Dynamic Approaches for Stochastic Gradient Methods in Reinforcement Learning

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

vorgelegt von

Sara Klein

aus Saarlouis

Mannheim, 2024

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: 20. Dezember 2024

Abstract

This work addresses the convergence behaviour of first-order optimization methods in the context of reinforcement learning. Specifically, we analyse the vanilla Policy Gradient (PG) method under softmax parametrization. Initially, we focus on Markov Decision Processes (MDPs) with finite-time horizons, demonstrating that the convergence rate of vanilla PG exhibits an unfavorable and imprecisely determinable dependence on the time horizon. To resolve this issue, we introduce a combination of dynamic programming and policy gradient called finite-time dynamic policy gradient. The use of the dynamic approach much better exploits the structure of the Markovian problem which is reflected in an improved, explicit dependence of the convergence rate on all relevant model parameters.
In the second part of this thesis, we extend this concept to discounted MDPs with an infinite-time horizon, where the convergence rate in vanilla PG cannot be explicitly determined with respect to the effective horizon. For the transferred dynamic PG method, we once again establish improved and explicit convergence guarantees.
In the third part of this thesis, we analyze the convergence of stochastic gradient methods independently of reinforcement learning. Under the assumption of (weak) gradient domination, we derive almost sure convergence rates in both the global and local settings. The new results find applications in the optimization of analytical neural networks, as well as in the previously discussed classical and dynamical PG methods under softmax parametrization.

Zusammenfassung

Diese Arbeit beschäftigt sich mit dem Konvergenzverhalten von Gradientenverfahren erster Ordnung im Kontext von Reinforcement Learning. Dazu wird das klassische Policy-Gradient (PG) Verfahren unter Softmax-Parametrisierung analysiert. Zunächst betrachten wir Markov-Entscheidungsprobleme (MDPs) mit endlichem Zeithorizont und zeigen, dass die Konvergenzrate des PG-Verfahrens eine ungünstige, nicht genau zu bestimmende Abhängigkeit vom Zeithorizont aufweist. Zur Lösung wird Dynamic Programming in das bestehende Verfahren integriert. Der dynamische Ansatz nutz die Markovsche Struktur des MDPs aus und liefert eine verbesserte, explizite Abhängigkeit der Konvergenzrate von allen beteiligten Modellparametern.
Im zweiten Teil der Arbeit wird dieses Konzept auf MDPs mit unendlichem Zeithorizont übertragen. Statt des deterministischen Zeithorizonts spielt der erwartete Zeithorizont des unendlichen Problems, abhängig vom Diskontierungsfaktor, eine analoge Rolle. Für das dynamische PG-Verfahren werden erneut bessere und explizite Konvergenzgarantien bewiesen.
Im dritten Teil der Arbeit analysieren wir die Konvergenz stochastischer Gradientenmethoden losgelöst vom Reinforcement Learning. Unter der Annahme (schwach) dominierter Gradienten werden fast sichere Konvergenzraten im globalen und lokalen Setting hergeleitet. Die neuen Resultate finden Anwendung in der Optimierung analytischer neuronaler Netze sowie in den zuvor diskutierten klassischen und dynamischen PG-Verfahren unter Softmax-Parametrisierung.

Danksagung

An dieser Stelle möchte ich all denjenigen danken, die mich während meiner Promotionszeit unterstützt und gefördert haben.
Mein größter Dank gilt meinem Doktorvater Prof. Dr. Leif Döring, der mir nicht nur in den letzten dreieinhalb Jahren, sondern bereits während des Studiums ein herausragender Mentor war. Ich hege große Bewunderung für seinen Mut, die bekannte Welt der Stochastischen Prozesse ein Stück weit zu verlassen und auf ein neues Forschungsgebiet umzusatteln. Obwohl Reinforcement Learning zu Beginn der Promotion in gewisser Weise „Neuland“ war, haben die vielen Diskussionen und seine Fähigkeit, die richtigen Fragen zu stellen, schnell Licht ins Dunkel gebracht. Ich bin sehr dankbar, dass ich ein Teil dieser Transformation sein durfte, für das entgegengebrachte Vertrauen und dafür, dass du mich ermutigt hast, diese Herausforderung anzunehmen!
Ein weiteres großes Dankeschön geht an Prof. Dr. Simon Weißmann, der seine Expertise in (stochastischer) Optimierung mit mir geteilt, und so zum Erfolg aller drei in dieser Dissertation enthaltenen Projekte beigetragen hat. Ohne sein Zutun würden einige Teile dieser Arbeit nicht in dieser Form existieren. Danke auch für deine emotionale Stütze, insbesondere in Deadline Phasen.
Ebenso herzlichen Dank an Prof. Dr. Steffen Dereich, der sich bereit erklärt hat, diese Dissertation als Korreferent zu begutachten.
Als nächstes möchte ich mich bei meinen Freundinnen und Mitdoktorandinnen Vicky und Khadija bedanken. Gemeinsam sind wir durch so einige emotionale Hochs und Tiefs des Promotionslebens gegangen und haben uns immer wieder gegenseitig aufgebaut! Mit Vicky habe ich ein Online-Office gegründet in dem wir vermutlich tausende von Stunden verbracht haben – Gott sei Dank zahlt die Uni einen unbegrenzten Zoom Zugang – und tatsächlich ist es uns oft gelungen auch produktiv zu arbeiten. Mit Khadija hingegen habe ich die meisten realen Kaffeepausen an der Uni geteilt – gut, dass es bei dir im Büro auch noch den besten Kaffee der Uni gibt. Danke für diese tolle Zeit, auf die ich dank euch beiden immer mit einem Lächeln zurückblicken werde.
Ein weiteres Danke geht an Svenja, Benedikt, André und Lukas für das harmonische und offene Miteinander im Büro. Besonderer Dank gebührt Felix, mit dem ich über die längste Zeit ein Büro teilen durfte und der immer für eine Mathe-Diskussion zu haben war. An dieser Stelle gehen auch liebe Grüße an den 4. Stock und alle anderen, die regelmäßig an den gemeinsamen Mensagängen beteiligt waren. Auch wenn das Essen nicht immer lecker war, so war es doch immer lustig mit euch! Danke auch an alle Mitglieder des Graduiertenkollegs *RTG 1593 „Statistical Modeling of Complex Systems“* für die schöne Zeit bei den gemeinsamen Workshops und Vorträgen. Ebenso ein Dankeschön an die Hanns-Seidel-Stiftung e.V., die mich sowohl finanziell als auch ideell in den letzten drei Jahren unterstützt hat.
Zum Schluss möchte ich versuchen meine unendliche Dankbarkeit gegenüber meiner Familie zum Ausdruck zu bringen. Wir kommen also zu den Menschen, ohne deren Unterstützung ich wohl nie so weit gekommen wäre. Mein Dank gegenüber meinem Verlobten Raphael ist wohl kaum in Worte zu fassen. Du musstest meine Nervenzusammenbrüche schon während des gesamten Studiums ertragen und hast mich auch in der Promotion weiterhin bedingungslos unterstützt. Danke, dass du mein Ruhepol bist und mich immer wieder daran erinnerst, was

die wirklich wichtigen Dinge im Leben sind. Mein großer Bruder David ist mir in vielen Dingen noch heute ein Vorbild, und auch wenn wir uns aufgrund räumlicher Distanz nicht so oft sehen, weiß ich, dass er immer für mich da ist. Mein Papa Elmar, meine Mama Petra und ihr Mann Christoph haben mich zu jedem Zeitpunkt meines Lebens selbstlos und großzügig unterstützt. Ich danke euch vor allem für die Werte, die ihr mir beigebracht und vorgelebt habt. Ihr habt mich geprägt und zu dem Menschen gemacht, der ich heute bin.
Ich weiß, dass ich immer auf meine Familie zählen kann, und ihnen ist diese Arbeit gewidmet.

Contents

Introduction

1

Reinforcement learning is applied to so-called Markov decision processes (MDPs), which mathematically formalize the interaction between a learning agent and its environment, with the objective of maximizing observable rewards. Algorithms that train an agent (or a policy) for an MDP are collectively referred to under the broader term of reinforcement learning (RL). Although MDPs have been known since the 1950s, computer scientists only achieved a breakthrough in the application of RL to video games in the past decade, garnering worldwide attention. Motivated by this success, researchers across various fields began applying RL to real-world problems. Virtually any control problem can be framed as an MDP and addressed using RL.

In this work, we distinguish between two classes of MDPs. The first class includes discounted infinite-time horizon MDPs, where decisions in the distant future become progressively less significant due to discounting, as they are further removed from the present. The second class includes finite-time horizon MDPs, which have a fixed, deterministic endpoint, where discounting may be applied but is not necessarily required. Typical infinite-time horizon problems are commonly found in robotics, video games, or scenarios where the endpoint is random and uncertain. Conversely, typical finite-time horizon MDPs include supply chain problems or optimal stopping problems in finance.

The ongoing AI hype and the increasing integration of RL into everyday applications have amplified concerns regarding the safety and interpretability of these algorithms' behavior. Still, the theoretical guarantees are limited. In this dissertation, we focus on a specific class of RL algorithms, namely the policy gradient (PG) algorithm. The goal is to analyze the convergence and convergence rate of this method to develop a deeper understanding of the causes of potential misbehavior. Our primary interest lies in the convergence towards the global optimum, ensuring that no sub-optimal decisions are made. When analysing the classical PG methods, we obtain limitations in the speed of convergence with respect to some model parameters and develop dynamic approaches for PG in finite-time and infinite-time MDPs to overcome these dependencies.

Naturally, RL algorithms are divided into value-based or policy-based approaches. In value-based methods, the optimal state-action function is learned, from which the optimal policy can be derived. In policy-based methods, however, the goal is to directly search for the optimal policy. PG belongs to the class of policy-based methods, and it trains a parameterized policy by maximizing the value function using first-order optimization methods such as (stochastic) gradient descent (GD/SGD).

Since the optimization problem of PG is non-convex, we quickly encounter limitations regarding convergence. In order to perform a complete theoretical analysis, certain assumptions are necessary. In this thesis, we will primarily assume that all policies are softmax-parameterized, an assumption that has become increasingly common in recent years. This serves as a preliminary step towards a full understanding of PG methods. In infinite-time horizon problems, vanilla softmax PG has already been analyzed by Mei et al. [Mei+20]. For the softmax-parameterized value function, a gradient domination property along the gradient ascent trajectory was demonstrated, which guarantees convergence to the global optimum under the exact gradient assumption, and a convergence rate for deterministic softmax PG was derived.

However, in the finite-time setting, there have been no results on the convergence or convergence rate of PG, even under the assumption of exact gradients and softmax parametrization. In parts, this may be due to the fact that the finite-time case is more complex, in the sense that a non-stationary optimal policy is required. Hence, the first research question we address in this thesis is:

RQ1: *Does softmax PG converge for finite-time MDPs, and if so, how quickly does it converge?*

In the first project of this work, we initially demonstrate that the finite-time value function under softmax parametrization also satisfies a gradient domination property along the gradient ascent scheme, through which we derive a convergence rate under the exact gradient assumption. While the proof structure and ideas are inspired by the result for infinite-time horizon PG, they cannot be directly extended from the infinite-time case due to the non-stationary policy required in the finite-time setting. The classical finite-time PG variant simultaneously trains the policies for all time epochs at once, which is why we refer to this algorithm as finite-time simultaneous policy gradient (FT-SimPG).
We observe that, although convergence can be guaranteed, an unknown model-dependent constant appears in the convergence rate. An analogous constant was also found by [Mei+20] in the infinite-time horizon case. This constant can become arbitrarily small and may depend on factors such as the time horizon or the number of states in the MDP. This prevents us from providing a satisfactory answer to the question of how quickly the algorithm converges. This leads us to the second research question:

RQ2: *Can we guarantee an explicit convergence rate for finite-time softmax PG?*

We provide an answer to this question by introducing finite-time dynamic policy gradient (FT-DynPG). The idea for this algorithm emerged as follows. PG is a gradient-based method aimed at maximizing the value function, but it overlooks the inherent structure of an MDP which can be leveraged to solve the problem more efficiently. Under perfect information, finite-time MDPs are typically solved using backward induction. The optimal policy for the last time epoch is determined first and then decisions are made progressively backward to the first epoch where the future optimal actions are already known. This approach is referred to as finite-time dynamic programming (DP). DP relies on the Markov property, a fundamental characteristic of MDPs and in interpretive terms it means that today's best action is independent of past decisions. By combining DP with finite-time PG we introduce the FT-DynPG algorithm, a dynamic approach to PG for solving finite-time MDPs. The non-stationary policy is parameterized such that each time epoch has its own parameterized policy. These policies are then solved by policy gradient backward in time, following the DP principle. Since we explicitly exploit the structure of the problem, we are able to derive a convergence rate with all constants explicitly specified.
We observe that the constants, such as those related to the time horizon, are significantly more favorable in FT-DynPG compared to FT-SimPG, even when ignoring the unknown constant in FT-SimPG. It is important to note that, while the actual rate in terms of gradient steps is $O(1/n)$ for both algorithms, constants play a crucial role in the slow convergence rate.
Thus far, in our discussion of convergence, we have assumed that the gradient method can be executed exactly—that is, we have assumed access to the true gradients of the value function with respect to the policy parameters. However, in practice, this is not feasible, as the MDP is

typically a black box from which we can only generate samples. Given a state at a particular time, an action is executed, and we observe a reward for that state-action pair. Using these samples and the so-called Policy Gradient Theorem, the gradients of the value function can be easily estimated. This makes the algorithm particularly appealing in practical applications. When sampled (i.e. stochastic) gradients are used in PG, we refer to these algorithms as stochastic policy gradient (SPG). This leads us to the third research question:

RQ3: *Can we guarantee convergence for stochastic FT-SimPG and stochastic FT-DynPG under softmax parametrization, and how fast is the convergence?*

The answer to the first part of the question is yes. In both cases, convergence to the global optimum can be guaranteed with high probability. A complexity analysis is derived, where we obtain again that FT-DynPG has a favorable dependence on the problem parameters compared to FT-SimPG. However, our proof technique (in both approaches) requires a very large batch size during gradient sampling to ensure that the stochastic gradient trajectories remain close to the deterministic ones. We want this to hold as the gradient domination property is only fulfilled along the exact gradient path.
Since the required batch sizes in both cases depend unfavorably on parameters such as the time horizon of the problem, we do not expect this result to yield practical insights. Nevertheless, we demonstrate for the first time that an SPG Algorithm, without additional regularization, can converge to the global optimum. In the final project of this thesis, we revisit the convergence properties of SPG methods and obtain the convergence without large batch sizes is still possible. For now, we conclude the section on convergence of PG in finite-time MDPs and move on to discounted infinite-time horizon MDPs.

In the second project of this work we consider discounted infinite-time horizon MDPs. As previously mentioned, Mei et al. [Mei+20] were the first to establish a convergence rate for vanilla PG under softmax parametrization. We have also noted that, similar to FT-SimPG, an unknown model-dependent constant appears in their analysis. By integrating dynamic programming into finite-time PG, we effectively eliminated the unknown constant. On a higher level, we have gained a new perspective on RL algorithms:
Rather than adhering to the traditional separation between value-based and policy-based methods, we distinguish between algorithms that leverage the model's structure and those that do not.
The first class exploits the dynamic programming principle, as seen in value iteration, policy iteration, or Q-learning. These algorithms essentially optimize single-step problems, using a single reward feedback in the update process and future payoffs are estimated through evaluation. The second class approaches the entire multi-step problem at once and solves it using classical optimization techniques, as seen in traditional PG methods.
Over time, hybrid approaches that integrate both methodologies, such as Actor-Critic (AC) methods, have consistently demonstrated superior performance in practical applications even though dynamic programming is rather indirectly used. In AC, the critic provides a baseline by estimating either the value function or the action-value function. DP becomes crucial during the critic's update process, as it effectively reduces the variance in gradient estimation. The essence of employing DP lies in the fact that parameter updates are not solely dependent on

sampling new trajectories; rather, new information is bootstrapped. The collective efforts of a broad research community have contributed to the remarkable success of AC-type algorithms (such as Natural Policy Gradient (NPG), Trust Region Policy Optimization (TRPO), and Proximal Policy Optimization (PPO) [Kak01; Sch+15b; Sch+17]), although these methods have yet to be sufficiently supported by a comprehensive theoretical understanding.
With FT-DynPG, we also developed a hybrid approach where dynamic programming (DP) is directly, rather than indirectly, embedded within the algorithm. This integration led to an improved convergence rate with explicit constants for finite-time MDPs. This brings us to the following two research questions:

RQ4: *To what extent can the Markovian property of an infinite-time MDP be exploited more directly to improve the convergence behavior of PG methods?*

RQ5: *How much improvement is gained compared to vanilla PG?*

To address these questions, we introduce the dynamic policy gradient (DynPG) algorithm. Similar to FT-DynPG, dynamic programming (this time for infinite-time MDPs) is combined with vanilla PG. DynPG can be viewed as a hybrid RL algorithm that utilizes policy gradient to optimize a sequence of contextual bandits. In each iteration, the algorithm extends the horizon of the MDP by adding an additional epoch at the beginning and shifting trained policies to the future. We will see that adding a new epoch is analog to the application of the Bellman operator. A policy for the newly added epoch at time step 0 is trained using policy gradient, while the already trained policies are employed to determine the future actions. DynPG trains a (non-statioary) sequence of policies, which would solve a corresponding finite-time MDP. Nevertheless, we will see how finitely many steps are sufficient to result at the stationary optimal policy for the infinite-time problem. Note that DynPG is not an AC method. The algorithm minimizes the variance in the gradient estimation as much as possible by utilizing previously trained and therefore fixed policies to generate trajectories. This results in stable Q-value estimates and improved convergence behavior.
We first present a general error analysis for DynPG, which is theoretically compatible with any optimization scheme capable of solving a contextual bandit problem. This includes not only PG, but also NPG or Policy Mirror Descent (PMD). Following this, we focus on PG and the softmax parametrization to derive an explicit convergence rate for DynPG. Compared to vanilla PG, this approach eliminates the unknown constant and addresses a lower-bound example in which vanilla PG exhibits an exponentially poor dependence on the discount factor.
The theoretical results for DynPG are all with respect to the exact gradient assumption. We have briefly discussed that the convergence analysis for FT-DynPG is particularly challenging with stochastic gradients due to the absence of a global gradient domination property. Since the convergence rates obtained for FT-DynPG are not tight, we did not carry out the analogous analysis for DynPG.

In the third and final project of this dissertation, we return to stochastic gradient methods where the gradient can just be accessed though a first order oracle. We address the convergence of gradient methods under weak gradient domination and also look into the case where gradient domination is only locally fulfilled. We focus on almost sure convergence, the strongest type of

convergence for stochastic algorithms, where each individual run of the algorithm converges. This project was motivated by improving the convergence rates known for SPG, but it can also stand on its own as a contribution in the field of non-convex first order optimization.
In the first part, we consider the case where the weak gradient domination (WGD) is globally satisfied.

RQ6: *Is WGD sufficient to ensure almost sure convergence of stochastic gradient methods and can we derive a rate of convergence?*

We derive asymptotic convergence rates almost surely and in expectation for stochastic gradient descent (SGD) and stochastic Heavy Ball (SHB). For SGD, the almost sure convergence rate we obtain is arbitrarily close to the one obtained in expecation under WGD in [FBD21; Fat+22]. For SHB, convergence in expectation and almost sure convergence under WGD are new contributions.
In the second part, we relax the weak gradient domination property and assume that the gradient domination property is only locally fulfilled.

RQ7: *Can we still assure convergence of SGD when WGD is only locally fulfilled?*

We distinguish between WGD locally around a stationary point or locally around the global minimum. In both cases we prove that SGD, initialized in the local region, remains within the gradient dominated region with high probability, given a small enough step size. Conditioned on this event we provide converges rates almost surely and in expectation towards the local or global minimum respectively with the same convergence speed as in the global case.
The local gradient domination around stationary points is applicable to the training of neural networks (NNs) with supervised learning. All analytical functions, and thereby NNs with analytical activation functions, satisfy this assumption.
The case of local gradient domination around the global minimum applies to stochastic softmax policy gradient algorithms. In the case of infinite-time horizon MDPs, we show that the local WGD is satisfied under softmax parametrization for both vanilla PG and entropy-regularized PG. If the initialization is close to the global optimum, almost sure convergence is guaranteed. The convergence holds without requiring a batch size, meaning only a simple gradient estimator is needed, which represents a significant improvement over previous stochastic results. In finite-time MDPs, we show similarly that each individual optimization step converges almost surely to the global optimum under good initialization and sufficient small step size. In all cases, we can theoretically characterize the local regions for initialization.
We cannot make a direct comparison between vanilla PG and DynPG or between FT-SimPG and FT-DynPG, as the almost sure convergence rates are asymptotic and do not provide explicit constants for comparison. It should be noted that the asymptotic convergence rates of the respective algorithms are equivalent.
To conclude this introduction, we provide a brief overview of the outline. As three projects are consolidated in this thesis we clarify which parts of the dissertation are the author's original contributions and which parts are contributed by co-authors.

1. The Chapters 2 and 3 are background chapters where we cover the basics on first order gradient methods and introduce MDPs and the PG framework.

2. In Chapter 4, we address the research questions RQ1-RQ3 regarding finite-time MDPs. The results in this chapter are already published at ICLR 2024 under the title "Beyond Stationarity: Convergence Analysis of Stochastic Softmax Policy Gradient Methods" [KWD24]. This project is joint work with Simon Weissmann and Leif Döring.
   The project idea, as well as the theorems and proofs were carried out by the author of this thesis and supervised by the co-authors.

3. Chapter 5 covers the research questions RQ4 and RQ5. This project is a preprint titled "Structure Matters: Dynamic Policy Gradient" [Kle+24] and joint work with Xiangyuan Zhang, Tamer Başar, Simon Weissmann and Leif Döring.
   The project idea, as well as the theorems and proofs were carried out by the author of this thesis and supervised by Leif Döring, Simon Weissmann and Tamer Başar. Figure 5.1 and the foundation of the Python code used in the example in Section 5.4.2 was established by Xiangyuan Zhang.

4. In Chapter 6, we discuss the research questions RQ6 and RQ7, concerning the almost sure convergence of stochastic gradient methods. This project is also published as a preprint titled "On Almost Sure Convergence Rates for Stochastic Gradient Methods under Gradient Domination" [Wei+24].
   The project idea, as well as the proof concept for Lemma 6.1, the example in Section 6.4, and the application to neural networks in Section 6.6 were derived by Simon Weissmann. The idea for the proof of Lemma 6.10 was contributed by Waïss Azizian. All other results and proofs in this project, especially the idea to apply the results in RL, were carried out by the author of this thesis.

Preliminaries: First Order Optimization

2

In this chapter, we cover the basic convergence results of first order methods where we aim to solve the problem

$$\min_{x\in\mathbb{R}^d} f(x). \tag{2.1}$$

Throughout this chapter we assume that the objective function $f : \mathbb{R}^d \to \mathbb{R}$ is bounded from below by $f^* = \inf_{x\in\mathbb{R}^d} f(x) > -\infty$ and we denote by $\|\cdot\|$ the euclidean norm on $\mathbb{R}^d$ induced by the standard euclidean scalar product $\langle x, y\rangle = x^T y$.

2.1 Gradient Descent

As a first step, we recall the deterministic iterative update generated by gradient descent with constant step size $\alpha > 0$, i.e.

$$x_{n+1} = x_n - \alpha\nabla f(x_n), \quad x_0 \in \mathbb{R}^d. \tag{GD}$$

before we cover the stochastic version in the final section of this chapter.
We are interested in the convergence behavior of the sequence $(x_n)_{n\in\mathbb{N}}$ and specify the following sets of interesting limit points:

- $x \in \mathbb{R}^d$ is called a stationary point if $\nabla f(x) = 0$.
- $x \in \mathbb{R}^d$ is called a local minimum if there exists $r > 0$ such that $f(x) \leq f(y)$ for all $y \in U_r(x) := \{y \in \mathbb{R}^d : ||x - y|| < r\}$.
- $x \in \mathbb{R}^d$ is called a global minimum if $f(x) \leq f(y)$ for all $y \in \mathbb{R}^d$.
- $x \in \mathbb{R}^d$ is called a saddle point if x is a stationary point but neither a (local) minimum of f nor a (local) minimum of $-f$.

All global minima are local minima and all local minima are stationary points. In order to derive convergence of $(x_n)_{n\in\mathbb{N}}$ towards stationary points, the objective f is assumed to satisfy the classical smoothness assumption:

Assumption 2.1. *The objective function $f : \mathbb{R}^d \to \mathbb{R}$ is differentiable and L-smooth, i.e the corresponding gradient ∇f is assumed to be L-Lipschitz continuous:*

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|, \quad \textit{for all } x, y \in \mathbb{R}^d. \tag{2.2}$$

From L-smoothness, the descent lemma is deduced. It is a fundamental instrument to analyze first order optimization methods.

Lemma 2.2. *[Bec17, Lem. 5.7] Let $f : \mathbb{R}^d \to \mathbb{R}$ fulfill Assumption 2.1, then for every $x, y \in \mathbb{R}^d$ we have*

$$f(y) \leq f(x) + \langle\nabla f(x), y - x\rangle + \frac{L}{2}\|y - x\|^2. \tag{2.3}$$

Applying the descent lemma to the iteration scheme equation (GD) with $y = x_{n+1}$, $x = x_n$ and step size $0 < \alpha \leq \frac{1}{L}$ yields the following iterative descent property

$$[f(x_{n+1}) - f^*] \leq [f(x_n) - f^*] - \frac{\alpha}{2}\|\nabla f(x_n)\|^2, \tag{2.4}$$

where f^* is subtracted on both sides of the inequality.
Rearranging this inequality and summing over the the iterations results in the following upper bound on the sum of the gradients

$$\sum_{n=1}^{N} \|\nabla f(x_n)\|^2 \leq \frac{2}{\alpha}[f(x_1) - f(x_{N+1})] \leq \frac{2}{\alpha}[f(x_1) - f^*].$$

We deduce directly that $\lim_{n\to\infty} \|\nabla f(x_n)\|^2 \to 0$ for $n \to \infty$ which proves the following theorem.

THEOREM 2.3. *Let $f : \mathbb{R}^d \to \mathbb{R}$ fulfill Assumption 2.1. Then, every accumulation point of the gradient descent scheme $(x_n)_{n\in\mathbb{N}}$ defined in equation* (GD) *with step size $0 < \alpha \leq \frac{1}{L}$ is a stationary point of f.*

Convergence towards stationary points is not what we originally aimed for. Instead we wish to converge towards global minima of f. To derive such a result from the descent inequality, equation (2.4), the term $\|\nabla f(x_n)\|^2$ has to be controlled. In the following subsections we will get to know two sufficient conditions to ensure convergence towards global minima of f.

2.1.1 *Convergence of GD under strong convexity*

The classical assumption to ensure that GD converges to a global minimum at a linear rate is the strong convexity assumption:

DEFINITION 2.4. A differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ is strongly convex if there exists $\mu > 0$ such that for every $x, y \in \mathbb{R}^d$ it holds that

$$f(y) \geq f(x) + \langle \nabla f(x), y - x\rangle + \frac{\mu}{2}\|y - x\|^2. \tag{2.5}$$

Note here, that strong convexity implies a unique global minimum $x^* \in \mathbb{R}^d$ such that $f^* = f(x^*)$. Minimizing equation (2.5) in y, we obtain that

$$f^* \geq f(x) - \frac{1}{2\mu}\|\nabla f(x)\|^2.$$

Rearranging the terms reveals that strong convexity implies the so called Polyak-Łojasiewicz (PL) inequality [Pol63]:

$$\|\nabla f(x)\|^2 \geq 2\mu(f(x) - f^*). \tag{PL}$$

Remark 2.5. If a function $f : \mathbb{R}^d \to \mathbb{R}$ is L-smooth and μ-strongly convex, then is has to hold that $L \geq \mu$. Combining equation (2.2) and equation (2.5) leads to

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

which induces that $L \geq \mu$ holds true.

Applying this property in the descent inequality, equation (2.4), results in the following convergence theorem.

THEOREM 2.6. *Let $f : \mathbb{R}^d \to \mathbb{R}$ fulfill Assumption 2.1 and be μ-strongly-convex. Let $(x_n)_{n\in\mathbb{N}}$ be the GD scheme from equation* (GD) *with step size $0 < \alpha < \min\{\frac{1}{L}, \frac{1}{\mu}\} \equiv \frac{1}{L}$. Then $(x_n)_{n\in\mathbb{N}}$ (and $(f(x_n))_{n\in\mathbb{N}}$, respectively) converges towards the unique global minimum x^* (f^*, respectively), at a linear rate. More precisely, it holds that*

$$\frac{\mu}{2}\|x_{n+1} - x^*\|^2 \leq f(x_{n+1}) - f(x^*) \leq (1 - \alpha\mu)^n [f(x_1) - f(x^*)].$$

Proof. First recall that $f(x^*) = f^*$ by the unique global minimum under strong convexity. The first inequality follows directly from the fact that $\nabla f(x^*) = 0$ and the definition of μ-strong-convexity in equation (2.5). We obtain

$$\frac{\mu}{2}\|x_{n+1} - x^*\|^2 = \langle \nabla f(x^*), x_{n+1} - x^* \rangle + \frac{\mu}{2}\|x_{n+1} - x^*\|^2 \leq f(x_{n+1}) - f(x^*). \tag{2.6}$$

Next, we apply equation (PL) to the descent inequality, equation (2.4), which results in

$$\begin{aligned}[f(x_{n+1}) - f^*] &\leq [f(x_n) - f^*] - \alpha\mu[f(x_n) - f^*] \\ &= (1 - \alpha\mu)[f(x_n) - f^*].\end{aligned}$$

We can iterate this inequality to deduce that

$$[f(x_{n+1}) - f^*] \leq (1 - \alpha\mu)^n [f(x_1) - f^*].$$

As $(1 - \alpha\mu) \in (0, 1)$ we conclude that $f(x_n) \to f^*$ for $n \to \infty$ at a linear rate. ∎

The proof demonstrates that strong convexity is only required to ensure the convergence of x_n to x^*. In contrast, the convergence of $f(x_n)$ is derived solely from the PL-inequality, which was itself a consequence of strong convexity (see also [KNS16, Thm. 1]).
It is noteworthy that convexity, and particularly strong convexity, is often not satisfied in practical applications. However, weaker gradient domination properties like the PL inequality are more readily met and can be established in certain ML and RL contexts which will be evidenced repeatedly throughout this thesis. Thus, in the subsequent section, we analyze the fundamental convergence properties of gradient descent under the weaker assumption of gradient domination and establish the convergence of $f(x_n)$ towards global optima.

2.1.2 *Convergence of GD under Gradient Domination*

In order to derive a convergence rate without assuming (strong) convexity of f one can use dominating relations of the gradient $\nabla f(x)$ with respect to the optimality gap $f(x) - f^*$.

DEFINITION 2.7. Let $f : \mathbb{R}^d \to \mathbb{R}$ be continuously differentiable with $f^* = \inf_{x\in\mathbb{R}^d} f(x) > -\infty$. We say that f satisfies the global gradient domination property with parameter $\beta \in [\frac{1}{2}, 1]$ if there exists $c > 0$ such that for all $x \in \mathbb{R}^d$ it holds true that

$$\|\nabla f(x)\| \geq c(f(x) - f^*)^\beta.$$

When the exponent $\beta = 1/2$ we recover the PL-inequality from equation (PL). If $\beta = \frac{1}{2}$, we will call the gradient domination *strong* since it is implied by strong convexity. In contrast, we call the gradient domination *weak* for $\beta \in (\frac{1}{2}, 1]$.

Remark 2.8. Note here, that gradient domination for some $\beta \in [\frac{1}{2}, 1]$ implies gradient domination for any weaker $\beta' \in [\beta, 1]$.

In [Fat+22; Att+10; BST14; ZWL18] examples of functions are discussed that fulfill the (weak) gradient domination property. For instance, one-dimensional monomials $f(x) = |x|^p$, $p \geq 2$, satisfy the weak global gradient domination property with $\beta = \frac{p-1}{p}$. We refer to [Fat+22, App. A] for a longer list of globally gradient dominated functions including convex and non-convex functions.
For $\beta = \frac{1}{2}$ we already obtained linear convergence in Theorem 2.6. In the following result, we obtain sub-linear convergence under the weakest form of gradient domination with $\beta = 1$. Note, that the same upper bound on the convergence rate holds also for any $\beta \in [\frac{1}{2}, 1]$ by Remark 2.8. The convergence relies on the following auxiliary lemma, which demonstrates the convergence of a deterministic sequence under the condition that a specific descent inequality is satisfied.

LEMMA 2.9. *[KWD24, Lem. B.7] Let $(d_n)_{n\in\mathbb{N}}$ be a positive sequence, such that $d_{n+1} \leq d_n - qd_n^2$ for some $q > 0$, then $d_n \leq \frac{1}{(n-1)q}$. If in addition $d_1 < \frac{1}{q}$, then $d_n \leq \frac{1}{qn}$.*

Proof. We use an argument similar to Nesterov [Nes13, Thm. 2.1.14]. It holds

$$\frac{1}{d_{n+1}} \geq \frac{1}{d_n} + \frac{qd_n}{d_{n+1}} \geq \frac{1}{d_n} + q,$$

where the first inequality is due to dividing by $d_n d_{n+1}$ and the second inequality follows by monotonicity. Using a telescope-sum argument we obtain

$$\frac{1}{d_n} = \frac{1}{d_1} + \sum_{k=1}^{n-1} \frac{1}{d_{k+1}} - \frac{1}{d_k} \geq \frac{1}{d_1} + (n-1)q.$$

Finally,

$$d_n \leq \frac{1}{(n-1)q + \frac{1}{d_0}} \leq \frac{1}{(n-1)q}.$$

and if $d_1 < \frac{1}{q}$, then

$$d_n \leq \frac{1}{(n-1)q + \frac{1}{d_1}} \leq \frac{1}{qn}.$$

■

THEOREM 2.10. *Let $f : \mathbb{R}^d \to \mathbb{R}$ fulfill Assumption 2.1 and the weak gradient domination property with $c > 0$ and $\beta = 1$. Let $(x_n)_{n\in\mathbb{N}}$ be the GD scheme from equation* (GD) *with step size $0 < \alpha \leq \frac{1}{L}$. Then $(f(x_n))_{n\in\mathbb{N}}$ converges towards the global minimum f^* at a sub-linear rate, i.e.*

$$f(x_n) - f(x^*) \leq \frac{2}{n\alpha c}.$$

Proof. Applying the general gradient domination property for $\beta = 1$ in the recursive descent property, equation (2.4) reads as

$$[f(x_{n+1}) - f^*] \leq [f(x_n) - f^*] - \frac{\alpha c}{2}[f(x_n) - f^*]^2.$$

We can apply Lemma 2.9 with $d_n = f(x_n) - f^*$ and define $q = \frac{\alpha c}{2}$, to deduce that

$$f(x_{n+1}) - f^* \leq \frac{2}{n\alpha c}.$$

■

We formulate the following stronger version of the theorem where it is sufficient that the gradient domination is fulfilled only along the gradient trajectory. We apply this result in the RL applications in Chapter 4.

THEOREM 2.11. *Let $f : \mathbb{R}^d \to \mathbb{R}$ fulfill Assumption 2.1 and denote by $(x_n)_{n\in\mathbb{N}}$ is the GD scheme with step size $0 < \alpha \leq \frac{1}{L}$. For $x_1 \in \mathbb{R}^d$ assume that $f(x_1) - f^* \leq \frac{2}{\alpha b}$ and additionally the gradient domination property $\|\nabla f(x_n)\|^2 \geq b(f^* - f(x_n))^2$ holds for every $n \in \mathbb{N}$. Then, for any $n \in \mathbb{N}$,*

$$f(x_n) - f^* \leq \frac{2}{\alpha b n}.$$

Proof. We deduce from L-smoothness and the descent inequality in equation (2.4) that

$$[f(x_{n+1}) - f^*] \leq [f(x_n) - f^*] - \frac{\alpha}{2}\|\nabla f(x_n)\|^2.$$

Together with the gradient domination assumption we obtain

$$[f(x_{n+1}) - f^*] \leq [f(x_n) - f^*] - \alpha b[f(x_n) - f^*]^2.$$

Applying Lemma 2.9 with the initial condition $f(x_1) - f^* \leq \frac{2}{\alpha b}$ results in the claim. ■

2.2 Stochastic Gradient Descent

In this section, we will drop the assumption of access to the exact gradient ∇f. Instead we consider access to a stochastic fist order oracle which provides us with unbiased samples of the gradient in any point.[1]

2.2.1 *Assumptions on the Stochastic First Order Oracle*

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be an underlying probability space. We assume that we can estimate the exact gradient $\nabla f(x)$ at any $x \in \mathbb{R}^d$ through a stochastic first order oracle $V : \mathbb{R}^d \times M \to \mathbb{R}^d$ defined by

$$V(x, m) = \nabla f(x) + Z(x, m), \quad x \in \mathbb{R}^d,\ m \in M, \tag{2.7}$$

where $(M, \mathcal{M})$ is a measurable space, $Z : \mathbb{R}^d \times M \to \mathbb{R}^d$ is a state dependent $\mathcal{B}(\mathbb{R}^d) \otimes \mathcal{M}/\mathcal{B}(\mathbb{R}^d)$-measurable mapping describing the error to the exact gradient ∇f. The stochastic gradient evaluation is then modelled through $V(x, \zeta)$, where the random variable $\zeta : \Omega \to M$ is independent of the state x. We make the following unbiasedness and second moment assumption:

[1] Section 2.2 is an extended version of [Wei+24, Sec. 2].

ASSUMPTION 2.12. *We assume that for each $x \in \mathbb{R}^d$ it holds that*

$$\mathbb{E}[Z(x,\zeta)] := \int_\Omega Z(x,\zeta(\omega))\mathrm{d}\mathbb{P}(\omega) = 0$$

and there exist non-negative constants A, B and C such that for all $x \in \mathbb{R}^d$ it holds that

$$\mathbb{E}[\|V(x,\zeta)\|^2] \leq A(f(x) - f^*) + B\|\nabla f(x)\|^2 + C\,. \tag{ABC}$$

It is worth noting that the (ABC) assumption is a generalization of the bounded variance assumption that appears for $A = B = 0$. It was introduced by Khaled and Richtárik [KR23] as expected smoothness condition and shown to be the weakest assumption among many others. We describe the stochastic gradient descent scheme as discrete time stochastic processes (X_n) driven by noisy gradient evaluations in equation (2.7). In each iteration, we assume that the stochastic first order oracle is accessed through the evaluation of ζ_{n+1} which is a copy of ζ independent from the current state X_n.
The stochastic gradient descent (SGD) scheme is given by the stochastic update

$$X_{n+1} = X_n - \alpha_n V(X_n, \zeta_{n+1})\,,$$

where X_0 is a $\mathbb{R}^d$-valued random vector which denotes the initial state. To keep the notation simple, we will introduce $V_{n+1}(X_n) := V(X_n, \zeta_{n+1})$ suppressing the explicit noise representation through (ζ_n) in the following. The iterative update formula then reads as

$$X_{n+1} = X_n - \alpha_n V_{n+1}(X_n). \tag{SGD}$$

Here, (α_n) denotes a sequence of positive step sizes and we denote by $(\mathcal{F}_n)_{n\in\mathbb{N}}$ the natural filtration induced by the process $(X_n)_{n\in\mathbb{N}}$.

Example 2.13 (Expected risk minimization). In order to give more insights into the considered setting we formulate a stochastic first order oracle based on expected risk minimization. In expected risk minimization we are interested in minimizing an objective function of the form

$$f(x) = \mathbb{E}[F(x,\zeta)] = \int_\Omega F(x,\zeta(\omega))\, d\mathbb{P}(\omega)$$

where $F : \mathbb{R}^d \times M \to \mathbb{R}$ is $\mathcal{B}(\mathbb{R}^d) \otimes \mathcal{M}/\mathcal{B}(\mathbb{R})$-measurable. In our notation the stochastic first order oracle then takes the form

$$V(x,\zeta) = \nabla f(x) + (\nabla_x F(x,\zeta) - \nabla f(x)) = \nabla_x F(x,\zeta)$$

and the iterative update of SGD reads as

$$X_{n+1} = X_n - \alpha_n \nabla_x F(X_n, \zeta_{n+1})$$

with a sequence of independent and identically distributed (ζ_n). Note that this scenario also includes empirical risk minimization where the objective function takes a finite sum form

$$f(x) = \frac{1}{N}\sum_{i=1}^N F(x,i) = \mathbb{E}[F(x,\zeta)]\,,$$

with $\zeta \sim \mathcal{U}(\{1,\dots,N\})$.

2.2.2 *Types of Convergence*

As in the deterministic scheme we are again interested in the convergence behaviour of the stochastic processes $(X_n)_{n\in\mathbb{N}}$ or $(f(X_n))_{n\in\mathbb{N}}$. Thereby we distinguish between the following two types of convergence:

- We say the stochastic process $(X_n)_{n\in\mathbb{N}}$ with values in $\mathbb{R}^d$ convergences in expectation (or in L^2) towards a point $x \in \mathbb{R}^d$, if $\mathbb{E}[\|X_n - x\|^2] \to 0$ for $n \to \infty$. Respectively, we say $(f(X_n))_{n\in\mathbb{N}}$ convergence in expectation (or in L^1) against a level $l \in \mathbb{R}$ if $\mathbb{E}[|f(X_n) - l|] \to 0$ for $n \to \infty$.

- We say the stochastic process $(X_n)_{n\in\mathbb{N}}$ with values in $\mathbb{R}^d$ convergences almost surely towards a point $x \in \mathbb{R}^d$, if there exists $A \in \mathcal{F}$ with $\mathbb{P}(A) = 1$ and $X_n(\omega) \to x$ for $n \to \infty$ and every $\omega \in A$. Respectively, we say $(f(X_n))_{n\in\mathbb{N}}$ convergence almost surely against a level $l \in \mathbb{R}$ if there exists $A \in \mathcal{F}$ with $\mathbb{P}(A) = 1$ and $f(X_n(\omega)) \to l$ for $n \to \infty$ and every $\omega \in A$.

It is worth noting that are also other types of convergences, e.g. L^p-convergence for $p \in (0, \infty]$ or convergence with high probability, which will not be discussed in the scope of this work. In the first part of this thesis we will mainly focus on convergence in expectation. In the final chapter we move on to almost sure convergence, the strongest convergence type for stochastic processes.

2.2.3 *Typical Steps to Derive Convergence Results*

We now outline the standard procedure for convergence analysis of stochastic gradient descent (SGD) to establish convergence in expectation. The analysis begins by leveraging the smoothness of the function f by applying the descent inequality in equation (2.3) to the iterative scheme,

$$\begin{aligned} f(X_{n+1}) &\leq f(X_n) - \langle \nabla f(X_n), X_{n+1} - X_n \rangle + \frac{L}{2}\|X_{n+1} - X_n\|^2 \\ &= f(X_n) - \alpha_n \langle \nabla f(X_n), V_{n+1}(X_n) \rangle + \frac{L\alpha_n^2}{2}\|V_{n+1}(X_n)\|^2 \end{aligned}$$

and subsequently taking conditional expectations,

$$\mathbb{E}[f(X_{n+1}) \mid \mathcal{F}_n] \leq f(X_n) - \alpha_n \|\nabla f(X_n)\|^2 + \frac{L\alpha_n^2}{2}\mathbb{E}[\|V_{n+1}(X_n)\|^2 \mid \mathcal{F}_n].$$

Next, f^* is subtracted on both sides and the gradient domination variance term of the stochastic gradient is controlled through the (ABC) assumption,

$$\mathbb{E}[f(X_{n+1}) - f^* \mid \mathcal{F}_n] \leq \left(1 + \frac{LA\alpha_n^2}{2}\right)(f(X_n) - f^*) - \left(\alpha_n - \frac{BL\alpha_n^2}{2}\right)\|\nabla f(X_n)\|^2 + \frac{LC\alpha_n^2}{2}. \tag{2.8}$$

Without further assumptions this inequality can now be used to show that the gradient $\nabla f(X_n)$ converges to zero almost surely and in expectation. In order to obtain convergence towards a global optimum additional assumptions, like convexity or gradient domination, are needed (similar to the deterministic case). For instance, incorporating the global gradient domination

property defined in Definition 2.7 yields an iterative inequality of the form

$$\begin{aligned}&\mathbb{E}[f(X_{n+1}) - f^* \mid \mathcal{F}_n] \\ &\leq \Big(1 + \frac{LA\alpha_n^2}{2}\Big)(f(X_n) - f^*) - \Big(\alpha_n - \frac{BL\alpha_n^2}{2}\Big)c^2(f(X_n) - f^*)^{2\beta} + \frac{LC\alpha_n^2}{2}.\end{aligned} \tag{2.9}$$

Now, taking the expectation on both sides of the inequality one can derive a sub-linear convergence rate in expectation by working with recursive inequalities [Fat+22; FBD21]. In Chapter 6 we push the argument further. We combine smoothness and gradient domination with a variant of the Robbins-Siegmund Theorem to derive almost sure convergence rates in the (weak) gradient dominated case.
We will refrain from delving further into the details at this point and instead direct the reader to Chapter 6 for an in-depth review of the relevant literature on the convergence properties of SGD.

2.3 Variants of Robbins-Siegmund Theorem

In the following section, we provide two specific convergence theorems used to prove almost sure convergence (Lemma 2.15) as well as convergence in expectation (Lemma 2.16). The former one is a direct consequence of the well-known Robbins-Siegmund theorem, provided here for completeness[2].

Theorem 2.14 (Theorem 1 in [RS71]). *Let $(\Omega, \mathcal{F}, (\mathcal{F}_n)_{n\in\mathbb{N}}, \mathbb{P})$ be a filtered probability space, $(Z_n)_{n\in\mathbb{N}}$, $(A_n)_{n\in\mathbb{N}}$, $(B_n)_{n\in\mathbb{N}}$ and $(C_n)_{n\in\mathbb{N}}$ be non-negative and adapted stochastic processes with*

$$\sum_{n=1}^{\infty} A_n < \infty \quad \text{and} \quad \sum_{n=1}^{\infty} B_n < \infty$$

almost surely. Suppose that for each $n \in \mathbb{N}$ the recursion

$$\mathbb{E}[Z_{n+1} \mid \mathcal{F}_n] \leq (1 + A_n)Z_n + B_n - C_n$$

is satisfied, then (i) there exists an almost surely finite random variable Z_∞ such that $Z_n \to Z_\infty$ almost surely as $n \to \infty$ and (ii) $\sum_{n=1}^{\infty} C_n < \infty$ almost surely.

Lemma 2.15. *Let $(\Omega, \mathcal{F}, (\mathcal{F}_n)_{n\in\mathbb{N}}, \mathbb{P})$ be a filtered probability space, $(Y_n)_{n\in\mathbb{N}}$, $(a_n)_{n\in\mathbb{N}}$, $(b_n)_{n\in\mathbb{N}}$ and $(r_n)_{n\in\mathbb{N}}$ be non-negative and adapted stochastic processes with*

$$\sum_{n=1}^{\infty} a_n = \infty, \quad \sum_{n=1}^{\infty} b_n < \infty \quad \text{and} \quad r_n > 0$$

almost surely. Suppose that for each $n \in \mathbb{N}$ the recursion

$$\mathbb{E}[r_{n+1}Y_{n+1} \mid \mathcal{F}_n] \leq (1 - a_n)r_nY_n + b_n$$

is satisfied, then we have $r_nY_n \to 0$ almost surely as $n \to \infty$.

[2]The results presented in this subsection are needed in Chapter 6 for the almost sure convergence analysis of SGD under gradient domination and this section is part of the preprint Weissmann et al. [Wei+24, Appendix A.4.]

Proof. We define $Z_n := r_n Y_n$, $B_n := b_n$ and $C_n := a_n r_n Y_n$ such that

$$\mathbb{E}[Z_{n+1} \mid \mathcal{F}_n] \leq Z_n - C_n + B_n$$

for $n \in \mathbb{N}$. Using Theorem 2.14 we observe that there exists Z_∞ almost surely finite such that $Z_n = r_n Y_n \to Z_\infty$ almost surely as $n \to \infty$. Recall that all sequences are positive and suppose that $Z_\infty > 0$ with positive probability. Then, for all $\omega \in \Omega$ with $Z_\infty > 0$ choose $m \in \mathbb{N}$ such that $Z_n(\omega) > \epsilon$ for all $n \geq m$ such that

$$\sum_{n=1}^{\infty} a_n r_n Y_n(\omega) \geq \sum_{n=m}^{\infty} a_n r_n Y_n(\omega) \geq \epsilon \sum_{n=m}^{\infty} a_n = \infty.$$

This contradicts that

$$\sum_{n=1}^{\infty} C_n = \sum_{n=1}^{\infty} a_n r_n Y_n < \infty$$

almost surely by Theorem 2.14 (ii). We conclude that

$$\lim_{n\to\infty} r_n Y_n = 0$$

almost surely has to hold. ∎

The following Lemma will be applied to prove convergence in expectation.

LEMMA 2.16. *Let $(w_n)_{n\in\mathbb{N}}$ be a non-negative sequence, such that $w_{n+1} \leq (1 - a_n)w_n + b_n$, where $(a_n)_{n\in\mathbb{N}}$ and $(b_n)_{n\in\mathbb{N}}$ are non-negative sequences satisfying*

$$\sum_{n=1}^{\infty} a_n = \infty \quad \text{and} \quad \sum_{n=1}^{\infty} b_n < \infty.$$

Then, $\lim_{n\to\infty} w_n = 0$.

Proof. W.l.o.g we assume that $w_{n+1} = (1 - a_n)w_n + b_n$, otherwise we could just increase a_n or decrease b_n which would have no effect on the summation tests. We obtain

$$-w_1 \leq w_n - w_1 = \sum_{k=1}^{n-1} (w_{k+1} - w_k) = \sum_{k=1}^{n-1} b_k - \sum_{k=1}^{n-1} w_k a_k.$$

Since $w_n - w_1$ is bounded below and $\sum_{k=1}^{\infty} b_k < \infty$, we deduce that $\sum_{k=1}^{n} w_k a_k$ is bounded. Since all summands are positive, the infinite sum converges. Thus, as a difference of two converging series also $(w_n)_{n\in\mathbb{N}}$ converges. Finally, the convergence of $\sum_{k=1}^{\infty} w_k a_k$ implies $\liminf_{n\to\infty} w_n = 0$ which, by the convergence of $(w_n)_{n\in\mathbb{N}}$, implies $\lim_n w_n = \liminf_n w_n = 0$. ∎

Preliminaries: Markov Decision Processes and Policy Gradient

3

In this chapter, we formally introduce Markov decision processes (MDPs) as a framework for modeling reinforcement learning (RL) problems and provide the necessary preliminaries on the policy gradient (PG) algorithm. We begin by focusing on discounted infinite-time horizon MDPs and subsequently address the distinct challenges and differences encountered in finite-time horizon problems. The definitions and results discussed here are well-established in standard MDP literature and can be found, for example, in [Put05; SB18].

3.1 Discounted infinite-time horizon MDPs

We denote by $\Delta(\mathcal{E})$ the probability simplex over a finite set $\mathcal{E}$ and for a function $f : \mathcal{E} \to \mathbb{R}$ or a point $x \in \mathbb{R}^d$, we denote its supremum norm by $\|f\|_\infty = \max_{x\in\mathcal{E}} |f(x)|$ or $\|x\|_\infty = \max_{i=1,\dots,d} |x_i|$, respectively. Just $\|\cdot\|$ always denotes the euclidean norm in $\mathbb{R}^d$ for $d \geq 1$.

Definition 3.1 (Discounted Markov decision process). The quintet $(\mathcal{S}, \mathcal{A}, \gamma, p, r)$ given by

- the finite state space $\mathcal{S}$,
- the finite action space $\mathcal{A}$,
- the discount factor $\gamma \in [0, 1)$,
- the transition function $p : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{A})$ and
- the reward function $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$

is called (discounted) Markov decision process (MDP). We write $p(s'|s, a)$ for the transition probability of state $s' \in \mathcal{S}$ given that we a currently in state $s \in \mathcal{S}$ and played action $a \in \mathcal{A}$ and denote by $r(s, a)$ the reward of playing action $a \in \mathcal{A}$ in state $s \in \mathcal{S}$.

A MDP is a mathematical framework to model the sequential decision making of an agent in an (unknown) environment. This is to say, the transition probabilities p are in general unknown to the agent. We model the agent through a so called policy, a function which determines the probability distribution over the action space.

Definition 3.2 (Policy). Let $(\mathcal{S}, \mathcal{A}, \gamma, p, r)$ be a MDP.

(i) A policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$ is a mapping from a state $s \in \mathcal{S}$ to a distribution over the action space, i.e $\pi(\cdot|s) \in \Delta(\mathcal{A})$. The set of all policies is denoted by Π.

(ii) A policy $\pi \in \Pi$ is deterministic if for every $s \in \mathcal{S}$ there exists $a \in \mathcal{A}$ such that $\pi(a|s) = 1$. Otherwise we call a policy stochastic.

(iii) A sequence of policies is denoted by $\boldsymbol{\pi} = (\pi_t)_{t=0}^\infty$ in Π^∞ and we call $\boldsymbol{\pi}$ a policy of an MDP.

(iv) A sequence of policies $\pi \in \Pi^\infty$ is stationary if $\pi_t = \pi$ for all $t \geq 0$ and we write just π instead of π.

For the remainder of this section we assume an underlying discounted MDP $(\mathcal{S}, \mathcal{A}, \gamma, p, r)$ with infinite-time horizon and $\gamma \in [0, 1)$ and let $\pi = (\pi_t)_{t=0}^\infty \in \Pi^\infty$ be a policy of the MDP.
The general goal in infinite-time horizon RL is to find a policy which maximizes the discounted sum of expected rewards. The quantity of interest in this optimization problem is called value function and defined in the following.

DEFINITION 3.3 (Discounted value, state-action value and advantage function).

(i) We define the value function $V^\pi : \mathcal{S} \to \mathbb{R}$ under policy π in state $s \in \mathcal{S}$ by

$$V^\pi(s) = \mathbb{E}_s^\pi\Big[\sum_{t=0}^\infty \gamma^t r(S_t, A_t)\Big], \tag{3.1}$$

where $\mathbb{E}_s^\pi$ denotes the expectation regarding the probability measure $\mathbb{P}_s^\pi$ induced by $S_0 = s$, $A_t \sim \pi_t(\cdot|S_t)$ and $S_{t+1} \sim p(\cdot|S_t, A_t)$ for all $t \geq 0$.

(ii) For an initial state distribution μ over the state space $\mathcal{S}$ we define $V^\pi(\mu) := \mathbb{E}_{S\sim\mu}[V^\pi(S)]$. We replace s with μ in the underlying probability measure and write $\mathbb{P}_\mu^\pi$ instead of $\mathbb{P}_s^\pi$.

(iii) The state-action value function (or Q-function) $Q^\pi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ under policy π in a state-action pair $(s, a) \in \mathcal{S} \times \mathcal{A}$ is defined by

$$Q^\pi(s, a) = \mathbb{E}_{s,a}^\pi\Big[\sum_{t=0}^\infty \gamma^t r(S_t, A_t)\Big], \tag{3.2}$$

where $\mathbb{E}_{s,a}^\pi$ denotes the expectation regarding the probability measure $\mathbb{P}_{s,a}^\pi$ induced by $S_0 = s$, $A_0 = a$, $A_t \sim \pi_t(\cdot|S_t)$ and $S_t \sim p(\cdot|S_{t-1}, A_{t-1})$ for all $t \geq 1$.

(iv) We define the advantage function $A^\pi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ under policy π in a state-action pair $(s, a) \in \mathcal{S} \times \mathcal{A}$ by

$$A^\pi(s, a) = Q^\pi(s, a) - V^\pi(s).$$

Remark 3.4. Note that every discounted MDP with $\gamma = 0$ indirectly implies a time horizon of length 1 as all factors γ^t in the value function are equal 0 besides when $t = 0$. We refer to this special case as contextual bandit problem, where the state s represents the context and the goal is to take an optimal action (arm) for every possible context.
In addition, we obtain from the definition of the value function the necessity of the discount factor $\gamma \in [0, 1)$. The finite state and action space ensures that the absolute values of the reward function are bounded by some $R^{\max} \in \mathbb{R}_+$ due to finitely many values. Thus, the value, state-action value and advantage function are well defined due to the discount factor $\gamma < 1$ and we have that

$$V^\pi(s) = \mathbb{E}_s^\pi\Big[\sum_{t=0}^\infty \gamma^t r(S_t, A_t)\Big] \leq R^{\max} \sum_{t=0}^\infty \gamma^t = \frac{R^{\max}}{1-\gamma}.$$

Definition 3.5 (Optimal value and state-action value function and optimal policy).

(i) The optimal value function is defined by $V^*(s) = \sup_{\mathbb{\pi} \in \Pi^\infty} V^{\mathbb{\pi}}(s)$ for all $s \in \mathcal{S}$.

(ii) The optimal state action value function is defined by $Q^*(s,a) = \sup_{\mathbb{\pi} \in \Pi^\infty} Q^{\mathbb{\pi}}(s,a)$ for all $s \in \mathcal{S}$, $a \in \mathcal{A}$.

(iii) Policies $\mathbb{\pi} \in \Pi^\infty$ which satisfy $V^{\mathbb{\pi}}(s) = V^*(s)$ are called optimal policies and denoted by $\mathbb{\pi}^*$.

It is a well-established result that stationary policies are sufficient for solving discounted MDPs.

Proposition 3.6. *[Put05, Thm. 6.2.7] Suppose a discounted MDP* $(\mathcal{S}, \mathcal{A}, \gamma, p, r)$*. Then it holds that*

$$V^*(\mu) = \sup_{\mathbb{\pi} \in \Pi^\infty} V^{\mathbb{\pi}}(\mu) = \sup_{\pi \in \Pi} V^{\pi}(\mu).$$

As a consequence of this result, the problem of solving discounted MDPs is typically reduced to identifying an optimal stationary policy $\pi^* \in \Pi$. Accordingly, we use the superscript π in V^π, Q^π or A^π to denote the (state-action) value function under a stationary policy.

Remark 3.7. Note that optimal policies do not need to be unique. Moreover, it can be shown that, due to the finite state and action space, at least one deterministic optimal policy exists [Put05, Thm. 6.2.10].

Next, we define the state visitation measure induced by a policy π, which quantifies the frequency or likelihood of visiting different states under a particular policy.

Definition 3.8 (State visitation measure and distribution).

(i) For an initial state distribution μ the state visitation measure under policy $\pi \in \Pi$ is defined by

$$\rho^\pi_\mu(s) = \sum_{t=0}^{\infty} \gamma^t \mathbb{P}^\pi_\mu(S_t = s).$$

(ii) The corresponding state visitation distribution is the defined by the induced probability measure

$$d^\pi_\mu(s) = (1-\gamma)\rho^\pi_\mu(s).$$

The performance difference lemma is a useful identity to compare policies. It turns out to be very useful to prove convergence of policy gradient methods [Aga+21].

Lemma 3.9. *[KL02, Lem. 6.1] For any two policies* $\pi, \pi' \in \Pi$ *and any initial state distribution* μ *it holds that*

$$V^{\pi'}(\mu) - V^\pi(\mu) = \frac{1}{1-\gamma} \mathbb{E}_{S \sim d^{\pi'}_\mu, A \sim \pi'(\cdot|s)} [A^\pi(S,A)]$$

3.1.1 *Dynamic programming principle*

Finding the optimal solution to an MDP under prefect information, i.e. when the transition functions are known, can be done using the dynamic programming (DP) principle.
Therefore we derive the following fix point relations for the value and state-action value function.

