epsilon(0)$. So a linear regression on samples $(\frac{1}{b_k}, Z_k)_{k\leq n}$ allows for the estimation of β_0 and β_1.

As the noise functions $\varsigma_k^{(i)}$ are all conditionally independent and therefore uncorrelated, we have

$$\operatorname{Cov}(\mathbf{Y}_k(x_k), \mathbf{Y}_l(x_l)) = \operatorname{Cov}(\mathbf{f}(x_k), \mathbf{f}(x_l)) = C\big(\tfrac{\|x_k - x_l\|^2}{2}\big)$$

Since the covariance kernel C is typically monotonously falling in the distance of parameters $\|x_k - x_l\|^2$, we want to select them as spaced out as possible to minimize the covariance of $\mathbf{Y}_k(x_k)$ to minimize the bias. Randomly selecting x_i with Glorot initialization[18] ensures a good spread.[19]

By the Gaussian assumption the variance of Z_n is the (centered) fourth moment of $\mathbf{Y}_n$, which is given by

$$\sigma_{b_n}^2 := \operatorname{Var}(Z_n) = \mathbb{E}[Z_n^4] - \mathbb{E}[Z_n^2]^2 = 2\operatorname{Var}(\mathbf{Y}_n(x))^2 = 2(\beta_0 + \tfrac{1}{b_n}\beta_1)^2.$$

In particular the variance of Z_n depends on the batch size b_n. The linear regression is therefore heteroskedastic. Weighted least squares (WLS)[20] is designed to handle this case, but for its application the variance of Z_n is needed. Since β_0, β_1 are the parameters we wish to estimate, we find ourselves in the paradoxical situation that we need β to obtain β. Our solution to this problem is to start with a guess of β_0, β_1, apply WLS to obtain a better estimate and repeat this bootstrapping procedure until convergence.

The same procedure can be applied to obtain $C'(0)$, where the counterpart of Equation (5.5) is given by

$$\operatorname{Var}(\partial_i \mathbf{Y}_n(x)) = \operatorname{Var}(\partial_i \mathbf{f}(x)) + \tfrac{1}{b_n}\operatorname{Var}(\partial_i \varsigma_n^{(1)}(x)) \overset{\text{isotropy}}{=} -(C'(0) + \tfrac{1}{b_n} C_\varsigma'(0)).$$

Remark 5.4.1. Under the isotropy assumption the partial derivatives are iid, so the expectation of $\|\nabla \mathbf{Y}_n(x)\|^2 = \sum_{i=1}^d (\partial_i \mathbf{Y}_n(x))^2$ is this variance scaled by d. In particular the variance needs to scale with $\frac{1}{d}$ to keep the gradient norms (and thus the Lipschitz constant of $\mathbf{f}$) stable. This observation is closely related to "isoperimetry",[21] for details see Benning and Döring, 2024. Removing the isotropy assumption and estimating the variance component-wise is most likely how "adaptive" step sizes,[22] like the ones used by Adam, work (cf. Sec. 5.E.1).

Batch size distribution Before we can apply linear regression to the samples $(\frac{1}{b_k}, Z_k)_{k\leq n}$, it is necessary to choose the batch sizes b_k. As this choice is left to us, we calculate the variance of our estimator $\hat{\beta}_0$ of β_0 explicitly (Lemma 5.B.2), in order to minimize this variance subject to a sample budget α over the selection of batch sizes

$$\min_{n, b_1, \dots, b_n} \operatorname{Var}(\hat{\beta}_0) \quad \text{s.t.} \quad \underbrace{\sum_{k=1}^n b_k}_{\text{samples used}} \leq \alpha. \tag{5.6}$$

Since this optimization problem is very difficult to solve, we rephrase it in terms of the empirical distribution of batch sizes $\nu_n = \frac{1}{n}\sum_{i=1}^n \delta_{b_i}$. Optimizing over distributions is still difficult, but we explain in Section 5.B how to heuristically arrive at the parametrization

$$\nu(b) \propto \exp\big(\lambda_1 \tfrac{1}{\sigma_b^2} - \lambda_2 b\big), \qquad b \in \mathbb{N}$$

where the parameters $\lambda_1, \lambda_2 \geq 0$ can then be used to optimize (5.6). Due to our usage of σ_b^2 this has to be bootstrapped.

[18] Glorot and Bengio, 2010.

[19] Note that Glorot initialization places all parameters approximately on the same sphere. This is because Glorot initialization initializes all parameters independently, therefore their norm is the sum of independent squares, which converges by the law of large numbers due to the normalization Glorot uses. Since stationary isotropy and non-stationary isotropy coincides on the sphere, this is an important effect to consider (cf. Section 4.2).

[20] e.g. Kay, 1993, Theorem 4.2.

[21] e.g. Bubeck and Sellke, 2021.

[22] e.g. Duchi et al., 2011; Diederik P. Kingma and Ba, 2015.

Covariance estimation While the variance estimates above ensure correct asymptotic learning rates, we motivated in Section 5.3.2 that asymptotic learning rates alone would result in too large step sizes at the beginning. We therefore use the estimates of $C(0)$ and $C'(0)$ to fit a covariance model, effectively acting as a gradient clipper while retaining the asymptotic guarantees. Note that covariance models with less than two parameters are generally fully determined by these values.

5.4.2 Stochastic RFD (S-RFD)

It is reasonable to ask whether there is a 'stochastic gradient descent'-like counterpart to the 'gradient descent'-like RFD. The answer is yes, and we already have all the required machinery.

Extension 5.4.2 (S-RFD). *For loss* $\mathbf{f} \sim \mathcal{N}(\mu, C)$ *and stochastic errors* $\varsigma_i \overset{\text{iid}}{\sim} \mathcal{N}(0, C_\varsigma)$ *we have*

$$\arg\min_{\mathbf{d}} \mathbb{E}[\mathbf{f}(x - \mathbf{d}) \mid \mathbf{Y}_n(x), \nabla \mathbf{Y}_n(x)] = \eta^*(\Theta) \frac{\nabla \mathbf{Y}_n(x)}{\|\nabla \mathbf{Y}_n(x)\|}$$

with the same step size function η^* *as for RFD, but modified* Θ

$$\Theta = \frac{C'(0)}{C'(0) + \frac{1}{b_n} C'_\varsigma(0)} \frac{C(0) + \frac{1}{b_n} C_\varsigma(0)}{C(0)} \frac{\|\nabla \mathbf{Y}_n(x)\|}{\mu - \mathbf{Y}_n(x)}.$$

Note, that our non-parametric covariance estimation already provides us with estimates of $C_\varsigma(0)$ and $C'_\varsigma(0)$, so no further adaptions are needed. The resulting asymptotic learning rate is given by

$$h_\infty = \frac{C(0) + \frac{1}{b_n} C_\varsigma(0)}{(C'(0) + \frac{1}{b_n} C'_\varsigma(0))(\mathbf{Y}_n(x_\infty) - \mu)}. \tag{5.7}$$

5.5 MNIST case study

For our case study we use the negative log likelihood loss to train a neural network[23] on the MNIST dataset.[24] We choose this model as one of the simplest state-of-the-art models at the time of selection, consisting only of convolutional layers with ReLU activation interspersed by batch normalization layers and a single dense layers at the end with softmax activation. Assuming isotropy, we estimate μ, $C(0)$ and $C'(0)$ as described in Section 5.4.1 and deduce the parameters σ^2 and s of the respective covariance model. We then use the step sizes listed in Table 5.1 for the 'squared exponential' and 'rational quadratic' covariance in our RFD algorithm.

[23] An et al., 2020, M7.

[24] LeCun et al., 2010.

In Figure 5.3, RFD is benchmarked against step size tuned Adam[25] and stochastic gradient descent (SGD). Even with early stopping, their tuning would typically require more than 1 epoch worth of samples, *in contrast to RFD* (Section 5.A.1). We highlight that A-RFD performs significantly worse than either of the RFD versions which effectively implement some form of learning rate warmup. This is despite the RFD learning rates converging to the asymptotic one within one epoch (ca. 30 out of 60 steps per epoch). The step sizes on the other hand are (up to noise) monotonously decreasing. This stands in contrast to the "wall next to the minimum" motivation of gradient clipping.

[25] Diederik P. Kingma and Ba, 2015.

Code availability: Our implementation of RFD can be found at `https://github.com/FelixBenning/pyrfd` and the package can also be installed from PyPI via '`pip install pyrfd`'.

5.6 Limitations and extensions

To cover the vast amount of ground that lays between the 'formulation of a general average case optimization problem' and the 'prototype of a working optimizer with theoretical backing',

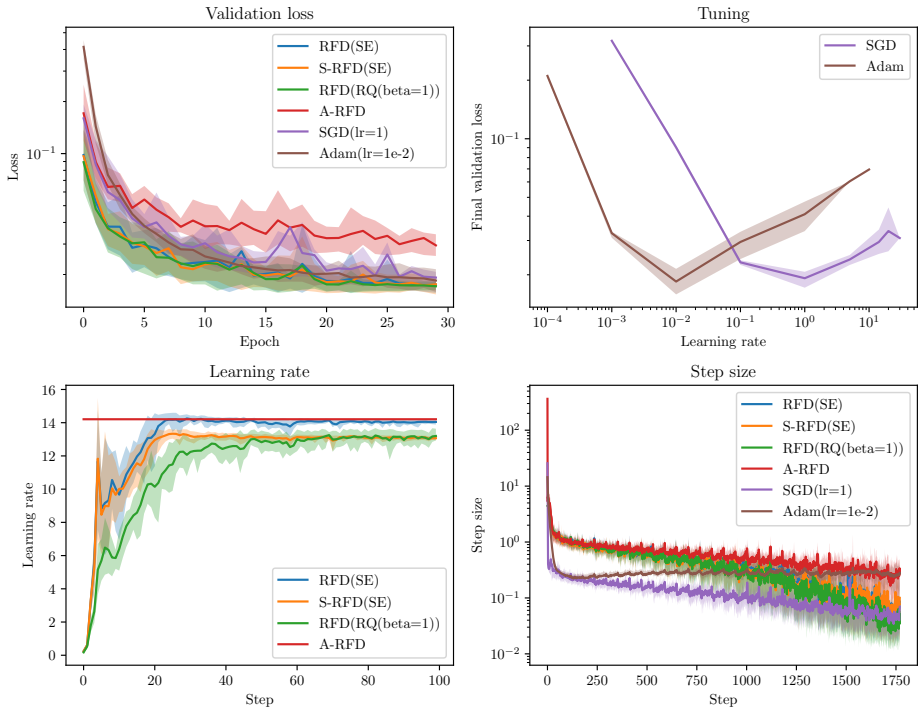


FIGURE 5.3: Training on the MNIST dataset (batch size 1024). Ribbons describe the range between the 10% and 90% quantile of 20 repeated experiments while lines represent their mean. SE stands for the squared exponential (5.11) and RQ for the rational quadratic (5.13) covariance. The validation loss uses the test data set, which provides a small advantage to Adam and SGD, as we also use it for tuning.

1. we used the common[26] *isotropic* and *Gaussian* distributional assumption for $\mathbf{f}$,
2. we used very *simple covariance models* for the actual implementation,
3. we used WLS in our variance estimation procedure despite the *violation of independence*.

[26] Frazier, 2018; Wang et al., 2016; Wu et al., 2017; Dauphin et al., 2014.

Since RFD is defined as the minimizer of an average instead of an upper bound – making it more risk affine – it naturally loses the improvement guarantee driving classical convergence proofs. It is therefore impossible to extend classical optimization proofs and new mathematical theory must be developed. This risk-affinity can also be observed in its comparatively large step sizes (cf. Fig. 5.3 and Sec. 5.A). On CIFAR-100,[27] the step sizes were *too* large and it is an open question whether assumptions were violated or whether RFD is simply too risk-affine. But since the variance of random functions vanishes asymptotic with high dimension[28] we highly suspect the former, see Remark 5.E.5 for further details.

[27] Krizhevsky, 2009.

[28] see Chapter 6

Future work will therefore have to target these assumptions. Some of the assumptions were already simplifications for the sake of exposition, and we deferred their relaxation to the appendix. The Gaussian assumption can be relaxed with a generalization to the 'BLUE' (Sec. 5.E.3), isotropy can be generalized to 'geometric anisotropies' (Sec. 5.E.1) and the risk-affinity of RFD can be reduced with confidence intervals (Sec. 5.E.2). Since simple random linear models already violate stationary isotropy (Sec. 7.4), we believe that stationarity is the most important assumption to attack in future work.

5.7 Conclusion

In this paper we have demonstrated the **viability** (computability and scalable complexity) and **advantages** (scale invariance, explainable step size schedule which does not require expensive tuning) of replacing the classical "convex function" framework with the "random function" framework. Along the way we bridged the gap between Bayesian optimization (not scalable so far) and classical optimization methods (scalable). This theoretical framework not only sheds light on existing step size heuristics, but can also be used to develop future heuristics.

We envision the following improvements to RFD in the future:

1. The *reliability* of RFD can be improved by generalizing the distributional assumptions to cover more real world scenarios. In particular we are interested in the generalization to non-stationary isotropy because we suspect that regularization such as weight and batch normalization[29] are used to patch violations of stationarity (cf. Section 7.4).

2. The *performance* of RFD can also be improved. Since RFD is forgetful while momentum methods retains some information it is likely fruitful to relax the full forgetfulness. Furthermore, we suspect that adaptive learning rates,[30] such as those used by Adam, can be incorporated with geometric anisotropies (cf. Sec. E.1). Performance could also be further improved by estimating the covariance (locally) online instead of globally at the start. Finally, the implementation itself can be made more performant.

[29] Salimans and Durk P Kingma, 2016; Ioffe and Szegedy, 2015.

[30] e.g. Duchi et al., 2011; Diederik P. Kingma and Ba, 2015.

Acknowledgement

We extend our sincere gratitude to our colleagues at the University of Mannheim, with special thanks to Rainer Gemulla and Julie Naegelen for insightful discussions and invaluable feedback. The Experiments in this work were partially carried out on the compute cluster of the state of Baden-Würtemberg (bwHPC).

References

Adler, Robert J. and Jonathan E. Taylor (2007). *Random Fields and Geometry*. Springer Monographs in Mathematics. New York, NY: Springer New York. ISBN: 978-0-387-48112-8. DOI: 10.1007/978-0-387-48116-6.

An, Sanghyeon et al. (2020). *An Ensemble of Simple Convolutional Neural Network Models for MNIST Digit Recognition*. DOI: 10.48550/arXiv.2008.10400. Pre-published.

Benning, Felix and Leif Döring (2024). *Gradient Span Algorithms Make Predictable Progress in High Dimension*. arXiv: 2410.09973. Pre-published.

Bubeck, Sebastien and Mark Sellke (2021). "A Universal Law of Robustness via Isoperimetry". In: *Advances in Neural Information Processing Systems*. Vol. 34. Virtual Event: Curran Associates, Inc., pp. 28811–28822. arXiv: 2105.12806 [cs, stat]. URL: https://proceedings.neurips.cc/paper/2021/hash/f197002b9a0853eca5e046d9ca4663d5-Abstract.html (visited on 2023-09-22).

Dauphin, Yann N et al. (2014). "Identifying and Attacking the Saddle Point Problem in High-Dimensional Non-Convex Optimization". In: *Advances in Neural Information Processing Systems*. Vol. 27. Montréal, Canada: Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper/2014/hash/17e23e50bedc63b4095e3d8204ce063b-Abstract.html (visited on 2022-06-10).

Deuflhard, P. and G. Heindl (1979). "Affine Invariant Convergence Theorems for Newton's Method and Extensions to Related Methods". In: *SIAM Journal on Numerical Analysis* 16.1, pp. 1–10. ISSN: 0036-1429. DOI: 10.1137/0716001.

Duchi, John, Elad Hazan, and Yoram Singer (2011). "Adaptive Subgradient Methods for Online Learning and Stochastic Optimization". In: *The Journal of Machine Learning Research* 12, pp. 2121–2159. ISSN: 1532-4435.

Frazier, Peter I. (2018). "Bayesian Optimization". In: *Recent Advances in Optimization and Modeling of Contemporary Problems*. INFORMS TutORials in Operations Research. Phoenix, Arizona, USA: INFORMS, pp. 255–278. ISBN: 978-0-9906153-2-3. DOI: 10.1287/educ.2018.0188. arXiv: 1807.02811 [cs, math, stat].

Gao, Fuchang and Lixing Han (2012). "Implementing the Nelder-Mead Simplex Algorithm with Adaptive Parameters". In: *Computational Optimization and Applications* 51.1, pp. 259–277. ISSN: 1573-2894. DOI: 10.1007/s10589-010-9329-3.

Glorot, Xavier and Yoshua Bengio (2010). "Understanding the Difficulty of Training Deep Feedforward Neural Networks". In: *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*. Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics. Sardinia, Italy: JMLR Workshop and Conference Proceedings, pp. 249–256. URL: https://proceedings.mlr.press/v9/glorot10a.html (visited on 2023-04-11).

Goyal, Priya et al. (2018). *Accurate, Large Minibatch SGD: Training ImageNet in 1 Hour*. arXiv. arXiv: 1706.02677 [cs]. URL: http://arxiv.org/abs/1706.02677 (visited on 2024-04-02).

Hansson, Anders (2023). *Optimization for Learning and Control*. Hoboken, New Jersey: John Wiley & Sons, Inc. ISBN: 978-1-119-80914-2.

Hinton, Geoffrey (2012). "Neural Networks for Machine Learning". Massive Open Online Course (Coursera). URL: https://www.cs.toronto.edu/~hinton/coursera_lectures.html (visited on 2021-11-16).

Ioffe, Sergey and Christian Szegedy (2015). "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift". In: *Proceedings of the 32nd International Conference on Machine Learning*. International Conference on Machine Learning. PMLR, pp. 448–456. arXiv: 1502.03167 [cs]. URL: https://proceedings.mlr.press/v37/ioffe15.html (visited on 2021-10-06).

Jaynes, E. T. (1957). "Information Theory and Statistical Mechanics". In: *Physical Review* 106.4, pp. 620–630. DOI: 10.1103/PhysRev.106.620.

Johnson, Richard Arnold and Dean W. Wichern (2007). *Applied Multivariate Statistical Analysis*. 6th ed. Upper Saddle River, N.J: Pearson College Div. 767 pp. ISBN: 978-0-13-187715-3.

Kay, Steven M. (1993). *Fundamentals of Statistical Signal Processing: Estimation Theory*. USA: Prentice-Hall, Inc. 595 pp. ISBN: 978-0-13-345711-7.

Kingma, Diederik P. and Jimmy Ba (2015). "Adam: A Method for Stochastic Optimization". In: *Proceedings of the 3rd International Conference on Learning Representations*. ICLR. San Diego. arXiv: 1412.6980.

Klenke, Achim (2014). *Probability Theory: A Comprehensive Course*. Universitext. London: Springer. ISBN: 978-1-4471-5360-3 978-1-4471-5361-0. DOI: 10.1007/978-1-4471-5361-0.

Krizhevsky, Alex (2009). *Learning Multiple Layers of Features from Tiny Images*. URL: https://www.cs.toronto.edu/%20kriz/learning-features-2009-TR.pdf (visited on 2024-05-21).

LeCun, Yann, Corinna Cortes, and Christopher J.C. Burges (2010). *THE MNIST DATABASE of Handwritten Digits*. URL: http://yann.lecun.com/exdb/mnist/ (visited on 2024-01-24).

Nesterov, Yurii Evgen'evič (2018). *Lectures on Convex Optimization*. Second edition. Springer Optimization and Its Applications; Volume 137. Cham: Springer. ISBN: 978-3-319-91578-4. DOI: 10.1007/978-3-319-91578-4.

Pascanu, Razvan, Tomas Mikolov, and Yoshua Bengio (2013). "On the Difficulty of Training Recurrent Neural Networks". In: *Proceedings of the 30th International Conference on Machine Learning*. International Conference on Machine Learning. Atlanta: PMLR, pp. 1310–1318. URL: https://proceedings.mlr.press/v28/pascanu13.html (visited on 2024-04-02).

Rasmussen, Carl Edward and Christopher K.I. Williams (2006). *Gaussian Processes for Machine Learning*. 2nd ed. Adaptive Computation and Machine Learning 3. Cambridge, Massachusetts: MIT Press. 248 pp. ISBN: 0-262-18253-X. URL: http://gaussianprocess.org/gpml/chapters/RW.pdf (visited on 2023-04-06).

Roos, Filip de, Alexandra Gessner, and Philipp Hennig (2021). "High-Dimensional Gaussian Process Inference with Derivatives". In: *Proceedings of the 38th International Conference on Machine Learning*. International Conference on Machine Learning. PMLR, pp. 2535–2545. URL: https://proceedings.mlr.press/v139/de-roos21a.html (visited on 2023-05-15).

Salimans, Tim and Durk P Kingma (2016). "Weight Normalization: A Simple Reparameterization to Accelerate Training of Deep Neural Networks". In: *Advances in Neural Information Processing Systems*. Vol. 29. Barcelona, Spain: Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper/2016/hash/ed265bc903a5a097f61d3ec064d96d2e-Abstract.html (visited on 2023-10-16).

Smith, Leslie N. (2018). *A Disciplined Approach to Neural Network Hyper-Parameters: Part 1 – Learning Rate, Batch Size, Momentum, and Weight Decay*. DOI: 10.48550/arXiv.1803.09820. arXiv: 1803.09820 [cs, stat]. Pre-published.

Stein, Michael L. (1999). *Interpolation of Spatial Data*. Springer Series in Statistics. New York, NY: Springer. ISBN: 978-1-4612-7166-6 978-1-4612-1494-6. DOI: 10.1007/978-1-4612-1494-6.

Wang, Ziyu et al. (2016). "Bayesian Optimization in a Billion Dimensions via Random Embeddings". In: *Journal of Artificial Intelligence Research* 55, pp. 361–387. ISSN: 1076-9757. DOI: 10.1613/jair.4806.

Wu, Jian et al. (2017). "Bayesian Optimization with Gradients". In: *Advances in Neural Information Processing Systems*. Vol. 30. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper/2017/hash/64a08e5f1e6c39faeb90108c430eb120-Abstract.html (visited on 2022-06-02).

Xiao, Han, Kashif Rasul, and Roland Vollgraf (2017). *Fashion-MNIST: A Novel Image Dataset for Benchmarking Machine Learning Algorithms*. DOI: 10.48550/arXiv.1708.07747. arXiv: 1708.07747 [cs, stat]. Pre-published.

Zeiler, Matthew D. (2012). "ADADELTA: An Adaptive Learning Rate Method". arXiv: 1212.5701 [cs].

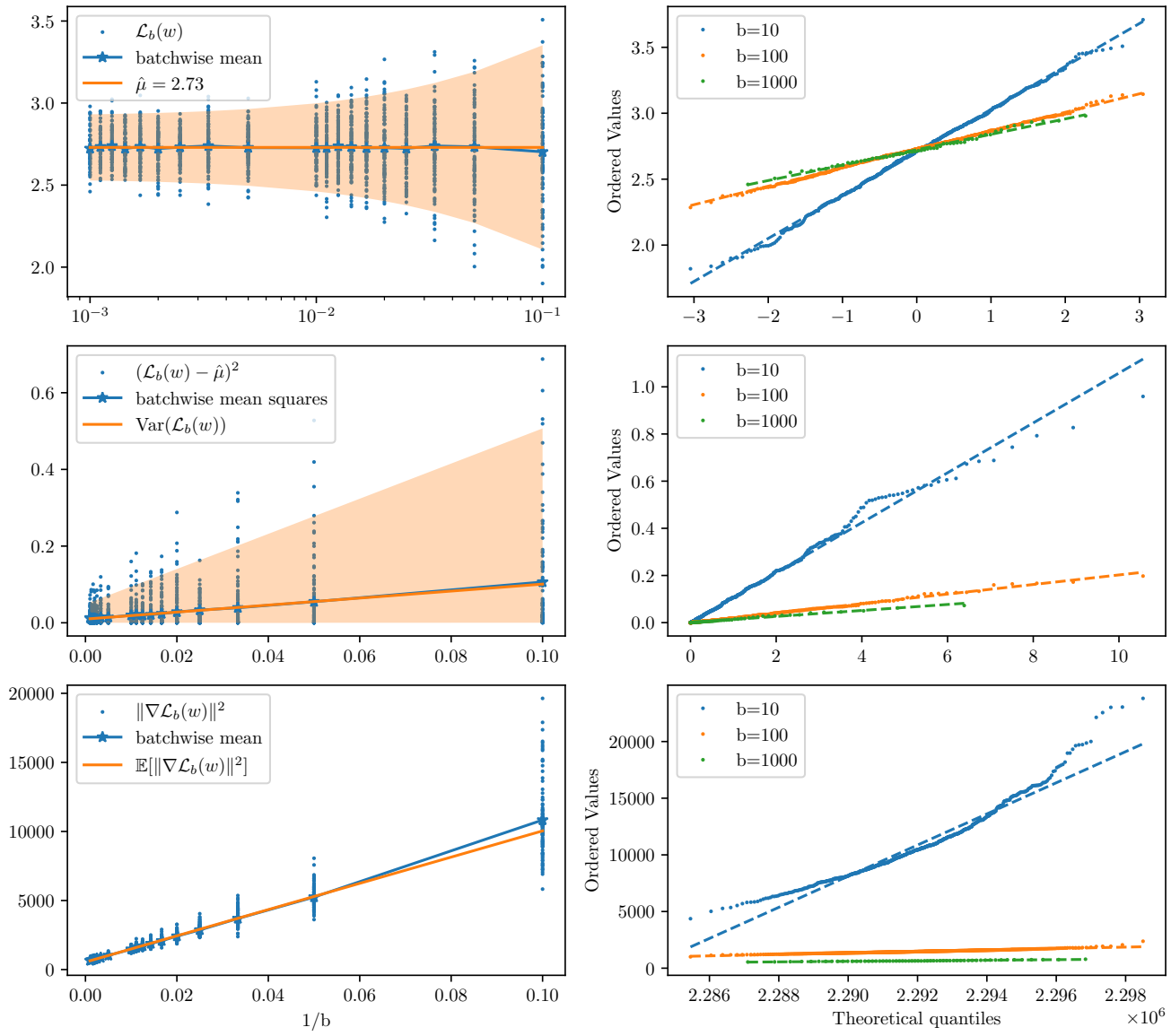


FIGURE 5.4: Visualization of the variance estimation (Section 5.4.1) with 95%-confidence intervals based on the assumed distribution. Here, $\mathcal{L}_b$ are the mini-batch losses of batch-size b that we more generally denote as the observations $\mathbf{Y}_n$ with batch-size b_n. Quantile-quantile (QQ) plots of the losses (against a normal distribution), squared losses (against a $\chi^2(1)$ distribution) and squared gradient norms (against a $\chi^2(d)$-distribution) are displayed on the right for a selection of batch sizes.

Appendix: Random Function Descent

5.A Experiments

5.A.1 Covariance estimation

In Figure 5.4 we visualize weighted least squares (WLS) regression of the covariance estimation from Section 5.4.1. Note, that we sampled much more samples per batch size for these plots than RFD would typically require by itself in order to be able to plot batch-wise means and batch-wise QQ-plots. The batch size distribution we described in Section 5.B would avoid sampling the same batch size multiple times to ensure better stability of the regression and generally requires much fewer samples than were used for this visualization (cf. 5.A.1)

We can observe from the QQ-plots on the right, that the Gaussian assumption is essentially justified for the losses, resulting in a $\chi^2(1)$ distribution for the squared losses and a $\chi^2(d)$ distribution for the gradient norms squared. The confidence interval estimate for the squared norms appears to be much too small (it is plotted, but too small to be visible). Perhaps this signifies a violation of the isotropy

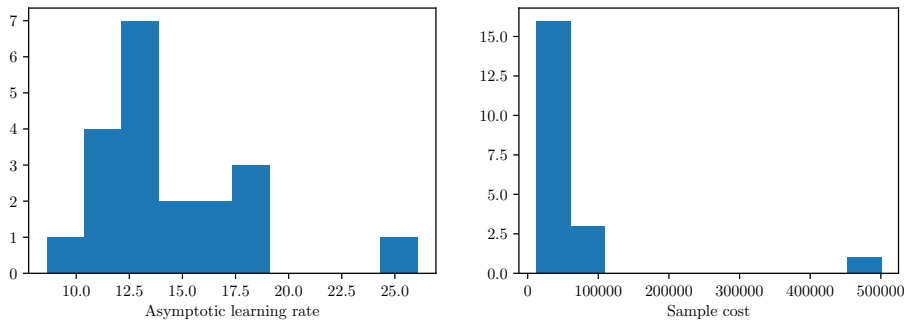


FIGURE 5.5: 20 repeated covariance estimations of model M7 (An et al., 2020) applied to the MNIST dataset. On the left are the resulting asymptotic learning rates (assuming a final loss of zero) and on the right are the samples used until the stopping criterion interrupted sampling.

assumption as the variance of

$$\|\nabla \mathbf{Y}_n(x)\|^2 = \sum_{i=1}^{d} (\partial_i \mathbf{Y}_n(x))^2$$

does not appear to be the variance of independent $\chi^2(d)$ Gaussian random variables, and the independence only follows from the isotropy assumption.

Sampling efficiency and stability

To evaluate the sampling efficiency and stability of our variance estimation process, we repeated the covariance estimation of the model model M7 (An et al., 2020) applied to the MNIST dataset 20 times (Figure 5.5). We used a tolerance of tol $= 0.3$ as a stopping criterion for the estimated relative standard deviation (5.10).

At this tolerance, the asymptotic learning rate already seems relatively stable (in the same order of magnitude) and the sample cost is quite cheap. The majority of runs (16/20 runs or 80%) required less than 60 000 samples (1 epoch). There was one large outlier which used 500 589 samples. A closer inspection revealed, that after the initial sample to estimate the optimal batch size distribution, it sampled almost exclusively at batch sizes 20 (which was the minimal cutoff to avoid instabilities caused by batch normalization) and batch sizes between 1700 and 1900. It therefore seems like the initial batch of samples caused a very unfavorable batch size distribution which then required a lot of samples to recover from. Our selection of an initial sample size of 6000 might therefore have been too small.

A more extensive empirical study is needed to tune this estimation process, but the process promises to be very sample efficient. Classical step size tuning would train models for a short duration in order to evaluate the performance of a particular learning rate (e.g. Smith, 2018), but a single epoch worth of samples is very hard to beat.

Our implementation of this process on the other hand is very inefficient as of writing. Piping data of differing batch sizes into a model is not a standard use case. We implement this by repeatedly initializing data loaders, which is anything but performance friendly.

5.A.2 Other models and datasets

To estimate the effect of the batch size on RFD, we trained the same model (M7 (An et al., 2020)) on MNIST with batch size 128 (Figure 5.6). We can see that the asymptotic learning rate of S-RFD is reduced at a smaller batch size (cf. Equation 5.7) but the performance is barely different. Overall, RFD seems to be slightly too risk-affine, selecting larger step sizes than the tuned SGD models.

We also trained a different model (M5 (An et al., 2020)) on the Fashion MNIST dataset (Xiao et al., 2017) with batch size 128 (Figure 5.7). Since the validation loss increases after epoch 5, early stopping would have been appropriate. We therefore include Adam with learning rate 10^{-3}, despite Adam with learning rate 10^{-4}

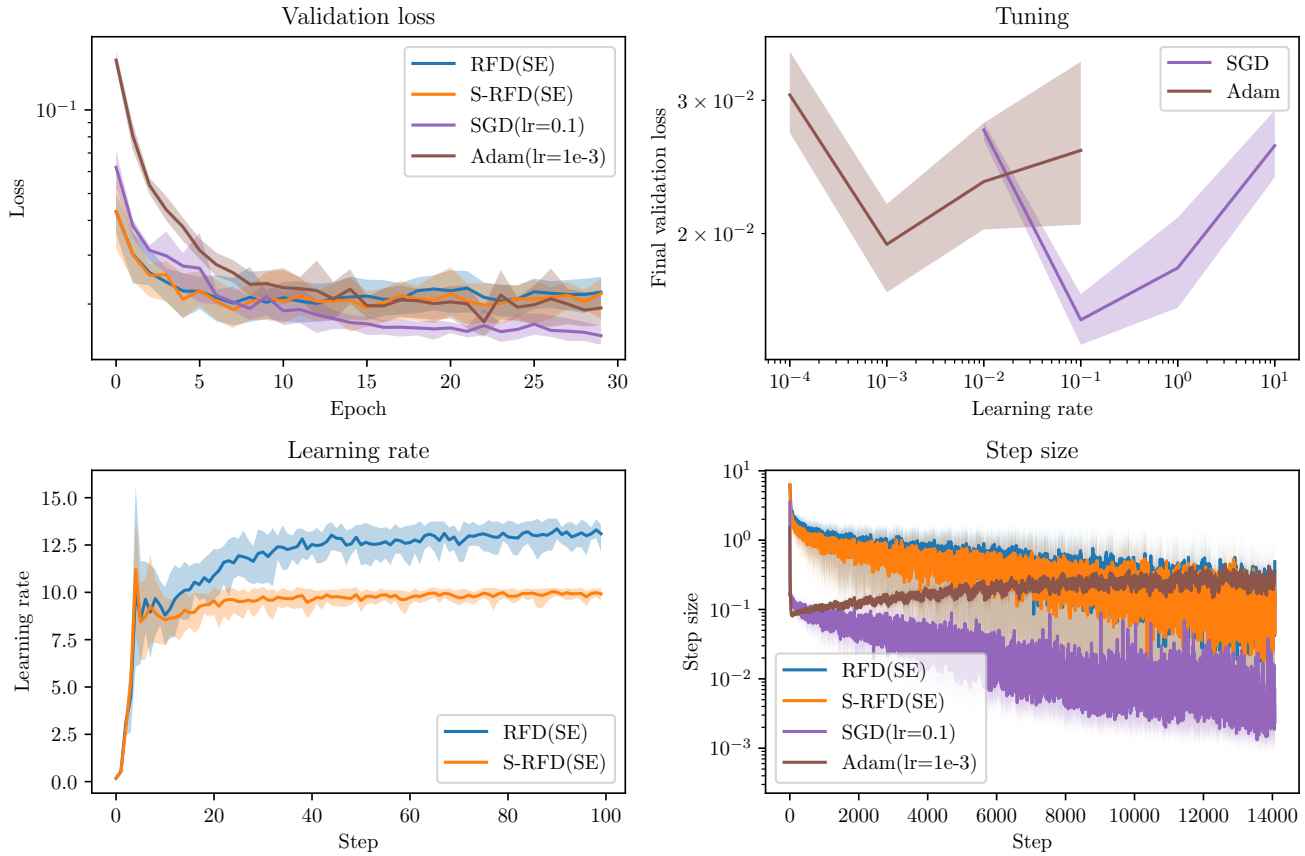


FIGURE 5.6: Training model M7 (An et al., 2020) with batch size 128 on MNIST (LeCun et al., 2010).

technically performing better at the end of training. We can generally see, that RFD comes very close to tuned performance at the time early stopping would have been appropriate. Again, learning rates seem to be slightly too large (risk-affine) in comparison to tuned SGD.

5.B Batch size distribution

Since we plan to use the data set $(\frac{1}{b_k}, \mathbf{Y}_k(x_k))_{k\leq n}$ for weighted least squares (WLS) regression and do not have a selection process for the batch sizes b_k yet, it might be appropriate to select the batch sizes b_k in such a way, that the variance of our estimator $\hat{\beta}_0$ of β_0 is minimized. Here we choose $\text{Var}(\hat{\beta}_0)$ and not $\text{Var}(\hat{\beta}_1)$ as our optimization target, since $\beta_0 = C(0)$ is used to fit the covariance model, while $\beta_1 = C_\epsilon(0)$ is only required for S-RFD. Without deeper analysis β_0 therefore seems to be more important.

Optimization over n parameters b_k is quite difficult, but we can simplify this optimization problem by considering the empirical batch size distribution

$$\nu_n = \frac{1}{n}\sum_{k=1}^{n} \delta_{b_k}.$$

Using a random variable B distributed according to ν_n, the total number of sample losses can then be expressed as

$$\sum_{k=1}^{n} b_k = n\mathbb{E}[B] = \text{samples used}.$$

Under an (unrealistic) independence assumption, the variance $\text{Var}(\hat{\beta}_0)$ also has a simple representation in terms of ν_n (Lemma 5.B.2). We now want to minimize

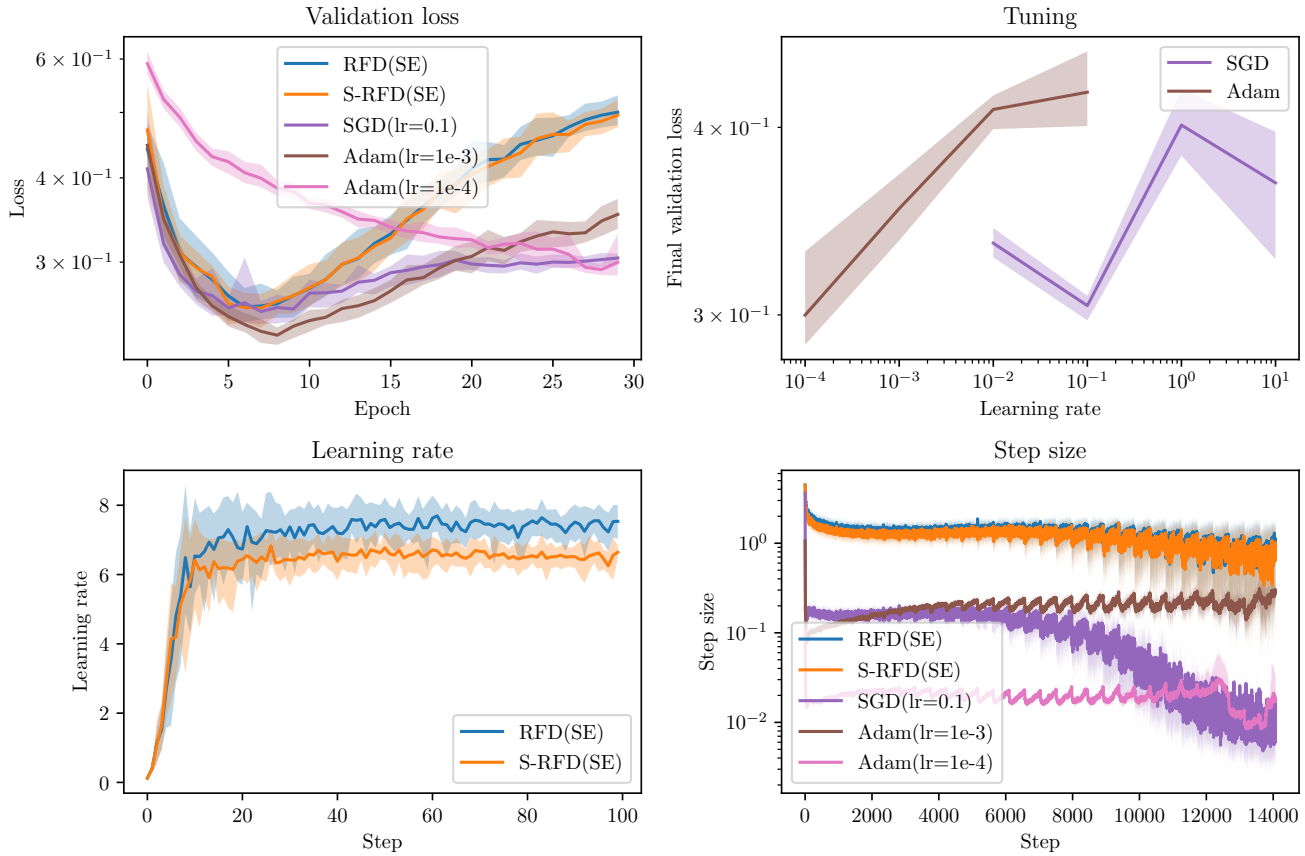


FIGURE 5.7: Model M5 (An et al., 2020) trained on Fashion MNIST (Xiao et al., 2017) with batch size 128.

this variance subject to compute constraint α limiting the number of sample losses we can use resulting in the optimization problem

$$\mathrm{Var}(\hat{\beta}_0) = \frac{1}{n} \underbrace{\frac{1}{\mathbb{E}[\frac{1}{\sigma_B^2}]}}_{\text{'variance of } Z_B\text{'}} \underbrace{\frac{\mathbb{E}[\frac{1}{\sigma_B^2 B^2}]}{\mathbb{E}\left[\frac{1}{\sigma_B^2}\left(\frac{1}{B} - \mathbb{E}[\frac{1}{B\sigma_B^2 \mathbb{E}[1/\sigma_B^2]}]\right)^2\right]}}_{\text{inverse of the 'spread' of } \frac{1}{B}} \quad \text{s.t.} \quad n\mathbb{E}[B] \leq \alpha. \tag{5.8}$$

where we recall that σ_B^2 is the variance of $Z_B = Z_n = (\mathbf{Y}_n(x) - \mu)^2$ if $B = b_n$. Intuitively, the first therm is therefore the average variance of Z_B. The second half is the fraction of a weighted second moment divided by the weighted variance. Unless the mean is at zero, the former will be larger. In particular we want a spread of data otherwise the variance would be zero. This is in some conflict with the variance of Z_B.

But first, let us get rid of n. Note that we would always increase n until our compute budget is used up, since this always reduces variance. So we approximately have $n\mathbb{E}[B] = \alpha$. Thus

$$\mathrm{Var}(\hat{\beta}_0) = \frac{\mathbb{E}[B]}{\alpha} \frac{1}{\mathbb{E}[\frac{1}{\sigma_B^2}]} \frac{\mathbb{E}[\frac{1}{\sigma_B^2 B^2}]}{\mathbb{E}\left[\frac{1}{\sigma_B^2}\left(\frac{1}{B} - \mathbb{E}[\frac{1}{B\sigma_B^2 \mathbb{E}[1/\sigma_B^2]}]\right)^2\right]}$$

Since α is now just resulting in a constant factor, it can be assumed to be 1 without loss of generality. Over batch size distributions ν we therefore want to solve the minimization problem

$$\min_{\nu} \underbrace{\frac{\mathbb{E}[B]}{\mathbb{E}[\frac{1}{\sigma_B^2}]}}_{\text{moments}} \underbrace{\frac{\mathbb{E}[\frac{1}{\sigma_B^2 B^2}]}{\mathbb{E}\left[\frac{1}{\sigma_B^2}\left(\frac{1}{B} - \mathbb{E}[\frac{1}{B\sigma_B^2 \mathbb{E}[1/\sigma_B^2]}]\right)^2\right]}}_{\text{spread}} \tag{5.9}$$

Example 5.B.1 (If we did not require spread)**.** If we were not concerned with the variance of batch sizes, we could select a constant $B = b$. Then it is straightforward to minimize the moments factor manually

$$\min_b \frac{\mathbb{E}[b]}{\mathbb{E}[\frac{1}{\sigma_b^2}]} = b\sigma_b^2 = 2b\big(\beta_0 + \tfrac{1}{b}\beta_1\big)^2,$$

resulting in $\frac{1}{b} = \frac{\beta_0}{\beta_1}$. In other words: If we did not have to be concerned with the spread of B there is one optimal selection to minimize the first factor. But in reality we have to trade-off this target with the spread of B.

To ensure a good spread of data, we use the maximum entropy distribution for B, with the moment constraints

$$\begin{aligned} \mathbb{E}[-B] &\geq -\frac{\alpha}{n} && \text{average sample usage} \\ \mathbb{E}\big[\tfrac{1}{\sigma_B^2}\big] &\geq \theta && Z_B \text{ variance} \end{aligned}$$

which capture the first factor. Maximizing entropy under moment constraints is known[31] to result in the Boltzmann (a.k.a. Gibbs) distribution

[31] Jaynes, 1957.

$$\nu(b) = \mathbb{P}(B = b) \propto \exp\Big(\lambda_1 \frac{1}{\sigma_b^2} - \lambda_2 b\Big),$$

where λ_1, λ_2 depend on the momentum constraints. We can now forget the origin of this distribution and use λ_1, λ_2 as parameters for the distribution ν in Equation (5.9) to get close to its minimum. In practice we use a zero order black box optimizer (Nelder-Mead[32]). One could calculate the expectations of (5.9) under this distribution explicitly and take manual derivatives with respect to λ_i to investigate this further, but we wanted to avoid getting too distracted by this tangent.

[32] Gao and Han, 2012.

We also use the estimated relative standard deviation

$$\text{rel\_std} = \frac{\sqrt{\widehat{\text{Var}}(\hat\beta_0)}}{\hat\beta_0} \tag{5.10}$$

as a stopping criterion for sampling. Without extensive testing we found a tolerance of rel_std $<$ tol $= 0.3$ to be reasonable, cf. Section 5.A.1.

Lemma 5.B.2 (Variance of $\hat\beta_0$ in terms of the empirical batch size distribution)**.** *Assuming independence of the samples $((\frac{1}{b_k}), Z_{b_k})_{k\leq n}$, the variance of $\hat\beta_0$ is given by*

$$\text{Var}(\hat\beta_0) = \frac{1}{n}\frac{1}{\mathbb{E}[\frac{1}{\sigma_B^2}]}\frac{\mathbb{E}[\frac{1}{\sigma_B^2 B^2}]}{\mathbb{E}\Big[\frac{1}{\sigma_B^2}\big(\frac{1}{B} - \mathbb{E}[\frac{1}{B\sigma_B^2\mathbb{E}[1/\sigma_B^2]}]\big)^2\Big]}$$

where B is distributed according to the empirical batch size distribution $\nu_n = \frac{1}{n}\sum_{k=1}^n \delta_{b_k}$.

Proof. With the notation $\sigma_k^2 = \sigma_{b_k}^2$ to describe the variance of Z_{b_k} it follows from[33] that the variance of the estimator $\hat\beta$ of β using n samples is given by

[33] cf. Kay, 1993, Thm. 4.2.

$$\begin{aligned} \text{Var}(\hat\beta) &= (H^T C^{-1} H)^{-1} \\ &= \frac{1}{\big(\sum_k \frac{1}{\sigma_k^2}\big)\big(\sum_k \frac{1}{(\sigma_k b_k)^2}\big) - \big(\sum_k \frac{1}{\sigma_k^2 b_k}\big)^2} \begin{pmatrix} \sum_k \frac{1}{(\sigma_k b_k)^2} & -\sum_k \frac{1}{\sigma_k^2 b_k} \\ -\sum_k \frac{1}{\sigma_k^2 b_k} & \sum_k \frac{1}{\sigma_k^2} \end{pmatrix} \end{aligned}$$

where

$$C := \begin{pmatrix} \sigma_1^2 & & \\ & \ddots & \\ & & \sigma_n^2 \end{pmatrix} \qquad H := \begin{pmatrix} 1 & \frac{1}{b_1} \\ \vdots & \\ 1 & \frac{1}{b_n} \end{pmatrix}.$$

In particular we have

$$\mathrm{Var}(\hat{\beta}_0) = \frac{\sum_k \frac{1}{\sigma_k^2 b_k^2}}{\left(\sum_k \frac{1}{\sigma_k^2}\right)\left(\sum_k \frac{1}{\sigma_k^2 b_k^2}\right) - \left(\sum_k \frac{1}{\sigma_k^2 b_k}\right)^2}.$$

With the help of $\theta := \sum_j \frac{1}{\sigma_j^2}$ and $\lambda_k := \frac{1}{\sigma_k^2 \theta}$, we can reorder the divisor. For this note that since the λ_k sum to 1 we have

$$\begin{aligned}\sum_k \lambda_k \Big(\frac{1}{b_k} - \sum_j \lambda_j \frac{1}{b_j}\Big)^2 &= \sum_k \lambda_k \Big(\frac{1}{b_k^2} - 2\frac{1}{b_k}\sum_j \lambda_j \frac{1}{b_j} + \Big(\sum_j \lambda_j \frac{1}{b_j}\Big)^2\Big) \\ &= \sum_k \lambda_k \frac{1}{b_k^2} - 2\Big(\sum_k \lambda_k \frac{1}{b_k}\Big) + \Big(\sum_k \lambda_k \frac{1}{b_j}\Big)^2 \\ &= \sum_k \lambda_k \frac{1}{b_k^2} - \Big(\sum_k \lambda_k \frac{1}{b_k}\Big)^2\end{aligned}$$

Where the above is essentially the well known statement $\mathbb{E}[(Y - \mathbb{E}[Y])^2] = \mathbb{E}[Y^2] - \mathbb{E}[Y]^2$ for an appropriate selection of Y. This implies that our divisor is given by a weighted variance

$$\theta^2 \sum_k \lambda_k \Big(\frac{1}{b_k} - \sum_j \lambda_j \frac{1}{b_j}\Big)^2 = \theta \sum_k \frac{1}{\sigma_k^2 b_k^2} - \Big(\sum_k \frac{1}{\sigma_k^2 b_k}\Big)^2,$$

where it is only necessary to plug in the definition of θ to see the right term is exactly our divisor. Expanding both the enumerator as well as the divisor by $\frac{1}{n}$, we obtain

$$\mathrm{Var}(\hat{\beta}_0) = \frac{1}{\theta} \frac{\frac{1}{n}\sum_k \frac{1}{\sigma_k^2 b_k^2}}{\frac{1}{n}\sum_k \frac{1}{\sigma_k^2}\left(\frac{1}{b_k} - \sum_j \lambda_j \frac{1}{b_j}\right)^2}$$

Since $\theta = n\mathbb{E}[1/\sigma_B^2]$ for $B \sim \frac{1}{n}\sum_{k=1}^n \delta_{b_k}$ and $\lambda_k = \frac{1}{n\sigma_k^2 \mathbb{E}[1/\sigma_B^2]}$, the above can thus be written as

$$\mathrm{Var}(\hat{\beta}_0) = \frac{1}{n} \frac{1}{\mathbb{E}[1/\sigma_B^2]} \frac{\mathbb{E}[\frac{1}{\sigma_B^2 B^2}]}{\mathbb{E}\Big[\frac{1}{\sigma_B^2}\left(\frac{1}{B} - \mathbb{E}[\frac{1}{B\sigma_B^2 \mathbb{E}[1/\sigma_B^2]}]\right)^2\Big]},$$

which proves our claim. □

5.C Covariance models

In this section we calculate the step sizes of the covariance models listed in Table 5.1 and plotted in Figure 5.2. Additionally we calculate the asymptotic learning rate of A-RFD and prove an Assumption of Corollary 5.3.3 for the squared exponential covariance (Prop. 5.C.3).

5.C.1 Squared exponential

The squared exponential covariance function is given by

$$C\big(\tfrac{\|x-y\|^2}{2}\big) = \sigma^2 \exp\big(-\tfrac{\|x-y\|^2}{2s^2}\big). \tag{5.11}$$

Note that σ^2 will play no role in the step sizes of RFD due to its scale invariance (cf. Advantage 5.1.3).

Theorem 5.C.1. *Let $\mathbf{f} \sim \mathcal{N}(\mu, C)$ where C is the **squared exponential** covariance function (5.11), then we have*

$$\eta^* \frac{\nabla \mathbf{f}(x)}{\|\nabla \mathbf{f}(x)\|} = \arg\min_{\mathbf{d}} \mathbb{E}[\mathbf{f}(x - \mathbf{d}) \mid \mathbf{f}(x), \nabla \mathbf{f}(x)]$$

with RFD step size

$$\eta^* = \frac{s^2\|\nabla \mathbf{f}(x)\|}{\sqrt{\left(\frac{\mu-\mathbf{f}(x)}{2}\right)^2 + s^2\|\nabla \mathbf{f}(x)\|^2} + \frac{\mu-\mathbf{f}(x)}{2}}.$$

Proof. The covariance function C is of the form

$$C(h) = \sigma^2 e^{-\frac{h}{s^2}}.$$

By Equation (5.2)

$$\eta^* = -\arg\min_{\eta} \frac{C(\frac{\eta^2}{2})}{C(0)} - \eta\frac{C'(\frac{\eta^2}{2})}{C'(0)}\Theta.$$

where $\Theta = \frac{\|\nabla \mathbf{f}(x)\|}{\mu-\mathbf{f}(x)}$. We calculate

$$-\frac{C\left(\frac{\eta^2}{2}\right)}{\kappa(0)} - \eta\frac{C'\left(\frac{\eta^2}{2}\right)}{C'(0)}\Theta = -e^{-\frac{\eta^2}{2s^2}}(1+\eta\Theta).$$

This results in the first order condition

$$0 \overset{!}{=} \frac{\eta}{s^2}e^{-\frac{\eta^2}{2s^2}}(1+\eta\Theta) - e^{-\frac{\eta^2}{2s^2}}\Theta = \frac{e^{-\frac{\eta^2}{2s^2}}}{s^2}(\eta^2\Theta + \eta - s^2\Theta).$$

Since the exponential can never be zero, we have to solve a quadratic equation. Its solution results in

$$\eta^*(\Theta) = \sqrt{\left(\tfrac{1}{2\Theta}\right)^2 + s^2} - \tfrac{1}{2\Theta}. \tag{5.12}$$

At this point we could stop, but the result is numerically unstable as it suffers from catastrophic cancellation. To solve this issue we set $x = \frac{1}{2\Theta}$ and reorder

$$\eta^* = \sqrt{x^2+s^2} - x = (\sqrt{x^2+s^2} - x)\frac{\sqrt{x^2+s^2}+x}{\sqrt{x^2+s^2}+x} = \frac{\cancel{x^2} + s^2 - \cancel{x^2}}{\sqrt{x^2+s^2}+x}.$$

Re-substituting $x = \frac{1}{2\Theta} = \frac{\mu-\mathbf{f}(x)}{2\|\nabla \mathbf{f}(x)\|}$, we finally get

$$\eta^* = \frac{s^2}{\sqrt{\left(\frac{\mu-\mathbf{f}(x)}{2\|\nabla \mathbf{f}(x)\|}\right)^2 + s^2} + \frac{\mu-\mathbf{f}(x)}{2\|\nabla \mathbf{f}(x)\|}} = \frac{s^2\|\nabla \mathbf{f}(x)\|}{\sqrt{\left(\frac{\mu-\mathbf{f}(x)}{2}\right)^2 + s^2\|\nabla \mathbf{f}(x)\|^2} + \frac{\mu-\mathbf{f}(x)}{2}}. \qquad \square$$

Proposition 5.C.2 (A-RFD for the Squared Exponential Covariance)**.** *If* $\mathbf{f}$ *is isotropic with squared exponential covariance* (5.11)*, then the step size of A-RFD is given by*

$$\hat{\eta} = \frac{s^2}{\mu - \mathbf{f}(x)}\|\nabla \mathbf{f}(x)\|,$$

Proof. By Definition 5.3.1 of A-RFD and $\Theta = \frac{\|\nabla \mathbf{f}(x)\|}{\mu-\mathbf{f}(x)}$ we have

$$\hat{\eta}(\Theta) = \frac{C(0)}{-C'(0)}\Theta = \frac{\sigma^2\exp(0)}{\frac{\sigma^2}{s^2}\exp(0)}\frac{\|\nabla \mathbf{f}(x)\|}{\mu-\mathbf{f}(x)} = s^2\frac{\|\nabla \mathbf{f}(x)\|}{\mu-\mathbf{f}(x)}. \qquad \square$$

Proposition 5.C.3. *If* $\mathbf{f}$ *is isotropic with squared exponential covariance* (5.11)*, then the RFD step sizes are strictly monotonously increasing in* Θ*.*

Proof. Since we know that $\Theta \to 0$ implies $\eta^* \sim \hat{\eta} \to 0$ strict monotonicity of η^* in Θ is sufficient to show that $\eta^* \to 0$ also implies $\Theta \to 0$. So we take the derivative of (5.12) resulting in

$$\frac{d}{d\Theta}\eta^* = \frac{1 - \frac{1}{\sqrt{1+s^2(2\Theta)^2}}}{2\Theta^2},$$

which is greater zero for all $\Theta > 0$. $\square$

5.C.2 Rational quadratic

The **rational quadratic** covariance function is given by

$$C\big(\tfrac{\|x-y\|}{2}\big) = \sigma^2 \left(1 + \frac{\|x-y\|^2}{\beta s^2}\right)^{-\beta/2} \qquad \beta > 0. \tag{5.13}$$

It can be viewed as a scale mixture of the squared exponential and converges to the squared exponential in the limit $\beta \to \infty$.[34]

[34] Rasmussen and Williams, 2006, p. 87.

Theorem 5.C.4 (Rational Quadratic)**.** *For $\mathbf{f} \sim \mathcal{N}(\mu, C)$ where C is the **rational quadratic covariance** we have for $\Theta = \frac{\|\nabla \mathbf{f}(x)\|}{\mu - \mathbf{f}(x)} \geq 0$ that the RFD step size is given by*

$$\eta^* = s\sqrt{\beta}\operatorname*{Root}_{\eta}\left(-1 + \tfrac{\sqrt{\beta}}{s\Theta}\eta + (1+\beta)\eta^2 + \tfrac{\sqrt{\beta}}{s\Theta}\eta^3\right).$$

The unique root of the polynomial in η can be found either directly with a formula for polynomials of third degree (e.g. using Cardano's method) or by bisection as it is contained in $[0, 1/\sqrt{1+\beta}]$.

Proof. By Theorem 5.2.2 we have

$$\eta^* = \arg\min_{\eta} -\frac{C\big(\frac{\eta^2}{2}\big)}{C(0)} - \eta\frac{C'\big(\frac{\eta^2}{2}\big)}{C'(0)}\Theta$$

for $C(x) = \sigma^2(1 + \frac{2x}{\beta s^2})^{-\beta/2}$. We therefore need to minimize

$$f\big(\tfrac{\eta}{\sqrt{\beta}s}\big) := -\left(1 + \frac{\eta^2}{\beta s^2}\right)^{-\beta/2} - \eta\left(1 + \frac{\eta^2}{\beta s^2}\right)^{-\beta/2-1}\Theta.$$

Substitute in $\tilde{\eta} := \frac{\eta}{\sqrt{\beta}s}$, then the first order condition is

$$0 \overset{!}{=} f'(\tilde{\eta}) = -\frac{d}{d\tilde{\eta}}\left(1+\tilde{\eta}^2\right)^{-\beta/2} + \sqrt{\beta}s\tilde{\eta}\left(1+\tilde{\eta}^2\right)^{-\beta/2-1}\Theta$$

Dividing both sides by $\sqrt{\beta}s\Theta$ we get

$$\begin{aligned} 0 = \frac{f'(\tilde{\eta})}{\sqrt{\beta}s\Theta} &= \tfrac{\beta}{2}(1+\tilde{\eta}^2)^{-\frac{\beta}{2}-1}2\tilde{\eta}\tfrac{1}{\sqrt{\beta}s\Theta} + (1+\tilde{\eta}^2)^{-\frac{\beta}{2}-2}\left[1+\tilde{\eta}^2 - (\tfrac{\beta}{2}+1)2\tilde{\eta}^2\right] \\ &= (1+\tilde{\eta}^2)^{-\frac{\beta}{2}-2}\underbrace{\left[\beta\tilde{\eta}\tfrac{1}{\sqrt{\beta}s\Theta}(1+\tilde{\eta}^2) - [1-\tilde{\eta}^2(1+\beta)]\right]}_{=-1+\frac{\sqrt{\beta}}{s\Theta}\tilde{\eta}+(1+\beta)\tilde{\eta}^2+\frac{\sqrt{\beta}}{s\Theta}\tilde{\eta}^3} \end{aligned}$$

Since $\Theta \geq 0$ and $\beta > 0$ all coefficients of the polynomial are positive except for the shift. The polynomial thus starts out at -1 in zero and only increases from there. Therefore there exists a unique positive critical point which is a minimum.

At the point $\tilde{\eta} = \sqrt{1+\beta}$ the quadratic term is already larger than 1 so the polynomial is positive and we have passed the root. The minimum is therefore contained in the interval $[0, \sqrt{1+\beta}]$.

After finding the minimum in $\tilde{\eta}$ we return to η by multiplication with $\sqrt{\beta}s$. □

Proposition 5.C.5 (A-RFD for the Rational Quadratic Covariance)**.** *If $\mathbf{f}$ is isotropic with rational quadratic covariance (5.13), then the step size of A-RFD is given by*

$$\hat{\eta} = \frac{s^2}{\mu - \mathbf{f}(x)}\|\nabla \mathbf{f}(x)\|.$$

Proof. $C(x) = \sigma^2(1 + \frac{2x}{\beta s^2})^{-\beta/2}$ implies by Definition 5.3.1 of A-RFD and $\Theta = \frac{\|\nabla \mathbf{f}(x)\|}{\mu - \mathbf{f}(x)}$

$$\hat{\eta}(\Theta) = \frac{C_{\mathbf{f}}(0)}{-C'_{\mathbf{f}}(0)}\Theta = \frac{\sigma^2(1+0)^{-\beta/2}}{\frac{\sigma^2}{s^2}(1+0)^{-\beta/2-1}}\frac{\|\nabla \mathbf{f}(x)\|}{\mu - \mathbf{f}(x)} = s^2\frac{\|\nabla \mathbf{f}(x)\|}{\mu - \mathbf{f}(x)}.$$ □

5.C.3 Matérn

Definition 5.C.6. The Matérn model parametrized by $s > 0, \nu \geq 0, \sigma^2 \geq 0$ is given by

$$C\big(\tfrac{\|x-y\|^2}{2}\big) = \sigma^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{\sqrt{2\nu}\|x-y\|}{s}\right)^{\nu} K_\nu \left(\frac{\sqrt{2\nu}\|x-y\|}{s}\right) \tag{5.14}$$

where K_ν is the modified Bessel function.

For $\nu = p + \frac{1}{2}$ with $p \in \mathbb{N}_0$, it can be simplified[35] to

$$C\big(\tfrac{\|x-y\|^2}{2}\big) = \sigma^2 e^{-\frac{\sqrt{2\nu}\|x-y\|}{s}} \frac{p!}{(2p)!} \sum_{k=0}^{p} \frac{(2p-k)!}{(p-k)!k!} \left(\frac{2\sqrt{2\nu}}{s}\|x-y\|\right)^k$$

[35] cf. Rasmussen and Williams, 2006, sec. 4.2.1.

The Matérn model encompasses Rasmussen and Williams (2006)

- the **nugget effect** for $\nu = 0$ (independent randomness)
- the **exponential model** for $\nu = \frac{1}{2}$ (Ornstein-Uhlenbeck process)
- the **squared exponential model** for $\nu \to \infty$ with the same scale s and variance σ^2.

The random functions induced by the Matérn model are a.s. $\lfloor \nu \rfloor$-times differentiable Rasmussen and Williams (2006), i.e. the smoothness of the model increases with increasing ν. While the exponential covariance model with $\nu = \frac{1}{2}$ results in a random function which is not yet differentiable, larger ν result in increasing differentiability. As differentiability starts with $\nu = \frac{3}{2}$ and we have a more explicit formula for $\nu = p + \frac{1}{2}$ the cases $\nu = \frac{3}{2}$ and $\nu = \frac{5}{2}$ are of particular interest.

> "[F]or $\nu \geq 7/2$, in the absence of explicit prior knowledge about the existence of higher order derivatives, it is probably very hard from finite noisy training examples to distinguish between values of $\nu \geq 7/2$ (or even to distinguish between finite values of ν and $\nu \to \infty$, the smooth squared exponential, in this case)".[36]

[36] Rasmussen and Williams, 2006, p. 85.