PROPOSITION 3.10. *For any policy $\pi \in \Pi$ and $s \in \mathcal{S}$ it holds that*

$$V^\pi(s) = \sum_{a \in \mathcal{A}} \pi(a|s) Q^\pi(s,a) = \sum_{a \in \mathcal{A}} \pi(a|s) \Big(r(s,a) + \gamma \sum_{s' \in \mathcal{S}} p(s'|s,a) V^\pi(s') \Big).$$

And similar for the state action value function, for any policy $\pi \in \Pi$ and $s \in \mathcal{S}$, $a \in \mathcal{A}$ it holds that

$$Q^\pi(s,a) = r(s,a) + \gamma \sum_{s' \in \mathcal{S}, a' \in \mathcal{A}} p(s'|s,a) \pi(a'|s') Q^\pi(s',a').$$

Proof. First of all, note that by definition

$$Q^\pi(s,a) = \mathbb{E}^\pi_{s,a}[\sum_{t=0}^{\infty} \gamma^t r(S_t, A_t)] = r(s,a) + \gamma \sum_{s' \in \mathcal{S}} p(s'|s,a) \mathbb{E}^\pi_{s'}[\sum_{t=0}^{\infty} \gamma^t r(S_t, A_t)].$$

Moreover, we have

$$\begin{aligned}
V^\pi(s) &= \mathbb{E}^\pi_s[\sum_{t=0}^{\infty} \gamma^t r(S_t, A_t)] \\
&= \mathbb{E}^\pi_s[r(S_0, A_0) + \gamma \sum_{t=0}^{\infty} \gamma^t r(S_{t+1}, A_{t+1})] \\
&= \sum_{a \in \mathcal{A}} \pi(a|s) \Big(r(s,a) + \gamma \sum_{s' \in \mathcal{S}} p(s'|s,a) \mathbb{E}^\pi_{s',a'}[\sum_{t=0}^{\infty} \gamma^t r(S_t, A_t)] \\
&= \sum_{a \in \mathcal{A}} \pi(a|s) Q^\pi(s,a) \\
&= \sum_{a \in \mathcal{A}} \pi(a|s) \Big(r(s,a) + \gamma \sum_{s' \in \mathcal{S}} p(s'|s,a) V^\pi(s') \Big).
\end{aligned}$$

On the other hand we obtain for Q that

$$\begin{aligned}
Q^\pi(s,a) &= r(s,a) + \gamma \sum_{s' \in \mathcal{S}} p(s'|s,a) \mathbb{E}^\pi_{s'}[\sum_{t=0}^{\infty} \gamma^t r(S_t, A_t)] \\
&= r(s,a) + \gamma \sum_{s' \in \mathcal{S}} p(s'|s,a) V^\pi(s') \\
&= r(s,a) + \gamma \sum_{s' \in \mathcal{S}} p(s'|s,a) \sum_{a' \in \mathcal{A}} \pi(a'|s') Q^\pi(s',a').
\end{aligned}$$

■

We define the Bellman operator and the Bellman optimality operator for the state value function based on these equations:

DEFINITION 3.11. For any function $V : \mathcal{S} \to \mathbb{R}$ we define

(i) the Bellman operator $T^\pi : \mathbb{R}^{|\mathcal{S}|} \to \mathbb{R}^{|\mathcal{S}|}$ by

$$T^\pi(V)(s) = \sum_{a\in\mathcal{A}} \pi(a|s)\big(r(s,a) + \gamma \sum_{s'\in\mathcal{S}} p(s'|s,a)V(s')\big), \quad \forall s \in \mathcal{S}. \tag{3.3}$$

(ii) the Bellman optimality operator $T^* : \mathbb{R}^{|\mathcal{S}|} \to \mathbb{R}^{|\mathcal{S}|}$ by

$$T^*(V)(s) = \max_{a\in\mathcal{A}} \big\{r(s,a) + \gamma \sum_{s'\in\mathcal{S}} p(s'|s,a)V(s')\big\}, \quad \forall s \in \mathcal{S}. \tag{3.4}$$

Remark 3.12. To define the Bellman operators we interpreted the functions $V : \mathcal{S} \to \mathbb{R}$ as vectors in $\mathbb{R}^{|\mathcal{S}|}$. We will interchangeably employ both interpretations throughout the thesis, depending on convenience and context.

Both operators are γ-contractions [Put05, Thm 6.2.4], and that V^π and V^* are unique fixed points of T^π and T^* respectively [Put05, Thm 6.2.5]. We deduce that a policy π^* is optimal if and only if $V^{\pi^*} = T^*V^{\pi^*}$ [Put05, Thm. 6.2.6].
As V^* is the unique fixed point of T^*, the Banach fixed-point Theorem (see Banach [Ban22, Thm. 6]) states that we can approximate V^* by iteratively applying the operator T^*. Thus, for any function $V : |\mathcal{S}| \to \mathbb{R}$ it holds that [Put05, Thm. 6.2.3]

$$\lim_{n\to\infty} (T^*)^n(V) = V^*.$$

For completeness, we should mention that analogously Bellman operators for the Q-function can be defined. As we will not need them throughout the thesis, they will not be introduced.

Remark 3.13. As V^* is a unique fixed point of T^* we deduce the relation

$$V^*(s) = \max_{a\in\mathcal{A}} \big\{r(s,a) + \gamma \sum_{s'\in\mathcal{S}} p(s'|s,a)V^*(s')\big\}, \quad \forall s \in \mathcal{S}.$$

Further, we already know that deterministic optimal policies exists, such that for all $s \in \mathcal{S}$,

$$\begin{aligned}
V^*(s) &= \sup_\pi V^\pi(s) \\
&= \sup_{(\pi_t)_{t\geq 1}} \sup_{\pi_0} V^{(\pi_t)_{t\geq 0}}(s) \\
&= \sup_{(\pi_t)_{t\geq 1}} \sup_{\pi_0} \sum_{a\in\mathcal{A}} \pi_0(a|s) Q^{(\pi_t)_{t\geq 1}}(s,a) \\
&= \sup_{(\pi_t)_{t\geq 1}} \max_{a\in\mathcal{A}} Q^{(\pi_t)_{t\geq 1}}(s,a) \\
&= \max_{a\in\mathcal{A}} \sup_{(\pi_t)_{t\geq 1}} Q^{(\pi_t)_{t\geq 1}}(s,a) \\
&= \max_{a\in\mathcal{A}} Q^*(s,a).
\end{aligned}$$

From $V^*(s) = \max_{a\in\mathcal{A}} Q^*(s,a)$ we can deduce an optimal deterministic policy from the optimal Q-function. For every $s \in \mathcal{S}$ the optimal action is given by $a^* = \operatorname{argmax}_a Q(s,a)$[1] and we set $\pi^*(a^*|s) = 1$ to identify an optimal stationary deterministic policy.

[1]When multiple actions are equally optimal, we select an arbitrary one among them.

DEFINITION 3.14 (Greedy Policy). For $V : \mathcal{S} \to \mathbb{R}$, a greedy policy π^V chooses the arbitrary action that maximizes the Q-matrix:

$$Q^V(s,a) := r(s,a) + \gamma \sum_{s' \in \mathcal{S}} p(s'|s,a)V(s'), \quad s \in \mathcal{S},\ a \in \mathcal{A}.$$

3.1.2 *Policy Gradient*

In practice, we do not have direct access to the Bellman operator, as the transition dynamics p are unknown. Consequently, the MDP is treated as a black-box model: in a given state, an action is taken, and a reward is observed from the underlying system. A reinforcement learning algorithm seeks to learn an optimal policy using these state-action-reward samples. Algorithms trained solely on such data are referred to as model-free as they make no assumptions about the underlying system. A prominent model-free approach is the policy gradient (PG) algorithm where the policy is parametrized by π^θ, with $\theta \in \mathbb{R}^d$. The indirectly parametrized value function V^{π^θ} is maximized via (stochastic) gradient ascent.
For any differential parametrization $(\pi^\theta)_{\theta \in \mathbb{R}^d}$, i.e. the mappings $\theta \mapsto \pi^\theta(s,a)$, $\theta \in \mathbb{R}^d$ are differentiable for all $(s,a) \in \mathcal{S} \times \mathcal{A}$. We use the following notation for the objective function,

$$J_\mu : \mathbb{R}^d \to \mathbb{R}, \quad \theta \mapsto J_\mu(\theta) = V^{\pi^\theta}(\mu).$$

We treat J_μ as the objective f from Chapter 2 and aim to maximize the function with gradient ascent. All results regarding gradient descent also apply to gradient ascent by replacing the objective f with $-f$. The optimum $J_\mu^* := \sup_\theta J_\mu(\theta)$ is finite due to the finite rewards induced by the finite state and action space.
Let us for now assume that we can access the exact gradient $\nabla J_\mu(\theta)$ for all $\theta \in \mathbb{R}^d$. Then, the gradient ascent algorithm for this problem is called (deterministic) vanilla policy gradient and summarized in Algorithm 1.

Algorithm 1: Deterministic Policy Gradient

Result: Approximation $\widehat{\pi}^*$ on the optimal policy π^*.
Input: Initial state distribution μ and class of policies $(\pi_\theta)_{\theta \in \mathbb{R}^d}$.
Initialize $\theta_0 \in \mathbb{R}^d$;
Choose step size $\alpha > 0$ and set $n = 0$;
while *Convergence criterion not met* **do**
  $\theta_{n+1} = \theta_n + \alpha \nabla J_\mu(\theta)$;
  $n = n + 1$;
end
Set $\widehat{\pi}^* = \pi^{\theta_{n-1}}$;

The original motivation for PG was to develop a data-driven algorithm capable of finding the optimal policy. Thus, we need a stochastic first order oracle for the gradient and perform stochastic gradient ascent instead. The policy gradient theorem, first derived in [Sut+99], forms the foundation for constructing an effective gradient estimator.

THEOREM 3.15 (Policy Gradient Theorem [Sut+99]). *Let $(\pi^\theta)_{\theta\in\mathbb{R}^d}$ be a differential class of policies. Then, it holds that*

$$\nabla J_\mu(\theta) = \sum_{s\in\mathcal{S}} \rho_\mu^{\pi^\theta}(s) \sum_{a\in\mathcal{A}} \nabla\pi^\theta(a|s) Q^{\pi^\theta}(s,a) \tag{3.5}$$

$$= \mathbb{E}_\mu^{\pi^\theta}\Big[\sum_{t=0}^{\infty} \gamma^t \nabla \log(\pi^\theta(A_t|S_t) Q^{\pi^\theta}(S_t,A_t)\Big]. \tag{3.6}$$

Proof. The first equality is Theorem 1 in [Sut+99].
The second equality follows from the definition of the state visitation measure in Definition 3.8 and the score function trick, i.e. $\nabla\pi^\theta(a|s) = \pi^\theta(a|s)\nabla\log(\pi^\theta(a|s))$. ∎

Remark 3.16. It is noteworthy that due to this theorem, we can deduce the differentiability of the objective function $J_\mu(\theta)$ from the differentiability of underlying policy parametrization.

The following variants of the PG Theorem can be derived:

$$\nabla J_\mu(\theta) = \mathbb{E}_\mu^{\pi^\theta}\Big[\sum_{t=0}^{\infty} \gamma^t \nabla \log(\pi^\theta(A_t|S_t) \sum_{k=0}^{\infty} \gamma^k r(S_k,A_k)\Big] \tag{3.7}$$

$$= \mathbb{E}_\mu^{\pi^\theta}\Big[\sum_{t=0}^{\infty} \gamma^t \nabla \log(\pi^\theta(A_t|S_t) \sum_{k=t}^{\infty} \gamma^k r(S_k,A_k)\Big] \tag{3.8}$$

$$= \mathbb{E}_\mu^{\pi^\theta}\Big[\sum_{t=0}^{\infty} \gamma^t \nabla \log(\pi^\theta(A_t|S_t)\, A^{\pi^\theta}(S_t,A_t)\Big]. \tag{3.9}$$

The typical approach to sample the gradient is to estimate one of the above expectations via a Monte Carlo estimator. The simplest estimator is the REINFORCE estimator introduced in [Wil92], where the infinite sum is truncated at a deterministic time: Let $T\in\mathbb{N}$ and $(s_0,a_0,s_1,\dots,s_T,A_T)$ be a trajectory of length T sampled from the MDP under the measure $\mathbb{P}_\mu^{\pi^\theta}$. Then the REINFORCE estimator for the gradient, $\nabla J_\mu(\theta)$, is given by

$$G_T^{\text{REINFORCE}} = \sum_{t=0}^{T} \gamma^t \nabla \log(\pi^\theta(a_t|s_t) \sum_{k=0}^{T} \gamma^k r(s_k,a_k), \quad T\in\mathbb{N}. \tag{3.10}$$

As the infinite sums in equation (3.7) are truncated by $T\in\mathbb{N}$, this estimator is biased and cannot fulfill the first order oracle conditions discussed in Section 2.2.1. Due to the bias, convergence towards the global optimum using this estimator in PG cannot be guaranteed.
The authors in [Zha+20] introduced a simple trick where independent geometric random variables are used to derive an unbiased estimator of the gradient. The trick is based on the following observation.

Remark 3.17. A discounted MDP can be seen as an undiscounted MDP stopped at an independent geometric random variable with mean $(1-\gamma)^{-1}$. It follows that for $T\sim Geom(1-\gamma)$ independent of the MDP it holds that

$$\rho_\mu^\pi(s) = \sum_{t=0}^{\infty} \gamma^t \mathbb{P}_\mu^\pi(S_t=s) = \sum_{t=0}^{\infty} \gamma^t \mathbb{E}_\mu^\pi[\mathbf{1}_{S_t=s}] = \sum_{t=0}^{\infty} \mathbb{P}(t\le T)\mathbb{E}_\mu^\pi[\mathbf{1}_{S_t=s}]$$

$$= \mathbb{E}_\mu^\pi\Big[\sum_{t=0}^{\infty} \mathbf{1}_{t\le T}\mathbf{1}_{S_t=s}\Big] = \mathbb{E}_\mu^\pi\Big[\sum_{t=0}^{T} \mathbf{1}_{S_t=s}\Big].$$

We deduce from the same trick that $V^{\pi^\theta}(\mu) = \mathbb{E}_\mu^{\pi^\theta}\big[\sum_{t=0}^{T}\gamma^t r(S_t, A_t)\big]$.
Therefore, we call $\mathbb{E}[T] = \frac{1}{1-\gamma}$ the effective or expected horizon of a discounted MDP.

In order to derive an unbiased estimator for the gradient, the trick can be applied twice in equation (3.8). Let $T \sim Geom(1-\gamma)$ and $T' \sim Geom(1-\gamma^{\frac{1}{2}})$ independent of each other and of the MDP, then a trajectory of the MDP with length $T+T'$, $(s_0, a_0, s_1, \dots, s_{T+T'}, A_{T+T'})$, can be used to define an unbiased estimator [Zha+20, Thm. 3.4.]

$$G^{\text{REINFORCE-UB}} = \frac{1}{1-\gamma}\nabla \log(\pi^\theta(a_T|s_T)\sum_{t=T}^{T+T'} \gamma^{(t-T)/2} r(s_t, a_t). \tag{3.11}$$

This gradient estimator is unbiased and aligns with the framework introduced in Section 2.2.1. However, the (ABC) condition in Assumption 2.12 does not universally hold for all parametrizations and the variance of the estimator must be controlled on a case-by-case basis. See for example [Zha+20, Ass. 3.1 (ii)] for sufficient conditions on the parametrization class which ensure almost surely bounded (estimated) gradients [Zha+20, Thm. 3.4]. To summarize we state the stochastic policy gradient (SPG) in Algorithm 2.

Algorithm 2: Stochastic Policy Gradient

Result: Approximation $\widehat{\pi}^*$ on the optimal policy π^*.
Input: Initial state distribution μ and class of policies $(\pi^\theta)_{\theta\in\mathbb{R}^d}$.
Initialize $\theta_0 \in \mathbb{R}^d$;
Choose step size $\alpha > 0$ and set $n = 0$;
while *Convergence criterion not met* **do**
  Sample the gradient G_n as in equation (3.10) or equation (3.11) under policy π^{θ_n};
  $\theta_{n+1} = \theta_n + \alpha\,\nabla G_n$;
  $n = n + 1$;
end
Set $\widehat{\pi}^* = \pi^{\theta_{n-1}}$;

Tabular Softmax Policy. A policy for which we can guarantee both the (ABC) condition (even with A=B=0) and a gradient domination property is the tabular softmax policy. We introduce the logit function $\theta : \mathcal{S}\times\mathcal{A} \to \mathbb{R}$ and the softmax policy parametrized by $\theta \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ as

$$\pi^\theta(a|s) = \frac{\exp(\theta(s,a))}{\sum_{a'\in\mathcal{A}}\exp(\theta(s,a'))}, \quad s\in\mathcal{S}, a\in\mathcal{A}. \tag{3.12}$$

The tabular softmax policy can approximate any deterministic policy arbitrarily close and is therefore suitable to converge to the optimal deterministic policy, whereas other parametrizations such as neural networks may induce an approximation error. This error needs to be considered in the convergence analysis for different parametrizations.

Due to the policy gradient theorem we compute the derivative of the log-softmax policy for every $s \in \mathcal{S}$ and $a \in \mathcal{A}$, with parameter $\theta \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$,

$$\frac{\partial \log(\pi^\theta(a|s))}{\partial \theta(a', s')} = \mathbf{1}_{\{s=s'\}}(\mathbf{1}_{\{a=a'\}} - \pi^\theta(a'|s')),$$

and obtain the score-function,

$$\nabla \log(\pi^\theta(a|s)) = \left(\mathbf{1}_{\{s=s'\}}(\mathbf{1}_{\{a=a'\}} - \pi^\theta(a'|s'))\right)_{s' \in \mathcal{S}, a' \in \mathcal{A}} \quad \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}. \tag{3.13}$$

Using equation (3.9) we arrive at the following form of the gradient.

LEMMA 3.18. *[Aga+21, Lem C.1] The gradient of the objective under the softmax policy parametrization has the form*

$$\frac{\partial J_\mu(\theta)}{\partial \theta(s,a)} = \rho_\mu^{\pi^\theta}(s)\pi^\theta(s,a)A^{\pi^\theta}(s,a).$$

In Agarwal et al. [Aga+21], the global asymptotic convergence of PG is demonstrated under tabular softmax parametrization, and convergence rates are derived using log-barrier regularisation and natural policy gradient. Building upon this work, Mei et al. [Mei+20] showed the first convergence rates for PG under softmax parametrization. The authors exploited that the smoothness property is globally fulfilled and a weak gradient domination property holds along the gradient ascent trajectory.

LEMMA 3.19.

(i) [YGL22, Lem. E.1] The objective function $J_\mu(\theta)$ is L-smooth under the tabular softmax parametrization with $L = \frac{R^*}{(1-\gamma)^2}\left(2 - \frac{1}{|\mathcal{A}|}\right)$.

(ii) [Mei+20, Lem. 8] Assume that $\mu(s) > 0$ for every $s \in \mathcal{S}$. Under the tabular softmax parametrization it holds that

$$\|\nabla J_\mu(\theta)\| \geq \frac{\min_{s\in\mathcal{S}} \pi^\theta(a^*(s)|s)}{\sqrt{|\mathcal{S}|}(1-\gamma)} \left\|\frac{d_\mu^{\pi^*}}{\mu}\right\|_\infty^{-1} [J_\mu^* - J_\mu(\theta),]$$

where π^* is a fixed deterministic stationary optimal policy and $a^*(s)$ is the action under π^* in state s.

(iii) [Mei+20, Lem. 9] Using the gradient ascent scheme $\theta_{n+1} = \theta_n + \alpha \nabla J_\mu(\theta)|_{\theta=\theta_n}$ with arbitrary $\theta_1 \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ and tabular softmax parametrization. It holds that

$$c_\gamma := \inf_{n\geq 1} \min_{s\in\mathcal{S}} \pi^{\theta_n}(a^*(s)|s) > 0.$$

Thus, $\|\nabla J_\mu(\theta)\| \geq \frac{c_\gamma}{\sqrt{|\mathcal{S}|}(1-\gamma)} \left\|\frac{d_\mu^{\pi^*}}{\mu}\right\|_\infty^{-1} [J_\mu^* - J_\mu(\theta).$

Remark 3.20. In Lemma 3.19 (ii) a so-called non uniform gradient domination property ([Mei+21]) is shown for tabular softmax parametrization, where the factor $\min_{s\in\mathcal{S}} \pi^\theta(a^*(s)|s)$ depends on the current parameter θ. The term $\left\|\frac{d_\mu^{\pi^*}}{\mu}\right\|_\infty$ denotes the so-called distribution mismatch coefficient introduced in [Aga+21]. In (iii), it is verified, that along the trajectory of gradient ascent the function $\theta \mapsto \min_{s\in\mathcal{S}} \pi^\theta(a^*(s)|s)$ stays strictly positive. This gives us a weak gradient domination property with $\beta = 1$ as exploited in Theorem 2.11. We can now use this result to obtain convergence.

THEOREM 3.21. *Assume that $\mu(s) > 0$ for every $s \in \mathcal{S}$. Let $\theta_{n+1} = \theta_n + \alpha \nabla J_\mu(\theta)|_{\theta=\theta_n}$ be the gradient ascent scheme with arbitrary $\theta_1 \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ and choose $\alpha = \frac{(1-\gamma)^2}{R^*(2-\frac{1}{|\mathcal{A}|})}$. Then,*

$$J_\mu^* - J_\mu(\theta_n) \geq \frac{16R^*(2-\frac{1}{|\mathcal{A}|})|\mathcal{S}|}{c_\gamma^2(1-\gamma)^4 n}\left\|\frac{d_\mu^{\pi^*}}{\mu}\right\|_\infty^2.$$

Proof. Proof is given by the application of Theorem 2.11 or can be obtained using the same proof as in [Mei+20, Thm. 4], but with tighter smoothness constant from [YGL22, Lem. E.1]. ∎

Remark 3.22. It is important to highlight that the constant c appearing in the convergence rate corresponds to the one from Lemma 3.19 (iii). This constant is typically unknown and may depend on γ or other model parameters, making the interpretation of the convergence rate challenging. In Chapter 5, we will explore the potential issues this presents and discuss a modification of the PG algorithm to derive tighter upper bounds on the convergence rate.

So, SPG can be applied in the tabular setting where each state-action pair is considered separately. Although the tabular setting is not used in practical applications, it is the most tractable setting for a complete mathematical analysis and sheds light on general principles. We will mainly deal with this parametrization throughout the thesis to ensure and compare convergence behaviour of different algorithms.

Convergence of PG and related Literature. Due to their high flexibility and model-free nature PG methods enjoy a great popularity in practice. But from the optimization perspective it is natural to ask for convergence guarantees of the algorithm. Despite the far-reaching history of PG [Wil92; Sut+99; KT99; Kak01], there were no proofs for the global convergence of these algorithms for a long time. Nevertheless, they have been very successful in many applications, which is why numerous variants have been developed in the last few decades, whose convergence analysis, if available, was mostly limited to convergence to stationary points [PRB13; Sch+15b; Pap+18; Cla+18; She+19; XGG20a; Hua+20; XGG20b; HGH22]. In finite state and action MDPs the smoothness property (Assumption 2.1) can often be easily justified due to bounded rewards, such that convergence (almost surely and in expectation) towards stationary points is directly implied. For convergence towards global optima, properties like convexity or gradient domination are required as discussed in Section 2.2.3. We have seen in the previous paragraph that such properties heavily depend on the choice of parametrization. Especially in deep learning scenarios, where neural networks are used to parameterize the policy, the objective function $J_\mu(\theta)$ is highly non-convex and does not fulfills such properties globally. Nevertheless, for specific parametrizations a (weak) gradient domination property can be derived:
First, the authors in [Faz+18] exploited gradient domination in the case of linear quadratic

regulator problems to derive convergence rates of (stochastic) PG methods in this specific reward setting. Notably, [Aga+21] were the first to show global convergence of PG in the general MDP setting under softmax parametrization but without relying on gradient domination. Their result is therefore without a rate. Further, the authors analyze the log-barrier regularized softmax PG with a gradient domination argument and for natural policy gradient (NPG) a convergence rate is derived even in the stochastic setting. Next to the vanilla PG method discussed in this section [Mei+20] also considered the entropy-regularized PG method under softmax parametrization. Under regularization the stronger PL-condition can be verified which leads to a linear convergence under exact gradients.
A growing community deals with the convergence of variants of PG methods like policy mirror descent (PMD) or natural policy gradient (NPG), where convergence is also partly due to gradient domination properties. Under exact gradients we refer the interested reader to [BR21; BR22; Cen+22] and in the stochastic case to [DZL22; Xia22; AR23; YGL22; Fat+23; JPBR23]. We want to point out, that all stochastic results consider regularized PG or variants like PMD or NPG and not much in known about vanilla SPG. This is partly due to the fact that regularization leads to *nicer* optimization problems which satisfy stronger regularity properties like the PL-condition under entropy regularization in [Mei+20]. The analysis of NPG or PMD differs from the gradient domination setting used for SGD methods and usually requires additional assumptions. See for example [Aga+21, Ass. 6.1], [Xia22, Ass. 11] and [JPBR23, Ass. 4.1].
In this thesis we mainly focus on deriving convergence rates for vanilla (softmax) PG in the stochastic and exact gradient setting without the need of regularization or additional assumptions.

3.1.3 *Other RL Algorithms*

Finally, we want to mention for completeness that there are multiple other RL algorithms besides the PG Method. We have seen from the dynamic programming principle, that it suffice to learn an optimal Q-function and that we can deduce an optimal policy from Q^* (Remark 3.13). Other very popular methods besides PG are Q-Learning, Temporal Difference Learning or Actor-Critic methods where Q-Learning and PG is combined.

3.2 Finite-time horizon MDPs

As the name suggests finite-time MDPs are MDPs over a deterministic finite-time horizon. Therefore, discounting is no longer necessary (but still possible) to assure a well-defined problem. In addition, we want to allow for a more general context, where the state space can change over time and the action space can depend on the state. Formally, we define the following.

Definition 3.23 (Finite-time MDP). The sextuplet $(\mathcal{H}, \mathcal{S}, \mathcal{A}, \gamma, p, r)$ given by

- the time index set $\mathcal{H} = \{0, \dots, H-1\} \subset \mathbb{N}$
- the finite state spaces $\mathcal{S} = \mathcal{S}_0 \cup \cdots \cup \mathcal{S}_{H-1}$,
- the finite action space $\mathcal{A} = \bigcup_{s \in \mathcal{S}} \mathcal{A}_s$,
- the discount factor $\gamma \in [0, 1]$,

- the transition function $p : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$, s.t. $p(\mathcal{S}_{h+1}|s,a) = 1$, for every $h < H-1$, $s \in \mathcal{S}_h$, $a \in \mathcal{A}_s$.
- the reward function $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$

is called a finite-time MDP. When $\gamma = 1$, we just write $(\mathcal{H}, \mathcal{S}, \mathcal{A}, p, r)$ for an undiscounted finite-time MDP.

First, note that the state spaces $\mathcal{S}_0, \dots, \mathcal{S}_{H-1}$ can and often does (partly) coincide and we recover the stationary state space scenario from discounted MDPs when $\mathcal{S}_h = \mathcal{S}$ for all $h \in \mathcal{H}$. Moreover, the transition function p is time-independent by definition given the domain $\mathcal{S} \times \mathcal{A}$, where $\mathcal{S}$ is the set of all possible states in the MDP. Hence, if state s is in $\mathcal{S}_{t_1}$ and in $\mathcal{S}_{t_2}$, then the transition probabilities under action a are the same for both time points.

Remark 3.24. The discount factor in finite-time MDPs is not necessary for a well-defined problem and can be incorporated into a time-dependent reward function. It should be noted that all results derived in this thesis for finite-time MDPs can be straightforwardly generalized to a time-dependent reward function, allowing discounting to be optional rather than required.

DEFINITION 3.25. [Finite-time Policy] Let $(\mathcal{H}, \mathcal{S}, \mathcal{A}, \gamma, p, r)$ be a finite-time MDP.

(i) A policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$ is a mapping from state $s \in \mathcal{S}$ to a distribution over the possible action space, i.e. $\pi(\cdot|s) \in \Delta(\mathcal{A}_s)$. The set of all policies is still denoted by Π.

(ii) A finite sequence of policies is denoted by $\mathbb{\pi}_H = (\pi_h)_{h=0}^{H-1} \in \Pi^H$, where $\pi_h : \mathcal{S}_h \to \Delta(\mathcal{A})$ is the policy in decision epoch $h \in \mathcal{H}$ and $\pi_h(\mathcal{A}_s|s) = 1$ for every $s \in \mathcal{S}_h$.

(iii) We denote by $\mathbb{\pi}_{(h)} := (\pi_t)_{t=h}^{H-1} \in \Pi^{H-h}$ sub-sequences from h to $H-1$ of $\mathbb{\pi}_H$.

It is well-known that in contrast to discounted infinite-time horizon MDPs non-stationary policies are needed to optimize finite-time MDPs. An optimal policy in time point h depends on the time horizon until the end of the problem (see for example Puterman [Put05]).
Given the pre-specified time horizon we definite the value function and analog the state-action value function as the (discounted) sum of expected rewards up to time $H-1$.

DEFINITION 3.26 (Finite-time value and state-action value function, Advantage function). Suppose a finite-time MDP $(\mathcal{H}, \mathcal{S}, \mathcal{A}, \gamma, p, r)$ and let $\mathbb{\pi}_H = (\pi_h)_{h=0}^{H-1} \in \Pi^H$ be a policy.

(i) The epoch-dependent value functions over the time horizon $\{h, \dots, H-1\}$ using the policy $\mathbb{\pi}_{(h)}$ are defined by

$$V_h^{\mathbb{\pi}_{(h)}}(s) = \mathbb{E}_{S_h=s}^{\mathbb{\pi}_{(h)}}\Big[\sum_{t=h}^{H-1} \gamma^t r(S_t, A_t)\Big], \quad \forall s \in \mathcal{S}_h. \tag{3.14}$$

We define $V^{\mathbb{\pi}_H} \equiv V_0^{\mathbb{\pi}_H}$ as the value function of the finite-time MDP under policy $\mathbb{\pi}_H = (\pi_h)_{h=0}^{H-1} \in \Pi^H$.

(ii) For any initial distribution μ_h over $\mathcal{S}_h$ we write $V_h^{\mathbb{\pi}_{(h)}}(\mu_h) = \mathbb{E}_{S_h \sim \mu_h}[V_h^{\mathbb{\pi}_{(h)}}(S_h)]$.

(iii) The h-state-action value function is defined by

$$Q_h^{\mathbb{m}_{(h+1)}}(s,a) = \mathbb{E}^{\mathbb{m}_{(h+1)}}_{S_h=s,A_h=a}\Big[\sum_{t=h}^{H-1}\gamma^t r(S_t,A_t)\Big], \quad \forall s\in\mathcal{S}_h, a\in\mathcal{A}_s. \tag{3.15}$$

(iv) The h-state-action advantage function is defined by

$$A_h^{\mathbb{m}_{(h)}}(s,a) := Q_h^{\mathbb{m}_{(h+1)}}(s,a) - V_h^{\mathbb{m}_{(h)}}(s), \quad s\in\mathcal{S}_h, a\in\mathcal{A}_s. \tag{3.16}$$

Remark 3.27. Note that Q_h is independent of policy π_h as the action in the initial time h is already determined by the input of the function. For $H-1$, this leads to $Q_{H-1}(s,a) = \gamma^{H-1}r(s,a)$ deterministic and independent of any policy.

Definition 3.28 (Optimal value, epoch-depended value and state-action value function).

(i) The optimal value function is defined by $V^*(s) = \sup_{\mathbb{m}_H\in\Pi^H} V_0^{\mathbb{m}_H}(s)$ for all $s\in\mathcal{S}_0$.

(ii) The optimal epoch dependent value functions are defined by $V_h^*(s) = \sup_{\mathbb{m}_{(h)}\in\Pi^{H-h}} V_h^{\mathbb{m}_{(h)}}(s)$ for all $s\in\mathcal{S}_h$.

(iii) The optimal state-action value function is defined by $Q^*(s) = \sup_{\mathbb{m}_H\in\Pi^H} V_0^{\mathbb{m}_H}(s)$ for all $s\in\mathcal{S}_0$.

Definition 3.29 (Optimal Policy). Policies $\mathbb{m}_H\in\Pi^H$ which satisfy $V_0^{\mathbb{m}_H}(s) = V^*(s)$ for all $s\in\mathcal{S}_0$ are called optimal policies and denoted by $\mathbb{m}_H^*$.

Remark 3.30. The optimal policies determined by the optimal value functions are consistent such that $\mathbb{m}_H^*$ restricted to the last $H-h$ time-points, i.e. $\mathbb{m}^*_{(h)}$, is an optimal policy for V_h. This can be deduced from the Markovian structure of the MDP and the resulting backward inductive solution which will be discussed in the following subsection.

In the remainder of this thesis we will drop the subscript in $\mathbb{m}_H$ or $\mathbb{m}_{(h)}$, when the horizon is clear from the indices in V_h, Q_h and A_h.

Definition 3.31.

(i) For an initial state distribution μ on $\mathcal{S}_0$ the state visitation measure under policy $\mathbb{m}\in\Pi^H$ is defined by

$$\rho_\mu^{\mathbb{m}}(s) = \sum_{h=0}^{H-1}\gamma^h\mathbb{P}_\mu^{\mathbb{m}}(S_h=s).$$

(ii) If $\gamma\in[0,1)$, then $d_\mu^{\mathbb{m}}(s) = \frac{1-\gamma}{1-\gamma^H}\rho_\mu^{\mathbb{m}}(s)$ is the normalized state-visitation distribution and when $\gamma=1$ (no discounting), then $d_\mu^{\mathbb{m}}(s) = \frac{1}{H}\rho_\mu^{\mathbb{m}}(s)$.

For finite-time MDPs we obtain the following version of the performance difference lemma.

Lemma 3.32. *[KWD24, Lem. A.3] For any $h\in\mathcal{H}$ and for any pair of policies $\mathbb{m}$ and $\mathbb{m}'\in\Pi^H$ the following holds true for every $s\in\mathcal{S}_h$:*

$$V_h^{\mathbb{m}}(s) - V_h^{\mathbb{m}'}(s) = \sum_{k=h}^{H-1}\mathbb{E}^{\mathbb{m}}_{S_h=s}\Big[A_k^{\mathbb{m}'}(S_k,A_k)\Big].$$

Proof. We derive

$$\begin{aligned}
V_h^{\pi}(s) - V_h^{\pi'}(s) &= \mathbb{E}_{S_h=s}^{\pi(h)}\Big[\sum_{k=h}^{H-1}\gamma^k r(S_k,A_k)\Big] - V_h^{\pi'}(s)\\
&= \mathbb{E}_{S_h=s}^{\pi(h)}\Big[\sum_{k=h}^{H-1}\gamma^k r(S_k,A_k) + \sum_{k=h}^{H-1}V_k^{\pi'}(S_k) - \sum_{k=h}^{H-1}V_k^{\pi'}(S_k)\Big] - V_h^{\pi'}(s)\\
&= \mathbb{E}_{S_h=s}^{\pi(h)}\Big[\sum_{k=h}^{H-1}\gamma^k r(S_k,A_k) + \sum_{k=h+1}^{H-1}V_k^{\pi'}(S_k) - \sum_{k=h}^{H-1}V_k^{\pi'}(S_k)\Big]\\
&= \mathbb{E}_{S_h=s}^{\pi(h)}\Big[\sum_{k=h}^{H-1}\gamma^k r(S_k,A_k) + \sum_{k=h}^{H-2}V_{k+1}^{\pi'}(S_{k+1}) - \sum_{k=h}^{H-1}V_k^{\pi'}(S_k)\Big]\\
&= \mathbb{E}_{S_h=s}^{\pi(h)}\Big[\sum_{k=h}^{H-1}\gamma^k r(S_k,A_k) + \sum_{k=h}^{H-2}V_k^{\pi'}(S_{k+1}) - \sum_{k=h}^{H-1}V_k^{\pi'}(S_k)\Big]\\
&= \mathbb{E}_{S_h=s}^{\pi(h)}\Big[\sum_{k=h}^{H-1}\big(\gamma^k r(S_k,A_k) + V_{k+1}^{\pi'}(S_{k+1}) - V_k^{\pi'}(S_k)\big)\Big]\\
&= \mathbb{E}_{S_h=s}^{\pi(h)}\Big[\sum_{k=h}^{H-1}A_k^{\pi'}(S_k,A_k)\Big]\\
&= \sum_{k=h}^{H-1}\mathbb{E}_{S_h=s}^{\pi(h)}\Big[A_k^{\pi'}(S_k,A_k)\Big],
\end{aligned}$$

where we have used that $\gamma^k r(S_k,A_k) + V_{k+1}^{\pi'}(S_{k+1}) = Q_k^{\pi'}(S_k,A_k)$. In the fifth equation we used the notation $V_H \equiv 0$ and note that $Q_{H-1} \equiv \gamma^{H-1} r$ independent of any policy. ∎

3.2.1 *Finite-time Dynamic Programming*

In contrast to discounted MDPs we can no longer derive fixed-point equations to optimally solve the finite-time MDPs under perfect information. Instead, the optimal solution is derived by backward induction over the finite-time horizon. In fact, this leverages from the following relation between the value and state-action value functions.

PROPOSITION 3.33. *Let $\pi \in \Pi^H$, then for any $h \le H-1$ and $s \in \mathcal{S}_h$ it holds that*

$$\begin{aligned}
V_h^{\pi(h)}(s) &= \sum_{a\in\mathcal{A}_s}\pi_h(a|s)Q_h^{\pi(h+1)}\\
&= \sum_{a\in\mathcal{A}_s}\pi_h(a|s)\big(\gamma^h r(s,a) + \sum_{s'\in\mathcal{S}}p(s'|s,a)V_{h+1}^{\pi(h+1)}(s')\big),
\end{aligned}$$

and for any $h \le H-1$, $s \in \mathcal{S}_h$ and $a \in \mathcal{A}_s$ it holds that

$$\begin{aligned}
Q_h^{\pi(h+1)}(s,a) &= \gamma^h r(s,a) + \sum_{s'\in\mathcal{S}}p(s'|s,a)V_{h+1}^{\pi(h+1)}(s')\\
&= \gamma^h r(s,a) + \sum_{s'\in\mathcal{S},a'\in\mathcal{A}_{s'}}p(s'|s,a)\pi_{h+1}(a'|s')Q_{h+1}^{\pi(h+2)}(s',a'),
\end{aligned}$$

where we introduce the convention $V_H \equiv 0$.

Proof. By definition, similar to the discounted case. ∎

First, it is important to notice that, due to the time dependence in V and Q, the equations in Proposition 3.33 are not fixed point equations. Instead they induce a backward inductive scheme, where we can derive $V_0^{\mathbb{\pi}}$ by starting with $V_H \equiv 0$ and iteratively apply the operators $T_h^{\pi_h}$ from $h = H-1, \dots, 0$ backwards in time. The operators are defined by

$$T_h^{\pi_h}(V_{h+1})(s) := \sum_{a \in \mathcal{A}_s} \pi_h(a|s)(\gamma^h r(s,a) + \sum_{s' \in \mathcal{S}} p(s'|s,a)V_{h+1}(s')), \quad \forall s \in \mathcal{S}_h. \tag{3.17}$$

We can deduce the following optimality relations from Proposition 3.33,

$$V_h^*(s) = \max_{a \in \mathcal{A}_s} Q_h^*(s,a), \quad \forall h \leq H-1,\ s \in \mathcal{S}_h$$

and

$$Q_h^*(s,a) = \gamma^h r(s,a) + \sum_{s' \in \mathcal{S}_{h+1}} V_{h+1}^*(s'), \quad \forall h \leq H-1, s \in \mathcal{S}_h, a \in \mathcal{A}_s.$$

So for finite-time MDPs we have the following optimality operators

$$T_h^*(V_{h+1})(s) := \max_{a \in \mathcal{A}_s}(\gamma^h r(s,a) + \sum_{s' \in \mathcal{S}} p(s'|s,a)V_{h+1}(s')), \quad \forall s \in \mathcal{S}_h. \tag{3.18}$$

Remark 3.34. From the relation $V_h^*(s) = \max_{a \in \mathcal{A}} Q_h^*(s,a)$, we can derive an optimal deterministic policy using the optimal Q-functions. Specifically, for every $s \in \mathcal{S}_h$ the optimal action is determined by $a_h^*(s) = \text{argmax}_a Q_h(s,a)$. Once this optimal action is identified, we can define a non-stationary deterministic optimal policy by $\pi_h^*(a_h^*|s) = 1$.
However, compared to the case of discounted MDPs, the policy π_h is not necessarily optimal at step h if $T_h^{\pi_h}(V_{h+1}^{\mathbb{\pi}_{(h+1)}}) = T_h^*(V_{h+1}^{\mathbb{\pi}_{(h+1)}})$. This condition holds only if the policy sequence $\mathbb{\pi}_{(h+1)}$ contains the optimal policies from step $h+1$ to the final time step $H-1$. In other words, π_h is an optimal policy for the current epoch h, if and only if $T_h^{\pi_h}(V_{h+1}^*) = T_h^*(V_{h+1}^*)$. This implies that the policy π_h must not only maximize the immediate reward but also depends on the decisions in the subsequent epochs. On the other hand, it is also straight forward to see and important to notice that the optimal policy of epoch h does not depend on the past, i.e. on decision which lead up to time point h. Thus, π_h^* is independent of π_l^* for $h > l$ but not vice versa.

With the convention $V_H \equiv 0$, we can solve for the optimal policy by using the dynamic programming algorithm in Algorithm 3. For more details on how to use backwards induction in learning algorithms we refer for instance to Bertsekas and Tsitsiklis [BT96b, Sec. 6.5].

3.2.2 *Finite-time Policy Gradient*

How to perform policy gradient in finite-time MDP is the first major questions in this thesis and will be discussed in detail in Chapter 4. We use this section to introduce the state-of-the-art algorithm mainly considered in practice (and theory) and provide references for further reading. Finite-time MDPs differ from discounted infinite-time MDPs in that the optimal policies are not stationary, i.e. optimal actions depend on the epochs. This requires a time-dependent

Algorithm 3: Dynamic Programming for finite-time MDPs

Result: Non-stationary optimal policy $\mathbb{\pi}_H^*$.
Set $V_H \equiv 0$;
for $h = H-1, \dots, 0$ **do**
  for $s \in \mathcal{S}_h$ **do**
    for $a \in \mathcal{A}_s$ **do**
      $Q_h(s,a) = r(s,a) + \gamma \sum_{s' \in \mathcal{S}} V_{h+1}^*(s')$;
    end
    $V_h(s) = \max_{a \in \mathcal{A}} Q_h(s,a)$;
  end
  Set $\pi_h^*(\cdot|s) = \mathbf{1}_{\operatorname{argmax}_a Q_h(s,a)}$;
end
Return $\mathbb{\pi}_H^* = (\pi_h^*)_{h=0}^{H-1}$;

parametrization of the policy when considering the policy gradient method. In fact, one can also consider PG algorithms that train stationary policies for finite-time MDPs. However, this violates the intrinsic nature of finite-time MDPs as optimal policies will only be stationary in trivial cases. One way to introduce time dependence to the parametrized policy class is to enlarge the state space.

DEFINITION 3.35 (Enlarged state-space). In comparison to $\mathcal{S}$, we define the enlarged state space, $\mathcal{S}^{[\mathcal{H}]}$, which encompasses all possible states across all epochs. Therefore initially, states are associated with their respective epochs, resulting in disjoint state spaces between epochs, which are subsequently fused into a single comprehensive state space $\mathcal{S}^{[\mathcal{H}]}$. Formally, this means that for every state space $\mathcal{S}_h = \{s^1, \dots, s^{L_h}\}$ one constructs disjoint sets $\mathcal{D}_h = \mathcal{S}_h \times \{h\} = \{s_h^1, \dots, s_h^{L_h}\}$ for $h = 0, \dots, H-1$. Then, $\mathcal{S}^{[\mathcal{H}]} := \mathcal{D}_0 \uplus \cdots \uplus \mathcal{D}_{H-1}$ contains all possible states associated with their epoch.

Now, we can consider artificially stationary policies on the enlarged state space, $\pi : \mathcal{S}^{[\mathcal{H}]} \to \Delta(\mathcal{A})$, which can decide epoch dependent in states due to the time horizon added to the state space. We write $\pi \in \Pi_{\mathcal{S}^{[\mathcal{H}]}}$ to denote stationary policies on the enlarged state space.

Remark 3.36. Any policy $\mathbb{\pi} = (\pi_h)_{h=0}^{H-1}$ can be reinterpreted an artificial stationary policy $\pi \in \Pi_{\mathcal{S}^{[\mathcal{H}]}}$ by

$$\pi(a|s_h) = \sum_{h=0}^{H-1} \mathbf{1}_{s_h \in \mathcal{D}_h} \pi_h(a|s_h).$$

Considering parametrized policies, $\pi_h = \pi^{\theta_h}$, leads to artificial stationary parametrized policies π^Θ with $\Theta = [\theta_0, \dots, \theta_{H-1}]^T$. But also other arbitrary parametrizations on the enlarged state space like neural networks can be considered. Seeing finite-time MDPs as artificial stationary MDPs leads to a training procedure in which parameters for all epochs are trained simultaneously, see for instance Guin and Bhatnagar [GB23]. Therefore, we call this approach the finite-time simultaneous PG (FT-SimPG) algorithm. Due to artificial stationary, the same PG algorithm as for discounted MDPs can be used and the objective function for a parametrized class of policies

$(\pi^\Theta)_{\Theta\in\mathbb{R}^d}$, with $\pi^\Theta \in \Pi_{\mathcal{S}[\mathcal{H}]}$, is given by

$$J(\cdot,\mu) : \mathbb{R}^d \to \mathbb{R}, \quad \Theta \mapsto J(\Theta,\mu) = V_0^{\pi^\Theta}(\mu).$$

We give a more detailed description of FT-SimPG including a policy gradient theorem for sample-based implementation and discuss convergence bounds for exact and inexact gradients under softmax parametrization in Chapter 4. While there are practical reasons to train all parameters jointly, we will also see that ignoring the structure of the problem yields worse convergence bounds.

Convergence and Literature for finite-time horizon PG. There are some recent articles studying PG of finite-time horizon MDPs. E.g. in [GHZ22] complexity bounds for finite-time MDPs are derived by considering a fictitious discounted algorithm and the stationary policy from the fictitious discounted problem is used as solution to the finite-time MDP. As non-stationary policies are required to obtain optimal solutions, this approach is biased. In [HXY21; HXY23] finite-time linear quadratic control problems are solved with (natural) policy gradient. A non-stationary policy is trained simultaneously, i.e. the policies in different time epochs are trained jointly. In comparison to these works, in [Zha+23; ZHB23; ZB23] the optimal control for finite-time linear quadratic regulator problems is derived in a backward inductive manner. The dynamic policy gradient algorithm, which we introduce in the following chapter, is based on a similar idea.

Policy Gradient for finite-time MDPs

4

Within this chapter we discuss how to solve finite-time MDPs by policy gradient. Firstly, we analyse the finite-time simultaneous PG (FT-SimPG) algorithm which was already partly introduced in Section 3.2.2. In FT-SimPG all parameters on the enlarged state space are trained jointly. This approach is usually considered in practical applications when a non stationary policy is searched. Secondly, we introduce finite-time dynamic policy gradient (FT-DynPG), a new approach where the dynamic programming structure is exploited. We view the MDP as a nested sequence of contextual bandits. Essentially, FT-DynPG performs a sequence of PG algorithms backwards in time with carefully chosen epoch dependent training steps. The algorithm can bee seen as a concrete policy search by dynamic programming (PSDP) algorithm, where policy gradient is used to solve the one-step MDP [Bag+03; Sch14]. We focus on theoretical convergence guarantees and compare both algorithms in the exact and stochastic gradient case.

In the exact gradient case in Section 4.2, the analysis goes along the gradient domination arguments for discounted MDPs discussed in Section 3.1.2. For FT-SimPG, we extend the results in [Aga+21; Mei+20] to non-stationary finite-time MDPs and for FT-DynPG we combine this analysis with the backward inductive dynamic programming approach. The unspecified dependence on the effective horizon $(1-\gamma)^{-1}$ in infinite-time horizon MDPs (cf. Section 3.1.2) transfers for finite-time MDPs in a dependence on the deterministic time horizon H (compare to Remark 3.17). Through careful analysis, we establish upper bounds involving H^5 for FT-SimPG, contrasting with H^3 for FT-DynPG. Essentially, FT-DynPG offers a clear advantage. Examining the PG theorem for finite-time MDPs reveals that early epochs should be trained less if policies for later epochs are still sub-optimal. A badly learned Q-function-to-go leads to badly directed gradients in early epochs. Thus, simultaneous training yields ineffective early epoch training, addressed by our dynamic algorithm, optimizing policies backward in time with more training steps. To illustrate this phenomenon we implemented a simple toy example in Section 4.3 where the advantage of FT-DynPG becomes visible.

In the stochastic analysis in Section 4.4, we abandon the assumption that the exact gradient is known and focus on the model-free stochastic PG method. For vanilla PG very little is known about convergence to global optima even in the discounted case (cf. final paragraph in Section 3.1.2). The authors in [DZL22] derive complexity bounds for entropy-regularized stochastic softmax PG. They use a well-chosen stopping time which measures the distance to the set of optimal parameters, and simultaneously guarantees convergence to the regularized optimum prior to the occurrence of the stopping time by using a small enough step size and large enough batch size. As we are interested in convergence to the unregularized optimum, we consider stochastic softmax PG without regularization. Similar to the previous idea, we construct a different stopping time, which allows us to derive complexity bounds for an approximation arbitrarily close to the global optimum that does not require a set of optimal parameters. This is relevant when considering softmax parametrization.

4.1 SIMULTANEOUS AND DYNAMIC POLICY GRADIENT

Throughout the chapter we assume a finite-time MDP $(\mathcal{H}, \mathcal{S}, \mathcal{A}, \gamma, p, r)$. In the following we will discuss two approaches to solve this MDP with PG:

- Finite-time Simultaneous PG (FT-SimPG): An algorithm that is often used in practice, where parametrized policies are trained simultaneously, i.e. the parameters for $\pi_0, \dots, \pi_{H-1}$ are trained at once using the objective V_0 (cf. Section 3.2.2).
- Finite-time dynamic PG (FT-DynPG): A new algorithm that trains the parameters sequentially starting at the last epoch. We call this scheme finite-time dynamic policy gradient because it combines dynamic programming (backwards induction) and PG.

In order to carry out a complete theoretical analysis, we will assume that all policies are softmax parametrized. It is a first step towards a full understanding and already indicates why PG methods should use the dynamic programming structure inherent in finite-time MDPs. The evaluations in this section should not be seen as limited to the softmax case, but more like a kick-off to analyse a new approach which is beneficial in many scenarios.

Finite-time Simultaneous Policy Gradient. Recall, that the action spaces may depend on the current state and we denote the numbers of possible actions in epoch h by $d_h := \sum_{s \in \mathcal{S}_h} |\mathcal{A}_s|$. To perform a PG algorithm the artificial stationary policy $\pi \in \Pi_{\mathcal{S}^{[\mathcal{H}]}}$ must be parametrized. While the algorithm does not require a particular policy we will analyse the artificial stationary tabular softmax parametrization on the enlarged state space (cf. Definition 3.35) $\pi^\Theta \in \Pi_{\mathcal{S}^{[\mathcal{H}]}}$,

$$\pi^\Theta(a|s) = \frac{\exp(\theta(s,a))}{\sum_{a'} \exp(\theta(s,a'))}, \quad \Theta = (\theta(s,a))_{s \in \mathcal{S}^{[\mathcal{H}]}, a \in \mathcal{A}_s} \in \mathbb{R}^{\sum_h d_h}. \tag{4.1}$$

The tabular softmax parametrization uses a single parameter for each possible state-action pair at all epochs. Other parametrized policies, e.g. neural networks, take states from all epochs, i.e. from the enlarged state space $\mathcal{S}^{[\mathcal{H}]}$, as input variables. FT-SimPG trains all parameters at once and solves the optimization problem (to maximize the state value function at time 0) by gradient ascent over all parameters (all epochs) simultaneously.

Algorithm 4: Finite-time Simultaneous Policy Gradient (FT-SimPG)

Result: Approximate policy $\widehat{\pi}^* \approx \mathbb{\pi}^*$.
Input: Initial state distribution μ and class of policies $(\pi^\Theta)_{\Theta \in \mathbb{R}^d}$.
Initialize $\Theta^{(1)} \in \mathbb{R}^{\sum_h d_h}$;
Choose fixed step sizes $\alpha > 0$ and number of training steps N;
for $n = 1, \dots, N-1$ **do**
  $\Theta^{(n+1)} = \Theta^{(n)} + \alpha \nabla_\Theta V_0^{\pi^{\Theta^{(n)}}}(\mu)\big|_{\Theta^{(n)}}$;
end
Set $\widehat{\pi}^* = \pi^{\Theta^{(N)}}$;

Most importantly, the algorithm does not treat epochs differently, the same training effort goes into all epochs. Further, recall the objective function in the simultaneous approach without

discounting for any parametrization $\pi^\Theta \in \Pi_{\mathcal{S}^{[\mathcal{H}]}}$

$$J(\Theta,\mu) := V_0^{\pi^\Theta}(\mu) = \mathbb{E}_\mu^{\pi^\Theta}\Big[\sum_{h=0}^{H-1}\gamma^h r(S_h,A_h)\Big]. \tag{4.2}$$

We denote by $J^*(\mu) = \sup_\Theta J(\Theta,\mu)$ the optimal value of the objective function and note that $J^*(\mu) = V_0^*(\mu) = \sup_{\mathbb{m}\in\Pi^H} V_0^{\mathbb{m}}(\mu)$ under the tabular softmax parametrization, as an optimal policy can be approximated arbitrarily well.

Definition 4.1 (State visitation measure and distribution on the enlarged state space).

(i) The state visitation measure on the enlarged state space is defined by

$$\widetilde{\rho}_\mu^{\pi^\Theta}(s) := \sum_{h=0}^{H-1}\gamma^h \mathbb{P}_\mu^{\pi^\Theta}(S_h = s), \quad s \in \mathcal{S}^{[\mathcal{H}]}. \tag{4.3}$$

(ii) If $\gamma \in [0,1)$, then $\widetilde{d}_\mu^{\pi^\Theta} = \frac{1-\gamma}{1-\gamma^H}\widetilde{\rho}_\mu^{\pi^\Theta}$ is the normalized state visitation distribution on $\mathcal{S}^{[H]}$ and if $\gamma = 1$, then $\widetilde{d}_\mu^{\pi^\Theta} = \frac{1}{H}\widetilde{\rho}_\mu^{\pi^\Theta}$.

We derive the following version of the policy gradient theorem.

Theorem 4.2 (Policy Gradient Theorem for (FT-SimPG)). *Consider any parametrization π^Θ on the enlarged state space $\mathcal{S}^{[\mathcal{H}]}$, then the gradient of the $J(\Theta,\mu)$ defined in equation* (4.2) *is given by*

$$\begin{aligned}\nabla J(\Theta,\mu) &= \mathbb{E}_\mu^{\pi^\Theta}\Big[\sum_{h=0}^{H}\nabla\log(\pi^\Theta(A_h|S_h))Q_h^{\pi^\Theta}(S_h,A_h)\Big]\\ &= \sum_{s\in\mathcal{S}^{[\mathcal{H}]}}\widetilde{\rho}_\mu^{\pi^\Theta}(s)\sum_{a\in\mathcal{A}_s}\pi^\Theta(a|s)\nabla\log(\pi^\Theta(a|s))Q_h^{\pi^\Theta}(s,a).\end{aligned}$$

Proof. The second equality follows directly from the definition of the state visitation measure in equation (4.3).
For the first equality consider the probability of a trajectory $\tau = (s_0,a_0,\dots,s_{H-1},a_{H-1})$ under the policy π^Θ and initial state distribution μ, i.e.

$$p_\mu^{\pi^\Theta}(\tau) = \mu(s_h)\pi^\Theta(a_0|s_0)\prod_{k=1}^{H-1}p(s_k|s_{k-1},a_{k-1})\pi^\Theta(a_k|s_k).$$

Then,

$$\begin{aligned}\nabla\log(p_\mu^{\pi^\Theta}(\tau)) &= \nabla\Big(\log(\mu(s_h)) + \log(\pi^\Theta(a_0|s_0))\\ &\qquad + \sum_{k=1}^{H-1}\log(p(s_k|s_{k-1},a_{k-1})) + \log(\pi^\Theta(a_k|s_k))\Big)\\ &= \nabla\sum_{k=0}^{H-1}\log(\pi^\Theta(a_k|s_k)),\end{aligned}$$

which is known as the log-trick. Let $\mathcal{W}$ be the set of all trajectories from 0 to $H-1$. Note that $\mathcal{W}$ is finite due to the assumption that state and action space is finite. Then,

$$\begin{aligned}
\nabla J(\Theta,\mu) &= \nabla \sum_{\tau\in\mathcal{W}} p_\mu^{\pi^\Theta}(\tau) \sum_{k=0}^{H-1} \gamma^k r(s_k,a_k) \\
&= \sum_{\tau\in\mathcal{W}} p_\mu^{\pi^\Theta}(\tau) \nabla \log(p_\mu^{\pi^\Theta}(\tau)) \sum_{k=0}^{H-1} \gamma^k r(s_k,a_k) \\
&= \sum_{\tau\in\mathcal{W}} p_\mu^{\pi^\Theta}(\tau) \sum_{h=0}^{H-1} \nabla \log(\pi^\Theta(a_h|s_h)) \sum_{k=0}^{H-1} \gamma^k r(s_k,a_k) \\
&= \sum_{\tau\in\mathcal{W}} p_\mu^{\pi^\Theta}(\tau) \sum_{h=0}^{H-1} \nabla \log(\pi^\Theta(a_h|s_h)) \sum_{k=h}^{H-1} \gamma^k r(s_k,a_k) \\
&= \mathbb{E}_\mu^{\pi^\Theta}\Big[\sum_{h=0}^{H-1} \nabla \log(\pi^\Theta(A_h|S_h)) \sum_{k=h}^{H-1} \gamma^k r(S_k,A_k) \Big] \\
&= \mathbb{E}_\mu^{\pi^\Theta}\Big[\sum_{h=0}^{H-1} \nabla \log(\pi^\Theta(A_h|S_h)) \mathbb{E}_{S_h}^{\pi^\Theta}\Big[\sum_{k=h}^{H-1} \gamma^k r(S_k,A_k) \Big| S_h, A_h \Big] \Big] \\
&= \mathbb{E}_\mu^{\pi^\Theta}\Big[\sum_{h=0}^{H-1} \nabla \log(\pi^\Theta(A_h|S_h))\, Q_h^{\pi^\Theta}(S_h,A_h) \Big].
\end{aligned}$$

In the forth equation we have used that for every $k < h$ it holds

$$\begin{aligned}
&\mathbb{E}_\mu^{\pi^\Theta}\Big[\nabla \log(\pi^\Theta(A_h|S_h)) \gamma^k r(S_k,A_k) \Big] \\
&= \mathbb{E}_\mu^{\pi^\Theta}\Big[\mathbb{E}_\mu^{\pi^\Theta}\Big[\nabla \log(\pi^\Theta(A_h|S_h)) \Big| S_0, A_0, \dots S_{h-1}, A_{h-1}, S_h \Big] \gamma^k r(S_k,A_k) \Big] = 0,
\end{aligned}$$

because

$$\begin{aligned}
&\mathbb{E}_\mu^{\pi^\Theta}\Big[\nabla \log(\pi^\Theta(A_h|S_h)) \Big| S_0, A_0, \dots S_{h-1}, A_{h-1}, S_h \Big] \\
&= \mathbb{E}_\mu^{\pi^\Theta}\Big[\nabla \log(\pi^\Theta(A_h|S_h)) \Big| S_h \Big] \\
&= \sum_{a\in\mathcal{A}_{S_h}} \pi^\Theta(a|S_h) \nabla \log(\pi^\Theta(A_h|S_h)) \\
&= \nabla\Big(\sum_{a\in\mathcal{A}_{S_h}} \pi^\Theta(a|S_h) \Big) = 0.
\end{aligned}$$

■

In finite-time unbiased estimators of the gradient can be easily obtained using trajectories of finite-time length in a Monte-Carlo estimator.

Finite-time Dynamic Policy Gradient. First of all, recall that the inherent structure of finite-time MDPs is a backwards induction principle (dynamic programming), see Section 3.2.1. In

a way, finite-time MDPs can be viewed as nested contextual bandits. The FT-DynPG approach suggested in this article builds upon this intrinsic structure and sets on top a PG scheme. We have discussed in Remark 3.36 how an artificial stationary policy relates to a sequence of policies and for the dynamic approach, we consider the sequence $(\pi^{\theta_h})_{h=0}^{H-1}$ such that the policy in epoch h depends only on the parameter $\theta_h \in \mathbb{R}^{d_h}$. Thus, the tabular softmax parametrization we consider for FT-DynPG is formulated slightly differently than above: For each decision epoch $h \in \mathcal{H}$ the tabular softmax parametrization is given by

$$\pi^{\theta_h}(a|s) = \frac{\exp(\theta_h(s,a))}{\sum_{a' \in \mathcal{A}} \exp(\theta_h(s,a'))}, \quad \theta_h = (\theta_h(s,a))_{s \in \mathcal{S}_h, a \in \mathcal{A}_s} \in \mathbb{R}^{d_h}. \tag{4.4}$$

The total dimension of the parameter tensor $[\theta_0, \dots, \theta_{H-1}]^T$ equals the one of Θ. The only difference lies in the notation and the epoch dependence is made more explicit in equation (4.4). The main idea of FT-DynPG is as follows. The dynamic programming perspective suggests to learn policies backwards in time. Thus, we start by training the last parameter vector θ_{H-1} on the sub-problem V_{H-1}, a one-step MDP which can be viewed as contextual bandit problem. After convergence up to some termination condition, it is known how to act near optimality in the last epoch and one can proceed to train the parameter vector from previous epochs by exploiting the knowledge of acting near optimal in the future. This is what the proposed FT-DynPG algorithm does. A policy is trained up to some termination condition and then used to optimize the earlier epochs.