Theorem 5.C.7. *Assuming $\mathbf{f} \sim \mathcal{N}(\mu, C)$ is a random function where C is the Matérn covariance such that $\nu = p + \frac{1}{2}$ with $p \in \{1, 2\}$. Then the RFD step is given for $\Theta := \frac{\|\nabla \mathbf{f}(x)\|}{\mu - \mathbf{f}(x)} \geq 0$ by*

- $p = 1$

$$\eta^* = \frac{s}{\sqrt{3}} \frac{1}{\left(1 + \frac{\sqrt{3}}{s\Theta}\right)}$$

- $p = 2$

$$\eta^* = \frac{s}{\sqrt{5}} \frac{(1-\zeta) + \sqrt{4 + (1+\zeta)^2}}{2(1+\zeta)} \qquad \zeta := \frac{\sqrt{5}}{3s\Theta}.$$

Proof. We define $\mathcal{C}(\eta) := C(\frac{\eta^2}{2})$, which implies

$$\mathcal{C}'(\eta) = C'\big(\tfrac{\eta^2}{2}\big)\eta$$

or conversely

$$C'\big(\tfrac{\eta^2}{2}\big) = \frac{1}{\eta}\mathcal{C}'(\eta). \tag{5.15}$$

By Theorem 5.2.2, we need to calculate

$$\eta^* = \arg\min_{\eta} -\frac{C(\frac{\eta^2}{2})}{C(0)} - \eta \frac{C'(\frac{\eta^2}{2})}{C'(0)}\Theta. \tag{5.16}$$

Discarding σ w.l.o.g. due to scale invariance (Advantage 5.1.3), we have in the case $p = 1$

$$\mathcal{C}(\eta) = \left(1 + \frac{\sqrt{3}}{s}\eta\right)\exp\left(-\frac{\sqrt{3}}{s}\eta\right).$$

The derivative is then given by

$$\mathcal{C}'(\eta) = -\left(\tfrac{\sqrt{3}}{s}\right)^2 \eta \exp\left(-\tfrac{\sqrt{3}}{s}\eta\right)$$

which implies using (5.15)

$$C'\left(\tfrac{\eta^2}{2}\right) = -\left(\tfrac{\sqrt{3}}{s}\right)^2 \exp\left(-\tfrac{\sqrt{3}}{s}\eta\right) \tag{5.17}$$

We therefore need to minimize (5.16) which is given by

$$\begin{aligned}&\arg\min_{\eta} -\left(1 + \tfrac{\sqrt{3}}{s}\eta\right)\exp\left(-\tfrac{\sqrt{3}}{s}\eta\right) - \eta\exp\left(-\tfrac{\sqrt{3}}{s}\eta\right)\Theta \\ &= \arg\min_{\eta} -\left(1 + (\tfrac{\sqrt{3}}{s} + \Theta)\eta\right)\exp\left(-\tfrac{\sqrt{3}}{s}\eta\right).\end{aligned}$$

The first order condition is

$$0 \overset{!}{=} \left(\tfrac{\sqrt{3}}{s}\left(\cancel{1} + (\tfrac{\sqrt{3}}{s} + \Theta)\eta\right) - (\cancel{\tfrac{\sqrt{3}}{s}} + \Theta)\right)$$

which (divided by Θ and noting that the exponential can never be zero) is equivalent to

$$0 \overset{!}{=} \tfrac{\sqrt{3}}{s}(\tfrac{\sqrt{3}}{s\Theta} + 1)\eta - 1$$

reordering for η implies

$$\eta \overset{!}{=} \frac{s}{\sqrt{3}}\frac{1}{\left(1 + \frac{\sqrt{3}}{s\Theta}\right)}.$$

It is also not difficult to see that this is the point where the derivative switches from negative to positive (i.e. a minimum).

Let us now consider the case $p = 2$, i.e.

$$\mathcal{C}(\eta) = \left(1 + \tfrac{\sqrt{5}}{s}\eta + \tfrac{5}{3s^2}\eta^2\right)\exp\left(-\tfrac{\sqrt{5}}{s}\eta\right),$$

which results in

$$\mathcal{C}'(\eta) = -\tfrac{5}{3s^2}\left(\eta + \tfrac{\sqrt{5}}{s}\eta^2\right)\exp\left(-\tfrac{\sqrt{5}}{s}\eta\right),$$

i.e. by (5.15)

$$C'\left(\tfrac{\eta^2}{2}\right) = -\tfrac{5}{3s^2}\left(1 + \tfrac{\sqrt{5}}{s}\eta\right)\exp\left(-\tfrac{\sqrt{5}}{s}\eta\right). \tag{5.18}$$

We therefore need to minimize (5.16) which is given by

$$\underbrace{\left(-\left(1 + \tfrac{\sqrt{5}}{s}\eta + \tfrac{5}{3s^2}\eta^2\right) - \eta\left(1 + \tfrac{\sqrt{5}}{s}\eta\right)\Theta\right)}_{=-\left(1+\left(\frac{\sqrt{5}}{s}+\Theta\right)\eta+\left(\frac{5}{3s^2}+\frac{\sqrt{5}}{s}\Theta\right)\eta^2\right)}\exp\left(-\tfrac{\sqrt{5}}{s}\eta\right).$$

The first order condition results in

$$\begin{aligned}0 &\overset{!}{=} \tfrac{\sqrt{5}}{s}\left(\cancel{1} + \left(\tfrac{\sqrt{5}}{s} + \Theta\right)\eta + \left(\tfrac{5}{3s^2} + \tfrac{\sqrt{5}}{s}\Theta\right)\eta^2\right) - \left((\cancel{\tfrac{\sqrt{5}}{s}} + \Theta) + 2\left(\tfrac{5}{3s^2} + \tfrac{\sqrt{5}}{s}\Theta\right)\eta\right) \\ &= -\Theta + \left(\tfrac{5}{3s^2} - \tfrac{\sqrt{5}}{s}\Theta\right)\eta + \tfrac{\sqrt{5}}{s}\left(\tfrac{5}{3s^2} + \tfrac{\sqrt{5}}{s}\Theta\right)\eta^2\end{aligned}$$

Dividing everything by Θ and using $\zeta := \frac{\sqrt{5}}{3s\Theta}$ we get

$$0 \overset{!}{=} -1 - (\zeta - 1)\left(\tfrac{\sqrt{5}}{s}\eta\right) + (\zeta + 1)\left(\tfrac{\sqrt{5}}{s}\eta\right)^2$$

Taking a closer look at the sign changes of the derivative it becomes obvious, that the positive root is the minimum, i.e.

$$\frac{\sqrt{5}}{s}\eta \overset{!}{=} \frac{(1-\zeta) + \sqrt{(1-\zeta)^2 + 4(1+\zeta)}}{2(1+\zeta)} = \frac{(1-\zeta) + \sqrt{4 + (1+\zeta)^2}}{2(1+\zeta)}. \qquad \square$$

Proposition 5.C.8 (A-RFD for the Matérn Covariance)**.** *If* $\mathbf{f}$ *is isotropic with Matérn covariance* (5.14) *such that* $\nu = p + \frac{1}{2}$*, then the step size of A-RFD for* $p \in \{1, 2\}$ *is given by*

- $p = 1$
$$\hat{\eta} = \frac{s^2}{3} \frac{\|\nabla \mathbf{f}(x)\|}{\mu - \mathbf{f}(x)}$$
- $p = 2$
$$\hat{\eta} = \frac{3s^2}{5} \frac{\|\nabla \mathbf{f}(x)\|}{\mu - \mathbf{f}(x)}$$

Proof. Noting $\Theta = \frac{\|\nabla \mathbf{f}(x)\|}{\mu - \mathbf{f}(x)}$, we have by Definition 5.3.1 of A-RFD for $p = 1$

$$\hat{\eta} = \frac{C(0)}{-C'(0)}\Theta \overset{(5.17)}{=} \frac{s^2}{3}\Theta,$$

and in the case $p = 2$

$$\hat{\eta} = \frac{C(0)}{-C'(0)}\Theta \overset{(5.18)}{=} \frac{3s^2}{5}\Theta. \qquad \square$$

5.D Proofs

In this section we prove all the claims made in the main body.

5.D.1 Section 5.1: Random function descent

Formal RFD

As we mentioned in a footnote at the definition of RFD, the fact that the parameters become random variables as they are selected by random gradients poses some mathematical challenges which would have been distracting to address in the main body. In following paragraphs leading up to Definition 5.D.1 we introduce and discuss the probability theory required to provide a mathematically sound definition.

For a fixed cost distribution $\mathbb{P}_{\mathbf{f}}$ and any weight vectors x and $\tilde{x}$ the conditional distribution

$$\mathbb{E}[\mathbf{f}(\tilde{x}) \mid \mathbf{f}(x), \nabla \mathbf{f}(x)]$$

is by its axiomatic definition a $(\mathbf{f}(x), \nabla \mathbf{f}(x))$-measurable random variable. By the factorization lemma,[37] there therefore exists a measurable function $(j, g) \mapsto \varphi_{x,\tilde{x}}(j, g)$ such that the following equation holds almost surely

[37] Klenke, 2014, Cor. 1.9.7.

$$\varphi_{x,\tilde{x}}(\mathbf{f}(x), \nabla \mathbf{f}(x)) = \mathbb{E}[\mathbf{f}(\tilde{x}) \mid \mathbf{f}(x), \nabla \mathbf{f}(x)]. \tag{5.19}$$

Since it is possible to calculate $\varphi_{x,\tilde{x}}$ explicitly in the Gaussian case (cf. 2.A.1), the function

$$\Phi_{\mathbb{P}_{\mathbf{f}}} : \begin{cases} (\mathbb{R}^d \times \mathbb{R} \times \mathbb{R}^d) \to \mathbb{R} \\ (x, j, g) \mapsto \arg\min_{\tilde{x}} \varphi_{x,\tilde{x}}(j, g), \end{cases}$$

which implements some tie-breaker rules for set valued $\arg\min$ is measurable when $\mathbf{f}$ is Gaussian and its covariance function is sufficiently smooth. To prove measurability in the general case is a difficult problem of its own, which we do not attempt to solve here, since we would not utilize the conditional expectation outside of the Gaussian case anyway (cf. Section 5.E.3). For deterministic x, we therefore have

$$\begin{aligned} \Phi_{\mathbb{P}_{\mathbf{f}}}(x, \mathbf{f}(x), \nabla \mathbf{f}(x)) &= \arg\min_{\tilde{x}} \varphi_{x,\tilde{x}}(\mathbf{f}(x), \nabla \mathbf{f}(x)) \\ &\overset{(5.19)}{=} \arg\min_{\tilde{x}} \mathbb{E}[\mathbf{f}(\tilde{x}) \mid \mathbf{f}(x), \nabla \mathbf{f}(x)]. \end{aligned}$$

So if the parameter vectors x_n were deterministic, our formal definition of RFD and our initial definition would coincide. But for random weights X (5.19) stops to hold in general, i.e.

$$\varphi_{X,\tilde{x}}(\mathbf{f}(X), \nabla \mathbf{f}(X)) \neq \mathbb{E}[\mathbf{f}(\tilde{x}) \mid \mathbf{f}(X), \nabla \mathbf{f}(X)].$$

If this equation does not need to hold, we similarly have in general

$$\Phi_{\mathbb{P}_{\mathbf{f}}}(X, \mathbf{f}(X), \nabla\mathbf{f}(X)) \neq \arg\min_{\tilde{x}} \mathbb{E}[\mathbf{f}(\tilde{x}) \mid \mathbf{f}(X), \nabla\mathbf{f}(X)].$$

So the following definition is not just a restatement of the original definition of RFD.

Definition 5.D.1 (Formal RFD)**.** For a Gaussian random cost function $\mathbf{f}$, we define the RFD algorithm with starting point $X_0 = x_0 \in \mathbb{R}^d$ by

$$X_{n+1} := \Phi_{\mathbb{P}_{\mathbf{f}}}(X_n, \mathbf{f}(X_n), \nabla\mathbf{f}(X_n))$$

This is what we effectively do in Theorem 5.2.2 under the additional isotropy assumption, where we calculate the arg min under the assumption that x is deterministic (i.e. we determine $\Phi_{\mathbb{P}_{\mathbf{f}}}$), before we plug-in the random variables X_n to obtain X_{n+1}. Similarly this is how the step size prescriptions of RFD actually work. We first assume deterministic weights and later plug the random variables into our formulas. For this reason, we avoided large letters indicating random variables for parameters x in the main body.

Scale invariance

Advantage 5.1.3 (Scale invariance and equivariance)**.** *RFD is invariant to additive shifts and positive scaling of the cost* $\mathbf{f}$*. RFD is also equivariant*[38] *with respect to transformations of the parameter input of* $\mathbf{f}$ *by differentiable bijections whose Jacobian is invertible everywhere (e.g. invertible linear maps).*

[38] a function ϕ is *invariant* with respect to a group G, if $\phi(\rho(g)x) = \phi(x)$, where $\rho(g)$ is a representation of the element $g \in G$. ϕ is *equivariant* with respect to the group G, if $\phi(\rho(g)x) = \rho(g)\phi(x)$ for all $g \in G$.

Before we get to the proof, let us quickly formulate the statement in mathematical terms. Let x_n be the parameters selected optimizing $\mathbf{f}$ starting in x_0 and $\tilde{x}_n$ the parameters selected by the same optimizer optimizing $\tilde{\mathbf{f}}$ starting in $\tilde{x}_0$.

If we apply affine linear scaling to cost $\mathbf{f}$ such that $\tilde{\mathbf{f}}(x) = a\mathbf{f}(x) + b$ and start optimization in the same point, i.e. $x_0 = \tilde{x}_0$, then we expect a scale invariant optimizer to select

$$x_n = \tilde{x}_n.$$

If we scale inputs on the other hand (or more generally map them with a bijection ϕ), then we expect for $\tilde{\mathbf{f}} := \mathbf{f} \circ \phi$ and starting point $\tilde{x}_0 = \phi^{-1}(x_0)$, that this relationship is retained by an equivariant optimizer, i.e.

$$\tilde{x}_n = \phi^{-1}(x_n).$$

Why do we use a different starting point? As an illustrating example, assume that ϕ maps miles into kilometers. Then $\tilde{\mathbf{f}}$ accepts miles, while $\mathbf{f}$ accepts kilometers. Then we have to map the initial starting point x_0 of $\mathbf{f}$ measured in kilometers into miles $\tilde{x}_0$. ϕ^{-1} is precisely this transformation from kilometers into miles. An equivariant optimizer should retain this relation, i.e. no matter if the input is measured in miles or kilometers the same points are selected.

Proof. The following proof will be split into three parts. The first two parts of the proof will address a more general audience and ignore the mathematical subtleties we discussed in Section 5.D.1. In the third part we explain to the interested probabilists how to resolve these issues.

1. **Invariance with regard to affine linear scaling**

Let $\tilde{\mathbf{f}}(x) := a\mathbf{f}(x) + b$ where $a > 0$ and $b \in \mathbb{R}$ and assume $\tilde{x}_0 = x_0$. With the induction start given, we only require the induction step to prove $\tilde{x}_n = x_n$.

For the induction step, we assume this equation holds up to n. Since $\phi(x) = ax + b$ is a measurable bijection, the sigma algebra[39] generated by

$$(\tilde{\mathbf{f}}(x_n), \nabla\tilde{\mathbf{f}}(x_n)) = (\phi \circ \mathbf{f}(x_n), a\nabla\mathbf{f}(x_n))$$

[39] if you are unfamiliar with sigma algebras read them as "information".

is therefore equal to the sigma algebra generated by $(\mathbf{f}(x_n), \nabla\mathbf{f}(x_n))$. This implies

$$\begin{aligned}
\tilde{x}_{n+1} &= \arg\min_x \mathbb{E}[\tilde{\mathbf{f}}(x) \mid \tilde{\mathbf{f}}(\tilde{x}_n), \nabla\tilde{\mathbf{f}}(\tilde{x}_n)] \\
&\overset{\text{induction}}{=} \arg\min_x \mathbb{E}[\tilde{\mathbf{f}}(x) \mid \tilde{\mathbf{f}}(x_n), \nabla\tilde{\mathbf{f}}(x_n)] \\
&\overset{\text{sigma alg.}}{=} \arg\min_x \mathbb{E}[\tilde{\mathbf{f}}(x) \mid \mathbf{f}(x_n), \nabla\mathbf{f}(x_n)] \\
&\overset{\text{linearity}}{=} \arg\min_x a\mathbb{E}[\mathbf{f}(x) \mid \mathbf{f}(x_n), \nabla\mathbf{f}(x_n)] + b \\
&\overset{\text{monotonicity}}{=} \arg\min_x \mathbb{E}[\mathbf{f}(x) \mid \mathbf{f}(x_n), \nabla\mathbf{f}(x_n)] \\
&\overset{\text{def.}}{=} x_{n+1}
\end{aligned} \tag{5.20}$$

Where we have used the linearity of the conditional expectation and the strict monotonicity of $\phi(x) = ax + b$.

2. **Equivariance with regard to certain input bijections**

Let ϕ be a differentiable bijection whose jacobian is invertible everywhere and assume $\tilde{\mathbf{f}} := \mathbf{f} \circ \phi$. Since ϕ is a bijection, $\phi(M)$ is the domain of $\mathbf{f}$ whenever M is the domain of $\tilde{\mathbf{f}}$.

For a starting point $x_0 \in \phi(M)$ we now assume $\tilde{x}_0 = \phi^{-1}(x_0) \in M$ and are again going to prove the claim

$$\tilde{x}_n = \phi^{-1}(x_n).$$

by induction. Assume that we have this claim up to n. Then we have by induction

$$\tilde{\mathbf{f}}(\tilde{x}_n) = \mathbf{f} \circ \phi(\phi^{-1}(x_n)) = \mathbf{f}(x_n) \tag{5.21}$$

and

$$\nabla\tilde{\mathbf{f}}(\tilde{x}_n) = \nabla_{\tilde{x}_n}(\mathbf{f} \circ \phi(\tilde{x}_n)) = \phi'(\tilde{x}_n)(\nabla\mathbf{f})(\phi(\tilde{x}_n)) = \phi'(\tilde{x}_n)\nabla\mathbf{f}(x_n).$$

Since $\phi'(\tilde{x}_n)$ is invertible by assumption, the σ-algebras generated by the vectors $(\tilde{\mathbf{f}}(\tilde{x}_n), \nabla\tilde{\mathbf{f}}(\tilde{x}_n))$ and $\mathbf{f}(x_n), \nabla\mathbf{f}(x_n)$ are identical. But this results in the induction step

$$\begin{aligned}
\tilde{x}_{n+1} &= \arg\min_{x \in M} \mathbb{E}[\tilde{\mathbf{f}}(x) \mid \tilde{\mathbf{f}}(\tilde{x}_n), \nabla\tilde{\mathbf{f}}(\tilde{x}_n)] \\
&\overset{\text{sigma alg.}}{=} \arg\min_{x \in M} \mathbb{E}[\tilde{\mathbf{f}}(x) \mid \mathbf{f}(x_n), \nabla\mathbf{f}(x_n)] \\
&\overset{\text{def.}}{=} \arg\min_{x \in M} \mathbb{E}[\mathbf{f} \circ \phi(x) \mid \mathbf{f}(x_n), \nabla\mathbf{f}(x_n)] \\
&= \phi^{-1}\Big(\underbrace{\arg\min_{\theta \in \phi(M)} \mathbb{E}[\mathbf{f}(\theta) \mid \mathbf{f}(x_n), \nabla\mathbf{f}(x_n)]}_{\overset{\text{def.}}{=} x_{n+1}}\Big).
\end{aligned} \tag{5.22}$$

where we simply optimize over $\theta = \phi(x)$ instead of x and correct the arg min at the end.

3. **Addressing the subtleties**

In equation (5.20) we have really proven for deterministic x

$$\Phi_{\mathbb{P}_{\tilde{\mathbf{f}}}}(x, \tilde{\mathbf{f}}(x), \nabla\tilde{\mathbf{f}}(x)) = \Phi_{\mathbb{P}_{\mathbf{f}}}(x, \mathbf{f}(x), \nabla\mathbf{f}(x)).$$

But this implies with the induction assumption $X_n = \tilde{X}_n$

$$\tilde{X}_{n+1} = \Phi_{\mathbb{P}_{\tilde{\mathbf{f}}}}(\tilde{X}_n, \tilde{\mathbf{f}}(\tilde{X}_n), \nabla\tilde{\mathbf{f}}(\tilde{X}_n)) \overset{\text{ind.}}{=} \Phi_{\mathbb{P}_{\mathbf{f}}}(X_n, \mathbf{f}(X_n), \nabla\mathbf{f}(X_n)) = X_{n+1}.$$

Similarly we have proven in (5.22) that

$$\Phi_{\mathbb{P}_{\tilde{\mathbf{f}}}}\big(\phi^{-1}(x), \tilde{\mathbf{f}}(\phi^{-1}(x)), \nabla\tilde{\mathbf{f}}(\phi^{-1}(x))\big) = \phi^{-1}\big(\Phi_{\mathbb{P}_{\mathbf{f}}}(x, \mathbf{f}(x), \nabla\mathbf{f}(x))\big).$$

By the induction assumption $\tilde{X} = \phi^{-1}(X_n)$, this implies

$$\begin{aligned}\tilde{X}_{n+1} &= \Phi_{\mathbb{P}_{\tilde{\mathbf{f}}}}(\tilde{X}_n, \tilde{\mathbf{f}}(\tilde{X}_n), \nabla\tilde{\mathbf{f}}(\tilde{X}_n)) \\ &\overset{\text{ind.}}{=} \Phi_{\mathbb{P}_{\tilde{\mathbf{f}}}}\big(\phi^{-1}(X_n), \tilde{\mathbf{f}}(\phi^{-1}(X_n)), \nabla\tilde{\mathbf{f}}(\phi^{-1}(X_n))\big) \\ &= \phi^{-1}\big(\Phi_{\mathbb{P}_{\mathbf{f}}}(X_n, \mathbf{f}(X_n), \nabla\mathbf{f}(X_n))\big) \\ &= \phi^{-1}(X_{n+1}).\end{aligned}$$ □

5.D.2 Section 5.2: Relation to gradient descent

Lemma 5.2.1 (Explicit first order stochastic Taylor approximation)**.** *For $\mathbf{f} \sim \mathcal{N}(\mu, C)$, the first order stochastic Taylor approximation is given by*

$$\mathbb{E}[\mathbf{f}(x-\mathbf{d}) \mid \mathbf{f}(x), \nabla\mathbf{f}(x)] = \mu + \frac{C(\frac{\|\mathbf{d}\|^2}{2})}{C(0)}(\mathbf{f}(x)-\mu) - \frac{C'(\frac{\|\mathbf{d}\|^2}{2})}{C'(0)}\langle\mathbf{d}, \nabla\mathbf{f}(x)\rangle.$$

Proof. $(\mathbf{f}(x), \nabla\mathbf{f}(x), \mathbf{f}(x-\mathbf{d}))$ is a Gaussian vector for which the conditional distribution is well known. It is only necessary to calculate the covariance matrix. The key ingredient here is to observe that $\mathbf{f}(x), \partial_1\mathbf{f}(x), \dots, \partial_d\mathbf{f}(x)$ are all independent, trivializing matrix inversion.

More formally, by Lemma 4.5.1 we have

$$\operatorname{Cov}\left(\begin{pmatrix}\mathbf{f}(x)\\ \nabla\mathbf{f}(x)\end{pmatrix}\right) = \begin{pmatrix}C(0) & \\ & -C'(0)\mathbb{I}_{d\times d}\end{pmatrix}$$

and

$$\operatorname{Cov}\left(\mathbf{f}(x-\mathbf{d}), \begin{pmatrix}\mathbf{f}(x)\\ \nabla\mathbf{f}(x)\end{pmatrix}\right) = \begin{pmatrix}C(\frac{\|\mathbf{d}\|^2}{2})\\ C'(\frac{\|\mathbf{d}\|^2}{2})\mathbf{d}\end{pmatrix}.$$

By Theorem 2.A.1 we therefore know that

$$\begin{aligned}&\mathbb{E}[\mathbf{f}(x-\mathbf{d}) \mid \mathbf{f}(x), \nabla\mathbf{f}(x)] \\ &= \mu + \begin{pmatrix}C(\frac{\|\mathbf{d}\|^2}{2})\\ C'(\frac{\|\mathbf{d}\|^2}{2})\mathbf{d}\end{pmatrix}^T \begin{pmatrix}C(0) & \\ & -C'(0)\mathbb{I}_{d\times d}\end{pmatrix}^{-1} \begin{pmatrix}\mathbf{f}(x)-\mu\\ \nabla\mathbf{f}(x)\end{pmatrix},\end{aligned}$$

which immediately yields the claim. □

Theorem 5.2.2 (Explicit RFD)**.** *Let $\mathbf{f} \sim \mathcal{N}(\mu, C)$, then RFD coincides with gradient descent*

$$x_{n+1} = x_n - \eta_n^* \tfrac{\nabla\mathbf{f}(x_n)}{\|\nabla\mathbf{f}(x_n)\|},$$

where the RFD step sizes are given by

$$\eta_n^* := \operatorname*{arg\,min}_{\eta\in\mathbb{R}} \frac{C(\frac{\eta^2}{2})}{C(0)}(\mathbf{f}(x_n)-\mu) - \eta\frac{C'(\frac{\eta^2}{2})}{C'(0)}\|\nabla\mathbf{f}(x_n)\|. \tag{5.1}$$

Proof. The explicit version of RFD follows essentially by fixing the step size $\eta = \|\mathbf{d}\|$ and optimizing over the direction first. With Lemma 5.2.1 we have

$$\begin{aligned}&\min_{\mathbf{d}} \mathbb{E}[\mathbf{f}(x-\mathbf{d}) \mid \mathbf{f}(x), \nabla\mathbf{f}(x)] \\ &= \min_{\eta\geq 0}\min_{\mathbf{d}:\|\mathbf{d}\|=\eta} \mu + \frac{C(\frac{\eta^2}{2})}{C(0)}(\mathbf{f}(x)-\mu) - \frac{C'(\frac{\eta^2}{2})}{C'(0)}\langle\mathbf{d}, \nabla\mathbf{f}(x)\rangle \\ &= \min_{\eta\geq 0} \mu + \frac{C(\frac{\eta^2}{2})}{C(0)}(\mathbf{f}(x)-\mu) - \frac{C'(\frac{\eta^2}{2})}{C'(0)} \begin{cases}\max\limits_{\mathbf{d}:\|\mathbf{d}\|=\eta}\langle\mathbf{d}, \nabla\mathbf{f}(x)\rangle & \frac{C'(\frac{\eta^2}{2})}{C'(0)} \geq 0\\ \min\limits_{\mathbf{d}:\|\mathbf{d}\|=\eta}\langle\mathbf{d}, \nabla\mathbf{f}(x)\rangle & \frac{C'(\frac{\eta^2}{2})}{C'(0)} < 0.\end{cases}\end{aligned}$$

By Lemma 1.A.2 and Corollary 1.A.3 the maximizing or minimizing step direction is then given by

$$\mathbf{d}(\eta) = \pm\eta\frac{\nabla\mathbf{f}(x)}{\|\nabla\mathbf{f}(x)\|}.$$

Where it is typically to be expected, that we have a positive sign. Since that depends on the covariance though, we avoid this problem with the following argument: Since η only appears as η^2 in the remaining equation, we can optimize over $\eta \in \mathbb{R}$ in the outer minimization instead of over $\eta \geq 0$ to move the sign into the step size η and set without loss of generality

$$\mathbf{d}(\eta) = \eta \frac{\nabla \mathbf{f}(x)}{\|\nabla \mathbf{f}(x)\|}.$$

Since $\langle \mathbf{d}(\eta), \nabla \mathbf{f}(x) \rangle = \eta \|\nabla \mathbf{f}(x)\|$ the remaining outer minimization problem over the step size is then given by

$$\min_{\eta \in \mathbb{R}} \frac{C\left(\frac{\eta^2}{2}\right)}{C(0)} (\mathbf{f}(x) - \mu) - \eta \frac{C'\left(\frac{\eta^2}{2}\right)}{C'(0)} \|\nabla \mathbf{f}(x)\|,$$

Its minimizer is by definition the RFD step size as given in the Theorem. □

5.D.3 Section 5.3: RFD-step sizes

Proposition 5.D.2 (Tayloring the step size optimization problem)**.** *The second order Taylor approximation of the step size optimization problem*

$$q_\Theta(\eta) = -\frac{C\left(\frac{\eta^2}{2}\right)}{C(0)} - \eta \frac{C'\left(\frac{\eta^2}{2}\right)}{C'(0)} \Theta$$

around zero is given by

$$T_2 q_\Theta(\eta) = -1 - \eta\Theta + \eta^2 \frac{-C'(0)}{2C(0)},$$

and minimized by

$$\hat{\eta} := \arg\min_\eta T_2 q_\Theta(\eta) = \tfrac{C(0)}{-C'(0)} \Theta.$$

Furthermore, the Taylor residual is bounded by

$$\left|q(\eta) - T_2 q(\eta)\right| \leq \eta^3 c_0 \left(\tfrac{\eta}{4} + \Theta\right)$$

with $c_0 = \frac{1}{2} \max\{\sup_{\theta \in [0,1]} |C''(\theta)|, |C'(0)|\} (\frac{1}{C(0)} + \frac{1}{|C'(0)|}) < \infty$.

Proof. Using the Taylor approximation with the mean value reminder for C, we get

$$\begin{aligned} C\left(\tfrac{\eta^2}{2}\right) &= C(0) + C'(0)\tfrac{\eta^2}{2} + C''(\theta_2) \frac{\left(\frac{\eta^2}{2}\right)^2}{2!} \\ C'\left(\tfrac{\eta^2}{2}\right) &= C'(0) + C''(\theta_1)\tfrac{\eta^2}{2} \end{aligned}$$

for some $\theta_1, \theta_2 \in [0, \frac{\eta^2}{2}]$. This implies

$$q(\eta) - \underbrace{\left(-\left(1 + \tfrac{C'(0)}{C(0)} \tfrac{\eta^2}{2}\right) - \eta\Theta\right)}_{=: T_2 q_\Theta(\eta)} = -\frac{C''(\theta_2)}{C(0)} \frac{\eta^4}{2^3} - \frac{C''(\theta_1)}{C'(0)} \frac{\eta^3}{2} \Theta$$

By the following error the optimistically defined $T_2 q_\Theta(\eta)$ is really the second Taylor approximation (which can be confirmed manually, but we deduce it by arguing that its residual is in $\mathcal{O}(\eta^3)$). More specifically,

$$\begin{aligned} \left|q(\eta) - T_2 q(\eta)\right| &\leq \eta^3 \left(\frac{\sup_{\theta \in [0, \frac{\eta^2}{2}]} |C''(\theta)|}{2C(0)} \frac{\eta}{4} + \frac{\sup_{\theta \in [0, \frac{\eta^2}{2}]} |C''(\theta)|}{2|C'(0)|} \Theta \right) \\ &\overset{\text{Lem. 5.D.8}}{\leq} \eta^3 c_0 \left(\tfrac{\eta}{4} + \Theta\right) \end{aligned}$$

It is easy to see for $\mathbf{f}(x) < \mu$ that $T_2 q(\eta)$ is a convex parabola due to $C'(0) < 0$. We thus have

$$\hat{\eta} := \arg\min_\eta T_2 q_\Theta(\eta) = \tfrac{C(0)}{-C'(0)} \Theta.$$ □

Theorem 5.D.3 (Details of Proposition 5.3.2)**.** *Let* $\mathbf{f} \sim \mathcal{N}(\mu, C)$ *and assume there exists* $\eta_0 > 0$ *such that the correlation for larger distances* $\eta \geq \eta_0$ *are bounded smaller than* 1*, i.e.* $\frac{C(\eta^2/2)}{C(0)} < \rho \in (0,1)$*. Then there exists* $K, \Theta_0 > 0$ *such that for all* $\Theta < \Theta_0$

$$1 - K\Theta \leq \frac{\eta^*(\Theta)}{\hat{\eta}(\Theta)} \leq 1 + K\Theta.$$

In particular we have $\eta^*(\Theta) \sim \hat{\eta}(\Theta)$ *as* $\Theta \to 0$ *or equivalently as* $\hat{\eta} \to 0$*.*

Proof. This follows immediately from Lemma 5.D.4, 5.D.5 and 5.D.6. □

Corollary 5.3.3. *Assume* $\eta^* \to 0$ *implies* $\Theta \to 0$*, the cost* $\mathbf{f}$ *is bounded, has continuous gradients and RFD converges to some point* x_∞*. Then* x_∞ *is a critical point and the RFD step sizes* η^* *are asymptotically equal to* $\hat{\eta}$*.*

Proof. Assuming RFD converges, its step sizes η^* converge to zero. But this implies $\Theta \to 0$ by assumption, i.e.

$$\Theta = \frac{\|\nabla \mathbf{f}(x)\|}{\mu - \mathbf{f}(x)} \to 0$$

Since $\mathbf{f}(x)$ is bounded, this implies $\|\nabla \mathbf{f}(x)\| \to 0$ and by continuity the of the gradient, it is zero in its limit. Thus we converge to a stationary point. The asymptotic equality follows by Lemma 5.D.4 and 5.D.5, as we know η^* converges so we do not require the assumptions of Lemma 5.D.6. □

Locating the Minimizer

In the following we want to rule out locations for the RFD step size η^* by proving $q_\Theta(\eta) > q_\Theta(\hat{\eta})$ for a wide range of η. For this endeavour the relative position of the step size η relative to $\hat{\eta}$ is a useful re-parametrization

$$\eta := \eta(\lambda) = \lambda\hat{\eta}.$$

Due to $\hat{\eta} = \frac{C(0)}{-C'(0)}\Theta$ we obtain

$$T_2 q_\Theta(\eta) = -1 - \eta\Theta + \tfrac{\eta^2}{2}\tfrac{-C'(0)}{C(0)} = -1 + \lambda(\tfrac{\lambda}{2} - 1)\hat{\eta}\Theta$$

On the other hand we have for the bound

$$|q_\Theta(\eta) - T_2 q_\Theta(\eta)| \leq \lambda^3 \hat{\eta}^3 c_0 \Big(\lambda \tfrac{C(0)}{4|C'(0)|} + 1\Big)\Theta$$

Since $\hat{\eta} = \eta(1)$ we thus obtain

$$\begin{aligned}
\frac{q_\Theta(\eta) - q_\Theta(\hat{\eta})}{\hat{\eta}\Theta} &\geq \frac{T_2 q_\Theta(\eta) - |q_\Theta(\eta) - T_2 q_\Theta(\eta)| - T_2 q_\Theta(\hat{\eta}) - |q_\Theta(\hat{\eta}) - T_2 q_\Theta(\hat{\eta})|}{\hat{\eta}\Theta} \\
&\geq \underbrace{\big(\lambda(\tfrac{\lambda}{2} - 1) - (-\tfrac{1}{2})\big)}_{=\frac{1}{2} - \lambda + \frac{\lambda^2}{2}} - \hat{\eta}^2 c_0 \Big[\lambda^3 \Big(\lambda \tfrac{C(0)}{4|C'(0)|} + 1\Big) + \Big(\tfrac{C(0)}{4|C'(0)|} + 1\Big)\Big] \\
&= \tfrac{1}{2}(1 - \lambda)^2 - \hat{\eta}^2 c_0 \Big[\lambda^3 \Big(\lambda \tfrac{C(0)}{4|C'(0)|} + 1\Big) + \Big(\tfrac{C(0)}{4|C'(0)|} + 1\Big)\Big]. \qquad (5.23)
\end{aligned}$$

This equation will be the basis of a number of lemmas ruling out various step sizes as minimizers.

Lemma 5.D.4 (Ruling out small step sizes)**.** *If the step size is (much) smaller than the asymptotic step size* $\hat{\eta} = \hat{\eta}(\Theta)$*, then it can not be a minimizer. More specifically*

$$\frac{\eta}{\hat{\eta}} \in [0, 1 - c_1\Theta) \implies q_\Theta(\eta) > q_\Theta(\hat{\eta})$$

where $c_1 := 2\frac{C(0)}{|C'(0)|}\sqrt{c_0\big(\frac{C(0)}{4|C'(0)|} + 1\big)} < \infty$*.*

Proof. Here we consider the case $\eta \leq \hat{\eta}$, i.e. $\lambda \in [0,1]$. By (5.23) we have

$$\begin{aligned}\frac{q_\Theta(\eta) - q_\Theta(\hat{\eta})}{\hat{\eta}\Theta} &\geq \tfrac{1}{2}(1-\lambda)^2 - \hat{\eta}^2 c_0 \Big[\lambda^3\Big(\lambda\tfrac{C(0)}{4|C'(0)|}+1\Big) + \Big(\tfrac{C(0)}{4|C'(0)|}+1\Big)\Big] \\ &\geq \tfrac{1}{2}(1-\lambda)^2 - 2\hat{\eta}^2 c_0\Big(\tfrac{C(0)}{4|C'(0)|}+1\Big) \\ &\overset{!}{>} 0\end{aligned}$$

for which

$$(1-\lambda)^2 > 4\hat{\eta}^2 c_0\Big(\tfrac{C(0)}{4|C'(0)|}+1\Big)$$

is sufficient or equivalently

$$\lambda < 1 - 2\hat{\eta}\sqrt{c_0\Big(\tfrac{C(0)}{4|C'(0)|}+1\Big)} = 1 - \Theta \underbrace{2\tfrac{C(0)}{|C'(0)|}\sqrt{c_0\Big(\tfrac{C(0)}{4|C'(0)|}+1\Big)}}_{=:c_1}$$

So for $\lambda \in [0, 1-\Theta c_1)$ we have $q_\Theta(\eta) > q_\Theta(\hat{\eta})$. □

Lemma 5.D.5 (Ruling out medium sized step sizes as minimizer)**.** *For $c_2 = 2c_1$ and $\Theta \leq \Theta_0 := \frac{1}{5c_1}$, we have*

$$\frac{\eta}{\hat{\eta}} \in \big(1 + c_2\Theta, \tfrac{1}{c_2\Theta}\big) \implies q_\Theta(\eta) > q_\Theta(\hat{\eta})$$

Proof. Here we consider the case $\lambda \geq 1$, i.e. $\eta > \hat{\eta}$. Again starting with (5.23) we get

$$\begin{aligned}\frac{q(\eta) - q(\hat{\eta})}{\hat{\eta}\Theta} &\geq \tfrac{1}{2}(1-\lambda)^2 - \hat{\eta}^2 c_0 \Big[\lambda^3\Big(\lambda\tfrac{C(0)}{4|C'(0)|}+1\Big) + \Big(\tfrac{C(0)}{4|C'(0)|}+1\Big)\Big] \\ &\geq \tfrac{1}{2}(\lambda-1)^2 - 2\lambda^4\hat{\eta}^2 c_0\Big(\tfrac{C(0)}{4|C'(0)|}+1\Big) \\ &\overset{!}{>} 0,\end{aligned}$$

for which

$$\lambda - 1 > 2\lambda^2\hat{\eta}\sqrt{c_0\Big(\tfrac{C(0)}{4|C'(0)|}+1\Big)} = c_1\Theta\lambda^2$$

or equivalently

$$\lambda - 1 - c_1\Theta\lambda^2 > 0$$

is sufficient. Note that this is a concave parabola in λ. So it is positive between its zeros which are characterized by

$$c_1\Theta\lambda^2 - \lambda + 1 = 0.$$

They are thus given by

$$\lambda_{1/2} = \frac{1 \pm \sqrt{1 - 4c_1\Theta}}{2c_1\Theta}.$$

So whenever $\lambda \in (\lambda_1, \lambda_2)$ we have that $q_\Theta(\eta) > q_\Theta(\hat{\eta})$. In particular for $4c_1\Theta \leq 1$ or equivalently $\Theta \leq \frac{1}{4c_1}$ we have

$$\lambda_2 \geq \frac{1}{2c_1\Theta} = \frac{1}{c_2\Theta}$$

To get a bound on λ_1 note that the original equation was essentially

$$\lambda \geq 1 + c_1\Theta\lambda^2$$

with equality for $\lambda = \lambda_1$, if Θ is reduced, the inequality remains, which implies that λ_1 is decreasing with Θ. So assuming the inequality is satisfied for a particular λ e.g. $\lambda = \sqrt{2}$ which requires

$$\sqrt{2} \geq 1 + 2c_1\Theta \iff \Theta \leq \tfrac{\sqrt{2}-1}{2c_1},$$

then we know that $\lambda_1 \leq \sqrt{2}$ for all smaller Θ. This implies for $\Theta \leq \Theta_0 = \frac{1}{5c_1} \leq \frac{\sqrt{2}-1}{2c_1}$

$$\lambda_1 = 1 + c_1\Theta\lambda_1^2 \leq 1 + \underbrace{2c_1}_{c_2}\Theta. \qquad \square$$

Lemma 5.D.6 (Ruling out large step sizes as minimizer)**.** *If there exists step size $\eta_0 > 0$ such that the correlation is bounded by some $\rho < 1$, i.e.*

$$\frac{C\left(\frac{\eta^2}{2}\right)}{C(0)} \leq \rho \in (0,1),$$

for larger step sizes $\eta \geq \eta_0$, then there exist $\Theta_0 > 0$ such that for all $\Theta < \Theta_0$

$$\frac{\eta}{\hat{\eta}} \in \left(1 + c_2\Theta, \infty\right) \implies q(\eta) > q(\hat{\eta}),$$

where c_2 is the constant from Lemma 5.D.5.

Proof. The upper bound $\frac{1}{c_2\Theta}$ in Lemma 5.D.5 is only due to the loss of precision of the Taylor approximation. To remove it, we take a closer look at the actual q_Θ itself. We have the following bound for our asymptotic minimum

$$\begin{aligned}\frac{q_\Theta(\hat{\eta})}{\Theta} &\leq \frac{T_2 q_\Theta(\hat{\eta}) + |q_\Theta(\hat{\eta}) - T_2 q_\Theta(\hat{\eta})|}{\Theta} = -\frac{1}{\Theta} - \frac{1}{2}\hat{\eta} + \hat{\eta}^3 \underbrace{c_0\Big(\tfrac{C(0)}{4|C'(0)|} + 1\Big)}_{=:c_3} \\ &\leq -\frac{1}{\Theta} + \hat{\eta}^3 c_3\end{aligned}$$

Which means we have for

$$\begin{aligned}\frac{q_\Theta(\eta) - q_\Theta(\hat{\eta})}{\Theta} &\geq \Big(1 - \frac{C\left(\frac{\eta^2}{2}\right)}{C(0)}\Big)\frac{1}{\Theta} - \eta\frac{C'\left(\frac{\eta^2}{2}\right)}{C'(0)} - \hat{\eta}^3 c_3 \\ &\overset{\text{Lemma 5.D.7}}{\geq} \Big(1 - \frac{C\left(\frac{\eta^2}{2}\right)}{C(0)}\Big)\frac{1}{\Theta} - \frac{\sqrt{C(0)}}{\sqrt{-C'(0)}} - \hat{\eta}^3 c_3 \\ &\geq \big(1 - M\big)\frac{1}{\Theta} - \frac{\sqrt{C(0)}}{\sqrt{-C'(0)}} - \hat{\eta}^3 c_3 \\ &\overset{!}{>} 0,\end{aligned}$$

where we use the assumption that there exists $\rho \in (0,1)$ such that $\rho \geq \frac{C\left(\frac{\eta^2}{2}\right)}{C(0)}$ for all $\eta \geq \eta_0$ and the fact that we only need to consider $\eta \geq \frac{1}{c_2\Theta}$ (due to Lemma 5.D.5) which allows a translation of η_0 into some maximal Θ_0. Note that $\hat{\eta} \sim \Theta$ vanishes as $\Theta \to 0$, so eventually the term $(1-M)\frac{1}{\Theta}$ dominates. Selecting Θ_0 small small enough is thus sufficient to cover everything that is not already covered by Lemma 5.D.5. $\square$

Technical bounds

Lemma 5.D.7 (Bound on the first derivative of the covariance)**.**

$$\sup_{\eta \geq 0} |C'\left(\tfrac{\eta^2}{2}\right)\eta| \leq \sqrt{-C'(0)C(0)}$$

Proof. Since we have

$$\operatorname{Cov}(D_v\mathbf{f}(x), \mathbf{f}(y)) = C'\left(\tfrac{\|x-y\|^2}{2}\right)\langle x - y, v\rangle$$

we have for a standardized vector $\|v\| = 1$ and $x - y = \eta v$ by Cauchy-Schwarz

$$|C'\left(\tfrac{\eta^2}{2}\right)\eta| = |\operatorname{Cov}(D_v\mathbf{f}(x), \mathbf{f}(y))| \overset{\text{C.S.}}{\leq} \sqrt{\operatorname{Var}(D_v\mathbf{f}(x))\operatorname{Var}(\mathbf{f}(y))} = \sqrt{-C'(0)C(0)}.$$

As the bound is independent of η this yields the claim. $\square$

Lemma 5.D.8 (Bound on the second derivative of the covariance)**.**

$$\sup_{\theta\geq 0}|C''(\theta)| \leq \max\Big\{\sup_{\theta\in[0,1]}|C''(\theta)|, |C'(0)|\Big\}$$

Proof. Note that

$$\mathrm{Cov}(D_v\mathbf{f}(x), D_w\mathbf{f}(y)) = -C''\big(\tfrac{\|x-y\|^2}{2}\big)\langle x-y, v\rangle\langle x-y, w\rangle - C'\big(\tfrac{\|x-y\|^2}{2}\big)\langle v, w\rangle$$

Selecting v, w as orthonormal vectors (e.g. $v = e_1, w = e_2$) and $x - y := \eta(v + w)$ for some $\eta > 0$ results in $\|x-y\|^2 = 2\eta^2$ and thus by the Cauchy-Schwarz inequality

$$\begin{aligned}\big|-C''(\eta^2)\eta^2\big| &= \big|\mathrm{Cov}(D_v\mathbf{f}(x), D_w\mathbf{f}(y))\big| \\ &\overset{\text{C.S.}}{\leq} \sqrt{\mathrm{Var}(D_v\mathbf{f}(x))\,\mathrm{Var}(D_w\mathbf{f}(y))} \\ &= \sqrt{(-C'(0))^2}\end{aligned}$$

This implies the claim. □

5.D.4 Section 5.4: Stochastic loss

Extension 5.4.2 (S-RFD)**.** *For loss* $\mathbf{f} \sim \mathcal{N}(\mu, C)$ *and stochastic errors* $\varsigma_i \overset{\text{iid}}{\sim} \mathcal{N}(0, C_\varsigma)$ *we have*

$$\arg\min_{\mathbf{d}} \mathbb{E}[\mathbf{f}(x-\mathbf{d}) \mid \mathbf{Y}_n(x), \nabla\mathbf{Y}_n(x)] = \eta^*(\Theta)\tfrac{\nabla\mathbf{Y}_n(x)}{\|\nabla\mathbf{Y}_n(x)\|}$$

with the same step size function η^* *as for RFD, but modified* Θ

$$\Theta = \frac{C'(0)}{C'(0)+\frac{1}{b_n}C'_\varsigma(0)}\frac{C(0)+\frac{1}{b_n}C_\varsigma(0)}{C(0)}\frac{\|\nabla\mathbf{Y}_n(x)\|}{\mu - \mathbf{Y}_n(x)}.$$

Proof. Since ϵ_i are conditionally independent between each other and to $\mathbf{f}$, as entire functions, the same holds true for $\nabla\epsilon_i$. As all the mixed covariances disappear we have

$$\begin{aligned}\mathrm{Cov}\Big(\begin{pmatrix}\mathbf{Y}_n(x)\\ \nabla\mathbf{Y}_n(x)\end{pmatrix}\Big) &= \mathrm{Cov}\Big(\begin{pmatrix}\mathbf{f}(x)\\ \nabla\mathbf{f}(x)\end{pmatrix}\Big) + \frac{1}{b_n^2}\sum_{i=1}^{b_n}\mathrm{Cov}\Big(\begin{pmatrix}\varsigma_n^{(i)}(x)\\ \nabla\varsigma_n^{(i)}(x)\end{pmatrix}\Big) \\ &= \begin{pmatrix}C(0) & \\ & -C'(0)\mathbb{I}_{d\times d}\end{pmatrix} + \frac{1}{b_n^2}\sum_{i=1}^{b_n}\begin{pmatrix}C_\varsigma(0) & \\ & -C'_\varsigma(0)\mathbb{I}_{d\times d}\end{pmatrix} \\ &= \begin{pmatrix}C(0)+\frac{1}{b_n}C_\varsigma(0) & \\ & -\Big(C'(0)+\frac{1}{b_n}C'_\varsigma(0)\Big)\mathbb{I}_{d\times d}.\end{pmatrix}\end{aligned}$$

by Lemma 4.5.1. If you want to break up the first step we recommend considering individual entries of the covariance matrix to convince yourself that all the mixed covariances disappear. Together with the fact

$$\begin{aligned}&\mathrm{Cov}\Big(\mathbf{f}(x-\mathbf{d}), \begin{pmatrix}\mathbf{Y}_n(x)\\ \nabla\mathbf{Y}_n(x)\end{pmatrix}\Big) \\ &= \mathrm{Cov}\Big(\mathbf{f}(x-\mathbf{d}), \begin{pmatrix}\mathbf{f}(x)\\ \nabla\mathbf{f}(x)\end{pmatrix}\Big) + \frac{1}{b_n^2}\sum_{i=1}^{b_n}\underbrace{\mathrm{Cov}\Big(\mathbf{f}(x-\mathbf{d}), \begin{pmatrix}\varsigma_n^{(i)}(x)\\ \nabla\varsigma_n^{(i)}(x)\end{pmatrix}\Big)}_{=0} \\ &= \begin{pmatrix}C(\frac{\|\mathbf{d}\|^2}{2})\\ C'(\frac{\|\mathbf{d}\|^2}{2})\mathbf{d}\end{pmatrix}.\end{aligned}$$

The rest is analogous to Lemma 5.2.1 and Theorem 5.2.2, so we only sketch the remaining steps.

Applying Theorem 2.A.1 as in Lemma 5.2.1 we obtain a stochastic version of the stochastic Taylor approximation ("stochastic^2 Taylor approximation" perhaps?)

$$\begin{aligned}&\mathbb{E}[\mathbf{f}(x-\mathbf{d}) \mid \mathbf{Y}_n(x), \nabla\mathbf{Y}_n(x)]\\ &= \mu + \frac{C\left(\frac{\|\mathbf{d}\|^2}{2}\right)}{C(0)+\frac{1}{b_n}C_\varsigma(0)}(\mathbf{Y}_n(x)-\mu) - \frac{C'\left(\frac{\|\mathbf{d}\|^2}{2}\right)}{C'(0)+\frac{1}{b_n}C'_\varsigma(0)}\langle \mathbf{d}, \mathbf{Y}_n(x)\rangle.\end{aligned}$$

Minimizing this subject to a constant step size as in Theorem 5.2.2 results in

$$\begin{aligned}\eta^* &= \arg\min_{\eta\in\mathbb{R}} \frac{C\left(\frac{\|\mathbf{d}\|^2}{2}\right)}{C(0)+\frac{1}{b_n}C_\varsigma(0)}(\mathbf{Y}_n(x)-\mu) - \eta\frac{C'\left(\frac{\|\mathbf{d}\|^2}{2}\right)}{C'(0)+\frac{1}{b_n}C'_\varsigma(0)}\|\mathbf{Y}_n(x)\|\\ &= \arg\min_{\eta\in\mathbb{R}} -\frac{C\left(\frac{\|\mathbf{d}\|^2}{2}\right)}{C(0)} - \eta\frac{C'\left(\frac{\|\mathbf{d}\|^2}{2}\right)}{C'(0)+\frac{1}{b_n}C'_\varsigma(0)}\frac{C(0)}{C(0)+\frac{1}{b_n}C_\varsigma(0)}\frac{\|\mathbf{Y}_n(x)\|}{\mu-\mathbf{Y}_n(x)},\end{aligned}$$

where we divided the term by $\frac{C(0)}{C(0)+\frac{1}{b_n}C_\varsigma(0)}\frac{1}{\mu-\mathbf{Y}_n(x)} \geq 0$ to obtain the last equation. The claim follows by definition of $\eta^*(\Theta)$ and our redefinition of Θ. □

5.E Extensions

In this section we present a few possible extensions to Theorem 5.2.2, which are all composable, i.e. it is possible to combine these extensions without any major problems (including S-RFD, i.e. Extension 5.4.2).

5.E.1 Geometric anisotropy/Adaptive step sizes

In this section, we generalize isotropy to "geometric anisotropies",[40] which provide good insights into the inner workings of adaptive learning rates (e.g. AdaGrad[41] and Adam[42]).

[40] Stein, 1999, p. 17.

[41] Duchi et al., 2011.

[42] Diederik P. Kingma and Ba, 2015.

Definition 5.E.1 (Geometric Anisotropy)**.** We say a random function $\mathbf{f}$ exhibits a "geometric anisotropy", if there exists an invertible matrix A such that $\mathbf{f}(x) = \mathbf{g}(Ax)$ for some isotropic random function $\mathbf{g}$.

This implies that the expectation of $\mathbf{f}$ is still constant ($\mathbb{E}[\mathbf{f}(x)] = \mathbb{E}[\mathbf{g}(Ax)] = \mu$) and the covariance function of $\mathbf{f}$ is given by

$$\operatorname{Cov}(\mathbf{f}(x), \mathbf{f}(y)) = \operatorname{Cov}(\mathbf{g}(Ax), \mathbf{g}(Ay)) = C\left(\frac{\|A(x-y)\|^2}{2}\right) = C\left(\frac{\|x-y\|^2_{A^TA}}{2}\right) \tag{5.24}$$

where $\|\cdot\|_\Sigma$ is the norm induced by $\langle x, y\rangle_\Sigma := \langle x, \Sigma y\rangle$ for some strictly positive definite matrix $\Sigma = A^TA$. Here (5.24) characterizes the set of random functions with a geometric anisotropy in the Gaussian case, because for an $\mathbf{f}$ with such a covariance we can always obtain an isotropic $\mathbf{g}$ by $\mathbf{g}(x) := \mathbf{f}(A^{-1}x)$. This is the whitening transformation we suggest looking for in order to ensure isotropy in the context of scale invariance (Section 5.1).

An important observation is, that Theorem 4.2.3 implies that $\mathbf{f}$ is still stationary, so the distribution of $\mathbf{f}$ is still invariant to translations. If stationarity is a problem, this is therefore not the solution. But geometric anisotropies are a beautiful model to explain preconditioning and adaptive step sizes. For this, we first determine the RFD steps.

Extension 5.E.2 (RFD steps under geometric anisotropy)**.** *Let $\mathbf{f}$ be a Gaussian random function which exhibits a "geometric anisotropy" A and is based on an isotropic random function $\mathbf{g} \sim \mathcal{N}(\mu, C)$. Then the RFD steps are given by*

$$\eta^*\frac{\Sigma^{-1}\nabla\mathbf{f}(x)}{\|\Sigma^{-1}\nabla\mathbf{f}(x)\|_\Sigma} = \arg\min_{\mathbf{d}} \mathbb{E}[\mathbf{f}(x-\mathbf{d}) \mid \mathbf{f}(x), \nabla\mathbf{f}(x)]$$

with

$$\eta^* = \arg\min_\eta q_\Theta(\eta) \quad \textit{where} \quad \Theta = \frac{\|\Sigma^{-1}\nabla\mathbf{f}(x)\|_\Sigma}{\mu - \mathbf{f}(x)}.$$

Proofsketch. There are two ways to see this. Either we apply scale invariance (Advantage 5.1.3) directly to translate the steps on $\mathbf{g}$ into steps on $\mathbf{f}$. Alternatively one can manually retrace the steps of the proof. Details in Subsection 5.E.1 □

The step direction is therefore

$$\Sigma^{-1}\nabla\mathbf{f}(x)$$

and Σ^{-1} acts as a preconditioner. So how would one obtain Σ? As it turns out the following holds true (by Lemma 4.5.1)

$$\mathbb{E}[\nabla\mathbf{f}(x)\nabla\mathbf{f}(x)^T] = A^T\mathbb{E}[\nabla\mathbf{g}(x)\nabla\mathbf{g}(x)^T]A = A^T(-C'(0)\mathbb{I})A = -C'(0)\Sigma$$

In their proposal of the first "adaptive" method, AdaGrad, Duchi et al. (2011) suggest to use the matrix

$$G_t = \sum_{k=1}^{t} \nabla\mathbf{f}(x_k)\nabla\mathbf{f}(x_k)^T,$$

which is basically already looking like an estimation method of Σ. They then restrict themselves to $\operatorname{diag}(G_t)$ due to the computational costs of a full matrix inversion. This results in entry-wise ("adaptive") learning rates. Later adaptive methods like RMSProp,[43] AdaDelta[44] and in particular Adam[45] replace this sum with an exponential mean estimate, i.e. in the case of Adam the decay rate β_2 is used to get an exponential moving average

$$v_t = \beta_2 v_{t-1} + (1-\beta_2)\operatorname{diag}(\nabla\mathbf{f}(x_t)\nabla\mathbf{f}(x_t)^T) = \beta_2 v_{t-1} + (1-\beta_2)(\nabla\mathbf{f}(x_t))^2.$$

[43] Hinton, 2012.

[44] Zeiler, 2012.

[45] Diederik P. Kingma and Ba, 2015.

They then take the expectation

$$\begin{aligned}\mathbb{E}[v_t] &= \mathbb{E}\Big[(1-\beta_2)\sum_{k=1}^{t}\beta_2^{t-k}\nabla\mathbf{f}(x_k)^2\Big] \\ &= \mathbb{E}\Big[(1-\beta_2)\sum_{k=1}^{t}\beta_2^{t-k}\nabla\mathbf{f}(x_k)^2\Big] = (1-\beta_2^t)\underbrace{\mathbb{E}[\nabla\mathbf{f}(x_t)^2]}_{\propto\operatorname{diag}(\Sigma)}\end{aligned}$$

So $\hat{v}_t = v_t/(1-\beta_2^t)$ in the Adam optimizer is essentially an estimator for $\operatorname{diag}(\Sigma)$. It is noteworthy, that Diederik P. Kingma and Ba (2015) already used the expectation symbol. This is despite the fact, that they did not yet model the optimization objective $\mathbf{f}$ as a random function.

We can not yet explain why they then use the square root of their estimate $\operatorname{diag}(\Sigma)^{-1/2}$ instead of $\operatorname{diag}(\Sigma)^{-1}$ itself. This might have something to do with the fact that the estimation of G_t happens online and the $\mathbf{f}(x_k)$ are therefore highly correlated. Another reason might be that the inverse of an estimator has different properties than the estimator itself. Finally, the fact that only the diagonal is used might also be the reason, if the preconditioner $\operatorname{diag}(\Sigma)^{-1/2}$ is simply better when we restrict ourselves to diagonal matrices.

Proof of Extension 5.E.2

Since the application of scale invariance provides no intuition, we provide a proof which retraces some of the steps of the original proof.

Recall, that for an isotropic random function $\mathbf{g}$ we have the stochastic Taylor approximation

$$\mathbb{E}[\mathbf{g}(x-\mathbf{d}) \mid \mathbf{g}(x), \nabla\mathbf{g}(x)] = \mu + \frac{C(\frac{\|\mathbf{d}\|^2}{2})}{C(0)}(\mathbf{g}(x)-\mu) + \frac{C'(\frac{\|\mathbf{d}\|^2}{2})}{C'(0)}\langle\mathbf{d}, \nabla\mathbf{g}(x)\rangle$$

This implies for a random function with geometric anisotropy $\mathbf{f}(x) = \mathbf{g}(Ax)$ that

$$\begin{aligned}
&\mathbb{E}[\mathbf{f}(x-\mathbf{d}) \mid \mathbf{f}(x), \nabla\mathbf{f}(x)] \\
&= \mathbb{E}[\mathbf{g}(A(x-\mathbf{d})) \mid \mathbf{g}(Ax), \nabla\mathbf{g}(Ax)] \\
&= \mu + \frac{C\left(\frac{\|A\mathbf{d}\|^2}{2}\right)}{C(0)}(\mathbf{g}(Ax)-\mu) - \frac{C'\left(\frac{\|A\mathbf{d}\|^2}{2}\right)}{C'(0)}\langle A\mathbf{d}, \nabla\mathbf{g}(Ax)\rangle \\
&= \mu + \frac{C\left(\frac{\|\mathbf{d}\|_\Sigma^2}{2}\right)}{C(0)}(\mathbf{f}(x)-\mu) - \frac{C'\left(\frac{\|\mathbf{d}\|_\Sigma^2}{2}\right)}{C'(0)}\langle \mathbf{d}, \underbrace{A^T\nabla\mathbf{g}(Ax)}_{=\nabla\mathbf{f}(x)}\rangle
\end{aligned}$$

with $\Sigma := A^TA$. As in the original proof, we now optimize over the direction first, while keeping the step size constant, although we now fix the step size with regard to the norm $\|\cdot\|_\Sigma$ (which basically means that we still do the optimization in the isotropic space). Note that

$$\max_{\mathbf{d}} \langle \mathbf{d}, \nabla\mathbf{f}(x)\rangle \quad \text{s.t.} \quad \|\mathbf{d}\|_\Sigma = \eta$$

is equivalent to

$$\max_{\mathbf{d}} \langle \mathbf{d}, \Sigma^{-1}\nabla\mathbf{f}(x)\rangle_\Sigma \quad \text{s.t.} \quad \|\mathbf{d}\|_\Sigma = \eta$$

which is solved by

$$\pm\eta\frac{\Sigma^{-1}\nabla\mathbf{f}(x)}{\|\Sigma^{-1}\nabla\mathbf{f}(x)\|_\Sigma}$$

The remainder of the proof is exactly the same as in the original.

5.E.2 Conservative RFD

In the first paragraph of Section 5.1 we motivated the relation between RFD and classical optimization with the observation, that gradient descent is the minimizer of a regularized first order Taylor approximation

$$\tfrac{1}{L}\nabla\mathbf{f}(x) = \arg\min_{\mathbf{d}} T[\mathbf{f}(x-\mathbf{d}) \mid \mathbf{f}(x), \nabla\mathbf{f}(x)] + \tfrac{L}{2}\|x\|^2.$$

This regularized Taylor approximation is in fact an upper bound on our function under the L-smoothness assumption,[46] i.e.

[46] Nesterov, 2018.

$$\mathbf{f}(x-\mathbf{d}) \leq T[\mathbf{f}(x-\mathbf{d}) \mid \mathbf{f}(x), \nabla\mathbf{f}(x)] + \tfrac{L}{2}\|\mathbf{d}\|^2$$

An improvement on of this upper bound compared to $\mathbf{f}(x)$ therefore guarantees an improvement of the loss. This guarantee was lost with the conditional expectation (on purpose, as we wanted to consider the average case). Losing this guarantee also makes convergence proofs more difficult as they typically make use of this improvement. In view of the confidence intervals of Figure 5.1, it is natural to ask for a similar upper bound in the random setting, where this can only be the top of an confidence interval. This is provided in the following theorem

Lemma 5.E.3 (An γ-upper bound)**.** *We have*

$$\mathbb{P}\Big(\mathbf{f}(x-\mathbf{d}) \leq \mathbb{E}[\mathbf{f}(x-\mathbf{d}) \mid \mathbf{f}(x), \nabla\mathbf{f}(x)] + \rho_\gamma(\|\mathbf{d}\|^2)\Big) \geq \gamma$$

for $\rho_\gamma(\eta^2) := \Phi^{-1}(\gamma)\sigma(\eta^2)$ *with*

$$\sigma^2(\eta^2) := C(0) - \frac{C\left(\frac{\eta^2}{2}\right)^2}{C(0)} - \frac{C'\left(\frac{\eta^2}{2}\right)^2}{-C'(0)}\eta^2$$

where Φ *is the cumulative distribution function (cdf) of the standard normal distribution.*

Proof. Note that the conditional variance is with the usual argument about the covariance matrices (cf. the proof of Thoerem 5.2.2) using Lemma 4.5.1 and an application of Theorem 2.A.1 given by

$$\sigma^2(\|x\|^2) := \operatorname{Var}[\mathbf{f}(x-\mathbf{d}) \mid \mathbf{f}(x), \nabla\mathbf{f}(x)] = C(0) - \frac{C\left(\frac{\|\mathbf{d}\|^2}{2}\right)^2}{C(0)} - \frac{C'\left(\frac{\|\mathbf{d}\|^2}{2}\right)^2}{-C'(0)}\|\mathbf{d}\|^2.$$

Since the conditional distribution is normal (by Theorem 2.A.1), we have

$$\frac{\mathbf{f}(x-\mathbf{d}) - \mathbb{E}[\mathbf{f}(x-\mathbf{d}) \mid \mathbf{f}(x), \nabla\mathbf{f}(x)]}{\sigma(\|x\|^2)} \sim \mathcal{N}(0,1).$$

But this implies the claim

$$\begin{aligned}
&\mathbb{P}\Big(\mathbf{f}(x-\mathbf{d}) \le \mathbb{E}[\mathbf{f}(x-\mathbf{d}) \mid \mathbf{f}(x), \nabla\mathbf{f}(x)] + \rho_\gamma(\|\mathbf{d}\|^2)\Big) \\
&= \mathbb{P}\Big(\frac{\mathbf{f}(x-\mathbf{d}) - \mathbb{E}[\mathbf{f}(x-\mathbf{d}) \mid \mathbf{f}(x), \nabla\mathbf{f}(x)]}{\sigma(\|x\|^2)} \le \Phi^{-1}(\gamma)\Big) \\
&= \Phi(\Phi^{-1}(\gamma)) = \gamma.
\end{aligned}$$

To avoid the Gaussian assumption, one could apply the Markov inequality instead, or another applicable concentration inequality. □

Using this upper bound, we obtain a natural conservative extension of RFD

Extension 5.E.4 (γ-conservative RFD)**.** *Let $\mathbf{f} \sim \mathcal{N}(\mu, C)$ be a Gaussian random function and $\rho_\gamma(\eta^2) := \Phi^{-1}(\gamma)\sigma(\eta^2)$, where σ is the conditional standard deviation as defined in Lemma 5.E.3. Then the conservative RFD step direction is given by*

$$\eta^* \frac{\nabla\mathbf{f}(x)}{\|\nabla\mathbf{f}(x)\|} = \arg\min_{\mathbf{d}} \mathbb{E}[\mathbf{f}(x-\mathbf{d}) \mid \mathbf{f}(x), \nabla\mathbf{f}(x)] + \rho_\gamma(\|\mathbf{d}\|^2)$$

and the γ-conservative RFD step size is given by

$$\eta^* = \arg\min_{\eta} \frac{C\left(\frac{\eta^2}{2}\right)}{C(0)}(\mathbf{f}(x)-\mu) - \eta\frac{C'\left(\frac{\eta^2}{2}\right)}{C'(0)}\|\nabla\mathbf{f}(x)\| + \rho_\gamma(\eta^2).$$

Proof. The proof is the same as in Theorem 5.2.2 with Lemma 5.2.1 replaced by Lemma 5.E.3. □

Taking multiple steps should generally have an averaging effect, so we expect faster convergence for almost risk neutral minimization of the conditional expectation (i.e. $\gamma \approx \frac{1}{2}$). Here γ is a natural parameter to vary conservatism. In a software implementation it might be a good idea to call this parameter 'conservatism' and rescale it to be in $[0,1]$ instead of $[\frac{1}{2},1]$. But formulas look cleaner with γ.

In Bayesian optimization it is much more common to reverse this approach and minimize a lower confidence bound ('conservatism' < 0 or $\gamma < \frac{1}{2}$) in order to encourage exploration. But since RFD is forgetful, this is not a good idea for RFD.

Remark 5.E.5 (Conservative RFD coincides asymptotically with RFD in high dimension). While conservative RFD might seem like a good approach to fix the instability of RFD under the isotropy assumption on some optimization problems, the variance generally vanishes in high dimension[47] and conservative RFD coincides asymptotically with RFD. We therefore believe that the underlying issue is not an overly risk-affine algorithm but rather that distributional assumptions, in particular the stationarity assumption, are violated when instabilities occur (cf. Section 7.4 and 4.3).

[47] see Chapter 6

Nevertheless, conservative RFD might be a good approach for lower dimensional, risk-sensitive applications.

5.E.3 Beyond the Gaussian assumption

In this section we sketch how the extension beyond the Gaussian case using the "best linear unbiased estimator" BLUE[48] works.[49]

For this we recapitulate what a BLUE is. A **linear estimator** $\hat{Y}$ of Y using $X_1, \dots, X_n$ is of the form

$$\hat{Y} \in \operatorname{span}\{X_1, \dots, X_n\} + \mathbb{R}.$$

The set of **linear unbiased estimators** is defined as

$$\begin{aligned} \operatorname{LUE} &:= \operatorname{LUE}[Y \mid X_1, \dots, X_n] \\ &:= \left\{\hat{Y} \in \operatorname{span}\{X_1, \dots, X_n\} + \mathbb{R} : \mathbb{E}[\hat{Y}] = \mathbb{E}[Y]\right\} \\ &= \{\hat{Y} + \mathbb{E}[Y] : \hat{Y} \in \operatorname{span}\{X_1 - \mathbb{E}[X_1], \dots, X_n - \mathbb{E}[X_n]\}\}. \end{aligned} \tag{5.25}$$

And the BLUE is the **best linear unbiased estimator**, i.e.

$$\operatorname{BLUE}[Y \mid X_1, \dots, X_n] := \operatorname*{arg\,min}_{\hat{Y} \in \operatorname{LUE}} \mathbb{E}[\|\hat{Y} - Y\|^2]. \tag{5.26}$$

Other risk functions to minimize are possible, but this is the usual one.

Lemma 5.E.6. *If $X, Y_1, \dots, Y_n$ are multivariate normal distributed, then we have*

$$\operatorname{BLUE}[Y \mid X_1, \dots, X_n] = \mathbb{E}[Y \mid X_1, \dots, X_n] \left(= \operatorname*{arg\,min}_{\hat{Y} \in \{f(X_1, \dots, X_n) : f \text{ meas.}\}} \mathbb{E}[\|Y - \hat{Y}\|^2] \right).$$

Proof. We observe that the conditional expectation of Gaussian random variables is linear (Theorem 2.A.1). So as a linear function its L^2 risk must be larger or equal to that of the BLUE. And as an L^2 projection[50] the conditional expectation was already optimal. □

If we now replace the conditional expectation with the BLUE, then all our theory remains the same because the result in Theorem 2.A.1 remains the BLUE for general distributions.[51] Instead of minimizing

$$\min_{\mathbf{d}} \mathbb{E}[\mathbf{f}(x - \mathbf{d}) \mid \mathbf{f}(x), \nabla \mathbf{f}(x)]$$

we can therefore always minimize

$$\min_{\mathbf{d}} \operatorname{BLUE}[\mathbf{f}(x - \mathbf{d}) \mid \mathbf{f}(x), \nabla \mathbf{f}(x)]$$

without the Gaussian assumption and all our results can be translated to this case. The reader only needs to replace all mentions of Theorem 2.A.1 with the BLUE equivalent and replace all "idependence" claims with "uncorrelated".

[48] e.g. Johnson and Wichern, 2007, ch. 7.

[49] this is the approach is common in geostatistics

[50] Klenke, 2014, Cor. 8.17.

[51] Johnson and Wichern, 2007.

Chapter 6

Predictable progress

In this chapter we prove that general first order optimization algorithms applied to realizations of high dimensional (Gaussian random functions) GRFs essentially make deterministic progress, regardless of the initialization and the realization. Since all quantities that appear are computable we call this property **predictable progress** of the optimizer.

Motivation 6.0.1 (Machine learning)**.** Our research is motivated by an astonishing phenomenon from the training of (large) machine learning models which this chapter tackles in a rigorous way. 'Training' is the process of minimizing a cost (also known as error or risk) function $f : \mathbb{R}^N \to \mathbb{R}$ over parameter vectors $x \in \mathbb{R}^N$ by selecting successively parameter vectors $x_n \in \mathbb{R}^N$ based on the evaluations $f(x_k)$ and gradients $\nabla f(x_k)$ at the previous parameter vectors $x_0, \dots x_{n-1}$. The parameters are typically the weights of a neural network and N is very large. The phenomenon is that over multiple initializations $x_0^{(1)}, x_0^{(2)}, \dots \in \mathbb{R}^N$ the progress a particular optimizer makes during the optimization is approximately equal over different optimization runs:

$$(f(x_0^{(i)}), f(x_1^{(i)}), \dots) \approx (f(x_0^{(j)}), f(x_1^{(j)}), \dots).$$

Figure 6.1 shows predictable progress in practice. We demonstrate predictable progress ourselves for the training on the MNIST dataset using a standard convolutional neural network (cf. Figure 6.1a). The key to predictable progress is high dimensionality of parameters (e.g. the neural network of Figure 6.1a has roughly $N = 2.3$ million parameters, which is small in comparison to neural networks used for tasks beyond the "toy-problem" MNIST). Figure 6.1b by[1] demonstrates that this phenomenon is well known and moreover relied upon for training heuristics.

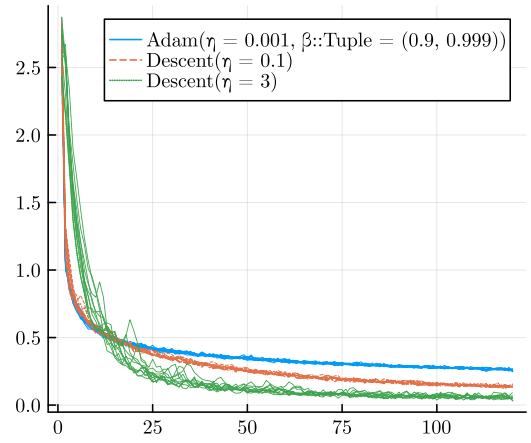
[1] Klein et al., 2017.

The predictable progress phenomenon is certainly surprising since there is no reason to believe that function values along an optimizer should be approximatively equal for different initializations. And, in view of handcrafted deterministic counterexamples, predictable progress is obviously not provable for arbitrary high-dimensional functions. Instead, we will prove this behavior under the assumption that the cost function f is the realization of a Gaussian random function $\mathbf{f}$ (GRF), because GRFs have turned out to be a fruitful model for the statistical properties of cost functions encountered in machine learning.[2] Modelling the cost function of Motivation 6.0.1 as a realization of a random function could for instance be motivated by the assumption that the available sequence of data is exchangeable. Since the cost is typically the infinite data limit over the average sample loss, where the loss measures the prediction error on a single sample of the data, de Finetti's representation theorem[3] implies the existence of an underlying random measure driving the randomness of the cost.

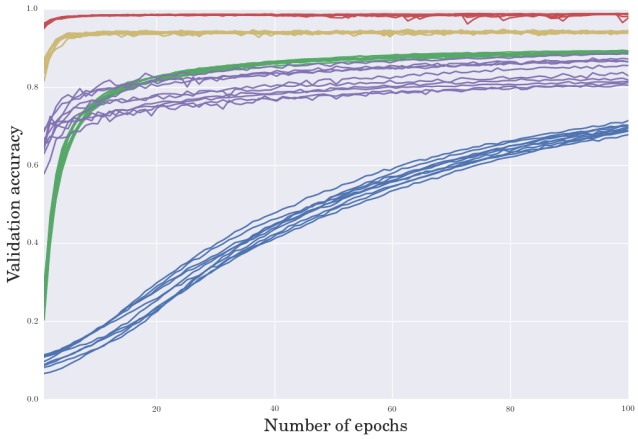
[2] e.g. Pascanu, Mikolov, et al., 2013; Yann N Dauphin et al., 2014; Choromanska et al., 2015.

[3] e.g. Schervish, 1995, Theorem 1.49.

While we have motivated predictable progress only for different initializations x_0 in Motivation 6.0.1, the mathematical results are much deeper. We show that predictable progress for Gaussian random functions is also independent of the particular realization, that is, for a given covariance model the values seen in an optimization run are also independent of the realization of the GRF $\mathbf{f}$.



(a) The plot shows an empirical approximation of the cost sequence resulting from the training of a standard convolutional neural network on the MNIST dataset.[a] We plot the objective $f(x_0^{(i)}), \dots, f(x_{120}^{(i)})$ against the steps $0, \dots, 120$ on the x-axis. The minimization is performed with three optimization algorithms: Adam[b] (with learning rate η and momentum β) in blue and two version of gradient descent (learning rate $\eta = 0.1$ and $\eta = 3$) in red and green. Each optimizer was run 10 times from randomly selected initializations $x_0^{(i)}$ using the (random) default initialization procedure known as Glorot initialization.[c]



(b) The plot, taken from Klein et al. (2017, Figure 3), shows 5 different optimization hyperparameter configurations each evaluated 10 times (grouped by color). The predictable progress of the validation accuracy per configuration is used in Klein et al. (2017) to argue that it is sufficient to try a configuration once. The overall goal of Klein et al. (2017) is furthermore to fit a parametric models to the 'learning curves' in order to stop training early and switch to a different configuration if the progression does not seem promising.
The training heuristics that aspire to be 'AutoML' (i.e. automatically fit data without human intervention) are therefore built on this phenomenon. And the fact that these training heuristics are so successful demonstrates how ubiquitous predictable progress is in practice.

[a] LeCun et al., 2010.
[b] Diederik P. Kingma and Ba, 2015.
[c] Glorot and Bengio, 2010.

FIGURE 6.1: Examples of predictable progress in machine learning practice

Recently, predictable progress has been proven for **random quadratic functions** generated from the mean squared error applied to linear models.[4] Under the term 'state evolution' the asymptotic behavior of 'approximate message passing' (AMP) algorithms has also been analyzed on random quadratic functions.[5] AMP algorithms can be related to gradient span algorithms in this quadratic setting induced by linear regression problems.[6] While quadratic functions are a simplified convex setting, these contributions offered a first explanation for the observed phenomenon of predictable behavior in high dimensions. We will extend these results to the much more general setting of Gaussian random functions. This setting was already used in the machine learning literature for instance by Pascanu, Yann N. Dauphin, et al. (2014) and Yann N Dauphin et al. (2014) to explain why the overwhelming share of critical points are saddle points in high dimension, whereas the critical points of low dimensional GRFs are dominated by minima and maxima.[7] Similarly, our results will show once again that high-dimensional behavior often eludes low dimensional intuitions. The specific modelling assumption of Pascanu, Yann N. Dauphin, et al. (2014) and Yann N Dauphin et al. (2014) were stationary isotropic GRFs.

[4] C. Paquette et al., 2022; E. Paquette and Trogdon, 2022; Deift and Trogdon, 2021; Pedregosa and Scieur, 2020; Scieur and Pedregosa, 2020.

[5] Bayati and Montanari, 2011; Javanmard and Montanari, 2013.

[6] Celentano et al., 2020.

[7] Rasmussen and Williams, 2006.

Similar to the convention of capital letters for random variables, we use bold font to denote random functions $\mathbf{f}$. Since our results are limit theorems in the dimension N we mark this important dependency in the index. While the parameter sequences would also have to be indexed by the dimension N we omit this index for notational clarity. In simplified terms we will prove the following theorem for predictable optimization in high dimension, precise theorems are given in Section 6.1:

> **If $\mathbf{f}_N$ is a (non-stationary) isotropic GRF on a high-dimensional domain $\mathbb{R}^N$, then the (random) sequence $\mathbf{f}_N(X_0), \mathbf{f}_N(X_1), \dots$ along the (random) parameter sequence $X_0, X_1, \dots$ selected by a standard first order optimization algorithm are close to a deterministic sequence $\mathfrak{f}_0, \mathfrak{f}_1, \dots$ with high probability.**

To be a bit more precise, we will prove the following in Theorem 6.1.10: Suppose $X_0, X_1, \dots$ is the (random) sequence of parameter points obtained by running an optimizer on the (random) function $\mathbf{f}_N : \mathbb{R}^N \to \mathbb{R}$ initialized at an independent,

possibly random point X_0. For 'gradient span algorithms' $\mathfrak{G}$ (e.g. gradient descent) and a sequence of (non-stationary) isotropic GRFs $(\mathbf{f}_N)_{N\in\mathbb{N}}$ we construct a sequence of deterministic real numbers $\mathfrak{f}_0, \mathfrak{f}_1, ...$ such that, for $n \in \mathbb{N}$,

$$\lim_{N\to\infty} \mathbb{P}\Big(|\mathbf{f}_N(X_n) - \mathfrak{f}_n| > \epsilon\Big) = 0, \quad \forall \epsilon > 0.$$

The proof is completely constructive and, in principle, allows for the numerical computation of the limiting values $\mathfrak{f}_n$ with complexity $\mathcal{O}(n^6)$ given an algorithm $\mathfrak{G}$, the mean and covariance kernel of the random functions $\mathbf{f}_N$ and the length of the initialization vector X_0. In Corollary 6.1.11 we will show that an application of the stochastic triangle inequality yields approximately equal optimization progress given different initialization points (as can be seen in Figure 6.1).

The general modelling assumption of (non-stationary) isotropy can be introduced in two different ways. The first is through the covariance function (Section 6.1.2), the second is axiomatically through certain invariances. To motivate the latter in the context of predictability let us highlight a case that must be avoided. Think about some one-dimensional function $f_1 : \mathbb{R} \to \mathbb{R}$ where the values along the optimization path (say of gradient descent) strongly depend on the initialization. Lift this one dimensional function into $\mathbb{R}^N$ by $f_N(x) := f_1(\langle v, x\rangle)$ for some direction vector v. Then f_N certainly does not share the features of high dimensional cost functions encountered in machine learning, since the cost along the optimization path depends heavily on the initialization again. To explain the phenomenon of predictable cost sequences (cf. Figure 6.1) the model for cost functions therefore should not be reliant on particular directions v. Thus, we assume that directions should be exchangeable, which suggests at least rotation and reflection invariant random functions as a model. This assumption was introduced in Chapter 4.2 as '(non-stationary) isotropy' to generalize the well known setting of stationary isotropy.

In spirit, this chapter is close to the **spin glass** community. (Non-stationary) isotropy captures the setting of spin-glasses, we use the same type of dimensional scaling and prove similar limiting results. Finally, there has also been recent progress in understanding optimization of spin glasses (see Section 6.2.1 for further details and e.g. Auffinger et al. (2023) for a review of optimization of spin glasses).

Outline In Sections 6.1.1-6.1.3 we formalize the setting (gradient span optimization algorithms and (non-stationary) isotropic GRFs) and state the main result. Our mathematical formulation of the original problem using limit formulations requires a dimensional scaling of the random functions that is also standard in spin glass theory. This scaling, crucial to our approach, is discussed in Section 6.2. The proof of our main result is given in Section 6.3, first as a sketch assuming stationary isotropy and then in detail.

6.1 Setting and main results

To enable our asymptotic analysis, a careful setting of the scene is required. Both, the optimization algorithms considered and the distribution of $\mathbf{f}_N$ need a representation independent of the dimension N to allow for an analysis of $N \to \infty$.

6.1.1 The class of optimization algorithms $\mathfrak{G}$

Given a sufficiently smooth function $f : \mathbb{R}^N \to \mathbb{R}$, a naive optimization algorithm is given by gradient descent, whose evaluation points are recursively defined as

$$x_n = x_{n-1} - \alpha \nabla f(x_{n-1})$$

with some initialization $x_0 \in \mathbb{R}^N$ and a learning rate α. The reader might want to keep gradient descent in mind as a toy example, but everything we prove holds for a much larger class of first order algorithms that contains many standard

optimizers. Gradient span algorithms (GSA) are this very general class of first order algorithms which pick the n-th point x_n from the previous span of gradients

$$x_n \in \operatorname{span}\{x_0, \nabla f(x_0), \dots, \nabla f(x_{n-1})\}. \tag{6.1}$$

GSAs contain classic gradient descent, momentum methods such as heavy-ball momentum[8] and Nesterov's momentum,[9] and also the conjugate gradient method.[10] GSAs do not only encompass minimizers but also maximizers and all sorts of other algorithms. While the initial point x_0 is typically not included in the span, we admit this generalization to allow for concepts such as 'weight normalization',[11] which project points back to the sphere (see Remark 6.1.2 for further details). Since the defining property (6.1) of gradient span algorithms is not sufficient to identify a particular GSA, we define a very general parametric family of GSAs in Definition 6.1.1. This family includes all the algorithms mentioned above. Importantly, the parametrization chosen does not use dimension specific information and a fixed gradient span algorithm (such as gradient descent) can therefore be used for all dimensions.

[8] Polyak, 1964.

[9] e.g. Nesterov, 2018.

[10] Hestenes and Stiefel, 1952.

[11] Salimans and Durk P Kingma, 2016.

Definition 6.1.1 (General gradient span algorithm)**.** For a starting point $x_0 \in \mathbb{R}^N$ and a function $f : \mathbb{R}^N \to \mathbb{R}$, a *gradient span algorithm* $\mathfrak{G}$ selects

$$x_n := \mathfrak{G}(f, x_0, n) = h_n^{(x)} x_0 + \sum_{k=0}^{n-1} h_{n,k}^{(g)} \nabla f(x_k), \tag{6.2}$$

where we assume the prefactors $h_n = (h_n^{(x)}) \cup (h_{n,k}^{(g)})_{k=0,\dots,n-1}$, using the union $\cup$ to indicate a concatenation of tuples, to be functions $h_n = h_n(I_{n-1})$ of the previous dimensionless information I_{n-1} available at time n, that is

$$I_n := \Big(f(x_k) : k \le n\Big) \cup \Big(\langle v, w\rangle : v, w \in (x_0) \cup G_n\Big)$$
$$\text{with} \quad G_n := \big(\nabla f(x_k) : k \le n\big).$$

The algorithm $\mathfrak{G}$ is called x_0-*agnostic*, if

- it remains in the gradient span shifted by x_0, i.e.

$$h_n^{(x)} = 1 \qquad \forall n \in \mathbb{N}.$$

- The prefactors h_n do not use the inner products with x_0, that is they are functions of the reduced information

$$I_n^{\setminus x_0} = \Big(f(x_k), \langle \nabla f(x_k), \nabla f(x_l)\rangle : k, l \le n\Big).$$

As mentioned above the reader might just think about ordinary gradient descent, which is also x_0-agnostic. We introduce the 'x_0-agnostic' property, to allow for arbitrary initialization distributions in the stationary isotropic case. Without this property, the algorithm could, for example, use the length $\|X_0\|$ as an indicator to switch between optimizers and therefore cause non-deterministic behavior.

Since the main objective of this article is a dimensional limit statement for fixed algorithms (i.e. fixed choice of prefactors h) it is important to note that every fixed gradient span algorithm $\mathfrak{G}$ is well-defined in all dimensions because the scalar product $\langle \cdot, \cdot \rangle$ is well defined for every dimension $N \in \mathbb{N}$.

It should perhaps be noted, that the initial point x_0 is not the only point that may be used by gradient span algorithms, since we have

$$\operatorname{span}\{(x_0) \cup G_{n-1}\} = \operatorname{span}\{x_0, \dots, x_{n-1}, \nabla f(x_0), \dots, \nabla f(x_{n-1})\}$$

by induction over n as x_n is selected from this span. We claimed that the inclusion of the initial point x_0 into the span, or equivalently the prefactor $h_n^{(x)}$, would allow for algorithms projecting back to the sphere or ball. In the following, we show that our general gradient span algorithms also contain gradient span algorithms with spherical projections.

Remark 6.1.2 (Projection). Assume that the point x_n is defined by

$$x_n := P\tilde{x}_n \quad \text{with} \quad \tilde{x}_n = \tilde{h}_n^{(x)} x_0 + \sum_{k=0}^{n-1} \tilde{h}_{n,k}^{(g)} \nabla f(x_k),$$

where P is either a projection to the sphere or the ball. Note that this implies either a division by $\|\tilde{x}_n\|$ in case of the sphere, or a division by $\max\{\|\tilde{x_n}\|, 1\}$ in case of the ball. But the norm

$$\|\tilde{x}_n\|^2 = (\tilde{h}_n^{(x)})^2 \underbrace{\|x_0\|^2}_{\in I_n} + 2\tilde{h}_n^{(x)} \sum_{k=0}^{n-1} \tilde{h}_{n,k}^{(g)} \underbrace{\langle \nabla f(x_k), x_0 \rangle}_{\in I_n} + \sum_{k,l=0}^{n-1} \tilde{h}_{n,k}^{(g)} \tilde{h}_{n,l}^{(g)} \underbrace{\langle \nabla f(x_k), \nabla f(x_l) \rangle}_{\in I_n}$$

is a function of the information in I_{n-1} and can therefore be used to define

$$h_n^{(x)} := \frac{\tilde{h}_n^{(x)}}{\|\tilde{x}_n\|} \qquad h_{n,k}^{(g)} := \frac{\tilde{h}_{n,k}^{(g)}}{\|\tilde{x}_n\|}$$

in the case of the projection to the sphere and similarly for the projection to the ball.

6.1.2 The class of random function distributions

For a random function[12] $\mathbf{f}_N : \mathbb{R}^N \to \mathbb{R}$ on some probability space $(\Omega, \mathcal{A}, \mathbb{P})$ recall we denote the mean and covariance functions by

$$\mu_{\mathbf{f}_N}(x) := \mathbb{E}[\mathbf{f}_N(x)] \quad \text{and} \quad \mathcal{C}_{\mathbf{f}_N}(x, y) := \mathrm{Cov}(\mathbf{f}_N(x), \mathbf{f}_N(y)), \qquad x, y \in \mathbb{R}^N.$$

As we want to prove a limit theorem for $N \to \infty$ to make predictions about the high dimensional cost functions found in practice, the mean $\mu_{\mathbf{f}_N}$ and covariance $\mathcal{C}_{\mathbf{f}_N}$ needs to be defined for every dimension. They should also remain constant in some sense to avoid arbitrary sequences. As the domain $\mathbb{R}^N$ changes over N this requires a parametric form provided by covariance kernels. To obtain non-trivial results, some scaling is also required (Section 6.2). The scaling used is consistent with the typical scaling for spin glasses as we discuss in Section 6.2.1.

[12] Recall that 'random process', 'stochastic process' and 'random field' are used synonymously for 'random function' in the literature, that their law is characterized by all finite dimensional marginals (e.g. Klenke, 2014, Thm. 14.36) and that Gaussian random functions are fully determined by their mean and covariance.

Definition 6.1.3 (Scaled sequence of isotropic Gaussian random functions)**.**

(i) Suppose $\mu : \mathbb{R} \to \mathbb{R}$ and $\kappa : D \to \mathbb{R}$ with $D = \{\lambda \in \mathbb{R}^2_{\geq 0} \times \mathbb{R} : |\lambda_3| \leq 2\sqrt{\lambda_1 \lambda_2}\} \subseteq \mathbb{R}^3$. A scaled sequence $(\mathbf{f}_N)_{N \in \mathbb{N}}$ of GRFs $\mathbf{f}_N : \mathbb{R}^N \to \mathbb{R}$ is called *(non-stationary) isotropic* if

$$\mu_{\mathbf{f}_N}(x) = \mu\big(\tfrac{\|x\|^2}{2}\big) \quad \text{and} \quad \mathcal{C}_{\mathbf{f}_N}(x, y) = \tfrac{1}{N}\kappa\big(\tfrac{\|x\|^2}{2}, \tfrac{\|y\|^2}{2}, \langle x, y \rangle\big), \quad x, y \in \mathbb{R}^N. \tag{6.3}$$

In that case we write $\mathbf{f}_N \sim \mathcal{N}(\mu, \kappa)$.

(ii) Suppose $C : \mathbb{R}_{\geq 0} \to \mathbb{R}$. A scaled sequence $(\mathbf{f}_N)$ of GRFs $\mathbf{f}_N : \mathbb{R}^N \to \mathbb{R}$ is called *stationary isotropic* if

$$\mu_{\mathbf{f}_N}(x) = \mu \in \mathbb{R} \qquad \text{and} \qquad \mathcal{C}_{\mathbf{f}_N}(x, y) = \tfrac{1}{N} C\big(\tfrac{\|x-y\|^2}{2}\big), \quad x, y \in \mathbb{R}^N.$$

In that case we write $\mathbf{f}_N \sim \mathcal{N}(\mu, C)$.

Remark 6.1.4 ("Function"). In the generality of this definition, $\mathbf{f}_N$ might only exist in the sense of distribution but for our results we will add continuity assumptions (cf. Assumption 6.1.9) that ensure $\mathbf{f}_N$ to exist as a continuous function.

Remark 6.1.5 (Alternative definition). It is also possible to define a scaled sequence of GRFs by scaling one big (non-stationary) isotropic centered Gaussian random function $\mathbf{f}$ defined on an infinite domain as

$$\mathbf{f}_N(x) := \mu\big(\tfrac{\|x\|^2}{2}\big) + \tfrac{1}{\sqrt{N}}\mathbf{f}(x), \quad x \in \mathbb{R}^N.$$

We refer the reader who prefers this point of view to sequences of random functions with varying domains to Remark 6.2.2 below.

Remark 6.1.6 (Scaling). The importance of the particular dimensional scaling $\frac{1}{N}$ in (6.3) will be discussed in Section 6.2. We only want to note that if we were to define a Hamiltonian $\mathbf{H}_N$ with the canonical scaling used for spin glasses

$$\mu_{\mathbf{H}_N}(x) = N\mu\big(\tfrac{\|x\|^2}{2}\big) \quad \text{and} \quad \mathcal{C}_{\mathbf{H}_N}(x,y) = N\kappa\big(\tfrac{\|x\|^2}{2}, \tfrac{\|y\|^2}{2}, \langle x, y\rangle\big), \quad x, y \in \mathbb{R}^N,$$

then $\mathbf{f}_N := \frac{\mathbf{H}_N}{N}$ would have the same distribution as in our definition. More details on the relation to spin glasses are given in Section 6.2.1.

(Non-stationary) isotropy as a distributional assumption for cost functions was introduced in Chapter 4 as a generalization of the common stationary isotropy assumption.[13] The axiomatic approach to isotropy in Definition 4.2.1 explains our characterization of (non-stationary) isotropy as a rotation and reflection invariant distribution resulting in exchangeable directions. Non-stationary isotropy is further motivated in Section 7.4 with the fact that simple linear models already break the stationary isotropy assumptions.

[13] e.g. Yann N Dauphin et al., 2014.

Before getting to the main results, let us recall one more concept. For a given mean function μ and covariance kernel κ we are interested in a sequence of (non-stationary) isotropic random functions $\mathbf{f}_N \sim \mathcal{N}(\mu, \kappa)$ over the dimension N. But it is not clear that a covariance function κ (resp. C) corresponds to a random function for every dimension. If it does we speak of validity in all dimensions. Since one can always restrict the domain of a random function $\mathbf{f}_N$ to a lower dimensional subspace it should be clear that the only possible type of restriction is an upper bound on the dimension (cf. Remark 4.4.2). Further recall, that the isotropic kernels valid in all dimensions correspond to the isotropic kernels defined on ℓ^2 (Lemma 4.4.3), which are characterized in Theorem 4.4.1. Further recall that the stationary isotropic covariance functions which are valid in all dimensions are given exactly by the kernels of the form[14]

$$C(r) = \int_{[0,\infty)} \exp(-t^2 r)\nu(dt), \qquad r \geq 0,$$

[14] e.g. Sasvári (2013, Theorem 3.8.5) or Schoenberg (1938), also listed in Table 4.1

for probability measure ν on $[0, \infty)$, which we refer to as the Schoenberg measure. An interesting property of the stationary isotropic covariance kernels valid in all dimensions is, is that they are all in C^∞ and the realization of $\mathbf{f}$ thereby smooth (Section 2.B).

6.1.3 Main result: predictable progress in high dimensions

Our main result proves that the optimization path $\mathbf{f}_N(X_0), \mathbf{f}_N(X_1), \ldots$ is asymptotically deterministic for large dimension N. In addition, we prove a similar statement about the gradient norms $\|\nabla \mathbf{f}_N(X_n)\|^2$ and more generally about their scalar products. As we assume continuity of the prefactors h_n for the gradient span algorithm $\mathfrak{G}$, they are also asymptotically deterministic due to continuous mapping. To reduce the initial complexity, we first state the simplified corollary for the stationary isotropic case. In this case some of the technical assumptions can be removed which should allow the reader to focus on the core message.

Corollary 6.1.7 (Predictable progress of gradient span algorithms, isotropic case)**.** *Let C be a stationary isotropic kernel valid in all dimensions (Lemma 4.4.3) and $\mathbf{f}_N \sim \mathcal{N}(\mu, C)$ be a sequence of scaled isotropic Gaussian random functions (Definition 6.1.3) in N. Let $\mathfrak{G}$ be a general gradient span algorithm (Definition 6.1.1),*

where we assume that its prefactors $h_n = h_n(I_{n-1})$ are continuous in the information I_{n-1}, and utilize the most recent gradients, i.e. $h^{(g)}_{n,n-1} \neq 0$ for all $n \in \mathbb{N}$. Let the algorithm $\mathfrak{G}$ be applied to $\mathbf{f}_N$ with independent, possibly random starting point $X_0 \in \mathbb{R}^N$ and denote by X_n the points resulting from $\mathfrak{G}$. Finally, assume that $\|X_0\| = \lambda$ almost surely. Then there exist characteristic real numbers

$$\mathfrak{f}_n = \mathfrak{f}_n(\mathfrak{G}, \mu, \kappa, \lambda) \in \mathbb{R} \quad \text{and} \quad \mathfrak{g}_{n,k} = \mathfrak{g}_{n,k}(\mathfrak{G}, \mu, \kappa, \lambda) \in \mathbb{R},$$

such that, for all $\epsilon > 0$, $n, k \in \mathbb{N}$,

$$\lim_{N\to\infty} \mathbb{P}\big(|\mathbf{f}_N(X_n) - \mathfrak{f}_n| > \epsilon\big) = 0 \quad \text{and}$$
$$\lim_{N\to\infty} \mathbb{P}\Big(\big|\langle \nabla \mathbf{f}_N(X_n), \nabla \mathbf{f}_N(X_k)\rangle - \mathfrak{g}_{n,k}\big| > \epsilon\Big) = 0.$$

If furthermore the algorithm is x_0-agnostic (e.g. gradient descent), then the limiting values do not depend on λ and the starting point X_0 can have arbitrary distribution independent of $\mathbf{f}_N$.

Since most algorithms are x_0-agnostic, this result explains the approximately deterministic behavior in high dimension for stationary isotropic GRFs. The general case is covered by the following remark.

Remark 6.1.8 (Initialization on the Sphere). Initialization procedures used in machine learning like Glorot initialization[15] select the entries of X_0 independent, essentially identically distributed, and scaled in such a way that the norm does not diverge. The norm therefore obeys a law of large numbers and is thus plausibly deterministic in high dimension. The assumption $\|X_0\| = \lambda$ almost surely is therefore realistic. This explains the phenomenon of (approximately) predictable progress of optimizers started in a randomly selected initial point using Glorot initialization, as observed in Figure 6.1.

[15] Glorot and Bengio, 2010.

The following proof of Corollary 6.1.7 should also serve as a reading aid for Theorem 6.1.10, which states the same result in greater generality. The proof shows how the general statements of Theorem 6.1.10 simplify to the more digestible statement of stationary isotropic GRFs.

Proof. Observe that $\mathbf{f}_N(X_k)$ and $\langle \nabla \mathbf{f}_N(X_k), \nabla \mathbf{f}_N(X_l)\rangle$ for $k, l \leq n$ are members of the random information vector $\mathbf{I}_n$ defined in Theorem 6.1.10. Their convergence therefore follows immediately from the convergence of the information vector $\mathbf{I}_n$ proven in Theorem 6.1.10. Since all stationary isotropic random functions are (non-stationary) isotropic (cf. Section 6.1.2), this Corollary therefore follows immediately from Theorem 6.1.10 once we verified the additional required assumptions.

The first assumption is the smoothness assumption on the covariance function (Assumption 6.1.9). It follows from the fact that *all* stationary isotropic random function, which are valid in all dimensions, are infinitely differentiable by the Schoenberg characterization (cf. Table 4.1 or (4.14) and Section 2.B). The second assumption is the strict positive definiteness of $(\mathbf{f}_N, \nabla \mathbf{f}_N)$, which also holds for all stationary isotropic random functions valid in all dimensions by Corollary 4.6.3 below. Note that Corollary 4.6.3 requires that the random function $\mathbf{f}_N$ is not almost surely constant. But if $\mathbf{f}_N$ were almost surely constant, then we get asymptotically deterministic behavior from the fact that $\mathbf{f}_N(x_0) \sim \mathcal{N}(\mu, \frac{1}{N}C(0))$, i.e. $\mathbf{f}_N(x_0) \to \mu$. Since $\mathbf{f}_N$ is almost surely constant, we thus obtain $\mathbf{f}_N \to \mu$ uniformly in probability. The x_0-agnostic case follows from the stationary case of Proposition 6.3.4. □

We now come to the main theorem of the paper, predictable progress of gradient span algorithms on (non-stationary) isotropic random functions in high dimensions. While kernels of stationary isotropic GRFs that are valid in all dimensions are always smooth, and the mean function is always constant, the same may not be the case for the more general situation of (non-stationary) isotropy. We will therefore assume the following smoothness properties:

Assumption 6.1.9 (Sufficiently smooth)**.** Let $\kappa_i(\lambda_1, \lambda_2, \lambda_3) := \frac{d}{d\lambda_i}\kappa(\lambda_1, \lambda_2, \lambda_3)$, we assume the partial derivatives

$$\kappa_{12}, \quad \kappa_{13}, \quad \kappa_{23} \quad \text{and} \quad \kappa_{33} \tag{6.4}$$

of the kernel κ exist and are continuous. Furthermore, we assume the derivative of the mean μ exists and is continuous.

In Section 2.B we have seen that the covariance of derivatives is directly related to the derivatives of the covariance function. For $\nabla \mathbf{f}_N$ to exist, it is therefore natural that $\mathcal{C}_{\mathbf{f}_N}$ has to be differentiable. The covariance has to be two times differentiable in a sense, as we require the following to be well defined

$$\operatorname{Cov}(\partial_i \mathbf{f}_N(x), \partial_j \mathbf{f}_N(y)) = \partial_{x_i}\partial_{x_j}\mathcal{C}_{\mathbf{f}_N}(x, y).$$

In the case of (non-stationary) isotropic functions, this requires the existence of (6.4) by Lemma 4.5.2. And the existence of the terms in (6.4) is in fact necessary and sufficient for $\nabla \mathbf{f}_N$ to exist in an L^2-sense (see Section 2.B for further details).

Here is our main result:

Theorem 6.1.10 (Predictable progress of gradient span algorithms, general case)**.** *Let κ be a kernel valid in all dimensions (Lemma 4.4.3) and $\mathbf{f}_N \sim \mathcal{N}(\mu, \kappa)$ be a sequence of scaled (non-stationary) isotropic Gaussian random functions (Definition 6.1.3) in N. Assume that μ and κ are sufficiently smooth (Assumption 6.1.9) and that the covariance of $(\mathbf{f}_N, \nabla \mathbf{f}_N)$ is strictly positive definite (Definition 4.6.1). Let $\mathfrak{G}$ be a general gradient span algorithm (Definition 6.1.1), where we assume that its prefactors $h_n = h_n(I_{n-1})$ are continuous in the information I_{n-1}, and utilize the most recent gradients, i.e. $h^{(g)}_{n,n-1} \neq 0$ for all $n \in \mathbb{N}$. Let the algorithm $\mathfrak{G}$ be applied to $\mathbf{f}_N$ with independent, possibly random starting point $X_0 \in \mathbb{R}^N$ and denote by X_n the points resulting from $\mathfrak{G}$. Finally, assume that $\|X_0\| = \lambda$ almost surely. Then, for any $n \in \mathbb{N}$, the random information vector*

$$\begin{aligned} \mathbf{I}_n &:= \Big(\mathbf{f}_N(X_k) : k \leq n\Big) \cup \Big(\langle v, w\rangle : v, w \in (x_0) \cup G_n\Big) \qquad \text{with} \\ G_n &:= \big(\nabla \mathbf{f}_N(X_k) : k \leq n\big) \end{aligned}$$

converges, as the dimension increases ($N \to \infty$), in probability against a deterministic information vector $\mathscr{I}_n = \mathscr{I}_n(\mathfrak{G}, \mu, \kappa, \lambda)$.

An application of the stochastic triangle inequality yields the following corollary, which perfectly describes Figure 6.1.

Corollary 6.1.11 (Asymptotically identical progress over multiple initializations)**.** *Assume the setting of Theorem 6.1.10 and let $X_0^{(1)}$, $X_0^{(2)}$ be random initialization points selected independently from $\mathbf{f}_N$. In the non-stationary case, we additionally assume that we we have almost surely $\|X_0^{(1)}\| = \|X_0^{(2)}\| = \lambda \in \mathbb{R}$. Let the sequences $(X_n^{(i)})_{n\in\mathbb{N}}$ be generated from the initialization $X_0^{(i)}$ by the same general gradient span algorithm $\mathfrak{G}$. Then we have, for all $n \in \mathbb{N}$ and all $\epsilon > 0$,*

$$\lim_{N\to\infty} \mathbb{P}\Big(\max_{k\leq n}\Big|\mathbf{f}_N(X_k^{(1)}) - \mathbf{f}_N(X_k^{(2)})\Big| > \epsilon\Big) = 0.$$

Proof. The proof is essentially an application of the stochastic triangle inequality

$$\Big\{|X - Z| > \epsilon\Big\} \subseteq \Big\{|X - Y| > \frac{\epsilon}{2}\Big\} \cup \Big\{|Y - Z| > \frac{\epsilon}{2}\Big\},$$

which yields by Theorem 6.1.10

$$\begin{aligned} &\lim_{N\to\infty} \mathbb{P}\Big(\max_{k\leq n}\Big|\mathbf{f}_N(X_k^{(1)}) - \mathbf{f}_N(X_k^{(2)})\Big| > \epsilon\Big) \\ &\leq \lim_{N\to\infty} \sum_{k\leq n}\Big[\mathbb{P}\Big(\Big|\mathbf{f}_N(X_k^{(1)}) - \mathfrak{f}_k\Big| > \frac{\epsilon}{2}\Big) + \mathbb{P}\Big(\Big|\mathbf{f}_N(X_k^{(2)}) - \mathfrak{f}_k\Big| > \frac{\epsilon}{2}\Big)\Big] \\ &= 0. \end{aligned}$$

□

Similar to C. Paquette et al. (2022), we obtain so-called asymptotically deterministic halting times as a corollary. The ϵ-halting is defined as

$$T_\epsilon := \inf\{n > 0 : \|\nabla \mathbf{f}_N(X_n)\|^2 \leq \epsilon\}.$$

We are going to prove it is asymptotically equal to the asymptotic ϵ-halting time

$$\tau_\epsilon := \inf\{n > 0 : \mathfrak{g}_n \leq \epsilon\} \in \mathbb{N},$$

where $\mathfrak{g}_n := \mathfrak{g}_{n,n}$ is the stochastic limit of $\|\nabla \mathbf{f}_N(X_n)\|^2$ in the dimension. In practical terms this means that optimization always stops at roughly the same time in high dimension.

Corollary 6.1.12 (Asymptotically deterministic halting times)**.** *If $\epsilon \notin (\mathfrak{g}_n)_{n\in\mathbb{N}}$, then the halting time is asymptotically deterministic*

$$\lim_{N\to\infty} \mathbb{P}(T_\epsilon = \tau_\epsilon) = 1.$$

If $\epsilon = \mathfrak{g}_n$ for some n, then

$$\lim_{N\to\infty} \mathbb{P}(T_\epsilon \in [\tau_\epsilon, \tau_\epsilon^+]) = 1. \quad \textit{with} \quad \tau_\epsilon^+ := \inf\{n > 0 : \mathfrak{g}_n < \epsilon\}.$$

Proof. The proof is identical to the proof of Theorem 4 of C. Paquette et al. (2022). □

In the random quadratic setting random matrix theory already allowed for the derivation of bounds on the limiting values $\mathfrak{f}_n$ and $\mathfrak{g}_n$. This in turn provides bounds on the asymptotic halting times in the random quadratic setting. In our much more general setting we cannot yet give such explicit convergence guarantees but want to emphasize again that all quantities appearing in our approach are in principle computable up to finite steps n.

6.2 A discussion of the dimensional scaling

Readers unfamiliar with the $\frac{1}{N}$ scaling of the covariance in (6.3) might hypothesize that this reduction of the variance simply collapses the random function $\mathbf{f}_N$ to the mean μ in the asymptotic limit. Since asymptotically deterministic behavior would then follow trivially, it is important to understand why this is not the case. The reader familiar with spin glasses will likely be satisfied by the explanation that we essentially consider $\mathbf{f}_N = \frac{H_N}{N}$ for a Hamiltonian H_N whose variance scales with N. Therefore the variance of $\mathbf{f}_N$ should scale with $\frac{1}{N}$. In the following we thus assume the reader to be unfamiliar with spin glasses.

To simplify the argument consider a stationary isotropic GRF, i.e. $\mathbf{f}_N \sim \mathcal{N}(\mu, C)$, the arguments work similarly for any (non-stationary) isotropic GRF. For any *fixed* parameter x we have $\mathbf{f}_N(x) \sim \mathcal{N}(\mu, \frac{1}{N}C(0))$, so the function value $\mathbf{f}_N(x)$ indeed collapses exponentially fast to the mean by a standard Chernoff-bound

$$\mathbb{P}\big(|\mathbf{f}_N(x) - \mu| \geq t\big) \leq 2\exp\big(-N\tfrac{t}{2C(0)}\big). \tag{6.5}$$

While the function value $\mathbf{f}_N(x)$ collapses to the mean, we will proceed to show that the gradient $\nabla \mathbf{f}_N(x)$ does not. And if the gradient does not collapse, it is sufficient to follow the gradient to find points where $\mathbf{f}_N$ stays away from μ. So the function $\mathbf{f}_N$ does not *uniformly* collapse to the mean.

In Section 2.B we explained that the covariance of derivatives of a random function is of the form

$$\mathrm{Cov}(\partial_{x_i}\mathbf{f}_N(x), \partial_{y_j}\mathbf{f}_N(y)) = \partial_{x_i}\partial_{y_j}\mathcal{C}_{\mathbf{f}_N}(x, y). \tag{6.6}$$

In the case of a stationary isotropic covariance $\mathcal{C}_{\mathbf{f}_N}(x, y) = \frac{1}{N}C\big(\frac{\|x-y\|^2}{2}\big)$ this implies

$$\partial_{x_i}\partial_{y_j}\mathcal{C}_{\mathbf{f}_N}(x, y) = \tfrac{1}{N}\Big[\underbrace{C''\big(\tfrac{\|x-y\|^2}{2}\big)(x_i - y_i)(y_j - x_j)}_{\text{(I)}} - \underbrace{C'\big(\tfrac{\|x-y\|^2}{2}\big)\delta_{ij}}_{\text{(II)}}\Big],$$

where δ_{ij} is the Kronecker delta. Part (I) is zero if $x = y$. The Kronecker delta in part (II) then implies that the entries of $\nabla \mathbf{f}(x)$ are uncorrelated and therefore independent by the Gaussian assumption. Specifically, we have

$$\partial_{x_i} \mathbf{f}_N(x) \overset{\text{iid}}{\sim} \mathcal{N}\big(0, -\tfrac{1}{N} C'(0)\big).$$

The gradient norm therefore experiences a law of large numbers

$$\|\nabla \mathbf{f}_N(x)\|^2 = \sum_{i=1}^{N} (\partial_{x_i} \mathbf{f}_N(x))^2 \underset{N\to\infty}{\overset{p}{\to}} -C'(0). \tag{6.7}$$

Thus, while the function value $\mathbf{f}_N(x)$ collapses to the mean exponentially fast (6.5), the slope of the function (counterintuitively) does not. Any other type of scaling would result in vanishing or exploding gradients. Since the supremum over gradient norms is exactly the Lipschitz constant of $\mathbf{f}_N$, this scaling is therefore necessary to stay in a class of Lipschitz functions, as is often assumed in optimization theory. And the Parisi formula (6.9) shows in the case of spin glasses, that this scaling also stabilizes the global maximum.

Remark 6.2.1 (Isoperimetry). In view of the counterintuitive observations above, one might ask whether the assumption of isotropy results in very peculiar Lipschitz functions. To understand why that is not the case, consider the concept of isoperimetry.[16] A random variable X on $\mathbb{R}^N$ is said to satisfy c-isoperimetry, if for any L-Lipschitz function $f : \mathbb{R}^N \to \mathbb{R}$ and any $t \geq 0$ an exponential concentration bound holds

$$\mathbb{P}\Big(|f(X) - \mathbb{E}[f(X)]| \geq t\Big) \leq 2\exp\big(-N\tfrac{t^2}{2cL}\big).$$

[16] e.g. Bubeck and Sellke, 2021.

Observe that this concentration bound is the mirror image of (6.5). While we consider random functions, isoperimetry is concerned with random input X to deterministic L-Lipschitz functions. In particular Gaussian or uniform input satisfies isoperimetry.[17] Intuitively, this paints the following picture of very high dimensional Lipschitz functions: Most of the function is equal to the mean except for a few peculiar points. In the case of a deterministic functions this requires the exclusion of specific deterministic points. In the case of random functions this implies any deterministic point is allowed but not the use gradient information. In either case the point we pick must be 'independent' of the function. And we show in Theorem 6.1.10 that if both function and input is random, independence is essentially a sufficient condition.

[17] Bubeck and Sellke, 2021.

The remark above shows that any Lipschitz function in high dimension appears to 'collapse to the mean'.

Our definition of scaled sequences of random functions can be also be reframed in terms of a single random function defined on ℓ^2, which is externally scaled.

Remark 6.2.2 (External scaling). One could define a a (non-stationary) isotropic GRF $\mathbf{f}$ with covariance

$$\mathcal{C}_{\mathbf{f}}(x, y) = \kappa\big(\tfrac{\|x\|^2}{2}, \tfrac{\|y\|^2}{2}, \langle x, y\rangle\big), \qquad x, y \in \ell^2$$

on the hilbertspace of sequences ℓ^2 which contains the eventually-zero sequences ℓ_0^2. These can also be viewed as the union of $\mathbb{R}^N$

$$\ell_0^2 := \bigcup_{N\in\mathbb{N}} \mathbb{R}^N \quad \text{with} \quad \mathbb{R}^N := \{(x_i)_{i\in\mathbb{N}} \in \ell^2 : x_i = 0 \quad \forall i > N\}.$$

Since the scaling $\frac{1}{\sqrt{N}}$ would scale away any mean the function $\mathbf{f}$ might possess, we assume $\mathbf{f}$ to be *centered* and define

$$\mathbf{f}_N(x) := \mu\big(\tfrac{\|x\|^2}{2}\big) + \tfrac{1}{\sqrt{N}} \mathbf{f}(x), \quad x \in \mathbb{R}^N.$$

The distribution of $\mathbf{f}_N$ is then equivalent to the one in Definition 6.1.3.

6.2.1 Relation to spin glasses

In our asymptotic viewpoint of scaled (non-stationary) isotropic GRFs is very close in spirit to the dimensional scaling of spin glasses. Spin glasses are modelled with a 'Hamiltonian' defined on the domain $\{-1,+1\}^N$ which represents {spin up, spin down}-configurations. This domain is embedded in the sphere $\mathbb{S}^{N-1}(\sqrt{N})$ of radius $\sqrt{N}$. It turned out that some mathematical question were easier to answer in this extended domain of spherical spin glasses. The notation further simplifies if spin glasses are defined on the unit sphere[18] and the domain can also be extended to $\mathbb{R}^N$. Specifically, define the pure p-spin Hamiltonian to be

$$\mathbf{H}_{p,N}(x) := \sqrt{N} \sum_{i_1,\dots,i_p=0}^{N-1} Z_{i_1\dots i_p} x^{(i_1)} \cdots x^{(i_p)}$$

for $x = (x^{(0)},\dots,x^{(N-1)}) \in \mathbb{R}^N$, where $(Z_{i_1\dots i_p})_{i_1\dots i_p}$ are identically distributed standard normal random variables. The covariance of a p-spin Hamiltonian $\mathbf{H}_{p,N}$ is then straightforward to calculate and given by

$$\mathcal{C}_{\mathbf{H}_{p,N}}(x,y) = \mathrm{Cov}(\mathbf{H}_{p,N}(x), \mathbf{H}_{p,N}(y)) = N\langle x,y\rangle^p.$$

The general mixed spin Hamiltonian $\mathbf{H}_N := \sum_{p=0}^{\infty} c_p \mathbf{H}_{p,N}$ then has the covariance

$$\mathcal{C}_{\mathbf{H}_N}(x,y) = N\xi\big(\langle x,y\rangle\big), \quad \text{with} \quad \xi(s) = \sum_{p=0}^{\infty} c_p^2 s^p. \tag{6.8}$$

While the covariance of the Hamiltonian $\mathbf{H}_N$ scales with N, the object under consideration is typically $\mathbf{f}_N = \frac{\mathbf{H}_N}{N}$. The covariance of $\mathbf{f}_N$ therefore scales with $\frac{1}{N}$, which is the scaling we chose in Definition 6.1.3. Setting

$$\mu \equiv 0 \quad \text{and} \quad \kappa(\lambda_1,\lambda_2,\lambda_3) := \xi(\lambda_3)$$

reveals that spin glasses are a subset of (non-stationary) isotropic Gaussian random functions. While the definition of spin glasses on $\mathbb{R}^N$ causes no issues, they are typically restricted to the sphere. Since the first two inputs to κ are the lengths of $x \in \mathbb{S}^{N-1}$, these are constant while the third input, the scalar product, can be reconstructed from the distance. The (non-stationary) isotropic and stationary isotropic functions therefore coincide on the sphere. By Schoenberg (1942) it follows that spin glasses actually exhaust all possible isotropic covariance functions on the sphere $\mathbb{S}^{N-1}$ which are valid in all dimensions N (cf. Remark 4.4.2).

Being the end goal of optimization, the analysis of the global optimum is complementary to our analysis of the optimization path. The Parisi formula[19] shows that the global optimum is also asymptotically deterministic

$$\lim_{N\to\infty} \max_{x\in\mathbb{S}^{N-1}} \mathbf{f}_N(x) = \lim_{N\to\infty} \mathbb{E}\Big[\max_{x\in\mathbb{S}^{N-1}} \mathbf{f}_N(x)\Big] = \inf_{\gamma} \mathcal{P}(\gamma), \tag{6.9}$$

where $\mathcal{P}$ is the Parisi functional. The first equation in (6.9) follows directly from the Borel-TIS inequality.[20] A formal proof of the Parisi formula can be found in the book by Panchenko (2013). In subsequent papers the proof was modified to be more constructive.[21] This allowed for the identification of an achievable algorithmic barrier,[22] which no Lipschitz continuous algorithms can pass[23]

$$\mathrm{ALG} = \int_0^1 \sqrt{\xi''(s)}ds.$$

Recall that ξ is the function from (6.8) and note that the algorithmic barrier assumes a restriction to the sphere. A recent review can be found in Auffinger et al. (2023), see also the lecture notes of Sellke (2024b).

So far, the algorithms used to show this barrier to be achievable have not been practical for optimization. But as we approach the topic from the viewpoint of

[18] The definition on the unit sphere removes the need for the definition of the 'overlap', a scaled scalar product which effectively turns the sphere of radius $\sqrt{N}$ into a unit sphere.

[19] Parisi, 1980.

[20] e.g. Adler and Taylor, 2007, Theorem 2.1.1.

[21] Subag, 2018; Huang and Sellke, 2024.

[22] Subag, 2021; Sellke, 2024a.

[23] Huang and Sellke, 2022.

machine learning with plausible first order optimizers, it might be possible to obtain sufficiently tight bounds on the limiting function values $\mathfrak{f}_n$ in order to prove convergence towards this algorithmic barrier. It is perhaps even possible to identify convergence rates and optimize these rates over the chosen optimizer.

Finally, building a bridge from the spin glass viewpoint towards machine learning, Choromanska et al. (2015) show, under strong simplifying assumptions, that the loss function of p-layer neural networks can be related to p-spin Hamiltonians. While they restricted the parameters of the networks to the sphere, it is likely that much of the spin glass theory could be extended from the sphere to (non-stationary) isotropic random functions on $\mathbb{R}^N$.

6.3 Proof of Theorem 6.1.10

We now present the proofs of our main result. First, we provide a sketch to motivate the overall picture (Section 6.3.1). After we explain how the covariance of derivatives are determined (Section 4.5), we reforge Theorem 6.1.10 into an even more general Theorem 6.3.2. This theorem is more natural to prove but requires the definition of a special orthonormal coordinate system (Definition 6.3.1). Finally, we give the main proof of Theorem 6.3.2. Proofs for some instrumental results are deferred to the appendix.

6.3.1 Sketch of the proof of Theorem 6.1.10

To better guide the reader, we first sketch the idea behind the proof. For simplicity, we will assume the random function to be centered and stationary isotropic $\mathbf{f}_N \sim \mathcal{N}(0, C)$.

Idea 1: Induction over Gaussian Conditionals

We interpret the objects of interest $\mathbf{g} = (\mathbf{f}_N, \nabla \mathbf{f}_N)$ (the heights joined with the gradients) as an auxiliary discrete-time stochastic process $(\mathbf{g}(X_n))_{n\in\mathbb{N}_0}$ and prove the claimed convergence in N by induction over the steps n. Please pretend for now that the inputs X_n were deterministic x_n. We will address this problem in Idea 3 of the sketch. The process $(\mathbf{g}(x_n))_{n\in\mathbb{N}_0}$ is then a Gaussian process and

$$\mathbf{g}(x_{[0:n]}) := (\mathbf{g}(x_0), \dots, \mathbf{g}(x_n))$$

is therefore a multivariate Gaussian vector. This might be a good time to recall our notation for discrete ranges

$$[n\!:\!m] := [n, m] \cap \mathbb{Z}, \qquad [n\!:\!m) := [n, m) \cap \mathbb{Z}, \qquad \text{etc.} \tag{6.10}$$

as we will make use of them throughout the following sections. It is well known that the conditional distribution $\mathbf{g}(x_n) \mid \mathbf{g}(x_{[0:n)})$ is then also normal distributed (cf. Theorem 2.A.1). By the induction hypothesis $\mathbf{g}(x_{[0:n)})$ is already converging in probability to something deterministic, so it is perhaps natural to decompose $\mathbf{g}(x_n)$ into

$$\mathbf{g}(x_n) = \mathbb{E}[\mathbf{g}(x_n) \mid \mathbf{g}(x_{[0:n)})] + \sqrt{\operatorname{Cov}[\mathbf{g}(x_n) \mid \mathbf{g}(x_{[0:n)})]} \begin{pmatrix} Y_0 \\ \vdots \\ Y_N \end{pmatrix} \tag{6.11}$$

for independent standard normal distributed Y_i independent of $\mathbf{g}(x_{[0:n)})$. The conditional expectation is then given (cf. Theorem 2.A.1) by

$$\mathbb{E}[\mathbf{g}(x_n) \mid \mathbf{g}(x_{[0:n)})] = \operatorname{Cov}(\mathbf{g}(x_{[0:n)}), \mathbf{g}(x_n))^T [\operatorname{Cov}(\mathbf{g}(x_{[0:n)}))]^{-1} \mathbf{g}(x_{[0:n)})$$

and conditional variance by

$$\begin{aligned}&\operatorname{Cov}[\mathbf{g}(x_n) \mid \mathbf{g}(x_{[0:n)})] \\ &= \operatorname{Cov}[\mathbf{g}(x_n)] - \operatorname{Cov}(\mathbf{g}(x_{[0:n)}), \mathbf{g}(x_n))^T [\operatorname{Cov}(\mathbf{g}(x_{[0:n)}))]^{-1} \operatorname{Cov}(\mathbf{g}(x_{[0:n)}), \mathbf{g}(x_n)).\end{aligned}$$

The vector $\mathbf{g}(x_{[0:n)})$ already converges by induction as the dimension N increases. What is therefore left to study are the covariance matrices.

Idea 2: Finding sparsity in the covariance matrices with a custom coordinate system

The covariance matrices which make up the conditional expectation and variance are increasing in size as the dimension N grows, because the gradient $\nabla \mathbf{f}_N$ contained in $\mathbf{g}$ increases in size. Additionally we somehow need to group the independent Y_i into something which converges. It turns out that the standard coordinate system is ill suited to achieve these goals and get sufficiently sparse (and therefore manageable) matrices. For this reason we will now phrase everything in terms of directional derivatives.

In Lemma 4.5.1 we show how the covariances of the directional derivatives of an isotropic $\mathbf{f}_N \sim \mathcal{N}(0, C)$ can be calculated explicitly. In particular we have for vectors $v, w \in \mathbb{R}^N$ and points $x, y \in \mathbb{R}^N$ with distance $\Delta = x - y$

$$\operatorname{Cov}(D_v \mathbf{f}_N(x), D_w \mathbf{f}_N(y)) = -\tfrac{1}{N}\Big[\underbrace{C''\big(\tfrac{\|\Delta\|^2}{2}\big)\langle \Delta, w\rangle\langle \Delta, v\rangle}_{\text{(I)}} + \underbrace{C'\big(\tfrac{\|\Delta\|^2}{2}\big)\langle w, v\rangle}_{\text{(II)}}\Big]. \tag{6.12}$$

We are now going to first explain how an orthogonal basis keeps (II) sparse before we explain where the use of the standard basis breaks down for (I). For simplicity, let us forget the height $\mathbf{f}_N$ for now and pretend $\mathbf{g}$ consists only of the gradient, i.e. $\mathbf{g} = \nabla \mathbf{f}_N$, and first consider the covariance matrix at a fixed point $\nabla \mathbf{f}_N(x_0)$ only. Since we only consider the derivatives at a single point we always have $x = y = x_0$ so the distance Δ vanishes. This completely removes the first part (I) (which will later require a special basis). The second part (II) is zero if v and w are orthogonal. Assuming we use an orthonormal coordinate system (e.g. the standard basis), we therefore get

$$\operatorname{Cov}[\mathbf{g}(x_0)] = \operatorname{Cov}[\nabla \mathbf{f}_N(x_0)] = \frac{1}{N}\begin{pmatrix} -C'(0) & & \\ & \ddots & \\ & & -C'(0) \end{pmatrix} = \tfrac{-C'(0)}{N}\mathbb{I}.$$

As $\mathbf{g}(x_{[0:0)})$ is the empty set we therefore have in the first step of the induction in n (cf. Equation (6.11))

$$\begin{aligned} \mathbf{g}(x_0) &= \mathbb{E}[\mathbf{g}(x_0) \mid \mathbf{g}(x_{[0:0)})] + \sqrt{\operatorname{Cov}[\mathbf{g}(x_0) \mid \mathbf{g}(x_{[0:0)})]}\begin{pmatrix} Y_0 \\ \vdots \\ Y_N \end{pmatrix} \\ &= \underbrace{\mathbb{E}[\mathbf{g}(x_0)]}_{=0} + \sqrt{\operatorname{Cov}[\mathbf{g}(x_0)]}\begin{pmatrix} Y_0 \\ \vdots \\ Y_N \end{pmatrix} \\ &= \sqrt{\tfrac{-C'(0)}{N}}\begin{pmatrix} Y_0 \\ \vdots \\ Y_N \end{pmatrix}. \end{aligned} \tag{6.13}$$

In particular, we have by the law of large numbers,

$$\langle \nabla \mathbf{f}_N(x_0), \nabla \mathbf{f}_N(x_0)\rangle = \frac{-C'(0)}{N}\sum_{i=1}^{N} Y_i^2 \underset{N\to\infty}{\xrightarrow{p}} -C'(0).$$

For the induction start and (II) there is therefore no need to change the coordinate system. But for the induction step with $n > 1$ the situation in (I) becomes more delicate.