Algorithm 5: Finite-time Dynamic Policy Gradient (FT-DynPG)

Result: Approximate policy $\widehat{\mathbb{\pi}}^* \approx \mathbb{\pi}^*$.
Input: Initial state distributions $(\mu_h)_{h=0}^{\infty}$ and class of policies $(\mathbb{\pi}^{\Theta})$.
Initialize $\theta^{(1)} = (\theta_0^{(1)}, \dots, \theta_{H-1}^{(1)}) \in \mathbb{R}^{\sum_h d_h}$;
for $h = H-1, \dots, 0$ **do**
  Choose fixed step size α_h and number of training steps N_h;
  for $n = 1, \dots, N_h - 1$ **do**
    $\theta_h^{(n+1)} = \theta_h^{(n)} + \alpha_h \nabla_{\theta_h} V_h^{(\pi^{\theta_h}, \widehat{\mathbb{\pi}}^*_{(h+1)})}(\mu_h)\big|_{\theta_h^{(n)}}$;
  end
  Set $\widehat{\pi}_h^* = \pi^{\theta_h^{(N_h)}}$;
end

Remark 4.3. Suppose that we have trained the first $h+1$ policies such that $\widehat{\pi}_k^* \approx \pi_k^*$ for $k = h+1, \dots, H-1$ and $V_{h+1}^{\mathbb{\pi}^*_{(h+1)}} \approx V_h^*$. In order to train parameter θ_h, we have to optimize $V_h^{(\pi^{\theta_h}, \widehat{\mathbb{\pi}}^*_{(h+1)})}(\mu_h)$ and by the dynamic programming principle (cf. equation (3.17)), it holds that

$$\begin{aligned} V_h^{(\pi^{\theta_h}, \widehat{\mathbb{\pi}}^*_{(h+1)})}(\mu_h) &= \sum_{s \in \mathcal{S}_h} \mu_h(s) \sum_{a \in \mathcal{A}_s} \pi^{\theta_h}(a|s) \gamma^h r(s,a) + \sum_{s' \in \mathcal{S}_{h+1}} V_{h+1}^{\widehat{\mathbb{\pi}}^*_{(h+1)}}(s') \\ &= T_h^{\pi^{\theta_h}}(V_{h+1}^{\widehat{\mathbb{\pi}}^*_{(h+1)}}). \end{aligned}$$

Hence, we train the next policy such that approximately $T_h^{\pi^{\theta_h}}(V_{h+1}^{\mathbb{\pi}^*_{(h+1)}}) \approx T_h^*(V_{h+1}^{\mathbb{\pi}^*_{(h+1)}})$. Note that the algorithm is just working fine when convergence close to the optimal policy is guaranteed in ever optimization step (cf. Remark 3.34).

A bit of notation is needed to analyse this approach. Given any fixed policy $\widetilde{\mathbb{\pi}} \in \Pi^H$, the objective function J_h in epoch h is defined to be the h-state value function in state under the extended policy $(\pi^{\theta_h}, \widetilde{\mathbb{\pi}}_{(h+1)}) := (\pi^{\theta_h}, \widetilde{\mathbb{\pi}}_{h+1}, \dots, \widetilde{\mathbb{\pi}}_{H-1})$,

$$J_h(\theta_h, \widetilde{\mathbb{\pi}}_{(h+1)}, \mu_h) := V_h^{(\pi^{\theta_h}, \widetilde{\mathbb{\pi}}_{(h+1)})}(\mu_h) = \mathbb{E}_{\mu_h}^{(\pi^{\theta_h}, \widetilde{\mathbb{\pi}}_{(h+1)})}\Big[\sum_{k=h}^{H-1} \gamma^k r(S_k, A_k)\Big]. \tag{4.5}$$

While the notation is a bit heavy the intuition behind is easy to understand. If the policy after epoch h is already trained (this is $\widetilde{\mathbb{\pi}}_{(h+1)}$) then J_h as a function of θ_h is the parametrized dependence of the value function when only the policy for epoch h is changed. Gradient ascent is then used to find a parameter θ_h^* that maximizes $J_h(\cdot, \widetilde{\mathbb{\pi}}_{(h+1)}, \delta_s)$, for all $s \in \mathcal{S}_h$, where δ_s the dirac measure on s. Note that θ_h^* depends on the fixed future policy $\widetilde{\mathbb{\pi}}_{(h+1)}$ and to train θ_h one chooses $\widetilde{\mathbb{\pi}}_{(h+1)} = \widehat{\mathbb{\pi}}^*_{(h+1)}$ in Algorithm 5. We define $J_h^*(\widetilde{\mathbb{\pi}}_{(h+1)}, \mu_h) := \sup_{\theta_h} J_h(\theta_h, \widetilde{\mathbb{\pi}}_{(h+1)}, \mu_h)$.

THEOREM 4.4 (Policy Gradient Theorem for (FT-DynPG)). *For a fixed policy $\widetilde{\mathbb{\pi}}$ and $h \in \mathcal{H}$ the gradient of $J_h(\theta_h, \widetilde{\mathbb{\pi}}_{(h+1)}, \delta_s)$ defined in equation* (4.5) *is given by*

$$\nabla J_h(\theta_h, \widetilde{\mathbb{\pi}}_{(h+1)}, \delta_s) = \mathbb{E}_{S_h=s, A_h \sim \pi^{\theta_h}(\cdot|s)}[\nabla \log(\pi^{\theta_h}(A_h|S_h)) Q_h^{\widetilde{\mathbb{\pi}}}(S_h, A_h)].$$

Proof. The probability of a trajectory $\tau = (s_h, a_h, \dots, s_{H-1}, a_{H-1})$ under the policy $(\pi^\theta, \widetilde{\mathbb{\pi}}_{(h+1)}) = (\pi^\theta, \widetilde{\mathbb{\pi}}_{h+1}, \dots, \widetilde{\mathbb{\pi}}_{H-1})$ and initial state distribution δ_s is given by

$$p_s^{(\pi^\theta, \widetilde{\mathbb{\pi}}_{(h+1)})}(\tau) = \delta_s(s_h)\pi^\theta(a_h|s_h) \prod_{k=h+1}^{H-1} p(s_k|s_{k-1}, a_{k-1}) \widetilde{\mathbb{\pi}}_k(a_k|s_k).$$

Then,

$$\begin{aligned}
\nabla \log(p_s^{(\pi^\theta, \widetilde{\mathbb{\pi}}_{(h+1)})}(\tau)) &= \nabla\Big(\log(\delta_s(s_h)) + \log(\pi^\theta(a_h|s_h)) \\
&\qquad + \sum_{k=h+1}^{H-1} \log(p(s_k|s_{k-1}, a_{k-1})) + \log(\widetilde{\mathbb{\pi}}_k(a_k|s_k))\Big) \\
&= \nabla \log(\pi^\theta(a_h|s_h)),
\end{aligned}$$

which is known as the log-trick. Let $\mathcal{W}$ be the set of all trajectories from h to $H-1$. Note that $\mathcal{W}$

is finite due to the assumption that state and action space is finite. Then for $s \in \mathcal{S}_h$

$$\begin{aligned}
\nabla J_h(\theta_h, \widetilde{\mathbb{m}}_{(h+1)}, \delta_s) &= \nabla \sum_{\tau \in \mathcal{W}} p_s^{(\pi^\theta, \widetilde{\mathbb{m}}_{(h+1)})}(\tau) \sum_{k=h}^{H-1} \gamma^k r(s_k, a_k) \\
&= \sum_{\tau \in \mathcal{W}} p_s^{(\pi^\theta, \widetilde{\mathbb{m}}_{(h+1)})}(\tau) \nabla \log(p_s^{(\pi^\theta, \widetilde{\mathbb{m}}_{(h+1)})}(\tau)) \sum_{k=h}^{H-1} \gamma^k r(s_k, a_k) \\
&= \sum_{\tau \in \mathcal{W}} p_s^{(\pi^\theta, \widetilde{\mathbb{m}}_{(h+1)})}(\tau) \nabla \log(\pi^\theta(a_h|s_h)) \sum_{k=h}^{H-1} \gamma^k r(s_k, a_k) \\
&= \mathbb{E}_{S_h=s}^{(\pi^\theta, \widetilde{\mathbb{m}}_{(h+1)})} \left[\nabla \log(\pi^\theta(A_h|S_h)) \sum_{k=h}^{H-1} \gamma^k r(S_k, A_k) \right] \\
&= \mathbb{E}_{S_h=s}^{(\pi^\theta, \widetilde{\mathbb{m}}_{(h+1)})} \left[\nabla \log(\pi^\theta(A_h|S_h)) \mathbb{E}_{S_h}^{\widetilde{\mathbb{m}}} \left[\sum_{k=h}^{H-1} \gamma^k r(S_k, A_k) \middle| S_h, A_h \right] \right] \\
&= \mathbb{E}_{S_h=s, A_h \sim \pi^\theta(\cdot|s)} \left[\nabla \log(\pi^\theta(A_h|S_h))\, Q_h^{\widetilde{\mathbb{m}}}(S_h, A_h) \right].
\end{aligned}$$

■

A priori it is not clear if simultaneous or dynamic programming inspired training is more efficient. FT-DynPG has an additional loop but trains less parameters at once. We give a detailed analysis for the tabular softmax parametrization but want to give a heuristic argument why simultaneous training is not favorable. The policy gradient theorem in the simultaneous approach (Theorem 4.2) states that

$$\nabla J(\Theta, \mu) = \sum_{s \in \mathcal{S}^{[\mathcal{H}]}} \widetilde{\rho}_\mu^{\pi^\Theta}(s) \sum_{a \in \mathcal{A}_s} \pi^\Theta(a|s) \nabla \log(\pi^\Theta(a|s)) Q_h^{\pi^\Theta}(s, a).$$

It implies that training policies at earlier epochs are massively influenced by estimation errors of $Q_h^{\pi^\Theta}$. Reasonable training of optimal decisions is only possible if all later epochs have been trained well, i.e. $Q_h^{\pi^\Theta} \approx Q_h^*$. This may lead to inefficiency in earlier epochs when training all epochs simultaneously. It is important to note that the policy gradient formula is independent of the parametrization. While our precise analysis is only carried out for tabular softmax parametrizations this general heuristic remains valid for all classes of policies.
For the rest of this chapter we assume undiscounted finite-time MDPs with positive finite rewards:

Assumption 4.5. *We assume that $\gamma = 1$ and the rewards are bounded in $[0, R^*]$, for some $R^* > 0$.*

Remark 4.6. Note that the assumption $\gamma = 1$ can be relaxed to $\gamma \in [0, 1]$ as discounting is not relevant in finite-time MDPs (see also Remark 3.24). More precisely, we always upper bound $\sum_{h=0}^{H-1} r(S_h, A_h) \leq HR^*$ (a.s.) and this upper bound holds also with a modified reward function including discounting as $\gamma^h r(S_h, A_h) \leq r(S_h, A_h)$ (a.s.) for any $h \in \{0, \dots, H-1\}$. Moreover, the positivity assumption on the rewards is no restriction of generality, bounded negative rewards can be shifted using the base-line trick, and boundedness can always be assumed by the finite state and action space.

In what follows we will always assume the tabular softmax parametrization and analyse both PG schemes. First under the assumption of exact gradients, then with sampled gradients à la REINFORCE.

4.2 CONVERGENCE OF SOFTMAX POLICY GRADIENT WITH EXACT GRADIENTS

In the following, we analyse the convergence behavior of the simultaneous as well as the dynamic approach under the assumption to have access to exact gradient computation. The presented convergence analysis in both settings is inspired from the discounted setting considered recently in Agarwal et al. [Aga+21] and Mei et al. [Mei+20]. In both scenarios the global convergence in based on smoothness of the objective (Lipschitz continuous gradients) and a (weak) gradient domination property.
In a maximization problem for a differential function $f : \mathbb{R}^d \to \mathbb{R}$ with $f^* = \sup_x f(x) < \infty$, the weak gradient domination ($\beta = 1$ in Definition 2.7) is globally satisfied when $c > 0$ exists, such that

$$\|\nabla f(x)\| \geq c(f^* - f(x)), \quad \forall x \in \mathbb{R}^d.$$

This assumption is weaker then convexity but, combined with smoothness, still ensures convergence of $f(x_n) \to f^*$ in a gradient ascent scheme (cf. Theorem 2.10). In the following subsections, we will first derive a so called non-uniform gradient domination property, where c is not a constant, but a function $c(x)$. In a second step, we ensure that $\inf_n c(x_n) \geq c$ bounded away from 0 along the gradient ascent trajectory. Finally, we derive convergence from these results.
Before we deal with the two approaches separately, we formulate the performance difference lemma for the two objectives $J(\Theta, \mu)$ and $J_h(\theta_h, \widetilde{\mathbb{m}}_{(h+1)}, \mu_h)$. We will use these to derive the non-uniform gradient domination properties.

COROLLARY 4.7. *For the objective $J(\Theta, \mu)$ defined in equation* (4.2) *and $J_h(\theta_h, \widetilde{\mathbb{m}}_{(h+1)}, \mu_h)$ defined in equation* (4.5) *it holds*

$$J^*(\mu) - J(\Theta, \mu) = \mathbb{E}_\mu^{\pi^*}\Big[\sum_{h=0}^{H-1} A_h^{\pi^\Theta}(S_h, A_h)\Big] = \sum_{s \in \mathcal{S}^{[\mathcal{H}]}} \widetilde{\rho}_\mu^{\pi^*}(s) A_h^{\pi^\Theta}(s, a^*(s))$$

and

$$J_h^*(\widetilde{\mathbb{m}}_{(h+1)}, \mu) - J_h(\theta_h, \widetilde{\mathbb{m}}_{(h+1)}, \mu_h) = \mathbb{E}_\mu^{\mathbb{m}^*}\Big[A_h^{(\pi^\theta, \widetilde{\mathbb{m}}_{(h+1)})}(S_h, A_h)\Big].$$

Proof. The first claim follows directly from Lemma 3.32 and the definition of the state visitation measure on $\mathcal{S}^{[\mathcal{H}]}$ in equation (4.3).
For the second claim, we proof a more general result: For any $h \in \mathcal{H}$ and two policies π and π': If $\mathbb{m}_{(h+1)} = \mathbb{m}'_{(h+1)}$, it holds that

$$V_h^{\mathbb{m}}(s) - V_h^{\mathbb{m}'}(s) = \mathbb{E}_{S_h=s}^{\mathbb{m}_{(h)}}\Big[A_h^{\mathbb{m}'}(S_h, A_h)\Big].$$

To see this, let $k > h$, then

$$\begin{aligned}
\mathbb{E}_{S_h=s}^{\mathbb{m}_{(h)}}\Big[A_k^{\mathbb{m}'}(S_k, A_k)\Big] &= \sum_{a \in \mathcal{A}} \pi_h(a|s) \sum_{s' \in \mathcal{S}} p(s'|s, a) \mathbb{E}_{S_{h+1}=s'}^{\mathbb{m}_{(h+1)}}\Big[Q_k^{\mathbb{m}'}(S_k, A_k) - V_k^{\mathbb{m}'}(S_k)\Big] \\
&= \sum_{a \in \mathcal{A}} \pi_h(a|s) \sum_{s' \in \mathcal{S}} p(s'|s, a) \mathbb{E}_{S_{h+1}=s'}^{\mathbb{m}'_{(h+1)}}\Big[Q_k^{\mathbb{m}'}(S_k, A_k) - V_k^{\mathbb{m}'}(S_k)\Big]
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{a\in\mathcal{A}} \pi_h(a|s) \sum_{s'\in\mathcal{S}} p(s'|s,a)\Big(\mathbb{E}^{\mathbb{T}'_{(h+1)}}_{S_{h+1}=s'}\Big[\mathbb{E}^{\mathbb{T}'}_{S_k}[Q^{\mathbb{T}'}_k(S_k,A_k)]\Big] - \mathbb{E}^{\mathbb{T}'_{(h+1)}}_{S_{h+1}=s'}\Big[V^{\mathbb{T}'}_k(S_k)\Big]\Big)\\
&= \sum_{a\in\mathcal{A}} \pi_h(a|s) \sum_{s'\in\mathcal{S}} p(s'|s,a)\Big(\mathbb{E}^{\mathbb{T}'_{(h+1)}}_{S_{h+1}=s'}\Big[V^{\mathbb{T}'}_k(S_k)\Big] - \mathbb{E}^{\mathbb{T}'_{(h+1)}}_{S_{h+1}=s'}\Big[V^{\mathbb{T}'}_k(S_k)\Big]\Big)\\
&= 0.
\end{aligned}
$$

The claim follows with Lemma 3.32. ∎

4.2.1 *Finite-time Simultaneous Policy Gradient*

To prove convergence in the simultaneous approach we will interpret the finite-time MDP as an undiscounted stationary problem with state-space $\mathcal{S}^{[H]}$ and deterministic absorption time H. This MDP is undiscounted but terminates in finite-time. Building upon [Aga+21; Mei+20; YGL22], we prove that the objective function defined in equation (4.2) is L-smooth with parameter $L = H^2R^*(2-\frac{1}{|\mathcal{A}|})$ and satisfies a non-uniform weak gradient domination with $\beta = 1$.

Remark 4.8. Note that we can drop the subscript h in the value function, state-action value function or advantage function, when we define them on the enlarged state-space $\mathcal{S}^{[\mathcal{H}]}$. Then, V is a vector of dimension $|\mathcal{S}^{[\mathcal{H}]}|$ and Q and A are matrices of dimension $|\mathcal{S}^{[\mathcal{H}]}| \times |\mathcal{A}|$. More precisely, using a state $s \in \mathcal{S}^{[\mathcal{H}]}$ then s is assigned to one specific epoch $h \in \mathcal{H}$ and we introduce the value function on $\mathcal{S}^{[\mathcal{H}]}$ by $V^{\pi^\Theta}(s) = \sum_{h=0}^{H-1} \mathbf{1}_{s \text{ belongs to epoch } h} V_h^{\pi^\Theta}(s)$. Similar also for Q and A.

Smoothness. To derive the smoothness of the objective function we have to prove that ∇J is L-Lipschitz. Therefore we use the explicit formula of the gradient under softmax parametrization.

Lemma 4.9. *The partial derivative of the objective defined in equation* (4.2) *under softmax parametrization is given by:*

$$
\frac{\partial J(\Theta,\mu)}{\partial\theta(s,a)} = \widetilde{\rho}_\mu^{\pi^\Theta}(s)\pi^\Theta(a|s)A^{\pi^\Theta}(s,a),
$$

for every $s \in \mathcal{S}^{[\mathcal{H}]}$ and $a \in \mathcal{A}_s$.

Proof. Let $s \in \mathcal{S}^{[\mathcal{H}]}$ and $a \in \mathcal{A}_s$. Using Theorem 4.2, it holds that

$$
\begin{aligned}
\frac{\partial J(\Theta,\mu)}{\partial\theta(s,a)} &= \mathbb{E}^{\pi^\Theta}_\mu\Big[\sum_{h=0}^{H-1} \frac{\partial}{\partial\theta(s,a)} \log(\pi^\Theta(A_h|S_h)) Q_h^{\pi^\Theta}(S_h,A_h)\Big]\\
&= \mathbb{E}^{\pi^\Theta}_\mu\Big[\sum_{h=0}^{H-1} \mathbf{1}_{\{S_h=s\}}\big(\mathbf{1}_{\{A_h=a\}} - \pi^\Theta(a|s)\big) Q_h^{\pi^\Theta}(S_h,A_h)\Big]\\
&= \sum_{h=0}^{H-1} \mathbb{P}^{\pi^\Theta}_\mu(S_h=s) \sum_{a'} \pi^\Theta(a'|s)\big(\mathbf{1}_{\{a'=a\}} - \pi^\Theta(a|s)\big) Q_h^{\pi^\Theta}(s,a')\\
&= \widetilde{\rho}_\mu^{\pi^\Theta}(s)\Big(\pi^\Theta(a|s)Q^{\pi^\Theta}(s,a) - \sum_{a'} \pi^\Theta(a'|s)\pi^\Theta(a|s)Q^{\pi^\Theta}(s,a')\Big)\\
&= \widetilde{\rho}_\mu^{\pi^\Theta}(s)\pi^\Theta(a|s)A^{\pi^\Theta}(s,a).
\end{aligned}
$$

∎

We deduce the smoothness directly from this result.

LEMMA 4.10. *The objective $J(\Theta, \mu)$ from equation* (4.2) *under softmax parametrization is smooth in Θ with parameter $L = H^2R^*(2 - \frac{1}{|\mathcal{A}|})$.*

Proof. We are going to bound the norm of the hessian. Therefore, we first calculate the first and second derivative of J for finite-time horizon stationary MDPs. So, let $\tau = (s_0, a_0, s_1, \dots, s_{H-1}, a_{H-1})$ be a trajectory of the MDP under policy π^Θ and denote by p_μ^Θ the discrete probability density. Then,

$$\begin{aligned}\nabla J(\Theta, \mu) &= \nabla\Big(\sum_\tau p_\mu^\Theta(\tau)\sum_{h=0}^{H-1} r(s_h, a_h)\Big)\\ &= \sum_\tau p_\mu^\Theta(\tau)\Big(\sum_{h=0}^{H-1}\nabla\log(\pi^\Theta(a_h|s_h))\sum_{h=0}^{H-1} r(s_h, a_h)\Big)\\ &= \mathbb{E}_\mu^{\pi^\Theta}\Big[\sum_{h=0}^{H-1}\nabla\log(\pi^\Theta(a_h|s_h))\sum_{h=0}^{H-1} r(s_h, a_h)\Big].\end{aligned}$$

For the second derivative we have

$$\begin{aligned}\nabla^2 J(\Theta, \mu) &= \nabla\Big(\sum_\tau p_\mu^\Theta(\tau)\big(\sum_{h=0}^{H-1}\nabla\log(\pi^\Theta(a_h|s_h))\sum_{h=0}^{H-1} r(s_h, a_h)\big)\Big)\\ &= \underbrace{\sum_\tau p_\mu^\Theta(\tau)\Big(\big(\sum_{h=0}^{H-1}\nabla\log(\pi^\Theta(a_h|s_h))\big)\big(\sum_{h=0}^{H-1}\nabla\log(\pi^\Theta(a_h|s_h))\big)^T\sum_{h=0}^{H-1} r(s_h, a_h)\Big)}_{(1)}\\ &\quad + \underbrace{\sum_\tau p_\mu^\Theta(\tau)\Big(\sum_{h=0}^{H-1}\nabla^2\log(\pi^\Theta(a_h|s_h))\sum_{h=0}^{H-1} r(s_h, a_h)\Big)}_{(2)}.\end{aligned}$$

Using the bounded reward assumption we get for the second term, that

$$||(2)|| \le \mathbb{E}_\mu^{\pi^\Theta}\Big[\sum_{h=0}^{H-1}\|\nabla^2\log(\pi^\Theta(a_h|s_h))\|\Big]HR^* = HR^*\sum_{h=0}^{H-1}\mathbb{E}_\mu^{\pi^\Theta}\Big[\|\nabla^2\log(\pi^\Theta(a_h|s_h))\|\Big].$$

By [YGL22, Lem. 4.8], we have for the softmax parametrization that $\mathbb{E}_\mu^{\pi^\Theta}\Big[\|\nabla^2\log(\pi^\Theta(a_h|s_h))\|\Big] \le 1$. Hence, $||(2)|| \le H^2R^*$.
Next for the first term,

$$\begin{aligned}||(1)|| &\le \mathbb{E}_\mu^{\pi^\Theta}\Big[\|\sum_{h=0}^{H-1}\nabla\log(\pi^\Theta(a_h|s_h))\|^2\Big]HR^*\\ &= HR^*\sum_{h=0}^{H-1}\mathbb{E}_\mu^{\pi^\theta}\Big[\|\nabla\log(\pi^\Theta(a_h|s_h))\|^2\Big]\end{aligned}$$

$$\leq H^2 R^* \big(1 - \frac{1}{|\mathbb{A}|}\big),$$

where we first used the bounded reward assumption, then Lemma 3.6 and again Lemma 4.8 from Yuan, Gower, and Lazaric [YGL22]. Finally, we obtain that

$$\|\nabla^2 J(\Theta, \mu)\| \leq H^2 R^* \big(2 - \frac{1}{|\mathbb{A}|}\big).$$

■

We can compare this result to the smoothness of a discounted MDP under softmax parametrization, where $L = \frac{R^*}{(1-\gamma)^2}\big(2 - \frac{1}{|\mathbb{A}|}\big)$ (Lemma 3.19 (i)). We obtain that $\frac{1}{1-\gamma}$, the expectation of a geometric r.v. and the expected length of a discounted MDP, is replaced by H, the expected length of the finite-time MDP.

Weak gradient domination. In the following let $\pi^* \in \Pi_{\mathbb{S}^{[\mathcal{H}]}}$ be the reinterpretation of a fixed but arbitrary deterministic optimal policy $\mathbb{\pi}^* = (\pi_h^*)_{h=0}^{H-1} \in \Pi^H$. Then $a^*(s) = \operatorname{argmax}_{a \in \mathbb{A}_s} \pi^*(a|s)$ is an arbitrary and fixed best action in state $s \in \mathbb{S}^{[\mathcal{H}]}$ (compare to Remark 3.34).

Lemma 4.11. *Under softmax parametrization, it holds that*

$$\|\nabla J(\Theta, \mu)\|_2 \geq \frac{\min_{s \in \mathbb{S}^{[\mathcal{H}]}} \pi^\Theta(a^*(s)|s)}{\sqrt{|\mathbb{S}^{[\mathcal{H}]}|}} \Big\| \frac{d_\mu^{\pi^*}}{d_\mu^{\pi^\Theta}} \Big\|_\infty^{-1} (J^*(\mu) - J(\Theta, \mu)).$$

Proof. The idea of the proof follows the outline of Mei et al. [Mei+20, Lem. 8] from the discounted setting. It holds

$$\begin{aligned}
\|\nabla J(\Theta, \mu)\|_2 &= \Big[\sum_{s \in \mathbb{S}^{[\mathcal{H}]}} \sum_a \Big(\frac{\partial V_0^{\pi^\Theta}(\mu)}{\partial \theta(s,a)} \Big)^2 \Big]^{1/2} \\
&\geq \Big[\sum_{s \in \mathbb{S}^{[\mathcal{H}]}} \Big(\frac{\partial V_0^{\pi^\Theta}(\mu)}{\partial \theta(s, a^*(s))} \Big)^2 \Big]^{1/2} \\
&\geq \frac{1}{\sqrt{|\mathbb{S}^{[\mathcal{H}]}|}} \sum_{s \in \mathbb{S}^{[\mathcal{H}]}} \Big| \frac{\partial V_0^{\pi^\Theta}(\mu)}{\partial \theta(s, a^*(s))} \Big| \\
&= \frac{1}{\sqrt{|\mathbb{S}^{[\mathcal{H}]}|}} \sum_{s \in \mathbb{S}^{[\mathcal{H}]}} \widetilde{\rho}_\mu^{\pi^\Theta}(s) \pi^\Theta(a^*(s)|s) |A^{\pi^\Theta}(s, a^*(s))| \\
&\geq \frac{\min_{s \in \mathbb{S}^{[\mathcal{H}]}} \pi^\Theta(a^*(s)|s)}{\sqrt{|\mathbb{S}^{[\mathcal{H}]}|}} \sum_{s \in \mathbb{S}^{[\mathcal{H}]}} \widetilde{\rho}_\mu^{\pi^*}(s) \Big\| \frac{d_\mu^{\pi^*}}{d_\mu^{\pi^\Theta}} \Big\|_\infty^{-1} A^{\pi^\Theta}(s, a^*(s)) \\
&= \frac{\min_{s \in \mathbb{S}^{[\mathcal{H}]}} \pi^\Theta(a^*(s)|s)}{\sqrt{|\mathbb{S}^{[\mathcal{H}]}|}} \Big\| \frac{d_\mu^{\pi^*}}{d_\mu^{\pi^\Theta}} \Big\|_\infty^{-1} \underbrace{\sum_{s \in \mathbb{S}^{[\mathcal{H}]}} \rho_\mu^{\pi^*}(s) A^{\pi^\Theta}(s, a^*(s))}_{= \mathbb{E}_\mu^{\pi^*}[\sum_{h=0}^{H-1} A_h^{\pi^\Theta}(S_h, A_h)]} \\
&= \frac{\min_{s \in \mathbb{S}^{[\mathcal{H}]}} \pi^\Theta(a^*(s)|s)}{\sqrt{||\mathbb{S}^{[\mathcal{H}]}|}} \Big\| \frac{d_\mu^{\pi^*}}{d_\mu^{\pi^\Theta}} \Big\|_\infty^{-1} (J^*(\mu) - J(\Theta, \mu)).
\end{aligned}$$

The third line is due to Cauchy-Schwarz, afterwards we used the derivative of the objective function from Lemma 4.9 and that $\left\|\frac{\widetilde{\rho}_\mu^{\pi^*}}{\widetilde{\rho}_\mu^{\pi^\theta}}\right\|_\infty = \left\|\frac{d_\mu^{\pi^*}}{d_\mu^{\pi^\theta}}\right\|_\infty$ by definition of the state visitation measures and the reinterpretation of measures. Finally, the last equation is due to Corollary 4.7 from the performance difference lemma. ∎

The term

$$\left\|\frac{d_\mu^{\mathbb{m}^*}}{d_\mu^{\pi^\Theta}}\right\|_\infty := \max_{s\in\mathcal{S}} \frac{d_\mu^{\pi^*}(s)}{d_\mu^{\pi^\Theta}(s)} \tag{4.6}$$

is again the distribution mismatch coefficient similar to the discounted case (cf. Lemma 3.19). To ensure that the distribution mismatch coefficient can be bounded from below uniformly in θ we make the following assumption.

ASSUMPTION 4.12. *For FT-SimPG we assume that the state space is constant over all epochs, i.e.* $\mathcal{S}_h = \mathcal{S}$ *for all epochs.*

Remark 4.13. Under Assumption 4.12 it holds $d_\mu^{\pi^\Theta}(s) \geq \frac{1}{H}\mu(s)$ by definition for any policy π^Θ on the enlarged state space, since

$$d_\mu^{\pi^\Theta}(s) = \frac{1}{H}\sum_{h=0}^{H-1} \mathbb{P}_\mu^{\pi^\Theta}(S_h = s) \geq \frac{1}{H}\mu(s).$$

Hence, we obtain that

$$\|\nabla J(\Theta,\mu)\|_2 \geq \frac{\min_{s\in\mathcal{S}^{[\mathcal{H}]}} \pi^\Theta(a^*(s)|s)}{H\sqrt{|\mathcal{S}|H}} \left\|\frac{d_\mu^{\pi^*}}{\mu}\right\|_\infty^{-1} (J^*(\mu) - J(\Theta,\mu)).$$

Without Assumption 4.12 for $s \in \mathcal{S}$ and $s \notin \mathcal{S}_0$ we cannot bound $\sum_{h=0}^{H-1} \mathbb{P}_\mu^{\pi^\Theta}(S_h = s)$ with μ, as μ is only defined on $\mathcal{S}_0$.

As already pointed out in Mei et al. [Mei+20] one key challenge in providing global convergence is to bound the term $\min_{s\in\mathcal{S}} \pi^\Theta(a_h^*(s)|s)$ from below uniformly in Θ appearing in the gradient ascent updates. Techniques introduced in Agarwal et al. [Aga+21] can be extended to the finite-horizon setting to prove asymptotic convergence towards global optima (see Appendix A). This will be used to bound $c = c(\Theta^{(1)}) = \inf_n \min_{s\in\mathcal{S}} \pi^{\Theta^{(n)}}(a_h^*(s)|s) > 0$.

LEMMA 4.14. *Let* μ *be a probability measure such that* $\mu(s) > 0$ *for all* $s \in \mathcal{S}$*, let Assumption 4.12 holds true and let* $0 < \alpha \leq \frac{1}{5H^2R^*}$*. Consider the sequence* $(\Theta^{(n)})$ *generated by FT-SimPG (Algorithm 4) under softmax parametrization with arbitrary* $\Theta^{(1)} \in \mathbb{R}^{\sum_h d_h}$*. Then,* $c = c(\theta^{(1)}) = \inf_n \min_{s\in\mathcal{S}^{[\mathcal{H}]}} \pi^{\Theta^{(n)}}(a^*(s)|s) > 0$.

Proof. The proof is adapted from the discounted setting in Mei et al. [Mei+20, Lem. 9] to the finite-time horizon setting.
We will drop the μ in $J(\Theta,\mu)$ for the rest of the proof to save notation.
Define for all $s \in \mathcal{S}^{[\mathcal{H}]}$,

$$\Delta^*(s) = Q^\infty(s, a_h^*(s)) - \max_{a\neq a^*(s)} Q^\infty(s,a) > 0, \quad \text{and} \quad \Delta^* = \min_{s\in\mathcal{S}^{[\mathcal{H}]}} \Delta^*(s) > 0,$$

where Q^{∞} is the optimal Q-function on the enlarged state space from Lemma A.2.
Now consider for any $s \in \mathcal{S}^{[\mathcal{H}]}$ the following sets

$$\mathcal{R}_1(s) = \left\{\Theta : \frac{\partial J(\Theta)}{\partial \theta(s, a^*(s))} \geq \frac{\partial J(\Theta)}{\partial \theta(s,a)}, \text{ for all } a \neq a^*(s)\right\},$$
$$\mathcal{R}_2(s) = \left\{\Theta : Q^{\pi^\Theta}(s, a^*(s)) \geq Q^{\infty}(s, a^*(s)) - \frac{\Delta^*(s)}{2}\right\},$$
$$\mathcal{R}_3(s) = \left\{\Theta^{(n)} : V^{\pi^{\Theta^{(n)}}}(s) \geq Q^{\pi^{\Theta^{(n)}}}(s, a^*(s)) - \frac{\Delta^*(s)}{2}, \text{ for all } n \geq 1 \text{ large enough}\right\}.$$

Furthermore, for any $s \in \mathcal{S}^{[\mathcal{H}]}$ we define $c(s) = \frac{|\mathcal{A}|HR^*}{\Delta^*(s)} - 1$ and

$$N_c(s) = \left\{\Theta : \pi^\Theta(a^*(s)|s) \geq \frac{c(s)}{c(s)+1}\right\}.$$

We divide the proof into the following Claims:

1. $\mathcal{R}(s) = \mathcal{R}_1(s) \cap \mathcal{R}_2(s) \cap \mathcal{R}_3(s)$ is a *nice* region, i.e.
   (i) $\Theta^{(n)} \in \mathcal{R}(s) \Rightarrow \Theta^{(n+1)} \in \mathcal{R}(s)$.
   (ii) $\pi^{\Theta^{(n+1)}}(a^*(s)|s) \geq \pi^{\Theta^{(n)}}(a^*(s)|s)$.
2. $\mathcal{N}_c(s) \cap \mathcal{R}_2(s) \cap \mathcal{R}_3(s) \subset \mathcal{R}_1(s) \cap \mathcal{R}_2(s) \cap \mathcal{R}_3(s)$.
3. For every $s \in \mathcal{S}^{[\mathcal{H}]}$, there exists a finite-time $n_0(s) \geq 1$, such that
$$\theta^{(n_0(s))} \in \mathcal{N}_c(s) \cap \mathcal{R}_2(s) \cap \mathcal{R}_3(s) \subset \mathcal{R}_1(s) \cap \mathcal{R}_2(s) \cap \mathcal{R}_3(s)$$
and thus
$$\inf_{n\geq 1} \pi^{\Theta^{(n)}}(a^*(s)|s) = \min_{1\leq n\leq n_0(s)} \pi^{\Theta^{(n)}}(a^*(s)|s)$$
.

If all three claims hold true, we can finally define $n_0 = \max_{s\in\mathcal{S}^{[\mathcal{H}]}} n_0(s)$, such that

$$\inf_{n\geq 1} \min_{s\in\mathcal{S}^{[\mathcal{H}]}} \pi^{\Theta^{(n)}}(a^*(s)|s) = \min_{1\leq n\leq n_0} \min_{s\in\mathcal{S}^{[\mathcal{H}]}} \pi^{\Theta^{(n)}}(a^*(s)|s) > 0.$$

Due to the positiveness of the softmax parametrization the assertion follows.
Claim 1. We first prove (i). Let $\Theta^{(n)} \in \mathcal{R}(s)$ and $a \neq a^*(s)$. Then $\Theta^{(n+1)} \in \mathcal{R}_3(s)$ by definition of $\mathcal{R}_3(s)$. To see that $\Theta^{(n+1)} \in \mathcal{R}_2(s)$ assume that s belongs to the epoch h such that

$$\begin{aligned}
Q^{\pi^{\Theta^{(n+1)}}}(s, a^*(s)) &= Q_h^{\pi^{\Theta^{(n+1)}}}(s, a^*(s)) \\
&= Q_h^{\pi^{\Theta^{(n)}}}(s, a^*(s)) + Q_h^{\pi^{\Theta^{(n+1)}}}(s, a^*(s)) - Q_h^{\pi^{\Theta^{(n)}}}(s, a^*(s)) \\
&= Q_h^{\pi^{\Theta^{(n)}}}(s, a^*(s)) + r(s, a^*(s)) + \sum_{s'\mathcal{S}^{[\mathcal{H}]}} p(s'|s, a^*(s)) V_{h+1}^{\pi^{\Theta^{(n+1)}}}(s') \\
&\quad - r(s, a^*(s)) - \sum_{s'\mathcal{S}^{[\mathcal{H}]}} p(s'|s, a^*(s)) V_{h+1}^{\pi^{\Theta^{(n)}}}(s')
\end{aligned}$$

$$
\begin{aligned}
&= Q_h^{\pi^{\Theta^{(n)}}}(s, a^*(s)) + \sum_{s' \mathcal{S}^{[\mathcal{H}]}} p(s'|s, a^*(s)) \Big(V_{h+1}^{\pi^{\Theta^{(n+1)}}}(s') - V_{h+1}^{\pi^{\Theta^{(n)}}}(s') \Big) \\
&\geq Q_h^{\pi^{\Theta^{(n)}}}(s, a^*(s)) = Q^{\pi^{\Theta^{(n)}}}(s, a^*(s)) \\
&\geq Q^{\infty}(s, a^*(s)) - \frac{\Delta^*(s)}{2},
\end{aligned}
$$

where the first inequality is due to monotonicity of $V^{\pi^{\Theta^{(n+1)}}}(s')$ in n for every $s' \in \mathcal{S}^{[\mathcal{H}]}$ and the last inequality follows from $\Theta^{(n)} \in \mathcal{R}_2(s)$.
Next we show $\Theta^{(n+1)} \in \mathcal{R}_1(s)$. Therefore we first show that

$$
Q^{\pi^{\Theta^{(n)}}}(s, a^*(s)) - Q^{\pi^{\Theta^{(n)}}}(s, a) \geq \frac{\Delta^*(s)}{2}, \tag{4.7}
$$

for all $a \neq a^*(s)$. This holds true, because

$$
\begin{aligned}
&Q^{\pi^{\Theta^{(n)}}}(s, a^*(s)) - Q^{\pi^{\Theta^{(n)}}}(s, a) \\
&= Q^{\pi^{\Theta^{(n)}}}(s, a^*(s)) - Q^{\infty}(s, a^*(s)) + Q^{\infty}(s, a^*(s)) - Q^{\pi^{\Theta^{(n)}}}(s, a) \\
&\geq -\frac{\Delta^*(s)}{2} + Q^{\infty}(s, a^*(s)) - Q^{\infty}(s, a) + Q^{\infty}(s, a) - Q^{\pi^{\Theta^{(n)}}}(s, a) \\
&\geq -\frac{\Delta^*(s)}{2} + \Delta^*(s) + \sum_{s' \in \mathcal{S}^{[\mathcal{H}]}} p(s'|s, a)(V^{\infty}(s') - V^{\pi^{\Theta^{(n)}}}(s')) \\
&\geq \frac{\Delta^*(s)}{2}.
\end{aligned}
$$

The first inequality follows from $\Theta^{(n)} \in \mathcal{R}_2(s)$, second by the definition of $\Delta^*(s)$ and the last from mononicity of $V^{\pi^{\Theta^{(n)}}}(s')$ for every s' and V^{∞} beeing the limit. Using Lemma 4.9 we obtain for any $a \neq a^*(s)$ that

$$
\begin{aligned}
&\frac{\partial J(\Theta^{(n)})}{\partial \theta(s, a^*(s))} \geq \frac{\partial J(\Theta^{(n)})}{\partial \theta(s, a)} \\
\Leftrightarrow\ &\pi^{\Theta^{(n)}}(a^*(s)|s) \big(Q^{\pi^{\Theta^{(n)}}}(s, a^*(s)) - V^{\pi^{\Theta^{(n)}}}(s) \big) \geq \pi^{\Theta^{(n)}}(a|s) \big(Q^{\pi^{\Theta^{(n)}}}(s, a) - V^{\pi^{\Theta^{(n)}}}(s) \big).
\end{aligned} \tag{4.8}
$$

We divide into two cases:

a) $\pi^{\Theta^{(n)}}(a^*(s)|s) \geq \pi^{\Theta^{(n)}}(a|s)$,

b) $\pi^{\Theta^{(n)}}(a^*(s)|s) < \pi^{\Theta^{(n)}}(a|s)$.

In a) the assumption $\pi^{\Theta^{(n)}}(a^*(s)|s) \geq \pi^{\Theta^{(n)}}(a|s)$ implies $\theta^{(n)}(s, a^*(s)) \geq \theta^{(n)}(s, a)$. Thus,

$$
\begin{aligned}
\theta^{(n+1)}(s, a^*(s)) &= \theta^{(n)}(s, a^*(s)) + \eta \frac{\partial J(\Theta^{(n)})}{\partial \theta^{(n)}(s, a^*(s))} \\
&\geq \theta^{(n)}(s, a) + \eta \frac{\partial J(\Theta^{(n)})}{\partial \theta^{(n)}(s, a)} \\
&= \theta^{(n+1)}(s, a),
\end{aligned}
$$

which implies $\pi^{\Theta^{(n+1)}}(a^*(s)|s) \geq \pi^{\Theta^{(n+1)}}(a|s)$. Moreover, we have

$$Q^{\pi^{\Theta^{(n+1)}}}(s,a^*(s)) - Q^{\pi^{\Theta^{(n+1)}}}(s,a) \geq \frac{\Delta^*(s)}{2} \geq 0,$$
$$Q^{\pi^{\Theta^{(n+1)}}}(s,a^*(s)) - V^{\pi^{\Theta^{(n+1)}}}(s) \geq Q^{\pi^{\Theta^{(n+1)}}}(s,a) - V^{\pi^{\Theta^{(n+1)}}}(s).$$

Thus, both together yields

$$\pi^{\Theta^{(n+1)}}(a^*(s)|s)\big(Q^{\pi^{\Theta^{(n+1)}}}(s,a^*(s)) - V^{\pi^{\Theta^{(n+1)}}}(s)\big) \geq \pi^{\Theta^{(n+1)}}(a|s)\big(Q^{\pi^{\Theta^{(n+1)}}}(s,a) - V^{\pi^{\Theta^{(n+1)}}}(s)\big),$$

which is by equation (4.8) equivalent to

$$\frac{\partial J(\Theta^{(n+1)})}{\partial \theta^{(n+1)}(s,a^*(s))} \geq \frac{\partial J(\Theta^{(n+1)})}{\partial \theta^{(n+1)}(s,a)}.$$

Hence, $\Theta^{(n+1)} \in \mathcal{R}_1(s)$.
In b) assume now that $\pi^{\Theta^{(n)}}(a^*(s)|s) < \pi^{\Theta^{(n)}}(a|s)$. As $\Theta^{(n)} \in \mathcal{R}_1(s)$, equation (4.8) is also true in this case and rearranging of terms gives

$$\begin{aligned}
&\frac{\partial J(\Theta^{(n)})}{\partial \theta^{(n)}(s,a^*(s))} \geq \frac{\partial J(\Theta^{(n)})}{\partial \theta^{(n)}(s,a)} \\
\Leftrightarrow\ & Q^{\pi^{\Theta^{(n)}}}(s,a^*(s)) - Q^{\pi^{\Theta^{(n)}}}(s,a) \geq \Big(1 - \frac{\pi^{\Theta^{(n)}}(a^*(s)|s)}{\pi^{\Theta^{(n)}}(a|s)}\Big)\big(Q^{\pi^{\Theta^{(n)}}}(s,a^*(s)) - V^{\pi^{\Theta^{(n)}}}(s)\big) \\
\Leftrightarrow\ & Q^{\pi^{\Theta^{(n)}}}(s,a^*(s)) - Q^{\pi^{\Theta^{(n)}}}(s,a) \geq \big(1 - \exp(\theta^{(n)}(s,a^*(s)) - \theta^{(n)}(s,a)\big)\big(Q^{\pi^{\Theta^{(n)}}}(s,a^*(s)) - V^{\pi^{\Theta^{(n)}}}(s)\big).
\end{aligned} \tag{4.9}$$

Note next that by $\Theta^{(n)} \in \mathcal{R}_1(s)$ and definition of $\mathcal{R}_1(s)$ we have

$$\begin{aligned}
&\theta^{(n+1)}(s,a^*(s)) - \theta^{(n+1)}(s,a) \\
&= \theta^{(n)}(s,a^*(s)) + \eta\frac{\partial J(\Theta^{(n)})}{\partial \theta^{(n)}(s,a^*(s))} - \theta^{(n)}(s,a) - \eta\frac{\partial J(\Theta^{(n)})}{\partial \theta^{(n)}(s,a)} \\
&\geq \theta^{(n)}(s,a^*(s)) - \theta^{(n)}(s,a)
\end{aligned}$$

and is follows $\big(1-\exp(\theta^{(n+1)}(s,a^*(s)) - \theta^{(n+1)}(s,a))\big) \leq \big(1-\exp(\theta^{(n)}(s,a^*(s)) - \theta^{(n)}(s,a))\big) < 1$ by assumption b). We already know $\Theta^{(n+1)} \in \mathcal{R}_3(s)$ and therefore $V^{\pi^{\Theta^{(n+1)}}}(s) \geq Q^{\pi^{\Theta^{(n+1)}}}(s,a^*(s)) - \frac{\Delta^*(s)}{2}$. This leads to

$$Q^{\pi^{\Theta^{(n+1)}}}(s,a^*(s)) - V^{\pi^{\Theta^{(n+1)}}}(s) \leq \frac{\Delta^*(s)}{2} \leq Q^{\pi^{\Theta^{(n+1)}}}(s,a^*(s)) - Q^{\pi^{\Theta^{(n+1)}}}(s,a),$$

where the last inequality is due to equation (4.7). Combining everything leads to

$$\begin{aligned}
&\big(1 - \exp(\theta^{(n+1)}(s,a^*(s)) - \theta^{(n+1)}(s,a))\big)\Big[Q^{\pi^{\Theta^{(n+1)}}}(s,a^*(s)) - V^{\pi^{\Theta^{(n+1)}}}(s)\Big] \\
&\leq Q^{\pi^{\Theta^{(n+1)}}}(s,a^*(s)) - Q^{\pi^{\Theta^{(n+1)}}}(s,a),
\end{aligned}$$

which is by equation (4.9) equivalent to $\Theta^{(n+1)} \in \mathcal{R}_1(s)$.
Now we come to Claim (ii).

$$\begin{aligned}
&\pi^{\Theta^{(n+1)}}(a^*(s)|s) \\
&= \frac{\exp(\theta^{(n+1)}(s,a^*(s)))}{\sum_{a\in\mathcal{A}} \exp(\theta^{(n+1)}(s,a))} \\
&= \frac{\exp(\theta^{(n)}(s,a^*(s)) + \eta \frac{\partial J(\Theta^{(n)})}{\partial \theta^{(n)}(s,a^*(s))})}{\sum_{a\in\mathcal{A}} \exp(\theta^{(n)}(s,a) + \eta \frac{\partial J(\Theta^{(n)})}{\partial \theta^{(n)}(s,a)})} \\
&\geq \frac{\exp(\theta^{(n)}(s,a^*(s)))\exp(\eta \frac{\partial J(\Theta^{(n)})}{\partial \theta^{(n)}(s,a^*(s))})}{\sum_{a\in\mathcal{A}} \exp(\theta^{(n)}(s,a))\exp(\eta \frac{\partial J(\Theta^{(n)})}{\partial \theta^{(n)}(s,a^*(s))})} \\
&= \pi^{\Theta^{(n)}}(a^*(s)|s),
\end{aligned}$$

where the inequality follows by $\Theta^{(n)} \in \mathcal{R}_1(s)$.
Claim 2. Assume $\Theta \in \mathcal{N}_c(s) \cap \mathcal{R}_2(s) \cap \mathcal{R}_3(s)$ and divide again in two cases. If *a)* $\pi^\Theta(a^*(s)|s) \geq \max_{a\in\mathcal{A}} \pi^\Theta(a|s)$, then for all $a \neq a^*(s)$ we have

$$\begin{aligned}
&\frac{\partial J(\Theta)}{\partial \theta(s,a^*(s))} \\
&= \widetilde{\rho}_\mu^{\pi^\Theta}(s)\pi^\Theta(a^*(s)|s)A^{\pi^\Theta}(s,a^*(s)) \\
&\geq \widetilde{\rho}_\mu^{\pi^\Theta}(s)\pi^\Theta(a|s)A^{\pi^\Theta}(s,a) \\
&= \frac{\partial J(\Theta)}{\partial \theta(s,a)}.
\end{aligned}$$

Where the inequality follows from $A^{\pi^\Theta}(s,a^*(s)) - A^{\pi^\Theta}(s,a) = Q^{\pi^\Theta}(s,a^*(s)) - Q^{\pi^\Theta}(s,a) \geq \frac{\Delta^*(s)}{2} > 0$ by equation (4.7). Hence, $\Theta \in \mathcal{R}_1(s)$.
The case *b)* where $\pi^\Theta(a^*(s)|s) < \max_{a\in\mathcal{A}} \pi^\Theta(a|s)$ is not possible for $\Theta \in \mathcal{N}_c(s)$. Assume there exists $a \neq a^*(s)$ such that $\pi^\Theta(a^*(s)|s) < \pi^\Theta(a|s)$. Then

$$\pi^\Theta(a^*(s)|s) + \pi^\Theta(a|s) > \frac{2c(s)}{c(s)+1} = \frac{\frac{2|\mathcal{A}|HR^*}{\Delta^*(s)} - 2}{\frac{|\mathcal{A}|HR^*}{\Delta^*(s)}} = 2 - \frac{2\Delta^*(s)}{|\mathcal{A}|HR^*} \geq 2 - \frac{2}{|\mathcal{A}|} \geq 1,$$

because $\Delta^*(s) \leq HR^*$ by definition and $|\mathcal{A}| \geq 2$. This is a contradiction as π^Θ is a probability distribution and Claim 2 is proven.
Claim 3. By the asymptotic convergence in Theorem A.1, we have that $\pi^{\Theta^{(n)}}(a^*(s)|s) \to 1$ for $n \to \infty$. Thus, there exists an $N_0(s) > 0$, such that $\pi^{\Theta^{(n)}}(a^*(s)|s) \geq \frac{c(s)}{c(s)+1}$ for all $n \geq N_0(s)$, i.e. $\Theta^{(n)} \in N_c(s)$ for all $n \geq N_0(s)$.
Furthermore, as $Q^{\pi^{\Theta^{(n)}}}(s,a^*(s)) \to Q^\infty(s,a^*(s))$ for $n \to \infty$ there exists $N_1(s)$ such that $\Theta^{(n)} \in \mathcal{R}_2(s)$ for all $n \geq N_1(s)$.
Moreover, as $Q^{\pi^{\Theta^{(n)}}}(s,a^*(s)) \to Q^\infty(s,a^*(s)) = V^\infty(s)$ and $V^{\pi^{\Theta^{(n)}}}(s) \to V^\infty(s)$ for $n \to \infty$ there

exists $N_2(s)$ such that $\Theta^{(n)} \in \mathcal{R}_3(s)$ for all $n \geq N_2(s)$.
We choose $n_0(s) = \max\{N_0(s), N_1(s), N_2(s)\}$ which proves Claim 3.

■

Global convergence. Combining smoothness and the gradient domination property results in the following global convergence result.

THEOREM 4.15. *Under Assumption 4.12, let μ be a probability measure such that $\mu(s) > 0$ for all $s \in \mathcal{S}$, let $\alpha = \frac{1}{5H^2R^*}$ and consider the sequence $(\Theta^{(n)})$ generated by FT-SimPG (Algorithm 4) under softmax parametrization with arbitrary initialization $\Theta^{(1)}$. For $\epsilon > 0$ choose the number of training steps as $N = \frac{10H^5R^*|\mathcal{S}|}{c^2\epsilon}\left\|\frac{d_\mu^{\pi^*}}{\mu}\right\|_\infty^2$. Then it holds that*

$$V_0^*(\mu) - V_0^{\pi^{\Theta^{(N)}}}(\mu) \leq \epsilon.$$

Proof. We will show that

$$J^*(\mu) - J(\Theta^{(n)}, \mu) = V_0^*(\mu) - V_0^{\pi^{\Theta^{(n)}}}(\mu) \leq \frac{10H^5R^*|\mathcal{S}|}{c^2 n}\left\|\frac{d_\mu^{\pi^*}}{\mu}\right\|_\infty^2,$$

then the claim follows immediately from this.

We apply Theorem 2.11 with $f = J(\cdot, \mu)$, $\alpha = 5H^2R^*$ and $b = \frac{c^2}{|\mathcal{S}|H^3}\|\frac{d_\mu^{\pi^*}}{\mu}\|^{-2}$. Note for b the evaluations in Remark 4.13.
It remains to check that for any $\Theta^{(1)}$ we have

$$J^*(\mu) - J(\Theta^{(1)}, \mu) \leq \frac{2H^2R^*5H^2H|\mathcal{S}|}{c^2}\left\|\frac{d_\mu^{\pi^*}}{\mu}\right\|_\infty^2.$$

This is directly given by the bounded rewards in Assumption 4.5 and the fact that $c < 1$ (π is a probability kernel) and $\left\|\frac{d_\mu^{\pi^*}}{\mu}\right\|_\infty^2 > 1$. Thus, we yield the claim

$$J^*(\mu) - J(\Theta^{(n)}, \mu) \leq \frac{10H^5R^*|\mathcal{S}|}{c^2 n}\left\|\frac{d_\mu^{\pi^*}}{\mu}\right\|_\infty^2.$$

■

Remark 4.16. One can compare this result to Theorem 3.21 for discounted infinite-time horizon MDPs. A discounted MDP can be seen as an undiscounted MDP stopped at an independent geometric random variable with mean $(1-\gamma)^{-1}$ (cf. Remark 3.17). Thus, it comes as no surprise that algorithms with deterministic absorption time H have analogous estimates with H instead of $(1-\gamma)^{-1}$.
In the discounted setting we obtained the factor $(1-\gamma)^{-4}$, where a power of 2 is due to the smoothness constant and a power of 2 is due to the distribution mismatch coefficient from the gradient domination property. Comparing to our results for FT-SimPG, the smoothness of order H^2R^* leads to a H^2 in the convergence rate, then the distribution mismatch coefficient adds another H^2 and the additional H comes from the enlarged state space, as the cardinality of the enlarged state space under Assumption 4.12 is $|\mathcal{S}^{[\mathcal{H}]}| = |\mathcal{S}|H$.
Moreover, it cannot be proven that c is independent of H. We omitted this dependency when we compare to the discounted case because the model dependent constant there could also depend on γ in the same sense.

4.2.2 *Finite-time Dynamic Policy Gradient*

We now come to the first main contribution of this thesis, an improved convergence bound for the FT-DynPG algorithm. The optimization objectives J_h are defined in equation (4.5). The structure of proving convergence is as follows. For each fixed $h \in \mathcal{H}$ we provide global convergence given that the policy after h is fixed and denoted by $\widetilde{\pi}$. After having established bounds for each decision epoch, we apply backwards induction to derive complexity bounds on the total error accumulated over all decision epochs. The L-smoothness for different J_h is then reflected in different training steps for different epochs. The backwards induction setting can be described as a nested sequence of contextual bandits (one-step MDPs) and thus, can be analysed using results from the discounted setting by choosing $\gamma = 0$.

Smoothness. Parallel to the simultaneous approach we start with the explicit derivative of the objective under softmax parametrization.

LEMMA 4.17. *For fix $h \in \mathcal{H}$, the partial derivative of the objective under softmax parametrization defined in equation* (4.5) *is given by:*

$$\frac{\partial J_h(\theta_h, \widetilde{\pi}_{(h+1)}, \mu_h)}{\partial \theta_h(s,a)} = \mu_h(s)\pi^{\theta_h}(a|s)A_h^{(\pi^{\theta_h}, \widetilde{\pi}_{(h+1)})}(s,a),$$

for every $s \in \mathcal{S}_h$ and $a \in \mathcal{A}_s$

Proof. By the policy gradient Theorem 4.4,

$$\begin{aligned}
\nabla J_h(\theta_h, \widetilde{\pi}_{(h+1)}, \mu_h) &= \nabla \mathbb{E}_{s\sim\mu_h}[J_h(\theta_h, \widetilde{\pi}_{(h+1)}, \delta_s)] \\
&= \sum_{s\in\mathcal{S}} \mu_h(s) \nabla J_h(\theta_h, \widetilde{\pi}_{(h+1)}, \delta_s) \\
&= \sum_{s\in\mathcal{S}} \mu_h(s) \mathbb{E}_{S_h=s, A_h\sim\pi^{\theta_h}(\cdot|s)}[\nabla \log(\pi^{\theta_h}(A_h|S_h)) Q_h^{\widetilde{\pi}}(S_h, A_h)].
\end{aligned}$$

Next we plug in the derivative of the softmax parametrization, i.e. equation (3.13), and obtain

$$\begin{aligned}
&\nabla J_h(\theta_h, \widetilde{\pi}_{(h+1)}, \mu_h) \\
&= \sum_{s\in\mathcal{S}} \mu_h(s) \mathbb{E}_{S_h=s, A_h\sim\pi^{\theta_h}(\cdot|s)} \left[\left(\mathbf{1}_{\{S_h=s'\}} (\mathbf{1}_{\{A_h=a'\}} - \pi^{\theta_h}(a'|s')) \right)_{s'\in\mathcal{S}_h, a'\in\mathcal{A}_{s'}} Q_h^{\widetilde{\pi}}(S_h, A_h) \right] \\
&= \left(\sum_{s\in\mathcal{S}} \mu_h(s) \sum_{a\in\mathcal{A}_s} \pi^{\theta_h}(a|s) \mathbf{1}_{\{s=s'\}} (\mathbf{1}_{\{a=a'\}} - \pi^{\theta_h}(a'|s')) Q_h^{\widetilde{\pi}}(s,a) \right)_{s'\in\mathcal{S}_h, a'\in\mathcal{A}_{s'}} \\
&= \left(\mu_h(s')\pi^{\theta_h}(a'|s') Q_h^{\widetilde{\pi}}(s',a') - \mu_h(s')\pi^{\theta_h}(a'|s') \sum_{a\in\mathcal{A}_{s'}} \pi^{\theta}(a|s') Q_h^{\widetilde{\pi}}(s',a) \right)_{s'\in\mathcal{S}_h, a'\in\mathcal{A}_{s'}} \\
&= \left(\mu_h(s')\pi^{\theta_h}(a'|s') (Q_h^{\widetilde{\pi}}(s',a') - V_h^{(\pi^{\theta_h}, \widetilde{\pi}_{(h+1)})}(s')) \right)_{s'\in\mathcal{S}_h, a'\in\mathcal{A}_{s'}} \\
&= \left(\mu_h(s')\pi^{\theta_h}(a'|s') A_h^{(\pi^{\theta_h}, \widetilde{\pi}_{(h+1)})}(s',a') \right)_{s'\in\mathcal{S}_h, a'\in\mathcal{A}_{s'}},
\end{aligned}$$

where we used that $\sum_{a\in\mathcal{A}_{s'}} \pi^{\theta_h}(a|s') Q_h^{\widetilde{\pi}}(s',a) = V_h^{(\pi^{\theta}, \widetilde{\pi}_{(h+1)})}(s') = J_h(\theta_h, \widetilde{\pi}_{(h+1)}, \delta_{s'})$. ∎

We deduce directly, that each objective J_h from equation (4.5) is a smooth function in θ_h.