We now want to understand where the issues in (I) with the standard coordinate system come form. Recall that we also want to determine limiting values of $\langle \nabla \mathbf{f}_N(x_0), \nabla \mathbf{f}_N(x_1)\rangle$.

Problem: Since x_1 will be x_0 plus a gradient step, $\Delta = x_1 - x_0$ is therefore a multiple of the gradient. If we continue to use the standard coordinate system,

(I) causes the covariance matrix to be dense. This is because the entries of Δ are proportional to $\partial_i \mathbf{f}_N(x_0)$, which are multiples of the Y_i and therefore almost surely never zero.

Solution: We use an adapted coordinate system. For this we select $\tilde{\mathbf{v}}_0 := \nabla \mathbf{f}_N(x_0)$ and normalize this vector $\mathbf{v}_0 = \frac{\tilde{\mathbf{v}}_0}{\|\tilde{\mathbf{v}}_0\|}$. Again, we are going to pretend that this vector was deterministic and mark this fact with the notation v_0. We then extend it by $w_{[1:N)}$ to an orthonormal basis. Since the w_i are orthogonal to the gradient span, they are orthogonal to the distances $x_0 - x_1$, i.e. $\langle \Delta, w_i \rangle = 0$ and therefore (I) is zero for almost all directions (except for v_0). As the coordinate system is orthonormal, (II) is sparse again. More generally, we chose a coordinate system $v_{[0:n)}$ capturing the span of the first n gradients $\nabla \mathbf{f}_N(x_{[0:n)})$ and therefore also the span of the first n parameters and extend this coordinate system by $w_{[n:N)}$ to an orthonormal coordinate system such that we have essentially

$$\begin{aligned}\mathbf{g}(x_{[0:n)}) &= (\nabla \mathbf{f}_N(x_0), \dots, \nabla \mathbf{f}_N(x_{n-1})) \\ &= U \begin{pmatrix} D_{v_0}\mathbf{f}_N(x_0) & \cdots & D_{v_0}\mathbf{f}_N(x_{n-1}) \\ \vdots & & \vdots \\ D_{v_{n-1}}\mathbf{f}_N(x_0) & \cdots & D_{v_{n-1}}\mathbf{f}_N(x_{n-1}) \\ D_{w_n}\mathbf{f}_N(x_0) & \cdots & D_{w_n}\mathbf{f}_N(x_{n-1}) \\ \vdots & & \vdots \\ D_{w_{N-1}}\mathbf{f}_N(x_0) & \cdots & D_{w_{N-1}}\mathbf{f}_N(x_{n-1}) \end{pmatrix} U^T.\end{aligned}$$

Here U is a basis change matrix which we are going to suppress for simplicity in the following. Note that the covariance matrix is dense for the vectors $v_{[0:n)}$ because they span the subspace of the evaluation points x_k and therefore cause (I) to be non-zero. But (I) remains zero for the vectors w_i and the covariance matrix is therefore sparse there. We want to group the chaos together (i.e. group the directional derivatives of $v_{[0:n)}$), so we view $\mathbf{g}(x_{[0:n)})$ as a row-major matrix, i.e. whenever we treat $\mathbf{g}(x_{[0:n)})$ as a vector we concatenate the rows. It then turns out that the covariance matrix is of the form

$$\operatorname{Cov}[\mathbf{g}(x_{[0:n)})] = \begin{pmatrix} \boxed{\operatorname{Cov}[D_{v_{[0:n)}}\mathbf{f}_N(x_{[0:n)})]} & & & \\ & \boxed{\Sigma_n} & & \\ & & \boxed{} & \\ & & & \ddots \\ & & & \boxed{\Sigma_{N-1}} \end{pmatrix},$$

where $\Sigma_i = \operatorname{Cov}[D_{w_i}\mathbf{f}_N(x_{[0:n)})]$ are $n \times n$ matrices and $\operatorname{Cov}[D_{[0:n)}\mathbf{f}_N(x_{[0:n)})]$ is a dense $n^2 \times n^2$ block matrix. We can assume $N - 1 > n$ without loss of generality as we let $N \to \infty$. In the proof we will make the observation that all the Σ_i are in fact equal. Note that the mixed covariance matrix $\operatorname{Cov}(\mathbf{g}(x_{[0:n)}), \mathbf{g}(x_n))$ used in the conditional expectation and covariance have similar block form and $\operatorname{Cov}[\mathbf{g}(x_n)]$ is always diagonal. The key is now to understand that as N grows to infinity, the scaling $\frac{1}{N}$ keeps the $n^2 \times n^2$ chaos in check while we get a similar law of large numbers as in the first step (cf. (6.13)) by grouping the identical Σ_i to get the length of the new gradient $\nabla \mathbf{f}_N(x_n)$ projected to the subspace of $w_{[n:N)}$ which is the orthogonal space of $v_{[0:n)}$. This projection makes up the new coordinate v_n and the induction continues.

Idea 3: Treating the random input with care

At the beginning of the sketch we asked the reader to ignore the fact that the evaluation points $(X_n)_{n \in \mathbb{N}_0}$ are random themselves. Moreover, we ignored that the adapted orthogonal basis $\mathbf{v}_{[0:n)}$ was constructed from random gradients $\nabla \mathbf{f}_N(X_k)$ which implies it is random as well. Instead we denoted it by $v_{[0:n)}$ and pretended it was deterministic. In the following, we sketch how to get rid of the randomness

of X_n and the randomness of the coordinate system $\mathbf{v}_{[0:n)}$. It is perhaps most obvious for the random coordinate system, that

$$D_{\mathbf{v}_0}\mathbf{f}_N(x_0) = \langle \mathbf{v}_0, \nabla \mathbf{f}_N(x_0)\rangle = \|\nabla \mathbf{f}_N(x_0)\|^2$$

is certainly not Gaussian by construction (recall $\mathbf{v}_0 = \nabla \mathbf{f}_N(x_0)$). Similarly, the random points X_n will typically change the distribution of $\mathbf{f}_N(X_n)$ and taking (conditional) expectations and covariances becomes quite delicate. The key to solve this issue is Corollary 3.1.8 which (assuming the choice of a consistent joint conditional distribution) states that random inputs can be treated as deterministic via

$$\mathbb{E}[h(\mathbf{g}(X_n)) \mid \mathcal{F}_{n-1}] = \Big((x_{[0:n]}) \mapsto \mathbb{E}[h(\mathbf{g}(x_n)) \mid \mathbf{g}(x_{[0:n)})]\Big)(X_{[0:n]}),$$

if X_n is previsible, i.e. X_{n+1} is measurable with regard to the filtration $\mathcal{F}_n = \sigma(\mathbf{g}(X_k) : k \leq n)$. Applying this result to the evaluations points X_n is relatively straightforward. But it is not obvious how to treat the random coordinate system. The key is to turn the coordinate system into an input. The definition

$$Z(v; x) := D_v \mathbf{f}_N(x)$$

allows us to apply Corollary 3.1.8 to both coordinate system and points. The delicate part is that the input must be previsible with respect to the filtration $\mathcal{F}_n$. So now the coordinate system must also be previsible. We therefore cannot use future gradients for our coordinate system. This forces us to define a custom coordinate system for every step. That means after n steps we have n different coordinate systems. We will carefully reduce this to just one coordinate system per iteration in the formal proof. Turning this strategy into a rigorous proof is a bit tedious and requires a careful induction including a few additional technical induction hypothesis.

6.3.2 Proof of Theorem 6.1.10

After the preparations of the previous sections we can now give the details of the proof sketched in Section 6.3.1. In the sketch of the proof, we highlighted the importance of an adapted coordinate system in order to keep the covariance matrices sparse. In particular, part (I) of (6.12) requires orthogonality to the walking directions. We also needed the coordinate system to be orthogonal to keep (II) as sparse as possible. In Idea 3, we sketched why the coordinate system had to be previsible. That is, why we needed a different coordinate system in every step. We therefore begin the proof by building these coordinate systems.

Let us start by defining the natural filtration

$$\mathcal{F}_n := \sigma\big(\mathbf{f}_N(X_k), \nabla \mathbf{f}_N(X_k) : 0 \leq k \leq n\big).$$

The following previsible *vector space of evaluation points*

$$V_n := \operatorname{span}\Big(\{x_0\} \cup \{\nabla \mathbf{f}_N(X_k) : 0 \leq k < n\}\Big), \qquad d_n := \dim(V_n), \tag{6.14}$$

contains $X_{[0:n]}$ by definition of the generalized gradient span algorithm (Definition 6.1.1) and is similarly measurable with respect to $\mathcal{F}_{n-1}$, i.e. previsible. Also recall that $[0{:}n] = [0, n] \cap \mathbb{Z}$ is notation for a discrete intervals as (re-)introduced in Equation 6.10. We will now use this chain of vector spaces to define previsible coordinate systems for every step n.

Definition 6.3.1 (Previsible orthonormal coordinate systems)**.** We define an $\mathcal{F}_{n-1}$-measurable orthonormal basis $\mathbf{v}_{[0:d_n)}$ of V_n inductively such that $\mathbf{v}_{[0:d_k)}$ is a basis of the subspace V_k.

Induction start Assuming $x_0 \neq 0$ and thus $d_0 = 1$, we define

$$\mathbf{v}_0 := \frac{x_0}{\|x_0\|}.$$

In any case this results in a basis $\mathbf{v}_{[0:d_0)}$ of V_0 since $[0 : d_0) = \emptyset$ if $d_0 = 0$.

Induction step (Gram-Schmidt) Assuming we have a basis $\mathbf{v}_{[0:d_n)}$ of V_n, let us construct a basis for V_{n+1}. For this we define the following Gram-Schmidt procedure candidate

$$\tilde{\mathbf{v}}_n := \nabla \mathbf{f}_N(X_n) - P_{V_n} \nabla \mathbf{f}_N(X_n) = \nabla \mathbf{f}_N(X_n) - \sum_{i<d_n} \langle \nabla \mathbf{f}_N(X_n), \mathbf{v}_i \rangle \mathbf{v}_i, \tag{6.15}$$

where P_{V_n} is the projection to V_n. If $\tilde{\mathbf{v}}_n = 0$, then $\nabla \mathbf{f}_N(X_n) \in V_n$ and we thus have $V_{n+1} = V_n$ by definition (6.14). In that case we already have a basis for V_{n+1}.

For any n where the dimension increases such that $d_{n+1} = d_n + 1$, we define

$$\mathbf{v}_{d_{n+1}-1} = \mathbf{v}_{d_n} := \frac{\tilde{\mathbf{v}}_n}{\|\tilde{\mathbf{v}}_n\|}. \tag{6.16}$$

We thus have obtained an orthonormal basis $\mathbf{v}_{[0:d_{n+1})}$ of V_{n+1} which is $\mathcal{F}_n$ measurable.

Basis extensions With this construction of basis elements of V_n done, we $\mathcal{F}_{n-1}$-measurably select an arbitrary orthonormal basis $\mathbf{w}^{(n)}_{[d_n:N)}$ of $V_n^\perp$ to obtain an $\mathcal{F}_{n-1}$-measurable orthonormal basis

$$B_n := (\mathbf{v}_{[0:d_n)}, \mathbf{w}^{(n)}_{[d_n:N)})$$

for every $n \in \mathbb{N}$. The coordinate systems B_n are thus previsible.

Using this specialized coordinate system we state an extension of Theorem 6.1.10. This extension looks less friendly but is more natural to prove. It implies Theorem 6.1.10, proves three additional claims and relaxes the assumptions on the prefactors of the gradient span algorithm. The additional claims cannot be stated separately as they are all proved in one laborious induction. And to prove the claim we are most interested in (Ind-I) we also require the other claims in the induction step. Finally, with the help of Proposition 6.3.4, we can (and will) assume deterministic starting points in Theorem 6.3.2 without loss of generality.

Theorem 6.3.2 (Predictable progress of gradient span algorithms [Extension of Theorem 6.1.10])**.** *Let κ be a kernel valid in all dimensions (i.e. defined on ℓ^2 by Lemma 4.4.3) and $\mathbf{f}_N \sim \mathcal{N}(\mu, \kappa)$ be a sequence of scaled (non-stationary) isotropic Gaussian random functions (Definition 6.1.3) in N. Assume that μ and κ are sufficiently smooth (Assumption 6.1.9) and assume the covariance of $(\mathbf{f}_N, \nabla \mathbf{f}_N)$ is strictly positive definite. Let $\mathfrak{G}$ be a general gradient span algorithm (Definition 6.1.1), which is* asymptotically *continuous and uses the most recent gradient* asymptotically *(cf. Assumption 6.3.3).*

Let $\mathfrak{G}$ be applied to $\mathbf{f}_N$ with starting points $x_0 \in \mathbb{R}^N$ such that $X_n = \mathfrak{G}(\mathbf{f}_N, x_0, n)$. We assume that the deterministic initialization point $x_0 \in \mathbb{R}^N$ is of constant length $\|x_0\| = \lambda$ over the dimension N. Using $G_n := \big(\nabla \mathbf{f}_N(X_k) : k \le n\big)$, we further define the modified random information vector

$$\hat{\mathbf{I}}_n := \Big(\mathbf{f}_N(X_k) : k \le n\Big) \cup \Big(\langle v, w \rangle : v \in \mathbf{v}_{[0:d_{n+1})}, w \in G_n\Big).$$

The following inductive claims hold for all $n \in \mathbb{N}$, where (Ind-II) implies that the vector $\hat{\mathbf{I}}_n$ is almost surely of constant length.

(Ind-I) ***Information convergence:*** *There exists some deterministic limiting information vector $\hat{\mathscr{I}}_n = \hat{\mathscr{I}}_n(\mathfrak{G}, \mu, \kappa)$, such that*

$$\hat{\mathbf{I}}_n \xrightarrow[N\to\infty]{p} \hat{\mathscr{I}}_n.$$

The limiting information is split into $\hat{\mathscr{I}}_n = (\mathfrak{f}_k, \gamma_k^{(i)})_{k,i}$, where the limiting elements $\mathfrak{f}_k$ of $\mathbf{f}_N(X_k)$ were already defined in Theorem 6.1.10. We further define the limiting inner products of gradients by

$$\gamma_k^{(i)} := \lim_{N\to\infty} \langle \nabla \mathbf{f}_N(X_k), \mathbf{v}_i \rangle. \tag{6.17}$$

(Ind-II) ***(Asymptotic) Full rank****: If the most recent gradient is always used (and not just asymptotically), then the previsible vector space of evaluation points V_{n+1} has almost surely full rank (assuming $x_0 \neq 0$). Specifically, for all $m \leq n+1$ we have almost surely*

$$d_m = m + \mathbf{1}_{x_0 \neq 0}.$$

This always holds in the limit (even when the most recent gradient is only used asymptotically), which means that for all $k \leq n$ the Gram-Schmidt candidate $\tilde{\mathbf{v}}_k$ defined in (6.16) is not zero in the limit

$$\gamma_k^{(d_k)} = \lim_{N\to\infty} \langle \nabla \mathbf{f}_N(X_k), \mathbf{v}_{d_k} \rangle = \lim_{N\to\infty} \|\tilde{\mathbf{v}}_k\| \neq 0. \tag{6.18}$$

(Ind-III) ***Representation:*** *For all $k \leq m \leq n+1$ there exist limiting representation vectors*

$$y_k = y_k(\mathfrak{G}, \mu, \kappa) \quad \textit{with} \quad y_k = (y_k^{(0)}, \dots y^{(d_m-1)}) \in \mathbb{R}^{d_m}$$

of X_k, such that for all $i < d_m$ we have

$$\langle X_k, \mathbf{v}_i \rangle \xrightarrow[N\to\infty]{p} y_k^{(i)} \quad \textit{and} \quad \|X_k\|^2 = \sum_{i=0}^{d_m-1} \langle X_k, \mathbf{v}_i \rangle^2 \xrightarrow[N\to\infty]{p} \|y_k\|^2. \tag{6.19}$$

The asymptotic distances are then given for all $k, l \leq m$ by

$$\|X_k - X_l\|^2 = \sum_{i=0}^{d_m-1} \langle X_k - X_l, \mathbf{v}_i \rangle^2 \xrightarrow[N\to\infty]{p} \rho_{kl}^2 := \|y_k - y_l\|^2. \tag{6.20}$$

(Ind-IV) ***The evaluation points are asymptotically different****, i.e.*

$$\rho_{kl}^2 > 0 \quad \textit{for all} \quad k, l \leq n+1 \quad \textit{with} \quad k \neq l.$$

Before we prove that Theorem 6.1.10 follows from Theorem 6.3.2 let us state the 'asymptotic continuity' and 'use of the most recent gradient' assumptions on the prefactors.

Assumption 6.3.3 (Assumptions on prefactors)**.** Recall, that we defined the prefactors h_n of a GSA to be functions of the information $\mathbf{I}_{n-1}$ (Definition 6.1.1). We assume that

- h_n is continuous in the point $\mathscr{I}_{n-1}$ for all n, and
- the most recent gradients are used at least in the asymptotic limit, i.e.

$$\hat{h}_{n,n-1}^{(g)} := h_{n,n-1}^{(g)}(\mathscr{I}_{n-1}) \Big(= \lim_{N\to\infty} h_{n,n-1}^{(g)}(\mathbf{I}_{n-1}) \Big) \neq 0.$$

By Lemma 6.3.7 it is apparent, that we can equivalently assume the prefactors h_n to be functions in $\hat{\mathbf{I}}_{n-1}$ continuous in $\hat{\mathscr{I}}_{n-1}$, which use the most recent gradient in the asymptotic limit $\hat{h}_{n,n-1} = h_{n,n-1}(\hat{\mathscr{I}}_{n-1})$.

At first it might seem circular to make use of limiting elements in an assumption which is necessary to prove the limiting elements exist. But a closer inspection reveals that h_n is only used for the convergence of $\mathbf{I}_k$ to $\mathscr{I}_k$ for times $k \geq n$. We can therefore interleave this assumption on h_n with the inductive existence proof of $\mathscr{I}_{n-1}$.

The proof of Theorem 6.1.10 now follows readily from Theorem 6.3.2 via a couple of preliminary results:

Proposition 6.3.4. *We can assume without loss of generality that the independent initialization X_0 is a deterministic $x_0 \in \mathbb{R}^N$ of length $\|x_0\| = \lambda$ over all dimensions N.*

Stationary case: *Assume the random functions $\mathbf{f}_N$ are furthermore stationary isotropic and the algorithm x_0-agnostic (Definition 6.1.1). Then the limiting information is independent of λ and the random initialization X_0 does not need to satisfy $\|X_0\| = \lambda$ almost surely.*

The proof of this Proposition is accomplished via two lemmas:

Lemma 6.3.5 (The 'information' is invariant to linear isometries)**.** *Let f be some function and $(x_0, \dots, x_n)$ evaluation points. We further define a change in coordinates*

$$g(y) := f \circ \phi \quad \text{and} \quad y_k := \phi^{-1}(x_k) \quad \text{for} \quad k \in \{0, \dots, n\}$$

via a linear isometry $\phi(x) = Ux$ for some orthonormal matrix U. Then the information

$$\begin{aligned} I_n &:= \Big(f(x_k) : k \leq n\Big) \cup \Big(\langle v, w\rangle : v, w \in (x_0) \cup G_n\Big) \quad \text{with} \\ G_n &:= \big(\nabla f(x_k) : k \leq n\big) \end{aligned}$$

is exactly equal to the information

$$\begin{aligned} J_n &:= \Big(g(y_k) : k \leq n\Big) \cup \Big(\langle v, w\rangle : v, w \in (y_0) \cup \tilde{G}_n\Big) \quad \text{with} \\ \tilde{G}_n &:= \big(\nabla g(y_k) : k \leq n\big). \end{aligned}$$

If ϕ is a general isometry of the form $\phi(x) = Ux + b$, then we still have equality for the the reduced information, i.e.

$$I_n^{\setminus x_0} = J_n^{\setminus y_0}.$$

Proof. We have by definition

$$f(x_k) = f \circ \phi(\phi^{-1}(x_k)) = g(y_k).$$

Let us therefore turn to the inner products. Since $\phi(x) = Ux$ or $\phi(x) = Ux + b$, we have for the gradient

$$\nabla g(y) = U^T \nabla f(\phi(y)). \tag{6.21}$$

this implies all the inner products are equal

$$\langle \nabla g(y_k), \nabla g(y_l)\rangle = \langle U^T \nabla f(x_k), U^T \nabla f(x_l)\rangle = \langle \nabla f(x_k), \nabla f(x_l)\rangle.$$

In the linear isometry case, where $\phi(x) = Ux$, we have $y_n = U^T x_n$ and thus

$$\begin{aligned} \langle y_0, \nabla g(y_l)\rangle &= \langle U^T x_0, U^T \nabla f(x_l)\rangle &&= \langle x_0, \nabla f(x_l)\rangle \\ \langle y_0, y_0\rangle &= \langle U^T x_0, U^T x_0\rangle &&= \langle x_0, x_0\rangle. \end{aligned}$$

The information is therefore exactly the same. □

Lemma 6.3.6 (Linear isometry equivariance of general gradient span algorithms)**.** *Let $\mathfrak{G}$ be a general gradient span algorithm (Definition 6.1.1), let f be a function, x_0 a starting point. We define a change of basis*

$$g(y) := f \circ \phi \quad \text{and} \quad y_0 := \phi^{-1}(x_0).$$

via a linear isometry $\phi(x) = Ux$ with orthonormal matrix U. Then the optimization paths

$$x_n := \mathfrak{G}(f, x_0, n) \quad \text{and} \quad y_n := \mathfrak{G}(g, y_0, n)$$

are equivariant, i.e. the simple basis change

$$y_n = \phi^{-1}(x_n)$$

is retained for all $n \in \mathbb{N}$. The same holds true for all isometries ϕ, if the algorithm is x_0-agnostic.

Proof. We proceed by induction over n, where the induction start is obvious.

For the induction step $(n-1) \to n$ we use the induction claim $y_k = \phi^{-1}(x_k)$ for all $k \leq n-1$ to obtain that the information I_{n-1} is invariant (by Lemma 6.3.5). This implies by definition of the gradient span algorithm and (6.21)

$$y_n = h_n^{(x)} y_0 + \sum_{k=0}^{n-1} h_{n,k}^{(g)} \nabla g(y_k) = U^T \Big(h_n^{(x)} x_0 + \sum_{k=0}^{n-1} h_{n,k}^{(g)} \nabla f(x_k) \Big) = \phi^{-1}(x_n),$$

where the invariance of the information was used implicitly as the h_n are functions of I_{n-1}.

In the x_0-agnostic case, the induction step is almost the same except we have for isometries $\phi(x) = Ux + b$ that $y_k = \phi^{-1}(x_k) = U^T(x_k - b)$ and thus

$$y_n = y_0 + \sum_{k=0}^{n-1} h_{n,k}^{(g)} \nabla g(y_k) = U^T \Big(x_0 - b + \sum_{k=0}^{n-1} h_{n,k}^{(g)} \nabla f(x_k) \Big) = \phi^{-1}(x_n),$$

using the fact that $h_n^{(x)} = 1$. We similarly use that the reduced information $I_{n-1}^{\setminus x_0}$ is retained for the prefactors. □

We are now ready to prove that we can assume without loss of generality that the initialization is deterministic.

Proof of Proposition 6.3.4. Since we have

$$\mathbb{P}(\hat{\mathbf{I}}_n \in A) = \mathbb{E}\Big[\mathbb{P}(\hat{\mathbf{I}}_n \in A \mid X_0) \Big],$$

it is sufficient to show that $\mathbb{P}(\hat{\mathbf{I}}_n \in A \mid X_0 = x_0)$ is only dependent on $\|x_0\| = \lambda$, which is constant over the distribution of X_0.

For any x_0, y_0 with $\|x_0\| = \|y_0\| = \lambda$, there exists a linear isometry ϕ such that $x_0 = \phi^{-1}(y_0)$.

Let $\mathbf{I}_n = \mathbf{I}_n(\mathbf{f}_N, y_0)$ be the information vector generated from running the gradient span algorithm on $\mathbf{f}_N$ with starting point y_0 for n steps. Since $\mathbf{f}_N \overset{(d)}{=} \mathbf{f}_N \circ \phi$ in distribution, due to (non-stationary) isotropy (cf. Definition 4.2.1), we have

$$\mathbf{I}_n(\mathbf{f}_N, y_0) \overset{(d)}{=} \mathbf{I}_n(\mathbf{f}_N \circ \phi, y_0) = \mathbf{I}_n(\mathbf{f}_N, x_0)$$

where we have used Lemma 6.3.5 and Lemma 6.3.6 for the last equation. With the note that $\hat{\mathbf{I}}_n$ is a deterministic map of $\mathbf{I}_n$ as it only requires Gram-Schmidt orthogonalization as outlined in Definition 6.3.1 we can conclude this proof

$$\begin{aligned} \mathbb{P}\Big(\hat{\mathbf{I}}_n(\mathbf{f}_N, X_0) \in A \mid X_0 = y_0 \Big) &= \mathbb{P}\Big(\hat{\mathbf{I}}_n(\mathbf{f}_N, y_0) \in A \Big) \\ &= \mathbb{P}\Big(\hat{\mathbf{I}}_n(\mathbf{f}_N, x_0) \in A \Big) \\ &= \mathbb{P}\Big(\hat{\mathbf{I}}_n(\mathbf{f}_N, X_0) \in A \mid X_0 = x_0 \Big). \end{aligned}$$

In the stationary case with x_0-agnostic algorithm, we can make use of arbitrary isometries ϕ. In particular we can chose $\phi(x) = x - x_0$ that maps x_0 to zero. With the same arguments as above (using the x_0-agnostic version of Lemma 6.3.5 and Lemma 6.3.6) we get

$$\mathbb{P}\Big(\hat{\mathbf{I}}_n(\mathbf{f}_N, X_0) \in A \mid X_0 = x_0 \Big) = \mathbb{P}\Big(\hat{\mathbf{I}}_n(\mathbf{f}_N, X_0) \in A \mid X_0 = 0 \Big).$$

The distribution is thus completely independent of x_0. In particular, this forces the limiting values to be independent of x_0 and therefore also independent of λ. □

Lemma 6.3.7. *For any fixed x_0, $\mathbf{I}_n$ is a continuous function of $\hat{\mathbf{I}}_n$ for all $n \in \mathbb{N}$.*

Proof. Let us first recall the definition of $\mathbf{I}_n$ and $\hat{\mathbf{I}}_n$. With $G_n := \big(\nabla \mathbf{f}_N(X_k) : k \leq n\big)$ we have in a direct comparison:

$$\mathbf{I}_n := \Big(\mathbf{f}_N(X_k) : k \leq n\Big) \cup \Big(\langle v, w\rangle : v, w \in (x_0) \cup G_n\Big)$$
$$\hat{\mathbf{I}}_n := \Big(\mathbf{f}_N(X_k) : k \leq n\Big) \cup \Big(\langle v, w\rangle : v \in \mathbf{v}_{[0:d_{n+1})}, w \in G_n\Big).$$

As the identity is a continuous map, we simply map the function values $\mathbf{f}_N(X_k)$ from $\hat{\mathbf{I}}_n$ to itself in $\mathbf{I}_n$. We therefore only need to find a way to continuously construct the inner products of $\mathbf{I}_n$ from $\hat{\mathbf{I}}_n$. Since $\nabla \mathbf{f}_N(X_k)$ is contained in V_{n+1} by its definition (6.14) and $\mathbf{v}_{[0:d_{n+1})}$ is a basis of V_{n+1} by construction (Definition 6.3.1), we have for all $k, l \leq n$

$$\begin{aligned}\langle \nabla \mathbf{f}_N(X_k), \nabla \mathbf{f}_N(X_l)\rangle &= \Big\langle \sum_{i=1}^{d_{n+1}-1} \langle \nabla \mathbf{f}_N(X_k), \mathbf{v}_i\rangle \mathbf{v}_i, \sum_{j=1}^{d_{n+1}-1} \langle \nabla \mathbf{f}_N(X_l), \mathbf{v}_j\rangle \mathbf{v}_j \Big\rangle \\ &= \sum_{i=0}^{d_{n+1}-1} \underbrace{\langle \nabla \mathbf{f}_N(X_k), \mathbf{v}_i\rangle}_{\in \hat{\mathbf{I}}_n} \underbrace{\langle \nabla \mathbf{f}_N(X_l), \mathbf{v}_i\rangle}_{\in \hat{\mathbf{I}}_n} \end{aligned} \tag{6.22}$$

Since d_{n+1} is almost surely constant by (Ind-II)[24], this covers most of the inner products of $\mathbf{I}_n$. What is left are the inner products using x_0. If $x_0 = 0$, then all those inner products are zero and the zero map does the job.

In the following we therefore assume $x_0 \neq 0$. Because x_0 is deterministic, we do not need to construct $\|x_0\|^2 = \lambda^2$ from $\hat{\mathbf{I}}_n$ as a constant map does the job. Since $x_0 \neq 0$ implies $\mathbf{v}_0 = \frac{x_0}{\|x_0\|}$ by construction (Definition 6.3.1), we have for all $k \leq n$

$$\langle x_0, \nabla \mathbf{f}_N(X_k)\rangle = \|x_0\| \underbrace{\langle \mathbf{v}_0, \nabla \mathbf{f}_N(X_k)\rangle}_{\in \hat{\mathbf{I}}_n}.$$

We have thus continuously constructed all inner products in $\mathbf{I}_n$ form $\hat{\mathbf{I}}_n$. □

Here is how the main theorem of the paper follows from Theorem 6.3.2.

Proof of Theorem 6.1.10. By Proposition 6.3.4, we can assume without loss of generality that the initial point is deterministic. Since the other assumptions of Theorem 6.3.2 are the same (or even more general), we only need to prove that for every $n \in \mathbb{N}$ there exists some $\mathscr{I}_n$ such that

$$\mathbf{I}_n \underset{N\to\infty}{\xrightarrow{p}} \mathscr{I}_n.$$

As $\mathbf{I}_n$ is a continuous function of $\hat{\mathbf{I}}_n$ by Lemma 6.3.7 this follows immediately from continuous mapping by the inductive claim (Ind-I) of Theorem 6.3.2. □

The rest of the section is an inductive proof of Theorem 6.3.2.

6.3.3 Proof of Theorem 6.3.2

The heart of the proof will be a lengthy induction. Before, we want to address the additional assumptions of

1. representable limit points (Ind-III) and
2. the claim that these points are different (Ind-IV),

which we introduced in Theorem 6.3.2 but did not use in the proof of Theorem 6.1.10. Together with the strictly positive definite covariance of $(\mathbf{f}_N, \nabla \mathbf{f}_N)$ these assumptions will allow us to argue for converging entries of covariance matrices and invertible limiting covariance matrices, which are used in the conditional expectation and conditional variance.

[24] We sketch a path to get rid of the strict positive definiteness assumption in the outlook (Section 6.4). In this case (Ind-II) and (Ind-IV) are lost. This argument then has to be replaced using an upper bound on the d_{n+1} and Lemma 6.3.10.

Complexity reduction

Before we start the induction, we will perform some complexity reductions.

1. We show that (Ind-III) follows from (Ind-I) in Lemma 6.3.8.
2. We show that (Ind-IV) follows from (Ind-II) and (Ind-I) in Lemma 6.3.9.
3. We reduce the work necessary to prove (Ind-I).

In the actual induction we will therefore be able to focus on the claims (Ind-I) and (Ind-II).

Lemma 6.3.8. *For all fixed $n \in \mathbb{N}$, the convergence of information (Ind-I) implies the representation (Ind-III).*

Proof. Assuming (Ind-I) we have that $\hat{\mathbf{I}}_n \to \hat{\mathscr{I}}_n$. By Lemma 6.3.7 this also implies $\mathbf{I}_n \to \mathscr{I}_n$. Since the prefactors h_m are functions of $\mathbf{I}_{m-1}$ continuous in $\mathscr{I}_{m-1}$, this implies for all $m \leq n+1$ that the prefactors converge

$$h_m = h_m(\mathbf{I}_{m-1}) \xrightarrow[N\to\infty]{p} h_m(\mathscr{I}_{m-1}) =: \hat{h}_m.$$

Recall that by (6.17) of (Ind-I) we have for all $k \leq n$ and all $i < d_{n+1}$

$$\langle \nabla \mathbf{f}_N(X_k), \mathbf{v}_i \rangle \xrightarrow[N\to\infty]{p} \gamma_k^{(i)}$$

for some $\gamma_k^{(i)} \in \mathbb{R}$. For all $m \leq n+1$ we therefore have by definition of X_m (Definition 6.1.1)

$$\begin{aligned} \langle X_m, \mathbf{v}_i \rangle &= h_m^{(x)} \langle x_0, \mathbf{v}_i \rangle + \sum_{k=0}^{m-1} h_{m,k}^{(g)} \langle \nabla \mathbf{f}_N(X_k), \mathbf{v}_i \rangle \\ &\xrightarrow[N\to\infty]{p} y_m^{(i)} \quad := \hat{h}_m^{(x)} \|x_0\| \delta_{0i} + \sum_{k=0}^{m-1} \hat{h}_{m,k}^{(g)} \gamma_k^{(i)}, \end{aligned} \tag{6.23}$$

where δ_{ij} denotes the Kronecker delta. Since X_k is contained in the vector space of evaluation points V_m for all $k \leq m$ by definition (6.14), its norms converges

$$\|X_k\|^2 = \sum_{i=0}^{d_m-1} \langle X_k, \mathbf{v}_i \rangle^2 \xrightarrow[N\to\infty]{p} \sum_{i=0}^{d_m-1} (y_k^{(i)})^2 = \|y_k\|^2,$$

and likewise their distances for all $k, l \leq m$

$$\|X_k - X_l\|^2 = \sum_{i=0}^{d_m-1} \langle X_k - X_l, \mathbf{v}_i \rangle^2 \xrightarrow[N\to\infty]{p} \rho_{kl}^2 := \|y_k - y_l\|^2.$$

This proves the limiting representation (Ind-III). □

In the following we will prove, assuming (Ind-I) and (Ind-II), that the limiting distances ρ_{kl} are greater zero, i.e. (Ind-IV). The main ingredients are (Ind-II) and the assumed asymptotic use of the last gradient.

Lemma 6.3.9. *For all fixed $n \in \mathbb{N}$, the convergence of information (Ind-I) together with the asymptotic full rank (Ind-II) imply asymptotically different evaluation points (Ind-IV).*

Proof. We will proceed by induction over m, where we assume the claim to be shown for all $k, l \leq m \leq n+1$. The induction start $m = 0$ is trivial since a single point is always distinct. For the induction step we assume to have the statement for m, that is

$$\rho_{kl} > 0 \quad \text{for all} \quad k, l \leq m \quad \text{with} \quad k \neq l.$$

To show the statement for $m+1 \leq n+1$, we only need to check the distances to the point y_{m+1} as we have the others by induction.

To show that y_{m+1} is distinct from any y_k with $k \leq m$, we use the fact that the most recent gradient is used asymptotically by assumption of the Theorem 6.3.2 (cf. Assumption 6.3.3), i.e.

$$\hat{h}^{(g)}_{m+1,m} = h^{(g)}_{m+1,m}(\mathscr{I}_m) \neq 0. \tag{6.24}$$

The claim (Ind-II) ensures that V_{m+1} has a larger dimension than V_m, that is $d_{m+1} = d_m + 1$. The last basis element of V_{m+1} is thus given by $\mathbf{v}_{d_{m+1}-1} = \mathbf{v}_{d_m}$. By (6.15) this element is produced from the last gradient $\nabla \mathbf{f}_N(X_m)$, which cannot be used for X_k with $k \leq m$ but must be used for X_{m+1} asymptotically by (6.24). Therefore X_{m+1} asymptotically contains a component of $\mathbf{v}_{d_m}$, that no other X_k has, which translates to the inequality of the asymptotic representations y_{m+1} and y_k.

Formally, since $X_0, \dots X_m$ is contained in V_m spanned by $\mathbf{v}_{[0:d_m)}$, we have for all $k \leq m$

$$y^{(d_m)}_k \overset{(6.23)}{=} \lim_{N\to\infty} \langle X_k, \mathbf{v}_{d_m} \rangle = 0. \tag{6.25}$$

For the asymptotic representation of the last evaluation point X_{m+1} on the other hand, we have

$$y^{(d_m)}_{m+1} \overset{(6.23)}{=} \hat{h}^{(x)}_{m+1} \|x_0\| \underbrace{\delta_{0 d_m}}_{=0} + \sum_{k=0}^{m} \hat{h}^{(g)}_{m+1,k} \gamma^{(d_m)}_k = \hat{h}^{(g)}_{m+1,m} \gamma^{(d_m)}_m. \tag{6.26}$$

The last equation is due to (6.27). This follows from the fact that for all $k < m$ the gradient $\nabla \mathbf{f}_N(X_k)$ is contained in V_m spanned by $\mathbf{v}_{[0:d_m)}$. By definition of $\gamma^{(i)}_k$ (6.17) we therefore have

$$\gamma^{(d_m)}_k \overset{(6.17)}{=} \lim_{N\to\infty} \langle \mathbf{f}_N(X_k), \mathbf{v}_{d_m} \rangle = 0. \tag{6.27}$$

Putting (6.25) and (6.26) together we have

$$\rho^2_{(m+1)k} = \|y_{m+1} - y_k\|^2 \geq (y^{(d_m)}_{m+1} - y^{(d_m)}_k)^2 = (\hat{h}^{(g)}_{m+1,m} \gamma^{(d_m)}_m)^2 > 0,$$

where the last inequality is due to the asymptotic use of the most recent gradient (6.24) and the limiting full rank claim (6.18) of (Ind-II). □

In (Ind-I) we claim convergence of $\hat{\mathbf{I}}_n$, where $\hat{\mathbf{I}}_n$ contains inner products of the form

$$\langle \mathbf{v}_i, \nabla \mathbf{f}_N(X_k) \rangle$$

for $k \leq n$ and $i < d_{n+1}$. The following lemma essentially implies that the restriction on i was unnecessary. So when we increase $n-1$ to n in the induction step, we do not have to revisit the inner products of old gradients and can focus solely on $\nabla \mathbf{f}_N(X_n)$.

Lemma 6.3.10. *For all $k \in \mathbb{N}$ and $i \geq d_{k+1}$ we have*

$$\langle \nabla \mathbf{f}_N(X_k), \mathbf{v}_i \rangle \overset{p}{\underset{N\to\infty}{\to}} 0 = \gamma^{(i)}_k.$$

Proof. For all $i \geq d_{k+1}$ the basis vector $\mathbf{v}_i$ is constructed to be orthogonal to $\nabla \mathbf{f}_N(X_k)$ contained in V_{k+1} spanned by $\mathbf{v}_{[0:d_{k+1})}$. This implies

$$\lim_{N\to\infty} \langle \nabla \mathbf{f}_N(X_k), \mathbf{v}_i \rangle = 0 = \gamma^{(i)}_k$$

where the last equation is simply the definition of $\gamma^{(i)}_k$ of (6.17). □

Remark 6.3.11 (Gamma are triangular). Lemma 6.3.10 can be visualized using a triangular matrix as follows. With the definition

$$\mathbb{R}^n := \{(x_i)_{i\in[0:\infty)} : x_i = 0 \quad \forall i \geq n\},$$

we can view $\mathbb{R}^m$ as a subspace of $\mathbb{R}^n$ for $m \leq n$. The vector

$$\gamma_k := (\gamma_k^{(i)})_{i\in[0:d_{k+1})} \in \mathbb{R}^{d_{k+1}},$$

is then (by Lemma 6.3.10) also a member of $\mathbb{R}^m$ for $m \geq d_{k+1}$. Concatenating the vectors

$$\gamma_{[0:n]} = \begin{pmatrix} \gamma_0^{(0)} & \dots & \gamma_n^{(0)} \\ \vdots & & \vdots \\ \gamma_0^{(d_{n+1}-1)} & \dots & \gamma_n^{(d_{n+1}-1)} \end{pmatrix} = \begin{pmatrix} \gamma_0^{(0)} & \dots & \gamma_{n-1}^{(0)} & \gamma_n^{(0)} \\ \gamma_0^{(1)} & \dots & \gamma_{n-1}^{(1)} & \gamma_n^{(1)} \\ 0 & & & \\ \vdots & \ddots & & \vdots \\ 0 & \dots & 0 & \gamma_n^{(d_{n+1}-1)} \end{pmatrix}$$

therefore results in a upper triangular matrix above an offset diagonal, since we always have $d_k \leq k+1$ due to the definition of V_k (6.14).

Note that, for visualization purposes, we assumed $d_{n+1} = d_n + 1$ in the second representation. If the dimension stays constant at some times k, then the zeros encroach above this diagonal.

Induction start with $n = 0$

We start the induction by proving the claim for $n = 0$. We have structured the induction start similar to the induction step. We therefore suggest the reader to familiarize themselves with this strategy here, as it is easier to get lost in the details of the induction step.

By Lemma 6.3.8 and Lemma 6.3.9 we only need to prove (Ind-I) and (Ind-II). For (Ind-I) we need to prove that

$$\hat{\mathbf{I}}_0 = \big(\mathbf{f}_N(x_0)\big) \cup \big(\langle \nabla \mathbf{f}_N(x_0), v\rangle : v \in \mathbf{v}_{[0:d_1)}\big) = \big(\mathbf{f}_N(x_0), \langle \nabla \mathbf{f}_N(x_0), \mathbf{v}_{[0:d_1)}\rangle\big)$$

converges to a limiting $\hat{\mathscr{I}}_0 = (\mathfrak{f}_0, \gamma_0^{(0)}, \dots, \gamma_0^{(d_1-1)})$ in probability. Note that $d_1 \leq 2$ as the vector space of evaluation points is defined in (6.14) to be previsible, that is

$$V_1 = \operatorname{span}\{x_0, \nabla \mathbf{f}_N(x_0)\}.$$

So we have $d_1 - 1 \leq 1$ depending on whether the starting point x_0 is zero. The vector

$$\gamma_0 = (\gamma_0^{(i)})_{i<d_1} = (\gamma_0^{(0)}, \dots, \gamma_0^{(d_1-1)})$$

might therefore collapse to a single entry $\gamma_0^{(0)}$. The approach we will now take mirrors the approach in the induction step. Beyond the pedagogical benefit, this order is also quite natural for the induction start when using the notation we introduced.

Step 1 First we prove that

$$\big(\mathbf{f}_N(x_0)\big) \cup \big(\langle \nabla \mathbf{f}_N(x_0), v\rangle : v \in \mathbf{v}_{[0:d_0)}\big)$$

converges. We highlight that the limit d_0 is purposefully different from the limit in $\hat{\mathbf{I}}_0$.

Step 2 We will then squeeze in the proof of (Ind-II), which ensures the dimension actually increases

$$d_1 = d_0 + 1.$$

Step 3 Finally we prove that $\langle \nabla \mathbf{f}_N(x_0), \mathbf{v}_{d_0}\rangle = \langle \nabla \mathbf{f}_N(x_0), \mathbf{v}_{d_1-1}\rangle$ converges.

While we could prove (Ind-II) before (Ind-I) in the induction start, we will require the results of *Step 1* for (Ind-II) in the induction step. The element $\mathbf{v}_{d_0}$ is not only different from $\mathbf{v}_{[0:d_0)}$ in the sense that the dimension increase needs to be shown, it is also the first truly random basis element here. In the induction step the difference will be between the previsible basis elements and the last non-previsible element, which needs to be treated differently.

Step 1 Since our random function is (non-stationary) isotropic with $\mathbf{f}_N \sim \mathcal{N}(\mu, \kappa)$, we have

$$\mathbf{f}_N(x_0) \sim \mathcal{N}\Big(\mu\big(\tfrac{\|x_0\|^2}{2}\big), \frac{1}{N}\kappa\big(\tfrac{\|x_0\|^2}{2}, \tfrac{\|x_0\|^2}{2}, \|x_0\|^2\big)\Big).$$

this immediately implies convergence of the first component

$$\mathbf{f}_N(x_0) \xrightarrow[N\to\infty]{p} \mu\big(\tfrac{\|x_0\|^2}{2}\big) = \mu\big(\tfrac{\lambda^2}{2}\big) =: \mathfrak{f}_0.$$

Let us now turn to the convergence of the inner products. We consider two cases.

Case ($x_0 \neq 0$): In this case we have $\mathbf{v}_0 = \frac{x_0}{\|x_0\|}$ by Definition 6.3.1. Therefore $\mathbf{v}_0$ is deterministic and by an application of Lemma 4.5.2 we have

$$D_{\mathbf{v}_0}\mathbf{f}_N(x_0) \sim \mathcal{N}\Big(\mu'\big(\tfrac{\|x_0\|^2}{2}\big)\|x_0\|, \frac{1}{N}\big[(\kappa_{12}+\kappa_{13}+\kappa_{32}+\kappa_{33})\|x_0\|^2 + \kappa_3\big]\Big),$$

using the notation

$$\kappa := \kappa\big(\tfrac{\|x_0\|^2}{2}, \tfrac{\|x_0\|^2}{2}, \|x_0\|^2\big) = \kappa\big(\tfrac{\lambda^2}{2}, \tfrac{\lambda^2}{2}, \lambda^2\big)$$

to omit inputs to the kernel κ. But this immediately implies

$$\langle \nabla \mathbf{f}_N(x_0), \mathbf{v}_0\rangle = D_{\mathbf{v}_0}\mathbf{f}_N(x_0) \xrightarrow[N\to\infty]{p} \mu'\big(\tfrac{\|x_0\|^2}{2}\big)\|x_0\| =: \gamma_0^{(0)}.$$

As $d_0 = \dim(V_0) = \dim(\operatorname{span}(x_0)) = 1$, we have covered all elements in $\mathbf{v}_{[0:d_0)}$.

Case ($x_0 = 0$): In this case, $d_0 = \dim(V_0) = 0$ and $\mathbf{v}_{[0:d_0)}$ is empty. We therefore do not have to do anything.

Step 2 To prove the dimension increases, we consider the Gram-Schmidt candidate

$$\begin{aligned}\tilde{\mathbf{v}}_0 &:= \nabla \mathbf{f}_N(x_0) - P_{V_0}\mathbf{f}_N(x_0) = \nabla \mathbf{f}_N(x_0) - \sum_{i<d_0}\langle \nabla \mathbf{f}_N(X_0), \mathbf{v}_i\rangle \mathbf{v}_i \\ &= \sum_{i=d_0}^{N-1}\langle \nabla \mathbf{f}_N(X_0), \mathbf{w}_i^{(0)}\rangle \mathbf{w}_i^{(0)},\end{aligned}$$

where $\mathbf{w}^{(0)}_{[d_0:N)}$ is the basis defined to be orthogonal to V_0 (Definition 6.3.1). Since V_0 is deterministic, this orthogonal basis is also deterministic. In the induction step, both will only be previsible. Note that in the case $x_0 = 0$, $\tilde{\mathbf{v}}$ is simply the entire gradient by definition of d_0.

Since the $\mathbf{w}_i^{(0)}$ are deterministic, and orthogonal to x_0 in either case, we can apply Lemma 4.5.2 to obtain

$$D_{\mathbf{w}_i^{(0)}}\mathbf{f}_N(x_0) \overset{\text{iid}}{\sim} \mathcal{N}\Big(0, \frac{\kappa_3}{N}\Big).$$

This implies that there exist $Y_i \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$ such that

$$D_{\mathbf{w}_i^{(0)}}\mathbf{f}_N(x_0) = \sqrt{\frac{\kappa_3}{N}}Y_i.$$

But since $d_0 \leq 1$ and in particular d_0 is finite, this implies with the law of large numbers

$$\|\tilde{\mathbf{v}}_0\|^2 = \sum_{i=d_0}^{N-1} \langle \nabla \mathbf{f}_N(x_0), w_i \rangle^2 = \kappa_3 \cdot \frac{1}{N} \sum_{i=d_0}^{N-1} Y_i^2 \xrightarrow[N\to\infty]{p} \kappa_3. \tag{6.28}$$

Since we have $\kappa_3 > 0$ by Lemma 6.A.1, $\|\tilde{\mathbf{v}}_0\|^2$ is almost surely strictly greater than zero, which implies $\dim(V_1) > \dim(V_0)$ almost surely. Additionally, the dimension also increases in the limit, i.e. $\lim_{N\to\infty} \|\tilde{\mathbf{v}}_0\| = \kappa_3 > 0$. We have therefore shown all of (Ind-II).

Step 3 We have by definition of $\mathbf{v}_{d_0}$ in (6.16) and the definition of $\tilde{\mathbf{v}}_0$

$$\langle \nabla \mathbf{f}_N(x_0), \mathbf{v}_{d_0} \rangle = \left\langle \nabla \mathbf{f}_N(x_0), \tfrac{\tilde{\mathbf{v}}_0}{\|\tilde{\mathbf{v}}_0\|} \right\rangle = \|\tilde{\mathbf{v}}_0\| \xrightarrow[N\to\infty]{p} \sqrt{\kappa_3} =: \gamma_0^{(d_0)}.$$

This finishes the last step and therefore the induction start.

Induction step $(n-1) \to n$

We now get to the main body of the proof. Before we start, let us recapitulate the lemmas we proved for complexity reduction. By Lemma 6.3.8 and Lemma 6.3.9 it is sufficient to prove the statements (Ind-I) and (Ind-II), as the other follow.

With $G_n := \big(\nabla \mathbf{f}_N(X_k) : k \leq n\big)$ the modified information $\hat{\mathbf{I}}_n$ of the information convergence (Ind-I) was given by

$$\hat{\mathbf{I}}_n := \Big(\mathbf{f}_N(X_k) : k \leq n\Big) \cup \Big(\langle v, w \rangle : v \in \mathbf{v}_{[0:d_{n+1})}, w \in G_n\Big).$$

By induction we already have $\hat{\mathbf{I}}_{n-1} \to \hat{\mathscr{I}}_{n-1}$. And due to our discussion of $\langle \mathbf{v}_i, \nabla \mathbf{f}_N(X_k) \rangle$ for $i \geq d_{n+1} \geq d_{k+1}$ in Lemma 6.3.10 we therefore only need to prove

$$\begin{pmatrix} \mathbf{f}_N(X_n) \\ \langle \mathbf{v}_{[0:d_{n+1})}, \nabla \mathbf{f}_N(X_n) \rangle \end{pmatrix} = \begin{pmatrix} \mathbf{f}_N(X_n) \\ D_{\mathbf{v}_0} \mathbf{f}_N(X_n) \\ \vdots \\ D_{\mathbf{v}_{d_{n+1}-1}} \mathbf{f}_N(X_n) \end{pmatrix} \xrightarrow[N\to\infty]{p} \begin{pmatrix} \mathfrak{f}_n \\ \gamma_n \end{pmatrix}, \tag{6.29}$$

where we used the definition $\gamma_n = (\gamma_n^{(i)})_{i \in [0:d_{n+1})}$ from Remark 6.3.11. In this remark we have also explained how $\gamma_k \in \mathbb{R}^{d_{k+1}}$ can be viewed as a member of $\mathbb{R}^{d_{n+1}}$ and laid out the concatenated vectors $\gamma_{[0:n]}$ in triangular matrix form. In the following we will arrange the random information into a matching matrix form by the definition of an auxiliary function Z.

This function has multiple purposes: First, it turns the direction vectors of the directional derivatives into inputs for an application of previsible sampling (Corollary 3.1.8), which we explained in Idea 3 of the proof sketch (Section 6.3.1). Second, it structures the information vector $\hat{\mathbf{I}}_n$ into a better readable matrix form for human digestion. Third, it allows us to rearrange the elements of the column vectors we concatenate into row vectors. This effectively rearranges the covariance matrix of Z into a more sparse block form.

With some abuse of notation we define Z for a varying number[25] of direction vectors $w_1, \dots, w_m \in \mathbb{R}^N$ and evaluation points $x_1, \dots, x_k \in \mathbb{R}^N$ for $m \leq N$ and $k \leq n$

$$Z(w_1, \dots, w_m; x_0, \dots, x_k) := \begin{pmatrix} \mathbf{f}_N(x_0) & \dots & \mathbf{f}_N(x_k) \\ D_{w_1} \mathbf{f}_N(x_0) & \cdots & D_{w_1} \mathbf{f}_N(x_k) \\ \vdots & & \vdots \\ D_{w_m} \mathbf{f}_N(x_0) & \cdots & D_{w_m} \mathbf{f}_N(x_k) \end{pmatrix}.$$

The number of inputs is therefore not variable and we separate their type by a semicolon. Note that we are not interested in this matrix as an operator. The two

[25] formally, one can view this as slices of a function $\mathbb{R}^{N^2} \times \mathbb{R}^{(n+1)N} \to \mathbb{R}^{(N+1)(n+1)}$.

dimensional layout is only used for better readability and we will generally view it as a vector. For this purpose we treat the matrix as 'row-major', i.e. whenever we treat it like a vector, we concatenate the rows. This grouping is purposefully different from the column layout of the γ_k as it will enable a block matrix sparsity in the covariance matrices later on.

By induction we have information convergence (Ind-I) for $n-1$, which can be expressed using Z as

$$Z(\mathbf{v}_{[0:d_n)}; X_{[0:n)}) = \begin{pmatrix} \mathbf{f}_N(X_0) & \cdots & \mathbf{f}_N(X_{n-1}) \\ D_{\mathbf{v}_0}\mathbf{f}_N(X_0) & \cdots & D_{\mathbf{v}_0}\mathbf{f}_N(X_{n-1}) \\ \vdots & & \vdots \\ D_{\mathbf{v}_{d_n-1}}\mathbf{f}_N(X_0) & \cdots & D_{\mathbf{v}_{d_n-1}}\mathbf{f}_N(X_{n-1}) \end{pmatrix} \xrightarrow[N\to\infty]{p} \begin{pmatrix} \mathfrak{f}_{[0:n)} \\ \gamma_{[0:n)} \end{pmatrix}. \tag{6.30}$$

Observe that the induction step simply adds the additional column given in (6.29) to the right of the matrix, where we use Lemma 6.3.10 to argue that we can arbitrarily extend the number of rows.

We will now follow the same strategy as in the induction start:

Step 1 First, we prove the convergence of the 'new column'

$$Z(\mathbf{v}_{[0:d_n)}; X_n) = \begin{pmatrix} \mathbf{f}_N(X_n) \\ \langle \mathbf{v}_{[0:d_n)}, \nabla \mathbf{f}_N(X_n) \rangle \end{pmatrix} \xrightarrow[N\to\infty]{p} \begin{pmatrix} \mathfrak{f}_n \\ \gamma_n^{([0:d_n))} \end{pmatrix}$$

Step 2 We squeeze in the proof of (Ind-II), which ensures asymptotically

$$d_{n+1} = d_n + 1.$$

Step 3 We finally prove convergence of the 'new corner element'

$$Z(\mathbf{v}_{d_n}; X_n) = \langle \mathbf{v}_{d_n}, \nabla \mathbf{f}_N(X_n) \rangle \xrightarrow[N\to\infty]{p} \gamma_n^{(d_n)}.$$

The reason for this split are twofold. First, we have that $\mathbf{v}_{[0:d_n)}$ is previsible, that is $\mathcal{F}_{n-1}$-measurable, while $\mathbf{v}_n$ is not. $\mathbf{v}_n$ must therefore be treated differently. Second, we need to construct $\mathbf{v}_{d_n}$ from $\tilde{\mathbf{v}}_n$, which will naturally prove (Ind-II) before we get to the convergence of the inner product in *Step 3*.

This strategy can get a bit lost, as we will spend the majority of our time with *Step 1* before wrapping up *Step 2* and *Step 3* fairly quickly. An additional reason for this strategy to get lost is, that we will not just consider $(\mathbf{f}_N(x_n), \langle \mathbf{v}_{[0:d_n)}, \nabla \mathbf{f}_N(X_n) \rangle)$, but instead consider the conditional distribution of $(\mathbf{f}_N(x_n), \langle B_n, \nabla \mathbf{f}_N(X_n) \rangle)$ for the entire previsible basis $B_n = (\mathbf{v}_{[0:d_n)}, \mathbf{w}^{(n)}_{[d_n:N)})$. We will aggregate the directional derivatives in the directions $\mathbf{w}^n_i$ later into $\tilde{\mathbf{v}}_n$, which means that we work towards *Step 1* and *Step 2* simultaneously. The first objective will be, to apply the previsible sampling lemma such that we can treat the previsible inputs as deterministic.

Getting rid of the random input In the proof sketch we outlined how we needed to treat the random input with care (cf. 6.3.1, Idea 3). In particular we need to consider the (basis; point) pairs

$$(B_0; X_0), \cdots, (B_n; X_n) := \left(\mathbf{v}_{[0:d_0)}, \mathbf{w}^{(0)}_{[d_0:N)}; X_0\right), \cdots, \left(\mathbf{v}_{[0:d_n)}, \mathbf{w}^{(n)}_{[d_n:N)}; X_n\right).$$

Recall that we defined the basis B_k in Definition 6.3.1 to be previsible just like X_k, i.e. measurable with regard $\mathcal{F}_{k-1}$. Moreover, since all basis changes are previsible and invertible it is straightforward to check inductively that there exists a measurable bijective map between different basis representations of the derivative information up to n

$$\left(\mathbf{f}_N(X_k), \nabla \mathbf{f}_N(X_k)\right)_{k<n} \leftrightsquigarrow \left(Z(B_k; X_k)\right)_{k<n}.$$

This implies that their generated sigma algebras are the same

$$\mathcal{F}_{n-1} \overset{\text{def.}}{=} \sigma\Big(\big(\mathbf{f}_N(X_k), \nabla\mathbf{f}_N(X_k)\big), k \leq n-1\Big) = \sigma\big(Z(B_k; X_k), k \leq n-1\big).$$

By Application of Corollary 3.1.8 we then have, for all bounded measurable h,

$$\mathbb{E}\Big[h\big(Z(B_n; X_n)\big) \mid \mathcal{F}_{n-1}\Big] = F(X_{[0:n]}; B_{[0:n]}),$$

for the function F defined with deterministic evaluation points x_k and basis b_k as

$$F(x_{[0:n]}; b_{[0,n]}) := \int h(z)\mathbb{P}\Big[Z(b_n; x_n) \in dz \mid Z(b_0; x_0), \ldots, Z(b_{n-1}; x_{n-1})\Big],$$

using the standard consistent joint conditional distribution for Gaussian distributions (cf. Example 3.1.7). In other words, we are allowed to treat the inputs $(B_k; X_k)$ as deterministic when calculating the conditional distribution. But keeping track of $n+1$ different basis B_k for every is evaluation point X_k is very inconvenient. So our next goal is to reduce the number of basis to one. For this we optimistically define for a single basis b

$$G(x_{[0:n]}; b) := \mathbb{E}\Big[h\big(Z(b; x_n)\big) \mid Z(b; x_0), \ldots, Z(b; x_{n-1})\Big],$$

where we assume here and in the following, that this conditional distribution is defined via the standard consistent joint conditional distribution of Example 3.1.7. We are now going to observe that F is constant in all but the most recent basis b_n, i.e.

$$F(x_{[0:n]}; b_{[0:n]}) = G(x_{[0:n]}; b_n). \tag{6.31}$$

That is because the sigma algebras generated by different basis representations are identical as they can be bijectively translated into each other, i.e. for any b

$$\sigma(Z(b_0; x_0), \ldots, Z(b_{n-1}; x_{n-1})) = \sigma(Z(b; x_0), \ldots, Z(b; x_{n-1})).$$

In particular this is true for $b = b_n$ and thus we have (6.31), i.e.

$$\begin{aligned}&\mathbb{E}\Big[h\big(Z(b_n; x_n)\big) \mid Z(b_0; x_0), \ldots, Z(b_{n-1}; x_{n-1})\Big]\\ &= \mathbb{E}\Big[h\big(Z(b_n; x_n)\big) \mid Z(b_n; x_0), \ldots, Z(b_n; x_{n-1})\Big].\end{aligned}$$

Since we only need to consider the most recent basis from now on, we drop the index and write $\mathbf{w}_k := \mathbf{w}_k^{(n)}$ and summarize our result as

$$\mathbb{E}\Big[h\big(Z(\mathbf{v}_{[0:d_n)}, \mathbf{w}_{[d_n:N)}; X_n)\big) \mid \mathcal{F}_{n-1}\Big] = G(X_0, \ldots, X_n; \mathbf{v}_{[0:d_n)}, \mathbf{w}_{[d_n:N)}). \tag{6.32}$$

Recall that we use the semicolon to separate the basis elements from the evaluation points, which should be treated as concatenated vectors respectively.

In essence, by definition of G we can now treat our evaluation points $X_{[0:n]}$ and coordinate system $B_n = (\mathbf{v}_{[0:d_n)}, \mathbf{w}_{[d_n:N)})$ as deterministic when calculating the conditional distribution.

The conditional distribution is known in the Gaussian case Since Z is a Gaussian random function, we have for non-random input (b, x) that its conditional distribution is also Gaussian (cf. Theorem 2.A.1). Equation (6.32) translates this result to the random input and we can therefore conclude that

$$Z(B_n; X_n) \mid \mathcal{F}_{n-1} \sim \mathcal{N}\Big(\mathbb{E}[Z(B_n; X_n) \mid \mathcal{F}_{n-1}], \mathrm{Cov}[Z(B_n; X_n) \mid \mathcal{F}_{n-1}]\Big).$$

In particular there exist independent $Y_1, \ldots, Y_N \sim \mathcal{N}(0,1)$ independent of $\mathcal{F}_{n-1}$ such that we have in distribution

$$Z(B_n; X_n) = \mathbb{E}[Z(B_n; X_n) \mid \mathcal{F}_{n-1}] + \sqrt{\mathrm{Cov}[Z(B_n; X_n) \mid \mathcal{F}_{n-1}]} \begin{pmatrix} Y_0 \\ \vdots \\ Y_N \end{pmatrix}, \tag{6.33}$$

where we denote the cholesky decomposition of a matrix A by the squareroot $\sqrt{A}$.

Remark 6.3.12. As we only wish to prove convergence in probability to deterministic numbers, this distributional equality will be enough for our purposes, but if the covariance was invertible one could get a almost sure equality (cf. Remark 2.A.2). Later, we will also see that the covariance is at least asymptotically invertible due to strict positive definiteness of Z and asymptotically different evaluation points (Ind-IV).

Calculating the conditional expectation and covariance Since the first two moments can be calculated by treating the inputs as deterministic (cf. Equation (6.32)), we can calculate the conditional expectation and variance using the well known formulas for Gaussian conditionals (cf. Theorem 2.A.1)

$$\mathbb{E}[Z(B_n; X_n) \mid \mathcal{F}_{n-1}] = \mu_n^N + \boldsymbol{\Sigma}^N_{[0:n),n}{}^T [\boldsymbol{\Sigma}^N_{[0:n)}]^{-1} (Z(B_n; X_{[0:n)}) - \mu^N_{[0:n)}) \tag{6.34}$$

$$\mathrm{Cov}[Z(B_n; X_n) \mid \mathcal{F}_{n-1}] = \tfrac{1}{N}\Big[\boldsymbol{\Sigma}^N_n - \boldsymbol{\Sigma}^N_{[0:n),n}{}^T [\boldsymbol{\Sigma}^N_{[0:n)}]^{-1} \boldsymbol{\Sigma}^N_{[0:n),n}\Big], \tag{6.35}$$

where we define the mixed covariance by

$$\tfrac{1}{N}\boldsymbol{\Sigma}^N_{[0:n),n} := \mathcal{C}_Z(B_n; X_{[0:n)}, X_n) \quad \mathcal{C}_Z(b; x_{[0:n)}, x_n) := \mathrm{Cov}(Z(b; x_{[0:n)}), Z(b; x_n)),$$

the autocovariance matrices by

$$\begin{aligned} \tfrac{1}{N}\boldsymbol{\Sigma}^N_{[0:n)} &:= \mathcal{C}_Z(B_n; X_{[0:n)}) & \mathcal{C}_Z(b; x_{[0:n)}) &:= \mathrm{Cov}[Z(b; x_{[0:n)})] \\ \tfrac{1}{N}\boldsymbol{\Sigma}^N_n &:= \mathcal{C}_Z(B_n; X_n) & \mathcal{C}_Z(b; x_n) &:= \mathrm{Cov}[Z(b; x_n)] \end{aligned}$$

and the expectations by

$$\begin{aligned} \mu_n^N &:= \mu_Z(B_n; X_n) & \mu_Z(b; x_n) &:= \mathbb{E}[Z(b; x_n)] \\ \mu^N_{[0:n)} &:= \mu_Z(B_n; X_n) & \mu_Z(b; x_{[0:n)}) &:= \mathbb{E}[Z(b; x_{[0:n)})] \end{aligned}$$

The detour over the functions $\mathcal{C}_Z$ and μ_Z was necessary to define the unconditional covariance matrices properly, because we may only treat previsible random variables as deterministic in the calculation of conditional distributions (using Corollary 3.1.8). So in general, since B_n and X_n are random variables, we have for the unconditional covariance

$$\mathcal{C}_Z(B_n; X_n) \neq \mathrm{Cov}[Z(B_n; X_n)].$$

To ensure the inputs are treated as deterministic as Equation (6.32) demands, we therefore had to make sure the covariance was already calculated before we plugged in our random input.

Note that we have moved the dimensional scaling $\frac{1}{N}$ of the covariances (cf. Lemma 4.5.2) outside of our definition of $\boldsymbol{\Sigma}^N_{[0:n)}$, $\boldsymbol{\Sigma}^N_{[0:n),n}$ and $\boldsymbol{\Sigma}^N_n$, as we are now going to prove their entries, and the entries of $\mu^N_{[0:n)}$ and μ^N_n, converge. This will eventually allow us to prove that the conditional expectation and covariance will converge, which leads to the stochastic convergence of Z we want to obtain for (Ind-I). Moving the dimensional scaling out also made the dimensional scaling of the conditional covariance in (6.35) much more visible.

Covariance matrix entries converge Recall that Z is made up of evaluations of $\mathbf{f}_N$ and $D_v\mathbf{f}_N$ for directions v. Thus the entries of its covariance matrices are given by Lemma 4.5.2, which we restate here for your convenience.

Lemma 4.5.2 (Covariance of derivatives under (non-stationary) isotropy)**.** *Let $\mathbf{f} \sim \mathcal{N}(\mu, \kappa)$ be (non-stationary) isotropic (Definition 4.2.1). Then the expectation of the directional derivative is given by*

$$\mathbb{E}[D_v\mathbf{f}(x)] = \mu'\big(\tfrac{\|x\|^2}{2}\big)\langle x, v\rangle, \tag{4.10}$$

and, using the abbreviation $\kappa := \kappa\big(\frac{\|x\|^2}{2}, \frac{\|y\|^2}{2}, \langle x, y\rangle\big)$ *to omit inputs, the covariance of directional derivatives is given by*

$$\operatorname{Cov}(D_v\mathbf{f}(x), \mathbf{f}(y)) = \frac{1}{N}\Big[\kappa_1\langle x, v\rangle + \kappa_3\langle y, v\rangle\Big] \tag{4.11}$$

$$\operatorname{Cov}(D_v\mathbf{f}(x), D_w\mathbf{f}(y)) = \frac{1}{N}\Big[\underbrace{\kappa_{12}\langle x, v\rangle\langle y, w\rangle + \kappa_{13}\langle x, v\rangle\langle x, w\rangle + \kappa_{32}\langle y, v\rangle\langle y, w\rangle + \kappa_{33}\langle y, v\rangle\langle y, w\rangle}_{(I)} + \underbrace{\kappa_3\langle v, w\rangle}_{(II)}\Big]. \tag{4.12}$$

In particular, if the directions v, w *are orthogonal to the points* x, y *then (I) of* (4.11) *and* (4.12) *are zero. (II) in turn is zero, if the directions* v, w *are orthogonal.*

Recall that we have (6.19) of (Ind-III) for $n-1$ by induction assumption, i.e. we have for all $k \leq n$ and all $i < d_n$

$$\langle X_k, \mathbf{v}_i\rangle \xrightarrow[N\to\infty]{p} y_k^{(i)} \quad \text{and} \quad \|X_k\|^2 = \sum_{i=0}^{d_n-1} \langle X_k, \mathbf{v}_i\rangle^2 \xrightarrow[N\to\infty]{p} \|y_k\|^2.$$

Since the X_k for $k \leq n$ are contained in V_n orthogonal to $\mathbf{w}_{[d_n:N)}$ we also have for all $i \geq d_n$

$$\langle X_k, \mathbf{w}_i\rangle = 0. \tag{6.36}$$

Put together, we have that all the inner products $\langle X_k, v\rangle$ for $k \leq n$ and $v \in B_n$ converge.

The entries of $\mathbf{\Sigma}^N_{[0:n)}$, $\mathbf{\Sigma}^N_{[0:n),n}$, $\mathbf{\Sigma}^N_n$, $\mu^N_{[0:n)}$ and μ^N_n calculated with Lemma 4.5.2 therefore all converge in probability by the continuous mapping theorem, since κ and μ are sufficiently smooth by Assumption 6.1.9 used in Theorem 6.3.2. Observe that it was very important to remove the dimensional scaling $\frac{1}{N}$ of (4.11) and (4.12) from the covariance matrices $\mathbf{\Sigma}^N_{[0:n)}$, $\mathbf{\Sigma}^N_{[0:n),n}$ and $\mathbf{\Sigma}^N_n$, as their entries would otherwise all converge to zero. As we want to invert $\mathbf{\Sigma}^N_{[0:n)}$ this would have been very inconvenient.

We further note that the sizes of these covariance matrices change with the dimension, because the number of directional derivatives increases with N which increases the size of $Z(\mathbf{v}_{[0:d_n)}, \mathbf{w}_{[d_n:N)}; X_{[0:n)})$ and therefore the size of its covariance matrix. The convergence of their entries is therefore not yet sufficient for a limiting object to be well defined.

Splitting the increasing matrices into block matrices of constant size This is the heart of the proof, which relies heavily on our custom coordinate system $B_n = (\mathbf{v}_{[0:d_n)}, \mathbf{w}_{[d_n:N)})$. To understand this, let us focus on the covariance of derivatives given in (4.12) of Lemma 4.5.2, i.e.

$$\operatorname{Cov}(D_v\mathbf{f}_N(x), D_w\mathbf{f}_N(y)) = \frac{1}{N}\Big[\underbrace{\kappa_{12}\langle x, v\rangle\langle y, w\rangle + \kappa_{13}\langle x, v\rangle\langle x, w\rangle + \kappa_{32}\langle y, v\rangle\langle y, w\rangle + \kappa_{33}\langle y, v\rangle\langle y, w\rangle}_{\text{(I)}} + \underbrace{\kappa_3\langle v, w\rangle}_{\text{(II)}}\Big]. \tag{4.12}$$

Notice that for the $\mathbf{w}_i$ the part (I) is always zero by (6.36). This is why we defined $\mathbf{w}_{[d_n:n)}$ to be orthogonal to V_n in Definition 6.3.1. Since we also defined our basis to be an orthonormal basis, (II) is only non-zero when covariances of the directional derivatives in the same direction are taken.

As we treat our Z matrix (6.30) as row-major, the directional derivatives are grouped by direction $\mathbf{w}_i$ in Z, which therefore results in the following block matrix

structure.

$$\begin{aligned}\tfrac{1}{N}\boldsymbol{\Sigma}^N_{[0:n)} &= \mathcal{C}_Z(\mathbf{v}_{[0:d_n)}, \mathbf{w}_{[d_n:N)}; X_{[0:n)}) \\ &= \begin{pmatrix} \mathcal{C}_Z(\mathbf{v}_{[0:d_n)}; X_{[0:n)}) & & & \\ & \mathcal{C}_{D_{\mathbf{w}_{d_n}}\mathbf{f}_N}(X_{[0:n)}) & & \\ & & \ddots & \\ & & & \mathcal{C}_{D_{\mathbf{w}_{N-1}}\mathbf{f}_N}(X_{[0:n)}) \end{pmatrix} \\ &=: \frac{1}{N}\begin{pmatrix} \boxed{\boldsymbol{\Sigma}^{v,N}_{[0:n)}} & & & \\ & \boxed{\boldsymbol{\Sigma}^{w,N}_{[0:n)}} & & \\ & & \ddots & \\ & & & \boxed{\boldsymbol{\Sigma}^{w,N}_{[0:n)}} \end{pmatrix}, \end{aligned} \tag{6.37}$$

For $\boldsymbol{\Sigma}^{w,N}_{[0:n)}$ to be well defined, we require the blocks $\mathcal{C}_{D_{\mathbf{w}_i}\mathbf{f}_N}(X_{[0:n)})$ to not depend on i. But since (I) is zero and $\langle \mathbf{w}_i, \mathbf{w}_i\rangle = 1$ we have by (4.12) of Lemma 4.5.2

$$\begin{aligned}&\mathcal{C}_{D_{\mathbf{w}_i}\mathbf{f}_N}(X_{[0:n)}) \\ &= \frac{1}{N}\begin{pmatrix} \kappa_3\big(\frac{\|X_0\|^2}{2}, \frac{\|X_0\|^2}{2}, \langle X_0, X_0\rangle\big) & \cdots & \kappa_3\big(\frac{\|X_0\|^2}{2}, \frac{\|X_{n-1}\|^2}{2}, \langle X_0, X_{n-1}\rangle\big) \\ \vdots & & \vdots \\ \kappa_3\big(\frac{\|X_{n-1}\|^2}{2}, \frac{\|X_0\|^2}{2}, \langle X_{n-1}, X_0\rangle\big) & \cdots & \kappa_3\big(\frac{\|X_{n-1}\|^2}{2}, \frac{\|X_{n-1}\|^2}{2}, \langle X_{n-1}, X_{n-1}\rangle\big) \end{pmatrix} \\ &=: \tfrac{1}{N}\boldsymbol{\Sigma}^{w,N}_{[0:n)}. \end{aligned}$$

In particular there is no dependence on $\mathbf{w}_i$ so $\boldsymbol{\Sigma}^{w,N}_{[0:n)}$ is well defined. Since these block matrices are of constant size, they converge if all their finitely many entries converge. But we already argued that the entries converge (in the previous paragraph) and we therefore have[26]

$$\boldsymbol{\Sigma}^{v,N}_{[0:n)} \xrightarrow[N\to\infty]{p} \boldsymbol{\Sigma}^{v,\infty}_{[0:n)} \in \mathbb{R}^{n(d_n+1)\times n(d_n+1)} \quad \text{and} \quad \boldsymbol{\Sigma}^{w,N}_{[0:n)} \xrightarrow[N\to\infty]{p} \boldsymbol{\Sigma}^{w,\infty}_{[0:n)} \in \mathbb{R}^{n\times n}.$$

For the mixed covariance we similarly have

$$\begin{aligned}&\tfrac{1}{N}\boldsymbol{\Sigma}^N_{[0:n),n} \\ &= \mathcal{C}_Z(\mathbf{v}_{[0:d_n)}, \mathbf{w}_{[d_n:N)}; X_{[0:n)}, X_n) \\ &= \begin{pmatrix} \mathcal{C}_Z(\mathbf{v}_{[0:n)}; X_{[0:n)}, X_n) & & & \\ & \mathcal{C}_{D_{w_n}\mathbf{f}_N}(X_{[0:n)}, X_n) & & \\ & & \ddots & \\ & & & \mathcal{C}_{D_{w_{N-1}}\mathbf{f}_N}(X_{[0:n)}, X_n) \end{pmatrix} \\ &=: \frac{1}{N}\begin{pmatrix} \boxed{\boldsymbol{\Sigma}^{v,N}_{[0:n),n}} & & & \\ & \boxed{\boldsymbol{\Sigma}^{w,N}_{[0:n),n}} & & \\ & & \ddots & \\ & & & \boxed{\boldsymbol{\Sigma}^{w,N}_{[0:n),n}} \end{pmatrix}, \end{aligned} \tag{6.38}$$

with a similar argument why $\boldsymbol{\Sigma}^{w,N}_{[0:n),n}$ is well defined as for $\boldsymbol{\Sigma}^{w,N}_{[0:n)}$. And again by the discussion of the previous paragraph establishing convergence of the entries,

[26] The increasing number of identical $\boldsymbol{\Sigma}^{w,N}_{[0:n)}$ will drive the law of large numbers of

$$\|P_{V_n^\perp}\nabla \mathbf{f}_N(x_0)\|^2 = \sum_{i=d_n}^{N-1}(D_{\mathbf{w}_i}\mathbf{f}_N(x_0))^2$$

similar to the first step (cf. Equation (6.28)). At the moment they are still matrices but this will change when combined with the mixed covariances $\boldsymbol{\Sigma}^{w,N}_{[0:n),n}$ finally resulting in (6.47).

we have that these block matrices of constant size converge

$$\mathbf{\Sigma}^{v,N}_{[0:n),n} \xrightarrow[N\to\infty]{p} \Sigma^{v,\infty}_{[0:n),n} \in \mathbb{R}^{n(d_n+1)\times(d_n+1)} \quad \text{and} \quad \mathbf{\Sigma}^{w,N}_{[0:n),n} \xrightarrow[N\to\infty]{p} \Sigma^{w,\infty}_{[0:n),n} \in \mathbb{R}^{n\times 1}.$$

We can also split up the autocovariance Σ^N_n in a similar fashion with

$$\mathbf{\Sigma}^{v,N}_{n} \xrightarrow[N\to\infty]{p} \Sigma^{v,\infty}_{n} \in \mathbb{R}^{(d_n+1)\times(d_n+1)} \quad \text{and} \quad \mathbf{\Sigma}^{w,N}_{n} \xrightarrow[N\to\infty]{p} \Sigma^{w,\infty}_{n} \in \mathbb{R}^{1\times 1}.$$

Finally, the expectation functions can also be split into a block containing the directional derivatives $\mathbf{v}_{[0:d_n)}$ spanning V_n and the expectations of directional derivatives in the $\mathbf{w}_i$ directions. More specifically we have

$$\begin{aligned} \mu^{v,N}_n = \mu_Z(\mathbf{v}_{[0:d_n)}; X_n) \xrightarrow[N\to\infty]{p} \mu^{v,\infty}_n \in \mathbb{R}^{d_n+1} \qquad \text{and} \\ \mu^{v,N}_{[0:n)} \xrightarrow[N\to\infty]{p} \mu^{v,\infty}_{[0:n)} \in \mathbb{R}^{n(d_n+1)}, \end{aligned} \tag{6.39}$$

which converge since the inner products and norms used in (4.10) converge. Since the directions $\mathbf{w}_i$ are selected orthogonal to all evaluation points $X_n \in V_n$, the inner product in the expectation of the directional derivative (4.10) is always zero and we therefore have

$$\mu^{w,N}_n = \mu_Z(\mathbf{w}_{[d_n:N)}; X_n) = 0 \in \mathbb{R} \quad \text{and} \quad \mu^{w,N}_{[0:n)} = 0 \in \mathbb{R}^n. \tag{6.40}$$

Convergence of the conditional expectation Let us take a step back and review what we have done and want to do. We have argued that $Z(B_n; X_{[0:n)})$ is conditionally Gaussian and that it is therefore enough to understand the conditional expectation and covariance (cf. (6.33)). We have then applied a well known result about the conditional distribution of Gaussian random variables to obtain explicit formulas for the conditional expectation (6.34) and covariance (6.35) made up of unconditional covariance matrices. We proved that their entries converged and split these covariance matrices into block form, such that these blocks of constant size converge in probability. What is left to do, is to put these results together to prove convergence results about $Z(B_n; X_{[0:n)})$. We start with the conditional expectation and then get to $Z(B_n; X_{[0:n)})$ itself by a consideration of the conditional covariance.

So let us take a look at the conditional expectation. Since the $\mathbf{w}_i$ are by definition orthonormal to the previsible running span of evaluation points V_n (6.14), we have $D_{\mathbf{w}_k}\mathbf{f}_N(X_{[0:n)}) = 0$. By applying the block matrix structure (6.37), (6.38) to the the representation of the conditional expectation (6.34) we therefore get using (6.40)

$$\begin{aligned} \mathbb{E}[D_{\mathbf{w}_i}\mathbf{f}_N(X_n) \mid \mathcal{F}_{n-1}] &= \mu^{w,N}_n + \mathbf{\Sigma}^{w,N}_{[0:n),n}{}^T [\mathbf{\Sigma}^{w,N}_{[0:n)}]^{-1}(D_{\mathbf{w}_k}\mathbf{f}_N(X_{[0:n)}) - \mu^{w,N}_{[0:n)}) \\ &= 0. \end{aligned} \tag{6.41}$$

The only interesting part of the vector $\mathbb{E}[Z(\mathbf{v}_{[0:d_n)}, \mathbf{w}_{[d_n:N)}; X_n) \mid \mathcal{F}_{n-1}]$ are therefore the first d_n entries $\mathbb{E}[Z(\mathbf{v}_{[d_n:n)}; X_n) \mid \mathcal{F}_{n-1}]$. For this we use the formula for the conditional expectation (6.34) and our block matrix decompositions (6.37), (6.38) again to obtain

$$\mathbb{E}[Z(\mathbf{v}_{[0:d_n)}; X_n) \mid \mathcal{F}_{n-1}] = \mu^{(v,N)}_n + \mathbf{\Sigma}^{v,N}_{[0:n),n}{}^T [\mathbf{\Sigma}^{v,N}_{[0:n)}]^{-1}(Z(\mathbf{v}_{[0:d_n)}; X_{[0:n)}) - \mu^{(v,N)}_{[0:n)}). \tag{6.42}$$

Since $Z(\mathbf{v}_{[0:d_n)}; X_{[0:n)})$ converges in probability by the induction assumptions (Ind-I) summarized in (6.30), and since we have spent the previous paragraphs proving that the covariance matrices and expectations converge, it almost seems like we are done. But while the conditional expectation (6.34) can handle non-invertible matrices with a generalized inverse (Theorem 2.A.1), an application of continuous mapping requires continuity and the matrix inverse is only continuous at invertible matrices.

Lemma 6.3.13. $\Sigma^{v,\infty}_{[0:n)}$ *is strictly positive definite and therefore invertible.*

Proof. The essential ingredients are that $(\mathbf{f}_N, \nabla\mathbf{f}_N)$ is strictly positive definite by assumption of Theorem 6.3.2 and that the evaluation points are asymptotically different $\rho_{ij} > 0$ by (Ind-IV). The technical complications are that $\Sigma^{v,\infty}_{[0:n)}$ is only defined as the limit of covariance matrices (which are themselves not necessarily strictly positive definite themselves) and that this limit involves changing the domain of $(\mathbf{f}_N, \nabla\mathbf{f}_N)$ as we increase its dimension.

Let us first address the problem of the changing domain. Since we are only interested in the block matrix of the derivatives in the directions contained in V_n and the evaluation points $X_{[0:n)}$ are also contained in V_n we can map them via an isometry to $\mathbb{R}^{d_n}$ which retains all distances and inner products and therefore retains the covariance matrices $\mathbf{\Sigma}^{v,N}_{[0:n)}$ (cf. Lemma 4.5.2). Note that the dimensional scaling $\frac{1}{N}$ is not an issue here, as we have already removed the scaling from $\mathbf{\Sigma}^{v,N}_{[0:n)}$. Therefore $\mathbf{\Sigma}^{v,N}_{[0:n)}$ can be viewed as a sequence of covariance matrices of $(\mathbf{f}_N, \nabla\mathbf{f}_N)$ with $\mathbf{f}_N : \mathbb{R}^{d_n} \to \mathbb{R}$. This solves the problem of the changing domain.

Finally, we need that the limiting matrix $\Sigma^{v,\infty}_{[0:n)}$ is in fact a covariance matrix of $(\mathbf{f}_N, \nabla\mathbf{f}_N)$ and not just a limit of covariance matrices. For this we apply (Ind-III), which ensures that the limiting scalar products (6.19) and distances (6.20) which make up $\Sigma^{v,\infty}_{[0:n)}$ (cf. Lemma 4.5.2) are realizable by points $y_0, \dots, y_{n-1} \in \mathbb{R}^{d_n}$. $\Sigma^{v,\infty}_{[0:n)}$ is therefore a covariance matrix of $(\mathbf{f}_N, \nabla\mathbf{f}_N)$ viewed as a random function with domain $\mathbb{R}^{d_n}$. Since the distances ρ_{ij} are positive by (Ind-IV), the asymptotic representations y_k are distinct and their covariance matrix $\Sigma^{v,\infty}_{[0:n)}$ is therefore strictly positive definite by the strict positive definiteness of $(\mathbf{f}_N, \nabla\mathbf{f}_N)$. □

The matrix inverse is therefore continuous in $\Sigma^{v,\infty}_{[0:n)}$ and by the convergence of the matrices in probability and the convergence of $Z(\mathbf{v}_{[0:d_n)}; X_{[0:n)})$ by induction assumptions (Ind-I) for $n-1$ restated in (6.30), the conditional expectation (6.42) converges in probability

$$\begin{aligned}&\mathbb{E}[Z(\mathbf{v}_{[0:d_n)}; X_n) \mid \mathcal{F}_{n-1}]\\ &\xrightarrow[N\to\infty]{p} \mu^{v,\infty}_n + {\Sigma^{v,\infty}_{[0:n),n-1}}^T [\Sigma^{v,\infty}_{[0:n)}]^{-1}\left(\begin{pmatrix}\mathfrak{f}_{[0:n)}\\ \gamma_{[0:n)}\end{pmatrix} - \mu^{v,\infty}_{[0:n)}\right) =: \begin{pmatrix}\mathfrak{f}_n\\ \gamma_n^{([0:d_n))}\end{pmatrix}. \qquad (6.43)\end{aligned}$$

Together with the conditional expectation of the directional derivatives in the directions $\mathbf{w}_i$ in (6.41), we have now fully determined the conditional expectation.

An analysis of the conditional variance will immediately give us an understanding of the distribution and we can therefore put these considerations together to obtain convergence of Z itself.

***Step 1*: Convergence of $Z(\mathbf{v}_{[0:d_n)}; X_n)$** Recall, we outlined at the start of the induction that we would first prove convergence of $Z(\mathbf{v}_{[0:d_n)}; X_n)$ (i.e. *Step 1*) before proving V_{n+1} to asymptotically have full rank (*Step 2*) and the convergence of the new corner element $D_{\mathbf{v}_{d_n}} \mathbf{f}_N(X_n)$ (*Step 3*).

In (6.43) we obtained the limit of the conditional expectation of $Z(\mathbf{v}_{[0:d_n)}; X_n)$. This limit is also going to be the limit of the new column $Z(\mathbf{v}_{[0:d_n)}; X_n)$ itself. To prove this we only need to control the variance. For this purpose we recall that the unconditional variance is given by

$$\mathbf{\Sigma}^N_n = \begin{pmatrix}\mathbf{\Sigma}^{v,N}_n & \\ & \kappa_3\left(\frac{\|X_n\|^2}{2}, \frac{\|X_n\|^2}{2}, \|X_n\|^2\right)\mathbb{I}_{(N-d_n)\times(N-d_n)}\end{pmatrix}.$$

The conditional covariance matrix is therefore given by

$$\begin{aligned}&\mathrm{Cov}[Z(\mathbf{v}_{[0:d_n)}, \mathbf{w}_{[d_n:N)}; X_n) \mid \mathcal{F}_{n-1}]\\ &= \frac{1}{N}\begin{bmatrix}\mathbf{\Sigma}^{v,N}_n - {\mathbf{\Sigma}^{v,N}_{[0:n),n}}^T [\mathbf{\Sigma}^{v,N}_{[0:n)}]^{-1} \mathbf{\Sigma}^{v,N}_{[0:n),n} & \\ & \sigma^2_{w,d}\mathbb{I}_{(N-d_n)\times(N-d_n)}\end{bmatrix}, \qquad (6.44)\end{aligned}$$

where the conditional variance in the direction of the $\mathbf{w}_i$ is given by

$$\sigma^2_{w,N} := \big(\kappa_3\big(\tfrac{\|X_n\|^2}{2}, \tfrac{\|X_n\|^2}{2}, \|X_n\|^2\big) - \mathbf{\Sigma}^{w,N}_{[0:n),n}{}^T[\mathbf{\Sigma}^{w,N}_{[0:n)}]^{-1}\mathbf{\Sigma}^{w,N}_{[0:n),n}\big). \tag{6.45}$$

We will need $\sigma^2_{w,N}$ for *Step 2* and *Step 3*, but for now it is not important. Due to the diagonal structure we have for our new column $Z(\mathbf{v}_{[0:d_n)}; X_n)$ by (6.33)

$$\begin{aligned}
&Z(\mathbf{v}_{[0:d_n)}, X_n)\\
&= \mathbb{E}[Z(\mathbf{v}_{[0:d_n)}, X_n) \mid \mathcal{F}_{n-1}] + \sqrt{\mathrm{Cov}[Z(\mathbf{v}_{[0:d_n)}, X_n) \mid \mathcal{F}_{n-1}]} \begin{pmatrix} Y_0 \\ \vdots \\ Y_{d_n} \end{pmatrix}\\
&= \mathbb{E}[Z(\hat{v}_{[0:d_n)}, X_n) \mid \mathcal{F}_{n-1}] + \frac{1}{\sqrt{N}}\sqrt{\mathbf{\Sigma}^{v,N}_n - \mathbf{\Sigma}^{v,N}_{[0:n),n}{}^T[\mathbf{\Sigma}^{v,N}_{[0:n)}]^{-1}\mathbf{\Sigma}^{v,N}_{[0:n),n}} \begin{pmatrix} Y_0 \\ \vdots \\ Y_{d_n} \end{pmatrix}\\
&\xrightarrow[N\to\infty]{p} \begin{pmatrix} \mathfrak{f}_n \\ \gamma_n^{([0:d_n))} \end{pmatrix}.
\end{aligned} \tag{6.46}$$

This is exactly the convergence of the 'new column' required for *Step 1*.

***Step 2*: The rank of V_{n+1}** The last two steps follow fairly quickly. Recall that V_{n+1} is the *previsible* running span of evaluation points defined in (6.14). Since it is previsible, it only includes gradients up to time n and the Gram-Schmidt candidate (6.15) of its most recent addition is therefore given by

$$\tilde{\mathbf{v}}_n := \nabla \mathbf{f}_N(X_n) - P_{V_n}\nabla \mathbf{f}_N(X_n) = \sum_{k=d_n}^{N-1} \langle \mathbf{w}_i, \nabla \mathbf{f}_N(X_n)\rangle \mathbf{w}_i = \sum_{k=d_n}^{N-1} D_{\mathbf{w}_i}\mathbf{f}_N(X_n)\mathbf{w}_i,$$

where P_{V_n} is the projection onto V_n.

Now we naturally want to analyze the directional derivatives $D_{\mathbf{w}_i}\mathbf{f}_N(X_n)$ which make up $\tilde{\mathbf{v}}_n$. By representation (6.33) and our formula for the conditional covariance (6.44) we have in distribution

$$D_{\mathbf{w}_i}\mathbf{f}_N(X_n) = \underbrace{\mathbb{E}[D_{\mathbf{w}_i}\mathbf{f}_N(X_n) \mid \mathcal{F}_{n-1}]}_{\overset{(6.41)}{=}0} + \sqrt{\tfrac{1}{N}\sigma_{w,N}}Y_{i+1}.$$

We now already see where our law of large numbers is going to come from. But we first need to take a closer look at the residual variance $\sigma^2_{w,N}$ defined in (6.45). We get convergence in probability of the residual variance to a value strictly greater zero

$$\begin{aligned}
\sigma^2_{w,N} &= \big(\kappa_3\big(\tfrac{\|X_n\|^2}{2}, \tfrac{\|X_n\|^2}{2}, \|X_n\|^2\big) - \mathbf{\Sigma}^{w,N}_{[0:n),n}{}^T[\mathbf{\Sigma}^{w,N}_{[0:n)}]^{-1}\mathbf{\Sigma}^{w,N}_{[0:n),n}\big)\\
&\xrightarrow[N\to\infty]{p} \big(\kappa_3\big(\tfrac{\|y_n\|^2}{2}, \tfrac{\|y_n\|^2}{2}, \|y_n\|^2\big) - \Sigma^{w,\infty}_{[0:n),n}{}^T[\Sigma^{w,\infty}_{[0:n)}]^{-1}\Sigma^{w,\infty}_{[0:n),n}\big) =: \sigma^2_{w,\infty} > 0,
\end{aligned} \tag{6.47}$$

using the convergence of the block matrices and the following lemma.

Lemma 6.3.14. *$\Sigma^{w,\infty}_{[0:n)}$ is strictly positive definite and $\sigma^2_{w,\infty} > 0$.*

Proof. We are going to show with a similar argument as in Lemma 6.3.13 that the matrix

$$\Sigma^{w,\infty}_{[0:n]} := \begin{bmatrix} \Sigma^{w,\infty}_{[0:n)} & \Sigma^{w,\infty}_{[0:n),n} \\ \Sigma^{w,\infty}_{[0:n),n}{}^T & \kappa_3\big(\tfrac{\|y_n\|^2}{2}, \tfrac{\|y_n\|^2}{2}, \|y_n\|^2\big) \end{bmatrix}$$

is strictly positive definite. Before we do so, let us quickly argue why this finishes the proof. Since $\Sigma^{w,\infty}_{[0:n]}$ is then strictly positive definite, it has a cholesky decomposition L such that

$$\Sigma^{w,\infty}_{[0:n]} = LL^T = \begin{bmatrix} L_n & 0 \\ l^T & \sigma \end{bmatrix}\begin{bmatrix} L_n^T & l \\ l & \sigma \end{bmatrix},$$

which implies that L_n is the cholesky decomposition of $\Sigma^{w,\infty}_{[0:n)}$, $l = L_n^{-1}\Sigma^{w,\infty}_{[0:n),n}$ and

$$\sigma = \sqrt{\kappa_3\left(\tfrac{\|y_n\|^2}{2}, \tfrac{\|y_n\|^2}{2}, \|y_n\|^2\right) - \Sigma^{w,\infty}_{[0:n),n}{}^T[\Sigma^{w,\infty}_{[0:n)}]^{-1}\Sigma^{w,\infty}_{[0:n),n}} = \sigma_{w,\infty}.$$

Since we have

$$\det(L) = \sigma_{w,\infty}\det(L_n),$$

strict positive definiteness of $\Sigma^{w,\infty}_{[0:n]}$ and therefore $0 \neq \det(\Sigma^{w,\infty}_{[0:n]}) = \det(L)^2$ implies

$$\sigma^2_{w,\infty} > 0 \quad \text{and} \quad \det(L_n) \neq 0.$$

But $\det(L_n) \neq 0$ also implies that $\Sigma^{w,\infty}_{[0:n)}$ has to be strictly positive definite, since L_n is its cholesky decomposition.

What is therefore left to prove is the strict positive definiteness of $\Sigma^{w,\infty}_{[0:n]}$. For this note that

$$\mathbf{\Sigma}^{w,N}_{[0:n]} := \begin{bmatrix} \mathbf{\Sigma}^{w,N}_{[0:n)} & \mathbf{\Sigma}^{w,N}_{[0:n),n} \\ \mathbf{\Sigma}^{w,N}_{[0:n),n}{}^T & \kappa_3\left(\tfrac{\|X_n\|^2}{2}, \tfrac{\|X_n\|^2}{2}, \|X_n\|^2\right) \end{bmatrix}$$

is the plug-in covariance matrix of $D_{\mathbf{w}_i}\mathbf{f}_N(X_{[0:n]})$. Where we use the term "plug-in" covariance to say that we treat the evaluation points $X_{[0:n]}$ and the direction $\mathbf{w}_i$ as deterministic. Since $X_{[0:n]}$ are contained in V_n, we again map them isometrically to $\mathbb{R}^{d_n}$. But this time we view $\mathbb{R}^{d_n}$ as a subspace of $\mathbb{R}^{d_n+1}$ and map the additional vector $\mathbf{w}_i$ to e_{d_n+1} such that $\mathbf{\Sigma}^{w,N}_{[0:n]}$ is a covariance matrix of $\nabla\mathbf{f}_N$ in $\mathbb{R}^{d_n+1}$.

Now we finish the proof with the same limiting argument as in Lemma 6.3.13. Since $\nabla\mathbf{f}_N$ is strictly positive definite by assumption, we get that $\Sigma^{w,\infty}_{[0:n]}$ is strictly positive definite as the covariance matrix of $(\partial_{d_n+1}\mathbf{f}_N)$ at the points $y_{[0:n]} = (y_0, \dots, y_n) \subseteq \mathbb{R}^{d_n+1}$ of (Ind-III). These are the limiting representations and non of the points y_k are equal by (Ind-IV). □

Using the convergence of the residual variance $\sigma^2_{w,N} \to \sigma^2_{w,\infty}$ allows us to finally make our law of large numbers argument

$$\|\tilde{\mathbf{v}}_n\|^2 = \sum_{i=d_n}^{N-1}(D_{\mathbf{w}_i}\mathbf{f}_N(X_n))^2 = \frac{\sigma^2_{w,N}}{N}\sum_{i=d_n+1}^{N} Y_i^2 \xrightarrow[N\to\infty]{p} \sigma^2_{w,\infty} > 0. \tag{6.48}$$

This implies that V_n has full rank asymptotically (Ind-II).

Assuming the last gradient is always used (and not just in the asymptotic limit), we can get $d_{n+1} = d_n + 1$ almost surely, since $\sigma^2_{w,N} > 0$ almost surely is sufficient for $\|\tilde{\mathbf{v}}_n\|^2 > 0$ almost surely. The use of the most recent gradient ensures the points X_k are different by an inductive argument similar to Lemma 6.3.9. This in turn ensures, using the strict positive definiteness of $(\mathbf{f}_N, \nabla\mathbf{f}_N)$, that $\sigma^2_{w,N} > 0$ almost surely with a similar argument as in Lemma 6.3.14.

***Step 3*: Convergence of the new corner element** The last step follows immediately by definition of $\mathbf{v}_{d_n} = \frac{\tilde{\mathbf{v}}_n}{\|\tilde{\mathbf{v}}_n\|}$ in Definition 6.3.1 and (6.48)

$$D_{\mathbf{v}_{d_n}}\mathbf{f}_N(X_n) = \left\langle \nabla\mathbf{f}_N(X_n), \tfrac{\tilde{\mathbf{v}}_n}{\|\tilde{\mathbf{v}}_n\|}\right\rangle = \|\tilde{\mathbf{v}}_n\| \xrightarrow[N\to\infty]{p} \sigma_{w,\infty} =: \gamma_n^{(d_n)} > 0.$$

6.4 Outlook

In this section we want to discuss possible generalizations and extensions of our results. Possible generalizations include

G1. A **removal of the strict positive definiteness** assumption. The strict positive definiteness was used in the proof of Theorem 6.3.2 to show that the limiting covariance matrices are invertible. It also ensured the full rank of V_n (Ind-II), which in turn was used with the asymptotic use of the last

gradient to ensure asymptotically different evaluation points (Ind-IV). The asymptotically different evaluation were used in turn to keep the covariance matrices strictly positive definite in the next iteration.

A removal of the strict positive definiteness assumption will therefore remove the need for the assumption of the asymptotic use of the most recent gradient, but it will also lose the implications (Ind-II) and (Ind-IV).

There are two avenues to remove this assumption:

a) the use of *generalized matrix inverses*, since the Gaussian conditional distribution can be formed with these (cf. Theorem 2.A.1). The key would be to show that these generalized inverses would still converge with $N \to \infty$.

b) a *perturbation argument*. One can slightly perturb the function $\mathbf{f}_N$ into $\mathbf{f}_{N,\epsilon}$ such that $(\mathbf{f}_{N,\epsilon}, \nabla \mathbf{f}_{N,\epsilon})$ is strictly positive definite (by adding an independent random function with this property with an ϵ weight). We then have $\lim_{\epsilon \to 0} \mathbf{f}_{N,\epsilon} = \mathbf{f}_N$. The difficulty is now to argue that we can exchange limits, as we are interested in

$$\lim_{N \to \infty} \mathbf{f}_N = \lim_{N \to \infty} \lim_{\epsilon \to 0} \mathbf{f}_{N,\epsilon}$$

but want to consider $\lim_{\epsilon \to 0} \lim_{N \to \infty} \mathbf{f}_{N,\epsilon}$, where we can apply the existing theory for the strict positive definite case to the inner limit.

G2. Removal of the **Gaussian** assumption. Our argument is essentially based on a law of large numbers applied to the norm of gradients. Since uncorrelated directional derivatives (obtained from isotropy) are sufficient for independence in the Gaussian case, the squared directional derivatives are also uncorrelated and the law of large numbers applies to the gradient norms. In the non-Gaussian case, the uncorrelated directional derivatives no longer guarantee uncorrelated squares and a more sophisticated argument is necessary.

G3. Relaxation of the **isotropy** assumption. In view of the general definition of input invariance (Definition 4.2.1) it is natural to ask how small the set of invariant transformations can be before our results break down. In place of rotation invariance, one might consider exchangeable directions for example (which is captured by the set of dimension permutations). Since the partial derivatives would then be exchangeable random variables, their norm should still converge by laws of large numbers for exchangeable random variables. Unfortunately, our custom adapted coordinate system would not work in this more general framework. Thus a different approach for a proof would be necessary.

Perhaps even more interesting than generalizations of our results might be the pursuit of the following extensions

E1. The derivation of more **explicit representations** for the limiting information $\mathscr{I}_n$ (which includes limiting function values $\mathfrak{f}_n = \mathfrak{f}_n(\mathfrak{G}, \mu, \kappa, \lambda)$).

E2. Ideally, the explicit representation E1 would allow for **convergence proofs** and **meta optimization** over the optimizer $\mathfrak{G}$ such that the best optimization algorithm for high dimensional problems can be found.

E3. An analysis of **stochastic gradient descent** and similar stochastic algorithms which do not have access to the true cost but rather noisy evaluations of $\mathbf{f}_N$. Since every evaluation is equipped with independent noise, this analysis is most likely similar but more difficult than the perturbation argument suggested to remove the strict positive definiteness assumption in G1.

E4. A very difficult extension would capture algorithms with **component-wise learning rates**.[27] Since component-wise learning rates cannot easily be expressed in a dimension-free manner, it is unclear how to argue that the

[27] such as Adam by Diederik P. Kingma and Ba, 2015.

algorithm converges in a sense. That is because the convergence of the algorithm follows from the key assumption of continuity in the dimensionless information in our proof. Nevertheless, practical experiments confirm that these algorithms exhibit the same behavior (cf. Adam in Figure 6.1a). Perhaps an analysis of the Hessian with tools from random matrix theory would be suitable for an attack on this problem.

E5. The derivation of **concentration bounds**. Given that the conditional variance is bounded by the unconditional variance which scales with $\frac{1}{N}$, we would expect concentration inequalities to contract exponentially in the dimension N. That is because both Gaussian random variables and Chi2 random variables (the squared gradient norms), contract exponentially in the variance.

Acknowledgements

This research was supported by the RTG 1953, funded by the German Research Foundation (DFG). The authors would like to thank Yan Fyodorov for his hospitality and helpful discussions at King's College. We thank him along with Mark Sellke and Antoine Maillard for their time and valuable insights on spin glasses.

References

Adler, Robert J. and Jonathan E. Taylor (2007). *Random Fields and Geometry*. Springer Monographs in Mathematics. New York, NY: Springer New York. ISBN: 978-0-387-48112-8. DOI: 10.1007/978-0-387-48116-6.

Auffinger, Antonio, Andrea Montanari, and Eliran Subag (2023). "Optimization of Random High-Dimensional Functions: Structure and Algorithms". In: *Spin Glass Theory and Far Beyond*. 5 Toh Tuck Link, Singapore: WORLD SCIENTIFIC, pp. 609–633. ISBN: 9789811273919. DOI: 10.1142/9789811273926_0029. arXiv: 2206.10217 [math.PR].

Bayati, Mohsen and Andrea Montanari (2011). "The Dynamics of Message Passing on Dense Graphs, with Applications to Compressed Sensing". In: *IEEE Transactions on Information Theory* 57.2, pp. 764–785. ISSN: 1557-9654. DOI: 10.1109/TIT.2010.2094817.

Bubeck, Sebastien and Mark Sellke (2021). "A Universal Law of Robustness via Isoperimetry". In: *Advances in Neural Information Processing Systems*. Vol. 34. Virtual Event: Curran Associates, Inc., pp. 28811–28822. arXiv: 2105.12806 [cs, stat]. URL: https://proceedings.neurips.cc/paper/2021/hash/f197002b9a0853eca5e046d9ca4663d5-Abstract.html (visited on 2023-09-22).

Celentano, Michael, Andrea Montanari, and Yuchen Wu (2020). *The Estimation Error of General First Order Methods*. DOI: 10.48550/arXiv.2002.12903. arXiv: 2002.12903 [stat]. Pre-published.

Choromanska, Anna et al. (2015). "The Loss Surfaces of Multilayer Networks". In: *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*. Artificial Intelligence and Statistics. San Diego, CA, USA: PMLR, pp. 192–204. arXiv: 1412.0233 [cs]. URL: https://proceedings.mlr.press/v38/choromanska15.html (visited on 2023-01-23).

Dauphin, Yann N et al. (2014). "Identifying and Attacking the Saddle Point Problem in High-Dimensional Non-Convex Optimization". In: *Advances in Neural Information Processing Systems*. Vol. 27. Montréal, Canada: Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper/2014/hash/17e23e50bedc63b4095e3d8204ce063b-Abstract.html (visited on 2022-06-10).

Deift, Percy and Thomas Trogdon (2021). "The Conjugate Gradient Algorithm on Well-Conditioned Wishart Matrices Is Almost Deterministic". In: *Quarterly of Applied Mathematics* 79.1, pp. 125–161. ISSN: 0033-569X, 1552-4485. DOI: 10.1090/qam/1574. arXiv: 1901.09007 [cs, math].

Glorot, Xavier and Yoshua Bengio (2010). "Understanding the Difficulty of Training Deep Feedforward Neural Networks". In: *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*. Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics. Sardinia, Italy: JMLR Workshop and Conference Proceedings, pp. 249–256. URL: https://proceedings.mlr.press/v9/glorot10a.html (visited on 2023-04-11).

Hestenes, Magnus R. and Eduard Stiefel (1952). "Methods of Conjugate Gradients for Solving Linear Systems". In: *Journal of Research of the National Bureau of Standards* 49.6, pp. 409–436. DOI: 10.6028/jres.049.044..

Huang, Brice and Mark Sellke (2022). "Tight Lipschitz Hardness for Optimizing Mean Field Spin Glasses". In: *2022 IEEE 63rd Annual Symposium on Foundations of Computer Science (FOCS)*. 2022 IEEE 63rd Annual Symposium on Foundations of Computer Science (FOCS), pp. 312–322. DOI: 10.1109/FOCS54457.2022.00037.

— (2024). *A Constructive Proof of the Spherical Parisi Formula*. DOI: 10.48550/arXiv.2311.15495. arXiv: 2311.15495 [cond-mat, physics:math-ph]. Pre-published.

Javanmard, Adel and Andrea Montanari (2013). "State Evolution for General Approximate Message Passing Algorithms, with Applications to Spatial Coupling". In: *Information and Inference: A Journal of the IMA* 2.2, pp. 115–144. ISSN: 2049-8764. DOI: 10.1093/imaiai/iat004.

Kingma, Diederik P. and Jimmy Ba (2015). "Adam: A Method for Stochastic Optimization". In: *Proceedings of the 3rd International Conference on Learning Representations*. ICLR. San Diego. arXiv: 1412.6980.

Klein, Aaron et al. (2017). "Learning Curve Prediction with Bayesian Neural Networks". In: International Conference on Learning Representations. URL: https://openreview.net/forum?id=S11KBYclx (visited on 2024-11-27).

Klenke, Achim (2014). *Probability Theory: A Comprehensive Course*. Universitext. London: Springer. ISBN: 978-1-4471-5360-3 978-1-4471-5361-0. DOI: 10.1007/978-1-4471-5361-0.

LeCun, Yann, Corinna Cortes, and Christopher J.C. Burges (2010). *THE MNIST DATABASE of Handwritten Digits*. URL: http://yann.lecun.com/exdb/mnist/ (visited on 2024-01-24).

Nesterov, Yurii Evgen'evič (2018). *Lectures on Convex Optimization*. Second edition. Springer Optimization and Its Applications; Volume 137. Cham: Springer. ISBN: 978-3-319-91578-4. DOI: 10.1007/978-3-319-91578-4.

Panchenko, Dmitry (2013). *The Sherrington-Kirkpatrick Model*. Springer Monographs in Mathematics. New York, NY: Springer. ISBN: 978-1-4614-6288-0 978-1-4614-6289-7. DOI: 10.1007/978-1-4614-6289-7.

Paquette, Courtney et al. (2022). "Halting Time Is Predictable for Large Models: A Universality Property and Average-Case Analysis". In: *Foundations of Computational Mathematics* 23.2, pp. 597–673. ISSN: 1615-3383. DOI: 10.1007/s10208-022-09554-y. arXiv: 2006.04299 [math, stat].

Paquette, Elliot and Thomas Trogdon (2022). "Universality for the Conjugate Gradient and MINRES Algorithms on Sample Covariance Matrices". In: *Communications on Pure and Applied Mathematics* 76.5, pp. 1085–1136. ISSN: 1097-0312. DOI: 10.1002/cpa.22081.

Parisi, G. (1980). "A Sequence of Approximated Solutions to the S-K Model for Spin Glasses". In: *Journal of Physics A: Mathematical and General* 13.4, p. L115. ISSN: 0305-4470. DOI: 10.1088/0305-4470/13/4/009.

Pascanu, Razvan, Yann N. Dauphin, et al. (2014). *On the Saddle Point Problem for Non-Convex Optimization*. DOI: 10.48550/arXiv.1405.4604. arXiv: 1405.4604. Pre-published.

Pascanu, Razvan, Tomas Mikolov, and Yoshua Bengio (2013). "On the Difficulty of Training Recurrent Neural Networks". In: *Proceedings of the 30th International Conference on Machine Learning*. International Conference on Machine Learning. Atlanta: PMLR, pp. 1310–1318. URL: https://proceedings.mlr.press/v28/pascanu13.html (visited on 2024-04-02).

Pedregosa, Fabian and Damien Scieur (2020). "Acceleration through Spectral Density Estimation". In: *Proceedings of the 37th International Conference on Machine Learning*. International Conference on Machine Learning. Virtual Event (formerly Vienna): PMLR, pp. 7553–7562. URL: https://proceedings.mlr.press/v119/pedregosa20a.html (visited on 2023-11-09).

Polyak, Boris T. (1964). "Some Methods of Speeding up the Convergence of Iteration Methods". In: *USSR Computational Mathematics and Mathematical Physics* 4.5, pp. 1–17. ISSN: 0041-5553. DOI: 10.1016/0041-5553(64)90137-5.

Rasmussen, Carl Edward and Christopher K.I. Williams (2006). *Gaussian Processes for Machine Learning*. 2nd ed. Adaptive Computation and Machine Learning 3. Cambridge, Massachusetts: MIT Press. 248 pp. ISBN: 0-262-18253-X. URL: http://gaussianprocess.org/gpml/chapters/RW.pdf (visited on 2023-04-06).

Salimans, Tim and Durk P Kingma (2016). "Weight Normalization: A Simple Reparameterization to Accelerate Training of Deep Neural Networks". In: *Advances in Neural Information Processing Systems*. Vol. 29. Barcelona, Spain: Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper/2016/hash/ed265bc903a5a097f61d3ec064d96d2e-Abstract.html (visited on 2023-10-16).

Sasvári, Zoltán (2013). *Multivariate Characteristic and Correlation Functions*. De Gruyter Studies in Mathematics 50. Berlin/Boston: Walter de Gruyter. 376 pp. ISBN: 978-3-11-022399-6.

Schervish, Mark J. (1995). *Theory of Statistics*. Springer Series in Statistics. New York, NY: Springer. ISBN: 978-1-4612-8708-7 978-1-4612-4250-5. DOI: 10.1007/978-1-4612-4250-5.

Schoenberg, Isaac J. (1938). "Metric Spaces and Positive Definite Functions". In: *Transactions of the American Mathematical Society* 44.3, pp. 522–536. URL: https://community.ams.org/journals/tran/1938-044-03/S0002-9947-1938-1501980-0/S0002-9947-1938-1501980-0.pdf (visited on 2024-04-18).

— (1942). "Positive Definite Functions on Spheres". In: *Duke Mathematical Journal* 9.1, pp. 96–108. ISSN: 0012-7094, 1547-7398. DOI: 10.1215/S0012-7094-42-00908-6.

Scieur, Damien and Fabian Pedregosa (2020). "Universal Average-Case Optimality of Polyak Momentum". In: *Proceedings of the 37th International Conference on Machine Learning*. International Conference on Machine Learning. Virtual Event (formerly Vienna): PMLR, pp. 8565–8572. URL: https://proceedings.mlr.press/v119/scieur20a.html (visited on 2023-11-09).

Sellke, Mark (2024a). "Optimizing Mean Field Spin Glasses with External Field". In: *Electronic Journal of Probability* 29 (none), pp. 1–47. ISSN: 1083-6489, 1083-6489. DOI: 10.1214/23-EJP1066.

— (2024b). "STAT 291, Random High-Dimensional Optimization: Landscapes and Algorithmic Barriers". Lecture Notes. URL: https://msellke.com/courses/STAT_291/course_page_website.html (visited on 2024-08-16).

Subag, Eliran (2018). *Free Energy Landscapes in Spherical Spin Glasses*. DOI: 10.48550/arXiv.1804.10576. arXiv: 1804.10576 [math]. Pre-published.

— (2021). "Following the Ground States of Full-RSB Spherical Spin Glasses". In: *Communications on Pure and Applied Mathematics* 74.5, pp. 1021–1044. ISSN: 1097-0312. DOI: 10.1002/cpa.21922.

6.A Appendix

Recall that $\kappa_3 > 0$ was used in (6.28) to prove that the gradient has positive length. This fact follows from strict positive definiteness, which we are now going to prove.

Lemma 6.A.1. *Let κ be a (non-stationary) isotropic covariance kernel valid in all dimensions (equivalently in ℓ^2, cf. Lemma 4.4.3). Let $\mathbf{f} \sim \mathcal{N}(\mu, \kappa)$ and assume $(\mathbf{f}, \nabla\mathbf{f})$ to be strict positive definite, then for any $x \in \mathbb{R}^N$*

$$\kappa_3\big(\tfrac{\|x\|^2}{2}, \tfrac{\|x\|^2}{2}, \|x\|^2\big) > 0.$$

Proof. Since the norm $\|\cdot\|^2$ is rotation invariant, we can assume without loss of generality $x = \lambda e_1$ for the standard basis vector $e_1 \in \mathbb{R}^N$. Since $\partial_2 \mathbf{f}$ is strict positive definite, we know that

$$0 < \operatorname{Var}(\partial_2 \mathbf{f}(x)) = \operatorname{Cov}(\partial_2 \mathbf{f}(\lambda e_1), \partial_2 \mathbf{f}(\lambda e_1)) = \tfrac{1}{N}\kappa_3\big(\tfrac{\|x\|^2}{2}, \tfrac{\|x\|^2}{2}, \|x\|^2\big),$$

where we have used (4.12) of Lemma 4.5.2 and $\|\lambda e_1\| = \|x\|$ in the last equation. □

In the stationary isotropic case, we can say more. We do not only have $\kappa_3 = -C'(0) > 0$, but we have $C'(r) < 0$ for all $r \geq 0$.

Lemma 6.A.2. *Let C be a stationary isotropic covariance kernel valid in all dimensions. If $\mathbf{f} \sim \mathcal{N}(\mu, C)$ is* not *almost surely constant, then $C'(r) < 0$ for all $r \geq 0$.*

Proof. By Corollary 4.6.4 we have $\nu((0,\infty)) > 0$ for the Schoenberg measure ν in Schoenbergs's characterization (4.14). Thus by the continuity of measures there exists $a, b > 0$ such that $\nu([a,b]) > 0$. Then by (4.14) we have

$$\begin{aligned} -C'(r) &= \int_{[0,\infty)} t^2 \exp(-t^2 r)\nu(dt) \\ &\geq \int_{[a,b]} t^2 \exp(-t^2 r)\nu(dt) \\ &\geq \nu([a,b]) \inf_{s\in[a,b]} s^2 \exp(-s^2 r) > 0. \end{aligned}$$

□

Part II

Supervised machine learning

Chapter 7

Random objective from exchangeable data

In this chapter we explain how random objective functions arise from exchangeable data in supervised machine learning. While this motivation is specific to this application, it offers a more elegant rationale compared to the assumption of random objectives out of necessity, given the intractability of worst-case optimization. Additionally, this derivation can inform distributional assumptions about the random objective. In particular we show for a random linear model in Section 7.4 that the stationarity assumption is untenable in contrast to the isotropy assumption.

> *"The biggest lesson that can be read from 70 years of AI research is that general methods that leverage computation are ultimately the most effective, and by a large margin."*
>
> — Richard Sutton, "The Bitter Lesson"

The supervised machine learning task is to find a model[1] ϕ which best predicts the labels Y based on input data X. For this purpose the distance of the prediction $\phi(X)$ from the labels Y is measured by a loss function ℓ. The goal is then to find a model ϕ which minimizes the accumulated cost over the usage time of the model. Assuming the model is used indefinitely after training the cost function[2] is thus given by

$$\mathbf{J}(\phi) = \lim_{n\to\infty} \sum_{k=1}^{n} \ell(\phi(X_k), Y_k).$$

The existence of this limit is usually guaranteed by the (strong) law of large numbers, (S)LLN, and the assumption of independent, identically distributed (iid) data $(Z_k)_{k\in\mathbb{N}} = (X_k, Y_k)_{k\in\mathbb{N}}$. Instead, we will utilize the more general assumption of *exchangeable* data and explain why the iid assumption is inconsistent with the typical choices of the loss function in machine learning.

We recall, that a sequence of random variables is called exchangeable if its distribution is invariant under finite permutation, i.e. the order of the data does not matter. By de Finetti's seminal representation theorem[3] the sequence $(Z_k)_{k\in\mathbb{N}}$ is exchangeable if and only if there exists a random probability measure $\mathbf{P}_Z$ such that, conditional on $\mathbf{P}_Z$, the data $(Z_k)_{k\in\mathbb{N}}$ is iid with distribution $\mathbf{P}_Z$, i.e.

$$\mathbb{P}(Z_k \in A \mid \mathbf{P}_Z) = \mathbf{P}_Z(A).$$

This also implies that the SLLN is applicable conditional on $\mathbf{P}_Z$, i.e.

$$\mathbf{J}(\phi) = \mathbb{E}[\ell(\phi(X_k), Y_k) \mid \mathbf{P}_Z].$$

In essence, de Finetti's representation theorem states that the exchangeable setting always corresponds to the two stage model: First, sample the distribution $\mathbf{P}_Z$, and then, in the second stage, sample iid data $(Z_k)_{k\in\mathbb{N}}$ from this distribution.[4] Observe that unless $\mathbf{P}_Z$ is deterministic, which corresponds to the iid case, the

[1] The model $\phi = \phi_w$ is often parametrized by parameters w, such as the weights of an artificial neural network.

[2] If the model is parametrized, then we define $\mathbf{J}(w) := \mathbf{J}(\phi_w)$.

[3] e.g. Schervish, 1995, Thm. 1.49.

[4] Details in Section 7.3.

[5] as a mnemonic device recall that we denote random functions and now also random measures in bold.

cost function $\mathbf{J}$ is a random function in general.[5] The optimization of random functions is precisely the subject of Bayesian optimization but leaves open the important question: Which distributional assumptions are reasonable for $\mathbf{J}$? I.e. what prior should be chosen?

7.1 Choice of cost function

The most common cost functions in machine learning are the crossentropy for categorical data or generative tasks and the mean squared error (MSE) for regression tasks. In this section we recapitulate that, not only is the MSE effectively a crossentropy loss for a Gaussian model, but both cost functions can be interpreted as maximum likelihood estimators of $\mathbf{P}_Z$.[6] A maximum likelihood estimation (MLE) is in turn effectively a maximum a posteriori estimation (MAP) with a uninformative (uniform) prior on $\mathbf{P}_Z$. In particular this contradicts dirac priors on $\mathbf{P}_Z$, or in other words deterministic $\mathbf{P}_Z$, which corresponds to iid data. A consistent Bayesian model *must* therefore generalize the iid assumption on the data to exchangeablility.

[6] e.g. Goodfellow et al., 2016, Section 5.6.

By Schervish (1995, Theorem 1.31) the posterior distribution is given by

$$\overbrace{\mathbb{P}(\mathbf{P}_Z \in dP_Z \mid Z_1, \dots, Z_n)}^{\text{posterior}} \propto \overbrace{\mathbb{P}(\mathbf{P}_Z \in dP_Z)}^{\text{prior}} \overbrace{\prod_{k=1}^{n} \frac{dP_Z}{d\nu}(Z_k)}^{\text{likelihood}}, \tag{7.1}$$

where ν is a reference measure such that $\mathbb{P}(Z_k \in dz \mid \mathbf{P}_Z) = \mathbf{P}_Z(dz)$ is almost surely continuous w.r.t. ν and we used that the Z_k are iid conditional on $\mathbf{P}_Z$ with distribution $\mathbf{P}_Z$ to factorize the likelihood. Note that the choice of reference measure ν only changes the multiplicative constant of the likelihood. This implies that this choice is irrelevant for the MLE. But since it is not possible to maximize non-discrete distributions without a reference measure, we cannot directly maximize the posterior to obtain the MAP. And if we construct densities of the posterior to obtain an MAP, the choice of reference measure does affect the outcome.

Bayesians would of course argue against reducing the posterior to a single estimator. And, if you have to choose, Bayesian decision theory asks for a loss function measuring the loss $\ell(\hat{P}_Z, P_Z)$ of selecting $\hat{P}_Z$ when the actual distribution is given by P_Z. The resulting minimization problem is then given by

$$\min_{\hat{P}_Z} \int \ell(\hat{P}_Z, P_Z)\, \mathbb{P}(\mathbf{P}_Z \in dP_Z \mid Z_1, \dots, Z_n).$$

To get back to the intuition of a single most likely P_Z, one can consider the 0-1 loss function $\ell(\hat{P}_Z, P_Z) = \mathbf{1}_{d(\hat{P}_Z, P_Z) > \epsilon}$ and let $\epsilon \to 0$ forcing the choice of the single most likely P_Z. Since ϵ-balls are intimately related to the Hausdorff (Lebesgue) measure, it is perhaps not surprising that the limiting choice can often be shown to be the MAP with the Hausdorff measure as the reference measure.[7]

[7] Bassett and Deride, 2019.

While the Hausdorff measure does not always exist on the space of $\mathbf{P}_Z$, in particular if Z has continuous components,[8] uniform priors also often do not exist as measures and are taken as improper priors.[9]

[8] Gelfand and Vilenkin, 1964, Sec. 5.3, Thm. 4 and following.

[9] A solution to this problem may be to replace the 'countable additivity' requirement on measures with 'finite additivity' (Schervish, 1995, p. 21).

Regardless, assuming a suitable (Hausdorff) measure exists to convert the distribution of $\mathbf{P}_Z$ in (7.1) into a density p, we can maximize the posterior to obtain the MAP. Since the logarithm and the multiplication with $\frac{1}{n}$ are strictly monotone transformations, the MAP is equivalently given by

$$\text{MAP} = \arg\min_{P_Z} \underbrace{-\frac{1}{n} \log(p(P_Z))}_{\text{prior}} + \underbrace{\frac{1}{n} \sum_{k=1}^{n} -\log\Big(\frac{dP_Z}{d\nu}(Z_k)\Big)}_{=H^\nu(\mathbf{P}_Z^n, P_Z)},$$

where $\mathbf{P}_Z^n = \frac{1}{n}\sum_{k=1}^n \delta_{Z_k}$ is the empirical measure and the ν-crossentropy is given by

$$H^\nu(\mu_1, \mu_2) := \int -\log\Big(\frac{d\mu_2}{d\nu}(z)\Big)\mu_1(dz).$$

Due to $H^{\tilde{\nu}}(\mu_1, \mu_2) = H^{\nu}(\mu_1, \mu_2) + H^{\tilde{\nu}}(\mu_1, \nu)$ the minimization of the crossentropy in μ_2 is independent of the reference measure ν. Further observe that the influence of the prior decays as more data is collected ($n \to \infty$). But note that the prior must have support on the space P_Z is selected from. This rules out dirac priors for $\mathbf{P}_Z$ and therefore the iid setting. Uninformative (uniform) priors imply that the MAP is equivalent to the MLE and the MLE is always equivalent to the minimization of the crossentropy.

The crossentropy derived above does not yet match the crossentropy used in supervised machine learning. The reason for this is that in supervised learning we only wish to learn the conditional distribution $\mathbf{P}_{Y|X}$ with

$$\mathbf{P}_Z(dz) = \mathbf{P}_{Y|X}(dy \mid x)\mathbf{P}_X(dx),$$

where $\mathbf{P}_X(dx) = \mathbf{P}_Z(dx \times \mathbb{Y}) = \mathbb{P}(X_k \in dx \mid \mathbf{P}_Z)$ is the marginal distribution of X_k conditional on $\mathbf{P}_Z$. So we only model the conditional distribution $P^{\phi}_{Y|X}$ parametrized by a model ϕ and turn this into a distribution over Z by selecting an arbitrary distribution ν_X for X. This can be something reasonable like $\mathbf{P}_X$ or its empirical version $\mathbf{P}^n_X$, but we will see that this selection has no effect on optimization. Assuming $P^{\phi}_{Y|X}$ is absolutely continuous w.r.t ν_Y for every x, we can plug the full distribution

$$P^{\phi}_Z(dz) = P^{\phi}_{Y|X}(dy \mid x)\nu_X(dx)$$

into the ν-crossentropy with reference measure $\nu = \nu_Y \times \nu_X$ to obtain

$$H^{\nu}(\mathbf{P}^n_Z, P^{\phi}_Z) = \frac{1}{n}\sum_{k=1}^{n} -\log\Big(\frac{P^{\phi}_{Y|X}}{d\nu_Y}(Y_k \mid X_k)\Big) + \underbrace{\frac{1}{n}\sum_{k=1}^{n} -\log\Big(\underbrace{\frac{d\nu_X}{d\nu_X}(X_k)}_{=1}\Big)}_{=0}.$$

7.1.1 Generative models and classification

Generative models are supposed to sample (e.g. images) from a distribution $\mathbf{P}_{Y|X}(dy \mid x)$ over the domain $\mathbb{Y}$ (of images) that depends on a prompt x. This is structurally similar to classification where labels Y must be predicted based on input data x. But in classification tasks the set of possible labels $\mathbb{Y}$ is generally a finite set of categories. The model for this conditional distribution is in both cases typically set up like this: A model ϕ maps inputs x from $\mathbb{X}$ to their 'energy' or 'logits' in $\mathbb{R}^{\mathbb{Y}}$. These are then turned into a conditional distribution via a 'softmax layer' with inverse temperature β

$$p^{\phi}_{Y|X}(y \mid x) = \frac{\exp\big(\beta\, \phi(x)[y]\big)}{\sum_{j\in\mathbb{Y}} \exp\big(\beta\, \phi(x)[j]\big)},$$

where $p^{\phi}_{Y|X}$ is the density of $P^{\phi}_{Y|X}$ w.r.t. the counting measure over the set of possible labels. For generative models with infinite sample space $\mathbb{Y}$, the sum in the denominator (known as the 'partition function') turns into a general integral and becomes intractable. Methods to treat this case are known as 'energy based models'.[10] Since the model typically includes parameters to determine the scale, β might seem superfluous. However, the temperature β is often used after training to tune 'creativity'.[11] During training however, we can assume without loss of generality $\beta = 1$.