LEMMA 4.18. *Let $h \in \mathcal{H}$, then the objective $J_h(\theta_h, \widetilde{\mathbb{m}}_{(h+1)}, \mu_h)$ under softmax parametrization from equation (4.5) is smooth in θ_h with parameter $L_h = 2(H-h)R^*$.*

Proof. Note that we can interpret the objective function $J_h(\theta_h, \widetilde{\mathbb{m}}_{(h+1)}, \mu_h)$ as a value function of a one-step discounted MDP with $\gamma = 0$ and bounded rewards between $[0, R^*(H-h)]$. Hence, we can use Yuan, Gower, and Lazaric [YGL22, Lem 4.4 and 4.8] to obtain that the softmax policy π^{θ_h} fulfills the desired properties with

$$\mathbb{E}_{A\sim\pi^{\theta_h}}\left[||\nabla \log \pi^{\theta_h}(A|s)||_2^2\right] \leq 1 - \frac{1}{|\mathcal{A}_s|} \leq 1 \quad \forall s \in \mathcal{S}$$
$$\mathbb{E}_{A\sim\pi^{\theta_h}}\left[||\nabla^2 \log \pi^{\theta_h}(A|s)||_2\right] \leq 1,$$

which leads to a smoothness constant $L_h = 2(H-h)R^*$ for the objective function J_h. ∎

It is crucial to keep in mind that classical theory from non-convex optimization tells us that less smooth (large L) functions must be trained with more gradient steps as they require a smaller step size (cf. Theorem 2.10). It becomes clear that the FT-DynPG algorithm should spend less training effort on later epochs (earlier in the algorithm) and more training effort on earlier epochs (later in the algorithm). In fact, we make use of this observation by applying backwards induction in order to improve the convergence behavior depending on H (see Theorem 4.24).

Weak gradient domination. In the following let $\mathbb{m}^* = (\pi_h^*)_{h=0}^{H-1} \in \Pi^H$ be a fixed but arbitrary deterministic optimal policy, such that $a_h^*(s) = \operatorname{argmax}_{a\in\mathcal{A}_s} \pi_h^*(a|s)$ is an arbitrary fixed best action in state $s \in \mathcal{S}_h$ (compare to Remark 3.34). We derive the following non-uniform gradient domination property for any $h \in \mathcal{H}$.

LEMMA 4.19. *Under softmax parametrization, it holds that*

$$\|\nabla J_h(\theta_h, \widetilde{\mathbb{m}}_{(h+1)}, \mu_h)\|_2 \geq \min_{s\in\mathcal{S}_h} \pi^{\theta_h}(a_h^*(s)|s)(J_h^*(\widetilde{\mathbb{m}}_{(h+1)}, \mu_h) - J_h(\theta_h, \widetilde{\mathbb{m}}_{(h+1)}, \mu_h)).$$

Proof. First note that by the definition of π_h^*, we have $J_h^*(\widetilde{\mathbb{m}}_{(h+1)}, \mu_h) = V_h^{(\pi_h^*, \widetilde{\mathbb{m}}_{(h+1)})}(\mu_h)$, because the tabular softmax parametrization can approximate any deterministic policy arbitrarily well. Using the performance difference lemma in the dynamic setting from Corollary 4.7 and the derivative of the objective given in Lemma 4.17, we obtain

$$\begin{aligned}
&\left\|\frac{\partial J_h(\theta_h, \widetilde{\mathbb{m}}_{(h+1)}, \mu_h)}{\partial \theta_h}\right\|_2 \\
&= \left\| \sum_{s\in\mathcal{S}_h} \mu_h(s) \frac{\partial J_h(\theta_h, \widetilde{\mathbb{m}}_{(h+1)}, \delta_s)}{\partial \theta_h}\right\|_2 \\
&= \Big[\sum_{s'\in\mathcal{S}_h} \sum_{a'\in\mathcal{A}_{s'}} \Big(\sum_{s\in\mathcal{S}_h} \mu_h(s) \frac{\partial J_h(\theta_h, \widetilde{\mathbb{m}}_{(h+1)}, \delta_s)}{\partial \theta_h(s', a')}\Big)^2 \Big]^{\frac{1}{2}} \\
&\geq \sum_{s\in\mathcal{S}_h} \mu_h(s) \Big| \frac{\partial J_h(\theta_h, \widetilde{\mathbb{m}}_{(h+1)}, \delta_s)}{\partial \theta_h(s, a_h^*(s))}\Big|
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{s\in\mathcal{S}_h} \mu_h(s)\pi^{\theta_h}(a_h^*(s)|s) A_h^{(\pi^{\theta_h},\widetilde{\pi}_{(h+1)})}(s, a_h^*(s)) \\
&= \sum_{s\in\mathcal{S}_h} \mu_h(s)\pi^{\theta_h}(a_h^*(s)|s)\Big(J_h^*(\widetilde{\pi}_{(h+1)}, \delta_s) - J_h(\theta_h, \widetilde{\pi}_{(h+1)}, \delta_s)\Big) \\
&\geq \min_{s\in\mathcal{S}_h} \pi^{\theta_h}(a_h^*(s)|s)\Big(J_h^*(\widetilde{\pi}_{(h+1)}, \mu_h) - J_h(\theta_h, \widetilde{\pi}_{(h+1)}, \mu_h)\Big).
\end{aligned}
$$

The first inequality is due to the non-negativity of all other terms, and we just drop them. ∎

The main challenge is again to bound $\min_{s\in\mathcal{S}} \pi^{\theta_h}(a_h^*(s)|s)$ from below uniformly in θ_h appearing in the gradient ascent updates from Algorithm 5. In this setting the required asymptotic convergence follows directly from the one-step MDP viewpoint using $\gamma = 0$ obtained in Agarwal et al. [Aga+21, Thm 5] and it holds $c_h = \inf_{n\in\mathbb{N}} \min_{s\in\mathcal{S}_h} \pi^{\theta_h^{(n)}}(a_h^*(s)|s) > 0$.

LEMMA 4.20. *Let μ_h be a probability measure such that $\mu_h(s) > 0$ for all $s \in \mathcal{S}_h$ and let $0 < \alpha_h \leq \frac{1}{2(H-h)R^*}$. Consider the sequence $(\theta_h^{(n)})_{n\in\mathbb{N}}$ generated by FT-DynPG (Algorithm 5) under softmax parametrization for arbitrary $\theta_h^{(1)} \in \mathcal{R}^{d_h}$ and future policies $\widetilde{\pi}$. Then, $c_h = \inf_{n\in\mathbb{N}} \min_{s\in\mathcal{S}_h} \pi^{\theta_h^{(n)}}(a_h^*(s)|s) > 0$.*

Proof. The idea of the proof is based on Mei et al. [Mei+20, Lemma 5] for bandits and extended to the contextual bandit case.
Throughout the proof we change notation as follows:

- As we consider a fixed time point h we will only write θ_n instead of $\theta_h^{(n)}$.
- We denote the objective function by $J_h(\theta)$ instead of $J_h(\theta, \widetilde{\pi}_{(h+1)}, \mu_h)$ for a fixed policy $\widetilde{\pi}$ and start distribution μ_h. Furthermore, we will just write J_h^* instead of $J_h^*(\widetilde{\pi}_{(h+1)}, \mu_h)$.
- We will write $J_{h,s}(\theta)$ for the objective function which starts almost surly in $s \in \mathcal{S}_h$, i.e. $J_{h,s}(\theta) = J_h(\theta, \widetilde{\pi}_{(h+1)}, \delta_s)$.

First note that

$$
J_{h,s}(\theta) = \sum_{a\in\mathcal{A}_s} \pi^\theta(a|s) Q_h^{\widetilde{\pi}}(s, a),
$$

where $Q_h^{\widetilde{\pi}}(s, a)$ is independent of θ. We will drop the subscript $\widetilde{\pi}$ in Q_h for the rest of the proof and define for all $s \in \mathcal{S}_h$,

$$
\Delta^*(s) = Q_h(s, a_h^*(s)) - \max_{a\neq a_h^*(s)} Q_h(s, a) > 0, \quad \text{and} \quad \Delta^* = \min_{s\in\mathcal{S}_h} \Delta^*(s) > 0.
$$

Consider the following sets

$$
\begin{aligned}
\mathcal{R}_h^1(s) &= \{\theta : \frac{\partial J_{h,s}(\theta)}{\partial\theta(s, a_h^*(s))} \geq \frac{\partial J_{h,s}(\theta)}{\partial\theta(s, a)} \forall a \neq a_h^*(s)\} \\
\mathcal{R}_h^2(s) &= \{\theta : \pi^\theta(a_h^*(s)|s) \geq \pi^\theta(a|s) \forall a \neq a_h^*(s)\}
\end{aligned}
$$

$$\mathcal{N}_h(s) = \{\theta : \pi^\theta(a_h^*(s)|s) \geq \frac{c_h(s)}{c_h(s)+1}\},$$

for $c_h(s) = \frac{|\mathcal{A}|(H-h)R^*}{\Delta_h^*(s)} - 1$ and $\Delta_h^*(s) = Q_h(s, a^*(s)) - \max_{a \neq a^*} Q_h(s,a)$. Then consider the following Claims:

1. $\theta_n \in \mathcal{R}_h^1(s) \Rightarrow \theta_{n+1} \in \mathcal{R}_h^1(s)$,
2. If $\theta_n \in \mathcal{R}_h^1(s)$, then $\pi^{\theta_{n+1}}(a_h^*(s)|s) \geq \pi^{\theta_n}(a_h^*(s)|s)$,
3. $\mathcal{N}_h(s) \subset \mathcal{R}_h^2(s) \subset \mathcal{R}_h^1(s)$.

Claim 1. Let $\theta_n \in \mathcal{R}_h^1(s)$ and $a \neq a_h^*(s)$. Using the derivative of the value function we obtain

$$\begin{aligned} &\frac{\partial J_{h,s}(\theta_n)}{\partial \theta(s, a_h^*(s))} \geq \frac{\partial J_{h,s}(\theta_n)}{\partial \theta(s,a)} \\ \Leftrightarrow\ &\pi^{\theta_n}(a_h^*(s)|s)\big(Q_h(s, a_h^*(s)) - J_{h,s}(\theta_n)\big) \geq \pi^{\theta_n}(a|s)\big(Q_h(s,a) - J_{h,s}(\theta_n)\big). \end{aligned} \tag{4.10}$$

We divide into two cases:

a) $\pi^{\theta_n}(a_h^*(s)|s) \geq \pi^{\theta_n}(a|s)$,

b) $\pi^{\theta_n}(a_h^*(s)|s) < \pi^{\theta_n}(a|s)$.

In a) the assumption $\pi^{\theta_n}(a_h^*(s)|s) \geq \pi^{\theta_n}(a|s)$ implies $\theta_n(s, a_h^*(s)) \geq \theta_n(s,a)$. Thus,

$$\begin{aligned} \theta_{n+1}(s, a_h^*(s)) &= \theta_n(s, a_h^*(s)) + \eta_h \mu_h(s) \frac{\partial J_{h,s}(\theta_n)}{\partial \theta_n(s, a_h^*(s))} \\ &\geq \theta_n(s,a) + \eta_h \mu_h(s) \frac{\partial J_{h,s}(\theta_n)}{\partial \theta_n(s,a)} \\ &= \theta_{n+1}(s,a), \end{aligned}$$

which implies $\pi^{\theta_{n+1}}(a_h^*(s)|s) \geq \pi^{\theta_{n+1}}(a|s)$. By the optimality of $a_h^*(s)$ we follow

$$\pi_t^{\theta_{n+1}}(a_h^*(s)|s)\big(Q_h(s, a_h^*(s)) - J_{h,s}(\theta_{n+1})\big) \geq \pi_t^{\theta_{n+1}}(a|s)\big(Q_h(s,a) - J_{h,s}(\theta_{n+1})\big),$$

which is by equation (4.10) equivalent to

$$\frac{\partial J_{h,s}(\theta_{n+1})}{\partial \theta_{n+1}(s, a_h^*(s))} \geq \frac{\partial J_{h,s}(\theta_{n+1})}{\partial \theta_{n+1}(s,a)}.$$

Hence, $\theta_{n+1} \in \mathcal{R}_h^1(s)$.
In b) assume now that $\pi^{\theta_n}(a_h^*(s)|s) < \pi^{\theta_n}(a|s)$. As $\theta_n \in \mathcal{R}_h^1(s)$, equation (4.10) is also true in this case and rearranging of terms gives

$$\begin{aligned} &\frac{\partial J_{h,s}(\theta_n)}{\partial \theta_n(s, a_h^*(s))} \geq \frac{\partial J_{h,s}(\theta_n)}{\partial \theta_n(s,a)} \\ \Leftrightarrow\ &Q_h(s, a_h^*(s)) - Q_h(s,a) \geq \Big(1 - \frac{\pi^{\theta_n}(a_h^*(s)|s)}{\pi^{\theta_n}(a|s)}\Big)\big(Q_h(s, a_h^*(s)) - J_{h,s}(\theta_n)\big) \\ \Leftrightarrow\ &Q_h(s, a_h^*(s)) - Q_h(s,a) \geq \big(1 - \exp(\theta_n(s, a_h^*(s)) - \theta_n(s,a)\big)\big(Q_h(s, a_h^*(s)) - J_{h,s}(\theta_n)\big). \end{aligned} \tag{4.11}$$

Note next that by $\theta^{(n)} \in \mathcal{R}_h^1(s)$ and definition of $\mathcal{R}_h^1(s)$ we have

$$\begin{aligned}
&\theta_{n+1}(s, a_h^*(s)) - \theta_{n+1}(s,a) \\
&= \theta_n(s, a_h^*(s)) + \eta_h \mu_h(s) \frac{\partial J_{h,s}(\theta_n)}{\partial \theta_n(s, a_h^*(s))} - \theta_n(s,a) - \eta_h \mu_h(s) \frac{\partial J_{h,s}(\theta_n)}{\partial \theta_n(s,a)} \\
&\geq \theta_n(s, a_h^*(s)) - \theta_n(s,a)
\end{aligned}$$

and is follows $\left(1 - \exp(\theta_{n+1}(s, a_h^*(s)) - \theta_{n+1}(s,a))\right) \leq \left(1 - \exp(\theta_n(s, a_h^*(s)) - \theta_n(s,a))\right) < 1$ by assumption b). By the ascent lemma for smooth functions we get monotonicity in the objective function, so

$$Q_h(s, a_h^*(s)) - J_{h,s}(\theta_{n+1}) \leq Q_h(s, a_h^*(s)) - J_{h,s}(\theta_n),$$

where the last inequality is due to the definition of $\Delta^*(s)$. Combining everything leads to

$$\begin{aligned}
&\left(1 - \exp(\theta_{n+1}(s, a_h^*(s)) - \theta_{n+1}(s,a))\right) \left[Q_h(s, a_h^*(s)) - J_{h,s}(\theta_{n+1})\right] \\
&\leq \left(1 - \exp(\theta_n(s, a_h^*(s)) - \theta_n(s,a))\right) \left[Q_h(s, a_h^*(s)) - J_{h,s}(\theta_n)\right] \\
&\leq Q_h(s, a_h^*(s)) - Q_h(s,a),
\end{aligned}$$

which is by equation (4.11) equivalent to $\theta_{n+1} \in \mathcal{R}_1(s)$.

Claim 2. If $\theta_n \in \mathcal{R}_h^1(s)$, then

$$\begin{aligned}
&\pi^{\theta_{n+1}}(a_h^*(s)|s) \\
&= \frac{\exp(\theta_{n+1}(s, a_h^*(s)))}{\sum_{a \in \mathcal{A}} \exp(\theta_{n+1}(s,a))} \\
&= \frac{\exp(\theta_n(s, a_h^*(s)) + \eta_h \mu_h(s) \frac{\partial J_{h,s}(\theta_n)}{\partial \theta_n(s, a_h^*(s))})}{\sum_{a \in \mathcal{A}_s} \exp(\theta_n(s,a) + \eta_h \mu_h(s) \frac{\partial J_{h,s}(\theta_n)}{\partial \theta_n(s,a)})} \\
&\geq \frac{\exp(\theta_n(s, a_h^*(s))) \exp(\eta_h \mu_h(s) \frac{\partial J_{h,s}(\theta_n)}{\partial \theta_n(s, a_h^*(s))})}{\sum_{a \in \mathcal{A}_s} \exp(\theta_n(s,a)) \exp(\eta_h \mu_h(s) \frac{\partial J_{h,s}(\theta_n)}{\partial \theta_n(s, a_h^*(s))})} \\
&= \pi^{\theta_n}(a_h^*(s)|s),
\end{aligned}$$

where the inequality follows by $\theta_n \in \mathcal{R}_h^1(s)$.

Claim 3. Let $\theta_n \in \mathcal{R}_h^2(s)$, then by the optimality of $a^*(s)$,

$$\pi^{\theta_n}(a^*(s)|s)(Q_h(s, a_h^*(s)) - J_{h,s}(\theta_n)) \geq \pi^{\theta_n}(a|s)(Q_h(s,a) - J_{h,s}(\theta_n)) \tag{4.12}$$

$$\Leftrightarrow \frac{\partial J_{h,s}(\theta_n)}{\partial \theta_n(s, a^*(s))} \geq \frac{\partial J_{h,s}(\theta_n)}{\partial \theta_n(s,a)}. \tag{4.13}$$

Hence, $\theta_n \in \mathcal{R}_h^1(s)$.
On the other hand, let $\theta_n \in \mathcal{N}_h(s)$, then assume there exists $a \neq a_h^*(s)$ such that $\pi^\theta(a_h^*(s)|s) < \pi^\theta(a|s)$. Then

$$\pi^\theta(a_h^*(s)|s) + \pi^\theta(a|s) > \frac{2c(s)}{c(s)+1} = \frac{\frac{2|\mathcal{A}|(H-h)R^*}{\Delta_h^*(s)} - 2}{\frac{|\mathcal{A}|(H-h)R^*}{\Delta_h^*(s)}} = 2 - \frac{2\Delta_h^*(s)}{|\mathcal{A}|(H-h)R^*} \geq 2 - \frac{2}{|\mathcal{A}|} \geq 1,$$

because $\Delta^*(s) \leq (H-h)R^*$ by definition and $|\mathcal{A}| \geq 2$. This is a contradiction as π^θ is a probability distribution and Claim 3 is proven.
To follow the claim of the lemma from the claims 1, 2 and 3, we need asymptotic convergence to the global optimum. This is given by Agarwal et al. [Aga+21, Theorem 5], since we can interpret the objective J_h as a one-step MDP with $\gamma = 0$. Then, assuring that the step size is smaller than one over the smoothness parameter is enough to use the same proof as provided in Agarwal et al. [Aga+21]. So, there exists a time $t_0 \geq 1$ such that $\theta \in \mathcal{N}_h(s)$ for all $s \in \mathcal{S}$. Finally,

$$\inf_n \min_s \pi^{\theta_n}(a^*(s)|s) = \min_{1 \leq n \leq t_0} \min_s \pi^{\theta_n}(a^*(s)|s) > 0.$$

■

There is another subtle advantage in the backwards induction point of view. The contextual bandit interpretation allows using refinements of estimates for the special case of contextual bandits. A slight generalization of work of Mei et al. [Mei+20] for stochastic bandits shows that the unpleasant unknown constants c_h simplify if the PG algorithm is uniformly initialized:

Proposition 4.21. *For fixed $h \in \mathcal{H}$, let μ_h be a probability measure such that $\mu_h(s) > 0$ for all $s \in \mathcal{S}_h$ and let $0 < \alpha_h \leq \frac{1}{2(H-h)R^*}$. Consider the sequence $(\theta_h^{(n)})_{n\in\mathbb{N}}$ generated in epoch h by FT-DynPG (Algorithm 5) under softmax parametrization for arbitrary future policies $\widetilde{\pi}$. Further, let $\theta_h^{(1)} \in \mathcal{R}^{d_h}$ be an initialization such that the initial policy is a uniform distribution, then $c_h = \inf_{n\in\mathbb{N}} \min_{s\in\mathcal{S}_h} \pi^{\theta_h^{(n)}}(a_h^*(s)|s) = \frac{1}{|\mathcal{A}|}$.*

Proof. From the proof of the previous Lemma 4.20, we obtain that a uniform initialization, implies that $\theta_h^{(1)} \in \mathcal{R}_h^2(s)$ for all $s \in \mathcal{S}_h$. Therefore, $\theta_h^{(n)} \in \mathcal{R}_h^1(s)$ for all $n \geq 1$ and from Claim 2 we have

$$c_h = \inf_n \min_s \pi^{\theta_h^{(n)}}(a^*(s)|s) = \min_s \pi^{\theta_h^{(1)}}(a^*(s)|s) = \frac{1}{|\mathcal{A}|}.$$

■

Remark 4.22. This property is in sharp contrast to the simultaneous approach, where to the best of our knowledge it is not known how to lower bound c explicitly. Comparing the proofs of $c > 0$ and $c_h > 0$ one can see that this advantage comes from the backward inductive approach and is due to fixed future policies which are not changing during training.

Global convergence. For fixed decision epoch h combining L-smoothness and weak gradient domination yields the following global convergence result for the FT-DynPG.

LEMMA 4.23. *For fixed $h \in \mathcal{H}$, let μ_h be a probability measure such that $\mu_h(s) > 0$ for all $s \in \mathcal{S}_h$, let $\alpha_h = \frac{1}{2(H-h)R^*}$ and consider the sequence $(\theta_h^{(n)})$ generated by FT-DynPG (Algorithm 5) under softmax parametrization with arbitrary initialization $\theta_h^{(1)}$ and future policies $\widetilde{\mathbb{m}}$. For $\epsilon > 0$ choose the number of training steps as $N_h = \frac{4(H-h)R^*}{c_h^2 \epsilon}$. Then, it holds that*

$$V_h^{(\pi_h^*, \widetilde{\mathbb{m}}_{(h+1)})}(\mu_h) - V_h^{(\pi^{\theta_h^{(N_h)}}, \widetilde{\mathbb{m}}_{(h+1)})}(\mu_h) \leq \epsilon$$

Moreover, if $\theta_h^{(1)}$ initializes the uniform distribution the constants c_h can be replaced by $\frac{1}{|\mathcal{A}|}$.

Proof. First, note that $V_h^{(\pi_h^*, \widetilde{\mathbb{m}}_{(h+1)})}(\mu_h) = J_h^*(\widetilde{\mathbb{m}}_{(h+1)}, \mu_h)$ and $V_h^{(\pi^{\theta_h^{(n)}}, \widetilde{\mathbb{m}}_{(h+1)})}(\mu_h) = J_h(\theta_h^{(n)}, \widetilde{\mathbb{m}}_{(h+1)}, \mu_h)$ by definition of J_h and choice of π_h^*. We will proof

$$J_h^*(\widetilde{\mathbb{m}}_{(h+1)}, \mu_h) - J_h(\theta_h^{(n)}, \widetilde{\mathbb{m}}_{(h+1)}, \mu_h) \leq \frac{4(H-h)R^*}{c_h^2 n},$$

Then the claim follows directly from this.
We use the same arguments as in the proof of Theorem 4.15 and apply Theorem 2.11, with objective function $f = J_h(\cdot, \widetilde{\mathbb{m}}_{(h+1)}, \mu_h)$ and step size $\alpha = \alpha_h$. By $L = \frac{1}{\alpha_h}$ smoothness and weak gradient domination with $b = c_h$ along the gradient trajectory, we only need to assure that

$$J_h^*(\widetilde{\mathbb{m}}_{(h+1)}, \mu_h) - J_h(\theta_h^{(1)}, \widetilde{\mathbb{m}}_{(h+1)}, \mu_h) \leq \frac{2}{\alpha b^2}.$$

It holds that

$$J_h^*(\widetilde{\mathbb{m}}_{(h+1)}, \mu_h) - J_h(\theta_h^{(1)}, \widetilde{\mathbb{m}}_{(h+1)}, \mu_h) \leq (H-h)R^* \leq \frac{4(H-h)R^*}{c_h^2} = \frac{2L_h}{c_h^2} = \frac{2}{\alpha b^2}$$

and the claim follows. ∎

The error bound depends on the time horizon up to the last time point, meaning intuitively that an optimal policy for earlier time points in the MDP (smaller h) is harder to achieve and requires a longer learning period then later time points (h near to H). We remark that the assumption on μ_h is not a sharp restriction and can be achieved by using a strictly positive start distribution μ on $\mathcal{S}_0$ followed by a uniformly distributed policy. Note that assuming a positive start distribution is common in the literature and Mei et al. [Mei+20] showed the necessity of this assumption. Accumulating errors over time we can now derive the analogous estimates to the simultaneous PG approach. We obtain a linear accumulation such that an $\frac{\epsilon}{H}$-error in each time point h results in an overall error of ϵ which appears naturally from the dynamic programming structure of the algorithm.

THEOREM 4.24. *For all $h \in \mathcal{H}$, let μ_h be probability measures such that $\mu_h(s) > 0$ for all $s \in \mathcal{S}_h$, let $\alpha_h = \frac{1}{2(H-h)R^*}$. For $\epsilon > 0$ choose the number of training steps as $N_h = \frac{4(H-h)HR^*}{c_h^2 \epsilon} \left\| \frac{1}{\mu_h} \right\|_\infty$. Then for the final policy generated by FT-DynPG (Algorithm 5) under softmax parametrization, $\widehat{\mathbb{m}}^* = (\pi^{\theta_0^{(N_0)}}, \dots, \pi^{\theta_{H-1}^{(N_{H-1})}})$, it holds for all $s \in \mathcal{S}_0$ that*

$$V_0^*(s) - V_0^{\widehat{\mathbb{m}}^*}(s) \leq \epsilon.$$

If $\theta_h^{(1)}$ initializes the uniform distribution the constants c_h can be replaced by $\frac{1}{|\mathcal{A}|}$.

Proof. First note that by our choice of the future policy $\widetilde{\mathbb{\pi}} = \widehat{\mathbb{\pi}}^*$ we have

$$J_h(\theta_h^{(N_h)}, \widetilde{\mathbb{\pi}}_{(h+1)}, \delta_s) = V_h^{\widehat{\mathbb{\pi}}^*}(s). \tag{4.14}$$

By Lemma 4.23 we obtain

$$J_h^*(\widetilde{\mathbb{\pi}}_{(h+1)}, \mu_h) - J_h(\theta_h^{(N_h)}, \widetilde{\mathbb{\pi}}_{(h+1)}, \mu_h) \leq \frac{4(H-h)R^*}{c_h^2 N_h}.$$

For every $s \in \mathcal{S}_h$,

$$\begin{aligned} J_h^*(\widetilde{\mathbb{\pi}}_{(h+1)}, \delta_s) - J_h(\theta_h^{(N_h)}, \widetilde{\mathbb{\pi}}_{(h+1)}, \delta_s) &= \sum_{s' \in \mathcal{S}_h} \mu_h(s') \frac{\delta_s(s')}{\mu_h(s')} J_h^*(\widetilde{\mathbb{\pi}}_{(h+1)}, \delta_s) - J_{h,s}(\theta_h^{(N_h)}, \widetilde{\mathbb{\pi}}_{(h+1)}, \delta_s) \\ &\leq \left\| \frac{1}{\mu_h} \right\|_\infty \left(J_h^*(\widetilde{\mathbb{\pi}}_{(h+1)}, \mu_h) - J_h(\theta_h^{(N_h)}, \widetilde{\mathbb{\pi}}_{(h+1)}, \mu_h) \right) \\ &\leq \frac{4(H-h)R^*}{c_h^2 N_h} \left\| \frac{1}{\mu_h} \right\|_\infty, \end{aligned} \tag{4.15}$$

where $\left\| \frac{1}{\mu_h} \right\|_\infty = \max_{s \in \mathcal{S}_h} \frac{1}{\mu_h(s)} > 0$ by assumption. As $N_h = \frac{4(H-h)HR^*}{c_h^2 \epsilon} \left\| \frac{1}{\mu_h} \right\|_\infty$, it holds that

$$J_h^*(\widetilde{\mathbb{\pi}}_{(h+1)}, \delta_s) - J_h(\theta_h^{(N_h)}, \widetilde{\mathbb{\pi}}_{(h+1)}, \delta_s) \leq \frac{\epsilon}{H} \tag{4.16}$$

for every $s \in \mathcal{S}_h$. For $h = H-1$ it follows directly by equation (4.14) and the specialty of the last time point that for all $s \in \mathcal{S}_{H-1}$,

$$V_{H-1}^*(s) - V_{H-1}^{\widehat{\mathbb{\pi}}^*}(s) = J_{H-1}^*(\delta_s) - J_{H-1}(\theta_{H-1}^{(N_{H-1})}, \delta_s) \leq \frac{\epsilon}{H}.$$

Note that the last epoch is independent of $\widetilde{\mathbb{\pi}}$. Assume now that for all $s \in \mathcal{S}_h$,

$$V_h^*(s) - V_h^{\widehat{\mathbb{\pi}}^*}(s) \leq \frac{\epsilon(H-h)}{H}.$$

Then it holds for all $s \in \mathcal{S}_{h-1}$ that,

$$\begin{aligned} J_{h-1}^*(\widetilde{\mathbb{\pi}}_{(h)}, \delta_s) &= \max_{a \in \mathcal{A}_s} \Big(r(s,a) + \sum_{s' \in \mathcal{S}_h} p(s'|s,a) V_h^*(s) - \sum_{s' \in \mathcal{S}_h} p(s'|s,a) (V_h^*(s) - V_h^{\widehat{\mathbb{\pi}}^*}(s)) \Big) \\ &\geq \max_{a \in \mathcal{A}_s} \Big(r(s,a) + \sum_{s' \in \mathcal{S}_h} p(s'|s,a) V_h^*(s) \Big) - \frac{\epsilon(H-h)}{H} \\ &= V_{h-1}^*(s) - \frac{\epsilon(H-h)}{H}, \end{aligned} \tag{4.17}$$

by the Bellman expectation equation for finite-time MDPs ([Put05]). We close the backward induction using equation (4.14) such that for all $s \in \mathcal{S}_{h-1}$,

$$\begin{aligned} V_{h-1}^*(s) - V_{h-1}^{\widehat{\mathbb{\pi}}^*}(s) &= V_{h-1}^*(s) - J_{h-1}^*(\widetilde{\mathbb{\pi}}_{(h)}, \delta_s) + J_{h-1}^*(\widetilde{\mathbb{\pi}}_{(h)}, \delta_s) - V_{h-1}^{\widehat{\mathbb{\pi}}^*}(s) \\ &\leq \frac{\epsilon(H-h)}{H} + \frac{\epsilon}{H} \\ &= \frac{\epsilon(H-(h-1))}{H}. \end{aligned} \tag{4.18}$$

Finally, it holds for $h = 0$ and all $s \in \mathcal{S}_0$ that

$$V_0^*(s) - V_0^{\widehat{\mathbb{\pi}}^*}(s) \leq \epsilon.$$

■

4.2.3 *Comparison of the algorithms*

Comparing the convergence rate for FT-DynPG in Theorem 4.24 to the convergence rate for FT-SimPG in Theorem 4.15, we first highlight that the constant c_h in the dynamic approach can be explicitly computed under uniform initialization. This has not yet been established in FT-SimPG (see Remark 4.22) and especially it cannot be guaranteed that c is independent of the time horizon. Second, we compare the overall dependence of the training steps on the time horizon. In the dynamic approach $\sum_h N_h$ scales with H^3 in comparison to H^5 in the convergence rate for the simultaneous approach. In particular for large time horizons the theoretical analysis shows that reaching a given accuracy is more costly for simultaneous training of parameters. In FT-DynPG the powers are due to the smoothness constant, the $\frac{\epsilon}{H}$ error which we have to achieve in every epoch and finally the sum over all epochs. In comparison to FT-SimPG, the smoothness constant is a power of 1 better and the gradient domination property does not depend on H at all.
Note that we just compare upper bounds. However, in the next section we provide a toy example visualising that the rate of convergence in both approaches is of order $\mathcal{O}(\frac{1}{n})$ and the constants in the dynamic approach are indeed better then for the simultaneous approach.

4.3 NUMERICAL EXAMPLE UNDER EXACT GRADIENTS

We enclose a numerical toy example of a very simple MDP problem of optimally stopping when throwing a dice $H = 5$ times. This is a non-trivial example for which exact policy gradients can be computed. The simulations visualize that the theoretical results (in the exact gradient setup) are sharp up to constants.
The finite-time undiscounted MDP corresponding to this example is defined as follows:

- $\mathcal{H} = \{0, 1, 2, 3, 4\}$
- a constant state space over the epochs $\mathcal{S} = \{1, \dots, 6, \Delta\}$ containing all sides of the dice $1, \dots, 6$ and a terminal state Δ,
- a constant action space $\mathcal{A} = \{0, 1\}$, where 1 indicates stopping and jumping into the terminal state and 0 indicates continuing to the next epoch,
- a transition function p

$$\begin{aligned} p(s' \,|\, s, a) &= \mathbb{P}(S_{h+1} = s' \,|\, S_h = a, A_h = a) \\ &= \begin{cases} \frac{1}{6}, & \text{if } s', s \in \{0, 1 \dots, 6\},\ a = 0, \\ 1, & \text{if } s' = \Delta,\ s \in \mathcal{S},\ a = 1 \text{ or } s' = s = \Delta,\ a = 0, \\ 0, & \text{otherwise.} \end{cases} \end{aligned}$$

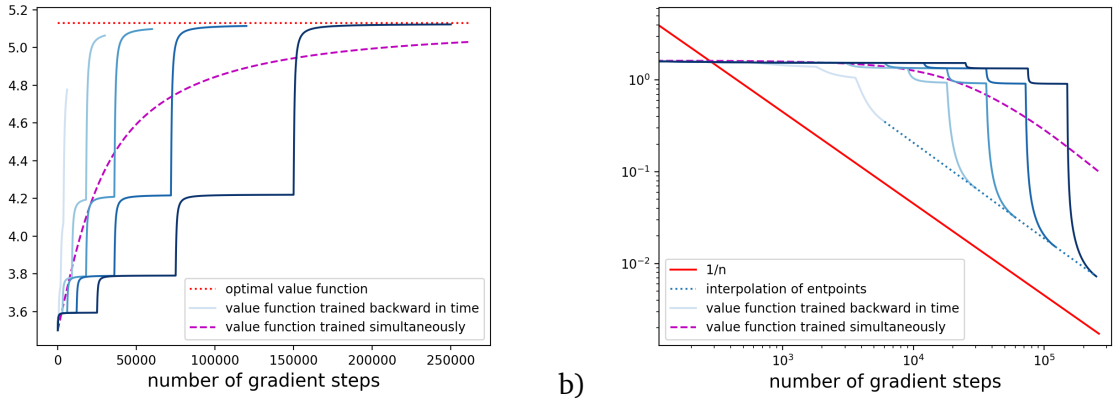
  Thus, we throw the dice iid until stopping for the first time, then we jump into the terminal state and stay there for the rest of the game.

- a reward function r

$$r(s,a) = \begin{cases} s, & \text{if } s \in \{0, 1 \ldots, 6\},\ a = 1, \\ 0, & \text{otherwise.} \end{cases}$$

We only observe a reward when we choose action 1 to top the game and the reward equals the number on the dice.

Having this model with known transition probabilities allows us to implement the simultaneous and FT-DynPG under the exact gradient assumption. In the simulation we always initialized the parameters equal to 0 to obtain a uniform initial distribution. Furthermore we chose the suggested learning rates from Theorem 4.15 in the simultaneous approach and from Theorem 4.24 in the dynamic approach.



a) b)

Figure 4.1: (a) shows the behavior of $V_0^{\pi^{\theta^{(n)}}}$ during the training steps over all epochs. (b) shows the log-log plot of the same simulation visualizing the convergence rate towards V_0^*.

The dotted red line in Figure 4.1 (a) shows the target: V_0^*. On the x-axis we count the number of gradient computations in the algorithms, a way of measuring the computational complexity. The dashed magenta curve shows the evolution of the estimated value function trained with the simultaneous training of all parameters. As c is unknown for the approach, we trained the parameters until an error of 0.1 was achieved. The blue curves show the evolution of the estimated value function trained with our algorithm backwards. Note that the number of gradient steps varies for different epochs, as suggested by Theorem 4.24, less training for later epochs. This can be seen in the plot by the different lengths of the plateaus of the blue lines. One plateau shows the training of one parameter. Just when the last parameter θ_0 is trained, the value function $V_0^{\pi^\theta}$ finally converges towards the target. In this simulation we chose $\epsilon = 5, 1, 0.5, 0.25, 0.12$ to define the length of the training steps according to Theorem 4.24. Note that the uniform initialization leads to $c_h = 0.5$ such that N_h could be explicitly calculated. From light to dark blue ϵ decreases. It can be seen that the final error is better than the chosen epsilon, indicating that the rate of convergence from the dynamic approach is tight up to constants.
In Figure 4.1 (b) for comparison the red line is a constant times $\frac{1}{n}$. The dashed magenta line is the optimal value minus the dashed magenta curve from (a) of the simultaneous approach. Also, the blue curves are the optimal value minus the blue curves from (a). The dotted blue line is the

linear interpolation of the end points of the blue lines. As the dotted blue line, the magenta line and the red line have the same slop, this shows the $\frac{1}{n}$-convergence rate in the accuracy level ϵ. The larger difference from the dashed magenta line to the red line in comparison the dotted blue line to the red line indicates the larger constant in the rate of convergence.
Both plots show that the FT-DynPG algorithm converges faster than the simultaneous one. As suggested by the upper bounds the effect gets much stronger for larger H.

4.4 Convergence of Stochastic Softmax Policy Gradient

In the previous sections, we have derived global convergence guarantees for solving a finite-time MDP via FT-SimPG as well as FT-DynPG with exact gradient computation. However, in practical scenarios assuming access to exact gradients is not feasible, since the transition function p of the underlying MDP is unknown. In the following section, we want to relax this assumption by replacing the exact gradient by a stochastic approximation. To be more precise, we view a model-free setting where we are only able to generate trajectories of the finite-time MDP. These trajectories are used to formulate the stochastic PG method for training the parameters in both the simultaneous and dynamic approach.
Although in both approaches we are able to guarantee almost sure asymptotic convergence similar to the exact PG scheme, we are no longer able to control the constants c and c_h respectively along trajectories of the stochastic PG scheme due to the randomness in our iterations. Therefore, the derived lower bound in the weak gradient domination may degenerate in general. In order to derive complexity bounds in the stochastic scenario, we make use of the crucial property that c (and c_h respectively) remain strictly positive along the trajectory of the exact PG scheme. To do so, we introduce the stopping times τ and τ_h stopping the scheme when the stochastic PG trajectory is too far away from the exact PG trajectory (under same initialization). Hence, conditioning on $\{\tau \geq n\}$ (and $\{\tau_h \geq n\}$ respectively) forces the stochastic PG to remain close to the exact PG scheme and hence, guarantees non-degenerated weak gradient domination. The proof structure in the stochastic setting is then two-fold:

1. We derive a rate of convergence of the stochastic PG scheme under non-degenerated weak gradient domination on the event $\{\tau \geq n\}$. Since we consider a constant step size, the batch size needs to be increased sufficiently fast for controlling the variance occurring through the stochastic approximation scheme.

2. We introduce a second rule for increasing the batch size depending on a tolerance $\delta > 0$ leading to $\mathbb{P}(\tau \leq n) < \delta$. This means, that one forces the stochastic PG to remain close to the exact PG with high probability.

A similar proof strategy has been introduced in Ding, Zhang, and Lavaei [DZL22] for proving convergence of entropy-regularized stochastic PG in infinite-time horizon MDPs. Their analysis heavily depends on the existence of an optimal parameter which is due to regularization. In the unregularized problem this is not the case since the softmax parameters usually diverge to $+/-\infty$ in order to approximate a deterministic optimal solution. Consequently, their analysis does not carry over straightforwardly to the unregularized setting. One of the main challenges in our proof is to construct a different stopping time, independent of optimal parameters, such

that the stopping time still occurs with small probability given a large enough batch size. We again first discuss the simultaneous approach followed by the dynamic approach.

4.4.1 *Simultaneous stochastic policy gradient*

We assume throughout this section that Assumption 4.12 holds true. Due to the finite-time horizon, it is not a challenge to define an unbiased estimator for the gradient compared to discounted MDPs (see Section 3.1.2). Using a mini-batch Monte-Carlo estimator in the representation of the gradient by the policy gradient theorem (Theorem 4.2) results in an unbiased estimator for which we can also guarantee bounded variance.

Unbiased gradient estimator. Consider K trajectories $(s_h^i, a_h^i)_{h=0}^{H-1}$, for $i = 1, \dots, K$, generated by $s_0^i \sim \mu$, $a_h^i \sim \pi^\Theta(\cdot|s_h^i)$ and $s_h^i \sim p(\cdot|s_{h-1}^i, a_{h-1}^i)$ for $0 \leq h < H$. The gradient estimator is defined by

$$\widehat{\nabla} J^K(\Theta, \mu) = \frac{1}{K} \sum_{i=1}^{K} \sum_{h=0}^{H-1} \nabla \log(\pi^\Theta(a_h^i|s_h^i)) \widehat{R}_h^i, \tag{4.19}$$

where $\widehat{R}_h^i = \sum_{k=h}^{H-1} r(s_k^i, a_k^i)$ is an unbiased estimator of the h-state-action value function in (s_h^i, a_h^i) under policy π^Θ. The number of trajectories K is often called batch size of the estimator. The stochastic PG updates are given by

$$\bar{\Theta}^{(n+1)} = \bar{\Theta}^{(n)} + \alpha \widehat{\nabla} J^K(\bar{\Theta}^{(n)}, \mu). \tag{4.20}$$

We first proof that the gradient estimator is unbiased (independent of the parametrization class) and has bounded variance (under tabular softmax parametrization in equation (4.1)).

LEMMA 4.25. *Consider the estimator from equation* (4.19). *For any $K > 0$ and parametrization $(\pi^\Theta)_{\Theta \in \mathbb{R}^d}$ it holds that*

$$\mathbb{E}_\mu^{\pi^\Theta}[\widehat{\nabla} J^K(\Theta, \mu)] = \nabla J(\Theta, \mu).$$

If $(\pi^\Theta)_{\Theta \in \mathbb{R}^d}$ the tabular softmax parametrization in equation (4.1)*, then*

$$\mathbb{E}_\mu^{\pi^\Theta}[\|\widehat{\nabla} J^K(\Theta, \mu) - \nabla J(\Theta, \mu)\|^2] \leq \frac{3H^4 \max\{R^*, 1\}^4}{K} =: \frac{\xi}{K}$$

Proof. By the definition of $\widehat{\nabla} J^K$ we have

$$\begin{aligned}
\mathbb{E}_\mu^{\pi^\Theta}[\widehat{\nabla} J^K(\Theta, \mu)] &= \mathbb{E}_\mu^{\pi^\Theta}\Big[\frac{1}{K} \sum_{i=1}^{K} \sum_{h=0}^{H-1} \nabla \log(\pi^\Theta(A_h^i|S_h^i)) \widehat{R}_h^i\Big] \\
&= \mathbb{E}_\mu^{\pi^\Theta}\Big[\sum_{h=0}^{H-1} \nabla \log(\pi^\Theta(A_h|S_h)) \widehat{R}_h\Big] \\
&= \mathbb{E}_\mu^{\pi^\Theta}\Big[\sum_{h=0}^{H-1} \nabla \log(\pi^\Theta(A_h|S_h)) \sum_{k=h}^{H-1} r(S_k, A_k)\Big],
\end{aligned}$$

where we used that we consider independent samples for $i = 1, \dots, K$. From the proof of the PG theorem (cf. Theorem 4.2), we deduce that

$$\mathbb{E}_\mu^{\pi^\Theta}[\widehat{\nabla} J^K(\Theta, \mu)] = \nabla J(\Theta, \mu).$$

For the second claim, we first obtain that

$$\begin{aligned}
\|\nabla J(\Theta, \mu)\| &= \Big(\sum_{s \in \mathcal{S}^{[\mathcal{H}]}} \sum_{a \in \mathcal{A}} (H d_\mu^{\pi^\Theta}(s) \pi^\theta(a|s) A^{\pi^\Theta}(s,a))^2 \Big)^{\frac{1}{2}} \\
&\leq H^2 R^* \Big(\sum_{s \in \mathcal{S}^{[\mathcal{H}]}} \sum_{a \in \mathcal{A}} (d_\mu^{\pi^\Theta}(s) \pi^\Theta(a|s))^2 \Big)^{\frac{1}{2}} \\
&\leq H^2 (R^*)^2,
\end{aligned}$$

because $\pi^\Theta(\cdot|s) \leq 1$ and $d_\mu^{\pi^\Theta}(s) \leq 1$, as both are probability distributions.
Next,

$$\begin{aligned}
\mathbb{E}_\mu^{\pi^\Theta}[\|\widehat{\nabla} J^1(\Theta, \mu)\|] &\leq \mathbb{E}_\mu^{\pi^\Theta} \Big[\sum_{h=0}^{H-1} \|\nabla \log(\pi^\Theta(A_h|S_h))\| |\widehat{R}_h| \Big] \\
&\leq H^2 R^* \mathbb{E}_\mu^{\pi^\Theta} \Big[\|\nabla \log(\pi^\Theta(A_h|S_h))\| \Big] \\
&\leq H^2 R^*,
\end{aligned}$$

where the last inequality follows with by Yuan, Gower, and Lazaric [YGL22, Lem 4.8] and Jensen's inequality. Thus,

$$\begin{aligned}
\mathbb{E}_\mu^{\pi^\Theta}[\|\widehat{\nabla} J^K(\Theta, \mu) - \nabla J(\Theta, \mu)\|^2] &\leq \frac{1}{K} \mathbb{E}_\mu^{\pi^\Theta} \Big[\|\widehat{\nabla} J^1(\Theta, \mu) - \nabla J(\Theta, \mu)\|^2 \Big] \\
&\leq \frac{1}{K} \mathbb{E}_\mu^{\pi^\Theta} \Big[\|\widehat{\nabla} J^1(\Theta)\|^2 + 2\|\widehat{\nabla} J^1(\Theta, \mu)\| \|\nabla J(\Theta)\| + \|\nabla J(\Theta, \mu)\|^2 \Big] \\
&\leq \frac{1}{K} \Big[H^4 (R^*)^2 + H^4 (R^*)^2 + H^4 (R^*)^4 \Big].
\end{aligned}$$

We define $\xi = 3H^4 \max\{R^*, 1\}^4 \geq H^4 (R^*)^2 + H^4 (R^*)^2 + H^4 (R^*)^4$ to prove the claim. ∎

Define the stopping time. We assume again the tabular softmax parametrization on the enlarged state space in equation (4.1) and denote by $(\Theta^{(n)})_{n \in \mathbb{N}}$ the deterministic sequence generated by Algorithm 4 under exact gradients and by $(\bar{\Theta}^{(n)})_{n \in \mathbb{N}}$ the stochastic sequence in equation (4.20). We assume that the initial parameter agree, i.e. $\Theta^{(1)} = \bar{\Theta}^{(1)}$, and the step size α is the same for both processes. The natural filtration of the stochastic process $(\bar{\Theta}_h^{(n)})_{n \in \mathbb{N}}$ is denoted by $(\mathcal{F}^{(n)})_{n \in \mathbb{N}}$. Recall that for the deterministic scheme we could assure that $c = \inf_n \min_{s \in \mathcal{S}^{[\mathcal{H}]}} \pi^{\Theta_n}(a^*(s)|s)$ is bounded away from 0 by Lemma 4.14. This cannot be guaranteed for the stochastic trajectory. The idea of the convergence analysis for stochastic softmax PG is now to define the following stopping time

$$\tau := \min\{n \geq 1 : \|\Theta^{(n)} - \bar{\Theta}^{(n)}\|_2 \geq \frac{c}{4}\}.$$

This means, τ is the first time when the stochastic process $(\bar{\Theta}^{(n)})_{n\in\mathbb{N}}$ is *too far away* from the PG trajectory $(\Theta^{(n)})_{n\in\mathbb{N}}$. Hence, all challenges encountered in the deterministic case transfer to the stochastic context, indicating that the model dependent constant c naturally appears in the error bounds of the stochastic case. We emphasize that τ is a stopping time with respect to the filtration $(\mathcal{F}^{(n)})_{n\in\mathbb{N}}$ by construction.

Event $\{n \le \tau\}$. First, consider the event $\{n \le \tau\}$, i.e. $\|\Theta^{(n)} - \bar{\Theta}^{(n)}\|_2 \le \frac{c}{4}$. Then, it follows from the $\sqrt{2}$-Lipschitz continuity of $\Theta \mapsto \pi^{\Theta}(a^*(s)|s)$ that $\min_{0\le n\le\tau}\min_{s\in\mathcal{S}} \pi^{\bar{\Theta}^{(n)}}(a^*(s)|s) \ge \frac{c}{2} > 0$.

LEMMA 4.26. *The softmax policy on the enlarged state space and analogously every softmax policy for FT-DynPG is $\sqrt{2}$-Lipschitz with respect to Θ or θ_h respectively.*

Proof. We consider a general softmax policy π^{θ} with parameter $(\theta(s,a))_{s\in\mathcal{S},a\in\mathcal{A}}$ of a finite state and action space such that

$$\pi^{\theta}(a|s) = \frac{\exp(\theta(s,a))}{\sum_{a'\in\mathcal{A}}\exp(\theta(s,a'))}.$$

The derivative of the softmax function is

$$\begin{aligned}\frac{\partial\pi^{\theta}(a|s)}{\partial\Theta(s',a')} &= \mathbf{1}_{s'=s}\Big[\frac{\mathbf{1}_{a'=a}\exp(\theta(s,a))\big(\sum_{\widetilde{a}\in\mathcal{A}_s}\exp(\theta(s,\widetilde{a}))\big) - \exp(\theta(s,a))\exp(\theta(s,a'))}{\big(\sum_{\widetilde{a}\in\mathcal{A}_s}\exp(\theta(s,\widetilde{a}))\big)^2}\Big] \\ &= \mathbf{1}_{s'=s}\Big[\mathbf{1}_{a'=a}\pi^{\theta}(a|s) - \pi^{\theta}(a|s)\pi^{\theta}(a'|s)\Big].\end{aligned}$$

Therefore,

$$\begin{aligned}\|\nabla\pi^{\theta}(a|s)\|_2 &= \sqrt{\sum_{\widetilde{a}\in\mathcal{A}_s}\Big(\mathbf{1}_{a'=a}\pi^{\Theta}(a|s) - \pi^{\theta}(a|s)\pi^{\theta}(a'|s)\Big)^2} \\ &\le \sqrt{\pi^{\theta}(a|s)^2 - 2\pi^{\theta}(a|s)^3 + \sum_{\widetilde{a}\in\mathcal{A}_s}\pi^{\theta}(a'|s)^2\pi^{\theta}(a|s)^2} \le \sqrt{2}.\end{aligned}$$

■

LEMMA 4.27. *Let μ be a probability measure such that $\mu(s) > 0$ for all $s \in \mathcal{S}$ and consider the sequence $(\bar{\Theta}^{(n)})_{n\in\mathbb{N}}$ generated by the stochastic recursion in equation* (4.20) *under softmax parametrization. Then, it holds almost surely that $\min_{0\le n\le\tau}\min_{s\in\mathcal{S}_h}\pi^{\bar{\Theta}^{(n)}}(a^*(s)|s) \ge \frac{c}{2}$ is strictly positive.*

Proof. For every $n \le \tau$ we obtain by the $\sqrt{2}$-Lipschitz continuity of softmax in Lemma 4.26 that

$$\begin{aligned}\pi^{\bar{\Theta}^{(n)}}(a^*(s)|s) &\ge \pi^{\Theta^{(n)}}(a^*(s)|s) - |\pi^{\Theta^{(n)}}(a^*(s)|s) - \pi^{\bar{\Theta}^{(n)}}(a^*(s)|s)| \\ &\ge \pi^{\Theta^{(n)}}(a^*(s)|s) - \sqrt{2}\|\bar{\Theta}^{(n)} - \Theta^{(n)}\|_2 \\ &> \frac{c}{2} > 0,\end{aligned}$$

holds almost surely. The claim follows directly. ■

This allows us to use the weak gradient domination of Lemma 4.11 to derive a convergence rate on the event $\{n \leq \tau\}$ in the following sense:

LEMMA 4.28. *Under Assumption 4.12, let μ be a probability measure such that $\mu(s) > 0$ for all $s \in \mathcal{S}$ and consider the sequence $(\bar{\Theta}^{(n)})_{n\in\mathbb{N}}$ generated by the stochastic recursion in equation* (4.20) *under softmax parametrization. Suppose that*

(i) the batch size $K^{(n)} \geq \frac{9}{8}\frac{c^2 \max\{R^*,1\}^2(1-\frac{1}{2\sqrt{N}})}{N^{3/2}|\mathcal{S}|H^{19}}\left\|\frac{d_\mu^{\pi^*}}{\mu}\right\|_\infty^{-2} n^2$ is increasing for fix $N \geq 1$ and

(ii) the step size $\alpha = \frac{1}{5H^2R^*\sqrt{N}}$.

Then, it holds true that

$$\mathbb{E}\big[(J^*(\mu) - J(\bar{\Theta}^{(n)}, \mu))\mathbf{1}_{\{n\leq\tau\}}\big] \leq \frac{20|\mathcal{S}|H^5R^*}{c^2\frac{1}{\sqrt{N}}(1-\frac{1}{2\sqrt{N}})n}\left\|\frac{d_\mu^{\pi^*}}{\mu}\right\|_\infty^2.$$

Proof. Throughout the proof we drop the μ in J and J^*.
First, we deduce from the L-smoothness of J, as in the proof of Theorem 4.15 that almost surely

$$J(\bar{\Theta}^{(n+1)}) \geq J(\bar{\Theta}^{(n)}) + \big(\nabla J(\bar{\Theta}^{(n)})\big)^T(\bar{\Theta}^{(n+1)} - \bar{\Theta}^{(n)}) - \frac{L}{2}\|\bar{\Theta}^{(n+1)} - \bar{\Theta}^{(n)}\|^2.$$

We continue with

$$\begin{aligned}
J(\bar{\Theta}^{(n+1)}) &\geq J(\bar{\Theta}^{(n)}) + \alpha\big(\nabla J(\bar{\Theta}^{(n)})\big)^T\widehat{\nabla}J^K(\bar{\Theta}^{(n)}) - \frac{L\alpha^2}{2}\|\widehat{\nabla}J^K(\bar{\Theta}^{(n)})\|^2 \\
&= J(\bar{\Theta}^{(n)}) + \alpha\big(\nabla J(\bar{\Theta}^{(n)})\big)^T\nabla J(\bar{\Theta}^{(n)}) + \alpha\big(\nabla J(\bar{\Theta}^{(n)})\big)^T\big(\widehat{\nabla}J^K(\bar{\Theta}^{(n)}) - \nabla J(\bar{\Theta}^{(n)})\big) \\
&\quad - \frac{L\alpha^2}{2}\|\big(\widehat{\nabla}J^K(\bar{\Theta}^{(n)}) - \nabla J(\bar{\Theta}^{(n)})\big) + \nabla J(\bar{\Theta}^{(n)})\|^2.
\end{aligned}$$

Thus,

$$J(\bar{\Theta}^{(n+1)}) \geq J(\bar{\Theta}^{(n)},) + \big(\alpha - \frac{L\alpha^2}{2}\big)\|\nabla J(\bar{\Theta}^{(n)})\|^2 + \big(\alpha - L\alpha^2\big)\langle\nabla J(\bar{\Theta}^{(n)}), \phi_n\rangle - \frac{L\alpha^2}{2}\|\phi_n\|^2,$$

where $\phi_n := \widehat{\nabla}J^K(\bar{\Theta}^{(n)}) - \nabla J(\bar{\Theta}^{(n)})$. Next we take the conditional expectation on $\mathcal{F}_n$. Then by Lemma 4.25 we obtain

$$E\Big[J(\bar{\Theta}^{(n+1)})|\mathcal{F}_n\Big] \geq J(\bar{\Theta}^{(n)}) + \Big(\alpha - \frac{L\alpha^2}{2}\Big)\|\nabla J(\bar{\Theta}^{(n)})\|^2 - \frac{L\alpha^2\xi}{2K_n}.$$

Subtracting this equation form J^* and taking the expectation under the event $\{n+1 \leq \tau\}$ results in:

$$\begin{aligned}
&\mathbb{E}\Big[(J^* - J(\bar{\Theta}^{(n+1)}))\mathbf{1}_{\{n+1\leq\tau\}}\Big] \\
&= \mathbb{E}\Big[\mathbb{E}\Big[(J^* - J(\bar{\Theta}^{(n+1)}))|\mathcal{F}_n\Big]\mathbf{1}_{\{n+1\leq\tau\}}\Big]
\end{aligned}$$

$$\begin{aligned}
&\leq \mathbb{E}\Big[\Big(J^* - \mathbb{E}\Big[J(\bar{\Theta}^{(n+1)})|\mathcal{F}_n\Big]\Big)\mathbf{1}_{\{n\leq\tau\}}\Big] \\
&\leq \mathbb{E}\Big[(J^* - J(\bar{\Theta}^{(n)}))\mathbf{1}_{\{n\leq\tau\}}\Big] - \Big(\alpha - \frac{L\alpha^2}{2}\Big)\mathbb{E}\Big[\|\nabla J(\bar{\Theta}^{(n)})\|^2\mathbf{1}_{\{n\leq\tau\}}\Big] + \frac{L\alpha^2\xi}{2K_n} \\
&\leq \mathbb{E}\Big[(J^* - J(\bar{\Theta}^{(n)}))\mathbf{1}_{\{n\leq\tau\}}\Big] - \alpha(1 - \frac{1}{2\sqrt{N}})\mathbb{E}\Big[\|\nabla J(\bar{\Theta}^{(n)})\|^2\mathbf{1}_{\{n\leq\tau\}}\Big] + \frac{L\alpha^2\xi}{2K_n},
\end{aligned}$$

where we used that $\{n+1 \leq \tau\} = \{\tau \leq n\}^C$ is $\mathcal{F}_n$-measurable and that $\mathbf{1}_{\{n+1\leq\tau\}} \leq \mathbf{1}_{\{n\leq\tau\}}$ a.s. With the gradient domination property in Lemma 4.11 and $\min_{1\leq n\leq\tau}\min_{s\in\mathcal{S}}\pi^{\bar{\Theta}^{(n)}}(a^*(s)|s) \geq \frac{c}{2}$ by Lemma 4.27 we deduce

$$\begin{aligned}
&\mathbb{E}\Big[(J^* - J(\bar{\Theta}^{(n+1)}))\mathbf{1}_{\{n+1\leq\tau\}}\Big] \\
&\leq \mathbb{E}\Big[(J^* - J(\bar{\Theta}^{(n)}))\mathbf{1}_{\{n\leq\tau\}}\Big] - \alpha(1 - \frac{1}{2\sqrt{N}})\frac{c^2}{|\mathcal{S}|H}\Big\|\frac{d_\mu^{\pi^*}}{d_\mu^{\pi^\theta}}\Big\|_\infty^{-2}\mathbb{E}\Big[(J^* - J(\bar{\Theta}^{(n)}))\mathbf{1}_{\{n\leq\tau\}}\Big]^2 + \frac{L\alpha^2\xi}{2K_n} \\
&\leq \mathbb{E}\Big[(J^* - J(\bar{\Theta}^{(n)}))\mathbf{1}_{\{n\leq\tau\}}\Big] - \alpha(1 - \frac{1}{2\sqrt{N}})\frac{c^2}{|\mathcal{S}|H^3}\Big\|\frac{d_\mu^{\pi^*}}{\mu}\Big\|_\infty^{-2}\mathbb{E}\Big[(J^* - J(\bar{\Theta}^{(n)}))\mathbf{1}_{\{n\leq\tau\}}\Big]^2 + \frac{L\alpha^2\xi}{2K_n},
\end{aligned}$$

where we used in the last inequality that Assumption 4.12 implies $d_\mu^{\pi^\Theta}(s) \geq \frac{1}{H}\mu(s)$ (see Remark 4.13). For $d_n := \mathbb{E}\Big[(J^* - J(\bar{\Theta}^{(n)}))\mathbf{1}_{\{n\leq\tau\}}\Big]$ we obtain the recursive inequality

$$d_{n+1} \leq d_n - \alpha(1 - \frac{1}{2\sqrt{N}})\frac{c^2}{|\mathcal{S}|H^3}\Big\|\frac{d_\mu^{\pi^*}}{\mu}\Big\|_\infty^{-2}d_n^2 + \frac{L\alpha^2\xi}{2K_n}.$$

We define $w := \alpha(1 - \frac{1}{2\sqrt{N}})\frac{c^2}{|\mathcal{S}|H^3}\Big\|\frac{d_\mu^{\pi^*}}{\mu}\Big\|_\infty^{-2}$ and $B = \frac{L\alpha^2\xi}{2} > 0$ such that

$$d_{n+1} \leq d_n(1 - wd_n) + \frac{B}{K_n}.$$

Note that $w > 0$ by the assumption $\mu(s) > 0$ for all $s \in \mathcal{S}$. Then by our choice of K_n it holds that

$$\begin{aligned}
&\frac{9}{4}wBn^2 = \frac{9}{8}\frac{c^2\alpha^3L(1 - \frac{1}{2\sqrt{N}})\xi}{|\mathcal{S}|H^3}\Big\|\frac{d_\mu^{\pi^*}}{\mu}\Big\|_\infty^{-2}n^2 \\
&\leq \frac{9}{8}\frac{c^2\alpha^2(1 - \frac{1}{2\sqrt{N}})\xi}{\sqrt{N}|\mathcal{S}|H^3}\Big\|\frac{d_\mu^{\pi^*}}{\mu}\Big\|_\infty^{-2}n^2 \leq \frac{9}{8}\frac{c^2\max\{R^*,1\}^2(1 - \frac{1}{2\sqrt{N}})}{N^{3/2}|\mathcal{S}|H^{19}}\Big\|\frac{d_\mu^{\pi^*}}{\mu}\Big\|_\infty^{-2}n^2 \leq K_n.
\end{aligned}$$

Furthermore, we have for $\alpha = \frac{1}{5H^2R^*\sqrt{N}}$ that

$$\frac{4}{3w} = \frac{4|\mathcal{S}|H^3}{3\alpha(1 - \frac{1}{2\sqrt{N}})c^2}\Big\|\frac{d_\mu^{\pi^*}}{\mu}\Big\|_\infty^2 = \frac{20|\mathcal{S}|H^5R^*}{c^2\frac{1}{\sqrt{N}}(1 - \frac{1}{2\sqrt{N}})}\Big\|\frac{d_\mu^{\pi^*}}{\mu}\Big\|_\infty^2.$$

We obtain that

$$d_1 \leq HR^* \leq \frac{4}{3w} \leq \frac{4}{3w\cdot 1},$$

because $c \leq 1$, $\left\| \frac{d_\mu^{\pi^*}}{\mu} \right\|_\infty^2 \geq 1$ and $\frac{1}{\sqrt{N}}(1 - \frac{1}{2\sqrt{N}}) < 1$ for all $N \geq 1$.
Suppose the induction assumption $d_n \leq \frac{4}{3wn}$ holds true. First, recall the recursive inequality

$$d_{n+1} \leq d_n - w d_n^2 + \frac{B}{K_n}.$$

The function $f(x) = x - wx^2$ is monotonically increasing in $[0, \frac{1}{2w}]$, and by induction assumption $d_n \leq \frac{1}{4wn} \leq \frac{1}{2w}$. Thus,

$$\begin{aligned} d_{n+1} &\leq d_n - w d_n^2 + \frac{B}{K_n} \leq \frac{4}{3wn} - \frac{16}{9wn^2} + \frac{B}{K_n} \\ &\leq \frac{4}{3wn} - \frac{16}{9wn^2} + \frac{4B}{9wBn^2} = \frac{4}{3wn} - \frac{12}{9wn^2} = \frac{4}{3w}\Big(\frac{1}{n} - \frac{1}{n^2}\Big) \\ &\leq \frac{4}{3wn}, \end{aligned}$$

by the choice of $K_n \geq \frac{9}{4} w B n^2$. We deduce the claim

$$d_n \leq \frac{4}{3wn} = \frac{20|\mathcal{S}|H^5 R^*}{c^2 \frac{1}{\sqrt{N}}(1 - \frac{1}{2\sqrt{N}})n} \left\| \frac{d_\mu^{\pi^*}}{\mu} \right\|_\infty^2.$$

■

Event $\{\tau \leq n\}$. Secondly, consider the complementary event $\{\tau \leq n\}$. We can bound the probability of this event by δ for a large enough batch size K. The proof is inspired by a similar result obtained by Ding, Zhang, and Lavaei [DZL22, Lem. 6.3] for infinite-time horizon discounted MDPs.