Using the counting measure for ν_Y, the sample cost induced by the crossentropy is then given by

$$\begin{aligned}\mathbf{J}_n(\phi) &= H^{\nu}(\mathbf{P}^n_Z, P^{\phi}_Z)\\ &= -\frac{1}{n}\sum_{k=1}^{n} \phi(X_k)[Y_k] + \frac{1}{n}\sum_{k=1}^{n} \log\Big(\sum_{y\in\mathbb{Y}} \exp\big(\phi(X_k)[y]\big)\Big).\end{aligned}$$

[10] see e.g. LeCun et al., 2007; Song and Kingma, 2021.

[11] With $\beta \to 0$ the softmax distribution approaches to the uniform distribution and thus becomes more 'creative'. Whereas with $\beta \to \infty$ it concentrates on the most probable classes and approaches the 'argmax', increasing its safety/reproducibility.

The long term cost is thus given by

$$\begin{aligned}\mathbf{J}(\phi) &= \lim_{n\to\infty} \mathbf{J}_n(\phi) = H^\nu(\mathbf{P}_Z, P_Z^\phi)\\ &= \mathbb{E}\Big[-\phi(X)[Y] \mid \mathbf{P}_Z\Big] + \mathbb{E}\Big[\log\Big(\sum_{y\in\mathbb{Y}} \exp\big(\phi(X)[y]\big)\Big) \mid \mathbf{P}_Z\Big],\end{aligned}$$

where we dropped the indices of (X, Y) since they are iid conditional on $\mathbf{P}_Z$.

7.1.2 Regression

In regression tasks, the space of labels $\mathbb{Y}$ is assumed to be a (Banach) vector space and for our purposes given by $\mathbb{Y} = \mathbb{R}^d$. While an expectation (averaging) would have been ill defined in the categorical case[12], this setting allows for the definition of a *target function* representing the best prediction of the label given the input x for a given relation $\mathbf{P}_Z$

$$\mathbf{f}(x) := \mathbb{E}[Y \mid X = x, \mathbf{P}_Z].$$

The model ϕ attempts to approximate this target function $\mathbf{f}$ and it is reasonable to model the conditional distribution with Gaussian noise around ϕ, i.e.

$$P_{Y|X}^{\phi,\sigma^2}(dy \mid x) = \mathcal{N}\big(\phi(x), \sigma^2\mathbb{I}\big)(dy).$$

We emphasize that this merely implies fitting a Gaussian model to the data without making a Gaussian assumptions about the true distribution.[13] With the Lebesgue measure as the reference measure ν_Y it is straightforward to verify, that minimization of the empirical or limiting MSE

$$\mathbf{J}_n(\phi) = \frac{1}{n}\sum_{k=1}^n \|\phi(X_k) - Y_k\|^2 \quad \overset{n\to\infty}{\longrightarrow} \quad \mathbf{J}(\phi) = \mathbb{E}\big[\|\phi(X) - Y\|^2 \mid \mathbf{P}_Z\big],$$

over the model ϕ is equivalent to minimizing the crossentropy between the empirical distribution $\mathbf{P}_Z^n = \frac{1}{n}\sum_{k=1}^n \delta_{Z_k}$ and the Gaussian model $P_{Y|X}^{\phi,\sigma^2}$ for the empirical MSE

$$H^\nu(\mathbf{P}_Z^n, P_Z^{\phi,\sigma^2}) = \frac{1}{2}\log(2\pi\sigma^2) + \frac{1}{n}\sum_{k=1}^n \frac{\|\phi(X_k) - Y_k\|^2}{2\sigma^2},$$

and equivalent to minimizing $H^\nu(\mathbf{P}_Z, P_Z^{\phi,\sigma^2})$ resp. for the limiting MSE. This is possible since in both cases the choice of σ^2 has no influence on the optimization over ϕ. Thus one can optimize over the model ϕ first to obtain ϕ^* and then analytically determine that $\sigma^2 = \mathbf{J}_n(\phi^*)$ (or $\sigma^2 = \mathbf{J}(\phi^*)$ resp.) is the minimizing choice for the variance σ^2 in the distributional model. The MSE can therefore be interpreted as the variance of the fitted Gaussian model.

7.2 Noise distribution

Observe that the stochastic losses $\ell(\phi(X), Y)$ are exactly of the form as the noisy observations in Definition 2.2.2. Specifically, the noise is given by

$$\varsigma_k(\phi) := \ell(\phi(X_k), Y_k) - \mathbf{J}(\phi)$$

such that the observations are given by

$$\ell_k(\phi) := \ell(\phi(X_k), Y_k) = \mathbf{J}(\phi) + \varsigma_k(\phi).$$

Clearly, the $\varsigma_k(\phi)$ are centered and identically distributed conditional on $\mathbf{P}_Z$.

[12] (Banach) vector spaces are the most general target space for integrals to be well defined (cf. Bochner or Pettis integral).

[13] Of course the quality of the fitted Gaussian model is best when the noise ς on the target function $\mathbf{f}$ is Gaussian. Indeed, considering linear models ϕ, Feng et al. (2024) find that the performance of the MSE minimizer deteriorates with non-Gaussian noise. They further device a method to automatically adapt the loss function (and thus the implicit model) to the noise distribution.

7.3 Parametric statistical model

We have motivated the random measure $\mathbf{P}_Z$ from the axiomatic assumption of exchangable random variables and de Finetti's theorem,[14] which essentially states that the exchangable case corresponds to the two stage model where we first sample $\mathbf{P}_Z$ and then iid data from $\mathbf{P}_Z$.

A different, equivalent way to set this up is to start with a parametric family of probability distributions, a 'statistical model'

$$\mathcal{P} = \{P_{\mathfrak{m}} : \mathfrak{m} \in \mathbb{M}\}.$$

Then one can choose a random variable $\mathbf{M}$ in $\mathbb{M}$ such that $P_{\mathbf{M}}$ becomes a random measure. To ensure no measurability problems arise here we may assume the statistical model is of the form $P_{\mathfrak{m}}(B) = \kappa(\mathfrak{m}; B)$ for some probability kernel κ. One may then select $(Z_k)_{k\in\mathbb{N}}$ iid conditional on $\mathbf{M}$ such that

$$\mathbb{P}(Z_k \in B \mid \mathbf{M}) = \kappa(\mathbf{M}; B) = P_{\mathbf{M}}(B).$$

But then the Z_k are clearly exchangeable and $P_{\mathbf{M}} = \mathbf{P}_Z$ and the non-parametric version follows from the parametric setup.

Similarly, we can turn the non-parametric setting into a parametric setting. Simply let $\mathbb{M}$ be the space of measures, define $P_{\mathfrak{m}} := \mathfrak{m}$ and let $\mathbf{M} = \mathbf{P}_Z$.

Observe that the parametric model circumvented the measure theoretical problem of proving the existence of a kernel κ with $\kappa(\mathbf{P}_Z; B) = \mathbf{P}_Z(B)$ by taking it as definition. And similarly it is a bit easier to prove statements that hold for all $\mathfrak{m} \in \mathbb{M}$ by proving things about $P_{\mathfrak{m}}$. For this reason this notation is used in Chapter 8.

The disadvantage of this parametric model is the additional notation that is introduced without a clear interpretation of what this variable $\mathfrak{m}$ actually represents.

[14] e.g. Schervish, 1995, Thm. 1.49.

7.4 Random linear model

Using a random linear model as an example we will show in this section, that the cost function typically cannot be assumed to be stationary. Isotropy however remains a reasonable assumption for cost functions.

We assume that X is sampled according to some probability distribution $\mathbb{P}_X$ and the labels $Y = \mathbf{f}(X)$ are the output of a random linear target function $\mathbf{f} = f_{\mathbf{M}}$. That is we define the parametric statistical model

$$P_{\mathfrak{m}}(dx, dy) = \mathbb{P}_X(dx)\, \delta_{f_{\mathfrak{m}}(x)}(dy)$$

and sample X_k, Y_k iid from $P_{\mathbf{M}}$, where $\mathbf{M} = (\mathbf{m}, \mathbf{b})$ defines the random linear target function

$$\mathbf{f}(x) = f_{\mathbf{M}}(x) = \mathbf{m}^T x + \mathbf{b}.$$

Using a linear model for regression

$$\phi_w(x) = w^T x + \beta,$$

where $w = (w, \beta)$ with some abuse of notation, the mean squared error cost function is then of the form

$$\begin{aligned}
\mathbf{J}(w) &= \mathbb{E}[(\phi_w(x) - \mathbf{f}(x))^2 \mid \mathbf{M}] \\
&= \mathbb{E}\Big[\big((w - \mathbf{m})^T X + (\beta - \mathbf{b})\big)^2 \mid \mathbf{M}\Big] \\
&= (w - \mathbf{m})^T \underbrace{\mathbb{E}[XX^T]}_{=\mathbb{I}} (w - \mathbf{m}) + (w - \mathbf{m})^T \underbrace{\mathbb{E}[X]}_{=0} (\beta - \mathbf{b}) + (\beta - \mathbf{b})^2 \\
&= \|w - \mathbf{m}\|^2 + (\beta - \mathbf{b})^2 \\
&= \|(w, \beta) - (\mathbf{m}, \mathbf{b})\|^2,
\end{aligned}$$

assuming 'whitened'[15] input in the third equation. Via some more abuse of notation, i.e. $\mathbf{m} := (\mathbf{m}, \mathbf{b})$ and $w = (w, \beta)$ we disregard the bias without loss of generality. Therefore we have

[15] i.e. centered, decorrelated and standardized input X such that $\mathbb{E}[X] = 0$ and $\mathbb{E}[XX^T] = \mathbb{I}$.

$$\mathbf{J}(w) = \|w - \mathbf{m}\|^2 = \underbrace{\overbrace{\|w\|^2 + \mathbb{E}\|\mathbf{m}\|^2}^{\text{deterministic}} - 2\langle w, \mathbf{m}\rangle}_{=:\tilde{\mathbf{J}}(w)} + \underbrace{\|\mathbf{m}\|^2 - \mathbb{E}\|\mathbf{m}\|^2}_{\text{random const.}}$$

For the purpose of optimization, constants are irrelevant and we may therefore pretend $\|\mathbf{m}\|^2 - \mathbb{E}\|\mathbf{m}\|^2 = 0$. Observe that by a change of the origin to $\mathbb{E}[\mathbf{m}]$ we may assume $\mathbb{E}[\mathbf{m}] = 0$, since $\mathbf{J}(w + \mathbb{E}[\mathbf{m}]) = \|w - (\mathbf{m} - \mathbb{E}[\mathbf{m}])\|$. The random function $\tilde{\mathbf{J}}$ is therefore centered with covariance function

$$\mathcal{C}_{\tilde{\mathbf{J}}}(w_1, w_2) = 4\mathbb{E}[\langle w_1, \mathbf{m}\rangle\langle \mathbf{m}, w_2\rangle] = 4w_1^T\mathbb{E}[\mathbf{m}\mathbf{m}^T]w_2.$$

And for $\mathbf{m} \sim \mathcal{N}(0, \sigma^2\mathbb{I})$ we further have that $\tilde{\mathbf{J}}$ is a Gaussian random function with

$$\begin{aligned} \mathbb{E}[\tilde{\mathbf{J}}(w)] &= c + \|w\|^2 && \text{(mean)} \\ \mathcal{C}_{\tilde{\mathbf{J}}}(w_1, w_2) &= 4\sigma^2\langle w_1, w_2\rangle. && \text{(covariance function)} \end{aligned}$$

where $c = \mathbb{E}\|\mathbf{m}\|^2$. Observe that the random function $\tilde{\mathbf{J}}$ is clearly non-stationary but still isotropic.

Proposition 7.4.1 (Distribution of the noise)**.** *In the bias free linear model with $\mathbf{m} \sim \mathcal{N}(0, \sigma^2\mathbb{I}_{d\times d})$ for whitened and Gaussian input $X \sim \mathcal{N}(0, \mathbb{I}_{d\times d})$ the noise*

$$\varsigma(w) := \ell(\phi_w(X), Y) - \mathbf{J}(w)$$

is centered, i.e. $\mathbb{E}[\varsigma(w) \mid \mathbf{m}] = 0$, with covariance

$$\begin{aligned} \mathbb{E}[\varsigma(w)\varsigma(\tilde{w}) \mid \mathbf{m}] &= 2\langle w - \mathbf{m}, \tilde{w} - \mathbf{m}\rangle^2 \\ \mathbb{E}[\varsigma(w)\varsigma(\tilde{w})] &= 2\big(\langle w, \tilde{w}\rangle + \sigma^2 d\big)^2. \end{aligned}$$

Proof. The noise is given by

$$\begin{aligned} \varsigma(w) &= \ell(\phi_w(X), Y) - \mathbf{J}(w) \\ &= (w^T X - \mathbf{m}^T X)^2 - \|w - \mathbf{m}\|^2 \\ &= (w - \mathbf{m})(XX^T - \mathbb{I})(w - \mathbf{m}). \end{aligned}$$

Since the noise is always centered by definition, we skip right to the calculation of the covariance

$$\begin{aligned} \mathbb{E}[\varsigma(w)\varsigma(\tilde{w}) \mid \mathbf{m}] &= \sum_{i,j=1}^{d}\sum_{l,k=1}^{d} \mathbb{E}\Big[(w - \mathbf{m})_i(X_iX_j - \delta_{ij})(w - \mathbf{m})_j \\ &\qquad \cdot (\tilde{w} - \mathbf{m})_k(X_kX_l - \delta_{kl})(\tilde{w} - \mathbf{m})_l \mid \mathbf{m}\Big] \\ &= \sum_{i,j,l,k=1}^{d} (w - \mathbf{m})_i(w - \mathbf{m})_j(\tilde{w} - \mathbf{m})_k(\tilde{w} - \mathbf{m})_l \\ &\qquad \cdot \mathbb{E}\big[(X_iX_j - \delta_{ij})(X_kX_l - \delta_{kl}) \mid \mathbf{m}\big] \end{aligned}$$

Since $X \sim \mathcal{N}(0, \mathbb{I}_{d\times d})$, independent of $\mathbf{m}$, a case by case analysis results in

$$\begin{aligned} &\mathbb{E}\big[(X_iX_j - \delta_{ij})(X_kX_l - \delta_{kl}) \mid \mathbf{m}\big] = \mathbb{E}[(X_iX_j - \delta_{ij})(X_kX_l - \delta_{kl})] \\ &= \begin{cases} \mathbb{E}[(X_i^2 - 1)^2] & = \mathrm{Var}(X_i^2) \quad = 2, & i = j = l = k \\ \mathbb{E}[X_i^2X_j^2] & = \mathbb{E}[X_i^2]\mathbb{E}[X_j^2] \quad = 1, & [i = k \neq j = l] \text{ or } [i = l \neq j = k] \\ 0, & & \text{else} \end{cases} \\ &= \delta_{ik}\delta_{jl} + \delta_{il}\delta_{jk}. \end{aligned}$$

With $\mathbf{m} \sim \mathcal{N}(0, \sigma^2 \mathbb{I})$ this implies

$$\begin{aligned}
&\mathbb{E}[\varsigma(w)\varsigma(\tilde{w}) \mid \mathbf{m}] \\
&= \sum_{i,j,l,k=1}^{d} (w-\mathbf{m})_i (w-\mathbf{m})_j (\tilde{w}-\mathbf{m})_k (\tilde{w}-\mathbf{m})_l [\delta_{ik}\delta_{jl} + \delta_{il}\delta_{jk}] \\
&= 2 \sum_{i,j=1}^{d} (w-\mathbf{m})_i (w-\mathbf{m})_j (\tilde{w}-\mathbf{m})_i (\tilde{w}-\mathbf{m})_j \\
&= 2\langle w-\mathbf{m}, \tilde{w}-\mathbf{m}\rangle^2.
\end{aligned}$$

And the unconditional expectation is given by

$$\begin{aligned}
&\mathbb{E}[\varsigma(w)\varsigma(\tilde{w})] \\
&= 2 \sum_{i,j=1}^{d} \mathbb{E}[(w-\mathbf{m})_i (w-\mathbf{m})_j (\tilde{w}-\mathbf{m})_i (\tilde{w}-\mathbf{m})_j] \\
&= 2 \sum_{i,j=1}^{d} \mathbb{E}\Big[\Big(w_i\tilde{w}_i - (w_i+\tilde{w}_i)\mathbf{m}_i + \mathbf{m}_i^2\Big)\Big(w_j\tilde{w}_j - (w_j+\tilde{w}_j)\mathbf{m}_j + \mathbf{m}_j^2\Big)\Big] \\
&= 2 \sum_{i,j=1}^{d} \mathbb{E}\Big[\Big(w_i\tilde{w}_i - (w_i+\tilde{w}_i)\mathbf{m}_i + \mathbf{m}_i^2\Big) \underbrace{\mathbb{E}\Big[w_j\tilde{w}_j - (w_j+\tilde{w}_j)\mathbf{m}_j + \mathbf{m}_j^2 \mid \mathbf{m}_i\Big]}_{=w_j\tilde{w}_j+\sigma^2}\Big] \\
&= 2 \sum_{i,j=1}^{d} (w_i\tilde{w}_i + \sigma^2)(w_j\tilde{w}_j + \sigma^2) \\
&= 2\left(\sum_{i=1}^{d} (w_i\tilde{w}_i + \sigma^2)\right)^2 = 2(\langle w, \tilde{w}\rangle + \sigma^2 d)^2.
\end{aligned}$$

□

References

Bassett, Robert and Julio Deride (2019). "Maximum a Posteriori Estimators as a Limit of Bayes Estimators". In: *Mathematical Programming* 174.1, pp. 129–144. ISSN: 1436-4646. DOI: 10.1007/s10107-018-1241-0.

Feng, Oliver Y. et al. (2024). *Optimal Convex M-Estimation via Score Matching*. DOI: 10.48550/arXiv.2403.16688. arXiv: 2403.16688 [math]. Pre-published.

Gelfand, I. M. and N. Ya Vilenkin (1964). *Generalized Functions, Volume 4: Applications of Harmonic Analysis*. Trans. by Amiel Feinstein. AMS Chelsea Publishing. Providence, Rhode Island: American Mathematical Society. 384 pp. ISBN: 978-1-4704-2662-0. URL: https://lccn.loc.gov/2015040021 (visited on 2023-12-04).

Goodfellow, Ian, Yoshua Bengio, and Aaron Courville (2016). *Deep Learning*. MIT Press. 801 pp. ISBN: 978-0-262-33737-3. Google Books: omivDQAAQBAJ.

LeCun, Yann et al. (2007). "A Tutorial on Energy-Based Learning". In: *Predicting Structured Data*. Cambridge, Massachusetts: MIT Press. ISBN: 978-0-262-52804-7. URL: http://yann.lecun.com/exdb/publis/orig/lecun-06.pdf (visited on 2025-01-09).

Schervish, Mark J. (1995). *Theory of Statistics*. Springer Series in Statistics. New York, NY: Springer. ISBN: 978-1-4612-8708-7 978-1-4612-4250-5. DOI: 10.1007/978-1-4612-4250-5.

Song, Yang and Diederik P. Kingma (2021). *How to Train Your Energy-Based Models*. DOI: 10.48550/arXiv.2101.03288. arXiv: 2101.03288 [cs]. Pre-published.

Chapter 8

Optimization landscape of shallow neural networks

In this chapter we examine the optimization landscape of a broad class of regression problems with squared error loss for shallow ANNs using analytic activation functions. The main finding is that, in an appropriate sense, almost all such problems exhibit a "nice" optimization landscape. More precisely, we show that the cost landscape is typically Morse on the domain of *non-degenerate* parameters. Conversely, the set of *degenerate* parameters – those for which the same response function can be accomplished with fewer neurons – has significantly smaller Hausdorff dimension. Such parameters do not fully exploit the network's representational capacity, and inherently have redundancies that prevent the optimization landscape from being Morse at those points.

In order to state our results we start with a formal definition of shallow neural networks. The definition uses a graph structure that will prove to be useful later.

Definition 8.0.1 (Shallow neural network)**.** A *(dense) shallow neural network* (briefly called *ANN*) is a tuple $\mathfrak{N} = (\mathbb{V}, \psi)$ consisting of

- a tuple $\mathbb{V} = (V_0, V_1, V_2)$ of finite disjoint sets V_0, V_1 and V_2 (the neurons of the input $V_{\text{in}} := V_0$, hidden V_1 and output layer $V_{\text{out}} := V_2$) and
- a measurable function $\psi : \mathbb{R} \to \mathbb{R}$ (the activation function).

Definition 8.0.2. Let $\mathfrak{N} = (\mathbb{V}, \psi)$ be an ANN.

1. We call the directed graph $G = (V, E)$ given by

$$V = V_0 \cup V_1 \cup V_2 \quad \text{and} \quad E = \bigcup_{k=0}^{1} V_k \times V_{k+1}$$

   the *ANN-graph* of $\mathfrak{N}$.

2. We call $\Theta = \Theta_{\mathfrak{N}} = \mathbb{R}^E \times \mathbb{R}^{V_1 \cup V_2}$ *parameter space of the network* $\mathfrak{N}$ and every tuple $\theta = (w, \beta) \in \Theta = \mathbb{R}^E \times \mathbb{R}^{V_1 \cup V_2}$ a *parameter of the network* $\mathfrak{N}$. We refer to w as the (edge) *weights* and to β as the *biases*.

3. For every parameter $\theta = (w, \beta) \in \Theta$ we call

$$\Psi_\theta : \mathbb{R}^{V_{\text{in}}} \to \mathbb{R}^{V_{\text{out}}},\ x \mapsto \Big(\beta_l + \sum_{j \in V_1} \psi\Big(\beta_j + \sum_{i \in V_{\text{in}}} x_i w_{ij}\Big) w_{jl}\Big)_{l \in V_{\text{out}}} \tag{8.1}$$

   the *response function* of the parameter θ.

Typically, the underlying network is clear from the context and it is therefore omitted in the notation.

We study regression problems where the input data lies in $\mathbb{R}^{V_{\text{in}}}$ and the labels in $\mathbb{R}^{V_{\text{out}}}$. These can be formally described by a distribution $\mathbb{P}_X$ on $\mathbb{R}^{V_{\text{in}}}$ (the distribution of the input data) and a probability kernel κ from $\mathbb{R}^{V_{\text{in}}}$ to $\mathbb{R}^{V_{\text{out}}}$ (the

conditional distribution of the label given the input data). Our aim is to show that for a fixed distribution $\mathbb{P}_X$ for "most" kernels κ the respective optimization landscape is Morse on the efficient domain. For this we analyze random regression problems where the kernel in the regression problem itself is random.

Definition 8.0.3. Let $\mathfrak{N}$ be an ANN. A *measurable family of regression problems* is a tuple $\mathfrak{R} = (\mathbb{P}_X, \kappa, \ell)$ consisting of

1. a distribution $\mathbb{P}_X$ on $\mathbb{R}^{V_{\text{in}}}$ (the distribution of the input data X)
2. a measurable set $(\mathbb{M}, \mathcal{M})$ (the statistical model space)
3. a probability kernel κ mapping model and input data from $\mathbb{M} \times \mathbb{R}^{V_{\text{in}}}$ to a probability distribution over labels in $\mathbb{R}^{V_{\text{out}}}$ and
4. a measurable function $\ell : \mathbb{R}^{V_{\text{out}}} \times \mathbb{R}^{V_{\text{out}}} \to [0, \infty]$ (the *loss*).

When dealing with measurable families of regression problems we will always associate the setting with a measurable space that is equipped with a family of distributions $(\mathbb{P}_{\mathfrak{m}})_{\mathfrak{m} \in \mathbb{M}}$ together with a $\mathbb{R}^{V_{\text{in}}}$-valued random variable X (the input data) and a $\mathbb{R}^{V_{\text{out}}}$-valued random variable Y (the label) such that under every distribution $\mathbb{P}_{\mathfrak{m}}$ with $\mathfrak{m} \in \mathbb{M}$, $\mathbb{P}_X$ is the distribution of X and $\kappa(\mathfrak{m}, \cdot; \cdot)$ is the conditional distribution of Y given X, i.e.,

$$\mathbb{P}_{\mathfrak{m}}(Y \in B \mid X) = \kappa(\mathfrak{m}, X; B), \quad \text{a.s.}$$

Then the loss that we incur when using a shallow ANN for the prediction defines a cost landscape in the sense of the following definition.

Definition 8.0.4 (Cost function)**.** Let $\mathfrak{N}$ be an ANN as in Definition 8.0.1 and $\mathfrak{R}$ a measurable family of regression problems as in Definition 8.0.3 and a function $R : \Theta \to \mathbb{R}$ (regularization). The family of functions $(J_{\mathfrak{m}})_{\mathfrak{m} \in \mathbb{M}}$ given by

$$J_{\mathfrak{m}} \colon \Theta \to (-\infty, \infty], \ \theta \mapsto \mathbb{E}_{\mathfrak{m}}\big[\ell(\Psi_\theta(X), Y)\big] + R(\theta)$$

the *(regularized) cost functions of* $(\mathfrak{N}, \mathfrak{R}, R)$.

A useful concept for the analysis of the MSE cost function is the 'target function' representing the best possible predictor.

Definition 8.0.5 (Family of L^p-integrable regression problems, target function)**.** Let $p \in [1, \infty)$. A measurable family of regression problems $\mathfrak{R}$ is said to be *L^p-integrable*, if for every $\mathfrak{m} \in \mathbb{M}$ the label is L^p-integrable, i.e.,

$$\forall \mathfrak{m} \in \mathbb{M} : \ \mathbb{E}_{\mathfrak{m}}\big[\|Y\|^p\big] < \infty.$$

For a family of L^1-integrable regression problems $\mathfrak{R}$, for every $\mathfrak{m} \in \mathbb{M}$ the function

$$f_{\mathfrak{m}}(x) := \int y \, \kappa(\mathfrak{m}, x; dy) \overset{\mathbb{P}_X\text{-a.s.}}{=} \mathbb{E}_{\mathfrak{m}}[Y \mid X = x]$$

is well-defined for $\mathbb{P}_X$-almost all $x \in \mathbb{R}^{V_{\text{in}}}$. We call $f_{\mathfrak{m}}$ the *target function of* $\mathfrak{m}$.

The conditions assumed for our main result are collected in the following definition.

Definition 8.0.6. The *standard setting* is a tuple $(\mathfrak{N}, \mathfrak{R}, R, \mathbf{M})$ consisting of

- an ANN $\mathfrak{N}$ with one dimensional output $\#V_{\text{out}} = 1$ and analytic activation function ψ,
- a family of L^2-integrable regression problems $\mathfrak{R}$ with squared-error loss $\ell(\hat{y}, y) = (\hat{y} - y)^2$ and compact support $\mathcal{X} := \operatorname{supp}(\mathbb{P}_X)$ of the input distribution X,
- analytic convex (regularization) function $R : \Theta \to \mathbb{R}$ and

- an $\mathbb{M}$-valued random variable $\mathbf{M}$ such that the random target function $\mathbf{f} := f_{\mathbf{M}}$ is an *weakly universal Gaussian random function* in the sense that for every continuous test function $\phi : \mathbb{R}^{V_{\text{in}}} \to \mathbb{R}$ the random variable

$$\langle \phi, f_{\mathbf{M}} \rangle_{\mathbb{P}_X} := \int \phi(x)\, f_{\mathbf{M}}(x)\, \mathbb{P}_X(dx)$$

is Gaussian and has strictly positive variance whenever $\phi \not\equiv 0$ on $\mathcal{X}$.[1]

To understand the cost landscape of models $\mathfrak{m} \in \mathbb{M}$ we need to divide the space of all parameters $\theta \in \Theta = \mathbb{R}^E \times \mathbb{R}^{V_1 \cup V_2}$ into two domains. For the activation functions sigmoid and tanh we define

- the *efficient domain* by

$$\mathcal{E}_0 := \left\{ \theta = (w, \beta) \in \mathbb{R}^{E \times (V \setminus V_0)} : \begin{array}{ll} w_{j\bullet} \neq 0 & \forall j \in V_1, \\ w_{\bullet j} \neq 0 & \forall j \in V_1, \\ (w_{\bullet i}, \beta_i) \neq \pm (w_{\bullet j}, \beta_j) & \forall i, j \in V_1 \text{ with } i \neq j \end{array} \right\}, \quad (8.2)$$

- the *redundant domain* by $(\mathbb{R}^E \times \mathbb{R}^{V_1 \cup V_2}) \setminus \mathcal{E}_0$.

Our main structural result, namely the fact that the cost is typically Morse, is only true on the efficient domain. The restriction onto the efficient domain is natural since any redundant parameter lies on a path of constant response such that no local minimum in the redundant domain can be a strict local minimum. In particular the cost cannot be Morse on the whole set of parameters.

Our main result states that for "most" statistical models $\mathfrak{m}$ the realization of the MSE is Morse on the efficient domain.

Theorem 8.0.7 (Almost all optimization landscapes are Morse on the efficient domain)**.** *Let $(\mathfrak{N}, \mathfrak{R}, R, \mathbf{M})$ be a standard setting (Definition 8.0.6). Assume $\psi \in \{\text{sigmoid}, \tanh\}$ about the activation function and that the support of $\mathbb{P}_X$ contains a non-empty open set. Almost surely, the regularized cost $J_{\mathbf{M}} \colon \mathcal{E}_0 \to \mathbb{R}$ with*

$$J_{\mathbf{M}}(\theta) = \mathbb{E}_{\mathbf{M}}\big[\ell(\Psi_\theta(X), Y)\big] + R(\theta)$$

is a Morse function. Equivalently, it holds that

$$\mathbb{P}\Big(\exists \theta \in \mathcal{E}_0 : \nabla J_{\mathbf{M}}(\theta) = 0,\, \det(\nabla^2 J_{\mathbf{M}}(\theta)) = 0\Big) = 0.$$

In Section 8.1 we actually prove a version of this theorem for general analytic activation functions (see Theorem 8.1.2). This requires a notion of the efficient domain that is an implicitly defined set. In Section 8.2 we then prove that for the activation functions sigmoid and tanh the implicitly defined version of the efficient domain agrees with the one used in the latter theorem.

Remark 8.0.8 (Generalization of the Gaussian assumption). While we assume that the target function is weakly universal Gaussian in the standard setting (Definition 8.0.6), our main theorem (Theorem 8.0.7) is a statement about null sets. Since null sets remain null sets for measures that are absolutely continuous with respect to such Gaussian measures and mixtures thereof, it is straight-forward to generalize the statement to significantly more general distributions of the random target function $f_{\mathbf{M}}$. That is, the statement remains true if the distribution of $f_{\mathbf{M}}$ can be written as a mixture of measures that are absolutely continuous with respect to weakly universal Gaussian measures!

Remark 8.0.9 (Weak universality). For a better understanding of weak universality consider the stronger assumption[2] that all continuous functions $\phi \colon \mathbb{R}^{V_{\text{in}}} \to \mathbb{R}^{V_{\text{out}}}$ lie in the support of $\mathbb{P}_{f_{\mathbf{M}}}$, when $f_{\mathbf{M}}$ is a random element in $L^2(\mathbb{P}_X)$. I.e. for every continuous $\phi \colon \mathbb{R}^{V_{\text{in}}} \to \mathbb{R}^{V_{\text{out}}}$ and $\epsilon > 0$, one has that

$$\mathbb{P}\big(\|f_{\mathbf{M}} - \phi\|_{\mathbb{P}_X} < \epsilon\big) > 0. \quad (8.3)$$

This is a *universality* assumption[3] and intuitively means that no continuous function

[1] Note that for every $\mathfrak{m} \in \mathbb{M}$, $f_{\mathfrak{m}}$ is in $L^2(\mathbb{P}_X) \subseteq L^1(\mathbb{P}_X)$ and ϕ is uniformly bounded on the compact support of $\mathbb{P}_X$ so that the integral $\langle \phi, f_{\mathbf{M}} \rangle_{\mathbb{P}_X}$ is for *all* realizations of $\mathbf{M}$ well-defined. It is further measurable since $f_\cdot(\cdot)$ is product measurable by Fubini's theorem.

[2] On the positive probability event in (8.3) we have

$$\begin{aligned} &|\langle \phi, f_{\mathbf{M}} \rangle_{\mathbb{P}_X} - \|\phi\|^2_{\mathbb{P}_X}| \\ &= |\langle \phi, f_{\mathbf{M}} - \phi \rangle_{\mathbb{P}_X}| \\ &\leq \epsilon \|\phi\|_{\mathbb{P}_X}. \end{aligned}$$

Since ϕ is continuous and non-zero on the support of $\mathbb{P}_X$ this implies $\|\phi\|_{\mathbb{P}_X} > 0$. Choosing $\epsilon \in (0, \|\phi\|_{\mathbb{P}_X})$ we conclude that $\langle \phi, f_{\mathbf{M}} \rangle_{\mathbb{P}_X} > 0$ with strictly positive probability. The same argument applied to $-\phi$ gives that $\langle \phi, f_{\mathbf{M}} \rangle_{\mathbb{P}_X} < 0$ with strictly positive probability. Consequently, $\langle \phi, f_{\mathbf{M}} \rangle_{\mathbb{P}_X}$ has positive variance.

[3] Micchelli et al., 2006; Carmeli et al., 2010; Bogachev, 1998, Thm. 3.6.1.

ϕ can be ruled out as the target function $f_{\mathbf{M}}$ ex ante. We believe this is a natural assumption for a learning problem.

In the proof of Theorem 8.0.7, we will actually work with an even weaker assumption than weak universality: It would suffice to assume the non-degeneracy for real-analytic test functions ϕ only.

A natural question is whether local minima on the efficient domain exist and whether the restriction to the efficient domain in Theorem 8.0.7 is an artefact of our proof. This will be the content of Sections 8.3-8.5. Intuitively we show, for the standard unregularized setting with activation $\psi \in \{\text{sigmoid}, \tanh\}$ and the additional regularity assumption (8.3) in Remark 8.0.9, that we have the following:

- For every open set $U \subset \Theta$ containing an efficient point, the probability is strictly positive that the loss has a local minimum in U, see Theorem 8.4.1.
- With strictly positive probability, there exist critical points in the set of redundant parameters (Theorem 8.5.1) and all redundant critical points have a direction of zero curvature (the determinant of the Hessian is zero), see Theorem 8.3.1.

It is therefore *impossible* to prove the MSE to be a Morse function on the redundant domain, since critical points may exist and those always violate the Morse condition.

Outline In Section 8.1 we prove a more general version of Theorem 8.0.7 which is applicable to all analytic activation functions. However, for general analytic activation functions the efficient domain has to be defined in an implicit way. Specifically, we will prove in Theorem 8.1.2 for the standard setting that the cost landscape is almost surely Morse on the set of *polynomially efficient parameters* (Definition 8.1.1). In Section 8.2 we show that the various definitions of efficient parameter domains coincide for $\psi \in \{\text{sigmoid}, \tanh\}$ (Theorem 8.2.3). With this result Theorem 8.0.7 becomes a direct corollary of Theorem 8.1.2. In Section 8.3 we prove for any redundant parameter θ that there exists a straight line of parameters $\theta(t)$, starting in θ, where the response, and therefore the cost, remains unchanged, i.e. $\Psi_{\theta(0)} = \Psi_{\theta(t)}$. In Section 8.4 we prove that efficient local minima exist with positive probability. We use this fact in Section 8.5 to prove that redundant critical points exist with positive probability. To this end we extend an efficient critical parameter of a smaller network to a redundant critical parameter of a larger network.

8.1 MSE is Morse on efficient domain

In this section, we will prove that for a standard model the random loss-landscape is Morse on the *polynomially*, efficient domain. For general activation functions ψ we have to work with a different notion of the efficient domain than $\mathcal{E}_0$ introduced in (8.2). As we will show in Section 8.2 the definition coincides with $\mathcal{E}_0$ whenever $\psi \in \{\text{sigmoid}, \tanh\}$ and the support of $\mathbb{P}_X$ contains an open set.

Definition 8.1.1 (Polynomial efficiency)**.** Let $\mathfrak{N} = (\mathbb{V}, \psi)$ be an ANN (Definition 8.0.1), $n \in \mathbb{N}_0$ and $m = (m_\emptyset, m_0, \dots, m_n) \in \mathbb{N}_0^{n+2}$.

(i) A parameter $\theta \in \Theta$ is called *m-polynomially independent on $\mathcal{X}$* if ψ is n-times differentiable and the equation

$$0 = P^{(\emptyset)}(x) + \sum_{j \in V_1} \sum_{k=0}^{n} P_j^{(k)}(x) \psi^{(k)}\Big(\beta_j + \sum_{i \in V_0} x_i w_{ij}\Big) \quad \forall x \in \mathcal{X}$$

considered in all polynomials $P^{(\emptyset)}$ and $(P_j^{(k)} : j \in V_1, k \in \{0, \dots, n\})$ of at most degree $m_\emptyset$ and m_k, respectively, has only the trivial solution where all polynomials are identically zero. Here, $\psi^{(k)}$ denotes the k-th derivative of the activation function ψ.

(ii) A parameter $\theta \in \Theta$ is called *m-polynomially efficient on $\mathcal{X}$*, if

(a) all neurons are used meaning that for all $k \in V_1$ one has

$$w_{k\bullet} = (w_{kl})_{l \in V_{\text{out}}} \not\equiv 0,$$

and

(b) it is m-polynomially independent.

We denote by $\mathcal{E}_P^m = \mathcal{E}_P^m(\mathcal{X})$ the set of all m-polynomially efficient parameters.

Theorem 8.1.2 (MSE is a Morse function on polynomially efficient parameters)**.** *Let $(\mathfrak{N}, \mathfrak{R}, R, \mathbf{M})$ be the standard setting (Definition 8.0.6). Then the MSE cost is almost surely a Morse function on the set $\mathcal{E}_P := \mathcal{E}_P^{(0,0,1,2)}(\mathcal{X})$ of $(0,0,1,2)$-polynomially efficient parameters on the support $\mathcal{X}$ of $\mathbb{P}_X$, i.e.*

$$\mathbb{P}\Big(\exists \theta \in \mathcal{E}_P : \nabla \mathbf{J}(\theta) = 0, \det(\nabla^2 \mathbf{J}(\theta)) = 0\Big) = 0$$

Before we explain the methodology of our proof we first derive a crucial representation for the MSE cost $J_{\mathfrak{m}}$ given by

$$J_{\mathfrak{m}}(\theta) = \mathbb{E}_{\mathfrak{m}}\big[\|\Psi_\theta(X) - Y\|^2\big] + R(\theta)$$

with convex regularizer R. Recall that Ψ_θ is the realization function of the ANN as introduced in (8.1).

Proposition 8.1.3 (Decomposition of the MSE cost)**.** *For $\mathfrak{m} \in \mathbb{M}$ and $\theta \in \Theta$ one has*

$$J_{\mathfrak{m}}(\theta) = R(\theta) + \|\Psi_\theta\|_{\mathbb{P}_X}^2 - 2\underbrace{\langle \Psi_\theta, f_{\mathfrak{m}} \rangle_{\mathbb{P}_X}}_{=:\hat{J}_{\mathfrak{m}}(\theta)} + \mathbb{E}_{\mathfrak{m}}[\|Y\|^2]. \tag{8.4}$$

Here $\|\cdot\|_{\mathbb{P}_X}$ is induced by $\langle \phi, \varphi \rangle_{\mathbb{P}_X} := \int \langle \phi(x), \varphi(x) \rangle \mathbb{P}_X(dx)$.

Proof. With the Pythagorean formula we have

$$\begin{aligned} J_{\mathfrak{m}}(\theta) &= R(\theta) + \mathbb{E}_{\mathfrak{m}}[\|\Psi_\theta(X) - Y\|^2] \\ &= R(\theta) + \mathbb{E}_{\mathfrak{m}}[\|\Psi_\theta(X)\|^2] - 2\mathbb{E}_{\mathfrak{m}}[\langle \Psi_\theta(X), Y \rangle] + \mathbb{E}_{\mathfrak{m}}[\|Y\|^2]. \end{aligned}$$

Since $f_{\mathfrak{m}}(X) = \mathbb{E}_{\mathfrak{m}}[Y|X]$ we conclude with the tower property

$$\mathbb{E}_{\mathfrak{m}}[\langle \Psi_\theta(X), Y \rangle] = \mathbb{E}_{\mathfrak{m}}[\langle \Psi_\theta(X), f_{\mathfrak{m}}(X) \rangle] \overset{\text{def.}}{=} \langle \Psi_\theta, f_{\mathfrak{m}} \rangle_{\mathbb{P}_X} = \hat{J}_{\mathfrak{m}}(\theta). \qquad \square$$

In view of (8.4) we observe that the term $\mathbb{E}_{\mathfrak{m}}[\|Y\|^2]$ does not depend on the parameter θ and it is thus irrelevant when it comes to deciding whether $J_{\mathfrak{m}}$ is Morse or not. In the proof we then argue that the event where the stochastic process $(R(\theta) + \|\Psi_\theta\|_{\mathbf{P}_X}^2 - 2\hat{J}_{\mathbf{M}}(\theta))_{\theta \in \Theta}$ is not Morse on the polynomially efficient parameters is a *"thin set"*.

Unfortunately, our setting is not immediately covered by the arguments of Adler and Taylor (2007). Roughly speaking, their approach is as follows. If there is a parameter θ that is critical with its Hessian having a zero eigenvalue, then this satisfies

$$\nabla \mathbf{J}(\theta) = 0 \quad \text{and} \quad \det(\nabla^2 \mathbf{J}(\theta)) = 0. \tag{8.5}$$

Note that the latter is a collection of $\dim(\Theta) + 1$ real equations in $\dim(\Theta)$ real variables and intuitively one would expect that, under appropriate non-degeneracy assumptions, the equation does not have solutions. The equations in (8.5) depend on the collection of first order differentials $\mathbf{g}_1(\theta)$ and of second order differentials $\mathbf{g}_2(\theta)$. As shown in Lemma 11.2.10 of Adler and Taylor (2007) solutions of (8.5) would not exist, if for every θ under consideration (in our case the efficient domain) the combined vector $(\mathbf{g}_1(\theta), \mathbf{g}_2(\theta))$ has locally uniformly bounded Lebesgue density.

Unfortunately, in our situation, many second order differentials are degenerate and the result is not applicable. To bypass this problem we proceed as follows. In the following subsection, we will first analyze the stochastic process

$$\hat{\mathbf{J}} = (\hat{J}_{\mathbf{M}}(\theta))_{\theta\in\Theta} = (\langle \Psi_\theta, f_{\mathbf{M}}\rangle_{\mathbb{P}_X})_{\theta\in\Theta}.$$

This process is obtained by applying a θ-dependent linear functional on the random target function $\mathbf{f} = f_{\mathbf{M}}$ and thus $\hat{\mathbf{J}}$ is a Gaussian process since $\mathbf{f}$ is Gaussian by assumption. We will collect in $\mathbf{g}_1(\theta)$ all first order differentials and in $\mathbf{g}_2(\theta)$ the *'centered'*[4] and *non-degenerate* second order differentials. The non-degeneracy of the combined collection $\mathbf{g} = (\mathbf{g}_1, \mathbf{g}_2)$ is shown in Proposition 8.1.5 and follows from the polynomial independence that is assumed in the polynomially efficient domain (cf. Definition 8.1.1).

[4] This is only true if $f_{\mathbf{M}}$ is centered, which we are not willing to assume. But this provides the right intuition, since we subtract a deterministic term (which is not necessarily the mean).

In Proposition 8.1.7 we show that the generalization of the volume argument of Adler and Taylor (2007, Lemma 11.2.10) given in Lemma 8.1.6 is applicable to $\mathbf{g}$. The generalization of the volume argument is necessary since we want to show that the process

$$\big(R(\theta) + \|\Psi_\theta\|^2_{\mathbb{P}_X} - 2\hat{\mathbf{J}}(\theta)\big)_{\theta\in\Theta}$$

never satisfies (8.5) on $\mathcal{E}_P$. While Adler and Taylor (2007) considered level sets, we move the model indepeendent term $R(\theta) + \|\Psi_\theta\|^2_{\mathbb{P}_X}$ to the other side in (8.5) and therefore need to consider function graph intersections. Proposition 8.1.7 would then immediately yield the Morse property if all second order derivatives would be contained in $\mathbf{g}_2(\theta)$.

The subsequent subsection (Section 8.1.2) finishes the proof of Theorem 8.1.2. To do so we carefully craft a thin set U as the zero set of a function F which $\mathbf{g}$ may not intersect. Although $\hat{\mathbf{J}}(\theta)$ has degenerate second order differentials in the last layer, the additional deterministic term $R(\theta) + \|\Psi_\theta\|^2_{\mathbb{P}_X}$ that is strictly convex in the last layer helps us out. More explicitly, we will design a real analytic function F taking an outcome of $(\theta, \mathbf{g}_2(\theta))$ to a real value in such a way that for all $\theta \in \mathcal{E}_P$

$$\nabla\mathbf{J}(\theta) = 0 \implies \det(\nabla^2\mathbf{J}(\theta)) = F(\theta, \mathbf{g}_2(\theta)).$$

Consequently, if there is a parameter $\theta \in \mathcal{E}_P$ with

$$\nabla\mathbf{J}(\theta) = 0 \quad \text{and} \quad \det(\nabla^2\mathbf{J}(\theta)) = 0,$$

then we also found a solution to

$$\mathbf{g}_1(\theta) = 0 \quad \text{and} \quad F(\theta, \mathbf{g}_2(\theta)) = 0.$$

This is again a collection of $\dim(\Theta) + 1$ real equations in $\dim(\Theta)$ variables and we formally conclude with Proposition 8.1.7 that, almost surely, no solutions exist. Note that we are now able to proceed since the latter equations only make use of the non-degenerate differentials of first and second order of $\hat{\mathbf{J}}(\theta)$.

8.1.1 Analysis of $(\hat{\mathbf{J}}(\theta))$

In this section we analyze the stochastic process $(\hat{\mathbf{J}}(\theta))_{\theta\in\Theta}$. We will

- derive representations for differentials of $(\hat{\mathbf{J}}(\theta))_{\theta\in\Theta}$ (Lemma 8.1.4)
- show non-degeneracy of the combined vector $(\mathbf{g}_1(\theta), \mathbf{g}_2(\theta))$ for all $\theta \in \mathcal{E}_P$, where $\mathbf{g}_1(\theta)$ and $\mathbf{g}_2(\theta)$ are constituted by all first order differentials and certain second order differentials of $\hat{\mathbf{J}}(\theta)$, respectively (Proposition 8.1.5)
- generalize the volume argument of Adler and Taylor (2007) to our needs (Lemma 8.1.6).

The combination of all these results leads to Proposition 8.1.7 which will allow us to show the Morse property in the subsequent section.

Lemma 8.1.4 (Differentiability)**.** *Let $k \in \mathbb{N}$, $\mathfrak{N} = (\mathbb{V}, \psi)$ be an ANN with a C^k activation function ψ and $\#V_{\text{out}} = 1$, let $\mathfrak{K}$ be a family of L^1-integrable regression problems with $\mathbb{P}_X$ having compact domain and let $\mathbf{m} \in \mathbf{M}$. Then*

$$\hat{J}_{\mathfrak{m}}(\theta) = \langle \Psi_\theta, f_{\mathfrak{m}} \rangle_{\mathbb{P}_X}$$

is in C^k and its partial derivatives satisfy

$$\partial_\theta^\alpha \hat{J}_{\mathfrak{m}}(\theta) = \langle \partial_\theta^\alpha \Psi_\theta, f_{\mathfrak{m}} \rangle_{\mathbb{P}_X}$$

for all multi-indices $|\alpha| \leq k$.

Note that the lemma implies that in the standard setting all differentials of $(\hat{\mathbf{J}}(\theta))_{\theta \in \Theta}$ define again Gaussian processes. This fact together with the representations for the differentials will be the basic tool in the analysis of $(\hat{\mathbf{J}}(\theta))_{\theta \in \Theta}$ and we mostly will not give reference to the lemma when using it.

Proof. By assumption, one has that $\mathbb{E}_{\mathfrak{m}}[\|Y\|] < \infty$. Recall that $f_{\mathfrak{m}}(x) = \mathbb{E}_{\mathfrak{m}}[Y \mid X = x]$ so that with the L^1-contraction property of the conditional expectation

$$\int |f_{\mathbf{m}}(x)| \, \mathbb{P}_X(dx) = \mathbb{E}_{\mathfrak{m}}\big[|\mathbb{E}_{\mathfrak{m}}[Y \mid X]|\big] \leq \mathbb{E}_{\mathfrak{m}}\big[\mathbb{E}_{\mathfrak{m}}[|Y| \mid X]\big] = \mathbb{E}_{\mathfrak{m}}[|Y|^2] < \infty. \tag{8.6}$$

Fix $\theta \in \Theta$ and $v \in \Theta$. By assumption, $\mathbb{P}_X$ has compact support $\mathcal{X}$ and the directional derivative $D_v^\theta \Psi$ in direction v in the θ component is a continuous mapping on $\Theta \times \mathcal{X}$. In particular, it is uniformly bounded on the compact set $\overline{B(\theta, \|v\|)} \times \mathcal{X}$, say by the constant C. Consequently, for every $t \in (0, 1]$, one has that

$$\begin{aligned} \frac{1}{t}(\hat{J}_{\mathfrak{m}}(\theta + tv) - \hat{J}_{\mathfrak{m}}(\theta)) &= \frac{1}{t} \int (\Psi_{\theta+tv}(x) - \Psi_\theta(x)) f_{\mathfrak{m}}(x) \, \mathbb{P}_X(dx) \\ &\overset{\text{FTC}}{=} \int \int_0^1 D_v^\theta \Psi_{\theta+stv}(x) \, f_{\mathfrak{m}}(x) \, ds \, \mathbb{P}_X(dx) \end{aligned}$$

Now note that $C|f_{\mathbf{m}}|$ is an integrable majorant due to (8.6). Using that for every $s \in [0, 1]$ and $x \in \mathbb{X}$, $\lim_{t \downarrow 0} D_v^\theta \Psi_{\theta+stv}(x) = D_v^\theta \Psi_\theta(x)$ by continuity of the differential it follows with dominated convergence that

$$\lim_{t \downarrow 0} \frac{1}{t}(\hat{J}_{\mathfrak{m}}(\theta + tv) - \hat{J}_{\mathfrak{m}}(\theta)) = \int D_v^\theta \Psi_\theta(x) \, f_{\mathfrak{m}}(x) \, \mathbb{P}_X(dx).$$

Recall that $(\theta, v) \mapsto D_v^\theta \Psi_\theta(x)$ is continuous and we get again with dominated convergence that the latter integral is continuous in the parameters θ and v. This proves that $\hat{J}_{\mathbf{m}}$ is C^1 and that the upper identity holds.

By induction, one obtains the general statement. The induction step can be carried out exactly as above by using that the assumptions imply that Ψ is k-times continuously differentiable as mapping on $\Theta \times \mathbb{X}$. □

As indicated before there are second order differentials that degenerate. However, for all first order and some second order differentials this is not the case. In the next step we will show this. We consider the stochastic processes $\mathbf{g}_1 = (\mathbf{g}_1(\theta))_{\theta \in \theta}$ and $\mathbf{g}_2 = (\mathbf{g}_2(\theta))_{\theta \in \theta}$ defined by

$$\mathbf{g}_1(\theta) := \nabla \mathbf{J}(\theta) \qquad \text{and} \tag{8.7}$$

$$\mathbf{g}_2(\theta) := \Big(\big(\partial^2_{\beta_j} \hat{\mathbf{J}}(\theta)\big)_{j \in V_1}, \big(\partial_{\beta_j} \partial_{w_{ij}} \hat{\mathbf{J}}(\theta)\big)_{\substack{j \in V_1 \\ i \in V_0}}, \big(\partial_{w_{ij}} \partial_{w_{kj}} \hat{\mathbf{J}}(\theta)\big)_{\substack{j \in V_1 \\ i,k \in V_0 \\ i \leq k}} \Big), \tag{8.8}$$

where we assume some total order on the input neurons V_0 such that $i \leq k$ makes sense for $i, k \in V_0$. Note that $\mathbf{g}_1$ utilizes the un-centered $\mathbf{J}$ to ensure $\nabla \mathbf{J}(\theta) = 0$ translates to $\mathbf{g}_1(\theta) = 0$ whereas $\mathbf{g}_2(\theta)$ utilizes the 'centered'[4] $\hat{\mathbf{J}}$. This is because $\mathbf{g}_2$ does not contain all second order differentials and a translation function F is necessary to get from $\mathbf{g}_2$ to $F(\theta, \mathbf{g}_2(\theta)) = \det(\nabla^2 \mathbf{J}(\theta))$. Constructing F in turn is more straightforward with the 'mean' $\|\Psi_\theta\|^2_{\mathbb{P}_X}$ built into F (cf. Section 8.1.2).

Proposition 8.1.5. *For an ANN $\mathfrak{N} = (\mathbb{V}, \psi)$ let $\#V_{out} = 1$. We consider $\mathbf{g}_i$ as defined in (8.7) and (8.8) based on the standard Gaussian setting (Definition 8.0.6). Then for every parameter $\theta \in \mathcal{E}_{\mathcal{P}}$ the Gaussian random vector $(\mathbf{g}_1(\theta), \mathbf{g}_2(\theta))$ is non-degenerate meaning that its covariance has full rank.*

Proof. Recall that $\hat{\mathbf{J}}(\theta) = \langle \Psi_\theta, \mathbf{f} \rangle_{\mathbb{P}_X}$. Since the variance does not depend on the mean we can assume without loss of generality $\mathbf{g}_1(\theta) = \nabla \hat{\mathbf{J}}(\theta)$ in this proof. Let I_1 and I_2 be index sets such that[5]

$$\mathbf{g}_1(\theta) = \nabla \hat{\mathbf{J}}(\theta) = (\partial_{\theta_i} \hat{\mathbf{J}}(\theta))_{i \in I_1} \quad \text{and} \quad \mathbf{g}_2(\theta) = (\partial_{\theta_i} \partial_{\theta_j} \hat{\mathbf{J}}(\theta))_{(i,j) \in I_2}.$$

[5] With slight misuse of the notation, we ignore the ordering of the differentials in the representation of $\mathbf{g}_2(\theta)$.

To show that $(\mathbf{g}_1(\theta), \mathbf{g}_2(\theta))$ is non-degenerate it suffices to show that the only vector $(\lambda_{\mathbf{i}})_{\mathbf{i} \in I_1 \cup I_2} \in \mathbb{R}^{I_1 \cup I_2}$ for which the linear combination

$$\sum_{i \in I_1} \lambda_i \partial_{\theta_i} \hat{\mathbf{J}}(\theta) + \sum_{(i,j) \in I_2} \lambda_{i,j} \partial_{\theta_i} \partial_{\theta_j} \hat{\mathbf{J}}(\theta)$$

has zero variance is $(\lambda_{\mathbf{i}})_{\mathbf{i} \in I_1 \cup I_2} \equiv 0$. Lemma 8.1.4 ensures that the differentials exist and that they can be moved into the inner product defining $\hat{\mathbf{J}}$. This implies

$$\begin{aligned} &\operatorname{Var}\left(\sum_{i \in I_1} \lambda_i \partial_{\theta_i} \hat{\mathbf{J}}(\theta) + \sum_{(i,j) \in I_2} \lambda_{i,j} \partial_{\theta_i} \partial_{\theta_j} \hat{\mathbf{J}}(\theta) \right) \\ &= \operatorname{Var}\left(\Big\langle \underbrace{\sum_{i \in I_1} \lambda_i \partial_{\theta_i} \Psi_\theta + \sum_{(i,j) \in I_2} \lambda_{i,j} \partial_{\theta_i} \partial_{\theta_j} \Psi_\theta}_{=:\phi}, \, \mathbf{f} \Big\rangle_{\mathbb{P}_X} \right). \end{aligned}$$

By weak universality (Definition 8.0.6) of the Gaussian process $\mathbf{f}$ on $\mathcal{X}$ it follows that the latter variance is zero if and only if $\phi \equiv 0$ on the support $\mathcal{X}$ of $\mathbb{P}_X$. It is therefore sufficient to prove that there exists no non-trivial linear combination of

$$\Big(\nabla \Psi_\theta, \big(\partial^2_{\beta_j} \Psi_\theta\big)_{j \in V_1}, \big(\partial_{\beta_j} \partial_{w_{ij}} \Psi_\theta\big)_{j \in V_1, i \in V_0}, \big(\partial_{w_{ij}} \partial_{w_{kj}} \Psi_\theta\big)_{j \in V_1, i,k \in V_0, i \le k} \Big),$$

which is zero on $\mathcal{X}$. We need to ensure that in θ all derivatives ∂_{θ_i} and $\partial_{\theta_i} \partial_{\theta_j}$ $(i \in I_1, (i,j) \in I_2)$ of the response Ψ_θ are linearly independent as functions on $\mathcal{X}$. We recall that

$$\Psi_\theta(x) = \beta_* + \sum_{j \in V_1} \psi\big(\beta_j + \langle x, w_{\bullet j} \rangle\big) w_{j*}$$

and note that we need to ensure linear independence of the following derivatives of the response Ψ_θ

$$\begin{aligned}
&x \mapsto 1 && && (\partial \beta_*) \\
&x \mapsto \psi\big(\beta_j + \langle x, w_{\bullet j} \rangle\big) && j \in V_1 && (\partial w_{j*}) \\
&x \mapsto \psi'\big(\beta_j + \langle x, w_{\bullet j} \rangle\big) w_{j*} && j \in V_1 && (\partial \beta_j) \\
&x \mapsto \psi'\big(\beta_j + \langle x, w_{\bullet j} \rangle\big) w_{j*} x_i && j \in V_1,\ i \in V_0 && (\partial w_{ij}) \\
&x \mapsto \psi''\big(\beta_j + \langle x, w_{\bullet j} \rangle\big) w_{j*} && j \in V_1 && (\partial \beta_j^2) \\
&x \mapsto \psi''\big(\beta_j + \langle x, w_{\bullet j} \rangle\big) w_{j*} x_i && i \in V_0, j \in V_1 && (\partial \beta_j \partial w_{ij}) \\
&x \mapsto \psi''\big(\beta_j + \langle x, w_{\bullet j} \rangle\big) w_{j*} x_i x_k && j \in V_1,\ i,k \in V_0 && (\partial w_{ij} \partial w_{kj})
\end{aligned}$$

Recall that $\phi = 0$ on $\mathcal{X}$ is a linear combination of the derivatives above with the prefactors $(\lambda_{\mathbf{i}})$. To distinguish between the types of the indices we write λ_{β_*}, λ_{β_j}, $\lambda_{\beta_j, \beta_j}$, $\lambda_{w_{i,j}}$ and so on to refer to the coefficients in front of the respective differentials. Note that in the above list all but the first differential can all be assigned to a particular neuron j in the first layer V_1. For a fixed neuron $j \in V_1$

we denote these contributions to ϕ by ϕ_j defined as

$$\begin{aligned}\phi_j(x) := \underbrace{\lambda_{w_{j*}}}_{=:P_j^{(0)}} \psi\big(\beta_j + \langle x, w_{\bullet j}\rangle\big) + \underbrace{\Big(\lambda_{\beta_j} + \sum_{i\in V_0} \lambda_{w_{ij}} x_i\Big)}_{=:P_j^{(1)}(x)} \psi'\big(\beta_j + \langle x, w_{\bullet j}\rangle\big) \\ + \underbrace{\Big(\lambda_{\beta_j^2} + \sum_{i\in V_0} \lambda_{\beta_j w_{ij}} x_i + \sum_{i,k\in V_0 : i\le k} \lambda_{w_{ij}w_{kj}} x_i x_k\Big)}_{=:P_j^{(2)}(x)} \psi''\big(\beta_j + \langle x, w_{\bullet j}\rangle\big).\end{aligned} \tag{8.9}$$

Together with $P^{(\emptyset)}(x) := \lambda_{\beta_*}$ we therefore obtain

$$\phi(x) = \lambda_{\beta_*} + \sum_{j\in V_1} \phi_j(x) = P^{(\emptyset)}(x) + \sum_{j\in V_1} \sum_{k=0}^{2} P_j^{(k)}(x) \psi^{(k)}\big(\beta_j + \langle x, w_{\bullet j}\rangle\big).$$

Since $\phi(x) = 0$ for all $x \in \mathcal{X}$ we can conclude with polynomial efficiency of θ that all the polynomials $P^{(\emptyset)}$ and $P_j^{(k)}$ $(j \in V_1, k = 0, 1, 2)$ are zero.

Now inspect the definition of the polynomials in (8.9) again. In the definition of every polynomial no monomial appears twice so that all coefficients have to be equal to zero. Since all coefficients appear in the polynomials, indeed all coefficients have to be equal to zero. This finishes the proof. □

We will combine the latter statement with a generalization of a result of Adler and Taylor (2007). It will allow us to conclude that the non-degeneracy of the distributions of $\mathbf{g}(\theta) := (\mathbf{g}_2(\theta), \mathbf{g}_1(\theta))$ will allow us to show that certain "thin" events (represented in terms of properties of the function graph of $\mathbf{g}$) have probability zero. In the final step, we will define appropriate "thin" events and verify the assumptions of the next lemma.

Lemma 8.1.6 (Generalization Adler and Taylor (2007, Lemma 11.2.10))**.** *Let $d, d' \in \mathbb{N}$, $U \subseteq \mathbb{R}^{d+d'}$ and $W \subseteq \mathbb{R}^d$ be measurable sets. Let $(\mathbf{g}(w))_{w\in W}$ be an $\mathbb{R}^{d'}$-valued, Lipschitz-continuous stochastic process and suppose that for some constants $C, \rho \in (0, \infty)$ one has for every $(w, v) \in U \cap (W \times \mathbb{R}^{d'})$ that the distribution of $\mathbf{g}(w)$ confined to $B_{\mathbb{R}^{d'}}(w, \rho)$ is absolutely continuous w.r.t. Lebesgue measure with the density being bounded by C. If*

$$\mathcal{H}_{d'}(U) = 0,$$

then the probability of the graph of $(\mathbf{g}(w))_{w\in W}$ intersecting U is zero, i.e.

$$\mathbb{P}\big(\exists w \in W : (w, \mathbf{g}(w)) \in U\big) = 0.$$

Proof. Without loss of generality we can assume that $U \subseteq W \times \mathbb{R}^{d'}$ (otherwise we replace U by $U \cap (W \times \mathbb{R}^{d'})$). Let $L, \epsilon \in (0, \infty)$. Since by assumption $\mathcal{H}_{d'}(U) = 0$ there exist a U-valued sequence $(w_i, v_i)_{i\in\mathbb{N}}$ and a $(0, \epsilon)$-valued sequence $(r_i)_{i\in\mathbb{N}}$ such that

$$U \subseteq \bigcup_{i\in\mathbb{N}} B((w_i, v_i), r_i) \quad \text{and} \quad \sum_{i\in\mathbb{N}} r_i^{d'} \le \epsilon.$$

Now note that for an arbitrary Lipschitz function $g : W \to \mathbb{R}^{d'}$ with Lipschitz constant L we have the following: if there exists $w \in W$ with $(w, g(w)) \in U$, then there exists an index $i \in \mathbb{N}$ with $\|g(w_i) - v_i\| < (1 + L)r_i$. Indeed, in that case there exists an index $i \in \mathbb{N}$ with $\|(w, g(w)) - (w_i, v_i)\| < r_i$, since balls of radius r_i around (w_i, v_i) cover U, and together with the Lipschitz continuity we get that

$$\begin{aligned}\|g(w_i) - v_i\| &\le \|g(w_i) - g(w)\| + \|g(w) - v_i\| \\ &\le L\|w_i - w\| + \|g(w) - v_i\| \le (1 + L)r_i.\end{aligned}$$

Consequently, we get that for the events

$$\mathbb{U} = \{\exists w \in W : (w, \mathbf{g}(w)) \in U\} \quad \text{and} \quad \mathbb{L}^{(L)} = \{\mathbf{g} \text{ is } L\text{-Lipschitz cont.}\}$$

one has that

$$\mathbb{U} \cap \mathbb{L}^{(L)} \subseteq \bigcup_{i \in \mathbb{N}} \{d(\mathbf{g}(w_i), v_i) < (1+L) r_i\}.$$

Now suppose that $\epsilon \in (0, \infty)$ was chosen sufficiently small to guarantee that $(1+L)\epsilon < \rho$ and conclude using the Lebesgue measure $\lambda^{d'}$ on $\mathbb{R}^{d'}$ that

$$\begin{aligned}
\mathbb{P}(\mathbb{U} \cap \mathbb{L}^{(L)}) &\leq \sum_{i=1}^{\infty} \mathbb{P}\big(\mathbf{g}(w_i) \in B_{(1+L)r_i}(v_i)\big) \\
&\leq \sum_{i=1}^{\infty} \int_{B_{(1+L)r_i}(v_i)} \frac{d\mathbb{P}_{\mathbf{g}(x_i)}}{d\lambda^{d'}} d\lambda^{d'} \\
&\leq C \sum_{i=1}^{\infty} \lambda^{d'}(B_{(1+L)r_i}(v_i)) \leq C\lambda^{d'}(B_1(0))(1+L)^{d'}\epsilon.
\end{aligned}$$

By letting ϵ go to zero we conclude that $\mathbb{P}(\mathbb{U} \cap \mathbb{L}^{(L)}) = 0$. This is true for every $L \in (0, \infty)$ and a union over rational L finishes the proof. □

Proposition 8.1.7. *Assume the standard setting (Definition 8.0.6). Let* $\mathbf{g} = (\mathbf{g}(\theta))_{\theta \in \theta}$ *be the process given by*

$$\mathbf{g}(\theta) = (\mathbf{g}_2(\theta), \mathbf{g}_1(\theta)),$$

where $\mathbf{g}_1$ *and* $\mathbf{g}_2$ *are as defined in (8.7) and (8.8). Let* $d, d' \in \mathbb{N}$ *be the dimensions of* θ *and the target space of* $\mathbf{g}$*. Moreover, let* $U \subset \mathbb{R}^{d+d'}$ *be a measurable set with*

$$\mathcal{H}_{d'}(U) = 0,$$

then

$$\mathbb{P}(\{\exists \theta \in \mathcal{E}_P : (\theta, \mathbf{g}(\theta)) \in U\}) = 0.$$

Proof. Since $\mathcal{E}_P$ is an open set in Θ which is separable as a finite dimensional real vector space, we can cover it by a countable number of compact balls contained in $\mathcal{E}_P$. Specifically, about any rational point in $\mathcal{E}_P$ we take a closed ball contained in Θ. Then it is sufficient to show the claim for any such ball as the countable union of zero sets is still a zero set. We thus consider such a compact subset $W \subseteq \mathcal{E}_P$ and aim to show

$$\mathbb{P}(\{\exists \theta \in W : (\theta, \mathbf{g}(\theta)) \in U\}) = 0.$$

Since we have that $\hat{\mathbf{J}}$ has continuous differentials up to third degree (Lemma 8.1.4), the process $\mathbf{g}$ is continuous as it only consists of first and second order differentials (and a continuous mean in the case of $\mathbf{g}_1$). This then implies that the covariance kernel of $\mathbf{g}$ is continuous.[6] Moreover the third order differentials are continuous and thereby bounded on compact sets, which implies $\mathbf{g}$ is almost surely Lipschitz continuous on V and since the covariance kernel of $\mathbf{g}$ is continuous, the function

$$\gamma(\theta) = \det(\mathrm{Cov}(\mathbf{g}(\theta)))$$

is continuous on W and therefore assumes a minimum in W as W is compact. Since the covariance is positive definite, this minimum must be greater or equal than zero and by Proposition 8.1.5 it cannot be zero since $W \subseteq \mathcal{E}_P$. And since $\mathbf{g}(\theta)$ is Gaussian by Lemma 8.1.4 and assumption on $f_{\mathbf{M}}$ its Lebesgue density is bounded by the density at its mean given by

$$(2\pi)^{-\dim(\Theta)/2} \det(\mathrm{Cov}(\mathbf{g}(\theta)))^{-1/2} \leq (2\pi)^{-\dim(\Theta)/2} (\inf_{\theta \in W} \gamma)^{-1/2} =: C < \infty.$$

We can now finish the proof by application of Lemma 8.1.6. □

[6] e.g. Talagrand, 1987, Theorem 3.

Remark 8.1.8. While we have assumed analytic activation functions in the standard setting (Definition 8.0.6) we only required the activation functions to be in C^3 so far. As we only work with the gradient and Hessian even the assumption of C^3 activation functions appears too strong. And indeed in the proof above we only used this fact to show that $\mathbf{g}$ is almost surely Lipschitz. Adler and Taylor (2007) highlights a similar issue after the proof of their Lemma 11.2.10, which we generalized in Lemma 8.1.6. Adler and Taylor (2007) proceed to generalize their result using a growth condition on the modulus of continuity in place of Lipschitz continuity (cf. Lemma 11.2.11). A similar generalization should be possible for Lemma 8.1.6. But since we need analytic activation functions anyway for Lemma 8.1.9 and therefore Theorem 8.1.2, we avoid the complications of this generalization.

8.1.2 Proof of Thm 8.1.2

The main task of this section is to prove existence of the function F announced in the end of the introduction to Section 8.1. We will show the following.

Lemma 8.1.9 (Definition of F)**.** *In the standard setting (Definition 8.0.6) let $\mathbf{g}_2 = (\mathbf{g}_2(\theta))_{\theta\in\theta}$ be the $\mathbb{R}^{I_2}$-valued process as defined in (8.8), with I_2 being the respective index set. Then there exists a non-zero, real-analytic function*

$$F : \mathcal{E}_P \times \mathbb{R}^{I_2} \to \mathbb{R},$$

such that for every $\theta \in \mathcal{E}_P = \mathcal{E}_P(\mathcal{X})$ with $\nabla\mathbf{J}(\theta) = 0$ we have that

$$\det(\nabla^2\mathbf{J}(\theta)) = F(\theta, \mathbf{g}_2(\theta)).$$

Before we prove this lemma, we show that it finishes the proof of Theorem 8.1.2. We have that

$$\begin{aligned}
&\mathbb{P}\Big(\exists\theta \in \mathcal{E}_P : \nabla\mathbf{J}(\theta) = 0,\ \nabla^2\mathbf{J}(\theta) = 0\Big)\\
&\overset{\text{Lemma 8.1.9}}{\leq} \mathbb{P}\Big(\exists\theta \in \mathcal{E}_P : \nabla\mathbf{J}(\theta) = 0,\ F(\theta, \mathbf{g}_2(\theta)) = 0\Big)\\
&\overset{\mathbf{g}_1(\theta)=\nabla\mathbf{J}(\theta)}{=} \mathbb{P}\Big(\exists\theta \in \mathcal{E}_P : (\theta, \mathbf{g}_2(\theta)) \in F^{-1}(0),\ \mathbf{g}_1(\theta) \in \{0\}\Big)\\
&= \mathbb{P}\Big(\exists\theta \in \mathcal{E}_P : (\theta, \mathbf{g}(\theta)) \in \underbrace{F^{-1}(0) \times \{0\}}_{=:U}\Big).
\end{aligned}$$

To apply Proposition 8.1.7 we only need that U has sufficiently small Hausdorff dimension (specifically smaller dimension than the target space of $\mathbf{g}$). And indeed, since F is a non-zero, real-analytic function its zero set is one dimension smaller than the dimension $d' := \#I_1 + \#I_2$ of its domain, see.[7] This entails that

[7] Mityagin, 2020.

$$\mathcal{H}_{d'}(F^{-1}(0)) = 0 \quad \text{and} \quad \mathcal{H}_{d'}(U) = 0.$$

By definition, d' is also the dimension of the target space of $\mathbf{g}$ and Proposition 8.1.7 entails that

$$\mathbb{P}\big(\exists\theta \in \mathcal{E}_P : (\theta, \mathbf{g}(\theta)) \in U\big).$$

This finishes the proof of Theorem 8.1.2.

Remark 8.1.10 (Analytic activation). Observe that the analytic activation function was only used to ensure F is analytic (cf. Remark 8.1.8). Lemma 8.1.9 is therefore the appropriate place to search for generalizations.

Remark 8.1.11 (Efficient is sufficient). The function in Lemma 8.1.9 can be defined on the efficient domain $\mathcal{E} = \mathcal{E}(\mathcal{X})$ (cf. Definition 8.2.1), i.e. $F : \mathcal{E} \times \mathbb{R}^{I_2} \to \mathbb{R}$. This is a superset of the polynomially efficient domain $\mathcal{E}_P$ in general and coincides with the efficient domain for certain activation functions (cf. Theorem 8.2.3). We conduct the proof with $\mathcal{E}$ but readers may replace this with $\mathcal{E}_P$.

Proof of Lemma 8.1.9. To prove Lemma 8.1.9 we need to construct a function F with the following properties

(P1) at any $\theta \in \mathcal{E}$ with $\nabla \mathbf{J}(\theta) = 0$ we have

$$\det(\nabla^2 \mathbf{J}(\theta)) = F(\theta, \mathbf{g}_2(\theta)),$$

(P2) F is analytic, and

(P3) F is non-zero, i.e. there exists an input to F where F is non-zero.

Recall that we have by definition for some total order on V_0

$$\mathbf{g}_2(\theta) = \Big(\big(\partial^2_{\beta_j} \hat{\mathbf{J}}(\theta)\big)_{j \in V_1}, \big(\partial_{\beta_j} \partial_{w_{ij}} \hat{\mathbf{J}}(\theta)\big)_{\substack{j \in V_1 \\ i \in V_0}}, \big(\partial_{w_{ij}} \partial_{w_{kj}} \hat{\mathbf{J}}(\theta)\big)_{\substack{j \in V_1 \\ i,k \in V_0 \\ i \le k}} \Big).$$

In order to be able to plug $\mathbf{g}_2$ into F it must thus be of the form

$$F : \mathcal{E} \times V_1 \times (V_1 \times V_0) \times (V_1 \times \mathrm{Sym}(V_0^2)) \to \mathbb{R},$$

where $\mathrm{Sym}(V_0^2) = \{(i,k) \in V_0^2 : i \le k\}$. To satisfy (P1) we need to reconstruct the determinant of the Hessian $\nabla^2 \mathbf{J}(\theta)$ on the basis of the differentials in $\mathbf{g}_2(\theta)$ (that originate from the Hessian $\nabla^2 \hat{\mathbf{J}}(\theta)$). First recall that by Proposition 8.1.3, one has that

$$\begin{aligned} \det(\nabla^2 \mathbf{J}(\theta)) &= \det\Big(\nabla^2 \Big[R(\theta) + \|\Psi_\theta\|^2_{\mathbb{P}_X} - 2\hat{\mathbf{J}}(\theta) + \mathbb{E}_{\mathbf{M}}[\|Y\|^2] \Big]\Big) \\ &= \det\big(\nabla^2 [R(\theta) + \|\Psi_\theta\|^2_{\mathbb{P}_X}] - 2\nabla^2 \hat{\mathbf{J}}(\theta)\big). \end{aligned}$$

Since $\theta \mapsto \nabla^2 [R(\theta) + \|\Psi_\theta\|^2_{\mathbb{P}_X}]$ is a deterministic function, we can absorb it into the definition of F and construct a function $\tilde{F}$ that reproduces $\nabla^2 \hat{\mathbf{J}}(\theta)$ from $\mathbf{g}_2(\theta)$. That is, assuming we had a function $\tilde{F}$ with $\tilde{F}(\theta, \mathbf{g}_2(\theta)) = \nabla^2 \hat{\mathbf{J}}(\theta)$ in all critical points, we can define

$$F(\theta, x) := \det\big(\nabla^2 [R(\theta) + \|\Psi_\theta\|^2_{\mathbb{P}_X}] - 2\tilde{F}(\theta, x)\big)$$

If all entries of the matrix valued function $\tilde{F}$ are analytic functions it is therefore sufficient for all entries of $\theta \mapsto \nabla^2 [R(\theta) + \|\Psi_\theta\|^2_{\mathbb{P}_X}]$ to be analytic for F to be analytic, because the determinant is a sum and product of these analytic components. And by the assumption that X has compact support in the standard setting (Definition 8.0.6) and Theorem 5.1 in Dereich and Kassing (2024), $\theta \mapsto \|\Psi_\theta\|^2_{\mathbb{P}_X}$ is analytic, and so are its differentials. (P2) thus follows if all entries of $\tilde{F}$ are analytic since we assumed Ψ to be analytic in the standard setting (Definition 8.0.6).

Our strategy is therefore to show that all entries/differentials of $\nabla^2 \hat{\mathbf{J}}(\theta)$ fall into one of the following categories:

(a) the partial differential is contained in $\mathbf{g}_2$ in which case the respective matrix entry in $\tilde{F}$ is identical to the related input (analytic)

(b) the partial differential is zero (analytic), or

(c) there is an (analytic) deterministic function of θ (that we still need to construct) such that the partial differential coincides with the function value whenever $\nabla \mathbf{J}(\theta) = 0$.

To enact this strategy, we now consider all the second order derivatives in $\nabla^2 \hat{\mathbf{J}}(\theta)$ that are not contained in $\mathbf{g}_2$ and categorize them into (b) or (c). Recall that by Lemma 8.1.4

$$\hat{\mathbf{J}}(\theta) = \langle \Psi_\theta, \mathbf{f} \rangle_{\mathbb{P}_X}, \quad \nabla \hat{\mathbf{J}}(\theta) = \langle \nabla \Psi_\theta, \mathbf{f} \rangle_{\mathbb{P}_X}, \quad \nabla^2 \hat{\mathbf{J}}(\theta) = \langle \nabla^2 \Psi_\theta, \mathbf{f} \rangle_{\mathbb{P}_X}.$$

We therefore need to consider the derivatives of the response function. Since the response Ψ_θ is essentially a weighted sum over the index set V_1 with all

parameters appearing only in one of the summands, we have that second order partial derivatives that belong to two different neurons $j \neq l$ of the hidden layer V_1 vanish. This then directly implies that these differentials also vanish for $\hat{\mathbf{J}}$. Specifically, we have that

$$\begin{aligned}
\partial_{\beta_j}\partial_{\beta_l}\Psi_\theta &= 0 && \forall j \neq l \in V_1\\
\partial_{\beta_j}\partial_{w_{il}}\Psi_\theta &= 0 && \forall j \neq l \in V_1,\ \forall i \in V_0\\
\partial_{w_{ij}}\partial_{w_{kl}}\Psi_\theta &= 0 && \forall j \neq l \in V_1,\ \forall i,k \in V_0\\
\partial_{w_{ij}}\partial_{w_{l*}}\Psi_\theta &= 0 && \forall j \neq l \in V_1,\ \forall i \in V_0\\
\partial_{w_{ij}}\partial_{\beta_*}\Psi_\theta &= 0 && \forall j \in V_1,\ \forall i \in V_0\\
\partial_{\beta_j}\partial_{\beta_*}\Psi_\theta &= 0 && \forall j \in V_1
\end{aligned}$$

Let us refer to "inner parameters" as the parameters that appear inside the activation functions (the connections from the input layer to the hidden layer and the biases of neurons in the hidden layer). All second order derivatives of these inner parameters are either included in $\mathbf{g}_2$ or mix derivatives of two different hidden neurons and therefore vanish. All second order differentials with respect to inner parameters (exclusively) are thus of type (a) or (b).

Since the response is linear in the outer (remaining) parameters w_{j*} and β_* taking two derivatives in these direction also results in zero, i.e.,

$$\partial^2_{\beta_*}\Psi_\theta = 0, \quad \partial_{\beta_*}\partial_{w_{j*}}\Psi_\theta = 0 \qquad \text{and} \qquad \partial^2_{w_{j*}}\Psi_\theta = 0 \quad \forall j \in V_1.$$

Most derivatives are thus in category (b).

The only derivatives left are therefore the mixtures of outer derivatives w_{j*} with inner derivatives β_j and w_{ij} of the same hidden neuron $j \in V_1$. These will be in category (c). For those observe:

$$\partial_{w_{ij}}\Psi_\theta(x) = w_{j*}\psi'\big(\beta_j + \langle x, w_{\bullet j}\rangle\big)x_i.$$

Using that w_{j*} is non-zero as $\theta \in \mathcal{E}$ we get that

$$\partial_{w_{j*}}\partial_{w_{ij}}\Psi_\theta(x) = \psi'\big(\beta_j + \langle x, w_{\bullet j}\rangle\big)x_i = \tfrac{1}{w_{j*}}\partial_{w_{ij}}\Psi_\theta(x).$$

Since we are allowed to move differentiation into the inner products by Lemma 8.1.4 we get

$$\partial_{w_{j*}}\partial_{w_{ij}}\hat{\mathbf{J}}(\theta) = \big\langle \partial_{w_{j*}}\partial_{w_{ij}}\Psi_\theta(x), \mathbf{f}\big\rangle = \tfrac{1}{w_{j*}}\big\langle \partial_{w_{ij}}\Psi_\theta(x), \mathbf{f}\big\rangle = \tfrac{1}{w_{j*}}\partial_{w_{ij}}\hat{\mathbf{J}}(\theta).$$

Now recall by Proposition 8.1.3 we have that

$$\nabla\mathbf{J}(\theta) = \nabla_\theta[R(\theta) + \|\Psi_\theta\|^2_{\mathbb{P}_X}] - 2\nabla\hat{\mathbf{J}}(\theta) \tag{8.10}$$

so that in every critical point θ of $\mathbf{J}$ with $\nabla\mathbf{J}(\theta) = 0$ we get that

$$\frac{1}{2w_{j*}}\partial_{w_{ij}}[R(\theta) + \|\Psi_\theta\|^2_{\mathbb{P}_X}] = \frac{1}{w_{j*}}\partial_{w_{ij}}\hat{\mathbf{J}}(\theta) = \partial_{w_{j*}}\partial_{w_{ij}}\hat{\mathbf{J}}(\theta).$$

For the definition of $\tilde{F}$ we therefore use the deterministic function

$$\theta \mapsto \frac{1}{2w_{j*}}\partial_{w_{ij}}[R(\theta) + \|\Psi_\theta\|^2_{\mathbb{P}_X}]$$

for the component where the differential $\partial_{w_{j*}}\partial_{w_{ij}}$ should be.

We proceed in complete analogy with the differential $\partial_{w_{j*}}\partial_{\beta_j}$ and obtain that for every critical efficient parameter θ

$$\partial_{w_{j*}}\partial_{\beta_j}\hat{\mathbf{J}}(\theta) = \frac{1}{2w_{j*}}\partial_{\beta_j}[R(\theta) + \|\Psi_\theta\|^2_{\mathbb{P}_X}]$$

and we use the respective deterministic function to define the corresponding component of $\tilde{F}$.

Note that $\theta \mapsto \|\Psi_\theta\|^2_{\mathbb{P}_X}$ is analytic, the deterministic functions that we use as substitutes for the remaining second order differentials in $\tilde{F}$ are therefore analytic on $\mathcal{E}$ (where $w_{j*} \neq 0$). Thus we constructed a function F satisfying properties (P1) and (P2).

It remains to show that F is a non-zero function (P3). To show this we will arrange the second order differentials appropriately. We put the outer parameters $\gamma := (\beta_*, (w_{j*})_{j\in V_1})$ at the end of the vector. Recall that all these components fall into category (b) so that, in particular,

$$\nabla^2_\gamma[R(\theta) + \|\Psi_\theta\|^2_{\mathbb{P}_X}] - 2\tilde{F}_\gamma(\theta, x) = \nabla^2_\gamma[R(\theta) + \|\Psi_\theta\|^2_{\mathbb{P}_X}],$$

where F_γ is F restricted to the output components with pairs from the γ coordinates.

Let the remaining inner parameters be given by

$$\alpha := \Big((\beta_j)_{j\in V_1}, (w_{ij})_{\substack{i\in V_0\\ j\in V_1}}\Big).$$

Recall that the second order differential with respect to two parameters from α belongs to category (a) or (b) and that the diagonal belongs to (a). The diagonal can therefore be fully controlled and the other entries are either zero naturally or can be set to zero. Thus, for every parameter $\theta \in \mathcal{E}$ and any given $\lambda \in \mathbb{R}$ we can find x_λ with

$$\tilde{F}_\alpha(\theta, x_\lambda) = -\tfrac{\lambda}{2}\mathbb{I},$$

where $\mathbb{I}$ is the identity matrix and F_α is F restricted to the output components with pairs from the α coordinates. Consequently, for this choice of x we have that

$$\nabla^2\|\Psi_\theta\|^2_{\mathbb{P}_X} - 2\tilde{F}(\theta, x_\lambda) = \begin{pmatrix} \nabla^2_\alpha[R(\theta) + \|\Psi_\theta\|^2_{\mathbb{P}_X}] + \lambda\mathbb{I} & B(\theta) \\ B(\theta)^T & \nabla^2_\gamma[R(\theta) + \|\Psi_\theta\|^2_{\mathbb{P}_X}] \end{pmatrix}.$$

Marked with a $B(\theta)$ are the mixed derivatives with parameters from α and γ. These are either of type (b) or (c) and in particular they are functions of θ. More than F being non-zero, we will show for any fixed θ there exists λ and thus x_λ such that $F(\theta, x_\lambda) \neq 0$. For this note that $F(\theta, x_\lambda)$ is given by the determinant of the equation above and the determinant of a block matrix is given by

$$\det \begin{bmatrix} A & B \\ B^T & D \end{bmatrix} = \det(D)\det(A - BD^{-1}B^T)$$

We will show that $D := D(\theta) := \nabla^2_\gamma[R(\theta) + \|\Psi_\theta\|^2_{\mathbb{P}_X}]$ is a strictly positive definite matrix and by doing so we show that it has full rank and therefore non-zero determinant. Since the eigenvalue λ of $\lambda\mathbb{I}$ can be directly controlled with the selection of x_λ, selecting a sufficiently large $\lambda \gg 0$ ensures that the eigenvalues of

$$A - BD^{-1}B^T = \lambda\mathbb{I} + \underbrace{\nabla^2_\alpha[R(\theta) + \|\Psi_\theta\|^2_{\mathbb{P}_X}] - B(\theta)D^{-1}B(\theta)}_{\text{some matrix}}$$

are all strictly positive. Thus we can ensure the second determinant is non-zero.

What is left is thus the proof that $\nabla^2_\gamma[R(\theta) + \|\Psi_\theta\|^2_{\mathbb{P}_X}]$ strict positive definite. Since $R(\theta)$ is assumed to be convex $\nabla^2_\gamma R(\theta)$ is positive semi-definite. It is thus sufficient to prove strict positive definiteness for $\nabla^2_\gamma\|\Psi_\theta\|^2_{\mathbb{P}_X}$.

Let $c \in \mathbb{R}^{\#V_1+1}$ and recall $\gamma = (\beta_*, (w_{j*})_{j\in V_1}) \in \mathbb{R}^{\#V_1+1}$. Using that the

iterated differentials from γ belong to category (b) we conclude that

$$\begin{aligned} c^T \nabla_\gamma^2 \|\Psi_\theta\|_{\mathbb{P}_X}^2 c &= \sum_{i,j=0}^{\#V_1} c_i c_j \partial_{\gamma_i}\partial_{\gamma_j} \|\Psi_\theta\|_{\mathbb{P}_X}^2 \\ &= \sum_{i,j=0}^{\#V_1} c_i c_j \partial_{\gamma_i}\partial_{\gamma_j} \int \Psi_\theta^2(x)\, \mathbb{P}_X(dx) \\ &= \int \sum_{i,j=0}^{\#V_1} c_i c_j \partial_{\gamma_i}\partial_{\gamma_j} \Psi_\theta(x)^2\, \mathbb{P}_X(dx). \\ &\overset{\partial_{\gamma_i}\partial_{\gamma_j}\Psi_\theta \equiv 0}{=} 2 \int \sum_{i,j=0}^{\#V_1} c_i c_j (\partial_{\gamma_i}\Psi_\theta(x))(\partial_{\gamma_j}\Psi_\theta(x))\, \mathbb{P}_X(dx). \\ &= 2 \Big\| \sum_{i=0}^{\#V_1} c_i \partial_{\gamma_i}\Psi_\theta \Big\|_{\mathbb{P}_X}^2. \end{aligned}$$

Consequently, $\nabla_\gamma^2 \|\Psi_\theta\|_{\mathbb{P}_X}^2$ is always positive definite. To see strict positive definiteness we analyze solutions c for which the latter norm is zero. This requires that

$$\sum_{i=0}^{\#V_1} c_i \partial_{\gamma_i}\Psi_\theta = 0, \qquad \mathbb{P}_X\text{-almost surely.} \tag{8.11}$$

Identifying V_1 with $\{1, \dots, \#V_1\}$ this implies that

$$\partial_{\gamma_0}\Psi_\theta = \partial_{\beta_*}\Psi_\theta = 1 \quad \text{and} \quad \partial_{\gamma_i}\Psi_\theta = \partial_{w_{i*}}\Psi_\theta = \psi\big(\beta_i + \langle w_{\bullet i}, x\rangle\big),$$

for every $i \in V_1$, so that (8.11) implies that

$$c_0 + \sum_{i=1}^{\#V_1} c_i\, \psi\big(\beta_i + \langle w_{\bullet i}, x\rangle\big) = 0 \qquad \mathbb{P}_X\text{-almost surely.}$$

Since the term above is continuous in x it is thus zero for all x in the support $\mathcal{X}$ of $\mathbb{P}_X$. Efficiency (Definition 8.2.1) of θ then implies that $c \equiv 0$ is the unique solution of (8.11). This proves that $\nabla_\gamma^2 \|\Psi_\theta\|_{\mathbb{P}_X}^2$ is strictly positive definite and therefore finishes the proof of (P3). □

8.2 Characterization of the efficient domain

In the following $\mathfrak{N} = (\mathbb{V}, \psi)$ is a fixed ANN. We start with a more natural axiomatic definition of efficient parameters than the polynomial efficiency we required for our main result.

Definition 8.2.1 (Efficient and redundant parameters)**.** A parameter $\theta = (w, \beta)$ is called *efficient*, if

(a) all neurons are used meaning that for all $k \in V_1$ one has

$$w_{k\bullet} = (w_{kl})_{l \in V_{\text{out}}} \not\equiv 0,$$

(b) the equation

$$\lambda_\emptyset + \sum_{j \in V_1} \lambda_j \psi\Big(\beta_j + \sum_{i \in V_0} x_i w_{ij}\Big) = 0 \qquad \forall x \in \mathcal{X}$$

has the unique solution $\lambda_j = 0$ for all $j \in V_1 \uplus \{\emptyset\}$.

We denote by $\mathcal{E} = \mathcal{E}(\mathcal{X})$ the *efficient domain*, which is the set of all parameters $\theta = (w, \beta)$ that are efficient. A parameter that is not efficient is called *redundant*.

In the remark below we introduce categories of redundant parameters and hope to convey the intuition of this definition of 'efficiency'.

Remark 8.2.2 (Taxonomy of redundant parameters). A parameter can be redundant for various reasons: If property (a) does not hold we call it a *deactivation redundancy* since the output of the hidden neuron k is ignored. If the property (b) does not hold there exists a neuron which can be linearly replicated by other neurons meaning that there exists a neuron $k \in V_1$ and $\lambda_j \in \mathbb{R}$ for all $j \in V_1 \uplus \{\emptyset\}$ with

$$\psi\Big(\beta_k + \sum_{i \in V_0} x_i w_{ik}\Big) = \lambda_\emptyset + \sum_{j \in V_1 \setminus \{k\}} \lambda_j \psi\Big(\beta_j + \sum_{i \in V_0} x_i w_{ij}\Big) \quad \forall x \in \mathcal{X}.$$

If $\lambda_\emptyset$ is the only λ_j not equal to zero in this linear combination, we speak of a *bias redundancy*, as the neuron k is constant like the bias (typically $w_{\bullet k} = 0$, cf. Lemma 8.5.6). If there is another neuron $j \in V_1$ such that $(\beta_j, w_{\bullet j}) = (\beta_k, w_{\bullet k})$ the neuron k can be trivially linearly combined from the others and we speak of a *duplication redundancy*. We call all other cases a *generalized duplication redundancy*. While deactivation, bias and duplication redundancies occur independently of the activation function (cf. Lemma 8.2.6), generalized duplication redundancies only occur due to symmetries of the activation function (cf. Lemma 8.2.7, Lemma 8.2.8 and Example 8.2.9).

Obviously, a configuration is $(0,0)$-polynomially efficient if and only if it is efficient in the sense of Definition 8.2.1. But in general the assumption of polynomial efficiency is stronger.

As we will show next both notions generally coincide in the case where the activation function is either sigmoid or tanh! Further we will show that they also coincide with the explicit set representation (8.2).

Theorem 8.2.3 (Characterization of efficient networks for sigmoid and tanh)**.** *Assume that in the considered ANN $\mathfrak{N}$ the activation function ψ is either* sigmoid *or* tanh*, further assume that $\mathcal{X}$ contains an open set. Then for every $n \in \mathbb{N}_0$ and $m = (m_\emptyset, m_0, \dots, m_n) \in \mathbb{N}_0^{n+2}$ one has*

$$\mathcal{E}(\mathcal{X}) = \mathcal{E}_P^m(\mathcal{X}) = \mathcal{E}_0,$$

where, as in (8.2)*,*

$$\mathcal{E}_0 := \left\{ \theta = (w,\beta) \in \mathbb{R}^{E \times (V \setminus V_0)} : \begin{array}{ll} w_{j\bullet} \neq 0 & \forall j \in V_1, \\ w_{\bullet j} \neq 0 & \forall j \in V_1, \\ (w_{\bullet i}, \beta_i) \neq \pm (w_{\bullet j}, \beta_j) & \forall i,j \in V_1 \text{ with } i \neq j \end{array} \right\}.$$

Proofsketch. Since the activation function is either sigmoid or tanh and therefore real-analytic, the equations in the definition of the efficient set (Definition 8.2.1) and the definition of polynomial independence (Definition 8.1.1) are real-analytic. Since analytic functions which are zero on an open set are zero anywhere, the open set contained in $\mathcal{X}$ therefore ensures that we can assume without loss of generality $\mathcal{X} = \mathbb{R}^{V_{\text{in}}}$

Suppressing $\mathcal{X}$ in the notation of $\mathcal{E}$ and $\mathcal{E}_P^m$ the proof is then established by showing that $\mathcal{E} \subseteq \mathcal{E}_0 \subseteq \mathcal{E}_P^m \subseteq \mathcal{E}$. Note that $\mathcal{E}_P^m \subseteq \mathcal{E}$ is trivial as polynomials can always chosen to be constant.

- To prove "$\mathcal{E} \subseteq \mathcal{E}_0$" we will show that any parameter $\theta \notin \mathcal{E}_0$ is not in $\mathcal{E}$. More explicitly, we will construct a redundancy and show that one of the properties (a) or (b) does not hold.

  Remark 8.2.4. As part of this proof in Subsection 8.2.1, we will prove that $\mathcal{E}$ is always contained in

$$\bar{\mathcal{E}} := \left\{ \theta = (w,\beta) \in \mathbb{R}^{E \times (V \setminus V_0)} : \begin{array}{ll} w_{j\bullet} \neq 0 & \forall j \in V_1, \\ w_{\bullet j} \neq 0 & \forall j \in V_1, \\ (w_{\bullet i}, \beta_i) \neq (w_{\bullet j}, \beta_j) & \forall i \neq j \in V_1 \end{array} \right\} \tag{8.12}$$

regardless of the activation function ψ (Lemma 8.2.6). With a counterexample (Example 8.2.9) we further show that the *sign-symmetric redundancy* $(w_{\bullet i}, \beta_i) = -(w_{\bullet j}, \beta_j)$ is specific to the activation functions $\psi \in \{\text{sigmoid}, \tanh\}$. This shows that the explicit definition of $\mathcal{E}_0$ does not generalize well.

- It will be harder to prove that "$\mathcal{E}_0 \subseteq \mathcal{E}_P^m$" and we will first consider the case with one dimensional input (i.e. $\#V_{\text{in}} = 1$) in Subsection 8.2.2. This proof relies on the complex poles of the meromorphic activation function $\psi \in \{\text{sigmoid}, \tanh\}$. We will then show that the one-dimensional result also implies the general result in Subsection 8.2.3.

  Remark 8.2.5. For the transfer from the one-dimensional result to the general result we we do not make use of the assumption $\psi \in \{\text{sigmoid}, \tanh\}$. We only use that the result holds for the 1-dimensional case. This suggests that this part of the proof should be transferrable to other activation functions except for the fact that $\mathcal{E}_0$ may be different for other activation functions in general and we use the specific structure of $\mathcal{E}_0$. □

8.2.1 Proof of $\mathcal{E} \subseteq \mathcal{E}_0$

Take $\theta \notin \mathcal{E}_0$, then it is sufficient to prove this parameter to be not efficient, i.e. $\theta \notin \mathcal{E}$. For this we are going to consider the possible constraints of $\mathcal{E}_0$ the parameter θ can violate and match them with the types of redundancies we classified in Remark 8.2.2. Recall

$$\mathcal{E}_0 \overset{\text{def}}{=} \left\{ \theta = (w, \beta) \in \mathbb{R}^{E \times (V \setminus V_0)} : \begin{array}{ll} w_{j\bullet} \neq 0 & \forall j \in V_1, \\ w_{\bullet j} \neq 0 & \forall j \in V_1, \\ (w_{\bullet i}, \beta_i) \neq \pm (w_{\bullet j}, \beta_j) & \forall i, j \in V_1 \text{ with } i \neq j \end{array} \right\}.$$

For $\theta \notin \mathcal{E}_0$ one of the inequalities must be violated. We consider all possibilities:

1. *Deactivation redundancy:* If there is a neuron $j \in V_1$ such that $w_{j\bullet} = 0$, then the parameter θ has a deactivation redundancy (Remark 8.2.2) as the output of the neuron is ignored and the parameter violates requirement (a) of Definition 8.2.1. Thus $\theta \notin \mathcal{E}$.

2. *Bias redundancy:* There is a neuron $k \in V_1$ such that $w_{\bullet k} = 0$. Since this implies that the output of neuron k is constant and equal to $\psi(\beta_k)$ irrespective of the input x it falls into the category of bias redundancies (Remark 8.2.2). Specifically, the realization function can be replicated by removing the neuron k and adjusting the bias. This allows for a non-trivial linear combination

$$\lambda_\emptyset + \sum_{j \in V_1} \lambda_j \psi\Big(\beta_j + \sum_{i \in V_0} x_i w_{ij}\Big) = 0 \tag{8.13}$$

   with

$$\lambda_j = \begin{cases} -\psi(\beta_k), & \text{if } j = \emptyset, \\ 1, & \text{if } j = k, \\ 0, & \text{if } j \in V_1 \setminus \{k\}. \end{cases}$$

   This is a violation of (b) from Definition 8.2.1 and we thus have $\theta \notin \mathcal{E}$.

3. *Duplication redundancy:* There are two neurons $k, \ell \in V_1$ such that their parameters are equal $(w_{\bullet k}, \beta_k) = (w_{\bullet \ell}, \beta_\ell)$. We call this a duplication redundancy since both neurons have identical parameters and therefore identical output. Here

$$\lambda_j = \begin{cases} 1, & \text{if } j = k, \\ -1, & \text{if } j = \ell, \\ 0, & \text{if } j \in \{0\} \cup (V_1 \setminus \{k, \ell\}) \end{cases}$$

defines a nontrivial solution for (8.13), violating (b) of Definition 8.2.1. Thus $\theta \notin \mathcal{E}$. Note that deactivation redundancies fall into the category of 'generalized deactivation redundancies' in Remark 8.2.2.

Observe that so far, we have not made use of $\psi \in \{\text{sigmoid}, \tanh\}$. This leads to the following upper bound on the set of efficient parameters regardless of the activation function.

Lemma 8.2.6 (General redundancies)**.** *Let $\mathfrak{N} = (G, \psi)$ with $G = (V, E)$ be a shallow neural network. Then the set of efficient parameters $\mathcal{E}$ satisfies*

$$\mathcal{E} \subseteq \left\{ \theta = (w, \beta) \in \mathbb{R}^{E \times (V \setminus V_0)} : \begin{array}{ll} w_{j\bullet} \neq 0 & \forall j \in V_1, \\ w_{\bullet j} \neq 0 & \forall j \in V_1, \\ (w_{\bullet i}, \beta_i) \neq (w_{\bullet j}, \beta_j) & \forall i \neq j \in V_1 \end{array} \right\} =: \bar{\mathcal{E}}.$$

Proof. The general arguments 1, 2 and 3 imply $\bar{\mathcal{E}}^{\mathfrak{C}} \subseteq \mathcal{E}^{\mathfrak{C}}$. □

Continuing with our proof of $\mathcal{E} \subseteq \mathcal{E}_0$ there is one possible constraint violation left for $\theta \notin \mathcal{E}_0$:

4. *Sign-symmetric redundancy:* There are two neurons $i, j \in V_1$ such that $(w_{\bullet i}, \beta_i) = -(w_{\bullet j}, \beta_j)$. The reason this results in a redundancy are symmetries of the activation function ψ. Details in Lemma 8.2.7 and 8.2.8.

Lemma 8.2.7 (sigmoid)**.** *Let $\psi = \text{sigmoid}$ and let θ be a parameter such that there exist $k, \ell \in V_1$ with $(w_{\bullet k}, \beta_k) = -(w_{\bullet \ell}, \beta_\ell)$. Then θ is not efficient.*

Proof. Observe that we have for all $x \in \mathbb{R}$ that

$$\text{sigmoid}(-x) = \frac{e^{-x}}{1 + e^{-x}} = \frac{1 + e^{-x}}{1 + e^{-x}} - \frac{1}{1 + e^{-x}} = 1 - \text{sigmoid}(x). \tag{8.14}$$

This allows for a non-trivial linear combination

$$\lambda_0 + \sum_{j \in V_1} \lambda_j \psi\Big(\beta_j + \sum_{i \in V_0} x_i w_{ij}\Big) = 0$$

via

$$\lambda_j = \begin{cases} 1, & \text{if } j \in \{k, \ell\}, \\ 0, & \text{if } j \in V_1 \setminus \{k, \ell\}, \\ -1, & \text{if } j = \emptyset. \end{cases}$$

This nontrivial solution violates (b) of Definition 8.2.1 and thus implies $\theta \notin \mathcal{E}$. □

Lemma 8.2.8 (tanh)**.** *Let $\psi = \tanh$ and let θ be a parameter such that there exist $k, \ell \in V_1$ with $(w_{\bullet k}, \beta_k) = -(w_{\bullet \ell}, \beta_\ell)$. Then θ is not efficient.*

Proof. Observe that we have for every $x \in \mathbb{R}$

$$\tanh(-x) = \frac{e^{-x} - e^{x}}{e^{-x} + e^{x}} = -\tanh(x). \tag{8.15}$$

Similarly, to the proof of Lemma 8.2.7 we use this symmetry to construct a non-trivial linear combination via

$$\lambda_j = \begin{cases} 1, & \text{if } j \in \{k, \ell\}, \\ 0, & \text{if } j \in \{\emptyset\} \cup V_1 \setminus \{k, \ell\} \end{cases}$$

violating (b) of Definition 8.2.1 and finishing the proof. □

The redundancy that is treated in the Lemmas 8.2.7 and 8.2.8 is caused by certain symmetries in the particular activation function. In general, the particular structure of the activation function can cause complex additional redundancies.

Example 8.2.9 (Softplus)**.** Consider the case of the softplus activation function $\psi(x) = \ln(1+\exp(x))$. If we have $(w_{\bullet i}, \beta_i) = -(w_{\bullet j}, \beta_j)$, then

$$\begin{aligned}\psi\big(\beta_i + \langle x, w_{\bullet i}\rangle\big) &= \ln\big(1+\exp\big[-\big(\beta_j + \langle x, w_{\bullet j}\rangle\big)\big]\big)\\ &= -\big(\beta_j + \langle x, w_{\bullet j}\rangle\big) + \ln\big(1+\exp\big[\beta_j + \langle x, w_{\bullet j}\rangle\big]\big)\\ &= \big(\beta_i + \langle x, w_{\bullet i}\rangle\big) + \psi\big(\beta_j + \langle x, w_{\bullet j}\rangle\big).\end{aligned}$$

So the neurons are equal up to a linear term, which can be cancelled out by another sign pair $(w_{\bullet k}, \beta_k) = -(w_{\bullet l}, \beta_l)$ and appropriate $w_{k\bullet}$ and $w_{l\bullet}$. For this it is important to keep in mind that we only need to take care of the first order term, as the zero order term can be absorbed by the bias β_m for $m \in V_2$.

Pruned networks in the case of the softplus function may therefore include one sign symmetry, but not more. The set of efficient parameters is therefore slightly larger. On the other hand polynomial efficiency results in the set $\mathcal{E}_0$ again if $P^{(\emptyset)}$ in Definition 8.1.1 may be of degree 1 (i.e. $m_\emptyset \geq 1$) as the linear term can be absorbed by a polynomial.

8.2.2 Proof of $\mathcal{E}_0 \subseteq \mathcal{E}_P^m$ in the case of one dimensional input

Let $\theta \in \mathcal{E}_0$. We need to show θ to be m-polynomially independent (Definition 8.1.1). In this first step we assume one dimensional input, i.e. $V_{\text{in}} = \{\blacklozenge\}$.

For ease of notation, we define $\alpha_j := w_{\blacklozenge j}$ for all $j \in V_1$ and write with slight misuse of notation $x := x_\blacklozenge$. We thus have to prove that if the representation

$$0 = P^{(\emptyset)}(x) + \sum_{j\in V_1}\sum_{k=0}^{n} P_j^{(k)}(x)\psi^{(k)}\big(\beta_j + \alpha_j x\big) \tag{8.16}$$

is true for all $x \in \mathbb{R}$,[8] then all polynomials $P_j^{(k)}$ in the equation are zero. We will use complex analysis to show this. Recall that the activation functions we consider are given by

$$\begin{aligned}\text{sigmoid}(x) &= \frac{1}{1+e^{-x}}\\ \tanh(x) &= \frac{e^x - e^{-x}}{e^x + e^{-x}} = \frac{1-e^{-2x}}{1+e^{-2x}}.\end{aligned}$$

[8] Recall that we could translate $x \in \mathcal{X}$ to $x \in \mathbb{R}^{V_{\text{in}}}$ without loss of generality since the activation functions tanh and sigmoid are analytic and Theorem 8.2.3 assumes an open set is contained in $\mathcal{X}$.

Both activation functions are given as quotients of entire functions and are thus meromorphic functions on $\{x \in \mathbb{C} : e^{-x} \neq -1\}$ and $\{x \in \mathbb{C} : e^{-2x} \neq -1\}$, respectively. The singularities form an infinite chain on the imaginary axis of the form

$$z_m = cmi \qquad (m \in \mathbb{Z}_{\text{odd}}),$$

where $\mathbb{Z}_{\text{odd}}$ denotes the odd integers and $c = \pi$ in the case of sigmoid and $c = \frac{1}{2}\pi$ in the case of tanh.

For every neuron $j \in V_1$ and $k = 0, \dots, n$ the singularities of the meromorphic function

$$x \mapsto \psi^{(k)}(\alpha_j x + \beta_j)$$

are of the form

$$z_m^{(j)} = \frac{cmi - \beta_j}{\alpha_j} \qquad (m \in \mathbb{Z}_{\text{odd}}).$$

In particular, the singularities $S_j := \{z_m^{(j)} : m \in \mathbb{Z}_{\text{odd}}\}$ do not depend on k and we call them the singularities of neuron j. Consequently, the right hand side of (8.16) always defines a meromorphic function on

$$\mathbb{C} \setminus \bigcup_{j\in V_1} S_j.$$

In order to have equality in (8.16), in particular, all singularities have to be liftable.

Now suppose we are given a nontrivial solution to (8.16). We will show that this would cause a contradiction as consequence of the following principles that we will state in detail and prove below:

1) We call a neuron $j \in V_1$ *active*, if one of the polynomials $P_j^{(k)}$ with $k \in \{0,\dots,n\}$ is nonzero. If the set of active neurons is non-empty, there is an active neuron $j^* \in V_1$ such that infinitely many of its singularities in S_{j^*} are not "served" by another active neuron meaning that the singularity does not lie in one of the singularity sets of the other active neurons.

2) If $j \in V_1$, $z \in S_j$ and $k \in \{0,\dots,n\}$ with $P_j^{(k)}(z) \neq 0$, then the merophormorphic function
$$\sum_{k=0}^{n} P_j^{(k)}(x)\psi^{(k)}(\alpha_j x + \beta_j)$$
has a non-liftable singularity in z.

Indeed, the two facts then almost immediately imply a contradiction if a non-trivial solution to (8.16) would exist: we fix j^* as in 1) and observe that for infinitely many $z \in S_{j^*}$, the singularity z is not served by other neurons. This entails that for all these z
$$x \mapsto \sum_{k=0}^{n} P_{j^*}^{(k)}(x)\psi^{(k)}(\alpha_{j^*} x + \beta_{j^*})$$
has a liftable singularity in z. Now 2) implies that for all these z and $k = 0,\dots,n$, $P_{j^*}^{(k)}(z) = 0$. Since there is no polynomial that satisfies this for an infinite number of points z, we produced a contradiction showing that j^* is not active.

It remains to prove the two statements.

Lemma 8.2.10. *Let $\theta \in \mathcal{E}_0$ and suppose there exists a nontrivial solution to (8.16). There is an active neuron $j^* \in V_1$ (meaning that there exists $k \in \{0,\dots,n\}$ with $P_k^{(j^*)} \not\equiv 0$) such that*
$$\#\Big(S_{j^*} \setminus \bigcup_{j \text{ active: } j \neq j^*} S_j\Big) = \infty.$$

Proof. First note that if none of the neurons $j \in V_1$ would be active then one has that $0 \equiv P^{(\emptyset)}$ which would imply that the solution is trivial.

Now fix an active neuron $j^* \in V_1$ that maximizes $|\alpha_j|$ over all active neurons j and denote by V_1^* the set of all active neurons with
$$\frac{\beta_j}{\alpha_j} = \frac{\beta_{j^*}}{\alpha_{j^*}}.$$
If there were another $j \in V_1^* \setminus \{v^*\}$ with $|\alpha_j| = |\alpha_{j^*}|$, then we get together with $\frac{\beta_j}{\alpha_j} = \frac{\beta_{j^*}}{\alpha_{j^*}}$ that (α_j, β_j) equals either $(\alpha_{j^*}, \beta_{j^*})$ or $(-\alpha_{j^*}, -\beta_{j^*})$. But this is not possible since $\theta \in \mathcal{E}_0$. Hence, for all $j \in V_1^* \setminus \{v^*\}$ we have that $|\alpha_j| < |\alpha_{j^*}|$.

Now let $j \in V_1^* \setminus \{v^*\}$. Then $z_m^{(j^*)}$ with $m \in \mathbb{Z}_{\text{odd}}$ is in S_j iff there exists $m' \in \mathbb{Z}_{\text{odd}}$ with $m'/\alpha_j = m/\alpha_{j^*}$ or, equivalently,
$$\frac{\alpha_{j^*}}{\alpha_j} m' = m. \tag{8.17}$$
This entails that whenever α_{j^*}/α_j is not a rational we have that $S_j \cap S_{j^*} = \emptyset$. Now suppose that α_{j^*}/α_j is a rational and let $p \in \mathbb{Z}$ and $q \in \mathbb{N}$ such that the representation $\alpha_{j^*}/\alpha_j = p/q$ is minimal (meaning that $|p|$ and q are minimal). Since $|\alpha_{j^*}| > |\alpha_j|$ we have that $|p| \geq 2$. The integers p and q have no common divisor and equation (8.17) can only be true if m' is a multiple of q. Consequently, for all but at most one odd prime number m (namely $|p|$) we have that $z_m^{(j^*)} \notin S_j$. Since there are only finitely many neurons in V_1^* we conclude that all but finitely many odd primes m correspond to singularities $z_m^{(j^*)}$ of neuron j^* that are not served by other active neurons. □

Lemma 8.2.11. *Let $\alpha \in \mathbb{R}\backslash\{0\}$, $\beta \in \mathbb{R}$ and $z^* \in \mathbb{C}$ be a singularity of a meromorphic function ψ. Let $m \in \mathbb{N}$ and for every $k \in \{0,\dots,m\}$ let $P^{(k)} : \mathbb{C} \to \mathbb{C}$ be a polynomial such that either $P^{(k)} \equiv 0$ or $P^{(k)}\big(\frac{z^*-\beta}{\alpha}\big) \neq 0$. Then the polynomial*

$$\Phi(x) = \sum_{k=0}^{m} P^{(k)}(x)\psi^{(k)}(\alpha x + \beta),$$

has a non-liftable singularity in $\frac{z^*-\alpha}{\beta}$ or $P^{(k)} \equiv 0$ for all $k = 0,\dots,m$.

Proof. By defining

$$\tilde{\Phi}(y) := \Phi\Big(\frac{y-\beta}{\alpha}\Big) = \sum_{k=0}^{m} P^{(k)}\Big(\frac{y-\beta}{\alpha}\Big)\psi^{(k)}(y),$$

and absorbing α, β into the polynomials we can assume without loss of generality that $\alpha = 1$ and $\beta = 0$. Since these polynomials are zero if and only if the modified polynomials are zero.

As a meromorphic function, all singularities are poles, i.e., there exists a smallest order r such that

$$(z - z^*)^r \psi(z)$$

can be continuously extended in z^*. Moreover its Laurent series is then of the form

$$\psi(z) = \sum_{l=-r}^{\infty} a_l (z - z^*)^l$$

with $a_{-m} \neq 0$. This representation allows us to argue that the order of the pole increases with every derivative and therefore we cannot cancel out the singularities with a simple linear combination. So either we retain the singularity, or we have to set everything to zero. More specifically, we have

$$\begin{aligned}\psi^{(k)}(z) &= \sum_{l=-r}^{\infty} a_l l \cdots (l-k+1)(z-z^*)^{l-k} \\ &= \sum_{l=-(r+k)}^{\infty} a_{l+k}(l+k)\cdots(l+1)(z-z^*)^l\end{aligned}$$

Let us assume that $P^{(m)} \neq 0$, then we have $P^{(m)}(z^*) \neq 0$ and thus

$$\begin{aligned}&\Big|(z-z^*)^{r+m-1}\Phi(z)\Big| \\ &\geq \Bigg|\underbrace{\big|(z-z^*)^{r+m-1}P^{(m)}(z)\psi^{(m)}(z)\big|}_{\to\infty} - \underbrace{\Big|(z-z^*)^{r+m-1}\sum_{k=0}^{m-1}P^{(k)}(z)\psi^{(k)}(z)\Big|}_{\to c\in\mathbb{R}}\Bigg| \\ &\to \infty \qquad (z \to z^*).\end{aligned}$$

Therefore Φ has a singularity at z^* (more specifically a pole of order at least $r+m$). If $P^{(m)} \equiv 0$, we repeat the same argument with $m-1$ until we have either a pole at z^* or $P^{(m)} \equiv \cdots \equiv P^{(0)} \equiv 0$. □

8.2.3 Proof of $\mathcal{E}_0 \subseteq \mathcal{E}_P^m$, general case

In this section we reduce the proof of $\mathcal{E}_0 \subseteq \mathcal{E}_P^m$ to the one dimensional result we have already proven in Subsection 8.2.2. For an element $\theta \in \mathcal{E}_0$ recall that this requires that the equation

$$0 = P^{(\emptyset)}(x) + \sum_{j\in V_1}\sum_{k=0}^{n} P_j^{(k)}(x)\psi^{(k)}\big(\beta_j + \langle x, w_{\bullet j}\rangle\big), \quad \forall x \in \mathbb{R}^{V_{\text{in}}}$$

has the unique solution of all polynomials being zero (Definition 8.1.1). Since this equation holds for all inputs $x \in \mathbb{R}^{V_{\text{in}}}$, it holds in particular for 1-dimensional slices

$$x_v(\lambda) := \lambda v, \qquad \lambda \in \mathbb{R}$$

for directions $v \in \mathbb{R}^{V_{\text{in}}}$. The following proof hinges on the fact that the mappings

$$\lambda \mapsto P_j^{(k)}(x_v(\lambda))$$

are polynomials in t, which we can prove to be zero with the one dimensional result. If sufficiently many appropriate directions v are chosen and all directional polynomials are zero, then Theorem 8.2.13 allows us to deduce that the original polynomial is zero. But to apply the one dimensional result, we need to ensure that we do not introduce degeneracies with the directions we choose. For example

$$\lambda \mapsto \langle x_v(\lambda), w_{\bullet j}\rangle = \lambda \langle v, w_{\bullet j}\rangle$$

may be constantly zero if the direction v is orthogonal to $w_{\bullet,j}$. This introduces a bias redundancy. For the number of directions $N = \binom{\max(m)+\#V_{\text{in}}-1}{\max(m)}$ we therefore want to select $v^{(1)}, \dots, v^{(N)} \in \mathbb{R}^{V_{\text{in}}}$ such that the following conditions hold at the same time:

1. the directions $v^{(l)}$ characterize polynomials in the sense of Theorem 8.2.13. This is required for us to deduce that the original polynomials are zero.
2. $\alpha_k^{(l)} := \langle v^{(l)}, w_{\bullet k}\rangle \neq 0$ for all $k \in V_1$.
3. $(\alpha_i^{(l)}, \beta_i) \neq \pm(\alpha_j^{(l)}, \beta_j)$ for all $i, j \in V_1$ with $i \neq j$.

The last two requirements ensure non-redundancy for the parameters of the surrogate neural networks.

Why is this possible? Select iid entries $v_i^{(l)} \sim \mathcal{N}(0,1)$ with $l \in \{1, \dots, N\}$ and $i \in V_{\text{in}}$. Then we satisfy the conditions of Theorem 8.2.13 and almost surely have the first condition. Since $w_{\bullet k} \neq 0$ for all $k \in V_1$ by assumption, we also have almost surely

$$\alpha_k^{(l)} = \langle v^{(l)}, w_{\bullet k}\rangle \neq 0.$$

That is we have the second condition almost surely. For the last condition, we use the assumption $(w_{\bullet i}, \beta_i) \neq \pm(w_{\bullet j}, \beta_j)$. Let us only consider the case where "$\pm$" is "$+$". The other case is analogous. Then by assumption, we have

$$(w_{\bullet i}, \beta_i) - (w_{\bullet j}, \beta_j) \neq 0$$

and thus either $w_{\bullet i} - w_{\bullet j} \neq 0$ (case 1) or $\beta_i - \beta_j \neq 0$ (case 2). This implies

$$(\alpha_i^{(l)}, \beta_i) - (\alpha_j^{(l)}, \beta_j) = \big(\langle v^{(l)}, \underbrace{w_{\bullet i} - w_{\bullet j}}_{\substack{\text{case 1}\\ \neq\ 0}}\rangle, \underbrace{\beta_i - \beta_j}_{\substack{\text{case 2}\\ \neq\ 0}}\big) \neq 0.$$

Note that due to the iid selection of the entries of $v^{(l)}$ the inequality of the tuple with zero only holds almost surely in case 1.

In summary, with the selection of random $v^{(l)}$ we can satisfy all three conditions almost surely. In particular, there *exist* directions $v^{(1)}, \dots, v^{(N)}$ which satisfy all three conditions simultaneously.

1-dimensional slices of the $\#V_{\text{in}}$-dimensional input For the network $\mathfrak{N} = (\mathbb{V}, \psi)$ we consider the 1-dimensional network $\tilde{\mathfrak{N}} := (\tilde{\mathbb{V}}, \psi)$ whose input layer is reduced to a single node

$$\tilde{\mathbb{V}} := (\{\blacklozenge\}, V_1, V_{\text{out}}).$$

From parameters $\theta = (w, \beta)$ of $\mathfrak{N}$ and direction $v^{(l)}$ we construct parameters $\theta^{(l)} = (w^{(l)}, \beta)$ of $\widetilde{\mathfrak{N}}$ by retaining the bias β and all connections from the hidden layer V_1 to the output V_{out} that is $w^{(l)}_{\bullet k} := w_{\bullet k}$ for all $k \in V_{\text{out}}$. For the input layer connections, we set

$$w^{(l)}_{\blacklozenge j} := \alpha^{(l)}_j = \langle v^{(l)}, w_{\bullet j} \rangle \qquad \forall j \in V_1.$$

Then $\theta^{(l)} \in \mathcal{E}_0$ since we have

$$\begin{aligned}
w^{(l)}_{\blacklozenge i} \neq 0 \quad \forall i \in V_1 \qquad & \text{due to } \alpha^{(l)}_j \neq 0, \\
w^{(l)}_{i\bullet} \neq 0 \quad \forall i \in V_1 \qquad & \text{due to } w^{(l)}_{i\bullet} = w_{i\bullet} \neq 0, \\
(w^{(l)}_{\blacklozenge i}, \beta_i) \neq \pm(w^{(l)}_{\blacklozenge j}, \beta_j) \quad \forall i \neq j \in V_1 \qquad & \text{due to } (\alpha^{(l)}_i, \beta_i) \neq \pm(\alpha^{(l)}_j, \beta_j).
\end{aligned}$$

Remark 8.2.12 (Response function slices). The response functions $\Psi_{\theta^{(l)}}$ of the new parameter $\theta^{(l)}$ has the following relation with the response Ψ_θ

$$\Psi_{\theta^{(l)}}(\lambda) = \Psi_\theta(\lambda v^{(l)}) = \Psi_\theta(x_{v^{(l)}}(\lambda)) \qquad \forall \lambda \in \mathbb{R}.$$

Using the slices for the proof of polynomial independence We are now finally ready to prove polynomial independence for multi-dimensional input. To do so, assume we have

$$0 = P^{(\emptyset)}(x) + \sum_{j \in V_1} \sum_{k=0}^{m} P^{(k)}_j(x) \psi^{(k)}\Big(\beta_j + \langle x, w_{\bullet j} \rangle\Big) \quad \forall x \in \mathbb{R}^{V_{\text{in}}}.$$

In particular we can select $x = \lambda v^{(l)}$ to obtain

$$0 = P^{(\emptyset)}(\lambda v^{(l)}) + \sum_{j \in V_1} \sum_{k=0}^{m} P^{(k)}_j(\lambda v^{(l)}) \psi^{(k)}\Big(\beta_j + w^{(l)}_{\blacklozenge j} \lambda\Big) \quad \forall \lambda \in \mathbb{R}.$$

By the polynomial independence of the 1-dimensional input network slices $\mathfrak{N}^{(l)}$, we then have that the 1-dimensional polynomial slices

$$\lambda \mapsto P^{(k)}_j(\lambda v^{(l)})$$

are all identically zero. As we selected the directions $v^{(1)}, \ldots, v^{(N)}$ to characterize polynomials in the sense of Theorem 8.2.13, we thus have $P^{(k)}_j \equiv 0$ for all $j \in V_1$ and all $k = \beta, 0, \ldots, n$. That is, polynomial independence for the case of multivariate input.

8.2.4 Polynomial slicing

The main tool to translate the 1-dimensional input result to the general case are slices of polynomials that characterize the full polynomial. This is formalized in the following theorem, which is proven in the remainder of this section.

Theorem 8.2.13 (Optimal polynomial slicing)**.** *Let $d, n \in \mathbb{N}$ and $N = \binom{n+d-1}{n}$.*

(I) ***Almost all selections of directions** $v_1, \ldots, v_N$ **characterize the** d**-variate polynomials of degree** n**.** If the matrix $(v_1, \ldots, v_N) \in \mathbb{R}^{d \times N}$ is selected randomly with a density with respect to the Lebesgue measure on $\mathbb{R}^{d \times N}$ (in particular there exist such v_i), then the following property holds almost surely:*

For any d-variate polynomial $p \in \mathbb{R}[x_1, \ldots, x_d]$ of order n we have $p \equiv 0$ if and only if all slices in the directions v_i are zero, i.e.

$$p_{v_i}(\lambda) := p(\lambda v_i) = 0 \quad \forall \lambda \in \mathbb{R}, \quad \forall i = 1, \ldots, N.$$

(II) ***The number N of directions is optimal.*** *That is, for any smaller selection of directions $v_1, \dots, v_M \in \mathbb{R}^d$ with $M < N$ there exists a* ***non-zero*** *d-variate polynomial $p \in \mathbb{R}[x_1, \dots, x_d]$ of order N such that all the slices p_{v_i} are identically zero.*

A crucial object in the proof of this theorem is the vector of all monomials of degree n. In the multivariate case, the definition of such a vector requires an ordering of the tuple of powers.

Definition 8.2.14 (Vector of monomials)**.** We define

$$\operatorname{mon}_n(x) = \left(\prod_{i=1}^{d} x_i^{r_i} \right)_{r \in R} \quad \text{for} \quad x \in \mathbb{R}^d$$

where R is a subset of d-tuples of non-negative integers that form monomials of exactly degree n

$$R = \left\{ r \in \mathbb{N}_0^d : \sum_{i=1}^{d} r_i = n \right\}.$$

With the following injection into the ordered set of non-negative integers $\mathbb{N}_0$

$$\phi : \begin{cases} R \to \mathbb{N}_0 \\ r \mapsto \sum_{i=1}^{d} r_i (n+1)^{i-1}, \end{cases}$$

we equip R with the pullback of this order. That is we define $r < \tilde{r}$ if and only if $\phi(r) < \phi(\tilde{r})$. In other words, we assume R has reverse lexicographic ordering.

This ensures $\operatorname{mon}_n(x)$ to be a vector and not just an unordered set.

Proposition 8.2.15 (Independent monomials)**.** *Let $d, n \in \mathbb{N}$ and $N = \binom{n+d-1}{n}$. Then there exist $v_1, \dots, v_N$ such that $\operatorname{mon}_n(v_i)$ are linearly independent.*

Almost all selections of $v_1, \dots, v_N$ have this property. That is, if the directions $(v_1, \dots, v_N) \in \mathbb{R}^{d \times N}$ are random variables with a density with respect to the Lebesgue measure on $\mathbb{R}^{d \times N}$, then $\operatorname{mon}_n(v_i)$ are almost surely linearly independent.