LEMMA 4.29. *Let μ be a probability measure such that $\mu(s) > 0$ for all $s \in \mathcal{S}$ and consider the sequence $(\bar{\Theta}^{(n)})_{n \in \mathbb{N}}$ generated by the stochastic recursion in equation (4.20) under softmax parametrization. For any $\delta > 0$, suppose that*

(i) the batch size $K \geq \frac{10 \max\{R^*, 1\}^2 n^3}{c^2 \delta^2}$ and

(ii) the step size $\alpha = \frac{1}{\sqrt{n} 5 H^2 R^*}$.

Then, it holds true that $\mathbb{P}(\tau \leq n) < \delta$.

Proof. By the definition of τ we have

$$\mathbb{P}(\tau \leq n) = \mathbb{P}\Big(\max_{1 \leq t \leq n} \|\Theta^{(t)} - \bar{\Theta}^{(t)}\| \geq \frac{c_h}{4}\Big),$$

so we first study $\|\Theta^{(t)} - \bar{\Theta}^{(t)}\|$. We emphasize that [DZL22, Lemma 6.3] established a similar recursive inequality.

$$\|\bar{\Theta}^{(t)} - \Theta^{(t)}\| = \|\bar{\Theta}^{(1)} + \sum_{k=1}^{t-1} \alpha \widehat{\nabla} J^K(\bar{\Theta}^{(k)}, \mu) - (\Theta^{(1)} + \sum_{k=1}^{t-1} \alpha \nabla J(\Theta^{(k)}, \mu))\|$$

$$\leq \sum_{k=1}^{t-1} \alpha \|\widehat{\nabla} J^K(\bar{\Theta}^{(k)}, \mu) - \nabla J(\Theta^{(k)}, \mu)\|$$
$$\leq \alpha \sum_{k=1}^{t-1} (\|\widehat{\nabla} J^K(\bar{\Theta}^{(k)}, \mu) - \nabla J(\bar{\Theta}^{(k)}, \mu)\| + \|\nabla J(\bar{\Theta}^{(k)}, \mu) - \nabla J(\Theta^{(k)}, \mu)\|).$$

We define again $\phi_k^K = \widehat{\nabla} J^K(\bar{\Theta}^{(k)}, \mu) - \nabla J(\bar{\Theta}^{(k)}, \mu)$ and continue using the L-Lipschitz continuity of $\nabla J(\Theta)$ such that

$$\|\Theta^{(t)} - \bar{\Theta}^{(t)}\| \leq \alpha \sum_{k=1}^{t-1} (\|\phi_k^K\| + L\|\Theta^{(k)} - \bar{\Theta}^{(k)}\|) = \alpha \sum_{k=1}^{t-1} \|\phi_k^K\| + \alpha L \sum_{k=1}^{t-1} \|\Theta^{(k)} - \bar{\Theta}^{(k)}\|.$$

Using this inequality sequentially leads to

$$\begin{aligned}
\|\Theta^{(t)} - \bar{\Theta}^{(t)}\| &\leq \alpha \sum_{k=1}^{t-1} \|\phi_k^K\| + \alpha L \sum_{k=1}^{t-1} \|\Theta^{(k)} - \bar{\Theta}^{(k)}\| \\
&\leq \alpha \sum_{k=1}^{t-1} \|\phi_k^K\| + \alpha L \sum_{k=1}^{t-2} \|\Theta^{(k)} - \bar{\Theta}^{(k)}\| + \alpha L \Big(\alpha \sum_{k=1}^{t-2} \|\phi_k^K\| + \alpha L \sum_{k=1}^{t-2} \|\Theta^{(k)} - \bar{\Theta}^{(k)}\| \Big) \\
&= \alpha \sum_{k=1}^{t-1} \|\phi_k^K\| + \alpha^2 L \sum_{k=1}^{t-2} \|\phi_k^K\| + (1 + \alpha L)\alpha L \sum_{k=1}^{t-2} \|\Theta^{(k)} - \bar{\Theta}^{(k)}\| \\
&= \alpha \|\phi_{t-1}^K\| + \alpha(1 + \alpha L) \sum_{k=1}^{t-2} \|\phi_k^K\| + (1 + \alpha L)\alpha L \sum_{k=1}^{t-2} \|\Theta^{(k)} - \bar{\Theta}^{(k)}\| \\
&\leq \sum_{k=1}^{t-1} \alpha (1 + \alpha L)^{t-k-1} \|\phi_k^K\|.
\end{aligned}$$

Applying Markov's inequality results in

$$\begin{aligned}
\mathbb{P}(\tau \leq n) &= \mathbb{P}(\max_{1 \leq t \leq n} \|\theta^{(t)} - \bar{\theta}^{(t)}\| \geq \frac{c}{4}) \\
&\leq \mathbb{P}(\sum_{k=1}^{n-1} \alpha (1 + \alpha L)^{n-k-1} \|\phi_k^K\| \geq \frac{c_h}{4}) \\
&\leq \frac{4 \sum_{k=1}^{n-1} \alpha (1 + \alpha L)^{n-k-1} \mathbb{E}[\|\phi_k^K\|]}{c} \\
&\leq \frac{4n\alpha (1 + \alpha L)^{n-1} \sqrt{\frac{\xi}{K}}}{c},
\end{aligned}$$

where in the last inequality $\mathbb{E}[\|\phi_k^K\|] \leq \sqrt{\mathbb{E}[\|\phi_k^K\|^2]} \leq \sqrt{\frac{\xi}{K}}$ by Jensen's inequality and Lemma 4.25. Now we plug in the choice of $\alpha = \frac{1}{\sqrt{n} 5H^2 R^*} < \frac{1}{\sqrt{n}L}$,

$$\mathbb{P}(\tau \leq n) \leq \frac{4n \frac{1}{\sqrt{n} 5H^2 R^*} (1 + \frac{1}{\sqrt{n}L} L)^{n-1} \sqrt{\frac{\xi}{K}}}{c} = \frac{4\sqrt{n} (1 + \frac{1}{\sqrt{n}})^{n-1} \sqrt{C_h}}{5H^2 R^* c \sqrt{K}} \leq \frac{4\sqrt{n} n \sqrt{\xi}}{5H^2 R^* c \sqrt{K}},$$

where the last step is due to $f(x) = (1+\frac{1}{\sqrt{x}})^{x-1} \leq x$ for all $x \geq 1$. We follow that $\mathbb{P}(\tau < n) < \delta$ if

$$\frac{16n^3\xi}{25H^4(R^*)^2c^2\delta^2} = \frac{16n^3H^4\max\{R^*,1\}^4 3}{25H^4(R^*)^2c^2\delta^2} \leq \frac{48\max\{R^*,1\}^2n^3}{5c^2\delta^2} \leq \frac{10\max\{R^*,1\}^2n^3}{c^2\delta^2} = K.$$

■

Convergence result. Our main result for stochastic FT-SimPG is stated in the following.

THEOREM 4.30. *Under Assumption 4.12, let μ be a probability measure such that $\mu(s) > 0$ for all $s \in \mathcal{S}$. Consider the final policy using Algorithm 4 with stochastic updates from equation* (4.20) *under softmax parametrization. Denote the final policy by $\widehat{\pi}^* = \pi^{\bar{\Theta}^{(N)}}$. Moreover, for any $\delta, \epsilon > 0$ assume that*

(i) the number of training steps satisfies $N \geq \left(\frac{21|\mathcal{S}|H^5R^*}{\epsilon\delta c^2}\right)^2 \left\|\frac{d_\mu^{\pi^*}}{\mu}\right\|_\infty^4$,

(ii) the step size $\alpha = \frac{1}{5H^2R^*\sqrt{N}}$ and

(iii) the batch size $K \geq \frac{10\max\{R^*,1\}^2N^3}{c^2\delta^2}$.

Then, it holds true that

$$\mathbb{P}\left(V_0^*(\mu) - V_0^{\widehat{\pi}^*}(\mu) < \epsilon\right) > 1-\delta\,.$$

Proof. First note again, that by definition $J^*(\mu) = V_0^*(\mu)$ and $J(\bar{\Theta}^{(N)}, \mu) = V_0^{\pi^{\bar{\Theta}^{(N)}}}(\mu)$. We separate the probability using the stopping time τ and obtain

$$\begin{aligned}
\mathbb{P}\Big((J^*(\mu) - J(\bar{\Theta}^{(N)},\mu)) \geq \epsilon\Big) &\leq \mathbb{P}\Big(\{\tau \geq N\} \cap \{(J^*(\mu) - J(\bar{\Theta}^{(N)},\mu)) \geq \epsilon\}\Big) \\
&\quad + \mathbb{P}\Big(\{\tau \leq N\} \cap \{(J^*(\mu) - J(\bar{\Theta}^{(N)},\mu)) \geq \epsilon\}\Big) \\
&\leq \frac{\mathbb{E}\Big[(J^*(\mu) - J(\bar{\Theta}^{(N)},\mu))\mathbf{1}_{\{\tau\geq N\}}\Big]}{\epsilon} + \mathbb{P}(\tau \leq N) \\
&\leq \frac{1}{\epsilon}\frac{20|\mathcal{S}|H^5R^*}{c^2\frac{1}{\sqrt{N}}(1-\frac{1}{2\sqrt{N}})N}\left\|\frac{d_\mu^{\pi^*}}{\mu}\right\|_\infty^2 + \frac{\delta}{2} \\
&\leq \frac{\delta}{2} + \frac{\delta}{2} \\
&= \delta,
\end{aligned}$$

where the second inequality holds due to Lemma 4.28 and Lemma 4.29. The last inequality follows by our choice of N:

$$\frac{20|\mathcal{S}|H^5R^*}{c^2\sqrt{N}(1-\frac{1}{2\sqrt{N}})}\left\|\frac{d_\mu^{\pi^*}}{\mu}\right\|_\infty^2 \leq \frac{\delta}{2}$$

if and only if $N \geq \left(\frac{20|\mathcal{S}|H^5R^*}{\epsilon\delta c^2}\left\|\frac{d_\mu^{\pi^*}}{\mu}\right\|_\infty^2 + \frac{1}{2}\right)^2$, which is satisfied if $N \geq \left(\frac{21|\mathcal{S}|H^5R^*}{\epsilon\delta c^2}\right)^2\left\|\frac{d_\mu^{\pi^*}}{\mu}\right\|_\infty^4$. Note that we can use Lemma 4.28 in the equation above with a constant batch size, because by our choice of α

$$\max\left\{\frac{9}{8}\frac{c^2\max\{R^*,1\}^2(1-\frac{1}{2\sqrt{N}})}{N^{3/2}|\mathcal{S}|H^{19}}\left\|\frac{d_\mu^{\pi^*}}{\mu}\right\|_\infty^{-2}n^2, \frac{10\max\{R^*,1\}^2N^3}{c^2\delta^2}\right\} = \frac{10\max\{R^*,1\}^2N^3}{c^2\delta^2},$$

for all $n \leq N$. The last equality holds, as $c < 1$, $\left\|\frac{d_\mu^{\pi^*}}{\mu}\right\|_\infty^{-2} < 1$. ∎

4.4.2 *Dynamic stochastic policy gradient*

Unbiased gradient estimator: For fixed h consider K_h trajectories $(s_k^i, a_k^i)_{k=h}^{H-1}$, for $i = 1, \dots, K_h$, generated by $s_h^i \sim \mu_h$, $a_h^i \sim \pi^{\theta_h}$ and $a_k^i \sim \widetilde{\pi}_k$ for $h < k < H$. The estimator is defined by

$$\widehat{\nabla} J_h^K(\theta_h, \widetilde{\mathbb{W}}_{(h+1)}, \mu_h) = \frac{1}{K_h}\sum_{i=1}^{K_h}\nabla\log(\pi^{\theta_h}(a_h^i|s_h^i))\widehat{R}_h^i, \tag{4.21}$$

where $\widehat{R}_h^i = \sum_{k=h}^{H-1} r(s_k^i, a_k^i)$ is an unbiased estimator of the h-state-action value function in (s_h^i, a_h^i) under future policy $\widetilde{\mathbb{W}}$. Then the stochastic PG update for training the parameter θ_h is given by

$$\bar{\theta}_h^{(n+1)} = \bar{\theta}_h^{(n)} + \alpha_h\widehat{\nabla} J_h^{K_h}(\bar{\theta}_h^{(n)}, \widetilde{\mathbb{W}}_{(h+1)}, \mu_h). \tag{4.22}$$

We start again by showing that the gradient estimator is unbiased and has bounded variance, independent of the chosen parametrization class.

Lemma 4.31. *For any $h \in \mathcal{H}$ consider the estimator in equation* (4.21)*. Let $K_h > 0$, then for any parametrization $(\pi^{\theta_h})_{\theta_h\in\mathbb{R}^d}$ it holds that*

$$\mathbb{E}_{\mu_h}^{(\pi^{\theta_h},(\widetilde{\mathbb{W}})_{(h+1)})}[\widehat{\nabla} J_h^{K_h}(\theta_h, \widetilde{\mathbb{W}}_{(h+1)}, \mu_h)] = \nabla J_h(\theta_h, \widetilde{\mathbb{W}}_{(h+1)}, \mu_h).$$

If $(\pi^{\theta_h})_{\theta_h\in\mathbb{R}^{d_h}}$ the tabular softmax parametrization in equation (4.4)*, then*

$$\mathbb{E}_{\mu_h}^{(\pi^{\theta_h},(\widetilde{\mathbb{W}})_{(h+1)})}[\|\widehat{\nabla} J_h^{K_h}(\theta_h, \widetilde{\mathbb{W}}_{(h+1)}, \mu_h) - \nabla J_h(\theta_h, \widetilde{\mathbb{W}}_{(h+1)}, \mu_h)\|^2] \leq \frac{5(H-h)^2(R^*)^2}{K_h} =: \frac{\psi_h}{K}.$$

Proof. We drop the subscript h in θ_h for this proof.
By the definition of $\widehat{\nabla} J_h^K$ we have

$$\begin{aligned}
\mathbb{E}_{\mu_h}^{(\pi^\theta,(\widetilde{\mathbb{W}})_{(h+1)})}[\widehat{\nabla} J_h^{K_h}(\theta, \widetilde{\mathbb{W}}_{(h+1)}, \mu_h)] &= \mathbb{E}_{\mu_h}^{(\pi^\theta,(\widetilde{\mathbb{W}})_{(h+1)})}\left[\frac{1}{K_h}\sum_{i=1}^{K_h}\nabla\log(\pi^\theta(A_t^i|S_t^i))\widehat{R}_h^i\right] \\
&= \mathbb{E}_{\mu_h}^{(\pi^\theta,(\widetilde{\mathbb{W}})_{(h+1)})}\left[\nabla\log(\pi^\theta(A_h^1|S_h^1))\widehat{R}_h^1\right] \\
&= \mathbb{E}_{\mu_h}^{(\pi^\theta,(\widetilde{\mathbb{W}})_{(h+1)})}\left[\nabla\log(\pi^\theta(A_h|S_h))\sum_{k=h}^{H-1} r(S_k, A_k)\right],
\end{aligned}$$

where we used that we consider independent samples for $i = 1, \dots, K_h$. From the proof of the dynamical PG theorem (Theorem 4.4), we obtain that

$$\begin{aligned}
&\mathbb{E}_{\mu_h}^{(\pi^\theta,(\widetilde{\mathbb{m}})_{(h+1)})}[\widehat{\nabla} J_h^{K_h}(\theta, \widetilde{\mathbb{m}}_{(h+1)}, \mu)] \\
&= \mathbb{E}_{\mu_h}^{(\pi^\theta,(\widetilde{\mathbb{m}})_{(h+1)})}\Big[\nabla \log(\pi^\theta(A_1|S_h)) \sum_{k=h}^{H-1} r(S_k, A_k)\Big] \\
&= \nabla J_h(\theta, \widetilde{\mathbb{m}}_{(h+1)}, \mu_h).
\end{aligned}$$

For the second claim, we have

$$\begin{aligned}
&\mathbb{E}_{\mu_h}^{(\pi^\theta,(\widetilde{\mathbb{m}})_{(h+1)})}\Big[\|\widehat{\nabla} J_h^{K_h}(\theta, \widetilde{\mathbb{m}}_{(h+1)}, \mu_h) - \nabla J_h(\theta, \widetilde{\mathbb{m}}_{(h+1)}, \mu_h)\|^2\Big] \\
&\le \frac{1}{K_h}\mathbb{E}_{\mu_h}^{(\pi^\theta,(\widetilde{\mathbb{m}})_{(h+1)})}\Big[\|\nabla \log(\pi^\theta(A_h|S_h))\widehat{Q}_h(S_h, A_h) - \nabla J_h(\theta)\|^2\Big] \\
&= \frac{1}{K_h}\mathbb{E}_{\mu_h}^{(\pi^\theta,(\widetilde{\mathbb{m}})_{(h+1)})}\Big[\sum_{s\in\mathcal{S}_h}\sum_{a\in\mathcal{A}_s}\Big(\mathbf{1}_{s=S_h}(\mathbf{1}_{a=A_h} - \pi^\theta(a|s)) \sum_{k=h}^{H-1} r(S_k, A_k) \\
&\quad - \mu_h(s)\pi^\theta(a|s) A_h^{(\pi^\theta,(\widetilde{\mathbb{m}})_{(h+1)})}(s, a)\Big)^2\Big],
\end{aligned}$$

by the definition of $\widehat{\nabla} J_h^{K_h}(\theta, \widetilde{\mathbb{m}}_{(h+1)}, \mu_h)$ and the derivative of $\nabla J_h(\theta, \widetilde{\mathbb{m}}_{(h+1)}, \mu_h)$ for the softmax parametrization. Further,

$$\begin{aligned}
&\mathbb{E}_{\mu_h}^{(\pi^\theta,(\widetilde{\mathbb{m}})_{(h+1)})}\Big[\|\widehat{\nabla} J_h^{K_h}(\theta, \widetilde{\mathbb{m}}_{(h+1)}, \mu_h) - \nabla J_h(\theta, \widetilde{\mathbb{m}}_{(h+1)}, \mu_h)\|^2\Big] \\
&\le \frac{1}{K_h}\mathbb{E}_{\mu_h}^{(\pi^\theta,(\widetilde{\mathbb{m}})_{(h+1)})}\Big[\sum_{a\in\mathcal{A}_s}(\mathbf{1}_{a=A_h} - \pi^\theta(a|S_h))^2\Big(\sum_{k=h}^{H-1} r(S_k, A_k)\Big)^2 \\
&\quad - 2\sum_{a\in\mathcal{A}_s}(\mathbf{1}_{a=A_h} - \pi^\theta(a|S_h)) \sum_{k=h}^{H-1} r(S_k, A_k)\mu_h(s)\pi^\theta(a|S_h) A_h^{(\pi^\theta,(\widetilde{\mathbb{m}})_{(h+1)})}(S_h, a) \\
&\quad + \sum_{s\in\mathcal{S}_h}\sum_{a\in\mathcal{A}_s}\mu_h(s)^2\pi^\theta(a|s)^2 A_h^{(\pi^\theta,(\widetilde{\mathbb{m}})_{(h+1)})}(s, a)^2\Big].
\end{aligned}$$

We consider all three terms separately. For the first term we have

$$\begin{aligned}
&\mathbb{E}_{\mu_h}^{(\pi^\theta,(\widetilde{\mathbb{m}})_{(h+1)})}\Big[\sum_{a\in\mathcal{A}_s}(\mathbf{1}_{a=A_h} - \pi^\theta(a|S_h))^2\Big(\sum_{k=h}^{H-1} r(S_k, A_k)\Big)^2\Big] \\
&= \mathbb{E}_{\mu_h}^{(\pi^\theta,(\widetilde{\mathbb{m}})_{(h+1)})}\Big[\Big(\sum_{k=h}^{H-1} r(S_k, A_k)\Big)^2\Big] - 2\mathbb{E}_{\mu_h}^{(\pi^\theta,(\widetilde{\mathbb{m}})_{(h+1)})}\Big[\pi^\theta(A_h|S_h)\Big(\sum_{k=h}^{H-1} r(S_k, A_k)\Big)^2\Big] \\
&\quad + \mathbb{E}_{\mu_h}^{(\pi^\theta,(\widetilde{\mathbb{m}})_{(h+1)})}\Big[\sum_{a\in\mathcal{A}_s}\pi^\theta(a|S_h)^2\Big(\sum_{k=h}^{H-1} r(S_k, A_k)\Big)^2\Big] \\
&\le ((H-h)R^*)^2 - 0 + ((H-h)R^*)^2 = 2((H-h)R^*)^2,
\end{aligned}$$

by bounded reward assumption and the fact that π^θ is a probability distribution. For the second term, we note that $A_h^{(\pi^\theta,(\widetilde{\mathbb{m}})_{(h+1)})}(S_h,a)$ can be negative, therefore we consider the absolute value and obtain

$$\begin{aligned}
&2\mathbb{E}_{\mu_h}^{(\pi^\theta,(\widetilde{\mathbb{m}})_{(h+1)})}\Big[\sum_{a\in\mathcal{A}_s}(\mathbf{1}_{a=A_h}-\pi^\theta(a|S_h))\sum_{k=h}^{H-1}r(S_k,A_k)\mu_h(s)\pi^\theta(a|S_h)\big|A_h^{(\pi^\theta,(\widetilde{\mathbb{m}})_{(h+1)})}(S_h,a)\big|\Big]\\
&\le 2\mathbb{E}_{\mu_h}^{(\pi^\theta,(\widetilde{\mathbb{m}})_{(h+1)})}\Big[\sum_{a\in\mathcal{A}_s}1\cdot(H-h)R^*\cdot 1\cdot\pi^\theta(a|S_h)\cdot(H-h)R^*\Big]\\
&=2((H-h)R^*)^2.
\end{aligned}$$

For the last term we have

$$\mathbb{E}_{\mu_h}^{(\pi^\theta,(\widetilde{\mathbb{m}})_{(h+1)})}\Big[\sum_{s\in\mathcal{S}_h}\sum_{a\in\mathcal{A}_s}\mu_h(s)^2\pi^\theta(a|s)^2A_h^{(\pi^\theta,(\widetilde{\mathbb{m}})_{(h+1)})}(s,a)^2\Big]\le((H-h)R^*)^2.$$

In total, it holds that

$$\mathbb{E}_{\mu_h}^{(\pi^\theta,(\widetilde{\mathbb{m}})_{(h+1)})}\Big[\|\widehat{\nabla}J_h^{K_h}(\theta,\widetilde{\mathbb{m}}_{(h+1)},\mu_h)-\nabla J_h(\theta,\widetilde{\mathbb{m}}_{(h+1)},\mu_h)\|^2\Big]\le\frac{5((H-h)R^*)^2}{K_h}.$$

■

Define the stopping time. Let $(\bar\theta_h^{(n)})_{n\in\mathbb{N}}$ be the stochastic process from equation (4.22) and let $(\theta_h^{(n)})_{n\in\mathbb{N}}$ be the deterministic sequence generated by FT-DynPG with exact gradients,

$$\theta_h^{(n+1)}=\theta_h^{(n)}+\alpha_h\nabla J_h(\theta_h^{(n)},\widetilde{\mathbb{m}}_{(h+1)},\mu_h)$$

such that the initial parameter agree, i.e. $\theta_h^{(1)}=\bar\theta_h^{(1)}$, and the step size α_h is the same for both processes. The natural filtration of $(\bar\theta_h^{(n)})_{n\in\mathbb{N}}$ is denoted by $(\mathcal{F}_h^{(n)})_{n\in\mathbb{N}}$.
For the deterministic scheme we could assure that $c_h=\min_{n\in\mathbb{N}}\min_{s\in\mathcal{S}}\pi^{\theta_h^{(n)}}(a^*(s)|s)$ is bounded away from 0 by Lemma 4.20. As for the simultaneous PG this cannot be guaranteed for the stochastic trajectory. Define for every epoch the following stopping time

$$\tau_h:=\min\{n\ge1:\|\theta_h^{(n)}-\bar\theta_h^{(n)}\|_2\ge\frac{c_h}{4}\}.$$

We emphasize that τ_h is a stopping time with respect to the filtration $(\mathcal{F}_h^{(n)})_{n\in\mathbb{N}}$ by construction.

Event $\{n\le\tau_h\}$. It follows again by the $\sqrt{2}$-Lipschitz continuity of the softmax policies (Lemma 4.26) that $\min_{0\le n\le\tau_h}\min_{s\in\mathcal{S}}\pi^{\bar\theta_h^{(n)}}(a^*(s)|s)\ge\frac{c_h}{2}>0$.

Lemma 4.32. *Let μ_h be probability measures such that $\mu_h(s)>0$ for all $s\in\mathcal{S}_h$ and consider the stochastic sequence $(\bar\theta_h^{(n)})_{n\in\mathbb{N}}$ generated by equation* (4.22) *under softmax parametrization. Then, it holds almost surely that* $\min_{0\le n\le\tau_h}\min_{s\in\mathcal{S}_h}\pi^{\bar\theta_h^{(n)}}(a^*(s)|s)\ge\frac{c_h}{2}$ *is strictly positive.*

Proof. For every $n \leq \tau_h$ we obtain by the $\sqrt{2}$-Lipschitz continuity in Lemma 4.26 that

$$\begin{aligned}\pi^{\bar{\theta}_h^{(n)}}(a^*(s)|s) &\geq \pi^{\theta_h^{(n)}}(a^*(s)|s) - |\pi^{\theta_h^{(n)}}(a^*(s)|s) - \pi^{\bar{\theta}_h^{(n)}}(a^*(s)|s)| \\ &\geq \pi^{\theta_h^{(n)}}(a^*(s)|s) - \sqrt{2}\|\bar{\theta}_h^{(n)} - \theta_h^{(n)}\|_2 > \frac{c_h}{2} > 0,\end{aligned}$$

holds almost surely. The claim follows directly. ∎

We derive a convergence rate on the event $\{n \leq \tau_h\}$ in the following sense:

LEMMA 4.33. *Let μ_h be probability measures such that $\mu_h(s) > 0$ for all $s \in \mathcal{S}_h$ and consider the stochastic sequence $(\bar{\theta}_h^{(n)})_{n\in\mathbb{N}}$ generated by equation* (4.22) *under softmax parametrization. Suppose that*

(i) the batch size $K_h^{(n)} \geq \frac{45c_h^2}{64N_h^{\frac{3}{2}}}(1 - \frac{1}{2\sqrt{N_h}})n^2$ is increasing for some $N_h \geq 1$ and

(ii) the step size $\alpha_h = \frac{1}{2(H-h)R^*\sqrt{N_h}}$.

Then, it holds true that

$$\mathbb{E}\big[(J_h^*(\widetilde{\mathbb{m}}_{(h+1)}, \mu_h) - J_h(\bar{\theta}_h^{(n)}, \widetilde{\mathbb{m}}_{(h+1)}, \mu_h))\mathbf{1}_{\{n\leq\tau_h\}}\big] \leq \frac{32\sqrt{N_h}(H-h)R^*}{3(1-\frac{1}{2\sqrt{N_h}})c_h^2 n}.$$

Proof. As in the proof of Theorem 4.30 we deduce from the L_h-smoothness and Lemma 4.31, that

$$\begin{aligned}&\mathbb{E}\Big[J(\bar{\theta}_h^{(n+1)}, \widetilde{\mathbb{m}}_{(h+1)}, \mu_h)|\mathcal{F}_h^{(n)}\Big] \\ &\geq J(\bar{\theta}_h^{(n)}, \widetilde{\mathbb{m}}_{(h+1)}, \mu_h) + \Big(\alpha_h - \frac{L_h\alpha_h^2}{2}\Big)\|\nabla J(\bar{\theta}_h^{(n)}, \widetilde{\mathbb{m}}_{(h+1)}, \mu_h)\|^2 - \frac{L_h\alpha_h^2\psi_h}{2K_h^{(n)}}.\end{aligned}$$

We take the expectation of this inequality on both sides under the event $\{n+1 \leq \tau_h\}$. Note that $\{n+1 \leq \tau_h\} = \{\tau_h \leq n\}^C$ is $\mathcal{F}_n$-measurable and that $\mathbf{1}_{\{n+1\leq\tau_h\}} \leq \mathbf{1}_{\{n\leq\tau_h\}}$ a.s., thus

$$\begin{aligned}&\mathbb{E}\Big[(J_h^*(\widetilde{\mathbb{m}}_{(h+1)}, \mu_h) - J_h(\bar{\theta}_h^{(n+1)}, \widetilde{\mathbb{m}}_{(h+1)}, \mu_h))\mathbf{1}_{\{n+1\leq\tau_h\}}\Big] \\ &= \mathbb{E}\Big[\mathbb{E}\Big[(J_h^*(\widetilde{\mathbb{m}}_{(h+1)}, \mu_h) - J_h(\bar{\theta}_h^{(n+1)}, \widetilde{\mathbb{m}}_{(h+1)}, \mu_h))|\mathcal{F}_h^{(n)}\Big]\mathbf{1}_{\{n+1\leq\tau_h\}}\Big] \\ &\leq \mathbb{E}\Big[\Big(J_h^*(\widetilde{\mathbb{m}}_{(h+1)}, \mu_h) - \mathbb{E}\Big[J_h(\bar{\theta}_h^{(n+1)}, \widetilde{\mathbb{m}}_{(h+1)}, \mu_h)|\mathcal{F}_h^{(n)}\Big]\Big)\mathbf{1}_{\{n\leq\tau_h\}}\Big] \\ &\leq \mathbb{E}\Big[(J_h^*(\widetilde{\mathbb{m}}_{(h+1)}, \mu_h) - J_h(\bar{\theta}_h^{(n)}, \widetilde{\mathbb{m}}_{(h+1)}, \mu_h))\mathbf{1}_{\{n\leq\tau_h\}}\Big] \\ &\quad - \Big(\alpha_h - \frac{L_h\alpha_h^2}{2}\Big)\mathbb{E}\Big[\|\nabla J_h(\bar{\theta}_h^{(n)}, \widetilde{\mathbb{m}}_{(h+1)}, \mu_h)\|^2\mathbf{1}_{\{n\leq\tau_h\}}\Big] + \frac{L_h\alpha_h^2\psi_h}{2K_h^{(n)}} \\ &= \mathbb{E}\Big[(J_h^*(\widetilde{\mathbb{m}}_{(h+1)}, \mu_h) - J_h(\bar{\theta}_h^{(n)}, \widetilde{\mathbb{m}}_{(h+1)}, \mu_h))\mathbf{1}_{\{n\leq\tau_h\}}\Big] \\ &\quad - \alpha_h\Big(1 - \frac{1}{2\sqrt{N_h}}\Big)\mathbb{E}\Big[\|\nabla J_h(\bar{\theta}_h^{(n)}, \widetilde{\mathbb{m}}_{(h+1)}, \mu_h)\|^2\mathbf{1}_{\{n\leq\tau_h\}}\Big] + \frac{5(H-h)R^*}{2K_h^{(n)}N_h}.\end{aligned}$$

By Lemma 4.19 we have that

$$\|\nabla J_h(\bar{\theta}_h^{(n)}, \widetilde{\mathbb{m}}_{(h+1)}, \mu_h)\|^2 \geq \min_{s\in\mathcal{S}} \pi^{\bar{\theta}_h^{(n)}}(a^*(s|s))^2 (J_h^*(\widetilde{\mathbb{m}}_{(h+1)}, \mu_h) - J_h(\bar{\theta}_h^{(n)}, \widetilde{\mathbb{m}}_{(h+1)}, \mu_h))^2$$

almost surely, and by Lemma 4.32 we have that $\min_{1\leq n\leq\tau_h} \min_{s\in\mathcal{S}} \pi^{\bar{\theta}_h^{(n)}}(a^*(s|s))^2 \geq \frac{c_h}{2} > 0$ almost surly. Therefore,

$$\begin{aligned}
&\mathbb{E}\Big[(J_h^*(\widetilde{\mathbb{m}}_{(h+1)}, \mu_h) - J_h(\bar{\theta}_h^{(n+1)}, \widetilde{\mathbb{m}}_{(h+1)}, \mu_h)\mathbf{1}_{\{n+1\leq\tau_h\}}\Big] \\
&\leq \mathbb{E}\Big[(J_h^*(\widetilde{\mathbb{m}}_{(h+1)}, \mu_h) - J_h(\bar{\theta}_h^{(n)}, \widetilde{\mathbb{m}}_{(h+1)}, \mu_h))\mathbf{1}_{\{n\leq\tau_h\}}\Big] \\
&\quad - \alpha_h\Big(1 - \frac{1}{2\sqrt{N_h}}\Big)\mathbb{E}\Big[\min_{s\in\mathcal{S}} \pi^{\bar{\theta}_h^{(n)}}(a^*(s|s))^2 (J_h^*(\widetilde{\mathbb{m}}_{(h+1)}, \mu_h) - J_h(\bar{\theta}_h^{(n)}, \widetilde{\mathbb{m}}_{(h+1)}, \mu_h))^2\mathbf{1}_{\{n\leq\tau_h\}}\Big] + \frac{5(H-h)R^*}{2K_h^{(n)}N_h}, \\
&\leq \mathbb{E}\Big[(J_h^*(\widetilde{\mathbb{m}}_{(h+1)}, \mu_h) - J_h(\bar{\theta}_h^{(n)}, \widetilde{\mathbb{m}}_{(h+1)}, \mu_h))\mathbf{1}_{\{n\leq\tau_h\}}\Big] \\
&\quad - \alpha_h\Big(1 - \frac{1}{2\sqrt{N_h}}\Big)\frac{c_h^2}{4}\mathbb{E}\Big[(J_h^*(\widetilde{\mathbb{m}}_{(h+1)}, \mu_h) - J_h(\bar{\theta}_h^{(n)}, \widetilde{\mathbb{m}}_{(h+1)}, \mu_h))\mathbf{1}_{\{n\leq\tau_h\}}\Big]^2 + \frac{5(H-h)R^*}{2K_h^{(n)}N_h},
\end{aligned}$$

where we used Jensen's inequality in the last step.
For $d_n := \mathbb{E}\Big[(J_h^*(\widetilde{\mathbb{m}}_{(h+1)}, \mu_h) - J_h(\bar{\theta}_h^{(n)}, \widetilde{\mathbb{m}}_{(h+1)}, \mu_h))\mathbf{1}_{\{n\leq\tau_h\}}\Big]$ we imply the recursive inequality

$$d_{n+1} \leq d_n - \alpha_h\Big(1 - \frac{1}{2\sqrt{N_h}}\Big)\frac{c_h^2}{4}d_n^2 + \frac{5(H-h)R^*}{2K_h^{(n)}N_h}.$$

Define $w := \alpha_h\Big(1 - \frac{1}{2\sqrt{N_h}}\Big)\frac{c_h^2}{4} > 0$ and $B = \frac{5(H-h)R^*}{2N_h} > 0$, then

$$d_{n+1} \leq d_n(1 - wd_n) + \frac{B}{K_h^{(n)}}$$

and by our choice of α_h,

$$K_h^{(n)} \geq \frac{45c_h^2}{64N_h^{\frac{3}{2}}}(1 - \frac{1}{2\sqrt{N_h}})n^2 = \frac{9}{4}wBn^2,$$

Moreover, it holds that

$$d_1 \leq (H-h)R^* \leq \frac{1}{\alpha_h} \leq \frac{4}{3w} \leq \frac{4}{3w\cdot 1},$$

because $c_h \leq 1$ and $\frac{1}{\sqrt{N_h}}(1 - \frac{1}{2\sqrt{N_h}}) < 1$ for all $N_h \geq 1$. Suppose the induction assumption $d_n \leq \frac{4}{3wn}$ holds true, then for d_{n+1},

$$d_{n+1} \leq d_n - wd_n^2 + \frac{B}{K_h^{(n)}}.$$

The function $f(x) = x - wx^2$ is monotonically increasing in $[0, \frac{1}{2w}]$ and by induction assumption $d_n \le \frac{1}{4wn} \le \frac{1}{2w}$. So $d_n - wd_n^2 \le \frac{4}{3wn}$ which implies

$$\begin{aligned} d_{n+1} &\le d_n - wd_n^2 + \frac{B}{K_h^{(n)}} \le \frac{4}{3wn} - \frac{16}{9wn^2} + \frac{B}{K_n} \\ &\le \frac{4}{3wn} - \frac{16}{9wn^2} + \frac{4B}{9wBn^2} = \frac{4}{3wn} - \frac{12}{9wn^2} = \frac{4}{3w}\Big(\frac{1}{n} - \frac{1}{n^2}\Big) \\ &\le \frac{4}{3w(n+1)}, \end{aligned}$$

where we used that $K_h^{(n)} \ge \frac{9}{4}wBn^2$. We deduce the claim

$$d_n \le \frac{4}{3wn} = \frac{32\sqrt{N_h}(H-h)R^*}{3(1-\frac{1}{2\sqrt{N_h}})c_h^2 n}.$$

■

Event $\{\tau_h \le n\}$. Secondly, consider the complementary event $\{\tau_h \le n\}$. We can bound the probability of this event by δ for a large enough batch size K_h. The proof is again inspired by similar results obtained in Ding, Zhang, and Lavaei [DZL22, Lem. 6.3] for discounted MDPs.

LEMMA 4.34. *Let μ_h be probability measures such that $\mu_h(s) > 0$ for all $s \in \mathcal{S}_h$ and consider the stochastic sequence $(\bar{\theta}_h^{(n)})_{n\in\mathbb{N}}$ generated by equation (4.22) under softmax parametrization. For any $\delta > 0$, suppose that*

(i) the batch size $K_h \ge \frac{5n^3}{c_h^2\delta^2}$

(ii) the step size $\alpha_h = \frac{1}{\sqrt{n}L_h}$.

Then, it holds true that $\mathbb{P}(\tau_h \le n) < \delta$.

Proof. The proof follows line by line the one of Lemma 4.29. One obtains

$$\mathbb{P}(\tau_h \le n) = \mathbb{P}\Big(\max_{1\le t\le n} \|\theta_h^{(t)} - \bar{\theta}_h^{(t)}\| \ge \frac{c_h}{4}\Big) \le \frac{4n\alpha_h(1+\alpha_h L_h)^{n-1}\sqrt{\frac{\psi_h}{K_h}}}{c_h},$$

where ψ_h from Lemma 4.31. Now we plug in the choice of $\alpha_h = \frac{1}{\sqrt{n}L_h} = \frac{1}{2(H-h)R^*\sqrt{n}}$,

$$\begin{aligned} \mathbb{P}(\tau_h \le n) &\le \frac{4n\frac{1}{\sqrt{n}L_h}(1+\frac{1}{\sqrt{n}L_h}L_h)^{n-1}\sqrt{\frac{\xi_h}{K_h}}}{c_h} \\ &= \frac{4\sqrt{n}(1+\frac{1}{\sqrt{n}})^{n-1}\sqrt{\psi_h}}{L_h c_h\sqrt{K_h}} \\ &\le \frac{2n\sqrt{n}\sqrt{\psi_h}}{L_h c_h\sqrt{K_h}} = \frac{n\sqrt{5n}}{c_h\sqrt{K_h}}, \end{aligned}$$

where the last step is due to $f(x) = (1+\frac{1}{\sqrt{x}})^{x-1} \le x$ for all $x \ge 1$. We conclude that $\mathbb{P}(\tau_h < n) < \delta$ if $K_h \ge \frac{5n^3}{c_h^2\delta^2}$. ■

Convergence result. We are now ready to proof the epoch wise statement.

LEMMA 4.35. *Let μ_h be probability measures such that $\mu_h(s) > 0$ for all $s \in \mathcal{S}_h$ and consider the stochastic sequence $(\bar{\theta}_h^{(n)})_{n\in\mathbb{N}}$ generated by equation (4.22) under softmax parametrization. Moreover, for any $\delta, \epsilon > 0$, assume that*

(i) the number of training steps $N_h \geq \big(\frac{12(H-h)R^*}{\epsilon\delta c_h^2}\big)^2$,

(ii) the step size $\alpha_h = \frac{1}{2(H-h)R^*\sqrt{N_h}}$ and

(iii) the batch size $K_h = \frac{5N_h^3}{c_h^2\delta^2}$.

Then, it holds true that $\mathbb{P}\big(J_h^*(\widetilde{\mathbb{m}}_{(h+1)}, \mu_h) - J_h(\bar{\theta}_h^{(N_h)}, \widetilde{\mathbb{m}}_{(h+1)}, \mu_h) \geq \epsilon\big) \leq \delta$.

Proof. We separate the probability using the stopping time τ_h and obtain

$$
\begin{aligned}
&\mathbb{P}\Big(J_h^*(\widetilde{\mathbb{m}}_{(h+1)}, \mu_h) - J_h(\bar{\theta}_h^{(N_h)}, \widetilde{\mathbb{m}}_{(h+1)}, \mu_h) \geq \epsilon\Big) \\
&\leq \mathbb{P}\Big(\{\tau_h \geq N_h\} \cap \{J_h^*(\widetilde{\mathbb{m}}_{(h+1)}, \mu_h) - J_h(\bar{\theta}_h^{(N_h)}, \widetilde{\mathbb{m}}_{(h+1)}, \mu_h) \geq \epsilon\}\Big) \\
&\quad + \mathbb{P}\Big(\{\tau_h \leq N_h\} \cap \{J_h^*(\widetilde{\mathbb{m}}_{(h+1)}, \mu_h) - J_h(\bar{\theta}_h^{(N_h)}, \widetilde{\mathbb{m}}_{(h+1)}, \mu_h) \geq \epsilon\}\Big) \\
&\leq \frac{\mathbb{E}\Big[(J_h^*(\widetilde{\mathbb{m}}_{(h+1)}, \mu_h) - J_h(\bar{\theta}_h^{(N_h)}, \widetilde{\mathbb{m}}_{(h+1)}, \mu_h))\mathbf{1}_{\{\tau_h \geq N_h\}}\Big]}{\epsilon} + \mathbb{P}(\tau_h \leq N_h) \\
&\leq \frac{1}{\epsilon}\frac{32\sqrt{N_h}(H-h)R^*}{3(1-\frac{1}{2\sqrt{N_h}})c_h^2 n} + \frac{\delta}{2} \\
&\leq \frac{\delta}{2} + \frac{\delta}{2} \\
&= \delta,
\end{aligned}
$$

where the second inequality it due to Lemma 4.33 and Lemma 4.34. The last inequality follows by our choice of N_h:

$$
\frac{32\sqrt{N_h}(H-h)R^*}{3\epsilon(1-\frac{1}{2\sqrt{N_h}})c_h^2 n} \leq \frac{11\sqrt{N_h}(H-h)R^*}{\epsilon(1-\frac{1}{2\sqrt{N_h}})c_h^2 n} \leq \frac{\delta}{2}
$$

for $N_h \geq \big(\frac{11(H-h)R^*}{\epsilon\delta c_h^2} + \frac{1}{2}\big)^2$, which is satisfied for $N_h \geq \big(\frac{12(H-h)R^*}{\epsilon\delta c_h^2}\big)^2$. Note further that we could use Lemma 4.33 in the equation above with a constant batch size K_h, because

$$
\max\Big\{\frac{45c_h^2}{64N_h^{\frac{3}{2}}}(1-\frac{1}{2\sqrt{N_h}})n^2, \frac{5N_h^3}{c_h^2\delta^2}\Big\} = \frac{5N_h^3}{c_h^2\delta^2},
$$

for all $n \leq N_h$, as $(1-\frac{1}{2\sqrt{N_h}}) < 1$ and $c_h < 1$. ∎

Our main result for the dynamic stochastic PG scheme is given as follows.

THEOREM 4.36. *For all $h \in \mathcal{H}$, let μ_h be probability measures such that $\mu_h(s) > 0$ for all $h \in \mathcal{H}$, $s \in \mathcal{S}_h$. Consider the final policy using Algorithm 5 with stochastic updates from equation* (4.22) *under softmax parametrization and denote by $\widehat{\mathbb{\pi}}^* = (\pi^{\bar{\theta}_0^{(N_0)}}, \dots, \pi^{\bar{\theta}_{H-1}^{(N_{H-1})}})$ the final policy. Moreover, for any $\delta, \epsilon > 0$ assume that*

(i) the numbers of training steps satisfy $N_h \geq \left(\frac{12(H-h)R^*H^2\left\|\frac{1}{\mu_h}\right\|_\infty}{\delta c_h^2 \epsilon}\right)^2$,

(ii) the step size $\alpha_h = \frac{1}{2(H-h)R^*\sqrt{N_h}}$ and

(iii) the batch size $K_h \geq \frac{5N_h^3H^2}{c_h^2\delta^2}$.

Then, it holds true that

$$\mathbb{P}\left(\forall s \in \mathcal{S}_0 : V_0^*(s) - V_0^{\widehat{\mathbb{\pi}}^*}(s) < \epsilon\right) > 1 - \delta.$$

Proof. As in the proof of the exact gradient case (Theorem 4.24, equation (4.14)) we have by our choice of the future policy $\widetilde{\mathbb{\pi}} = \widehat{\mathbb{\pi}}^*$ that

$$J_h(\bar{\theta}_h^{(N_h)}, \widetilde{\mathbb{\pi}}_{(h+1)}, \delta_s) = V_h^{\widehat{\mathbb{\pi}}^*}(s). \tag{4.23}$$

By Lemma 4.35 we have that

$$\mathbb{P}\left(J_h^*(\widetilde{\mathbb{\pi}}_{(h+1)}, \mu_h) - J_h(\bar{\theta}_h^{(N_h)}, \widetilde{\mathbb{\pi}}_{(h+1)}, \mu_h) \geq \frac{\epsilon}{H\left\|\frac{1}{\mu_h}\right\|_\infty}\right) \leq \frac{\delta}{H},$$

by our choice of N_h, α_h and K_h.
For every $s \in \mathcal{S}_h$, denote by δ_s the dirac measure on state s, then as in equation (4.15)

$$J_h^*(\widetilde{\mathbb{\pi}}_{(h+1)}, \delta_s) - J_h(\bar{\theta}_h^{(N_h)}, \widetilde{\mathbb{\pi}}_{(h+1)}, \delta_s) \leq \left\|\frac{1}{\mu_h}\right\|_\infty (J_h^*(\widetilde{\mathbb{\pi}}_{(h+1)}, \mu_h) - J_h(\theta_h^{(N_h)}, \widetilde{\mathbb{\pi}}_{(h+1)}, \mu_h)) \quad \text{a.s.}$$

Thus, for all $h \in \mathcal{H}$ it holds that

$$\begin{aligned}
&\mathbb{P}\left(\exists s \in \mathcal{S}_h : J_h^*(\widetilde{\mathbb{\pi}}_{(h+1)}, \delta_s) - J_h(\bar{\theta}_h^{(N_h)}, \widetilde{\mathbb{\pi}}_{(h+1)}, \delta_s) \geq \frac{\epsilon}{H}\right) \\
&\leq \mathbb{P}\left(J_h^*(\widetilde{\mathbb{\pi}}_{(h+1)}, \mu_h) - J_h(\theta_h^{(N_h)}, \widetilde{\mathbb{\pi}}_{(h+1)}, \mu_h) \geq \frac{\epsilon}{H\left\|\frac{1}{\mu_h}\right\|_\infty}\right) \leq \frac{\delta}{H}.
\end{aligned} \tag{4.24}$$

Define the event $A_h := \{J_h^*(\widetilde{\mathbb{\pi}}_{(h+1)}, \delta_s) - J_h(\bar{\theta}_h^{(N_h)}, \widetilde{\mathbb{\pi}}_{(h+1)}, \delta_s) < \frac{\epsilon}{H}, \ \forall s \in \mathcal{S}_h\}$. Then equation (4.24) states that $\mathbb{P}(A_h^C) \leq \frac{\delta}{H}$. For $h = H-1$ it follows directly with equation (4.23) and the special property of the last time point that

$$\begin{aligned}
&\mathbb{P}\left(\exists s \in \mathcal{S}_h : V_{H-1}^*(s) - V_{H-1}^{\widehat{\mathbb{\pi}}^*}(s) \geq \frac{\epsilon}{H}\right) \\
&= \mathbb{P}\left(\exists s \in \mathcal{S}_h : J_{H-1}^*(\delta_s) - J_{H-1}(\bar{\theta}_h^{(N_h)}, \delta_s) \geq \frac{\epsilon}{H}\right) \leq \frac{\delta}{H}.
\end{aligned}$$

We close the proof by induction. Assume for some $0 < h < H$ that

$$\mathbb{P}\Big(\exists s \in \mathcal{S}_h : V_h^*(s) - V_h^{\widehat{\mathbb{m}}^*}(s) \geq \frac{\epsilon(H-h)}{H}\Big) \leq \frac{\delta(H-h)}{H}. \tag{4.25}$$

Define $B_h := \{V_h^*(s) - V_h^{\widehat{\mathbb{m}}^*}(s) < \frac{\epsilon(H-h)}{H}, \forall s \in \mathcal{S}_h\}$. Similar to equation (4.17), on the event B_h it holds that

$$\begin{aligned} J_{h-1}^*(\widetilde{\mathbb{m}}_{(h)}, \delta_s) &= \max_{a \in \mathcal{A}_s}\Big(r(s,a) + \sum_{s' \in \mathcal{S}_h} p(s'|s,a)V_h^*(s) - \sum_{s' \in \mathcal{S}_h} p(s'|s,a)(V_h^*(s) - V_h^{\widehat{\mathbb{m}}^*}(s))\Big) \\ &> \max_{a \in \mathcal{A}_s}\Big(r(s,a) + \sum_{s' \in \mathcal{S}_h} p(s'|s,a)V_h^*(s)\Big) - \frac{\epsilon(H-h)}{H} \\ &= V_{h-1}^*(s) - \frac{\epsilon(H-h)}{H}. \end{aligned}$$

We obtain on the event $A_{h-1} \cap B_h$ that (compare to equation (4.18))

$$\begin{aligned} V_{h-1}^*(s) - V_{h-1}^{\widehat{\mathbb{m}}^*}(s) &= V_{h-1}^*(s) - J_{h-1}^*(\widetilde{\mathbb{m}}_{(h)}, \delta_s) + J_{h-1}^*(\widetilde{\mathbb{m}}_{(h)}, \delta_s) - V_{h-1}^{\widehat{\mathbb{m}}^*}(s) \\ &< \frac{\epsilon(H-h)}{H} + \frac{\epsilon}{H} \\ &= \frac{\epsilon(H-(h-1))}{H}, \end{aligned}$$

for every $s \in \mathcal{S}_{h-1}$. Hence, $A_{h-1} \cap B_h \subset B_{h-1}$. Finally, we close the induction by

$$\begin{aligned} &\mathbb{P}\Big(\exists s \in \mathcal{S}_{h-1} : V_{h-1}^*(s) - V_{h-1}^{\widehat{\mathbb{m}}^*}(s) \geq \frac{\epsilon(H-(h-1))}{H}\Big) \\ &= 1 - \mathbb{P}(B_{h-1}) \leq 1 - \mathbb{P}(A_{h-1} \cap B_h) = \mathbb{P}(A_{h-1}^C \cup B_h^C) \leq \mathbb{P}(A_{h-1}^C) + \mathbb{P}(B_h^C) \\ &= \mathbb{P}\Big(\exists s \in \mathcal{S}_{h-1} : J_{h-1}^*(\widetilde{\mathbb{m}}_{(h)}, \delta_s) - J_{h-1}(\theta_{h-1}^{(N_h-1)}, \widetilde{\mathbb{m}}_{(h)}, \delta_s) \geq \frac{\epsilon}{H}\Big) \\ &\quad + \mathbb{P}\Big(\exists s \in \mathcal{S}_h : V_h^*(s) - V_h^{\widehat{\mathbb{m}}^*}(s) \geq \frac{\epsilon(H-h)}{H}\Big) \\ &\leq \frac{\delta}{H} + \frac{\delta(H-h)}{H} \\ &= \frac{\delta(H-(h-1))}{H}. \end{aligned}$$

Finally, for $h = 0$ we have shown the assertion $\mathbb{P}\Big(\exists s \in \mathcal{S}_0 : V_0^*(s) - V_0^{\widehat{\mathbb{m}}^*}(s) \geq \epsilon\Big) \leq \delta$. ∎

4.4.3 *Comparison*

In both scenarios the derived complexity bounds for the stochastic algorithms use a very large batch size and small step size and we do not expect these rates to be tight. Similar large batch size and step size is also needed to prove convergence in entropy regularized infinite-time horizon SPG [DZL22]. It should also be noted that the choice of step size and batch size are closely connected and both strongly depend on the number of training steps N. Specifically, as N increases, the batch size increases, while the step size tends to decrease to prevent exceeding

the stopping time with high probability. However, it is possible to increase the batch size even further and simultaneously benefit from choosing a larger step size, or vice versa.
An advantage of the dynamic approach is that c_h can be explicitly known for uniform initialization. Hence, the complexity bounds for the dynamic approach results in a practicable algorithm, while c is unknown and possibly arbitrarily small for the simultaneous approach such that we can not determine K.
Finally, we compare the complexity with respect to the time horizon. For the simultaneous approach the number of training steps scales with H^{10}, and the batch size with H^{30}, while in the dynamic approach the overall number of training steps scale with H^7 and the batch size with H^{20}. We are aware that these bounds are far from tight and irrelevant for practical implementations. Nevertheless, these bounds highlight once more the advantage of the dynamic approach in comparison to the simultaneous approach and show (the non-trivial fact) that the algorithms can be made to converge without knowledge of exact gradients and without regularization.

Dynamic Policy Gradient for discounted MDPs

5

In the previous chapter we discussed a dynamic approach for policy gradient to tackle finite-time MDPs. The motivation to utilize dynamic programming stemmed from seeking non-stationary optimal policies in finite-time horizons. In this chapter however, our objective is the discounted infinite-time horizon MDP and we aim to identify a stationary optimal policy. Nevertheless, we explain in the following why dynamic policy gradient (DynPG), a combination of dynamic programming and PG, is sometimes a good idea also for discounted MDPs.

The discount factor $\gamma \in [0, 1)$ plays an essential role in the convergence behavior of RL algorithms, as convergence explicitly depends on the contraction property of the Bellman operator. The closer γ to one, the slower the Bellman operator contracts. Similarly, the convergence of PG methods also heavily depends on γ but establishing a clear dependence is generally challenging due to the non-convex optimization landscape. In Section 3.1.2, we obtained a sub-linear convergence rate for vanilla softmax PG such that dependencies on other salient but essential parameters of the MDP crucially affect the overall convergence behavior of PG methods. Recall, that a model-dependent and unknown constant $c_g g$ appeared in the convergence rate (cf. Theorem 3.21), which can in general depend on γ and other model-dependent parameters. Notably, [Li+23a] constructed a counterexample such that vanilla PG could take an exponential time with respect to $(1-\gamma)^{-1}$ to converge. Although the optimal solution can be reached in just $|\mathcal{S}|$ steps of exact value iteration in the counter example, vanilla softmax PG is very inefficient by not leveraging the inherent structure of the MDP.

We tackle this issue by introducing DynPG. The algorithm is motivated by FT-DynPG and modified to the infinite-time horizon. The idea is based on two observations: First recall that $(1-\gamma)^{-1}$, the expected horizon of infinite-time problems, correlates with the deterministic time horizon H in finite-time MDPs (Remark 3.17). Second recall that we could delineate the dependence on the unknown parameter c under uniform initialization in softmax FT-DynPG and established an explicit dependency on the finite-time horizon H (Remark 4.22). We will see in this chapter that the convergence rate of DynPG scales with $(1-\gamma)^{-4}$ up to logarithmic factors and the unknown model-dependent constant in the rate of vanilla softmax PG can be omitted. As a result, DynPG can efficiently address the counterexample of [Li+23a]; see Section 5.4.1. We summarize the complexity bounds for softmax PG and softmax DynPG in Table 5.1.

algorithm	complexity bounds	reference
softmax PG upper bound	$O\big(c_\gamma (1-\gamma)^{-4}\epsilon^{-1}\big)$	Theorem 3.21
softmax PG lower bound	$\lvert\mathcal{S}\rvert^{2^{\Omega((1-\gamma)^{-1})}}$ gradient steps for $\epsilon = 0.15$	[Li+23a, Thm. 1]
softmax DynPG upper bound	$O\big((1-\gamma)^{-4}\epsilon^{-1}\log((1-\gamma)^{-2}\epsilon^{-1})\big)$	[**new** Thm. 5.19]

Table 5.1: Comparison of convergence rate under exact gradients

The main contributions in this chapter are three-fold:

1. In Section 5.2, we introduce the algorithm DynPG that directly combines dynamic programming and policy gradient. The algorithm (provably) circumvents recent worst case lower bound problems for standard PG.

2. In Section 5.3, we provide a detailed error decomposition to outline the convergence behavior of DynPG in general frameworks. Afterwards, we derive rigorous upper bounds on the convergence rate under tabular softmax parametrization. The γ-dependence is explicit and turns out to be more suitable compared to vanilla PG.

3. In Section 5.4, we discuss the application of DynPG in the the lower bound example constructed in [Li+23a] and provide a numerical toy example to verify the performance of DynPG in stochastic settings. Afterwards, we discuss the limitations and modifications of DynPG for practical usage and finally compare the performance of DynPG to NPG.