Proof. Using $0 < a_1 < \cdots < a_N$ with $a_i \in \mathbb{R}$, we define

$$v_k = \left(a_k, a_k^{n+1}, \dots, a_k^{(n+1)^{d-1}} \right)$$

Then we have by definition of v_k, mon_n and ϕ

$$\operatorname{mon}_n(v_k) = \left(\prod_{i=1}^{d} (v_k^{(i)})^{r_i} \right)_{r \in R} = \left(\prod_{i=1}^{d} a_k^{r_i (n+1)^{i-1}} \right)_{r \in R} = (a_k^{\phi(r)})_{r \in R}.$$

Therefore we have

$$(\operatorname{mon}_n(v_1), \dots, \operatorname{mon}_n(v_N)) = \begin{pmatrix} a_1^{\lambda_1} & \cdots & a_N^{\lambda_1} \\ \vdots & & \vdots \\ a_1^{\lambda_N} & \cdots & a_N^{\lambda_N} \end{pmatrix}, \tag{8.18}$$

with $(\lambda_1, \dots, \lambda_N) = (\phi(r))_{r \in R} \subseteq \mathbb{N}_0$, where the size of the set R is given by Lemma 8.2.16 and we have $\lambda_1 < \cdots < \lambda_N$ by the ordering defined for R in Definition 8.2.14. But the matrix (8.18) is a generalized Vandermonde matrix as in Lemma 8.2.17 and its determinant is thus not equal to zero by Lemma 8.2.17. The monomial vectors are therefore linearly independent.

Observe that $\det(\operatorname{mon}_n(v_1), \dots, \operatorname{mon}_n(v_N))$ is a multivariate polynomial in the entries of v_i. In particular it is a (real) analytic function. By Mityagin (2020) the zero set of a real analytic which is not identically zero is a Lebesgue zero set. Since we found an example above where this determinant is non-zero, we ruled out that the function is identically zero. Thus almost all selections of $v_1, \dots, v_N$ result in a non-zero determinant. □

Lemma 8.2.16 (Number of monomials)**.** $|R| = N = \binom{n+d-1}{n}$

Proof. The number N is also sometimes referred to as "d multichoose n" and denoted by

$$\left(\!\binom{d}{n}\!\right) = \binom{n+d-1}{n}$$

as it is equal to the number of ways to create a multiset of size n from d elements. In our case, we are picking an n-sized multiset of x_i to finally multiply all elements together to obtain a monomial. Details of the proof can be found in textbooks such as Stanley (2011, pp. 25–26) or Riordan (2002, Sec. 3.2) □

Lemma 8.2.17 (Generalized Vandermonde)**.** *Let $0 < a_1 < \cdots < a_N$ for $a_i \in \mathbb{R}$ and $\lambda_1 < \cdots < \lambda_N$ with $\lambda_i \in \mathbb{R}$. Then we have that the following generalized Vandermonde matrix has non-zero determinant, i.e.*

$$\det \begin{pmatrix} a_1^{\lambda_1} & \cdots & a_N^{\lambda_1} \\ \vdots & & \vdots \\ a_1^{\lambda_N} & \cdots & a_N^{\lambda_N} \end{pmatrix} \neq 0.$$

Proof. The proof is adapted from a stack exchange answer.[9] We conduct an induction over N and note that for the induction start $N = 1$ the conclusion obviously holds.

[9] Szwarc, 2022.

For the induction step $N - 1 \to N$, assume that there exist $c_1, \ldots, c_N \in \mathbb{R}$ such that the rows weighted by c_i of the generalized Vandermonde sum to zero, i.e. we have for all columns k

$$c_1 a_k^{\lambda_1} + \cdots + c_N a_k^{\lambda_N} = 0.$$

In order to prove linear independence of these rows and thus that the determinant is zero, we only need to show that these equations imply $c_i = 0$ for all $i = 1, \ldots, N$.

Dividing the equations above by $a_k^{\lambda_1}$, we observe that the a_k are zeros of the function

$$f(x) = c_1 + c_2 x^{\lambda_2 - \lambda_1} + \cdots + c_N x^{\lambda_N - \lambda_1}$$

Since $\lambda_k - \lambda_1 > 0$ by assumption, f is a continuously differentiable function. Between any two points where f is zero there is therefore a point where its derivative

$$f'(x) = c_2(\lambda_2 - \lambda_1)x^{\lambda_2 - \lambda_1 - 1} + \cdots + c_N(\lambda_N - \lambda_1)x^{\lambda_N - \lambda_1 - 1}$$

is zero by the mean value theorem. In the gaps of $a_1 < \cdots < a_n$ are thus $N - 1$ points $0 < u_1 < \ldots, u_{N-1}$ such that f' is zero at all u_i. Since by induction hypothesis we have

$$\det \begin{pmatrix} u_1^{\tilde{\lambda}_1} & \cdots & u_{N-1}^{\tilde{\lambda}_1} \\ \vdots & & \vdots \\ u_1^{\tilde{\lambda}_{N-1}} & \cdots & u_N^{\tilde{\lambda}_{N-1}} \end{pmatrix} \neq 0$$

for $\tilde{\lambda}_i = \lambda_{i+1} - \lambda_1 - 1$, we have that the rows of this matrix are linearly independent. And since we have that all u_k are zeros of f', we have for the weighted colum sums

$$0 = c_2(\lambda_2 - \lambda_1)u_k^{\tilde{\lambda}_1} + \cdots + c_N(\lambda_N - \lambda_1)u_k^{\tilde{\lambda}_{N-1}}$$

for all k. By linear independence of the rows this implies that the coefficients $c_i(\lambda_i - \lambda_1)$ for $i \geq 2$ have to be be zero. Since $\lambda_i - \lambda_1 > 0$, this implies $c_2 = \cdots = c_N = 0$. We thus have $f \equiv c_1$ and since the a_k are zeros of f this also implies $c_1 = 0$. Thus all c_i are zero which is what we needed to prove. □

Proof of polynomial slicing (Theorem 8.2.13)

(I). For the proof of the first statement of the theorem we intend to use the directions $v_1, \dots, v_N$ of Proposition 8.2.15 which result in linearly independent monomials. Note that these are only monomials of *exactly* degree n, while the polynomials of degree n admit all monomials *up to* degree n.

We address this difference with an induction over the degree and by sorting the monomials of the polynomials into buckets with the same degree. The induction step is enabled by the following lemma.

Lemma 8.2.18. *If the monomials* $\mathrm{mon}_m(v_1), \dots, \mathrm{mon}_m(v_N)$ *span the space, then the lower degree monomials* $\mathrm{mon}_k(v_1), \dots, \mathrm{mon}_k(v_N)$ *also span the space for all degrees* $k \leq m$.

Proof. Without loss of generality assume $k = m - 1$. Let K be the length of $\mathrm{mon}_k(x)$ and M be the length of $\mathrm{mon}_m(x)$ for $x \in \mathbb{R}^d$. Choose any $y \in \mathbb{R}^K$. We now have to prove that there is a linear combination of $\mathrm{mon}_k(v_i)$ equal to y.

Observe that the vector

$$x_1 \,\mathrm{mon}_k(x) = x_1 \,\mathrm{mon}_{m-1}(x) \qquad x \in \mathbb{R}^d \tag{8.19}$$

contains a subset of the entries of $\mathrm{mon}_m(x)$. We obtain the vector $\tilde{y} \in \mathbb{R}^M$ from $y \in \mathbb{R}^K$ by setting the positions of all other entries to zero. Since the monomials $\mathrm{mon}_m(v_1), \dots, \mathrm{mon}_m(v_N)$ span the space, there exists a linear combination

$$\tilde{y} = \sum_{i=1}^{N} c_i \,\mathrm{mon}_m(v_i).$$

By the observation (8.19) this implies

$$y = \sum_{i=1}^{N} \underbrace{c_i v_i^{(1)}}_{=:\tilde{c}_i} \mathrm{mon}_k(v_i).$$

Thus the $\mathrm{mon}_k(v_i)$ span the space. □

With this lemma we can now finish the proof of the first statement of the theorem. As already mentioned we take the directions $v_1, \dots, v_N$ of Proposition 8.2.15 and note that with this lemma we have that $\mathrm{mon}_m(v_1), \dots, \mathrm{mon}_m(v_N)$ span the space for all $m \leq n$. We proceed by induction over m up to n. That is, we assume that the polynomial p is of degree m and assume that $\mathrm{mon}_m(v_1), \dots, \mathrm{mon}_m(v_N)$ span the space but are not necessarily linearly independent (this is only the case for $m = n$).

The base case $m = 0$ is trivially true as one direction is enough to figure out if a constant polynomial is zero.

For the induction step $m - 1 \to m$ let p be a d-variate polynomial of degree m. We decomposition the polynomial into

$$p(x) = \sum_{k=0}^{n} p^{(k)}(x),$$

where the polynomials $p^{(k)}$ consist of all monomials of exactly degree k. For all $k < m$ we then have for all i

$$\lim_{\lambda \to \infty} \frac{p^{(k)}(\lambda v_i)}{\lambda^m} = 0. \tag{8.20}$$

With the assumption that the slices of p are zero, i.e. $p_{v_i}(\lambda) = 0$ we thus obtain

$$\begin{aligned} 0 = \lim_{\lambda \to \infty} \frac{p_{v_i}(\lambda)}{\lambda^m} &\overset{\text{def.}}{=} \lim_{\lambda \to \infty} \frac{p(\lambda v_i)}{\lambda^m} = \lim_{\lambda \to \infty} \sum_{k=0}^{n} \frac{p^{(k)}(\lambda v_i)}{\lambda^m} \overset{(8.20)}{=} \lim_{\lambda \to \infty} \frac{p^{(m)}(\lambda v_i)}{\lambda^m} \\ &= p^{(m)}(v_i). \end{aligned} \tag{8.21}$$

For the last equation we used that $p^{(m)}$ consists only of monomials of exactly degree m, for which we have

$$\operatorname{mon}_m(\lambda x) = \lambda^m \operatorname{mon}_m(x). \tag{8.22}$$

Since $p^{(m)}$ consists only of monomials of exactly degree m, there exists a vector $q \in \mathbb{R}^M$ with M the length of $\operatorname{mon}_m(x)$ for $x \in \mathbb{R}^d$ such that

$$p(x) = q^T \operatorname{mon}_m(x). \tag{8.23}$$

With (8.21) we thus have

$$q^T \underbrace{(\operatorname{mon}_m(v_1), \dots, \operatorname{mon}_m(v_N))}_{\in \mathbb{R}^{M \times N}} = 0.$$

As the monomials $\operatorname{mon}_m(v_1), \dots, \operatorname{mon}_m(v_N)$ span the space $\mathbb{R}^M$, the matrix has rank M and thus $q = 0$. By (8.23) we thus have $p^{(m)} \equiv 0$. Therefore p is of degree $m-1$ and we finish the proof of the first claim of the theorem using the induction assumption.

(II). Let $N = \binom{n+d-1}{n}$ and let $v_1, \dots, v_M$ be fewer directions $M < N$. To prove the statement, we construct a non-zero d-variate polynomial of degree n (more specifically it only consists of monomials of exactly degree n), which is zero in all directions v_i. To do so, consider the matrix

$$A := (\operatorname{mon}_n(v_1), \dots, \operatorname{mon}_n(v_M)) \in \mathbb{R}^{N \times M}.$$

Since $M < N$ it is at most of rank M there exists $0 \neq q \in \mathbb{R}^N$ such that $q^T A = 0$. We then define the polynomial $p(x) = q^T \operatorname{mon}_n(x)$. Then by construction this polynomial is zero at all v_i. Moreover, by the scaling property of the monomials (8.22) we have

$$p_{v_i}(\lambda) = q^T \operatorname{mon}_n(\lambda v_i) = \lambda^n \underbrace{q^T \operatorname{mon}_n(v_i)}_{= p(v_i)} = 0. \quad \forall \lambda \in \mathbb{R}.$$

Thus we have $p_{v_i} \equiv 0$ for all $i = 1, \dots, M$ but $p \neq 0$ since $q \neq 0$.

The intuition is in essence, that the scaling property of the monomials (8.22) implies that we only collect a single information point for each direction v_i. To ensure the polynomial is zero, we thus have to collect enough points v_i to ensure the polynomial has to be zero. As the space of polynomials of exactly degree n has dimension N, this is the number of points required.

8.3 The neighborhood of redundant parameters

In this section, we analyze the redundant domain. We will show that every redundant parameter lies on a line of (redundant) parameters for which the realization function is identical. In the setting with *no regularization*, i.e., $R \equiv 0$, this entails that redundant parameters always have a degenerate Hessian (in the sense that its determinant is zero) and it can never be a strict local minimum.

In a second step, we will show that for redundancies that are not deactivation redundancies typically either all or no points on the latter line are critical points of the optimization landscape.

Theorem 8.3.1 (Neighborhood of redundant critical points)**.** *Let $\mathfrak{N}$ be an ANN and assume θ is redundant, i.e. $\theta \in \Theta \setminus \mathcal{E}(\mathcal{X})$. Then there exists a straight line $\ell \subset \Theta$ containing θ such that for all $\vartheta \in \ell$*

$$\Psi_\vartheta = \Psi_\theta, \quad \text{on } \mathcal{X}.$$

Proof. If θ has a deactivation redundancy (Definition 8.2.1 (a)) and $w_{k\bullet} = 0$ for a $k \in V_1$, then changing the parameters $w_{i,j}$ and β_j ($i \in V_0, j \in V_1$) does have no impact on the response and clearly the respective set contains a line.

Now suppose that there is $\lambda \not\equiv 0$ such that

$$\lambda_\emptyset + \sum_{j \in V_1} \lambda_j \psi\Big(\beta_j + \sum_{i \in V_0} x_i w_{ij}\Big) = 0 \quad \forall x \in \mathcal{X}. \tag{8.24}$$

We define $\theta(t) = (w(t), \beta(t))$ for $t \in \mathbb{R}$ as follows: We retain the weights connecting the input to the first layer and its biases, i.e.

$$w_{ij}(t) := w_{ij} \quad \text{and} \quad \beta_j(t) := \beta_j \qquad \forall i \in V_0, j \in V_1,$$

and in the second layer we add multiples of λ in an appropriate way:

$$w_{jk}(t) := w_{jk} + t\lambda_j \quad \text{and} \quad \beta_k(t) := \beta_k + t\lambda_\emptyset \qquad \forall j \in V_1, k \in V_2.$$

Since $\lambda \neq 0$, the function $(\theta(t))_{t \in \mathbb{R}}$ parametrizes a line ℓ that contains θ and basic linear algebra implies with (8.24) that the response does not depend on the choice of t: In terms of

$$\psi_j(x) := \psi\Big(\beta_j + \sum_{i \in V_0} x_i w_{ij}\Big) \qquad \forall j \in V_1, x \in \mathcal{X},$$

one has for every $k \in V_2$ and $x \in \mathcal{X}$ that

$$\begin{aligned} (\Psi_{\theta(t)}(x))_k &= \beta_k(t) + \sum_{j \in V_1} w_{jk}(t)\psi_j(x) \\ &= \beta_k + \sum_{j \in V_1} w_{jk}\psi_j(x) + t \underbrace{\Big(\lambda_0 + \sum_{j \in V_1} \lambda_j \psi_j(x)\Big)}_{\overset{(8.24)}{=} 0} \\ &= (\Psi_\theta(x))_k. \end{aligned} \tag{8.25}$$

□

Theorem 8.3.2. *Let $\mathfrak{N}$ be an ANN, X and Y be $\mathbb{R}^{V_{in}}$- and $\mathbb{R}^{V_{out}}$-valued random variables, $\ell : \mathbb{R}^{V_{out}} \times \mathbb{R}^{V_{out}} \to [0, \infty)$ a C^1-function such that the cost landscape*

$$J(\theta) = \mathbb{E}[\ell(\Psi_\theta(X), Y)]$$

is C^1 on Θ, and differentiation and integration can be interchanged. Let the parameter $\theta = (w, \beta) \in \Theta$ exhibit a bias or duplication redundancy (cf. Remark 8.2.2) and let $\theta(t)$ be the parametrization of the line as introduced after (8.24). Then either

- *for all $t \in \mathbb{R}$, $\theta(t)$ is a critical parameter or*
- *there it at most one $t \in \mathbb{R}$, for which $\theta(t)$ is critical.*

If $\#V_{out} = 1$ and there are no deactivation redundancies, then θ being critical implies that $\theta(t)$ is critical for all $t \in \mathbb{R}$.

Proof. In the following, $i \in V_0, j \in V_1, k, l \in V_2, t \in \mathbb{R}$ and $x \in \mathcal{X}$ are arbitrary. Consider

$$\psi_j(x) := \psi\Big(\beta_j + \sum_{i \in V_0} x_i w_{ij}\Big) \quad \text{and} \quad \psi'_j(x) := \psi'\Big(\beta_j + \sum_{i \in V_0} x_i w_{ij}\Big)$$

Then we get for the inner differentials of the realization function

$$\partial_{w_{i,j}} \big(\Psi_{\theta(t)}(x)\big)_l = \psi'_j(x) x_i w_{j,l}(t) \quad \text{and} \quad \partial_{\beta_j} \big(\Psi_{\theta(t)}(x)\big)_l = \psi'_j(x) w_{j,l}(t)$$

and for the outer differentials

$$\partial_{w_{j,k}} \big(\Psi_{\theta(t)}(x)\big)_l = \delta_{k,l}\psi_j(x) \quad \text{and} \quad \partial_{\beta_k} \big(\Psi_{\theta(t)}(x)\big)_l = \delta_{k,l},$$

where δ denotes the Kronecker-Delta. By assumption, we have that

$$\nabla J(\theta(t)) = \mathbb{E}\Big[\sum_{l \in V_{\text{out}}} \partial_{\hat{y}_l} \ell(\Psi_{\theta(t)}(X), Y) \nabla (\Psi_{\theta(t)}(x))_l \Big]. \tag{8.26}$$

With the above identities we thus get for the inner and outer differentials

$$\begin{aligned}
\partial_{w_{i,j}} J(\theta(t)) &= \sum_{l \in V_{\text{out}}} w_{j,l}(t)\, \mathbb{E}\Big[\partial_{\hat{y}_l} \ell(\Psi_\theta(X), Y)\, \psi_j'(X)\, X_i\Big] \\
\partial_{\beta_j} J(\theta(t)) &= \sum_{l \in V_{\text{out}}} w_{j,l}(t)\, \mathbb{E}\Big[\partial_{\hat{y}_l} \ell(\Psi_\theta(X), Y)\, \psi_j'(X)\Big], \\
\partial_{w_{j,k}} J(\theta(t)) &= \mathbb{E}\Big[\partial_{\hat{y}_k} \ell(\Psi_\theta(X), Y)\, \psi_j(X)\Big] = \partial_{w_{j,k}} J(\theta), \\
\partial_{\beta_k} J(\theta(t)) &= \mathbb{E}\Big[\partial_{\hat{y}_k} \ell(\Psi_\theta(X), Y)\Big] = \partial_{\beta_k} J(\theta).
\end{aligned}$$

By the latter two identities, the derivatives with respect to the outer parameters do not depend on t. The inner derivatives can be expressed in terms of

$$a_{i,j,l} = \mathbb{E}\Big[\partial_{\hat{y}_l} \ell(\Psi_\theta(X), Y)\, \psi_j'(X)\, X_i\Big] \quad \text{and} \quad b_{j,l} = \mathbb{E}\Big[\partial_{\hat{y}_l} \ell(\Psi_\theta(X), Y)\, \psi_j'(X)\, X_i\Big],$$

specifically

$$\begin{aligned}
\partial_{w_{i,j}} J(\theta(t)) &= \sum_{l \in V_{\text{out}}} w_{j,l}(t)\, a_{i,j,l} = \sum_{l \in V_{\text{out}}} (w_{j,l} + t\lambda_j) a_{i,j,l}, \\
&= \sum_{l \in V_{\text{out}}} w_{j,l} a_{i,j,l} + t\lambda_j \sum_{l \in V_{\text{out}}} a_{i,j,l} \\
\partial_{\beta_j} J(\theta(t)) &= \sum_{l \in V_{\text{out}}} w_{j,l}(t)\, b_{j,l} = \sum_{l \in V_{\text{out}}} (w_{j,l} + t\lambda_j) b_{j,l} \\
&= \sum_{l \in V_{\text{out}}} w_{j,l} b_{j,l} + t\lambda_j \sum_{l \in V_{\text{out}}} b_{j,l}.
\end{aligned}$$

Note that all differentials $\partial_{w_{i,j}} J(\theta(t))$ and $\partial_{\beta_j} J(\theta(t))$ are affine functions in t. Hence, each differential is either zero for all $t \in \mathbb{R}$ or at most one point t. Consequently, on the line $(\theta(t))_{t \in \mathbb{R}}$ either all points are critical or there is at most one point that is critical. In the case of $\#V_{\text{out}} = 1$, and no deactivation redundancies, i.e. $w_{j\bullet} \neq 0$, a critical point in $t = 0$ implies $a_{i,j,l} = b_{j,l} = 0$ for all i, j and thereby all $\theta(t)$ are critical. □

8.4 Existence of efficient critical points

In this section, we analyze the existence of local minima in the standard setting (Def. 8.0.6). More explicitly, we prove that for every open set $U \subseteq \Theta$ containing a polynomially efficient parameter θ one has with strictly positive probability that the random unregularized squared error loss contains a local minimum in U. The result illustrates that local minima may exist in the unregularized setting. In the case of non-trivial regularization the cost typically tends to infinity when the parameter θ tends to infinity. In this case the existence of (local) minima is trivial.

Theorem 8.4.1 (Efficient minima exist with positive probability)**.** *Assume that we are in the unregularised (i.e., $R \equiv 0$) standard setting (Definition 8.0.6) and that the random target function $\mathbf{f} = (f_{\mathbf{M}}(\theta))_{\theta \in \Theta}$ additionally satisfies that for all continuous functions $\phi : \mathbb{R}^{V_{in}} \to \mathbb{R}$ and $\delta \in (0, \infty)$ one has*

$$\mathbb{P}(\|\mathbf{f} - \phi\|_{\mathbb{P}_X} < \delta) > 0.$$

Then every non-empty, open set U of $(0,0,1)$-polynomially efficient parameters contains a local minimum of the MSE cost with positive probability, i.e.

$$\mathbb{P}\Big(\exists \theta \in U : \theta \text{ is a local minimum of } \mathbf{J}\Big) > 0.$$

Proof of Theorem 8.4.1. Recall that for every $\mathfrak{m} \in \mathbb{M}$ the MSE cost function is of the form

$$\begin{aligned} J_{\mathfrak{m}}(\theta) &= \mathbb{E}_{\mathfrak{m}}[\|\Psi_\theta(X) - Y\|^2] \\ &\overset{(*)}{=} \mathbb{E}_{\mathfrak{m}}[\|\Psi_\theta(X) - f_{\mathfrak{m}}(X)\|^2] + \underbrace{\mathbb{E}_{\mathfrak{m}}[\|f_{\mathfrak{m}}(X) - Y\|^2]}_{\text{'noise' const. in } \theta}. \end{aligned}$$

For $(*)$ we note that $\mathbb{E}_{\mathfrak{m}}[Y - f_{\mathfrak{m}}(X) \mid X] = 0$ by definition of the target function $f_{\mathfrak{m}}(x) = \mathbb{E}_{\mathfrak{m}}[Y \mid X = x]$, and the mixed term therefore disappears. Since we assumed $\#V_{\text{out}} = 1$, the norm is simply a square. Consequently, we get by interchanging differentiation and integration (this can be justified in complete analogy to Lemma 8.1.4) that

$$\nabla J_{\mathfrak{m}}(\theta) = 2\int (\Psi_\theta(x) - f_{\mathbf{m}}(x))\nabla_\theta \Psi_\theta(x)\, \mathbb{P}_X(dx) \quad \text{and} \tag{8.27}$$

$$\nabla^2 J_{\mathfrak{m}}(\theta) = 2\mathbb{E}\big[\nabla_\theta \Psi_\theta(X)\nabla_\theta \Psi_\theta(X)^T\big] + 2\int (\Psi_\theta(x) - f_{\mathbf{m}}(x))\nabla^2 \Psi_\theta(x)\mathbb{P}_X(dx). \tag{8.28}$$

Note that if the realization $f_{\mathbf{m}}$ is very close to the response Ψ_θ for an efficient parameter $\theta \in U$, then we informally have that

$$\nabla J_{\mathfrak{m}}(\theta) \approx 0 \quad \text{and} \quad \nabla^2 J_{\mathfrak{m}}(\theta) \approx 2\mathbb{E}\big[\nabla_\theta \Psi_\theta(X)\nabla_\theta \Psi_\theta(X)^T\big] =: 2G_\theta.$$

Our proof strategy is therefore to

- show that G_θ is strictly positive definite,
- carry out a spectral analysis for $\nabla^2 J_{\mathfrak{m}}(\theta)$
- show that in the case that the target function $f_{\mathbf{m}}$ is close to Ψ_θ, there exists a local minimum in the neighborhood of an efficient parameter θ_0.

Strict positive definiteness of G_θ. Let $v \in \mathbb{R}^{\dim(\theta)}$. We need to show that

$$v^T G_\theta v \geq 0 \quad \text{and} \quad [v^T G_\theta v = 0 \;\Rightarrow\; v = 0].$$

One has

$$v^T G_\theta v = v^T \mathbb{E}\big[\nabla_\theta \Psi_\theta(X)\nabla_\theta \Psi_\theta(X)^T\big] v = \Big\|\langle v, \nabla_\theta \Psi_\theta(\cdot)\rangle\Big\|^2_{\mathbb{P}_X} \geq 0.$$

Now suppose that $v^T G_\theta v = 0$. Then

$$\langle v, \nabla_\theta \Psi_\theta(\cdot)\rangle = 0, \quad \mathbb{P}_X\text{-almost surely.}$$

The latter function is analytic (since Ψ_θ is analytic and $\mathbb{P}_X$ is compactly supported). Consequently, it is zero on the entire support $\mathcal{X}$ of $\mathbb{P}_X$.

Recall that $\#V_{\text{out}} = 1$ and let $V_{\text{out}} = \{*\}$. The derivatives of Ψ_θ are then given by

$$\begin{aligned} &x \mapsto 1 && && (\partial\beta_*) \\ &x \mapsto \psi\big(\beta_j + \langle x, w_{\bullet j}\rangle\big) && j \in V_1 && (\partial w_{j*}) \\ &x \mapsto \psi'\big(\beta_j + \langle x, w_{\bullet j}\rangle\big) w_{j*} && j \in V_1 && (\partial\beta_j) \\ &x \mapsto \psi'\big(\beta_j + \langle x, w_{\bullet j}\rangle\big) w_{j*} x_i && j \in V_1,\ i \in V_0 && (\partial w_{ij}) \end{aligned}$$

We thus get the representation

$$\langle v, \nabla_\theta \Psi_\theta(x)\rangle = v_{\beta_*} + \sum_{j\in V_1} \Phi_j(x) \qquad \forall x \in \mathcal{X}$$

with

$$\Phi_j(x) := \underbrace{v_{w_{j*}}}_{=:P_0^{(j)}} \psi\big(\beta_j + \langle x, w_{\bullet j}\rangle\big) + \underbrace{\Big(v_{\beta_j} w_{j*} + \sum_{i\in V_0} v_{w_{ij}} w_{j*} x_i\Big)}_{=:P_1^{(j)}(x)} \psi'\big(\beta_j + \langle x, w_{\bullet j}\rangle\big),$$

where we write $v_{\theta_i} := v_i$ (to easily refer to the particular parameters). Recall that since θ is $(0,0,1)$-polynomially efficient the function $\langle v, \nabla_\theta \Psi_\theta(\cdot)\rangle$ can only be zero on $\mathcal{X}$, if all polynomials ('coefficients') are zero. By assumption, every $w_{j*} \neq 0$ so that we get that indeed all entries of the vector v are zero. Thereby we proved that G_θ is strictly positive definite and that the minimal eigenvalue $\underline{\lambda}_\theta$ of θ is strictly positive.

Spectral analysis of $\nabla^2 J_{\mathfrak{m}}(\theta)$. Observe that in terms of

$$\overline{\lambda}_\theta := \sup_{\|v\|=1} \big\|v^T \nabla_\theta^2 \Psi_\theta(\cdot) v\big\|_{\mathbb{P}_X} \leq \Big(\int \|\nabla^2 \Psi_\theta(x)\|_{\mathrm{op}}\, \mathbb{P}_X(dx)\Big)^{1/2}$$

one has for $v \in \Theta$ that by the Cauchy-Schwarz inequality

$$\begin{aligned} v^T \nabla^2 J_{\mathfrak{m}}(\theta)\, v &\overset{(8.28)}{=} 2\Big(v^T G_\theta v + \int (\Psi_\theta(x) - f_{\mathfrak{m}}(x)) v^T \nabla_\theta^2 \Psi_\theta(x) v\, \mathbb{P}_X(dx)\Big) \\ &\geq 2\Big(\underline{\lambda}_\theta - \|\Psi_\theta - f_{\mathfrak{m}}\|_{\mathbb{P}_X} \overline{\lambda}_\theta\Big)\|v\|^2. \end{aligned} \tag{8.29}$$

Finding a local minimum. We will prove that there exists a local minimum in $B_\delta(\theta_0)$ for $\delta > 0$, if there exists a lower bound on the spectrum of the Hessian $\rho > 0$ such that

$$\|\nabla J_{\mathfrak{m}}(\theta_0)\| < \tfrac{\delta}{2}\rho \qquad \text{and} \qquad \big[\nabla^2 J_{\mathfrak{m}}(\theta) \succeq \rho \quad \forall \theta \in B_\delta(\theta_0)\big]. \tag{8.30}$$

Since $J_{\mathfrak{m}}$ is thereby ρ-strongly convex on $B_\delta(\theta_0)$[10] we have for all $\theta \in B_\delta(\theta_0)$[11]

$$\begin{aligned} J_{\mathfrak{m}}(\theta) &\geq J_{\mathfrak{m}}(\theta_0) + \langle \nabla J_{\mathfrak{m}}(\theta_0), \theta - \theta_0\rangle + \tfrac{\rho}{2}\|\theta - \theta_0\|^2 \\ &\geq J_{\mathfrak{m}}(\theta_0) - \|\nabla J_{\mathfrak{m}}(\theta_0)\|\|\theta - \theta_0\| + \tfrac{\rho}{2}\|\theta - \theta_0\|^2. \end{aligned}$$

[10] Nesterov, 2018, Thm. 2.1.11.

[11] Nesterov, 2018, Def. 2.1.3.

For all parameters θ on the boundary $\partial B_\delta(\theta_0)$ of the ball we therefore have

$$J_{\mathfrak{m}}(\theta) \geq J_{\mathfrak{m}}(\theta_0) - \|\nabla J_{\mathfrak{m}}(\theta_0)\|\delta + \tfrac{\rho}{2}\delta^2 \overset{(8.30)}{>} J_{\mathfrak{m}}(\theta_0).$$

The minimum, which the cost J_θ assumes on the (compact) closed ball $\overline{B_\delta(\theta_0)}$, can therefore not be on the boundary. Consequently, there must be a local minimum in $B_\delta(\theta_0)$.

Finishing the proof By reducing the size of the open set U if necessary, we can assume without loss of generality that its closure $\overline{U}$ is compact and also contained in the set of polynomially efficient parameters $\mathcal{E}_P^{(0,0,1)}$. Now suppose that θ_0 is an arbitrary efficient element of U. The continuous maps $(\underline{\lambda}_\theta)_{\theta\in\Theta}$ and $(\overline{\lambda}_\theta)_{\theta\in\Theta}$ both attain their minimum and maximum on the compact set $\overline{U}$, where all $\underline{\lambda}_\theta > 0$ and all $\overline{\lambda}_\theta < \infty$ so that

$$\underline{\lambda} = \min_{\theta\in\overline{U}} \underline{\lambda}_\theta > 0 \quad \text{and} \quad \overline{\lambda} = \max_{\theta\in\overline{U}} \overline{\lambda}_\theta < \infty.$$

To satisfy (8.30) with $\rho := \underline{\lambda}$, let $\epsilon := \underline{\lambda}/(2\overline{\lambda})$ and select δ sufficiently small such that

- $B_\delta(\theta_0) \subseteq U \subseteq \mathcal{E}_P^{(0,0,1)}$ (this ensures a minimum in U)
- $\|\Psi_\theta - \Psi_{\theta_0}\|_{\mathbb{P}_X} \leq \epsilon/2$ for all $\theta \in B_\delta(\theta_0)$ (using continuity of Ψ_θ).

Then choose $r \in (0, \epsilon/2)$ such that $\|\nabla_\theta \Psi_{\theta_0}\|_{\mathbb{P}_X} < \frac{\delta}{2r}\rho$. Consequently the inequality $\|\Psi_{\theta_0} - f_{\mathfrak{m}}\| \leq r$ implies

1. by (8.27) and Cauchy's inequality
$$\|\nabla J_{\mathfrak{m}}(\theta_0)\| \leq \|\Psi_{\theta_0} - f_{\mathfrak{m}}\|_{\mathbb{P}_X} \|\nabla \Psi_{\theta_0}\|_{\mathbb{P}_X} < \tfrac{\delta}{2}\rho,$$
2. and for all $\theta \in B_\delta(\theta_0)$
$$\|\Psi_\theta - f_{\mathfrak{m}}\|_{\mathbb{P}_X} \leq \|\Psi_\theta - \Psi_{\theta_0}\|_{\mathbb{P}_X} + \|\Psi_{\theta_0} - f_{\mathfrak{m}}\|_{\mathbb{P}_X} \leq \epsilon.$$
Using $\epsilon = \underline{\lambda}/(2\overline{\lambda})$ and (8.29) we can lower bound spectrum by ρ, i.e. for all v such that $\|v\| = 1$
$$v^T \nabla^2 J_{\mathfrak{m}}(\theta)\, v \geq 2(\underline{\lambda} - \epsilon\overline{\lambda}) = \underline{\lambda} \overset{\text{def.}}{=} \rho.$$

This means we satisfy (8.30) if $\|\Psi_{\theta_0} - f_{\mathfrak{m}}\| \leq r$. By assumption on the random statistical model $\mathbf{M}$ the latter property holds with strict positive probability. □

8.5 Existence of redundant critical points

Since the set of redundant parameters is generally a thin set with respect to the Lebesgue measure (e.g. $\mathcal{E}_0$ in Theorem 8.2.3), one may reasonably hope that this set does not contain any critical points of the MSE with probability one. In that case the MSE would be a Morse function over the entire set of parameters with probability one. Unfortunately, this hypothesis is wrong in general as the following theorem shows. We further break down the set of redundant parameters, using the taxonomy introduced in Remark 8.2.2, to make more precise statements about the existence of redundancies which are required to be of a certain type.

Theorem 8.5.1 (Redundancies cannot be ruled out in general)**.** *Assume the standard setting (Definition 8.0.6) without regularization, i.e.* $R \equiv 0$.

Assume $\mathcal{E}_P^{(0,0,1)}$ *contains an open set*[12] *and that there is at least one hidden neuron* ($\#V_1 \geq 1$)*, then, with* ***positive probability****, critical points of the MSE* $\mathbf{J}$ *do* ***exist*** *in the sets of*

(A) *redundant parameters,*

(B) *redundant parameters that only admit duplication redundancies (assuming* $\#V_1 \geq 2$*)*

(C) *redundant parameters that only admit bias* and *deactivation redundancies,*

[12] e.g. $\psi \in \{\text{sigmoid}, \tanh\}$ and an open set in the support of $\mathbb{P}_X$ by Theorem 8.2.3

Proof (outline). Clearly, the existence of redundant critical points (A) follows from the existence of critical points with more specific redundancies, i.e. (B) or (C). So we only need to prove (B) and (C). To do so, we make use of the fact that we have proven efficient critical points exist with positive probability (Theorem 8.4.1). Using $\#V_1 \geq 1$ we can therefore find a critical point of a smaller network with $\#V_1 - 1$ hidden neurons with positive probability. We then carefully **extend** this network and its parameters by a redundancy in a fashion that retains the criticality of the parameters. But for a duplication redundancy we obviously need at least two hidden neurons. Details follow in Section 8.5.1. □

Conversely, pure bias redundancies can be ruled out.

Proposition 8.5.2 (Pure bias redundancies can be ruled out)**.** *Assume the standard unregularized setting (Definition 8.0.6). If* $\psi'(x) \neq 0$ *for all* $x \in \mathbb{R}$*, the support* $\mathcal{X}$ *of* $\mathbb{P}_X$ *contains an open set and an efficient parameter is automatically* $(1, 0, 1)$*-polynomially efficient,*[12] *then, with* ***probability one****, critical points of the MSE* $\mathbf{J}$ *do* ***not exist*** *in the set of*

(D) *redundant parameters that only admit bias redundancies.*

Proof (outline). The proof of (D) relies on **pruning** the bias redundancies to obtain an efficient parameter for a smaller network. This efficient parameter must then also be a critical point of the cost and satisfy an additional condition. We then show that there are almost surely no efficient critical points which satisfy this additional condition and thereby rule out critical bias redundancies. Details follow in Section 8.5.3 after we outline a general pruning process in Section 8.5.2. □

8.5.1 Extending (Proof of Theorem 8.5.1)

Using the following lemma to extend an efficient critical point of a smaller network, (B) and (C) clearly follow from the existence of such an efficient critical point in the smaller network positive probability. This follows from Theorem 8.4.1, for which we require the existence of an open set in the set $\mathcal{E}_P^{(0,0,1)}$.

Lemma 8.5.3 (Extension)**.** *Assume the setting of Theorem 8.5.1 and without loss of generality $V_1 = \{1, \dots, \#V_1\}$. Define the reduced ANN to be $\tilde{\mathfrak{N}} = (\tilde{\mathbb{V}}, \psi)$ with neurons $\tilde{\mathbb{V}} := \big(V_0, V_1 \setminus \{1\}, V_2\big)$. Assume that the parameter $\tilde{\theta}$ of the network $\tilde{\mathfrak{N}}$ is a critical point of $J_{\mathfrak{m}}$. Then there exists a parameter θ of the network $\mathfrak{N}$ such that it is a critical point of $J_{\mathfrak{m}}$ and either*

1. *θ only has a single duplication redundancy and no other redundancies (if we further assume $\#V_1 \geq 2$), or*
2. *θ only has a deactivation and bias redundancy at the same neuron and no other redundancies.*

Remark 8.5.4. We will not make use of the fact that the loss ℓ is the squared error. We only require sufficient regularity such that derivatives may be moved into the expectation (e.g. Lemma 8.1.4).

Proof of 1. We are going to construct a parameter θ with a single duplication redundancy from the parameter $\tilde{\theta}$ of the reduced network. Assume that the parameters we do not mention are kept as is. Our plan is to duplicate the neuron 2 so we define for all $i \in V_0$

$$w_{i1} := \tilde{w}_{i2} \qquad \text{and} \qquad \beta_1 := \tilde{\beta}_2. \tag{duplication}$$

To ensure neither neuron is deactivated pick $\lambda \in \mathbb{R} \setminus \{0, 1\}$ and define

$$w_{1l} := \lambda \tilde{w}_{2l} \qquad \text{and} \qquad w_{2l} := (1-\lambda)\tilde{w}_{2l}. \tag{'convex' combination}$$

Clearly, θ is in the set of parameters which only admit duplication redundancies.

It is straightforward to see, that the response must remain the same, i.e. $\Psi_\theta = \Psi_{\tilde{\theta}}$, as we have just split one identical neuron into a 'convex' combination of two identical ones. That is

$$\begin{aligned}
&(\Psi_\theta(x))_l \\
&= \beta_l + \sum_{j=1}^{\#V_1} \psi\big(\beta_j + \langle x, w_{\bullet j}\rangle\big) w_{jl} \\
&= \beta_l + \psi\big(\beta_2 + \langle x, w_{\bullet 2}\rangle\big) \underbrace{\big((1-\lambda)w_{1l} + \lambda w_{2l}\big)}_{=\tilde{w}_{2l}} + \sum_{j=3}^{\#V_1} \psi\big(\beta_j + \langle x, w_{\bullet j}\rangle\big) w_{jl} \\
&= \tilde{\beta}_l + \sum_{j \in V_1 \setminus \{1\}} \psi\big(\tilde{\beta}_j + \langle x, \tilde{w}_{\bullet j}\rangle\big) \tilde{w}_{jl} \\
&= (\Psi_{\tilde{\theta}}(x))_l.
\end{aligned}$$

With this fact under our belt, we can now consider the derivatives of the cost. Recall that we denote by $\partial_{\hat{y}_l} \ell$ the partial derivative of the loss $\ell(\hat{y}, y)$ with respect

to the l-th component of the prediction $\hat{y}$. For $l \in V_2$, $j \in V_1$ and $i \in V_1$ we then have

$$\partial_{\beta_l} J_{\mathfrak{m}}(\theta) = \mathbb{E}_{\mathfrak{m}}\big[\partial_{\hat{y}_l}\ell(\Psi_\theta(X), Y)\underbrace{\partial_{\beta_l}(\Psi_\theta(X))_l}_{=1}\big] \overset{\Psi_\theta=\Psi_{\tilde{\theta}}}{=} \partial_{\tilde{\beta}_l} J_{\mathfrak{m}}(\tilde{\theta}) = 0,$$

$$\partial_{w_{jl}} J_{\mathfrak{m}}(\theta) = \mathbb{E}_{\mathfrak{m}}\big[\partial_{\hat{y}_l}\ell(\Psi_\theta(X), Y)\underbrace{\partial_{w_{jl}}(\Psi_\theta(X))_l}_{=\psi(\tilde{\beta}_j + \langle \tilde{w}_{\bullet j}, X\rangle)}\big] \overset{\Psi_\theta=\Psi_{\tilde{\theta}}}{=} \partial_{\tilde{w}_{jl}}\mathbf{J}(\tilde{\theta}) = 0,$$

$$\begin{aligned}
\partial_{\beta_j} J_{\mathfrak{m}}(\theta) &= \mathbb{E}_{\mathfrak{m}}\Big[\sum_{l\in V_2} \partial_{\hat{y}_l}\ell(\Psi_\theta(X), Y)\ \underbrace{w_{jl}}\ \psi'(\beta_j + \langle X, w_{\bullet j}\rangle)\Big] \\
&\qquad = \begin{cases} (1-\lambda\delta_{j2})\tilde{w}_{jl} & j \neq 1 \\ \lambda \tilde{w}_{2l} & j = 1 \end{cases} \\
&\overset{\Psi_\theta=\Psi_{\tilde{\theta}}}{=} \begin{cases} (1-\lambda\delta_{j2})\partial_{\tilde{\beta}_j} J_{\mathfrak{m}}(\tilde{\theta}) & j \neq 1 \\ \lambda \partial_{\tilde{\beta}_2} J_{\mathfrak{m}}(\tilde{\theta}) & j = 1 \end{cases} \\
&= 0,
\end{aligned}$$

$$\begin{aligned}
\partial_{w_{ij}} J_{\mathfrak{m}}(\theta) &= \mathbb{E}\Big[\sum_{l\in V_2} \partial_{\hat{y}_l}\ell(\Psi_\theta(X), Y)\psi'(\beta_j + \langle X, w_{\bullet j}\rangle) w_{jl} X_i\Big] \\
&\overset{\Psi_\theta=\Psi_{\tilde{\theta}}}{=} \begin{cases} (1-\lambda\delta_{j2})\partial_{\tilde{w}_{ij}} J_{\mathfrak{m}}(\theta) & j \neq 1 \\ \lambda \partial_{\tilde{w}_{i2}} J_{\mathfrak{m}}(\theta) & j = 1 \end{cases} \\
&= 0.
\end{aligned}$$

the parameter θ is thereby clearly a critical point with no other redundancies except for a single duplication.

Proof of 2. To construct a parameter θ with a deactivation and bias redundancy from the reduced network, define

$$w_{\bullet 1} = 0, \qquad w_{1\bullet} = 0,$$

select $\beta_1 \in \mathbb{R}$ arbitrarily and retain all parameters of $\tilde{\theta}$ for the other neurons.

Since the additional neuron is deactivated, the response remains the same, i.e. $\Psi_\theta = \Psi_{\tilde{\theta}}$. And since $\tilde{\theta}$ is a critical point, it is straightforward to check that the derivatives with respect to the old parameters remain the same and are thereby zero. For the derivatives with respect to the new parameters let us consider the outer derivatives first

$$\begin{aligned}
\partial_{w_{1l}} J_{\mathfrak{m}}(\theta) &= \mathbb{E}_{\mathfrak{m}}\big[\partial_{\hat{y}_l}\ell(\Psi_\theta(X), Y)\psi\big(\beta_1 + \langle X, w_{\bullet 1}\rangle\big)\big] \\
&\overset{w_{\bullet 1}=0}{=} \underbrace{\mathbb{E}_{\mathfrak{m}}\big[\partial_{\hat{y}_l}\ell(\Psi_\theta(X), Y)\big]}_{=\partial_{\tilde{\beta}_l} J_{\mathfrak{m}}(\tilde{\theta})=0}\psi(\beta_1).
\end{aligned}$$

In the last equation we used that the response remains the same. The derivatives with respect to the inner derivatives on the other hand are all zero due to the deactivation with the outer parameter $w_{1\bullet} = 0$.

8.5.2 Pruning

The following result shows that for any redundant parameter there exists an efficient parameter of a smaller network with equal response function on the support $\mathcal{X}$ of the input X. We also show that criticality is retained under the standard assumption of $\#V_{\text{out}} = 1$ (cf. Definition 8.0.6).

Proposition 8.5.5. *Let $\mathfrak{N} = (\mathbb{V}, \psi)$ with $\mathbb{V} = (V_0, V_1, V_2)$ be a shallow ANN as in Definition 8.0.1. Assume that the parameter θ of this network is redundant. Then there exists a pruned network $\tilde{\mathfrak{N}} = ((V_0, \tilde{V}_1, V_2), \psi)$ with $\tilde{V}_1 \subseteq V_1$ and an **efficient** parameter $\tilde{\theta}$ of this pruned network such that the response remains the same, i.e.*

$$\Psi_\theta(x) = \Psi_{\tilde{\theta}}(x) \qquad \forall x \in \mathcal{X}.$$

Furthermore, if $\#V_{out} = 1$ *and* $\tilde{\theta}$ *was a critical point of some cost function*

$$J_{\mathfrak{m}}(\theta) = \mathbb{E}_{\mathfrak{m}}[\ell(\Psi_\theta(X), Y)]$$

for some loss function ℓ *and expectation* $\mathbb{E}_{\mathfrak{m}}$ *induced by some distribution* $\mathbb{P}_{\mathfrak{m}}$*, then (assuming suitable regularity on* ℓ *and* ψ *such that derivatives may be moved into the expectation* $\mathbb{E}_{\mathfrak{m}}$*, cf. Lemma 8.1.4) the pruned parameter* $\tilde{\theta}$ *is also a critical point of* $J_{\mathfrak{m}}$*.*

Proof. A parameter is redundant if either of the two criterions (a) or (b) in the Definition of Efficiency 8.2.1 are violated. Recall, that we called the violation of criterion (a) a deactivation redundancy (Remark 8.2.2).

Deactivation pruning It is straightforward to see that the response of an ANN does not change if all the deactivated hidden neurons, i.e. all $j \in V_1$ where $w_{j\bullet} = 0$, are removed from V_1. Similarly, it is straightforward to show that critical parameters remain critical since the previous gradient contains all the partial derivatives with respect to the remaining parameters in the pruned network.

Pruning the other redundancies Until the parameter is efficient we will iteratively remove a single neuron, while ensuring that the response stays the same and critical points remain critical. Since there only a finite number of neurons, this procedure will eventually terminate – if there are no hidden neurons left, then there is only a bias on the output which is clearly efficient (Definition 8.2.1). We therefore only describe the procedure of a single step.

Note that if a pruning step reintroduces deactivation redundancies, we interject a deactivation pruning step. We can therefore always assume there are no deactivation redundancies at the beginning of a pruning step.

If the parameter θ is redundant without deactivation redundancies (a), then there must be a hidden neuron $k \in V_1$ that can be linearly combined from the others (cf. Remark 8.2.2), i.e.

$$\psi\Big(\beta_k + \sum_{i \in V_0} x_i w_{ik}\Big) = \lambda_\emptyset + \sum_{j \in V_1 \setminus \{k\}} \lambda_j \psi\Big(\beta_j + \sum_{i \in V_0} x_i w_{ij}\Big) \quad \forall x \in \mathcal{X}.$$

We define a parameter $\tilde{\theta}$ for the pruned network $\tilde{\mathfrak{N}} = ((V_0, V_1 \setminus \{k\}, V_2), \psi)$ using the parameter θ from the old network. Specifically, we retain all inner parameters and define the outer parameters to be

$$\tilde{w}_{jl} := w_{jl} + w_{kl}\lambda_j \qquad \tilde{\beta}_l := \beta_l + w_{kl}\lambda_\emptyset \qquad \forall l \in V_2,\ j \in V_1 \setminus \{k\}.$$

Using this definition it is straightforward to check that the response remains the same, i.e. $\Psi_\theta = \Psi_{\tilde{\theta}}$.

In the case $\#V_{\text{out}} = 1$, i.e. $V_{\text{out}} = \{*\}$, we need to show that this pruning step retains criticality. Recall that the derivatives are given by

$$\partial_{\theta_i} J_{\mathfrak{m}}(\theta) = \mathbb{E}_{\mathfrak{m}}\big[\partial_{\hat{y}}(\Psi_\theta(X), Y)\partial_{\theta_i}\Psi_\theta(X)\big]. \tag{8.31}$$

Since the derivatives of the response $\partial_{\tilde{\beta}_*}\Psi_{\tilde{\theta}}$ and $\partial_{\tilde{w}_{j*}}\Psi_{\tilde{\theta}}$ with respect to the outer parameters only contain inner parameters (which we have not changed) and the response remains the same $\Psi_{\tilde{\theta}} = \Psi_\theta$, we immediately get the criticality of the partial derivatives with respect to the outer parameters. What is left to consider are the derivatives with respect to the inner derivatives. Since θ had no deactivation redundancies, we have $w_{j*} \neq 0$ for all $j \in V_1$. In particular we have for all $j \in V_1 \setminus \{k\}$

$$\partial_{\tilde{\beta}_j}\Psi_{\tilde{\theta}}(x) = \psi'(\beta_j + \langle x, w_{\bullet j}\rangle)\tilde{w}_{j*} = \partial_{\beta_j}\Psi_\theta(x)\frac{\tilde{w}_{j*}}{w_{j*}}$$

$$\partial_{\tilde{w}_{ij}}\Psi_{\tilde{\theta}}(x) = \psi'(\beta_j + \langle x, w_{\bullet j}\rangle)x_i\tilde{w}_{j*} = \partial_{w_{ij}}\Psi_\theta(x)\frac{\tilde{w}_{j*}}{w_{j*}}$$

The derivatives of the response therefore only change up to a constant that can be moved out of the expectation in (8.31). Together with $\Psi_{\tilde{\theta}} = \Psi_\theta$ this yields criticality. □

8.5.3 Bias redundancies (Proof of Proposition 8.5.2)

Recall that a bias redundancy as defined in Remark 8.2.2 implies that $\psi_k(x) := \psi(\beta_k + \langle x, w_{\bullet k}\rangle)$ is constant on $\mathcal{X}$.

Lemma 8.5.6 (Bias redundancy characterization)**.** *Let the activation function ψ be injective and assume*

$$\{x - y : x, y \in \mathcal{X}\}^\perp = \{0\}. \tag{8.32}$$

Then a bias redundancy at neuron k implies $w_{\bullet k} = 0$.

Observe that (8.32) is satisfied as soon as $\mathcal{X}$ contains an open set.

Proof. Let there be a bias redundancy at neuron k. If there were $x, y \in \mathcal{X}$ such that $\langle x - y, w_{\bullet k}\rangle \neq 0$, then $\psi_k(x) \neq \psi_k(y)$ due to injectivity. Consequently $\langle x - y, w_{\bullet k}\rangle = 0$ for all $x, y \in \mathcal{X}$ and (8.32) thereby implies $w_{\bullet k} = 0$. □

Recall that we assumed in Proposition 8.5.2 $\#V_{\text{out}} = 1$, $\psi'(x) \neq 0$ for all $x \in \mathbb{R}$, which also implies ψ is injective as it is strictly monotonous, and an open set in $\mathcal{X}$ such that Lemma 8.5.6 is satisfied.

In Section 8.5.2 we discussed a general pruning procedure that proceeds in steps, removing one neuron at a time. Consequently, this procedure is path dependent. If a different neuron were removed first one might end up with a different pruned network and parameter. In the case where we only have bias redundancies we can do better. Assume the requirements of Lemma 8.5.6 are satisfied, let $I \subseteq V_1$ be the maximal set such that $w_{\bullet j} = 0$ for all $j \in I$. We then define the pruned network $\tilde{\mathfrak{N}} := ((V_0, V_1 \setminus I, V_2), \psi)$ in a single step: The parameter $\tilde{\theta}$ retains all the weights and biases from θ restricted to the pruned ANN-graph except for

$$\tilde{\beta}_l := \beta_l + \sum_{j \in I} w_{jl}\psi(\beta_j) \qquad l \in V_2.$$

It is straightforward to show that the response then remains the same. In the following lemma we will relate the criticality of $\tilde{\theta}$ to that of θ.

Lemma 8.5.7 (Characterization of critical bias redundancies)**.** *Assume the setting of Proposition 8.5.2. If a critical point θ of the cost $J_{\mathfrak{m}}$ only has bias redundancies, then the parameter $\tilde{\theta}$ of the pruned network is also a critical point of $J_{\mathfrak{m}}$ and the following equation is satisfied*

$$\mathbb{E}_{\mathfrak{m}}\big[\partial_{\hat{y}}\ell(\Psi_{\tilde{\theta}}(X), Y)X_i\big] = 0 \qquad \forall i \in V_0. \tag{8.33}$$

This is furthermore sufficient, i.e. if $\tilde{\theta}$ is critical and (8.33) is satisfied then the original parameter θ is critical.

Remark 8.5.8. We do not make use of the squared error loss function and only require sufficient regularity that derivatives may be moved into the expectation.

Proof. "⇒": That $\nabla J_{\mathfrak{m}}(\theta) = 0$ implies $\nabla J_{\mathfrak{m}}(\tilde{\theta}) = 0$ is a straightforward exercise since the response remains the same and we retain almost all parameters except for the outer bias which does not occur in any of the partial derivatives of the response. Since we assume $V_1 = \{*\}$ in this section we furthermore have

$$0 = \partial_{w_{ij}} J_{\mathfrak{m}}(\theta) = w_{j*}\mathbb{E}_{\mathfrak{m}}\big[\partial_{\hat{y}}\ell(\Psi_\theta(X), Y)X_i\big]\psi'(\beta_j). \tag{8.34}$$

And since we assume $\psi'(x) \neq 0$ for all $x \in \mathbb{R}$ in this section and $w_{j*} \neq 0$ since we ruled out deactivation redundancies, we obtain (8.33).

"⇐": Using (8.33) and $\nabla J_{\mathfrak{m}}(\tilde{\theta}) = 0$ we now have to prove $\nabla J_{\mathfrak{m}}(\theta) = 0$. The directional derivatives of the parameters that remained on the pruned ANN-graph are zero because they coincide with those of $\tilde{\theta}$. What is left are therefore the directional derivatives of the parameters attached to the nodes $j \in I$. Using (8.34) in reverse with (8.33) we obtain that $\partial_{w_{ij}} J_{\mathfrak{m}}(\theta) = 0$ What is left are therefore the biases β_j with $j \in I$. Those are given by

$$\partial_{\beta_j} J_{\mathfrak{m}}(\theta) = \mathbb{E}\big[\partial_{\hat{y}}\ell(\Psi_\theta(X), Y)\psi'\big(\beta_j\big)\big]w_{j*} = \underbrace{\partial_{\tilde{\beta}_*} J_{\mathfrak{m}}(\tilde{\theta})}_{=0}\psi'\big(\beta_j\big)w_{j*}. \qquad \square$$

The pruned condition a.s. never happens (Proof of (D))

With the characterization of critical bias redundancies (Lemma 8.5.7), proving the non-existence of bias redundancies is equivalent to proving that a parameter vector of an efficient network can never be a critical point which also satisfies (8.33). To prove (D) we therefore simply have to show that (8.33) almost surely never coincides with an efficient critical point.

Since the efficient parameters are automatically $(1,0,1)$-polynomially efficient by assumption, we need to show that the set $\mathcal{E}_P^{(1,0,1)}$ almost surely does not contain critical points that satisfy (8.33). Recall that we assume the squared error $\ell(\hat{y},y) = (\hat{y}-y)^2$ in (D) and therefore (8.33) reduces to

$$0 = \mathbb{E}_{\mathfrak{m}}\big[(\Psi_\theta(X) - Y)X_i\big] \overset{\text{tower}}{=} \mathbb{E}_{\mathfrak{m}}\big[(\Psi_\theta(X) - f_{\mathfrak{m}}(X))X_i\big] \qquad \forall i \in V_0.$$

For the random cost $\mathbf{J} = J_{\mathbf{M}}$ this means

$$0 = \int (\Psi_\theta(x) - \mathbf{f}(x))x_i \mathbb{P}_X(dx) = \langle \Psi, \pi_i\rangle_{\mathbb{P}_X} - \langle \mathbf{f}, \pi_i\rangle_{\mathbb{P}_X}$$

with random target $\mathbf{f} = f_{\mathbf{M}}$ and projection $\pi_i : x \mapsto x_i$. This combination can be captured by the level set $\mathbf{g}^{-1}(0)$ of

$$\mathbf{g} : \begin{cases} \mathcal{E}_P^{(1,0,1)} \to \mathbb{R}^{\dim(\theta)} \times \mathbb{R}^{V_{\text{in}}} \\ \theta \mapsto \Big(\nabla \mathbf{J}(\theta), \big(\langle \Psi, \pi_i\rangle_{\mathbb{P}_X} - \langle \mathbf{f}, \pi_i\rangle_{\mathbb{P}_X}\big)_{i \in V_{\text{in}}}\Big). \end{cases}$$

Since $\mathcal{E}_P^{(1,0,1)} \subseteq \mathbb{R}^{\dim(\theta)}$, $\mathbf{g}$ is a mapping into a larger dimension. Its level sets are therefore empty with probability one by Lemma 8.1.6, assuming we can prove it to be non-degenerate for every $\theta \in \mathcal{E}_P^{(1,0,1)}$. This will therefore be the finial step of the proof. Since the variance of $\mathbf{g}$ is independent of the mean, we may consider $\hat{\mathbf{J}}(\theta) = \langle \Psi_\theta, \mathbf{f}\rangle_{\mathbb{P}_X}$ as introduced in Proposition 8.1.3 instead and similarly prune $\langle \Psi_\theta, \pi_i\rangle_{\mathbb{P}_X}$, i.e. we may consider

$$\hat{\mathbf{g}}(\theta) = \Big(\nabla_\theta \langle \Psi_\theta, \mathbf{f}\rangle_{\mathbb{P}_X}, \big(\langle \mathbf{f}, \pi_i\rangle_{\mathbb{P}_X}\big)_{i \in V_{\text{in}}}\Big) = \Big(\langle \nabla_\theta \Psi_\theta, \mathbf{f}\rangle_{\mathbb{P}_X}, \big(\langle \pi_i, \mathbf{f}\rangle_{\mathbb{P}_X}\big)_{i \in V_{\text{in}}}\Big)$$

Similar to our argument in the proof of Proposition 8.1.5 it is therefore sufficient to prove the linear independence of the following functions

$$\begin{aligned}
&x \mapsto 1 && && (\partial \beta_*)\\
&x \mapsto \psi\big(\beta_j + \langle x, w_{\bullet j}\rangle\big) && j \in V_1 && (\partial w_{j\bullet})\\
&x \mapsto \psi'\big(\beta_j + \langle x, w_{\bullet j}\rangle\big) w_{j\bullet} x_i && j \in V_1, i \in V_0 && (\partial w_{ij})\\
&x \mapsto \psi'\big(\beta_j + \langle x, w_{\bullet j}\rangle\big) w_{j\bullet} && j \in V_1 && (\partial \beta_j)\\
&x \mapsto x_i && i \in V_0, && (\pi_i)
\end{aligned}$$

where the last equation is the extra condition that follows from (8.33). But their linear independence follows from the $(1,0,1)$-polynomial independence (Definition 8.1.1).

Note, that we did not require second order derivatives, but first order polynomials in the affine term due to the extra condition (8.33). This explains why we required $(1,0,1)$-polynomial independence instead of $(0,0,1,2)$-polynomial independence as in Proposition 8.1.5.

References

Adler, Robert J. and Jonathan E. Taylor (2007). *Random Fields and Geometry*. Springer Monographs in Mathematics. New York, NY: Springer New York. ISBN: 978-0-387-48112-8. DOI: 10.1007/978-0-387-48116-6.

Bogachev, Vladimir I. (1998). *Gaussian Measures*. Vol. 62. Mathematical Surveys and Monographs. Providence, RI: American Mathematical Society. 433 pp. ISBN: 0-8218-1054-5.

Carmeli, C. et al. (2010). "Vector Valued Reproducing Kernel Hilbert Spaces and Universality". In: *Analysis and Applications* 08.01, pp. 19–61. ISSN: 0219-5305. DOI: 10.1142/S0219530510001503.

Dereich, Steffen and Sebastian Kassing (2024). "Convergence of Stochastic Gradient Descent Schemes for Łojasiewicz-Landscapes". In: *Journal of Machine Learning* 3.3, pp. 245–281. ISSN: 2790-203X, 2790-2048. DOI: 10.4208/jml.240109.

Micchelli, Charles A., Yuesheng Xu, and Haizhang Zhang (2006). "Universal Kernels". In: *Journal of Machine Learning Research* 7.12. URL: http://www.jmlr.org/papers/volume7/micchelli06a/micchelli06a.pdf (visited on 2025-03-21).

Mityagin, B. S. (2020). "The Zero Set of a Real Analytic Function". In: *Mathematical Notes* 107.3, pp. 529–530. ISSN: 1573-8876. DOI: 10.1134/S0001434620030189.

Nesterov, Yurii Evgen'evič (2018). *Lectures on Convex Optimization*. Second edition. Springer Optimization and Its Applications; Volume 137. Cham: Springer. ISBN: 978-3-319-91578-4. DOI: 10.1007/978-3-319-91578-4.

Riordan, John (2002). *Introduction to Combinatorial Analysis*. dover edition. Mineola, N.Y: Dover Publications Inc. 256 pp. ISBN: 978-0-486-42536-8.

Stanley, Richard P. (2011). *Enumerative Combinatorics: Volume 1*. 2nd edition. Cambridge, NY: Cambridge University Press. 642 pp. ISBN: 978-1-107-01542-5.

Szwarc, Ryszard (2022). *How to Prove That Generalized Vandermonde Matrix Is Invertible?* Mathematics Stack Exchange. eprint: https://math.stackexchange.com/q/4549001. URL: https://math.stackexchange.com/a/4549001 (visited on 2024-11-10).

Talagrand, Michel (1987). "Regularity of Gaussian Processes". In: *Acta Mathematica* 159 (none), pp. 99–149. ISSN: 0001-5962, 1871-2509. DOI: 10.1007/BF02392556.