As a final remark, we want to point out that the idea of using dynamic programming in searching the optimal policy is not new and dates back to the early 2000s, having been revisited several times in recent decades. Policy search by dynamic programming (PSDP) [Bag+03; Kak03; Sch14] is most closely related to DynPG and searches for optimal policies in a restricted policy class without explicitly using gradient ascent to find the optimal policy. There is a line of similar algorithms like approximate policy iteration (API), gradient temporal difference, policy dynamic programming and numerous other variations [BT96a; SMS08; Sut+09; AGK12; KL02; Sch+15a]. In these works, however, PG was not used as policy optimization step.

5.1 Preliminaries and Notation

In the following we consider an infinite-time horizon MDP, $(\mathcal{S}, \mathcal{A}, \gamma, p, r)$, as introduced in Section 3.1. Due to a finite state and action space we can assume bounded rewards, more precisely we assume $r(s,a) \in [-R^*, R^*]$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$. In Section 3.1.1, we discussed dynamic programming in discounted MDPs and we have seen that applying the Bellman optimality operator infinitely often to any function leads to converge towards the optimal value function, i.e. $\lim_{n\to\infty} T^*(V) = V^*$. In comparison for finite-time MDPs, we need to apply the operators T_h^* for $h = H-1, \dots, 0$ backwards in time to an initial $V_H \equiv 0$ (cf. Section 3.2.1).
The idea for FT-DynPG was to approximate these optimal operators by policy gradient in a backward inductive manner (cf. Remark 4.3). The idea of DynPG for discounted MDPs is to train a sequence of policies $(\pi^{\theta_t})_{t=0}^{\infty}$ forward in time (from $t = 0, 1, \dots$) such that at time point t, $\theta_t^* \approx \operatorname{argmax}_\theta T^{\pi^\theta} T^{\pi^{\theta_{t-1}^*}} \cdots T^{\pi^{\theta_0^*}}(\mathcal{V}_0)$ with the convention $\mathcal{V}_0 \equiv 0$. As the training of a new policy happens by adding an additional Bellman operator, we can interpret the procedure as a forward dynamic programming principle, which adds a new time step at the beginning. We will explain this in more detail in Section 5.2.
To analyze this forward inductive procedure, we introduce time dependent non-stationary policies with time-horizon h, denoted by $\pi_h := (\pi_{h-1}, \dots, \pi_0) \in \Pi^h$. Note that the time-indexing is reversed compared to Definition 3.25. From now on, we always consider this reverse ordering and refer to $h \in \mathbb{N}$ as the deterministic finite-time horizon, where the case $h = \infty$ corresponds to the standard infinite-time horizon MDP.

DEFINITION 5.1. Let μ be an initial state distribution on $\mathcal{S}$. Define $\mathcal{V}_0 \equiv 0$ and for all $h > 0$,$h = \infty$ allowed, we define the truncated h-step value function under the policies $\mathbb{\pi}_h = (\pi_{h-1}, \dots, \pi_0) \in \Pi^h$ as

$$\mathcal{V}_h^{\mathbb{\pi}_h}(\mu) = \mathcal{V}_h^{(\pi_{h-1},\dots,\pi_0)}(\mu) := \mathop{\mathbb{E}}_{\substack{S_0\sim\mu, A_t\sim\pi_{h-t-1}(\cdot|S_t)\\ S_{t+1}\sim p(\cdot|S_t,A_t)}}\Big[\sum_{t=0}^{h-1}\gamma^t r(S_t, A_t)\Big]. \tag{5.1}$$

When μ is a Dirac measure at s we let $\mathcal{V}_h^{\mathbb{\pi}_h}(s) := \mathcal{V}_h^{\mathbb{\pi}_h}(\delta_s)$.

Similar as for the infinite-time horizon MDP, we use $\mathcal{V}_h^{\pi}(\mu)$ to denote the value function of the stationary policy $\pi \in \Pi$ being applied h times in a row. Note that $\mathcal{V}_\infty$ equals the originally defined value function V in Definition 3.3. To be consistent, we will always use $\mathcal{V}_\infty$ in the following. We use the calligraphic notation to distinguish from the finite-time value function in Definition 3.26.

Remark 5.2. The finite-time value function V_{H-h} defined in Definition 3.26, where h time steps from $H-h$ to $H-1$ are considered, differs from $\mathcal{V}_h$ in Definition 5.1 by an index-shift in the time space and therefore also different discounting. More precisely, if the state space $\mathcal{S}$ is stationary in a finite-time MDP and a policy $\mathbb{\pi}_h = (\pi_{h-1}, \dots, \pi_0) \in \Pi^h$ for discounted MDPs and a policy $\widetilde{\mathbb{\pi}}_{(H-h)} = (\widetilde{\pi}_{H-h}, \dots, \widetilde{\pi}_{H-1})$ for finite-time MDPs agree, i.e. $\pi_j = \widetilde{\pi}_{H-1-j}$, then the value function $\mathcal{V}_h^{\mathbb{\pi}_h}(\mu)$ defined in Definition 5.1 is equal to $\gamma^{h-H} V_{H-h}^{\widetilde{\mathbb{\pi}}_{(H-h)}}(\mu)$ defined in Definition 3.26. Thus, the functions only differ by a constant which is due to the different discounting.

Recall from Section 3.1, that for $h = \infty$ the resulting infinite-time horizon discounted MDP admits a stationary optimal policy. We define $\mathcal{V}_\infty^*(\mu) := \sup_{\pi\in\Pi} \mathcal{V}_\infty^{\pi}(\mu)$ and use π^* to denote a stationary policy that achieves $\mathcal{V}_\infty^{\pi^*}(\mu) = \mathcal{V}_\infty^*(\mu)$. In contrast, when h is finite, the finite-horizon MDP optimization problem needs non-stationary optimal policies; thus, we define $\mathcal{V}_h^*(\mu) := \sup_{\mathbb{\pi}_h\in\Pi^h} \mathcal{V}_h^{\mathbb{\pi}_h}(\mu)$ and use $\mathbb{\pi}_h^* := (\pi_{h-1}^*, \dots, \pi_0^*) \in \Pi^h$ to denote a sequence of policies that achieves $\mathcal{V}_h^{\mathbb{\pi}_h^*}(\mu) = \mathcal{V}_h^*(\mu)$.
Recall the dynamic programming principle for discounted MDPs described in Section 3.1.1, where we established that $\mathcal{V}_\infty$ can be approximated by applying the Bellman optimality operator T^* iteratively. In addition, the optimal h-step value functions $\mathcal{V}_h^*$ can also be obtained by applying T^* h-times to $\mathcal{V}_0 \equiv 0$ and as $h \to \infty$, $\mathcal{V}_h^*$ converges to $\mathcal{V}_\infty^*$.

LEMMA 5.3. *For any $h \geq 1$, it holds that*

(i) $\mathcal{V}_h^*(s) = T^*(\mathcal{V}_{h-1}^*)(s)$*, for all* $s \in \mathcal{S}$*,*

(ii) $\|\mathcal{V}_\infty^* - \mathcal{V}_h^*\|_\infty \leq \frac{\gamma^h}{1-\gamma} R^*$.

Proof. The claim and proof is similar to [Ber01, Prop. 1.2.1].
The first claim, follows directly from the definition of the Bellman optimality operator (Definition 3.11)

$$\begin{aligned}\mathcal{V}_h^*(s) &= \sup_{\mathbb{\pi}_h\in\Pi^h} \sum_{a\in\mathcal{A}} \pi_0(a|s)\left(r(s,a) + \gamma \sum_{s'\in\mathcal{S}} p(s'|s,a)\mathcal{V}_{h-1}^{\mathbb{\pi}_{h-1}}(s')\right)\\ &= \max_{a\in\mathcal{A}}\left(r(s,a) + \gamma \sum_{s'\in\mathcal{S}} p(s'|s,a) \sup_{\mathbb{\pi}_{h-1}\in\Pi^{h-1}} \mathcal{V}_{h-1}^{\mathbb{\pi}_{h-1}}(s')\right) = T^*(\mathcal{V}_{h-1}^*)(s).\end{aligned}$$

For the second part, we fix $h \geq 0$. Recall that $\pi^* \in \Pi$ denotes a stationary optimal policy for the infinite-time problem and $\mathbb{\pi}_h^* \in \Pi_h$ an optimal non-stationary policy.
We divide the proof of the second claim into two cases.
Case 1: Assume that $\mathcal{V}_\infty^*(s) - \mathcal{V}_h^*(s) \leq 0$ for all $s \in \mathcal{S}$. Note that $\mathcal{V}_\infty^*(s) \geq \mathcal{V}_\infty^{(\mathbb{\pi}_h^*)_\infty}(s)$, where $(\mathbb{\pi}_h^*)_\infty$ denotes that we apply the finite time policy $\mathbb{\pi}_h^*$ in a loop for the infinite time problem. We have

$$\begin{aligned}\mathcal{V}_h^{\mathbb{\pi}_h^*}(s) - \mathcal{V}_\infty^*(s) &\leq \mathcal{V}_h^{\mathbb{\pi}_h^*}(s) - \mathcal{V}_\infty^{(\mathbb{\pi}_h^*)_\infty}(s) \\ &= - \mathop{\mathbb{E}}_{\substack{S_0=s, A_t \sim \pi_{h-(t \bmod h)-1}(\cdot|S_t) \\ S_{t+1} \sim p(\cdot|S_t, A_t)}} \left[\sum_{t=h}^{\infty} \gamma^t r(S_t, A_t) \right] \\ &\leq \sum_{t=h}^{\infty} \gamma^t R^* = \frac{\gamma^h}{1-\gamma} R^*,\end{aligned}$$

where R^* bounds the absolute value of rewards.
Case 2: Assume that $\mathcal{V}^*(s) - \mathcal{V}_h^*(s) > 0$ for all $s \in \mathcal{S}$. Note that $\mathcal{V}_h^{\mathbb{\pi}_h^*}(s) \geq \mathcal{V}_h^{\pi^*}(s)$ due to the optimality of $\mathbb{\pi}_h^*$ over the finite-time horizon. Further, by the definition of $\mathcal{V}_\infty^*(s) = \mathcal{V}_\infty^{\pi^*}(s)$, we have

$$\begin{aligned}\mathcal{V}_\infty^*(s) - \mathcal{V}_h^{\mathbb{\pi}_h^*}(s) &\leq \mathcal{V}_\infty^{\pi^*}(s) - \mathcal{V}_h^{\pi^*}(s) \\ &= \mathop{\mathbb{E}}_{\substack{S_0=s, A_t \sim \pi^*(\cdot|S_t) \\ S_{t+1} \sim p(\cdot|S_t, A_t)}} \left[\sum_{t=h}^{\infty} \gamma^t r(S_t, A_t) \right] \\ &\leq \sum_{t=h}^{\infty} \gamma^t R^* = \frac{\gamma^h}{1-\gamma} R^*.\end{aligned}$$

Hence, we arrive at

$$\|\mathcal{V}_\infty^* - \mathcal{V}_h^{\pi_h^*}\|_\infty = \max_{s \in \mathcal{S}} |\mathcal{V}_\infty^*(s) - \mathcal{V}_h^{\pi_h^*}(s)| \leq \frac{\gamma^h}{1-\gamma} R^*.$$

■

Further, we define for any $h \geq 1$ the truncated state-action value under policy $\mathbb{\pi}_{h-1} \in \Pi^{h-1}$ for every $s \in \mathcal{S}, a \in \mathcal{A}$ by

$$\begin{aligned}\mathcal{Q}_h^{\mathbb{\pi}_{h-1}}(s, a) &:= \mathop{\mathbb{E}}_{\substack{S_{t+1} \sim p(\cdot|S_t, A_t) \\ A_t \sim \pi_{h-t-1}(\cdot|S_t)}} \left[\sum_{t=0}^{h-1} \gamma^t r(S_t, A_t) \middle| S_0 = s, A_0 = a \right] \\ &= r(s, a) + \gamma \sum_{s' \in \mathcal{S}} p(s'|s, a) \mathcal{V}_{h-1}^{\mathbb{\pi}_{h-1}}(s').\end{aligned}$$

For $h \geq 1$ and $\mathbb{\pi}_h \in \Pi^h$ we also define the truncated advantage functions

$$\mathcal{A}_h^{\mathbb{\pi}_h}(s, a) := \mathcal{Q}_h^{\mathbb{\pi}_{h-1}}(s, a) - \mathcal{V}_h^{\mathbb{\pi}_h}(s), \forall s \in \mathcal{S}, a \in \mathcal{A}. \tag{5.2}$$

We derive the following performance difference lemma for the truncated discounted value functions.

LEMMA 5.4. *Let $\mathbb{m}_{h+1}, \mathbb{m}'_{h+1} \in \Pi^{h+1}$, then it holds that*

$$(i) \quad \mathcal{V}_{h+1}^{\mathbb{m}_{h+1}}(s) - \mathcal{V}_{h+1}^{\mathbb{m}'_{h+1}}(s) = \mathop{\mathbb{E}}_{\substack{S_0=s, A_t\sim\pi_{h-t}(\cdot|S_t) \\ S_{t+1}\sim p(\cdot|S_t,A_t)}} \Big[\sum_{t=0}^{h} \gamma^t \mathcal{A}_{h+1-t}^{\mathbb{m}'_{h+1-t}}(S_t, A_t) \Big].$$

(ii) If both policies only differ in the first policy, i.e. $\mathbb{m}_h = \mathbb{m}'_h$, then the above equation simplifies to

$$\mathcal{V}_{h+1}^{\mathbb{m}_{h+1}}(s) - \mathcal{V}_{h+1}^{\mathbb{m}'_{h+1}}(s) = \mathop{\mathbb{E}}_{S_0=s, A_0\sim\pi_h(\cdot|S_0)} \Big[\mathcal{A}_{h+1}^{\mathbb{m}'_{h+1}}(S_0, A_0) \Big]. \tag{5.3}$$

Proof. The proof is similar to Lemma 3.32. However, we need to adapt to the time shift (see Remark 5.2). First, let $\mathbb{m}_{h+1}, \mathbb{m}'_{h+1} \in \Pi^{h+1}$ be two arbitrary policies. We have

$$\begin{aligned}
&\mathcal{V}_{h+1}^{\mathbb{m}_{h+1}}(s) - \mathcal{V}_{h+1}^{\mathbb{m}'_{h+1}}(s) \\
&= \mathop{\mathbb{E}}_{\substack{S_0=s, A_t\sim\pi_{h-t}(\cdot|S_t) \\ S_{t+1}\sim p(\cdot|S_t,A_t)}} \Big[\sum_{t=0}^{h} \gamma^t r(S_t, A_t) \Big] - \mathcal{V}_{h+1}^{\mathbb{m}'_{h+1}}(s) \\
&= \mathop{\mathbb{E}}_{\substack{S_0=s, A_t\sim\pi_{h-t}(\cdot|S_t) \\ S_{t+1}\sim p(\cdot|S_t,A_t)}} \Big[\sum_{t=0}^{h} \gamma^t r(S_t, A_t) + \sum_{t=0}^{h} \gamma^t \mathcal{V}_{h+1-t}^{\mathbb{m}'_{h+1-t}}(S_t) - \sum_{t=0}^{h} \gamma^t \mathcal{V}_{h+1-t}^{\mathbb{m}'_{h+1-t}}(S_t) \Big] - \mathcal{V}_{h+1}^{\mathbb{m}'_{h+1}}(s) \\
&= \mathop{\mathbb{E}}_{\substack{S_0=s, A_t\sim\pi_{h-t}(\cdot|S_t) \\ S_{t+1}\sim p(\cdot|S_t,A_t)}} \Big[\sum_{t=0}^{h} \gamma^t r(S_t, A_t) + \sum_{t=1}^{h} \gamma^t \mathcal{V}_{h+1-t}^{\mathbb{m}'_{h+1-t}}(S_t) - \sum_{t=0}^{h} \gamma^t \mathcal{V}_{h+1-t}^{\mathbb{m}'_{h+1-t}}(S_t) \Big] \\
&\quad = \mathop{\mathbb{E}}_{\substack{S_0=s, A_t\sim\pi_{h-t}(\cdot|S_t) \\ S_{t+1}\sim p(\cdot|S_t,A_t)}} \Big[\sum_{t=0}^{h} \gamma^t r(S_t, A_t) + \sum_{t=0}^{h-1} \gamma^{t+1} \mathcal{V}_{h-t}^{\mathbb{m}'_{h-t}}(S_{t+1}) - \sum_{t=0}^{h} \gamma^t \mathcal{V}_{h+1-t}^{\mathbb{m}'_{h+1-t}}(S_t) \Big] \\
&\quad = \mathop{\mathbb{E}}_{\substack{S_0=s, A_t\sim\pi_{h-t}(\cdot|S_t) \\ S_{t+1}\sim p(\cdot|S_t,A_t)}} \Big[\sum_{t=0}^{h} \gamma^t \Big(r(S_t, A_t) + \gamma \mathcal{V}_{h-t}^{\mathbb{m}'_{h-t}}(S_{t+1}) - \mathcal{V}_{h+1-t}^{\mathbb{m}'_{h+1-t}}(S_t) \Big) \Big] \\
&\quad = \mathop{\mathbb{E}}_{\substack{S_0=s, A_t\sim\pi_{h-t}(\cdot|S_t) \\ S_{t+1}\sim p(\cdot|S_t,A_t)}} \Big[\sum_{t=0}^{h} \gamma^t \Big(\mathcal{Q}_{h+1-t}^{\mathbb{m}'_{h-t}}(S_t, A_t) - \mathcal{V}_{h+1-t}^{\mathbb{m}'_{h+1-t}}(S_t) \Big) \Big] \\
&\quad = \mathop{\mathbb{E}}_{\substack{S_0=s, A_t\sim\pi_{h-t}(\cdot|S_t) \\ S_{t+1}\sim p(\cdot|S_t,A_t)}} \Big[\sum_{t=0}^{h} \gamma^t \mathcal{A}_{h+1-t}^{\mathbb{m}'_{h+1-t}}(S_t, A_t) \Big],
\end{aligned}$$

where in the fifth equation we used the convention $\mathcal{V}_0 \equiv 0$, in the sixth equation the definition of the Q-function, and in the last equation the definition of the advantage function. Second, suppose that $\mathbb{m}_{h+1}$ and $\mathbb{m}'_{h+1}$ agree on all policies besides π_h, i.e. $\mathbb{m}_h = \mathbb{m}'_h$. Then, for any $t > 0$, it holds that

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\substack{S_0=s, A_t\sim\pi_{h-t}(\cdot|S_t) \\ S_{t+1}\sim p(\cdot|S_t,A_t)}} \Big[\mathcal{A}_{h+1-t}^{\mathbb{m}'_{h+1-t}}(S_t, A_t) \Big] \\
&= \sum_{a\in\mathcal{A}} \pi_h(a|s) \sum_{s'\in\mathcal{S}} p(s'|s, a) \mathop{\mathbb{E}}_{\substack{S_0=s', A_t\sim\pi_{h-t-1}(\cdot|S_t) \\ S_{t+1}\sim p(\cdot|S_t,A_t)}} \Big[\mathcal{A}_{h+1-t}^{\mathbb{m}'_{h+1-t}}(S_t, A_t) \Big]
\end{aligned}$$

$$= \sum_{a\in\mathcal{A}} \pi_h(a|s) \sum_{s'\in\mathcal{S}} p(s'|s,a) \left(\mathop{\mathbb{E}}_{\substack{S_0=s', A_t\sim\pi'_{h-t-1}(\cdot|S_t) \\ S_{t+1}\sim p(\cdot|S_t,A_t)}} \left[Q^{\mathbb{\pi}'_{h-t}}_{h+1-t}(S_t,A_t) - V^{\mathbb{\pi}'_{h+1-t}}_{h+1-t}(S_t) \right] \right)$$
$$= 0.$$

This proves the claim. ∎

Policy Search Framework. Inspired by dynamic programming, [Bag+03; Kak03] proposed policy search by dynamic programming (PSDP) to search policies $(\pi_{H-1}, \cdots, \pi_0) \in \widetilde{\Pi}^H$, where H is the problem horizon given as an input to the algorithm and $\widetilde{\Pi} \subset \Pi$ is the set of all deterministic policies. Formally, given H and $\widetilde{\Pi}$, PSDP computes

$$\widetilde{\pi}^*_h = \operatorname{argmax}_{\pi_h\in\widetilde{\Pi}} V^{(\pi_h,\widetilde{\mathbb{\pi}}^*_h)}_{h+1}(\mu) = \operatorname{argmax}_{\pi_h\in\widetilde{\Pi}} V^{(\pi_h,\widetilde{\pi}^*_{h-1},\cdots,\widetilde{\pi}^*_0)}_{h+1}(\mu), \tag{5.4}$$

for $h = 0, \ldots, H-1$. PSDP solves an $(h+1)$-step MDP initialized at an $S_0 \sim \mu$ in the iteration indexed by h. Specifically, it finds the optimal deterministic policy for selecting the first action A_0, denoted by $\widetilde{\pi}^*_h$, but then all the remaining actions in the episode $\{A_1, \cdots, A_{h-1}\}$ are selected according to the sequence of deterministic policies $\widetilde{\mathbb{\pi}}^*_h := (\widetilde{\pi}^*_{h-1}, \cdots, \widetilde{\pi}^*_0)$. Note that for all h, $\widetilde{\mathbb{\pi}}^*_h$ have been computed in previous iterations and are kept fixed. Compared to vanilla PG, PDSP exploits the Markovian property of the environment, rendering each iteration into solving a contextual bandit problem. While considering deterministic policies may suffice in tabular MDPs, no explicit computational procedures have been provided in [Bag+03; Kak03] to solve the optimization problem in equation (5.4). Further, determining the policy to be applied post-training is not immediately evident. In [Sch14] the author proposes to apply the non-stationary policy $\widetilde{\mathbb{\pi}}^*_H$ in a loop. Still, to solve the discussed infinite horizon MDP, a stationary policy is sufficient. We provide answers to these issues using DynPG.

5.2 The DynPG Algorithm

DynPG starts by solving a one-step contextual bandit problem and then incrementally extends the problem horizon by one in each iteration, which is done by appending the new decision epoch in front of the current problem horizon (see Algorithm 6 and the illustration in Figure 5.1).

Algorithm 6: DynPG for discounted MDPs

Result: Approximation of π^*, denoted as $\widehat{\pi}^*$.
Input: Initial state distribution μ and class of policies $(\pi^\theta)_{\theta\in\mathbb{R}^d}$.

Set $h = 0$ and initialize $\Lambda = []$;
while *Convergence criterion not met* **do**
  Initialize θ_0 (e.g., $\theta_0 \equiv 0$);
  Choose α_h and N_h (cf., Remark 5.11);
  for $n = 0, \dots, N_h - 1$ **do**
    Sample $G \approx \nabla_{\theta_n} \mathcal{V}_{h+1}^{(\pi^\theta,\Lambda)}(\mu)$;
    Update $\theta_{n+1} = \theta_n + \alpha_h G$;
  end
  Set $\widehat{\pi}_h^* = \pi^{\theta_{N_h}}$;
  Attach $\widehat{\pi}_h^*$ at the beginning of Λ;
  Set $h = h + 1$;
end
Return $\widehat{\pi}^* = \widehat{\pi}_{h-1}^*$ (the first element of Λ);

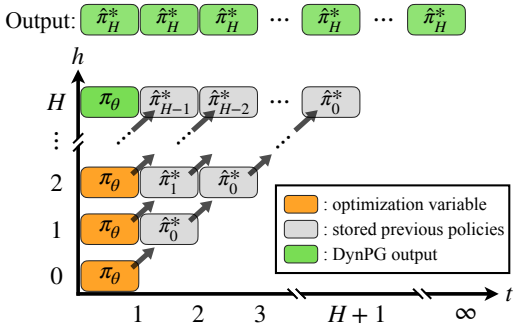


Figure 5.1: DynPG solves a sequence of contextual bandit problems, iteratively storing the convergent policies to memory and applying them accordingly as fixed policies in later iterations.

In each iteration, the parametrized (stochastic) policy responsible for sampling action A_0 from the newly-added epoch is optimized using gradient ascent. In subsequent time steps, DynPG applies the previous convergent policies to sample actions. Upon convergence of the current iteration, the policy is stored in a designated memory location, which will be utilized in future iterations. We define $\widehat{\mathbb{\pi}}_h^* := (\widehat{\pi}_{h-1}^*, \cdots, \widehat{\pi}_0^*) \in \Pi^h$ with convention $\widehat{\mathbb{\pi}}_0^* = \emptyset$ to denote the learned policies. DynPG returns the first policy in Λ when certain user-defined convergence criteria have been met. For example, this is the case if the relative value improvements between two consecutive DynPG iterations are small i.e., $\|\mathcal{V}_{h+1}^{\widehat{\mathbb{\pi}}_{h+1}^*} - \mathcal{V}_h^{\widehat{\mathbb{\pi}}_h^*}\|_\infty \leq \epsilon$.
Recall the definition of $\mathcal{V}_0 \equiv 0$, then the construction of DynPG implies

$$\mathcal{V}_{h+1}^{\widehat{\mathbb{\pi}}_{h+1}^*}(s) = T^{\widehat{\pi}_h^*}(\mathcal{V}_h^{\widehat{\mathbb{\pi}}_h^*})(s) = \cdots = (T^{\widehat{\pi}_h^*} \circ \cdots \circ T^{\widehat{\pi}_0^*})(\mathcal{V}_0^{\widehat{\mathbb{\pi}}_0^*})(s), \quad s \in \mathcal{S}, \tag{5.5}$$

which will be employed in the analyzes of the convergence rate.

Remark 5.5. Lets discuss the similarities and differences of DynPG to FT-DynPG, Algorithm 5, proposed in Chapter 4.

1. FT-DynPG has a prefixed time horizon H and trains policy backwards in time. In contrast, DynPG adds arbitrarily many policies in the beginning and can therefore be applied without prefixed H.

2. Given a fixed time horizon H, FT-DynPG returns a non-stationary policy $\widehat{\mathbb{\pi}}_H^*$ for an H-step MDP but with $\gamma = 1$, so without discounting. When DynPG is run for the same fixed number of iterations H as FT-DynPG and we include the discount factor in the rewards of FT-DynPG (cf. Remark 3.24 and Remark 4.6), then DynPG only differs by the time shift discussed in Remark 5.2.

In order to sample the gradient $G \approx \nabla_{\theta_n} \mathcal{V}_{h+1}^{(\pi^\theta,\Lambda)}(\mu)$ we use the following modified version of the policy gradient theorem.

THEOREM 5.6. *Suppose that $\mathbb{m}_h \in \Pi^h$ is fixed for some $h \geq 0$, with the convention $\mathbb{m}_0 = \emptyset$. Then for any differentiable parametrized family, say $(\pi^\theta)_{\theta \in \mathbb{R}^d}$, it holds that*

$$\nabla_\theta \mathcal{V}_{h+1}^{(\pi^\theta, \mathbb{m}_h)}(s) = \mathop{\mathbb{E}}_{S=s, A\sim\pi^\theta(\cdot|S)} \Big[\nabla_\theta \log(\pi^\theta(A|S)) \mathcal{Q}_{h+1}^{\mathbb{m}_h}(S, A) \Big].$$

Proof. The proof is the same as for dynamic policy gradient in finite-time horizon MDPs (Theorem 4.4). However, we will need an index shift and additionally to consider the discount factor γ and thus go throuth the arguments again.
An $(h+1)$-step trajectory $\tau_{h+1} = (s_0, a_0, \dots, s_h, a_h)$ under policy $(\pi^\theta, \mathbb{m}_h)$ and initial state distribution δ_s occurs with probability

$$p_s^{(\pi^\theta, \mathbb{m}_h)}(\tau_h) = \delta_s(s_0)\pi^\theta(a_0|s_0) \prod_{k=1}^{h} p(s_k|s_{k-1}, a_{k-1}) \pi_{h-k}(a_k|s_k).$$

Then, the log trick yields that

$$\begin{aligned}
&\nabla_\theta \log(p_s^{(\pi^\theta, \mathbb{m}_h)}(\tau_h)) \\
&= \nabla_\theta \Big(\log(\delta_s(s_0)) + \log(\pi^\theta(a_0|s_0)) + \sum_{k=1}^{h} \log(p(s_k|s_{k-1}, a_{k-1})) + \log(\pi_{h-k}(a_k|s_k)) \Big) \\
&= \nabla_\theta \log(\pi^\theta(a_0|s_0)).
\end{aligned}$$

Let $\mathcal{W}_{h+1}$ be the set of all trajectories from 0 to h. Then, the set $\mathcal{W}_{h+1}$ is finite due to the assumption that state and action space are finite. For $s \in \mathcal{S}$ we have

$$\begin{aligned}
&\nabla_\theta \mathcal{V}_{h+1}^{(\pi^\theta, \mathbb{m}_h)}(s) \\
&= \nabla_\theta \sum_{\tau_{h+1} \in \mathcal{W}_{h+1}} p_s^{(\pi^\theta, \mathbb{m}_h)}(\tau_{h+1}) \sum_{k=0}^{h} \gamma^k r(s_k, a_k) \\
&= \sum_{\tau_{h+1} \in \mathcal{W}_{h+1}} p_s^{(\pi^\theta, \mathbb{m}_h)}(\tau_h) \nabla_\theta \log(p_s^{(\pi^\theta, \mathbb{m}_h)}(\tau_{h+1})) \sum_{k=0}^{h} \gamma^k r(s_k, a_k) \\
&= \sum_{\tau_{h+1} \in \mathcal{W}_{h+1}} p_s^{(\pi^\theta, \mathbb{m}_h)}(\tau_{h+1}) \nabla_\theta \log(\pi^\theta(a_0|s_0)) \sum_{k=0}^{h} \gamma^k r(s_k, a_k) \\
&= \mathop{\mathbb{E}}_{\substack{S_0=s, A_0\sim\pi^\theta(\cdot|S) \\ S_{t+1}\sim p(\cdot|A_t, S_t), A_t\sim\pi_{h-t}(\cdot|S_t)}} \Big[\nabla_\theta \log(\pi^\theta(A_0|S_0)) \sum_{k=0}^{h} \gamma^k r(S_k, A_k) \Big] \\
&= \mathop{\mathbb{E}}_{S_0=s, A_0\sim\pi^\theta(\cdot|S)} \Big[\nabla_\theta \log(\pi^\theta(A_0|S_0)) \mathop{\mathbb{E}}_{S_{t+1}\sim p(\cdot|A_t, S_t), A_t\sim\pi_{h-t}(\cdot|S_t)} \Big[\sum_{k=0}^{h} \gamma^k r(S_k, A_k) \Big| S_0, A_0 \Big] \Big] \\
&= \mathop{\mathbb{E}}_{S=s, A\sim\pi^\theta(\cdot|S)} \Big[\nabla_\theta \log(\pi^\theta(A|S)) \mathcal{Q}_{h+1}^{\mathbb{m}_h}(S, A) \Big].
\end{aligned}$$

■

Compared to vanilla PG, which performs rollouts up to a geometric time horizon or until termination to ensure an unbiased gradient estimate (Section 3.1.2), DynPG dynamically adjusts the episode horizon during the execution of the algorithm. This results in three notable advantages. First, we observe that DynPG effectively reduces the variance in gradient estimation through the utilization of non-changing future policies. In contrast, the estimation of the gradient in vanilla PG involves assessing Q-values based on the stationary policy π^θ, which changes during training. Second, DynPG requires significantly fewer samples, a topic we discuss in detail at the end of this section. Moreover, each DynPG iteration has more benign optimization landscapes, as they are essentially contextual bandit problems. Specifically, analysing the convergence behaviour of DynPG in the next section reveals that DynPG under softmax parametrization has a smaller smoothness constant, enabling the selection of a more aggressive step size, α_h, to enhance convergence. Third, DynPG is less likely to suffer from committal behaviour, more exploration in the policy space is achieved by consistent re-initialization of the newly added policy.
In contrast to PSDP, we specify the set of policies $\widetilde{\Pi}$ to be a parameterized class of differentiable policies and a computational procedure, namely PG, for the inner loop. Hence, using deep networks with great approximation behavior works in DynPG and non-tabular MDPs can be considered in general. This preserves the model-free optimization characteristic of gradient methods, while still exploiting the underlying structure of MDPs by dynamic programming. DynPG can be further modified with additional enhancements such as regularization, natural policy gradient or policy mirror descent in ever optimization epoch. In addition, we show that applying the policy of the last training epoch as stationary policy is sufficient. This is a non-trivial result and our analysis in Section 5.3 reveals that in general it requires additional training to obtain a good stationary policy compared to using the non-stationary policy $\widehat{\mathbb{m}}_H^*$ in a loop as proposed in [Sch14]. In the tabular softmax case we specify these additional computational cost explicitly.

On the total sample complexity. For each gradient step in vanilla policy gradient, one must run the MDP until termination or up to a stochastic horizon $H \sim \text{Geom}(1-\gamma)$ to obtain an unbiased sample of the gradient (see Remark 3.17). DynPG, on the other hand, only requires h interactions with the environment to sample an unbiased estimator of the gradient in epoch h. Thus, comparing other policy gradient methods to DynPG solely based on the number of gradient steps is inadequate. Instead, for fairness, the number of samples (interactions with the environment) should be compared in practical implementations. For DynPG the total sample complexity is given by $\sum_{h=0}^{H-1}(h+1)N_h$. This results in a trade-off between increasing samples required for estimation and more accurate training to obtain convergence (cf. Section 5.3.1).

5.3 Convergence Analysis of DynPG

We analyze the convergence of DynPG under the tabular softmax parametrization. Specifically, Section 5.3.1 introduces four different layers of approximations that DynPG employed in solving the infinite-time horizon discounted MDP. These results are general and do not require a specific parametrization. Based on this we present the asymptotic global convergence of DynPG under the assumption of small enough optimization errors and rich enough parametrization class. Section 5.3.2 establishes the non-asymptotic global convergence rate of DynPG under the softmax parametrization and provides suitable parameters that achieve the theoretical limit.

5.3.1 *Error Decomposition and General Convergence*

To show the error decomposition we need the following Lemma to establish validation of applying the last policy trained in DynPG as stationary policy. The result is inspired by error bounds presented in [Ber01, Sec. 1.3].

LEMMA 5.7. *For any $V \in \mathbb{R}^{|\mathcal{S}|}$ and policy $\pi \in \Pi$ it holds that*

$$\|V - \mathcal{V}_\infty^\pi\|_\infty \leq \frac{\|T^\pi(V) - V\|_\infty}{1-\gamma}.$$

Proof. Consider any state $s \in \mathcal{S}$. Since $\mathcal{V}_\infty^\pi$ is the unique fixed point of the operator T^π, we have

$$\mathcal{V}_\infty^\pi(s) = T^\pi(\mathcal{V}_\infty^\pi)(s) = \sum_{a\in\mathcal{A}} \pi(a|s) r(s,a) + \gamma \sum_{a\in\mathcal{A}} \pi(a|s) \sum_{s'\in\mathcal{S}} p(s'|s,a)\mathcal{V}_\infty^\pi(s').$$

This implies that

$$\begin{aligned}
&\mathcal{V}_\infty^\pi(s) - V(s) \\
&= \sum_{a\in\mathcal{A}} \pi(a|s) r(s,a) + \gamma \sum_{a\in\mathcal{A}} \pi(a|s) \sum_{s'\in\mathcal{S}} p(s'|s,a)\mathcal{V}_\infty^\pi(s') - V(s) \\
&= \sum_{a\in\mathcal{A}} \pi(a|s) r(s,a) + \gamma \sum_{a\in\mathcal{A}} \pi(a|s) \sum_{s'\in\mathcal{S}} p(s'|s,a)(\mathcal{V}_\infty^\pi(s') - V(s') + V(s')) - V(s) \\
&= \sum_{a\in\mathcal{A}} \pi(a|s) r(s,a) + \gamma \sum_{a\in\mathcal{A}} \pi(a|s) \sum_{s'\in\mathcal{S}} p(s'|s,a)V(s') \\
&\quad + \gamma \sum_{a\in\mathcal{A}} \pi(a|s) \sum_{s'\in\mathcal{S}} p(s'|s,a)(\mathcal{V}_\infty^\pi(s') - V(s')) - V(s) \\
&= T^\pi(V)(s) - V(s) + \gamma \sum_{a\in\mathcal{A}} \pi(a|s) \sum_{s'\in\mathcal{S}} p(s'|s,a)(\mathcal{V}_\infty^\pi(s') - V(s')),
\end{aligned}$$

for any $s \in \mathcal{S}$. We define the mappings $s \mapsto g_\pi(s) = T^\pi(V)(s) - V(s)$ and $s \mapsto J_\pi(s) = \mathcal{V}_\infty^\pi(s) - V(s)$, $s \in \mathcal{S}$. Then, the above equation simplifies to

$$J_\pi(s) = g_\pi(s) + \gamma \sum_{a\in\mathcal{A}} \sum_{s'\in\mathcal{S}} \pi(a|s) p(s'|s,a) J_\pi(s').$$

By definition J_π satisfies

$$J_\pi(s) = \sum_{a\in\mathcal{A}} \pi(a|s) \left(g_\pi(s) + \gamma \sum_{s'\in\mathcal{S}} p(s'|s,a) J_\pi(s') \right)$$

for all $s \in \mathcal{S}$, and therefore is a solution of the Bellman equation with an auxiliary reward function $(s,a) \mapsto \widetilde{r}(s,a) = g_\pi(s)$. Note that this reward function is also bounded and by the uniqueness of the solution of the Bellman equation it has to hold that

$$J_\pi(s) = \mathop{\mathbb{E}}_{\substack{S_0=s, A_t\sim\pi(\cdot|S_t) \\ S_{t+1}\sim p(\cdot|S_t,A_t)}} [\sum_{k=0}^{\infty} \gamma^k g_\pi(S_k)] = g_\pi(s) + \sum_{k=1}^{\infty} \gamma^k \mathop{\mathbb{E}}_{\substack{S_0=s, A_t\sim\pi(\cdot|S_t) \\ S_{t+1}\sim p(\cdot|S_t,A_t)}} [g_\pi(S_k)].$$

Define $\underline{\beta} = \min_{s\in\mathcal{S}} g_\pi(s)$ and $\overline{\beta} = \max_{s\in\mathcal{S}} g_\pi(s)$. Then,

$$\frac{\underline{\beta}}{1-\gamma} \leq g_\pi(s) + \frac{\gamma\underline{\beta}}{1-\gamma} \leq J_\pi(s) \leq g_\pi(s) + \frac{\gamma\overline{\beta}}{1-\gamma} \leq \frac{\overline{\beta}}{1-\gamma}.$$

It follows that for any $s \in \mathcal{S}$,

$$|J_\pi(s)| \leq \frac{\max_{s\in\mathcal{S}} |g_\pi(s)|}{1-\gamma}.$$

By definition of g_π and J_π this yields the claim. ∎

DynPG employs four different layers of approximations, including

1. adopting parametrizations incapable of modeling the optimal policy perfectly (approximation error),
2. truncating the infinite problem horizon to a finite time window (truncation error),
3. utilizing a finite number of gradient updates to approximately solve each optimization problem (accumulated optimization error), and
4. applying the first policy of $\widehat{\mathbb{\pi}}_h$ as a stationary one in solving finite-horizon MDP, where non-stationary policies are required for optimality (stationary policy error).

We formally quantify these errors in the following lemma where the terms on the right-hand side of equation (5.6) correspond to the aforementioned errors, respectively.

Proposition 5.8. *The overall error of DynPG after H iterations can be decomposed as follows*

$$\begin{aligned}
\|\mathcal{V}_\infty^* - \mathcal{V}_\infty^{\widehat{\pi}_H^*}\|_\infty &\leq \left\|\mathcal{V}_\infty^* - \sup_\theta \mathcal{V}_\infty^{\pi_\theta}\right\|_\infty + \left\|\sup_\theta \mathcal{V}_\infty^{\pi_\theta} - \sup_{\theta_0,\dots,\theta_{H-1}} \mathcal{V}_H^{(\pi_{\theta_{H-1}},\dots\pi_{\theta_0})}\right\|_\infty \\
&\quad + \left\|\sup_{\theta_0,\dots,\theta_{H-1}} \mathcal{V}_H^{(\pi_{\theta_{H-1}},\dots\pi_{\theta_0})} - \mathcal{V}_H^{\widehat{\mathbb{\pi}}_H^*}\right\|_\infty + \|\mathcal{V}_H^{\widehat{\mathbb{\pi}}_H^*} - \mathcal{V}_\infty^{\widehat{\pi}_H^*}\|_\infty \\
&\leq \left\|\mathcal{V}_\infty^* - \sup_\theta \mathcal{V}_\infty^{\pi_\theta}\right\|_\infty + \frac{\gamma^H R^*}{1-\gamma} \\
&\quad + \sum_{h=0}^{H-1} \gamma^{H-h-1} \left\|\sup_\theta T^{\pi_\theta}(\mathcal{V}_h^{\widehat{\mathbb{\pi}}_h^*}) - \mathcal{V}_{h+1}^{\widehat{\mathbb{\pi}}_{h+1}^*}\right\|_\infty + \frac{1}{1-\gamma}\|\mathcal{V}_{H+1}^{\widehat{\mathbb{\pi}}_{H+1}^*} - \mathcal{V}_H^{\widehat{\mathbb{\pi}}_H^*}\|_\infty.
\end{aligned} \tag{5.6}$$

Proof. The first inequality follows directly from the triangle inequality of the supremum norm. Note here that we cannot simplify the first error further as it will depend on the chosen parametrization. Before we prove the second inequality, note that for $f, g : \mathbb{R}^d \to \mathbb{R}$

$$\left|\sup_\theta f(\theta) - \sup_\theta g(\theta)\right| \leq \sup_\theta |f(\theta) - g(\theta)|$$

for any two functions $f, g : \mathbb{R}^d \to \mathbb{R}$. If $\sup_\theta f(\theta) - \sup_\theta g(\theta) \geq 0$, then

$$\left|\sup_\theta f(\theta) - \sup_\theta g(\theta)\right| = \sup_\theta f(\theta) - \sup_\theta g(\theta) \leq \sup_\theta (f(\theta) - g(\theta)) \leq \sup_\theta |f(\theta) - g(\theta)|.$$

If $\sup_\theta f(\theta) - \sup_\theta g(\theta) < 0$, then the role of f and g are swapped. Thus, for $V, G \in \mathbb{R}^{|\mathcal{S}|}$ we deduce that

$$\|\sup_\theta T^{\pi^\theta}(V) - \sup_\theta T^{\pi^\theta}(G)\|_\infty \le \|\sup_\theta \left(T^{\pi^\theta}(V) - T^{\pi^\theta}(G)\right)\|_\infty \le \gamma \|V - G\|_\infty. \tag{5.7}$$

We treat the second, third and fourth terms separately.
For the second term: We prove that $\left\| \sup_\theta \mathcal{V}_\infty^{\pi^\theta} - \sup_{\theta_0,\dots,\theta_{H-1}} \mathcal{V}_H^{(\pi_{\theta_{H-1}},\dots\pi_{\theta_0})} \right\|_\infty \le \frac{\gamma^H R^*}{1-\gamma}$ similar to Lemma 5.3. Let $s \in \mathcal{S}$ be arbitrary but fixed. If $\sup_\theta \mathcal{V}_\infty^{\pi^\theta}(s) - \sup_{\theta_0,\dots,\theta_{H-1}} \mathcal{V}_H^{(\pi_{\theta_{H-1}},\dots\pi_{\theta_0})}(s) \ge 0$, then

$$\begin{aligned}
\sup_\theta \mathcal{V}_\infty^{\pi^\theta}(s) - \sup_{\theta_0,\dots,\theta_{H-1}} \mathcal{V}_H^{(\pi_{\theta_{H-1}},\dots\pi_{\theta_0})}(s) &\le \sup_\theta \mathcal{V}_\infty^{\pi^\theta}(s) - \sup_\theta \mathcal{V}_H^{\pi^\theta}(s) \\
&\le \sup_\theta (\mathcal{V}_\infty^{\pi^\theta}(s) - \mathcal{V}_H^{\pi^\theta}(s)) \\
&= \sup_\theta \mathop{\mathbb{E}}_{\substack{S_0=s, A_t\sim\pi^\theta(\cdot|S_t) \\ S_{t+1}\sim p(\cdot|S_t,A_t)}} \left[\sum_{t=H}^\infty \gamma^t r(S_t, A_t)\right] \\
&\le \sum_{t=H}^\infty \gamma^t R^* = \frac{\gamma^H}{1-\gamma} R^*,
\end{aligned}$$

where we used equation (5.7) in the second inequality.
On the other hand, if $\sup_\theta \mathcal{V}_\infty^{\pi^\theta}(s) - \sup_{\theta_0,\dots,\theta_{H-1}} \mathcal{V}_H^{(\pi_{\theta_{H-1}},\dots\pi_{\theta_0})}(s) < 0$, then

$$\begin{aligned}
&\sup_{\theta_0,\dots,\theta_{H-1}} \mathcal{V}_H^{(\pi_{\theta_{H-1}},\dots\pi_{\theta_0})}(s) - \sup_\theta \mathcal{V}_\infty^{\pi^\theta}(s) \\
&\le \sup_{\theta_0,\dots,\theta_{H-1}} \mathcal{V}_H^{(\pi_{\theta_{H-1}},\dots\pi_{\theta_0})}(s) - \sup_{\theta_0,\dots,\theta_{H-1}} \mathcal{V}_\infty^{((\pi_{\theta_{H-1}},\dots\pi_{\theta_0}))_\infty}(s) \\
&\le \sup_{\theta_0,\dots,\theta_{H-1}} \left(\mathcal{V}_H^{(\pi_{\theta_{H-1}},\dots\pi_{\theta_0})}(s) - \mathcal{V}_\infty^{((\pi_{\theta_{H-1}},\dots\pi_{\theta_0}))_\infty}(s)\right) \\
&= \sup_{\theta_0,\dots,\theta_{H-1}} - \mathop{\mathbb{E}}_{\substack{S_0=s, A_t\sim\pi_{\theta_{H-(t \bmod H)-1}}(\cdot|S_t) \\ S_{t+1}\sim p(\cdot|S_t,A_t)}} \left[\sum_{t=H}^\infty \gamma^t r(S_t, A_t)\right] \\
&\le \sum_{t=H}^\infty \gamma^t R^* = \frac{\gamma^H}{1-\gamma} R^*.
\end{aligned}$$

Collecting these together, we obtain that $\left\| \sup_\theta \mathcal{V}_\infty^{\pi^\theta} - \sup_{\theta_0,\dots,\theta_{H-1}} \mathcal{V}_H^{(\pi_{\theta_{H-1}},\dots\pi_{\theta_0})} \right\|_\infty \le \frac{\gamma^H R^*}{1-\gamma}$, since $s \in \mathcal{S}$ was chosen arbitrary.
For the third term: We show that

$$\| \sup_{\theta_0,\dots,\theta_{H-1}} \mathcal{V}_H^{(\pi_{\theta_{H-1}},\dots\pi_{\theta_0})} - \mathcal{V}_H^{\widehat{\mathbb{m}}_H^*} \|_\infty \le \sum_{h=0}^{H-1} \gamma^{H-h-1} \|\sup_\theta T^{\pi^\theta}(\mathcal{V}_h^{\widehat{\mathbb{m}}_h^*}) - T^{\widehat{\pi}_h^*}(\mathcal{V}_h^{\widehat{\mathbb{m}}_h^*})\|_\infty \tag{5.8}$$

holds for all $H \ge 1$ by induction. For $H = 1$ we have by equation (5.5) that

$$\|\sup_\theta \mathcal{V}_1^{\pi^\theta} - \mathcal{V}_1^{\widehat{\mathbb{m}}_1^*}\|_\infty = \left\| \sup_\theta T^{\pi^\theta}(\mathcal{V}_0^{\widehat{\mathbb{m}}_0^*}) - T^{\widehat{\pi}_0^*}(\mathcal{V}_0^{\widehat{\mathbb{m}}_0^*}) \right\|_\infty,$$

with the convention $\mathcal{V}_0 \equiv 0$ and $\widehat{\mathbb{m}}_0^* = \emptyset$. So assume that equation (5.8) holds for some $H \geq 1$; then for $H+1$ we have

$$\begin{aligned}
&\Big\| \sup_{\theta_0,\dots,\theta_H} \mathcal{V}_{H+1}^{\{\pi_{\theta_H},\dots\pi_{\theta_0}\}} - \mathcal{V}_{H+1}^{\widehat{\mathbb{m}}_{H+1}^*} \Big\|_\infty \\
&= \Big\| \sup_{\theta_H} T^{\pi_{\theta_H}} \Big(\sup_{\theta_0,\dots,\theta_{H-1}} \mathcal{V}_H^{(\pi_{\theta_{H-1}},\dots\pi_{\theta_0})} \Big) - T^{\widehat{\pi}_H^*}(\mathcal{V}_H^{\widehat{\mathbb{m}}_H^*}) \Big\|_\infty \\
&\leq \Big\| \sup_{\theta_H} T^{\pi_{\theta_H}} \Big(\sup_{\theta_0,\dots,\theta_{H-1}} \mathcal{V}_H^{(\pi_{\theta_{H-1}},\dots\pi_{\theta_0})} \Big) - \sup_{\theta_H} T^{\pi_{\theta_H}}(\mathcal{V}_H^{\widehat{\mathbb{m}}_H^*}) \Big\|_\infty \\
&\quad + \Big\| \sup_{\theta_H} T^{\pi_{\theta_H}}(\mathcal{V}_h^{\widehat{\mathbb{m}}_H^*}) - T^{\widehat{\pi}_H^*}(\mathcal{V}_H^{\widehat{\mathbb{m}}_H^*}) \Big\|_\infty \\
&\leq \gamma \Big\| \sup_{\theta_0,\dots,\theta_{H-1}} \mathcal{V}_H^{(\pi_{\theta_{H-1}},\dots\pi_{\theta_0})} - \mathcal{V}_H^{\widehat{\mathbb{m}}_H^*} \Big\|_\infty + \Big\| \sup_{\theta_H} T^{\pi_{\theta_H}}(\mathcal{V}_H^{\widehat{\mathbb{m}}_H^*}) - T^{\widehat{\pi}_H^*}(\mathcal{V}_H^{\widehat{\mathbb{m}}_H^*}) \Big\|_\infty \\
&\leq \gamma \sum_{h=0}^{H-1} \gamma^{H-h-1} \| \sup_\theta T^{\pi^\theta}(\mathcal{V}_h^{\widehat{\mathbb{m}}_h^*}) - T^{\widehat{\pi}_h^*}(\mathcal{V}_h^{\widehat{\mathbb{m}}_h^*}) \|_\infty + \Big\| \sup_\theta T^{\pi^\theta}(\mathcal{V}_H^{\widehat{\mathbb{m}}_H^*}) - T^{\widehat{\pi}_H^*}(\mathcal{V}_H^{\widehat{\mathbb{m}}_H^*}) \Big\|_\infty \\
&= \sum_{h=0}^{H} \gamma^{H-h} \| \sup_\theta T^{\pi^\theta}(\mathcal{V}_h^{\widehat{\mathbb{m}}_h^*}) - T^{\widehat{\pi}_h^*}(\mathcal{V}_h^{\widehat{\mathbb{m}}_h^*}) \|_\infty,
\end{aligned}$$

where we used equation (5.5), as well as equation (5.7), and the induction assumption. This yields the desired claim in equation (5.8) for all $H \geq 1$.
For the fourth error term: We have to deal with the error of applying the final policy as stationary policy. We use Lemma 5.7 to arrive at

$$\begin{aligned}
\| \mathcal{V}_H^{\widehat{\mathbb{m}}_H^*} - \mathcal{V}_\infty^{\widehat{\pi}_H^*} \|_\infty &\leq \frac{1}{1-\gamma} \| T^{\widehat{\pi}_H^*}(\mathcal{V}_H^{\widehat{\mathbb{m}}_H^*}) - \mathcal{V}_H^{\widehat{\mathbb{m}}_H^*} \|_\infty \\
&= \frac{1}{1-\gamma} \| \mathcal{V}_{H+1}^{\widehat{\mathbb{m}}_{H+1}^*} - \mathcal{V}_H^{\widehat{\mathbb{m}}_H^*} \|_\infty,
\end{aligned}$$

where we used again equation (5.5) in the last line. ∎

The error decomposition gives clear insights into algorithm design, for which it is necessary to discuss the practical implications of the four summands.

1. To control the approximation error a rich enough policy parametrization is required.
2. To keep the truncation error small, DynPG should run more than $H \approx \frac{\log(1-\gamma)}{\log(\gamma)}$ rounds.
3. The third summand shows that approximation errors in earlier iterations are discounted more than approximation errors in later iterations. To achieve optimal training efficiency, we will thus require a geometrically decreasing optimization error across the iterations of DynPG. Recall the sample complexity discussion at the end of Section 5.2 to note that a trade-off between more accurate training and increasing samples required for estimating the gradient must be made to obtain the best performance of DynPG.
4. DynPG approximates the value function by truncation at a fixed time H and then replaces the optimal time-dependent policy by a stationary policy. As mentioned for PSDP, one could also apply the non-stationary policy $\widehat{\mathbb{m}}_h^*$ to approximate the value function. This would cause

the fourth error term to vanish. Thus, in what follows, we distinguish between the **overall error** in equation (5.6) and the **value function error**:

$$\|\mathcal{V}_\infty^* - \mathcal{V}_H^{\widehat{\mathbb{m}}_H^*}\|_\infty \leq \big\|\mathcal{V}_\infty^* - \sup_{\theta\in\mathbb{R}^d} \mathcal{V}_\infty^{\pi^\theta}\big\|_\infty + \frac{\gamma^H R^*}{1-\gamma} + \sum_{h=0}^{H-1} \gamma^{H-h-1} \Big\| \sup_{\theta\in\mathbb{R}^d} T^{\pi^\theta}(\mathcal{V}_h^{\widehat{\mathbb{m}}_h^*}) - \mathcal{V}_{h+1}^{\widehat{\mathbb{m}}_{h+1}^*}\Big\|_\infty. \quad (5.9)$$

It is important to note that the factor $(1-\gamma)^{-1}$ in the stationary policy error causes an additional dependence on the effective horizon in the complexity bounds for softmax DynPG.

General convergence result. For the remainder of this section, we will proceed under the assumption of zero approximation error stemming from the parametrization. This condition generally holds true when the class of policies (π^θ) can effectively approximate all deterministic policies, such as achieved through tabular softmax. Under this circumstance, it follows that $T^* = \sup_\theta T^{\pi^\theta}$ and $\|\mathcal{V}_\infty^* - \sup_\theta \mathcal{V}_\infty^{\pi^\theta}\|_\infty = 0$. Subsequently, we demonstrate that a certain reduction in the optimization error is adequate for achieving global convergence within the parametrized policy space.

ASSUMPTION 5.9. *The class (π^θ) has zero approximation error and there exists a positive sequence ϵ_h such that*

- $\sum_{t=0}^{H-1} \gamma^{H-h-1}\epsilon_h \to 0$ *for* $H \to \infty$,
- *the policies obtained by DynPG satisfy* $\|T^*(\mathcal{V}_h^{\widehat{\mathbb{m}}_h^*}) - \mathcal{V}_{h+1}^{\widehat{\mathbb{m}}_{h+1}^*}\|_\infty \leq \epsilon_h$ *for all* $h \geq 0$.

As an example for such a sequence one might think of $\epsilon_h = c\gamma^h$ for some $c > 0$. The second condition in the assumption holds true, when each contextual bandit problem can be sufficiently well addressed. We deduce convergence directly from the error decomposition in Proposition 5.8.

COROLLARY 5.10. *Let Assumption 5.9 hold. Then Algorithm 6 generates a sequence of non-stationary policies $\widehat{\mathbb{m}}_H^* \in \Pi^H$ that satisfy*

$$\|\mathcal{V}_\infty^* - \mathcal{V}_H^{\widehat{\mathbb{m}}_H^*}\|_\infty \to 0 \quad \text{for } H \to \infty.$$

Furthermore, the overall error vanishes in the limit:

$$\|\mathcal{V}_\infty^* - \mathcal{V}_\infty^{\widehat{\pi}_H^*}\|_\infty \to 0 \quad \text{for } H \to \infty.$$

Proof. Adjusting equation (5.9) to zero approximation error and exploiting Assumption 5.9 we obtain

$$\|\mathcal{V}_\infty^* - \mathcal{V}_H^{\widehat{\mathbb{m}}_H^*}\|_\infty \leq \frac{\gamma^H R^*}{1-\gamma} + \sum_{h=0}^{H-1} \gamma^{H-h-1}\epsilon_h \to 0 \quad \text{for } H \to \infty.$$

For the second part of the theorem note that $\mathcal{V}_H^{\mathbb{m}_H^*}$ converges in the supremum norm to $\mathcal{V}_\infty^*$ by the first part. This implies that $\mathcal{V}_H^{\mathbb{m}_H^*}$ is a Cauchy sequence with respect to the supremum norm, i.e. $\|\mathcal{V}_{H+1}^{\mathbb{m}_{H+1}^*} - \mathcal{V}_H^{\mathbb{m}_H^*}\|_\infty \to 0$ for $H \to \infty$. The claim follows directly from Proposition 5.8. ∎

Remark 5.11. The condition $\sum_{h=0}^{H-1} \gamma^{H-h-1}\epsilon_h \to 0$ for $H \to \infty$ implies that the optimization error must decrease with increasing h to guarantee convergence. Hence, training must become more precise over time. It is therefore advisable to increase the number of gradient steps N_h and decrease the step size α_h when h increases. In FT-DynPG, the error accumulated linearly over time and an accuracy of $\frac{\epsilon}{H}$ in every optimization epoch lead to an overall error of ϵ. In DynPG the error is discounted over time, by adding new epochs in the beginning.

5.3.2 *Convergence Rates under Tabular Softmax Parametrization*

One might inquire whether Assumption 5.9 is reasonable and whether there are situations in which the condition on the algorithm holds. Indeed, we will show that this is the case for the softmax class of policies. More precisely, for any $\epsilon_h > 0$, we can specify a step size α_h and a number of gradient steps N_h such that Assumption 5.9 is satisfied. In a second step, we optimize H and the error sequence (ϵ_h) to obtain the optimal sample complexity for DynPG under softmax parametrization, where we distinguish again between the value function error and the overall error.

Assumption 5.12. *Suppose that $\mu(s) > 0$ for all $s \in \mathcal{S}$. Furthermore, the parametrization in DynPG $(\pi^\theta)_{\theta\in\mathbb{R}^{|\mathcal{S}||\mathcal{A}|}}$ is chosen to be the tabular softmax parametrization introduced in equation* (3.12). *We assume further that θ_0 in Algorithm 6 is initialized such that the softmax policy forms a uniform distribution over the action space and the gradient can be accessed exactly.*

The proof will be similar to the ones presented in Section 4.2 and relies on the smoothness and gradient domination property. Before we come to the results, recall the differences of DynPG and FT-DynPG discussed in Remark 5.5. We have to adapt the proofs in Section 4.2.2 to an additional discount factor. Note, that we cannot apply [Mei+20, Thm. 2] for bandits, as we consider contextual bandits. Further, we cannot apply the proof of [Mei+20, Thm. 4] with $\gamma = 0$ as they just consider positive rewards in $[0, 1]$ which is inconvenient for our contextual bandit setting where the maximal rewards grows when we add a new time-epoch in the beginning. Therefore, we have to adapt the proofs in [Mei+20] and Section 4.2.2 to our setting.

Smoothness. First, we derive the smoothness of our objective functions.

Lemma 5.13. *Let $(\pi^\theta)_{\theta\in\mathbb{R}^{|\mathcal{S}||\mathcal{A}|}}$ be the softmax parametrization. Then, for arbitrary $\mathbb{m}_h \in \Pi^h$ and $\mu \in \Delta(\mathcal{S})$, the function $\theta \mapsto \mathcal{V}_{h+1}^{(\pi^\theta,\mathbb{m}_h)}(\mu)$ is L_h-smooth with $L_h = \frac{2R^*(1-\gamma^{h+1})}{(1-\gamma)}$.*

Proof. The proof is similar to the one in Lemma 4.18. Note that we can interpret $\mathcal{V}_{h+1}^{(\pi^\theta,\mathbb{m}_h)}(\mu)$ as a contextual bandit problem, i.e. a discounted (infinite-time) MDP with discount factor $\gamma = 0$. The reward of the contextual bandit problem is almost surely bounded in $[-\frac{1-\gamma^{h+1}}{1-\gamma}R^*, \frac{1-\gamma^{h+1}}{1-\gamma}R^*]$, because

$$\sum_{t=0}^{h} \gamma^t R^* = \frac{1-\gamma^{h+1}}{1-\gamma} R^*.$$

We can apply [YGL22, Lem. 4.4 and Lem. 4.8] with $R_{\max} = \frac{1-\gamma^{h+1}}{1-\gamma}R^*$, $G^2 = 1 - \frac{1}{|\mathcal{A}|} \leq 1$ and $F = 1$ to obtain the smoothness constant $L_h = \frac{2R^*(1-\gamma^{h+1})}{(1-\gamma)}$. ∎

Weak gradient domination. Second, we obtain the following non-uniform gradient domination property.

LEMMA 5.14. *Under Assumption 5.12 it holds for any* $\mathbb{m}_h \in \Pi^h$ *that*

$$\|\nabla_\theta \mathcal{V}_{h+1}^{(\pi^\theta,\mathbb{m}_h)}(\mu)\|_\infty \leq \min_{s\in\mathcal{S}} \pi^\theta(a^*(s)|s)\big(T^*(\mathcal{V}_h^{\mathbb{m}_h})(\mu) - \mathcal{V}_{h+1}^{(\pi^\theta,\mathbb{m}_h)}(\mu)\big),$$

where $a^*(s)$ *denotes the (unique) action taken after the greedy policy* $\pi^{\mathcal{V}_h^{\mathbb{m}_h}}$.

Remark 5.15. As for finite-time MDPs we assume again, without loss of generality, that the action $a^*(s)$ is unique for any fixed future policy $\mathbb{m}_h$.

Proof. First note from Theorem 5.6 that under the tabular softmax parametrization we have

$$\nabla_\theta \mathcal{V}_{h+1}^{(\pi^\theta,\mathbb{m}_h)}(\mu) = \sum_{s\in\mathcal{S}} \mu(s) \mathop{\mathbb{E}}_{S=s,A\sim\pi^\theta(\cdot|S)} \big[\nabla_\theta \log(\pi^\theta(A|S)) \mathcal{Q}_{h+1}^{\mathbb{m}_h}(S,A)\big].$$

Hence, with the derivative of the softmax function, equation (3.13)

$$\frac{\partial \log(\pi^\theta(a|s))}{\partial\theta(s',a')} = \mathbf{1}_{\{s=s'\}}\Big(\mathbf{1}_{\{a=a'\}} - \pi^\theta(a'|s)\Big),$$

it holds that

$$\begin{aligned}
\frac{\partial \mathcal{V}_{h+1}^{(\pi^\theta,\mathbb{m}_h)}(s)}{\partial\theta(s',a')} &= \mathbf{1}_{\{s=s'\}} \mathop{\mathbb{E}}_{S=s',A\sim\pi^\theta(\cdot|S)} \Big[\Big(\mathbf{1}_{\{A=a'\}} - \pi^\theta(a'|s')\Big)\mathcal{Q}_{h+1}^{\mathbb{m}_h}(S,A)\Big] \\
&= \mathbf{1}_{\{s=s'\}}\Big(\pi^\theta(a'|s')\mathcal{Q}_{h+1}^{\mathbb{m}_h}(s',a') - \pi^\theta(a'|s') \mathop{\mathbb{E}}_{S=s',A\sim\pi^\theta(\cdot|S)} \big[\mathcal{Q}_{h+1}^{\mathbb{m}_h}(S,A)\big]\Big) \qquad (5.10) \\
&= \mathbf{1}_{\{s=s'\}}\pi^\theta(a'|s')\Big(\mathcal{Q}_{h+1}^{\mathbb{m}_h}(s',a') - \mathcal{V}_{h+1}^{(\pi^\theta,\mathbb{m}_h)}(s')\Big) \\
&= \mathbf{1}_{\{s=s'\}}\pi^\theta(a'|s')\mathcal{A}_{h+1}^{(\pi^\theta,\mathbb{m}_h)}(s',a').
\end{aligned}$$

We deduce from Lemma 5.4, equation (5.3), that

$$T^*(\mathcal{V}_h^{\mathbb{m}_h})(s) - T^{\pi_h}(\mathcal{V}_h^{\mathbb{m}_h})(s) = \mathop{\mathbb{E}}_{S_0=s,A_0\sim\pi_h^*(\cdot|S_0)} \Big[\mathcal{A}_{h+1}^{\mathbb{m}_{h+1}}(S_0,A_0)\Big].$$

Finally we can derive that

$$\begin{aligned}
\|\nabla_\theta \mathcal{V}_{h+1}^{(\pi^\theta,\mathbb{m}_h)}(\mu)\|_2 &= \Big\| \sum_{s\in\mathcal{S}} \mu(s) \frac{\partial \mathcal{V}_{h+1}^{(\pi^\theta,\mathbb{m}_h)}(s)}{\partial\theta} \Big\|_2 \\
&= \Big[\sum_{s'\in\mathcal{S}} \sum_{a'\in\mathcal{A}} \Big(\sum_{s\in\mathcal{S}} \mu(s) \frac{\partial \mathcal{V}_{h+1}^{(\pi^\theta,\mathbb{m}_h)}(s)}{\partial\theta(s',a')} \Big)^2 \Big]^2 \\
&= \Big[\sum_{a'\in\mathcal{A}} \Big(\sum_{s\in\mathcal{S}} \mu(s)\pi^\theta(a'|s)\mathcal{A}_{h+1}^{(\pi^\theta,\mathbb{m}_h)}(s,a') \Big)^2 \Big]^2
\end{aligned}$$

$$
\begin{aligned}
&\geq \Big| \sum_{s\in\mathcal{S}} \mu(s)\pi^{\theta}(a^*(s)|s)\mathcal{A}_{h+1}^{(\pi^{\theta},\mathbb{m}_h)}(s,a^*(s))\Big| \\
&= \Big| \sum_{s\in\mathcal{S}} \mu(s)\pi^{\theta}(a^*(s)|s) \mathop{\mathbb{E}}_{S=s,A\sim\pi^{\mathcal{V}_h^{\mathbb{m}_h}}(\cdot|s)} \Big[\mathcal{A}_{h+1}^{(\pi^{\theta},\mathbb{m}_h)}(S,A)\Big]\Big| \\
&= \sum_{s\in\mathcal{S}} \mu(s)\pi^{\theta}(a^*(s)|s)\Big(T^*(\mathcal{V}_h^{\mathbb{m}_h})(s) - T^{\pi^{\theta}}(\mathcal{V}_h^{\mathbb{m}_h})(s)\Big) \\
&\geq \min_{s\in\mathcal{S}} \pi^{\theta}(a^*(s)|s)\Big(T^*(\mathcal{V}_h^{\mathbb{m}_h})(\mu) - T^{\pi^{\theta}}(\mathcal{V}_h^{\mathbb{m}_h})(\mu)\Big),
\end{aligned}
$$

where $a^*(s)$ denotes the action taken after the greedy policy $\pi^{\mathcal{V}_h^{\mathbb{m}_h}}$ (cf. Remark 5.15). ∎

The next step is to show that the term $\min_{s\in\mathcal{S}} \pi^{\theta}(a^*(s)|s)$ can be bounded (uniformly in $s\in\mathcal{S}$) from below by $\frac{1}{|\mathcal{A}|}$ along the gradient ascent trajectory, when softmax is initialized uniformly.

LEMMA 5.16. *Let Assumption 5.12 hold and denote by $(\theta_n)_{n\geq 0}$ the gradient ascent sequence in epoch $h\geq 0$ of DynPG under the fixed future policy $\widehat{\mathbb{m}}_h^*\in\Pi^h$. Suppose further that the step size $\alpha_h = \frac{1-\gamma}{2R^*(1-\gamma^{h+1})}$, then for any $s\in\mathcal{S}$ it holds that*

$$\min_{n\geq 0} \pi^{\theta_n}(a^*(s)|s) = \frac{1}{|\mathcal{A}|}.$$

where $a^*(s)$ denotes the (unique) action taken after the greedy policy $\pi^{\mathcal{V}_h^{\widehat{\mathbb{m}}_h^*}}$.

Proof. The proof is adapted from [Mei+20, Lem. 5] and Lemma 4.20.
First, we define the sets

$$
\begin{aligned}
\mathcal{R}_1(s) &= \{\theta : \pi^{\theta}(a^*(s)|s) \geq \pi^{\theta}(a|s)\forall a\neq a^*(s)\} \\
\mathcal{R}_2(s) &= \Big\{\theta : \frac{\partial V_{h+1}^{(\pi^{\theta},\widehat{\mathbb{m}}_h^*)}(s)}{\partial\theta(s,a^*(s)} \geq \frac{\partial V_{h+1}^{(\pi^{\theta},\widehat{\mathbb{m}}_h^*)}(s)}{\partial\theta(s,a^*(s)}\forall a\neq a^*(s)\Big\}.
\end{aligned}
$$

Claim 1: It holds that $\theta_n\in\mathcal{R}_2 \implies \theta_{n+1}\in\mathcal{R}_2$.
To prove this claim, we first deduce from equation (5.10) that

$$
\begin{aligned}
&\frac{\partial \mathcal{V}_{h+1}^{(\pi^{\theta},\widehat{\mathbb{m}}_h^*)}(s)}{\partial\theta(s,a^*(s))} \geq \frac{\partial \mathcal{V}_{h+1}^{(\pi^{\theta},\widehat{\mathbb{m}}_h^*)}(s)}{\partial\theta(s,a^*(s))} \qquad (5.11)\\
\iff\ &\pi_{\theta_n}(a^*(s)|s)\mathcal{A}_{h+1}^{(\pi^{\theta},\widehat{\mathbb{m}}_h^*)}(s,a^*(s)) \geq \pi_{\theta_n}(a|s)\mathcal{A}_{h+1}^{(\pi^{\theta},\widehat{\mathbb{m}}_h^*)}(s,a).
\end{aligned}
$$

Let $a\neq a^*(s)$ be arbitrary. We consider two cases to proof claim 1:

1. Suppose that $\pi_{\theta_n}(a^*(s)|s)\geq\pi_{\theta_n}(a|s)$ for any $a\neq a^*(s)$. Then, is holds that $\theta_n(s,a^*(s))\geq\theta_n(s,a)$ by the definition of softmax. Next, as $\theta_n\in\mathcal{R}_2^*$, we derive

$$\theta_{n+1}(s,a^*(s)) = \theta_n(s,a^*(s)) + \eta_h\mu_h(s)\frac{\partial \mathcal{V}_{h+1}^{(\pi_{\theta_n},\widehat{\mathbb{m}}_h^*)}(s)}{\partial\theta_n(s,a^*(s))}$$

$$\geq \theta_n(s,a) + \eta_h \mu_h(s) \frac{\partial \mathcal{V}_{h+1}^{(\pi^\theta, \widehat{\mathbb{m}}_h^*)}(s)}{\partial \theta(s, a^*(s))}$$
$$= \theta_n(s,a).$$

Thus, by the definition of the softmax function, $\pi_{\theta_{n+1}}(a^*(s)|s) \geq \pi_{\theta_{n+1}}(a|s)$ for any $a \neq a^*(s)$. Further, because $a^*(s)$ is the greedy action, we arrive at

$$\pi_{\theta_{n+1}}(a^*(s)|s) \mathcal{A}_{h+1}^{(\pi_{\theta_{n+1}}, \widehat{\mathbb{m}}_h^*)}(s, a^*(s)) \geq \pi_{\theta_n}(a|s) \mathcal{A}_{h+1}^{(\pi_{\theta_{n+1}}, \widehat{\mathbb{m}}_h^*)}(s,a),$$

such that $\theta_{n+1} \in \mathcal{R}_2(s)$ by equation (5.11).

2. Suppose that $\pi_{\theta_n}(a^*(s)|s) < \pi_{\theta_n}(a|s)$ for any $a \neq a^*(s)$. Then, $\theta_n(s, a^*(s)) - \theta_n(s,a) < 0$ by definition of the softmax function. As $\theta_n \in \mathcal{R}_2(s)$, it holds that

$$\begin{aligned}
&\theta_{n+1}(s, a^*(s)) - \theta_{n+1}(s,a) \\
&\geq \theta_n(s, a^*(s)) + \eta_h \mu(s) \frac{\partial \mathcal{V}_{h+1}^{(\pi^\theta, \widehat{\mathbb{m}}_h^*)}(s)}{\partial \theta(s, a^*(s))} - \theta_n(s,a) - \eta_h \mu(s) \frac{\partial \mathcal{V}_{h+1}^{(\pi^\theta, \widehat{\mathbb{m}}_h^*)}(s)}{\partial \theta(s,a)} \\
&\geq \theta_n(s, a^*(s)) - \theta_n(s,a).
\end{aligned}$$

Thus, we obtain

$$(1 - \exp(\theta_{n+1}(s, a^*(s)) - \theta_{n+1}(s,a))) \leq (1 - \exp(\theta_n(s, a^*(s)) - \theta_n(s,a))) < 1.$$

By the ascent lemma for smooth functions [Mei+20, Lem. 18] it follows monotonicity in the objective function (due to small enough step size $\eta_h = \frac{1}{\beta_h}$ with β_h the smoothness constant in Lemma 5.13) such that $\mathcal{V}_{h+1}^{(\pi_{\theta_{n+1}}, \widehat{\mathbb{m}}_h^*)}(s) \geq \mathcal{V}_{h+1}^{(\pi_{\theta_n}, \widehat{\mathbb{m}}_h^*)}(s)$. So,

$$\begin{aligned}
&(1 - \exp(\theta_{n+1}(s, a^*(s)) - \theta_{n+1}(s,a))) \Big(\mathcal{Q}_{h+1}^{\widehat{\mathbb{m}}_h^*}(s, a^*(s)) - \mathcal{V}_{h+1}^{(\pi_{\theta_{n+1}}, \widehat{\mathbb{m}}_h^*)}(s) \Big) \\
&\leq (1 - \exp(\theta_n(s, a^*(s)) - \theta_n(s,a))) \Big(\mathcal{Q}_{h+1}^{\widehat{\mathbb{m}}_h^*}(s, a^*(s)) - \mathcal{V}_{h+1}^{(\pi_{\theta_n}, \widehat{\mathbb{m}}_h^*)}(s) \Big)
\end{aligned}$$

We rearrange equation (5.11) and obtain that $\theta \in \mathcal{R}_2(s)$ is equivalent to

$$\mathcal{Q}_{h+1}^{\widehat{\mathbb{m}}_h^*}(s, a^*(s)) - \mathcal{Q}_{h+1}^{\widehat{\mathbb{m}}_h^*}(s,a) \geq (1 - \exp(\theta_n(s, a^*(s)) - \theta_n(s,a))) \mathcal{A}_{h+1}^{\{\pi_\theta, \widehat{\mathbb{m}}_h^*\}}(s, a^*(s)).$$

We deduce by $\theta_n \in \mathcal{R}_2(s)$ that

$$\begin{aligned}
&(1 - \exp(\theta_{n+1}(s, a^*(s)) - \theta_{n+1}(s,a))) \Big(\mathcal{Q}_{h+1}^{\widehat{\mathbb{m}}_h^*}(s, a^*(s)) - \mathcal{V}_{h+1}^{(\pi_{\theta_{n+1}}, \widehat{\mathbb{m}}_h^*)}(s) \Big) \\
&\leq (1 - \exp(\theta_n(s, a^*(s)) - \theta_n(s,a))) \Big(\mathcal{Q}_{h+1}^{\widehat{\mathbb{m}}_h^*}(s, a^*(s)) - \mathcal{V}_{h+1}^{(\pi_{\theta_n}, \widehat{\mathbb{m}}_h^*)}(s) \Big) \\
&\leq \mathcal{Q}_{h+1}^{\widehat{\mathbb{m}}_h^*}(s, a^*(s)) - \mathcal{Q}_{h+1}^{\widehat{\mathbb{m}}_h^*}(s,a),
\end{aligned}$$

and thus $\theta_{n+1} \in \mathcal{R}_2(s)$.

This proves claim 1.
Claim 2: If $\theta_n \in \mathcal{R}_2(s)$, then it holds that $\pi_{\theta_{n+1}}(a^*(s)|s) \geq \pi_{\theta_n}(a^*(s)|s)$.
To see this we compute

$$\begin{aligned}
\pi_{\theta_{n+1}}(s, a^*(s)) &= \frac{\exp(\theta_{n+1}(s, a^*(s))}{\sum_{a' \in \mathcal{A}} \exp(\theta_{n+1}(s, a')} \\
&= \frac{\exp(\theta_n(s, a^*(s)) \exp\left(\eta_h \mu(s) \frac{\partial \mathcal{V}_{h+1}^{(\pi_{\theta_n}, \widehat{\mathbb{m}}_h^*)}(s)}{\partial \theta_n(s, a^*(s))}\right)}{\sum_{a' \in \mathcal{A}} \exp(\theta_n(s, a') \exp\left(\eta_h \mu(s) \frac{\partial \mathcal{V}_{h+1}^{(\pi_{\theta_n}, \widehat{\mathbb{m}}_h^*)}(s)}{\partial \theta_n(s, a')}\right)} \\
&\geq \frac{\exp(\theta_n(s, a^*(s)) \exp\left(\eta_h \mu(s) \frac{\partial \mathcal{V}_{h+1}^{(\pi_{\theta_n}, \widehat{\mathbb{m}}_h^*)}(s)}{\partial \theta_n(s, a^*(s))}\right)}{\sum_{a' \in \mathcal{A}} \exp(\theta_n(s, a') \exp\left(\eta_h \mu(s) \frac{\partial \mathcal{V}_{h+1}^{(\pi_{\theta_n}, \widehat{\mathbb{m}}_h^*)}(s)}{\partial \theta_n(s, a^*(s))}\right)} \\
&= \pi_{\theta_n}(s, a^*(s)).
\end{aligned}$$

Claim 3: It holds that $\theta_n \in \mathcal{R}_1(s) \implies \theta_n \in \mathcal{R}_2(s)$.
Let $\theta_n \in \mathcal{R}_1(s)$. As $a^*(s)$ is optimal we have for any $a \neq a^*(s)$ that

$$\mathcal{A}_{h+1}^{(\pi_{\theta_n}, \widehat{\mathbb{m}}_h^*)}(s, a^*(s)) \geq \mathcal{A}_{h+1}^{(\pi_{\theta_n}, \widehat{\mathbb{m}}_h^*)}(s, a).$$

Further by $\theta_n \in \mathcal{R}_1(s)$ it holds

$$\pi_{\theta_n}(a^*(s)|s) \mathcal{A}_{h+1}^{(\pi_{\theta_n}, \widehat{\mathbb{m}}_h^*)}(s, a^*(s)) \geq \pi_{\theta_n}(a|s) \mathcal{A}_{h+1}^{(\pi_{\theta_n}, \widehat{\mathbb{m}}_h^*)}(s, a).$$

Hence, by equation (5.10), we deduce that $\theta_n \in \mathcal{R}_2(s)$.
Conclusion of the proof by combining claim 1 to 3:
Since θ_0 is initialized such that softmax is the uniform distribution, we have that $\theta_0 \in \mathcal{R}_1(s)$ for all $s \in \mathcal{S}$. By claim 3, we have that $\theta_0 \in \mathcal{R}_2(s)$ and by claim 1 it follows that $\theta_n \in \mathcal{R}_2(s)$ for all $n \geq 0$ and all $s \in \mathcal{S}$. Finally, by claim 2, it follows that $\min_{n \geq 0} \pi_{\theta_n}(s, a^*(s)) = \pi_{\theta_0}(s, a^*(s)) = \frac{1}{|\mathcal{A}|}$ for any $s \in \mathcal{S}$. ■

Remark 5.17. When the softmax policy is not uniformly initialized, we suffer, as in vanilla PG, from the existence of an unknown constant c_γ. The constant can depend on γ as well as on other MDP parameters like the size of the state and action space.

Global convergence. Finally we combine all results and derive the global convergence of DynPG in every optimization step under exact gradients and softmax parametrization.
By Corollary 5.10, we obtain from Theorem 5.18 convergence for softmax DynPG by choosing a sufficient decreasing error sequence (ϵ_h).

THEOREM 5.18. *Let Assumption 5.12 hold. Let $h \geq 0$, $\epsilon_h > 0$ and $\Lambda = \widehat{\mathbb{m}}_h^* \in \Pi^h$ be a collection of h arbitrary policies. Then, using step size $\alpha_h = \frac{1-\gamma}{2R^*(1-\gamma^{h+1})}$ and gradient steps $N_h = \left\lceil \frac{4R^*(1-\gamma^{h+1})|\mathcal{A}|^2}{(1-\gamma)\epsilon_h} \left\| \frac{1}{\mu} \right\|_\infty \right\rceil$ in DynPG (Algorithm 6) guarantees that the policy $\widehat{\pi}_h^*$ of iteration h achieves*

$$\|T^*(\mathcal{V}_h^{\widehat{\mathbb{m}}_h^*}) - \mathcal{V}_{h+1}^{\widehat{\mathbb{m}}_{h+1}^*}\|_\infty \leq \epsilon_h.$$

Proof. First, recall that the tabular softmax parametrization can approximate any deterministic policy arbitrarily well. As optimal policies in finite horizon MDPs are deterministic, we have that $\sup_{\theta\in\mathbb{R}^{|\mathcal{S}||\mathcal{A}|}} \mathcal{V}_{h+1}^{\{\pi_\theta,\widehat{\mathbb{m}}_h^*\}}(s) = T^*(\mathcal{V}_h^{\widehat{\mathbb{m}}_h^*})(s)$ for all $s \in \mathcal{S}$ (zero approximation error induced by the parametrization). Moreover, for any $s \in \mathcal{S}$

$$\begin{aligned} T^*(\mathcal{V}_h^{\widehat{\mathbb{m}}_h^*})(s) - T^{\widehat{\pi}_h^*}(\mathcal{V}_h^{\widehat{\mathbb{m}}_h^*})(s) &= \sum_{s'\in\mathcal{S}} \mu(s') \frac{\mathbf{1}_{s'=s}}{\mu(s')} \Big(T^*(\mathcal{V}_h^{\widehat{\mathbb{m}}_h^*})(s') - T^{\widehat{\pi}_h^*}(\mathcal{V}_h^{\widehat{\mathbb{m}}_h^*})(s') \Big) \\ &\leq \left\| \frac{1}{\mu} \right\|_\infty \Big(T^*(\mathcal{V}_h^{\widehat{\mathbb{m}}_h^*})(\mu) - T^{\widehat{\pi}_h^*}(\mathcal{V}_h^{\widehat{\mathbb{m}}_h^*})(\mu) \Big). \end{aligned}$$

By Lemma 5.13 the function $\theta \mapsto \mathcal{V}_{h+1}^{\{\pi_\theta,\widehat{\mathbb{m}}_h^*\}}(\mu)$ is L_h-smooth and fulfills the weak gradient domination property along the gradient ascent steps with constant $\frac{1}{|\mathcal{A}|}$ (Lemma 5.14 and Lemma 5.16). Moreover, for any $\theta \in \mathbb{R}^d$ it holds that

$$T^*(\mathcal{V}_h^{\widehat{\mathbb{m}}_h^*})(s) - T^{\pi_\theta}(\mathcal{V}_h^{\widehat{\mathbb{m}}_h^*})(s) \leq 2\frac{1-\gamma^{h+1}}{1-\gamma} R^* = L_h \leq \frac{2|\mathcal{A}|^2}{\alpha}.$$

Hence, we can apply Theorem 2.11 with $b = \frac{1}{|\mathcal{A}|}$, $\alpha = \alpha_h = \frac{1-\gamma}{2R^*(1-\gamma^{h+1})}$ and $k = N_h$, so

$$T^*(\mathcal{V}_h^{\widehat{\mathbb{m}}_h^*})(\mu) - T^{\pi_{\theta_{N_h}}}(\mathcal{V}_h^{\widehat{\mathbb{m}}_h^*})(\mu) \leq \frac{4|\mathcal{A}|^2 R^*(1-\gamma^{h+1})}{(1-\gamma)N_h}.$$

By the choice of N_h we deduce for any $s \in \mathcal{S}$

$$T^*(\mathcal{V}_h^{\widehat{\mathbb{m}}_h^*})(s) - T^{\widehat{\pi}_h^*}(\mathcal{V}_h^{\widehat{\mathbb{m}}_h^*})(s) \leq \left\| \frac{1}{\mu} \right\|_\infty \Big(T^*(\mathcal{V}_h^{\widehat{\mathbb{m}}_h^*})(\mu) - T^{\pi_{\theta_{N_h}}}(\mathcal{V}_h^{\widehat{\mathbb{m}}_h^*})(\mu) \Big) \leq \epsilon_h,$$

which proves the claim. ∎

5.3.3 *Complexity bounds for Softmax DynPG*

In order to obtain complexity bounds from Theorem 5.18 for a given accuracy ϵ we optimize the total number of gradient steps $\sum_{h=0}^H N_h(\epsilon_h)$ with respect to H and (ϵ_h) under the constraint that the overall error in equation (5.6) or the value function error in equation (5.9) is bounded by ϵ. We summarize the complexity bounds for both error types in the following and then deal with both case separately to provide detailed proof with explicit selections for H, $(N_h)_{h=0}^H$, and $(\alpha_h)_{h=0}^H$..

THEOREM 5.19. *[cf. Theorem 5.22 and Theorem 5.24 for detailed versions] Let Assumption 5.12 hold and choose $\epsilon > 0$.*

1. *Overall error: We can specify H, $(N_h)_{h=0}^H$ and $(\alpha_h)_{h=0}^H$ such that,*

$$\sum_{h=0}^H N_h = \left\lceil \frac{24R^*|\mathcal{A}|^2}{(1-\gamma)^2\epsilon} \left\| \frac{1}{\mu} \right\|_\infty \right\rceil \left\lceil \frac{\log(6R^*(1-\gamma)^{-2}\epsilon^{-1})}{\log(\gamma^{-1})} \right\rceil^2 + \left\lceil \frac{24R^*(1-\gamma^{H+1})|\mathcal{A}|^2}{(1-\gamma)^3\epsilon} \left\| \frac{1}{\mu} \right\|_\infty \right\rceil$$

accumulated gradient steps are required to achieve $\|\mathcal{V}_\infty^* - \mathcal{V}_\infty^{\widehat{\pi}_H^*}\|_\infty \leq \epsilon$.

2. *Value function error: We can specify* H, $(N_h)_{h=0}^{H-1}$ *and* $(\alpha_h)_{h=0}^{H-1}$ *such that,*

$$\sum_{h=0}^{H-1} N_h = \left\lceil \frac{8R^*|\mathcal{A}|^2}{(1-\gamma)\epsilon} \left\| \frac{1}{\mu} \right\|_\infty \right\rceil \left\lceil \frac{\log(2R^*(1-\gamma)^{-1}\epsilon^{-1})}{\log(\gamma^{-1})} \right\rceil^2$$

accumulated gradient steps are required to achieve $\|\mathcal{V}_\infty^* - \mathcal{V}_H^{\widehat{\mathbb{m}}_H^*}\|_\infty \leq \epsilon$.

To obtain the convergence behavior of DynPG in terms of γ close to 1, we derive in the following asymptotic equivalency.

LEMMA 5.20. *The term* $\log\left(\gamma^{-1}\right) = -\log(\gamma)$ *is asymptotically equivalent to* $(1-\gamma)$ *for* $\gamma \uparrow 1$.

Proof. By the definition of asymptotic equivalence for two functions f and g we have to show that

$$\lim_{x \uparrow 1} \frac{f(x)}{g(x)} = 1.$$

It holds by L'Hospital rule that

$$\lim_{\gamma \uparrow 1} \frac{-\log(\gamma)}{(1-\gamma)} = \lim_{\gamma \uparrow 1} \frac{-\gamma^{-1}}{-1} = 1.$$

■

Combined with Theorem 5.19, we find that the required gradient steps for γ close to 1 behave like

$$\begin{cases} O\big((1-\gamma)^{-4}\epsilon^{-1}\log((1-\gamma)^{-2}\epsilon^{-1})\big), \text{ for the overall error,} \\ O\big((1-\gamma)^{-3}\epsilon^{-1}\log((1-\gamma)^{-1}\epsilon^{-1})\big), \text{ for the value function error.} \end{cases} \tag{5.12}$$

Note for the overall error that the first summand is the dominating term in terms of $\gamma \to 1$. Compared to softmax PG, we observe an additional $\log(\epsilon^{-1})$ factor in the convergence and future research could explore the possibility of eliminating the log-factor. In terms of γ, however, DynPG offers a resilient upper bound, which, in comparison to vanilla PG, remains at most polynomial in the effective horizon.
We prove the two claims in Theorem 5.19 separately.

Value function error. Note that under the tabular softmax parametrization we have zero approximation error and the upper bound in equation (5.9) simplifies to

$$\|\mathcal{V}_\infty^* - \mathcal{V}_H^{\widehat{\mathbb{m}}_H^*}\|_\infty \leq \frac{\gamma^H R^*}{1-\gamma} + \sum_{h=0}^{H-1} \gamma^{H-h-1}\epsilon_h,$$

where ϵ_h are upper bounds on one step optimization errors $\left\| \sup_{\theta \in \mathbb{R}^d} T^{\pi_\theta}(\mathcal{V}_h^{\widehat{\mathbb{m}}_h^*}) - \mathcal{V}_{h+1}^{\widehat{\mathbb{m}}_{h+1}^*} \right\|_\infty$. In order to minimize the accumulated number of gradient steps $\sum_{h=0}^{H-1} N_h$, recall that the gradient

steps in epoch h are given by $N_h = \left\lceil \frac{4R^*(1-\gamma^{h+1})|\mathcal{A}|^2}{(1-\gamma)\epsilon_h} \left\| \frac{1}{\mu} \right\|_\infty \right\rceil$ in Theorem 5.18. To optimize, we can drop the gaussian brackets and minimize

$$\frac{4R^*|\mathcal{A}|^2}{(1-\gamma)} \left\| \frac{1}{\mu} \right\|_\infty \sum_{h=0}^{H-1} \frac{(1-\gamma^{h+1})}{\epsilon_h}$$

with respect to H and $(\epsilon_h)_{h=0}^{H-1}$. The resulting optimization problem has the form:

$$\begin{aligned} &\min_{H, (\epsilon_h)_{h=0}^{H-1}, \epsilon_h > 0} \sum_{h=0}^{H-1} \frac{(1-\gamma^{h+1})}{\epsilon_h} \\ &\text{subject to} \quad \frac{\gamma^H R^*}{1-\gamma} + \sum_{h=0}^{H-1} \gamma^{H-h-1} \epsilon_h \leq \epsilon. \end{aligned} \tag{5.13}$$

In [WWZ22, Sec. 3.2] a similar optimization problem is considered and it is shown that asymptotically (as $\epsilon \to 0$) it suffices to bound both error terms in the constraint by $\frac{\epsilon}{2}$. The first condition, i.e. $\frac{\gamma^H R^*}{1-\gamma} \leq \frac{\epsilon}{2}$, leads to the criterion $H \geq \log_\gamma \left(\frac{(1-\gamma)\epsilon}{2R^*} \right)$. To minimize the number of gradient steps we fix $H = \left\lceil \log_\gamma \left(\frac{(1-\gamma)\epsilon}{2R^*} \right) \right\rceil$. It remains to solve the following optimization problem

$$\min_{(\epsilon_h)_{h=0}^{H-1}, \epsilon_h > 0} \sum_{h=0}^{H-1} a_h \epsilon_h^{-1} \quad \text{subject to} \quad \sum_{h=0}^{H-1} \gamma^{H-h-1} \epsilon_h \leq \frac{\epsilon}{2}, \tag{5.14}$$

with $a_h = (1-\gamma^{h+1})$.
We provide a solution of this optimization problem by solving the following more general one:

LEMMA 5.21. *Fix $H > 0$. Let $(a_h)_{h=0}^{H-1}$ and $(b_h)_{h=0}^{H-1}$ be strictly positive sequences. For any $d > 0$ the optimization problem*

$$\min_{(c_h)_{h=0}^{H-1}} \sum_{h=0}^{H-1} a_h c_h^{-1} \quad \textit{subject to} \quad \sum_{h=0}^{H-1} b_h c_h \leq d,$$

is optimally solved for $c_h = C_{d,H} \left(\frac{b_h}{a_h} \right)^{-\frac{1}{2}}$, with $C_{H,d} = d \left(\sum_{h=0}^{H-1} (a_h b_h)^{\frac{1}{2}} \right)^{-1}$.

Hence, the minimum of the optimization problem is given by $\frac{1}{d} \left(\sum_{h=0}^{H-1} (a_h b_h)^{\frac{1}{2}} \right)^2$.

Proof. The result and the proof is inspired by [WWZ22, Lem. 3.8].
We solve the constrained optimization problem by employing the Lagrange method, i.e

$$\min_{\lambda, (c_h)_{h=0}^{H-1}} \sum_{h=0}^{H-1} a_h c_h^{-1} + \lambda \left(\sum_{h=0}^{H-1} b_h c_h - d \right).$$

The first order conditions are given by

$$-a_h c_h^{-2} + \lambda b_h = 0 \quad \forall h = 0, \dots, H-1, \quad \text{and} \quad \sum_{h=0}^{H-1} b_h c_h - d = 0.$$

We deduce from the first equations, that there exists a constant $C_{H,d}$ such that $c_h = C_{H,d}\left(\frac{b_h}{a_h}\right)^{-\frac{1}{2}}$ for all $h = 0, \dots, H-1$. Using this in the second equation we can solve for the constant

$$C_{H,d} = d\Big(\sum_{h=0}^{H-1} b_h^{\frac{1}{2}} a_h^{\frac{1}{2}}\Big)^{-1}.$$

Using the minima $c_h = C_{H,d}\left(\frac{b_h}{a_h}\right)^{-\frac{1}{2}}$ in the optimization function $\sum_{h=0}^{H-1} a_h c_h^{-1}$, we obtain the minimum

$$\sum_{h=0}^{H-1} a_h c_h^{-1} = \sum_{h=0}^{H-1} a_h C_{H,d}^{-1}\left(\frac{b_h}{a_h}\right)^{\frac{1}{2}} = C_{H,\epsilon}^{-1} \sum_{h=0}^{H-1} (a_h b_h)^{\frac{1}{2}} = \frac{1}{d}\left(\sum_{h=0}^{H-1} (a_h b_h)^{\frac{1}{2}}\right)^2.$$

■

Finally, we are ready to state the detailed version of Theorem 5.19 (2.) for the value function error.

THEOREM 5.22. *[Detailed version of Theorem 5.19 (2.)]*
Let Assumption 5.12 hold and choose $\epsilon > 0$. Set

$$\begin{aligned}
H &= \Big\lceil \log_\gamma\Big(\frac{(1-\gamma)\epsilon}{2R^*}\Big)\Big\rceil,\\
\epsilon_h &= \frac{\epsilon}{2}\Big(\sum_{t=0}^{H-1}((1-\gamma^{t+1})\gamma^{H-t-1})^{\frac{1}{2}}\Big)^{-1}\Big(\frac{\gamma^{H-h-1}}{(1-\gamma^{h+1})}\Big)^{-\frac{1}{2}},\\
\alpha_h &= \frac{1-\gamma}{2R^*(1-\gamma^{h+1})},\\
N_h &= \Big\lceil \frac{4R^*(1-\gamma^{h+1})|\mathcal{A}|^2}{(1-\gamma)\epsilon_h}\Big\|\frac{1}{\mu}\Big\|_\infty\Big\rceil,
\end{aligned}$$

for $h = 0, \dots H-1$. Then, the non-stationary policy $\widehat{\mathbb{m}}_H^*$ obtained by DynPG under exact gradients achieves $\|\mathcal{V}_\infty^* - \mathcal{V}_H^{\widehat{\mathbb{m}}_H^*}\|_\infty \le \epsilon$. The total number of gradient steps are given by

$$\sum_{h=0}^{H-1} N_h = \Big\lceil \frac{8R^*|\mathcal{A}|^2}{(1-\gamma)\epsilon}\Big\|\frac{1}{\mu}\Big\|_\infty\Big\rceil\Big\lceil \frac{\log(2R^*(1-\gamma)^{-1}\epsilon^{-1})}{\log(\gamma^{-1})}\Big\rceil^2.$$

Proof. First, note that the choice of ϵ_h in the theorem is the solution of the optimization problem

$$\min_{(\epsilon_h)} \frac{4R^*|\mathcal{A}|^2}{(1-\gamma)}\Big\|\frac{1}{\mu}\Big\|_\infty \sum_{h=0}^{H-1} \frac{(1-\gamma^{h+1})}{\epsilon_h} \text{ subject to } \sum_{h=0}^{H-1} \gamma^{H-h-1}\epsilon_h \le \frac{\epsilon}{2}.$$

To see this, choose $a_h = (1-\gamma^{h+1})$, $b_h = \gamma^{H-h-1}$, $c_h = \epsilon_h$ and $d = \frac{\epsilon}{2}$ in Lemma 5.21, where we excluded the constant $\frac{4R^*|\mathcal{A}|^2}{(1-\gamma)}\Big\|\frac{1}{\mu}\Big\|_\infty$. Then,

$$\epsilon_h = C_{H,d}\left(\frac{b_h}{a_h}\right)^{-\frac{1}{2}} = \frac{\epsilon}{2}\Big(\sum_{t=0}^{H-1}((1-\gamma^{t+1})\gamma^{H-t-1})^{\frac{1}{2}}\Big)^{-1}\Big(\frac{\gamma^{H-h-1}}{(1-\gamma^{h+1})}\Big)^{-\frac{1}{2}},$$

is the optimal solution to this problem. For $H = \lceil \frac{\log(\frac{(1-\gamma)\epsilon}{2R^*})}{\log(\gamma)} \rceil$ we have that $\frac{\gamma^H R^*}{1-\gamma} \leq \frac{\epsilon}{2}$. Using these $(\epsilon_h)_{h=0}^{H-1}$ in Theorem 5.18 results in a value function error for DynPG bounded by

$$\|\mathcal{V}_\infty^* - \mathcal{V}_H^{\widehat{\mathbb{M}}_H^*}\|_\infty \leq \frac{\gamma^H R^*}{1-\gamma} + \sum_{h=0}^{H-1} \gamma^{H-h-1}\epsilon_h \leq \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon. \tag{5.15}$$

For the optimal complexity bound we derive

$$\begin{aligned}
\sum_{h=0}^{H-1} N_h &= \sum_{h=0}^{H-1} \Big\lceil \frac{4R^*(1-\gamma^{h+1})|\mathcal{A}|^2}{(1-\gamma)\epsilon_h} \Big\| \frac{1}{\mu} \Big\|_\infty \Big\rceil \\
&= \sum_{h=0}^{H-1} \Big\lceil \frac{8R^*|\mathcal{A}|^2 \big(\sum_{t=0}^{H-1} ((1-\gamma^{t+1})\gamma^{H-t-1})^{\frac{1}{2}} \big)}{(1-\gamma)\epsilon} \Big\| \frac{1}{\mu} \Big\|_\infty ((1-\gamma^{h+1})\gamma^{H-h-1})^{\frac{1}{2}} \Big\rceil \\
&\leq \Big\lceil \frac{8R^*|\mathcal{A}|^2}{(1-\gamma)\epsilon} \Big\| \frac{1}{\mu} \Big\|_\infty \Big\rceil \Big(\sum_{h=0}^{H-1} \lceil ((1-\gamma^{t+1})\gamma^{H-t-1})^{\frac{1}{2}} \rceil \Big)^2.
\end{aligned}$$

As $((1-\gamma^{t+1})\gamma^{H-t-1})^{\frac{1}{2}} \leq 1$ for any $\gamma \in (0,1)$ we have

$$\sum_{h=0}^{H-1} N_h \leq \Big\lceil \frac{8R^*|\mathcal{A}|^2}{(1-\gamma)\epsilon} \Big\| \frac{1}{\mu} \Big\|_\infty \Big\rceil H^2.$$

Finally, note that we can rewrite H to be

$$H = \Big\lceil \log_\gamma \Big(\frac{(1-\gamma)\epsilon}{2R^*} \Big) \Big\rceil = \Big\lceil \frac{\log\left(\frac{(1-\gamma)\epsilon}{2R^*} \right)}{\log(\gamma)} \Big\rceil = \Big\lceil \frac{\log((1-\gamma)^{-1}\epsilon^{-1}2R^*)}{\log(\gamma^{-1})} \Big\rceil.$$

■

Overall error. To deal with the overall error, we first derive the following upper bound.

LEMMA 5.23. *Assume zero approximation error, i.e. $T^* = \sup_\theta T^{\pi_\theta}$. Further, suppose bounded optimization errors $\|T^*(\mathcal{V}_h^{\widehat{\mathbb{M}}_h^*}) - T^{\widehat{\pi}_h^*}(\mathcal{V}_h^{\widehat{\mathbb{M}}_h^*})\|_\infty \leq \epsilon_h$ for all $h = 0, \ldots, H$. If $\epsilon_H \leq (1-\gamma)\sum_{h=0}^{H-1} \gamma^{H-h-1}\epsilon_h$, then the overall error of DynPG is bounded by*

$$\|\mathcal{V}_\infty^* - \mathcal{V}_\infty^{\widehat{\pi}_H^*}\|_\infty \leq \frac{3}{1-\gamma}\Big(\frac{\gamma^H R^*}{1-\gamma} + \sum_{h=0}^{H-1} \gamma^{H-h-1}\epsilon_h \Big).$$

Proof. First we have by triangle inequality, that

$$\|\mathcal{V}_{H+1}^{\widehat{\mathbb{M}}_{H+1}^*} - \mathcal{V}_H^{\widehat{\mathbb{M}}_H^*}\|_\infty \leq \|\mathcal{V}_{H+1}^{\widehat{\mathbb{M}}_{H+1}^*} - \mathcal{V}_\infty^*\|_\infty + \|\mathcal{V}_\infty^* - \mathcal{V}_H^{\widehat{\mathbb{M}}_H^*}\|_\infty.$$

By Proposition 5.8 we obtain under zero approximation error that

$$\|\mathcal{V}_{H+1}^{\widehat{\mathbb{M}}_{H+1}^*} - \mathcal{V}_\infty^*\|_\infty + \|\mathcal{V}_\infty^* - \mathcal{V}_H^{\widehat{\mathbb{M}}_H^*}\|_\infty \leq \frac{\gamma^{H+1}R^*}{1-\gamma} + \sum_{h=0}^{H} \gamma^{H-h}\epsilon_h + \frac{\gamma^H R^*}{1-\gamma} + \sum_{h=0}^{H-1} \gamma^{H-h-1}\epsilon_h.$$

By the assumption $\epsilon_H \leq (1-\gamma)\sum_{h=0}^{H-1}\gamma^{H-h-1}\epsilon_h$ it holds further that

$$\sum_{h=0}^{H}\gamma^{H-h}\epsilon_h \leq \gamma\sum_{h=0}^{H-1}\gamma^{H-h-1}\epsilon_h + \epsilon_H \leq \sum_{h=0}^{H-1}\gamma^{H-h-1}\epsilon_h.$$

We obtain

$$\|\mathcal{V}_{H+1}^{\widehat{\mathbb{m}}_{H+1}^*} - \mathcal{V}_H^{\widehat{\mathbb{m}}_H^*}\|_\infty \leq \frac{2\gamma^H R^*}{1-\gamma} + 2\sum_{h=0}^{H-1}\gamma^{H-h-1}\epsilon_h.$$

We deduce for the overall error (by Proposition 5.8 and under zero approximation error) that

$$\begin{aligned}\|\mathcal{V}_\infty^* - \mathcal{V}_\infty^{\widehat{\pi}_H^*}\|_\infty &\leq \frac{\gamma^H R^*}{1-\gamma} + \sum_{h=0}^{H-1}\gamma^{H-h-1}\epsilon_h + \frac{1}{1-\gamma}\left(\frac{2\gamma^H R^*}{1-\gamma} + 2\sum_{h=0}^{H-1}\gamma^{H-h-1}\epsilon_h\right)\\ &\leq \frac{3}{1-\gamma}\Big(\frac{\gamma^H R^*}{1-\gamma} + \sum_{h=0}^{H-1}\gamma^{H-h-1}\epsilon_h\Big).\end{aligned}$$

■

It is important to notice that the overall error is upper bounded by the same error terms, $\frac{\gamma^H R^*}{1-\gamma} + \sum_{h=0}^{H-1}\gamma^{H-h-1}\epsilon_h$, as the value function error (under zero approximation error) up to the constant $\frac{3}{1-\gamma}$. Thus, we can obtain the result for the overall error by substituting ϵ with $\frac{(1-\gamma)\epsilon}{3}$ in the result for the value function error:

THEOREM 5.24. *[Detailed version of Theorem 5.19 (1.)]*
Let Assumption 5.12 hold and choose $\epsilon > 0$. Set

$$\begin{aligned}H &= \Big\lceil \log_\gamma\big(\frac{(1-\gamma)^2\epsilon}{6R^*}\big)\Big\rceil,\\ \epsilon_h &= \frac{\epsilon(1-\gamma)}{6}\Big(\sum_{h=0}^{H-1}(((1-\gamma^{h+1})\gamma^{H-h-1})^{\frac{1}{2}}\Big)^{-1}\Big(\frac{\gamma^{H-h-1}}{(1-\gamma^{h+1})}\Big)^{-\frac{1}{2}}, \quad \forall h = 0\ldots, H-1,\\ \epsilon_H &= \frac{(1-\gamma)^2\epsilon}{6},\\ \alpha_h &= \frac{1-\gamma}{2R^*(1-\gamma^{h+1})}, \quad \forall h = 0\ldots, H,\\ N_h &= \Big\lceil \frac{4R^*(1-\gamma^{h+1})|\mathbb{A}|^2}{(1-\gamma)\epsilon_h}\Big\|\frac{1}{\mu}\Big\|_\infty\Big\rceil, \quad \forall h = 0\ldots, H.\end{aligned}$$

Then, the stationary policy $\widehat{\pi}_H^*$ obtained by DynPG under exact gradients achieves $\|\mathcal{V}_\infty^* - \mathcal{V}_\infty^{\widehat{\pi}_H^*}\|_\infty \leq \epsilon$. The total number of gradient steps are given by

$$\sum_{h=0}^{H}N_h = \Big\lceil \frac{24R^*|\mathbb{A}|^2}{(1-\gamma)^2\epsilon}\Big\|\frac{1}{\mu}\Big\|_\infty\Big\rceil\Big\lceil \frac{\log(6R^*(1-\gamma)^{-2}\epsilon^{-1})}{\log(\gamma^{-1})}\Big\rceil^2 + \Big\lceil \frac{24R^*(1-\gamma^{H+1})|\mathbb{A}|^2}{(1-\gamma)^3\epsilon}\Big\|\frac{1}{\mu}\Big\|_\infty\Big\rceil.$$

Proof. The optimization procedure is the same as for the value function error by substituting ϵ with $\frac{(1-\gamma)\epsilon}{3}$. Moreover, note that

$$\sum_{h=0}^{H-1} \gamma^{H-h-1}\epsilon_h = \frac{\epsilon(1-\gamma)}{6}.$$

Thus, $\epsilon_H \leq \frac{(1-\gamma)^2\epsilon}{6}$ is enough for Lemma 5.23 to hold and we obtain that using the (ϵ_h)'s defined in the claim results in an overall error for DynPG given by

$$\|\mathcal{V}_\infty^* - \mathcal{V}_\infty^{\widehat{\pi}_H^*}\|_\infty \leq \frac{3}{1-\gamma}\Big(\frac{\gamma^H R^*}{1-\gamma} + \sum_{h=0}^{H-1}\gamma^{H-h-1}\epsilon_t\Big) \leq \frac{3}{1-\gamma}\frac{\epsilon(1-\gamma)}{3} = \epsilon.$$

Note that we can again rewrite H to be

$$H = \Big\lceil \log_\gamma\Big(\frac{(1-\gamma)^2\epsilon}{6R^*}\Big)\Big\rceil = \Big\lceil \frac{\log((1-\gamma)^{-2}\epsilon^{-1}6R^*)}{\log(\gamma^{-1})}\Big\rceil.$$

Thus, the complexity bounds for the optimization epochs $h = 0, \dots, H$ are given by

$$\begin{aligned}
\sum_{h=0}^{H} N_h &= \sum_{h=0}^{H-1} N_h + N_H \\
&\leq \Big\lceil \frac{24R^*|\mathcal{A}|^2}{(1-\gamma)^2\epsilon}\Big\|\frac{1}{\mu}\Big\|_\infty\Big\rceil\Big\lceil \frac{\log(6R^*(1-\gamma)^{-2}\epsilon^{-1})}{\log(\gamma^{-1})}\Big\rceil^2 + \Big\lceil \frac{4R^*(1-\gamma^{H+1})|\mathcal{A}|^2}{(1-\gamma)\epsilon_H}\Big\|\frac{1}{\mu}\Big\|_\infty\Big\rceil \\
&\leq \Big\lceil \frac{24R^*|\mathcal{A}|^2}{(1-\gamma)^2\epsilon}\Big\|\frac{1}{\mu}\Big\|_\infty\Big\rceil\Big\lceil \frac{\log(6R^*(1-\gamma)^{-2}\epsilon^{-1})}{\log(\gamma^{-1})}\Big\rceil^2 + \Big\lceil \frac{24R^*(1-\gamma^{H+1})|\mathcal{A}|^2}{(1-\gamma)^3\epsilon}\Big\|\frac{1}{\mu}\Big\|_\infty\Big\rceil,
\end{aligned}$$

where we used the calculations in the proof of Theorem 5.22 and substituted ϵ by $\frac{(1-\gamma)\epsilon}{3}$ in the first inequality. In the second inequality we used $\epsilon_H = \frac{(1-\gamma)^2\epsilon}{6}$. ∎

5.4 Advantages and Limitations of DynPG

5.4.1 *Breaking the lower bound example with DynPG*

In [Li+23a] a lower bound example was given for which softmax PG takes exponential time in the expected horizon $(1-\gamma)^{-1}$ to converge. More precisely, it is shown that at least $|\mathcal{S}|^{2^{\Omega((1-\gamma)^{-1})}}$ gradient steps are required to approximate the optimal value function with $\epsilon = 0.15$ accuracy. The constructed MDP is designed such that the unknown constant c_γ in the upper bound on softmax PG (see Table 5.1) is exponential in the effective time horizon $(1-\gamma)^{-1}$. To understand this behavior, we take a closer look at the definition of $c_\gamma = \big(\min_{s\in\mathcal{S}} \inf_{n\geq 1} \pi^{\theta_n}(a^*(s)|s)\big)^{-1}$ in Lemma 3.19, where $a^*(s)$ denotes the optimal action in state s. To ensure small c_γ the probability of choosing the best action should not get close to 0 during training. But by construction in [Li+23a], finding the best action in state s decreases as long as the probability of choosing the best action in a previous state $s' < s$ is not close enough to 1. This results in the phenomenon that the gradient steps required to converge towards $a^*(s)$ grows at least geometrically as s increases. However, using DP one can solve the MDP within $|\mathcal{S}|$ steps of exact value iteration [Li+23a, Lem.

1]. This already implies that DynPG easily circumvents this exponential convergence time by employing DP. As future actions are determined by the previously trained policies, DynPG can evaluate the MDP under a non-stationary policy during training and thereby avoid the above described phenomenon.
The upper bound on the complexity of softmax DynPG in Theorem 5.19 provides a theoretical proof that the needed gradient steps scale at most with $(1-\gamma)^{-4}$ up to logarithmic factors.

5.4.2 *Numerical Example of DynPG*

To further demonstrate DynPG's effectiveness in the more general sampled-based gradient setting, we present a numerical study of a canonical example that has a similar flavor as the counterexample in [Li+23a]. This example is an extension of Example 6.7 in [SB18], which has been used to compare different variants of Q-learning algorithms, as it suffers from overestimation of the Q-values. Since the overestimation problem in Q-learning and the committal behavior problem in policy gradient are closely related [FHM18], we decided to use this example. The example is certainly artificial, as the number of actions and the rewards in the MDP are chosen to trap policy gradient from convergence. If we vary the MDP parameters from the current setting, DynPG consistently performs well but the advantage over vanilla PG will become less significant. The MDP is defined as follows:

- The state space is given by $\mathcal{S} := \{0, \dots, 6\}$; States $0, 3, 6$ are the terminal states and states $1, 2, 4, 5$ are the initial states. We sample s_0 uniformly from $\{1, 2, 4, 5\}$.
- The action space is given by $\mathcal{A} := \{0, \dots, 299\}$.
- The state transitions and state-dependent actions are visualized in Figure 5.2. Each node represents a state, with squared ones being terminal states and elliptical ones being initial states. Each arrow represents an action that deterministically transits from one state to another. From state 1 to state 0, there are a total of 300 possible actions, succinctly visualized using the dots.

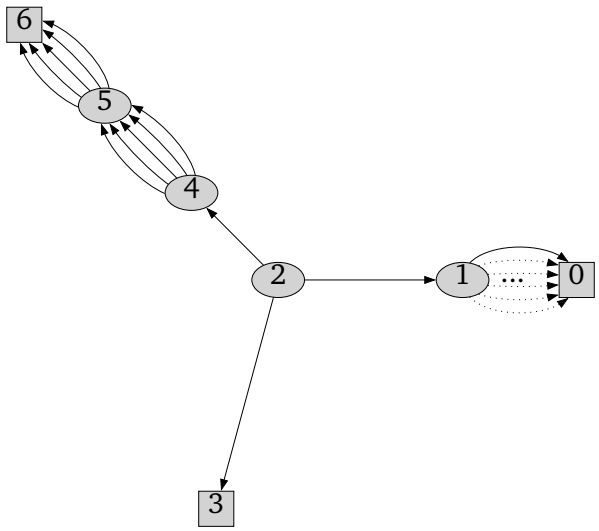


Figure 5.2: Visualization of the MDP state transitions.

- Taking any $a \in \mathcal{A}$ from state 1 reaches state 0 and receives a reward of $r(1,a) \sim \mathcal{N}(-0.3, 10)$.
- Taking any of the 5 possible actions from state 4 reaches state 5 and receives the reward of $r(4,a) \sim \mathcal{N}(1.25, 1.25)$.
- Taking any of the 5 possible actions from state 5 reaches state 6 and receives the reward of $r(5,a) \sim \mathcal{N}(1.25, 1.25)$.
- Taking any of the 3 possible actions from state 2 receives the reward of $r(2,a) = 0$.

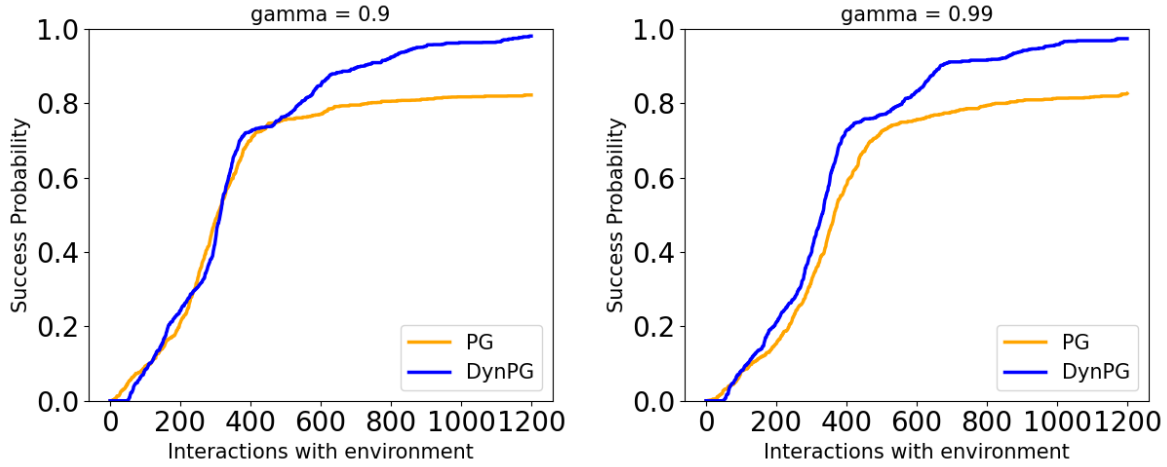
Experimental setup: We evaluated the performance of (stochastic) vanilla PG and (stochastic) DynPG under two different discount factors, $\gamma = 0.9$ and $\gamma = 0.99$. We used the tabular softmax parametrization studied in the convergence analysis for both algorithms. In DynPG, we used the 1-batch Monte-Carlo estimator to sample the gradient according to Theorem A.3. In vanilla PG, we chose the classical REINFORCE 1-batch estimator with truncation horizon 3 (cf. equation (3.10)), such that the estimator is also unbiased due to the episodic setting (the maximum episode length in our example is 3).
In DynPG, we chose the step size η_h and number of training steps N_h according to Theorem 5.22 and Theorem 5.24 and only fine-tuned the constants 2 and 45:

$$\eta_h = 2\frac{1-\gamma}{1-\gamma^h}, \qquad N_h = \left\lceil 45\frac{1-\gamma^{h+1}}{1-\gamma} \right\rceil.$$

We want to emphasize that the choices of η_h and N_h consistently perform well under different choices of γ, which underscores that the algorithmic parameters developed in our theory also provide good guidance in practice. For a fair comparison, we fine-tuned $\eta = 2\frac{1-\gamma}{1-\gamma^6}$ for stochastic vanilla PG, which is much larger than the pessimistic $\eta = c * (1-\gamma)^2$ suggested in Theorem 3.21. In Figure 5.3, we plotted the success probability in 2000 individual runs of the algorithm that achieves the overall value function error of less than $\epsilon = 0.01$ from the optimal, with the x-axis being the number of interactions with the environment. We observe that vanilla PG had a hard time solving this MDP, suffering from the high variance in the rewards from state 1 to state 0. The sample-based algorithm tends to concentrate on large rewards samples. DynPG circumvented this committal behaviour by more accurate estimated of the future Q-values.
For the case of $\gamma = 0.9$ presented in Figure 5.3 (a), we observe that the performance of DynPG and vanilla PG are similar for the first 400 interactions with the environment. However, vanilla PG fails to converge to the optimal policy and can only reach the success probability of around 0.8 under the given training budget. DynPG is shown to converge much faster and can solve the MDP with a probability of nearly 1 after 1200 samples. In Figure 5.3 (b), where we presented the case of $\gamma = 0.99$, one can observe a similar performance gain from DynPG compared to vanilla PG. DynPG converges consistently and faster, while vanilla PG fails to converge. These experiments support the theoretical findings from the previous section.

5.4.3 *Limitations and Modifications*

In the previous subsections we have seen reasonable examples, where DynPG outperforms vanilla PG. Still, there are examples, where vanilla PG already performs quite well and does not suffer



(a) (b)

Figure 5.3: For $\gamma = 0.9$ (a) and $\gamma = 0.99$ (b), we show the success probability of achieving the sub-optimality gap of $\epsilon = 0.01$ in the overall error. The x-axis is the number of interactions with the environment used by both sample-based vanilla PG and sample-based DynPG. The success probability is calculated by executing each algorithm 2000 times with randomly sampled initial states.

.

from committal behavior or high variance in the Q-value estimation, such that using DynPG as a more complex algorithm cannot surpass vanilla PG. We want to emphasize that DynPG should be kept in mind for cases where vanilla PG is particularly slow or even fails to converge.
In the following, we aim to discuss a challenge that may arise when applying DynPG in practice. In very complex MDPs, where the policy parametrization needs to be rich, the classical DynPG approach in Algorithm 6 can suffer from a storage problem. The number of policy parameters that need to be stored for future decisions scale linearly with the number of training steps. In order to circumvent this problem we introduce an actor-critic based modification of DynPG.

Dynamic Actor-Critic (DynAC). The general idea behind actor-critic in vanilla PG is to introduce a so called critic, a second parametrized class of function $(Q^w)_{w\in\mathbb{R}^l}$, which approximates the Q-function Q^{π_θ} in the classical PG Theorem (Theorem 3.15). Using the critic as estimator of the Q-values no more roll-outs are needed to estimate the rewards-to-go; compare to the (unbiased) REINFORCE estimator discussed in Section 3.1.2. For further reading regarding the classical actor-critic algorithm we refer the interested reader to [SB18, Sec. 13.5].
In the actor-critic variant of DynPG, we include an additional training procedure after optimizing policy $\widehat{\pi}_h^*$ to update the critic, $Q^{w_{h+1}} \approx Q_{h+1}^{\widehat{\pi}_h^*,\dots,\widehat{\pi}_0^*}$, based on the old critic, $Q^{w_h} \approx Q_h^{\widehat{\pi}_{h-1}^*,\dots,\widehat{\pi}_0^*}$, and the newly trained policy, $\widehat{\pi}_h^*$, using the relation:

$$Q_{h+1}^{\widehat{\pi}_h^*,\dots,\widehat{\pi}_0^*}(s,a) = r(s,a) + \gamma \sum_{s'} p(s'|s,a) \sum_{a'} \widehat{\pi}_h^*(a'|s') Q_h^{\widehat{\pi}_{h-1}^*,\dots,\widehat{\pi}_0^*}. \tag{5.16}$$

The critic is used in the policy gradient theorem as an estimator for $Q_{h+1}^{\pi_h}$ such that the new

gradient in DynAC is given by

$$G^{\text{Dyn-AC}} = \nabla_\theta \log(\pi^\theta(A|S))Q^{w_{h+1}}(S,A), \tag{5.17}$$

where $S \sim \mu$ and $A \sim \pi^\theta(\cdot|s)$ and π^θ is the policy we are currently training for epoch $h+1$. In this case, we do no longer have to store all trained policies and just store the currently trained policy and the current critic. For applications where the performance of vanilla PG is poor and the parametrized policy needs to be very rich such that storing many policies might cause storage issues, we emphasize to keep this modification in mind. We call this approach DynAC for dynamic actor-critic and summarize the steps in Algorithm 7.

Algorithm 7: DynAC for discounted MDPs

Result: Approximation of π^*, denoted as $\widehat{\pi}^*$.
Input: Initial state distribution μ, class of policies $(\pi^\theta)_{\theta\in\mathbb{R}^d}$, class of parametrization for the critic $(Q_w)_{w\in\mathbb{R}^l}$.

Set $h = 0$;
Train Q^{w_1} to approximate the reward function r;
while *Convergence criterion not met* **do**
  Initialize θ_0 (e.g., $\theta_0 \equiv 0$);
  Choose α_h and N_h (cf., Remark 5.11);
  for $n = 0, \dots, N_h - 1$ **do**
    Sample $G^{\text{Dyn-AC}} = \nabla_\theta \log(\pi^\theta(A|S))Q^{w_{h+1}}(S,A)$ with $S \sim \mu$ and $A \sim \pi^\theta(\cdot|S)$;
    Update $\theta_{n+1} = \theta_n + \alpha_h G$;
  end
  Set $\widehat{\pi}^*_h = \pi^{\theta_{N_h}}$;
  Train w_{h+2} s.t. $Q^{w_{h+2}}(s,a) \approx \mathbb{E}_{S\sim p(\cdot|s,a), A\sim\widehat{\pi}^*_h(\cdot|S)}[r(s,a) + \gamma Q^{w_{h+1}}(S,A)]$ for all s, a;
  Set $h = h + 1$;
end
Return $\widehat{\pi}^* = \widehat{\pi}^*_{h-1}$;

A theoretical analysis of this approach would require an assumption on the approximation error to train the Q-functions. As the gradients are no longer unbiased, convergence towards the global optimum cannot be theoretically guaranteed and the bias errors will appear in the overall error. We leave further investigations to analyze this approach in practise and theory to future work.

5.4.4 *Comparison to NPG*

Finally, we compare the theoretical performance of DynPG to Natural Policy Gradient (NPG). NPG is a version of PG where the natural gradient of the objective function is used in gradient ascent instead of the gradient. In [Aga+21] it is shown that softmax NPG has a convergence rate $O((1-\gamma)^{-2}\epsilon^{-1})$ and so the algorithm achieves a dependency in the number of required gradient steps which is also explicit and even better in terms of the effective horizon compared to DynPG. Thus, in the exact gradient setting softmax DynPG cannot compete with softmax NPG. However, in the sample based setting it is noteworthy that it requires much more computational power to estimate the natural gradient compared to the gradient used in DynPG. DynPG reduces the variance in the gradient estimation by using fixed future policies (cf. Theorem 5.6). Additionally,

fewer samples are required in DynPG compared to PG and even more so compared to NPG. NPG is a quasi second order method, such that implementation is expensive. Thus, it is plausible that DynPG will outperform NPG in sample based settings.

Remark 5.25. We do not carry out an analysis in the stochastic setting as we did in Section 4.4, because the results rely on unrealistic very large batch sizes and we cannot expect that a comparison gives a realistic inside to the sample based setting. Instead a sample based implementation on a range of bench mark problems can give a more realistic comparison. As this thesis is of theoretical nature, we leave it as future work to examine this research question.

It is natural to check how the upper bounds change when vanilla PG is replaced by NPG in dynamic policy gradient. Replacing the sample of the gradient in Algorithm 6 with a sample of the natural gradient results in the DynNPG Algorithm and there is no need to restate the algorithm here. In the following, we will verify that the upper bounds on the performance of softmax DynPG and softmax DynNPG are the same.
By [Aga+21, Lem. 5.1] with $\gamma = 0$ (contextual bandit interpretation of $\mathcal{V}_{h+1}^{\{\pi_\theta,\Lambda\}}$) we can obtain that the natural gradient is given by $\mathcal{A}_{h+1}^{\{\pi_\theta,\Lambda\}}$, where $\mathcal{A}_{h+1} \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ is the advantage function defined in equation (5.2). Under exact gradient assumption the natural gradient update for softmax DynNPG to train policy $\widehat{\pi}_h^*$ is therefore given by

$$\theta_{n+1} = \theta_n + \alpha\, \mathcal{A}_{h+1}^{\{\pi_\theta,\Lambda\}}.$$

We suppose throughout this section that the following assumption holds.

Assumption 5.26. *Suppose that $\mu(s) > 0$ for all $s \in \mathcal{S}$. Furthermore, the parametrization in DynNPG $(\pi_\theta)_{\theta\in\mathbb{R}^{|\mathcal{S}||\mathcal{A}|}}$ is chosen to be the tabular softmax parametrization introduced in equation* (3.12). *We assume further that the natural gradient can be accessed exactly.*

We obtain the following theorem for the optimization error in every iteration of DynNPG. This is the corresponding result to Theorem 5.18 of softmax DynPG.

Theorem 5.27. *Let Assumption 5.26 hold. Let $h \geq 0$, $\epsilon_h > 0$ and $\Lambda = \widehat{\mathbb{\pi}}_h^* \in \Pi^h$ be a collection of h arbitrary policies. Then, using step size $\alpha_h = \frac{(1-\gamma)\log(|\mathcal{A}|)}{(1-\gamma^{h+1})R^*}$ and gradient steps $N_h = \frac{2R^*(1-\gamma^{h+1})\log(|\mathcal{A}|)}{(1-\gamma)\epsilon_h}\left\|\frac{1}{\mu}\right\|_\infty$ in DynNPG guarantees that the policy $\widehat{\pi}_h^*$ of iteration h achieves*

$$\|T^*(\mathcal{V}_h^{\widehat{\mathbb{\pi}}_h^*}) - \mathcal{V}_{h+1}^{\widehat{\mathbb{\pi}}_{h+1}^*}\|_\infty \leq \epsilon_h.$$

Sketch of Proof. The proof follows by adapting the convergence result [Aga+21, Thm. 5.3] for softmax NPG to contextual bandits with bounded rewards. Although in [Aga+21] rewards in $[0, 1]$ are assumed, the proof of [Aga+21, Thm. 5.3] can be generalized to the more general reward assumption $r \in [-\mathbf{r}, \mathbf{r}]$ by changing the upper bound used on the value functions in the final inequality of the proof, i.e.

$$V^{(T)} - V^{(0)} \leq (1-\gamma)^{-1}$$

to be

$$V^{(T)} - V^{(0)} \leq 2\mathbf{r}(1-\gamma)^{-1}.$$

For MDPs with rewards in $[-\mathbf{r}, \mathbf{r}]$ we obtain the global convergence of NPG

$$V_\infty^{\pi_{\theta_n}} \geq V_\infty^* - \frac{\log(|\mathcal{A}|)}{\alpha n} - \frac{2\mathbf{r}}{(1-\gamma)^2 n}$$

directly from [Aga+21, Thm. 5.3]. For contextual bandit ($h = 0$ and $\gamma = 0$) with rewards in $[-\mathbf{r}, \mathbf{r}]$ we obtain that

$$V_0^{\pi_{\theta_n}} \geq V_0^* - \frac{\log(|\mathcal{A}|)}{\alpha n} - \frac{2\mathbf{r}}{n}.$$

We apply this result to the contextual bandit with value function $V_{h+1}^{\{\pi_\theta, \Lambda\}}$, the objective function we aim to optimize in epoch h. Note that the rewards in our setting are bounded in $[-\frac{R^*(1-\gamma^{h+1})}{1-\gamma}, \frac{R^*(1-\gamma^{h+1})}{1-\gamma}]$ (cf. Lemma 5.13). Thus, we get

$$V_{h+1}^{\{\pi_{\theta_n}, \widehat{\mathbb{T}}_h^*\}} \geq T^*(V_h^{\widehat{\mathbb{T}}_h^*}) - \frac{\log(|\mathcal{A}|)}{\alpha_h n} - \frac{2R^*(1-\gamma^{h+1})}{(1-\gamma)n}$$

and still obtain a dependence on $(1-\gamma)^{-1}$ steaming from the horizon of the h-step value function. Choosing $\alpha_h = \frac{(1-\gamma)\log(|\mathcal{A}|)}{(1-\gamma^{h+1})R^*}$, we have that

$$T^*(V_h^{\widehat{\mathbb{T}}_h^*}) - V_{h+1}^{\{\pi_{\theta_n}, \widehat{\mathbb{T}}_h^*\}} \leq \frac{3R^*(1-\gamma^{h+1})}{(1-\gamma)n}.$$

For $N_h = \frac{3R^*(1-\gamma^{h+1})}{(1-\gamma)\epsilon_h}$ and $\widehat{\pi}_h^* = \pi_{\theta_{N_h}}$ we arrive at the theorem. ∎

We note directly that the gradient step N_h and the step size α_h agrees in terms of ϵ_h and γ with the ones obtained in Theorem 5.18 for softmax DynPG. Thus, optimizing over (ϵ_h) and H as in Section 5.3.2 would results in the same complexity bounds for DynNPG as for DynPG.
Although no improvement of bounds is obtained in DynNPG under exact gradients we leave it to future work to compare the convergence behavior of DynPG and DynNPG/NPG in sample based settings.

Almost Sure Convergence Rates under Gradient Domination

6

Until now, the thesis focused on convergence rates for different variants of policy gradient methods in RL. Considering the practically used stochastic versions of the algorithms, the results regarding convergence are very rare and unsatisfying, compare to Remark 5.25, as unrealistic large batch sizes are required. This is due to the non-uniform gradient domination property (cf. Lemma 3.19, Lemma 4.11, Lemma 4.19) which can just be controlled through large batch sizes or small step sizes (cf. Section 4.4). Nevertheless, in practical applications we often observe strictly better performance of these algorithms even when different parametrizations like neural networks are considered. The settings examined in this chapter (especially the local setting) are motivated by the aim of gaining a deeper understanding of this practical behavior.
After a brief literature review in Section 6.1, where we classify the contributions in this chapter, we dive into a preliminary discussion on a super-martingale convergence result in Section 6.2.
The first contribution, in Section 6.3, focuses on almost sure convergence rates for the error $f(X_n) - f^*$ in stochastic gradient schemes under (weak) gradient domination with parameter $\beta \in [\frac{1}{2}, 1]$ (Definition 2.7). We prove that stochastic gradient descent (SGD) and stochastic heavy ball (SHB) converge almost surely and in expectation towards the global optimum with rate arbitrarily close to $o(n^{-\frac{1}{4\beta-1}})$. The rate of convergence that we obtain (a.s. and in expectation) depends on the gradient domination parameter β and is the same for both algorithms and convergence types. For SGD this rate is arbitrarily close to the tight upper bound known in expectation [Fat+22], while the almost sure convergence rate is new for the (weak) gradient domination assumption (see Theorem 6.2 and discussion afterwards). To the best of our knowledge for SHB this is the first convergence result towards global optima under (weak) gradient domination, for both the almost sure convergence and convergence in expectation (see Theorem 6.3).
Second, in Section 6.5,we assume that the gradient domination property is only locally fulfilled, where we distinguish between locally around a stationary point or locally around the global minimum. In both cases we prove that SGD remains within the good local region with high probability, given a small enough step size. Conditioned on this event we provide converges rates almost surely and in expectation towards the local or global minimum respectively with the same convergence speed as in the global case (see Theorem 6.7).
The local setting around stationary points is especially of interest for machine learning applications with (deep) neural networks, as they fulfill this property [DK21]. In particular, we demonstrate in Section 6.6 that it encompasses the training task of deep neural networks with analytic activation functions in supervised learning. Our result illustrates that the iterates of SGD are likely to become trapped in areas of local minima when the step size is is small. We verify under mild conditions, that SGD converges to local minima with given convergence speed (see Corollary 6.20).
Finally in Section 6.7, we apply the results obtained in the local setting around the global minimum to policy gradient training in reinforcement learning. For infinite-time horizon MDPs, we verify the local gradient domination around the global optimum for softmax vanilla PG and entropy-regularized softmax PG. In both cases, we prove that the local rate of convergence under local gradient domination applies to stochastic (regularized) softmax PG (see Corollary 6.23

β	**Step size**	**Rate**	**Dom.**	**Algo.**	**Conv.**	**Ref.**
$\frac{1}{2}$	$\Theta\left(n^{-1+\epsilon}\right)$	$o\left(n^{-1+\epsilon}\right)$	global	SGD	a.s.	Thm. 6.2 (i); e.g. [LY22, Thm. 1]
					$\mathbb{E}$	Thm. 6.2 (ii); e.g. [KR23, Thm. 3]
				SHB	a.s	Thm. 6.3 (i); e.g. [LY22, Thm. 2]
					$\mathbb{E}$	Thm. 6.3 (ii); e.g. [LLX23, Thm. 4.3]
			local*	SGD	a.s.	Thm. 6.7 (ii); Thm. 6.16 (ii)
					$\mathbb{E}$	Thm. 6.7 (iii); Thm. 6.16 (iii); e.g. [Mer+20, Thm. 4]
$(\frac{1}{2}, 1]$	$\Theta\left(n^{-\frac{2\beta}{4\beta-1}}\right)$	$o\left(n^{-\frac{1}{4\beta-1}+\epsilon}\right)$	global	SGD	a.s.	Thm. 6.2 (i)
					$\mathbb{E}$	Thm. 6.2 (ii); e.g. [Fat+22, Cor. 1]
				SHB	a.s.	Thm. 6.3 (i)
					$\mathbb{E}$	Thm. 6.3 (ii)
			local*	SGD	a.s.	Thm. 6.7 (ii); Thm. 6.16 (ii)
					$\mathbb{E}$	Thm. 6.7 (iii); Thm. 6.16 (iii)

Table 6.1: Summary of known and new results. Table presents convergence rates for tuned step size ($\epsilon > 0$ arbitrarily small). Dom.: gradient domination holds locally or globally; local*: additional assumption on α_1 required and results holds only locally. a.s.: almost surely; $\mathbb{E}$: in expectation. Ref.: for some cited results minor adjustments are necessary.

and Corollary 6.25). For finite-time horizon MDPs, we verify the local gradient domination around the global optimum for every optimization step in FT-DynPG. We obtain almost sure convergence under good initialization (Corollary 6.27) and improve upon the very large batch sizes required in Theorem 4.36. Although these rates hold only under good initialization, we can characterize the initialization regions for the infinite- and finite-time cases explicitly.
We summarize the contributions of this chapter in Table 6.1. These findings are also illustrated in a numerical toy experiment in Section 6.4, where the performance of SGD and SHB for monomials with increasing degree is implemented.

6.1 Literature Review and Classification of the Contribution

The roots of stochastic gradient methods trace back to Robbins and Monro [RM51]. Since then, various variants of SGD have been established as fundamental algorithms for optimizing complex models in the realm of machine learning. We refer to Bottou, Curtis, and Nocedal [BCN18] and

Garrigos and Gower [GG24] for a detailed overview.
We start the review with the literature deriving convergence rates in expectation for SGD. Under the assumptions of smoothness and (strong) convexity Polyak [Pol87], Moulines and Bach [MB11], Nguyen et al. [Ngu+18], Wang et al. [Wan+21], and Liu et al. [Liu+23] studied convergence rates towards global optima. Moreover, many articles additionally analyze the non-convex case and prove convergence rates for the gradient norm towards zero [GL13; Li+23b; Liu+23; Ngu+23].
Notably, several other results regarding convergence of SGD towards global optima have been established under the gradient domination setting [KNS16]. Bassily, Belkin, and Ma [BBM18] demonstrate exponential convergence rates in expectation in the overparameterized setting under strong gradient domination. See also Madden, Dall'Anese, and Becker [MDB21] and Liu, Zhu, and Belkin [LZB22], who show convergence rates of order $O(\frac{1}{n})$ for neural networks using the (strong) gradient domination property. Scaman, Malherbe, and Santos [SMS22] provide high-probability bounds on the approximation error under a generalized gradient domination property, the so-called Separable-Łojasiewicz assumption, fulfilled by smooth neural networks. Lei et al. [Lei+20] also assume strong gradient domination and weaken the smoothness assumption through α-Hölder continuity, achieving a rate of $O(\frac{1}{n^\alpha})$ in expectation. Khaled and Richtárik [KR23] introduce the (ABC) condition and show $O(\frac{1}{n})$ convergence under strong gradient domination. Furthermore, Fatkhullin et al. [Fat+22] and Fontaine, Bortoli, and Durmus [FBD21] consider generalizations of gradient domination that include our definition as a special case. They derive convergence rates in expectation which we encompass with our result and extend to almost sure convergence (see also the discussion behind Theorem 6.2).
All the results mentioned so far consider convergence in expectation or high-probability bounds, although originally, motivated by Robbins and Siegmund [RS71], research commenced with the quest for almost sure convergence rates for gradient methods. In recent years, Sebbouh, Gower, and Defazio [SGD21] and, building upon it, Liu and Yuan [LY22] derive almost sure convergence rates towards global optima under strong convexity. Sebbouh, Gower, and Defazio [SGD21] also analyzed almost sure convergence rates for SHB but under the assumption of convexity and Liu and Yuan [LY22] study SHB under (strong) convexity and in a non-convex setting. Returning the attention back to SGD with respect to gradient domination also some almost sure convergence results have been established. As an extension to the PL-type gradient domination, Chouzenoux, Fest, and Repetti [CFR23] assume the so-called KL property, which contains gradient domination as a special case. The authors demonstrate almost sure convergence to a critical point, though without a rate. To conclude, to the best of our knowledge the derived almost sure convergence rate under gradient domination in Theorem 6.2 is novel.
Next, we want to provide further insights to the literature regarding SHB. In the realm of momentum methods, Polyak's Heavy-Ball Method (HBM) [Pol64] and Nesterov's accelerated gradient method [Nes83] stand out as a foundational contribution. The authors of [GPS18] provide a detailed description of the stochastic formulation of HBM and establish almost sure convergence but without giving a rate. In Yang, Lin, and Li [YLL16], Orvieto, Kohler, and Lucchi [OKL20], Yan et al. [Yan+18], Mai and Johansson [MJ20], and Zhou et al. [Zho+20] convergence rates in expectation are shown in (strongly) convex and non-convex settings, where the non-convex analysis covers convergence of the norm of the gradient. Gess and Kassing [GK23] show convergence of momentum methods under the strong gradient domination property and prove linear convergence due to an overparametrized machine learning setting. In [LLX23] the

authors determine $O(\frac{1}{n})$ convergence rate for SHB under strong gradient domination. Our main result for SHB presented in Theorem 6.3 describes almost sure convergence and convergence in expectation under global gradient domination. Both result are quantified with a given rate of convergence.

Finally, we aim to differentiate the present article from existing results on the convergence of SGD under the assumption of local gradient domination. Dereich and Kassing [DK21] demonstrate almost sure convergence of SGD to a stationary point under the local gradient domination (for x^*), provided that the process (X_n) remains local, albeit without a rate. Fehrman, Gess, and Jentzen [FGJ20] present a local analysis of SGD towards minima without any gradient domination assumption. Instead, a rank assumption is imposed on the Hessian, and mini-batches, along with resampling, are leveraged to ensure convergence to the global optimum with high probability. The resulting rate does not converge to zero and requires an increasing batch size. Mertikopoulos et al. [Mer+20] demonstrate, under the global Lipschitz assumption on the objective, that SGD almost surely converges to a stationary point and the authors derive a local convergence analysis under local strong convexity. Our analysis in Section 6.5 builds upon Mertikopoulos et al. [Mer+20] and generalizes their results to the local gradient domination property. In our analysis we distinguish the cases where the local gradient domination property holds in a neighbourhood of a local minimum or in the neighbourhood of the global optimum respectively. Finally, we would like to acknowledge that related results have been independently obtained in the recent preprint [QMM24].

For the application in the training of DNNs, it is worth noting that local convergence of SGD has been analyzed under stronger variants of gradient domination [Woj23]; [AL24]. Due to the stronger form of gradient domination, specific sub-classes of DNNs need to be considered to verify these assumptions whereas our result is only constrained to analytic activation functions. Under the machine learning noise conditions in [Woj23], convergence toward zero loss with high probability is shown, provided that the initial loss is sufficiently small. In contrast, [AL24] demonstrate convergence towards zero loss under initialization in a local (strong) Łojasiewicz region. Indeed, one can construct DNNs satisfying the latter condition [Cha22].

For the application in reinforcement learning, we have seen in Section 3.1.2, Chapter 4 and Chapter 5 that choosing the tabular softmax parametrization in PG algorithms results in objective functions which fulfill a non-uniform gradient domination property. The convergence of PG for exact gradients is quite well understood, but convergence rates for stochastic PG are rare and mostly require very large batch sizes (see [DZL22] and Theorem 4.36). In Section 6.7, we consider both the unregularized and entropy regularized setting and observe that one can also achieve convergence arbitrarily close to $o(\frac{1}{n})$ without the need for an increasing batch size. Moreover, the local convergence occurs almost surely on an event with high probability. It is noteworthy that a similar local analysis for stochastic policy gradient under entropy regularization is presented in [DZL23]. Their local result is also based on [Mer+20], but requires an increasing batch size sequence to obtain $O(\frac{1}{n})$-convergence towards the regularized optimum with high probability. In contrast, we consider both the unregularized and entropy regularized setting and observe that one can also achieve convergence arbitrarily close to $o(\frac{1}{n})$ without the need for an increasing batch size. Moreover, the local convergence occurs almost surely on an event with high probability.

6.2 Preliminary Discussion on Super-Martingale Convergence Rates

Let $f : \mathbb{R}^d \to \mathbb{R}$ be a differentiable and L-smooth function, i.e. f fulfills Assumption 2.1. In Section 2.2.3, we have sketched how to combine the global gradient domination property with smoothness to derive a recursive inequality of the form

$$\mathbb{E}[Y_{n+1} \mid \mathcal{F}_n] \leq (1 + c_1\alpha_n)Y_n - c_2\alpha_n Y_n^{2\beta} + c_3\alpha_n^2,$$

where $Y_n := f(X_n) - f^*$ (cf. equation (2.9)). For analysing these inequalities, we must deal separately with the strong gradient domination case ($\beta = \frac{1}{2}$) and the weak gradient domination case ($\beta > \frac{1}{2}$) to avoid divisions by zero. For the former case the recursive inequality simplifies, whereas a more complex analysis is required for the latter. To establish almost sure convergence rates we employ convergence lemmas for super-martingales based on the Robbins-Sigmund Theorem. This methodology has been introduced in [SGD21] and further utilized in [LY22] to analyze SGD and SHB under (strong) convexity. In the following, we illustrate how to extend the arguments to convergence under the global gradient domination property. While the extension to the strong gradient domination case is straightforward, we have to invest more work in the weak case.
Here is our super-martingale result that also encompasses [LY22, Lem. 1] when $\beta = \frac{1}{2}$ for completeness:

Lemma 6.1. *Let $(Y_n)_{n\in\mathbb{N}}$ be a sequence of non-negative random variables on an underlying probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with natural filtration $(\mathcal{F}_n)_{n\in\mathbb{N}}$ and suppose there exists $\beta \in [\frac{1}{2}, 1]$, $c_1, c_3 \geq 0$ and $c_2 > 0$ such that*

$$\mathbb{E}[Y_{n+1} \mid \mathcal{F}_n] \leq (1 + c_1\alpha_n^2)Y_n - c_2\alpha_n Y_n^{2\beta} + c_3\alpha_n^2,$$

for all $n \geq 1$, where $\alpha_n = \Theta(\frac{1}{n^\theta})$ for some fixed $\theta \in \left(\frac{1}{2}, 1\right)$. Then, for any

$$\eta \in \begin{cases} \left(\max\{2 - 2\theta, \frac{\theta+2\beta-2}{2\beta-1}\}, 1\right) & : \beta \in (\frac{1}{2}, 1] \\ (2 - 2\theta, 1) & : \beta = \frac{1}{2} \end{cases},$$

$(Y_n)_{n\in\mathbb{N}}$ vanishes almost surely with $Y_n \in o\left(\frac{1}{n^{1-\eta}}\right)$.

Proof. In the following, we treat both cases $\beta = \frac{1}{2}$ and $\beta \in (\frac{1}{2}, 1]$ separately.
$\underline{\beta = \frac{1}{2}}$: In this case, the inequality reduces to

$$\mathbb{E}[Y_{n+1} \mid \mathcal{F}_n] \leq (1 + c_1\alpha_n^2 - c_2\alpha_n)Y_n + c_3\alpha_n^2.$$

By the choice of α_n, there exists some $N > 0$ and $\widetilde{c}_1 > 0$ such that $c_2\alpha_n - c_1\alpha_n^2 \geq \widetilde{c}_1\alpha_n$ for all $n \geq N$. Hence, for all $n \geq N$

$$\mathbb{E}[Y_{n+1} \mid \mathcal{F}_n] \leq (1 - \widetilde{c}_1\alpha_n)Y_n + c_3\alpha_n^2$$

such that the claim follows by [LY22, Lem. 1].
$\underline{\beta \in (\frac{1}{2}, 1]}$: The proof uses the elementary inequality

$$(n+1)^{1-\eta} \leq n^{1-\eta} + (1-\eta)n^{-\eta}, \tag{6.1}$$

which was also applied and proved in [LY22, Lem. 1]. The aim is to apply the Robbins-Siegmund corollary, lemma 2.15, in order to derive the almost sure convergence rate. Let $1 \leq q < 2$ be arbitrary for now. The key step of the proof is the following computation

$$\begin{aligned}\mathbb{E}[Y_{n+1} \mid \mathcal{F}_n] &\leq (1 + c_1\alpha_n^2)Y_n - c_2\alpha_n Y_n^{2\beta} + c_3\alpha_n^2 \\ &= (1 + c_1\alpha_n^2)Y_n - c_2\alpha_n^q Y_n + c_2\alpha_n^q Y_n - c_2\alpha_n Y_n^{2\beta} + c_3\alpha_n^2 \\ &= (1 + c_1\alpha_n^2 - c_2\alpha_n^q)Y_n + c_2\alpha_n\left(\alpha_n^{q-1}Y_n - Y_n^{2\beta}\right) + c_3\alpha_n^2.\end{aligned} \tag{6.2}$$

Similar to the case $\beta = \frac{1}{2}$ there exists some $N > 0$ and $\widetilde{c}_1 > 0$ such that $c_2\alpha_n^q - c_1\alpha_n^2 \geq \widetilde{c}_1\alpha_n^q$ for all $n \geq N$. Hence, for all $n \geq N$ we obtain the iterative inequality of the form

$$\mathbb{E}[Y_{n+1} \mid \mathcal{F}_n] \leq (1 - \widetilde{c}_1\alpha_n^q)Y_n + c_2\alpha_n\left(\alpha_n^{q-1}Y_n - Y_n^{2\beta}\right) + c_3\alpha_n^2. \tag{6.3}$$

The function $x \mapsto ax - bx^{2\beta}$ takes it maximum at $\bar{x} = \left(\frac{a}{2b\beta}\right)^{\frac{1}{2\beta-1}}$ such that

$$\begin{aligned}\alpha_n(\alpha_n^{q-1}Y_n - Y_n^{2\beta}) &\leq \frac{\alpha_n^{q+\frac{q-1}{2\beta-1}}}{(2\beta)^{\frac{1}{2\beta-1}}} - \frac{\alpha_n^{1+\frac{(q-1)2\beta}{2\beta-1}}}{(2\beta)^{\frac{2\beta}{2\beta-1}}} \\ &= \frac{1}{(2\beta)^{\frac{1}{2\beta-1}}}\alpha_n^{\frac{2q\beta-1}{2\beta-1}} - \frac{1}{(2\beta)^{\frac{2\beta}{2\beta-1}}}\alpha_n^{\frac{2q\beta-1}{2\beta-1}} \\ &= (2\beta)^{-\frac{1}{2\beta-1}}(1 - \frac{1}{2\beta})\alpha_n^{\frac{2q\beta-1}{2\beta-1}}\end{aligned} \tag{6.4}$$

holds almost surely. We define $\widetilde{c}_2 = c_2(2\beta)^{-\frac{1}{2\beta-1}}(1 - \frac{1}{2\beta}) \in (0, \infty)$ for $\beta \in (\frac{1}{2}, 1)$ and proceed with

$$\mathbb{E}[Y_{n+1} \mid \mathcal{F}_n] \leq (1 - \widetilde{c}_1\alpha_n^q)Y_n + \widetilde{c}_2\alpha_n^{\frac{2\beta q-1}{2\beta-1}} + c_3\alpha_n^2. \tag{6.5}$$

Next, we apply the elementary inequality, equation (6.1), and choose q such that $\frac{1}{2} < \theta \leq \frac{1}{q} \leq 1$. Moreover by the choice of α_n, there exists some $c_4 > 0$ such that $\widetilde{c}_1\alpha_n^q \geq \frac{c_4}{t^{q\theta}}$ for all $n \geq N$. It follows that for all $n \geq N$

$$\begin{aligned}&\mathbb{E}[(n+1)^{1-\eta}Y_{n+1} \mid \mathcal{F}_n] \\ &\leq (n+1)^{1-\eta}(1 - \widetilde{c}_1\alpha_n^q)Y_n + (n+1)^{1-\eta}\widetilde{c}_2\alpha_n^{\frac{2\beta q-1}{2\beta-1}} + (n+1)^{1-\eta}c_3\alpha_n^2 \\ &\leq (n^{1-\eta} + (1-\eta)n^{-\eta})(1 - \frac{c_4}{n^{q\theta}})Y_n + (n+1)^{1-\eta}\widetilde{c}_2\alpha_n^{\frac{2\beta q-1}{2\beta-1}} + (n+1)^{1-\eta}c_3\alpha_n^2 \\ &= \left(1 + \frac{1-\eta}{n} - \frac{c_4}{n^{q\theta}} - \frac{c_4(1-\eta)}{n^{q\theta+1}}\right)n^{1-\eta}Y_n + (n+1)^{1-\eta}\widetilde{c}_2\alpha_n^{\frac{2\beta q-1}{2\beta-1}} + (n+1)^{1-\eta}c_3\alpha_n^2.\end{aligned}$$

We set $\widetilde{c}_3 = \max\{\widetilde{c}_2, c_3\}$ such that for all $n \geq N$

$$\begin{aligned}&\mathbb{E}[(n+1)^{1-\eta}Y_{n+1} \mid \mathcal{F}_n] \\ &\leq \left(1 + \frac{1-\eta}{n} - \frac{c_4}{n^{q\theta}} - \frac{c_4(1-\eta)}{n^{q\theta+1}}\right)n^{1-\eta}Y_n + \widetilde{c}_3(n+1)^{1-\eta}(\alpha_n^{\frac{2\beta q-1}{2\beta-1}} + \alpha_n^2).\end{aligned}$$

Observe that $q\theta \leq 1$ by condition $\theta \leq \frac{1}{q}$. Hence, there exists $\widetilde{c}_4 > 0$ and $\widetilde{N} > N$for sufficiently large $\widetilde{N} \geq N$ such that for all $n \geq \widetilde{N}$ we have

$$\mathbb{E}[(n+1)^{1-\eta} Y_{n+1} \mid \mathcal{F}_n] \leq (1 - \widetilde{c}_4 \frac{1}{n^{q\theta}}) n^{1-\eta} Y_n + c_3 (n+1)^{1-\eta} (\alpha_n^{\frac{2\beta q-1}{2\beta-1}} + \alpha_n^2) \tag{6.6}$$

In order to apply Robbins-Siegmund, more precisely Lemma 2.15, we are going to verify the following three sufficient conditions:

$$\sum_{n=\widetilde{N}}^{\infty} \frac{1}{n^{q\theta}} = \infty, \tag{6.7}$$

$$\sum_{n=\widetilde{N}}^{\infty} n^{1-\eta-2\theta} < \infty, \tag{6.8}$$

$$\sum_{n=\widetilde{N}}^{\infty} n^{1-\eta-\frac{\theta(2\beta q-1)}{2\beta-1}} < \infty. \tag{6.9}$$

Then, $Y_n \in o\left(\frac{1}{n^{1-\eta}}\right)$ almost surely.
The first condition, equation (6.7), is obviously satisfied, since we assume $\theta \leq \frac{1}{q}$. For the second condition, equation (6.8), we may choose $\theta > 1 - \frac{\eta}{2}$ such that $1-\eta-2\theta < -1$. The third condition, equation (6.9), gives $1 - \eta - \frac{\theta(2\beta q-1)}{2\beta-1} < -1$ which leads to the condition $\theta > \frac{(2-\eta)(2\beta-1)}{2\beta q-1}$. Hence, all together we obtain the sufficient condition

$$\theta \in \left(\max\left\{\frac{(2-\eta)(2\beta-1)}{2\beta q-1}, 1-\frac{\eta}{2}\right\}, \frac{1}{q}\right].$$

In the following, we consider the two cases separately that correspond to the maximum being either $1 - \frac{\eta}{2}$ or $\frac{(2-\eta)(2\beta-1)}{2\beta q-1}$. The first case occurs precisely for $\frac{1}{q} \leq \frac{2\beta}{4\beta-1}$, the latter one for $\frac{1}{q} \geq \frac{2\beta}{4\beta-1}$.
Firstly, let $\frac{1}{q} \leq \frac{2\beta}{4\beta-1}$. In this situation the sufficient condition on θ simplifies to

$$\theta \in \left(1-\frac{\eta}{2}, \frac{1}{q}\right].$$

The interval is non-empty for $\frac{1}{q} > \frac{2-\eta}{2}$, which requires $\eta \in (\frac{4\beta-2}{4\beta-1}, 1)$.
Secondly, let $\frac{1}{q} \geq \frac{2\beta}{4\beta-1}$. In this situation the sufficient condition on θ simplifies to

$$\theta \in \left(\frac{(2-\eta)(2\beta-1)}{2\beta q-1}, \frac{1}{q}\right],$$

the interval is non-empty for $\frac{1}{q} < 2\beta\eta - 2\beta + 2 - \eta$. Hence, $\frac{1}{q} \in (\frac{2\beta}{4\beta-1}, 2\beta\eta - 2\beta + 2 - \eta)$ which requires the condition $\eta \in (\frac{4\beta-2}{4\beta-1}, 1)$.
Either case yields sufficient conditions on θ and η (depending on the auxiliary variable q) under which $Y_n \in o\left(\frac{1}{n^{1-\eta}}\right)$ holds almost surely. We will now utilize the free variable q to prove the claim.

- Let $\theta \in (\frac{1}{2}, \frac{2\beta}{4\beta-1})$: We set $q = \frac{4\beta-1}{2\beta}$ and use the first case. The assumption $\eta > 2 - 2\theta = \max\{2-2\theta, \frac{\theta+2\beta-2}{2\beta-1}\}$ implies $\theta \in \left(1 - \frac{\eta}{2}, \frac{1}{q}\right]$. (Note that $\eta > \frac{4\beta-2}{4\beta-1}$ is automatically fulfilled by $2 - 2\theta > \frac{4\beta-2}{4\beta-1}$ for this choice of θ.)
- Let $\theta \in [\frac{2\beta}{4\beta-1}, 1)$: By assumption we have $\eta > \frac{\theta+2\beta-2}{2\beta-1} = \max\{2-2\theta, \frac{\theta+2\beta-2}{2\beta-1}\}$. We choose some $\frac{1}{q} \in (\theta, 2\beta\eta - 2\beta + 2 - \eta)$ and use the second case. (Note that $\eta > \frac{4\beta-2}{4\beta-1}$ again is automatically fulfilled by $\frac{\theta+2\beta-2}{2\beta-1} > \frac{4\beta-2}{4\beta-1}$ for this choice of θ.)

All in all we have proved that $\theta \in (\frac{1}{2}, 2)$ implies $Y_n \in o\left(\frac{1}{n^{1-\eta}}\right)$ almost surely for all $\eta \in (\max\{2 - 2\theta, \frac{\theta+2\beta-2}{2\beta-1}\}, 1)$. ■

6.3 Almost Sure Convergence under Global Gradient Domination

6.3.1 *Convergence Rates for SGD*

Combining the recursive inequality in equation (2.9) with the super-martingale convergence result from Lemma 6.1 leads to the following theorem, which to the best of our knowledge our theorem presents the first convergence rate for SGD under weak gradient domination with respect to almost sure convergence. This is in contrast to [LY22; SGD21] where the authors derive almost sure convergence rates in non-convex settings but only for the gradient norm to zero.

THEOREM 6.2. *Suppose Assumption 2.1 and Assumption 2.12 are fulfilled and let f satisfy the global gradient domination property from Definition 2.7 with $\beta \in [\frac{1}{2}, 1]$. Denote by (X_n) the sequence generated by equation* (SGD) *using a step size $\alpha_n = \Theta(\frac{1}{n^\theta})$ with $\theta \in \left(\frac{1}{2}, 1\right)$. For any*

$$\eta \in \begin{cases} \left(\max\{2-2\theta, \frac{\theta+2\beta-2}{2\beta-1}\}, 1\right) & : \beta \in (\frac{1}{2}, 1] \\ (2-2\theta, 1) & : \beta = \frac{1}{2} \end{cases}$$

it holds that

(i) $f(X_n) - f^* \in o\left(\frac{1}{n^{1-\eta}}\right)$, *almost surely, and*

(ii) $\mathbb{E}[f(X_n) - f^*] \in o\left(\frac{1}{n^{1-\eta}}\right)$.

Proof. Recall, in Section 2.2 we derived equation (2.9),

$$\begin{aligned} &\mathbb{E}[f(X_{n+1}) - f^* \mid \mathcal{F}_n] \\ &\leq \left(1 + \frac{LA\alpha_n^2}{2}\right)(f(X_n) - f^*) - \left(\alpha_n - \frac{BL\alpha_n^2}{2}\right)c^2(f(x) - f^*)^{2\beta} + \frac{LC\alpha_n^2}{2}, \end{aligned}$$

which will be the basis of the proof.
We treat again both cases for $\beta = \frac{1}{2}$ and $\beta \in (\frac{1}{2}, 1]$ separately:

$\underline{\beta = \frac{1}{2}}$: In this case, equation (2.9) results in the super-martingale inequality

$$\mathbb{E}[Y_{n+1} \mid \mathcal{F}_n] \leq \Big(1 + \frac{LA\alpha_n^2}{2} - \alpha_n c^2 + \frac{BLc^2\alpha_n^2}{2}\Big)Y_n + \frac{LC\alpha_n^2}{2},$$

with $Y_n = f(X_n) - f^*$. By the choice of α_n there exists $N > 0$ and a constant $\widetilde{c} > 0$ such that $\alpha_n c^2 - \frac{LA\alpha_n^2}{2} - \frac{BLc^2\alpha_n^2}{2} \geq \widetilde{c}\alpha_n$ for all $n \geq N$. Thus,

$$\mathbb{E}[Y_{n+1} \mid \mathcal{F}_n] \leq \Big(1 - \widetilde{c}\alpha_n\Big)Y_n + \frac{LC\alpha_n^2}{2},$$

for all $n \geq N$. Then, claim (i) follows by applying Lemma 6.1 with $c_1 = 0, c_2 = \widetilde{c}, c_3 = \frac{LC}{2}$ and $\beta = \frac{1}{2}$.
To prove claim (ii) we multiply $(n+1)^{1-\eta}$ on both sides and take the expectation. It follows that

$$\begin{aligned}
\mathbb{E}[(n+1)^{(1-\eta)}Y_{n+1}] &\leq (1 - \widetilde{c}\alpha_n)(n+1)^{1-\eta}\mathbb{E}[Y_n] + \frac{LC}{2}(n+1)^{1-\eta}\alpha_n^2 \\
&\leq (1 - \widetilde{c}\alpha_n)(n^{1-\eta} + (1-\eta)n^{-\eta})\mathbb{E}[Y_n] + \frac{LC}{2}(n+1)^{1-\eta}\alpha_n^2 \\
&= \left(1 - \widetilde{c}\alpha_n + \frac{1-\eta}{n} - \frac{\widetilde{c}(1-\eta)\alpha_n}{n}\right)n^{1-\eta}\mathbb{E}[Y_n] + \frac{LC}{2}(n+1)^{1-\eta}\alpha_n^2.
\end{aligned}$$

As $\theta_n \in \Theta(\frac{1}{n^\theta})$ we obtain that $\widetilde{c}\alpha_n$ is the dominating term. Hence, there exists a constant $\widetilde{c}_1 > 0$ and $\widetilde{N} > N$ such that $\widetilde{c}\alpha_n - \frac{1-\eta}{n} + \frac{\widetilde{c}(1-\eta)\alpha_n}{n} \geq \widetilde{c}_1\alpha_n$ for all $n \geq \widetilde{N}$. Thus, for all $n \geq \widetilde{N}$

$$\mathbb{E}[(n+1)^{(1-\eta)}Y_{n+1}] \leq (1 - \widetilde{c}_1\alpha_n)n^{1-\eta}\mathbb{E}[Y_n] + \frac{LC}{2}(n+1)^{1-\eta}\alpha_n^2.$$

We apply Lemma 2.16 with $w_n = n^{1-\eta}\mathbb{E}[Y_n]$, $a_n = \widetilde{c}_1\alpha_n$ and $b_n = (n+1)^{1-\eta}\alpha_n^2$ and obtain that $n^{1-\eta}\mathbb{E}[Y_n] \to 0$ for $n \to \infty$ which yields claim (ii). Note that $\sum_n b_n < \infty$ as $1 - \eta < 2\theta - 1$ for $\eta \in (2 - 2\theta, 1)$.
$\underline{\beta \in (\frac{1}{2}, 1]}$: In this case, equation (2.9) results in the super-martingale inequality

$$\mathbb{E}[Y_{n+1} \mid \mathcal{F}_n] \leq \Big(1 + \frac{LA\alpha_n^2}{2}\Big)Y_n - \Big(\alpha_n - \frac{BL\alpha_n^2}{2}\Big)c^2 Y_n^{2\beta} + \frac{LC\alpha_n^2}{2},$$

with $Y_n = f(X_n) - f^*$. By the choice of α_n there exists $c_2 > 0$ and $N_1 > 0$ such that $c^2\alpha_n - \frac{BLc^2\alpha_n^2}{2} \geq c_2\alpha_n$ for all $n \geq N_1$,

$$\mathbb{E}[Y_{n+1} \mid \mathcal{F}_n] \leq \Big(1 + \frac{LA\alpha_n^2}{2}\Big)Y_n - c_2\alpha_n Y_n^{2\beta} + \frac{LC\alpha_n^2}{2}\,.$$

We deduce claim (i) from Lemma 6.1 with $c_1 = \frac{LA}{2}, c_2 = c_2, c_3 = \frac{LC}{2}$ and $\beta \in (\frac{1}{2}, 1]$.
For claim (ii) we firstly proceed as in the proof of Lemma 6.1. Therefore, one can choose the auxiliary parameter $1 < q \leq \frac{1}{\theta}$ and find constants $c_4, c_3, \widetilde{N}_1 > 0$ such that for all $n \geq \widetilde{N}_1$ by equation (6.6) we have

$$\mathbb{E}[(n+1)^{1-\eta}Y_{n+1} \mid \mathcal{F}_n] \leq n^{1-\eta}Y_n - c_4\frac{1}{n^{q\theta}}n^{1-\eta}Y_n + c_3(n+1)^{1-\eta}(\alpha_n^{\frac{2\beta q-1}{2\beta-1}} + \alpha_n^2)\,.$$

Next, we take the expectation to obtain

$$\mathbb{E}[(n+1)^{1-\eta}Y_{n+1}] \leq (1 - c_4\frac{1}{n^{q\theta}})\mathbb{E}[n^{1-\eta}Y_n] + c_3(n+1)^{1-\eta}(\alpha_n^{\frac{2\beta q-1}{2\beta-1}} + \alpha_n^2)$$

for all $n \geq \widetilde{N}_1$, implying that $\omega_n = \mathbb{E}[n^{1-\eta}Y_n] \to 0$ as $n \to \infty$ by Lemma 2.16. Note that we have chosen θ, η and q as in Lemma 6.1, such that $\sum_n \frac{1}{n^{q\theta}} = \infty$, $\sum_n (n+1)^{1-\eta}\alpha_n^{\frac{2\beta q-1}{2\beta-1}} < \infty$, and $\sum_n (n+1)^{1-\eta}\alpha_n^2 < \infty$ (see equation (6.7), equation (6.8) and equation (6.9)). Therefore, the assumptions of Lemma 2.16 are met. ∎

It is natural to ask which θ leads to the best convergence rate. First, it is important to notice, that is not possible for η to approach 0 for fixed $\beta \in (\frac{1}{2}, 1]$. Optimising for η yields an optimal choice $\theta = \frac{2\beta}{4\beta-1}$ to achieve the best possible rate of convergence. This specific choice yields a lower bound of the interval given by $2 - 2\theta = \frac{\theta+2\beta-2}{2\beta-1} = 1 - \frac{1}{4\beta-1}$ and therefore an almost sure convergence of the form $o(\frac{1}{n^p})$ where p is arbitrarily close to $\frac{1}{4\beta-1}$ (see also table 6.1).
Roughly speaking, our result guarantees a faster convergence rate for "stronger" gradient domination properties (i.e. for smaller β). Indeed, as $2 - 2\theta > \frac{\theta+2\beta-2}{2\beta-1}$ for β sufficiently close to $\frac{1}{2}$ our result is consistent to the one presented in [LY22, Thm. 1] by replacing the μ-strongly convex assumption with the strong gradient domination property with $\beta = \frac{1}{2}$.
Note that the global gradient domination property using $\beta = \frac{1}{2}$ mirrors the strongly convex case and is also covered in [LY22, Theorem 1] by replacing the μ-strongly convex assumption by the weaker global gradient domination property with $\beta = \frac{1}{2}$. We emphasize that the rate we obtain is arbitrarily close to the one obtained in [FBD21; Fat+22] in expectation and is tight according to [Fat+22, Prop. 2]. With respect to almost sure convergence, we get arbitrarily close to the rates obtained in [SGD21] in the convex setting.

6.3.2 *Convergence Rates for SHB*

The following section deals with the stochastic heavy ball (SHB) scheme, where a momentum term is added to the classical optimization algorithm. Recall the noisy gradient evaluations in equation (2.7), where we still assume that the stochastic first order oracle is accessed through the evaluation of ζ_{n+1} which is a copy of ζ independent from the current state X_n.
The iterative scheme of stochastic heavy ball (SHB) is defined by

$$X_{n+1} = X_n - \alpha_n V_{n+1}(X_n) + \nu(X_n - X_{n-1}), \tag{SHB}$$

with initial $\mathbb{R}^d$-valued random vector X_0. The additional summand is called the momentum term with momentum parameter $\nu \in [0, 1)$. Similar arguments as for SGD can be used to derive almost sure convergence rates for SHB under global gradient domination:

THEOREM 6.3. *Suppose Assumption 2.1 and Assumption 2.12 are fulfilled and let f satisfy the global gradient domination property from Definition 2.7 with $\beta \in [\frac{1}{2}, 1]$. Denote by $(X_n)_{n\in\mathbb{N}}$ the sequence generated by equation* (SHB) *using a step size $\alpha_n = \Theta(\frac{1}{n^\theta})$ for $\theta \in (\frac{1}{2}, 1)$. For any*

$$\eta \in \begin{cases} \left(\max\{2 - 2\theta, \frac{\theta+2\beta-2}{2\beta-1}\}, 1\right) & : \beta \in (\frac{1}{2}, 1] \\ (2 - 2\theta, 1) & : \beta = \frac{1}{2} \end{cases}$$

it holds that

(i) $f(X_n) - f^* \in o\Big(\frac{1}{n^{1-\eta}}\Big)$, *almost surely, and*

(ii) $\mathbb{E}[f(X_n) - f^*] \in o\Big(\frac{1}{n^{1-\eta}}\Big)$.

To proof this result, we rewrite the iteration scheme in equation (SHB) by using the definitions

$$Z_n := X_n + \frac{\nu}{1-\nu} W_n, \quad \text{and} \quad W_n := X_n - X_{n-1}. \tag{6.10}$$

We derive the following iterative evolution from SHB

$$W_{n+1} = \nu W_n - \alpha_n V(X_n) \tag{6.11}$$

$$Z_{n+1} = Z_n - \frac{\alpha_n}{1-\nu} V(X_n). \tag{6.12}$$

We will utilize these auxiliary variables in the proof.

Proof. The proof begins as in the proof of [LY22, Thm. 2]. Using only L-smoothness and assumption (ABC), they show that for any $c_3 \in (0, \frac{1}{1-\nu})$, $\lambda \in (\nu, 1)$ there exist constants $c_1, c_2, c_4 > 0$ such that choosing step size $\alpha_n \sim \frac{1}{n^\theta}$, for some $\theta \in (\frac{1}{2}, 1)$ results in [LY22, eq. (21)]

$$\begin{aligned} &\mathbb{E}[f(Z_{n+1}) - f^* + \|W_{n+1}\|^2 \mid \mathcal{F}_n] \\ &\leq (1 + c_1\alpha_n^2)(f(Z_n) - f^*) + (\lambda + c_2\alpha_n^2)\|W_n\|^2 - c_3\alpha_n\|\nabla f(Z_n)\|^2 + c_4\alpha_n^2 \end{aligned} \tag{6.13}$$

for all $n \geq N$ and some $N > 0$ sufficiently large. Next, we apply the global gradient domination property for any $\beta \in [\frac{1}{2}, 1]$ to derive

$$\begin{aligned} &\mathbb{E}[f(Z_{n+1}) - f^* + \|W_{n+1}\|^2 \mid \mathcal{F}_n] \\ &\leq (1 + c_1\alpha_n^2)(f(Z_n) - f^*) - cc_3\alpha_n(f(Z_n) - f^*)^{2\beta} + (\lambda + c_2\alpha_n^2)\|W_n\|^2 + c_4\alpha_n^2. \end{aligned} \tag{6.14}$$

For the remaining proof, we denote $Q_n := f(Z_n) - f^*$. Similar as before, we treat both cases for $\beta = \frac{1}{2}$ and $\beta \in (\frac{1}{2}, 1]$ separately:
$\beta = \frac{1}{2}$: Instead of μ-strong convexity we use the gradient domination inequality $\|\nabla f(x)\|^2 \geq c(f^* - f(x))$, as the same inequality is implied by strong convexity using $c = \mu$. Then, Claim (i) follows using the same proof as [LY22, Thm. 2b)]. Note that the inequality

$$\frac{1}{2L}\|\nabla f(x)\|^2 \leq f(x) - f^*, \tag{6.15}$$

used in the last step only requires the L-smoothness assumption [Nes13, Sec. 1.2.3].
For Claim (ii) we consider equation (6.14) which simplifies for $\beta = \frac{1}{2}$ to

$$\mathbb{E}[Q_{n+1} + \|W_{n+1}\|^2 \mid \mathcal{F}_n] \leq (1 + c_1\alpha_n^2 - cc_3\alpha_n)Q_n + (\lambda + c_2\alpha_n^2)\|W_n\|^2 + c_4\alpha_n^2.$$

By the choice of α_n there exists $N > 0$ and $\widetilde{c}_1, \widetilde{c}_2 > 0$, such that $cc_3\alpha_n - c_1\alpha_n^2 \geq \widetilde{c}_1\alpha_n$ and $\lambda + c_2\alpha_n^2 \leq \widetilde{c}_2\alpha_n$ for all $n \geq N$. Hence, for $n \geq N$

$$\begin{aligned} \mathbb{E}[Q_{n+1} + \|W_{n+1}\|^2 \mid \mathcal{F}_n] &\leq (1 - \widetilde{c}_1\alpha_n)Q_n + (1 - \widetilde{c}_2\alpha_n)\|W_n\|^2 + c_4\alpha_n^2 \\ &\leq (1 - \min\{\widetilde{c}_1, \widetilde{c}_2\})\Big(Q_n + \|W_n\|^2\Big) + c_4\alpha_n^2. \end{aligned}$$

Let $c_5 = \min\{\widetilde{c}_1, \widetilde{c}_2\}$, multiply by $(n+1)^{1-\eta}$ on both sides and use equation (6.1) to obtain for $n \geq N$

$$\begin{aligned}
&\mathbb{E}[(n+1)^{1-\eta}\Big(Q_{n+1} + \|W_{n+1}\|^2\Big) \mid \mathcal{F}_n] \\
&\leq (n+1)^{1-\eta}(1 - c_5\alpha_n)\Big(Q_n + \|W_n\|^2\Big) + c_4\alpha_n^2(n+1)^{1-\eta} \\
&\leq (n^{1-\eta} + (1-\eta)n^{-\eta})(1-c_5)\Big(Q_n + \|W_n\|^2\Big) + c_4\alpha_n^2(n+1)^{1-\eta} \\
&= \left(1 - c_5\alpha_n + \frac{1-\eta}{n} - \frac{c_5(1-\eta)\alpha_n}{n}\right)n^{1-\eta}\Big(Q_n + \|W_n\|^2\Big) + c_4\alpha_n^2(n+1)^{1-\eta}.
\end{aligned}$$

Taking expectation and using that there exists $\widetilde{c}_5 > 0$ and $\widetilde{N} > N$ such that $c_5\alpha_n - \frac{1-\eta}{n} + \frac{c_5(1-\eta)\alpha_n}{n} \geq \widetilde{c}_5\alpha_n$, we have for all $n \geq \widetilde{N}$

$$\begin{aligned}
&\mathbb{E}[(n+1)^{1-\eta}\Big(Q_{n+1} + \|W_{n+1}\|^2\Big)] \\
&\leq (1 - \widetilde{c}_5\alpha_n)\mathbb{E}\Big[n^{1-\eta}\Big(Q_n + \|W_n\|^2\Big)\Big] + c_4\alpha_n^2(n+1)^{1-\eta}.
\end{aligned}$$

Note that $\sum_n \alpha_n^2(n+1)^{1-\eta} < \infty$ because $\eta \in (2-2\theta, 1)$ implies $1 - \eta < 2\theta - 1$. We can apply Lemma 2.16 which yields that $\mathbb{E}\big[n^{1-\eta}\big(Q_n + \|W_n\|^2\big)\big] \to 0$. Hence, $\mathbb{E}[\big(Q_n + \|W_n\|^2\big)] \in o(\frac{1}{n^{1-\eta}})$. To finish the proof, one can derive

$$f(X_n) - f^* \leq Q_n + \frac{1}{2}\|\nabla f(X_n)\|^2 + \frac{\nu^2 + L\nu^2}{1(1-\nu)^2}\|W_n\|^2 \tag{6.16}$$

see [LY22, eq. (19)] for more details. Using inequality (6.15), we get almost surely

$$\left(1 - \frac{1}{4L}\right)f(X_n) - f^* \leq Q_n + \frac{\nu^2 + L\nu^2}{1(1-\nu)^2}\|W_n\|^2. \tag{6.17}$$

implying that $\mathbb{E}[(f(X_n) - f^*)] \in o(\frac{1}{n^{1-\eta}})$ which proves Claim (ii).
$\beta \in (\frac{1}{2}, 1]$: For Claim (i), note that in equation (6.14) $\lambda < 1$, such that

$$\begin{aligned}
&\mathbb{E}[Q_{n+1} + \|W_{n+1}\|^2 \mid \mathcal{F}_n] \\
&\leq (1 + c_1\alpha_n^2)Q_n + (1 + c_2\alpha_n^2)\|W_n\|^2 + cc_3\alpha_n Q_n^{2\beta} + c_4\alpha_n^2 \\
&\leq (1 + \max\{c_1, c_2\}\alpha_n^2)(Q_n + \|W_n\|^2) + cc_3\alpha_n(Q_n + \|W_n\|^2)^{2\beta} + c_4\alpha_n^2.
\end{aligned}$$

By Lemma 6.1 we obtain that $Q_n + \|W_n\|^2 = f(Z_n) - f^* + \|W_n\|^2 \in o\Big(\frac{1}{n^{1-\eta}}\Big)$ for all $\eta \in \Big(\max\{2-2\theta, \frac{\theta+2\beta-2}{2\beta-1}\}, 1\Big)$. We apply the inequality in equation (6.17) to conclude that also $f(X_n) - f^* \in o\Big(\frac{1}{n^{1-\eta}}\Big)$ for all $\eta \in \Big(\max\{2-2\theta, \frac{\theta+2\beta-2}{2\beta-1}\}, 1\Big)$. This proves Claim (i).
For Claim (ii), we again use the q-trick from Lemma 6.1 in equation (6.14). For $1 < q < \frac{1}{\theta} < 2$ we have that

$$\begin{aligned}
&\mathbb{E}[Q_{n+1} + \|W_{n+1}\|^2 \mid \mathcal{F}_n] \\
&\leq (1 + c_1\alpha_n^2 - cc_3\alpha_n^q)Q_n + cc_3\alpha_n\Big(\alpha_n^{q-1}Q_n - Q_n^{2\beta}\Big) + (\lambda + c_2\alpha_n^2)\|W_n\|^2 + c_4\alpha_n^2.
\end{aligned}$$

Now with equation (6.4) in Lemma 6.1 there exists $\widetilde{c}_3 \geq 0$ such that

$$\mathbb{E}[Q_{n+1} + \|W_{n+1}\|^2 \mid \mathcal{F}_n] \leq (1 + c_1\alpha_n^2 - cc_3\alpha_n^q)Q_n + \widetilde{c}_3\alpha_n^{\frac{2\beta q-1}{2\beta-1}} + (\lambda + c_2\alpha_n^2)\|W_n\|^2 + c_4\alpha_n^2.$$

By the choice of α_n there exists $\widetilde{c}_1 > 0$ and $N > 0$ such that $c_1\alpha_n^2 - cc_3\alpha_n^q \geq \widetilde{c}_1\alpha_n^q$ and $\lambda + c_2\alpha_n^2 \leq \widetilde{c}_1\alpha_n^q$ for all $n \geq N$. Thus, for all $n \geq N$,

$$\mathbb{E}[Q_{n+1} + \|W_{n+1}\|^2 \mid \mathcal{F}_n] \leq (1 - \widetilde{c}_1\alpha_n^q)(Q_n + \|W_n\|^2) + \max\{\widetilde{c}_3, c_4\}\left(\alpha_n^{\frac{2\beta q-1}{2\beta-1}} + \alpha_n^2\right).$$

For $\max\{\widetilde{c}_3, c_4\} =: \widetilde{c}_2$, we multiply on both sides with $(n+1)^{1-\eta}$ and take the expectation to obtain for $n \geq N$

$$\begin{aligned}
&\mathbb{E}[(n+1)^{1-\eta}\left(Q_{n+1} + \|W_{n+1}\|^2\right)] \\
&\leq (n+1)^{1-\eta}(1 - \widetilde{c}_1\alpha_n^q)\mathbb{E}[(Q_n + \|W_n\|^2)] + \widetilde{c}_2(n+1)^{1-\eta}\left(\alpha_n^{\frac{2\beta q-1}{2\beta-1}} + \alpha_n^2\right) \\
&\leq (n^{1-\eta} + (1-\eta)n^{-\eta})(1 - \widetilde{c}_1\alpha_n^q)\mathbb{E}[(Q_n + \|W_n\|^2)] + \widetilde{c}_2(n+1)^{1-\eta}\left(\alpha_n^{\frac{2\beta q-1}{2\beta-1}} + \alpha_n^2\right) \\
&= \left(1 - \widetilde{c}_1\alpha_n^q + \frac{1-\eta}{n} - \frac{\widetilde{c}_1(1-\eta)\alpha_n^q}{n}\right)\mathbb{E}[n^{1-\eta}(Q_n + \|W_n\|^2)] \\
&\quad + \widetilde{c}_2(n+1)^{1-\eta}\left(\alpha_n^{\frac{2\beta q-1}{2\beta-1}} + \alpha_n^2\right).
\end{aligned}$$

Next, there exists $\widetilde{N} > N$ and $\widetilde{c}_5 > 0$ such that for all $n \geq \widetilde{N}$

$$\begin{aligned}
&\mathbb{E}[(n+1)^{1-\eta}\left(Q_{n+1} + \|W_{n+1}\|^2\right)] \\
&\leq \left(1 - \widetilde{c}_5\alpha_n^q\right)\mathbb{E}[n^{1-\eta}(Q_n + \|W_n\|^2)] + \widetilde{c}_2(n+1)^{1-\eta}\left(\alpha_n^{\frac{2\beta q-1}{2\beta-1}} + \alpha_n^2\right).
\end{aligned}$$

From the proof of Lemma 6.1, we choose the auxiliary parameter q such that $\sum_n (n+1)^{1-\eta}\left(\alpha_n^{\frac{2\beta q-1}{2\beta-1}} + \alpha_n^2\right) < \infty$ (see equation (6.8) and equation (6.9)). By applying again Lemma 2.16 we obtain $\mathbb{E}[n^{1-\eta}(Q_n + \|W_n\|^2)] \to 0$, i.e. $\mathbb{E}[Q_n + \|W_n\|^2] \in o(\frac{1}{n^{1-\eta}})$. Finally, Claim (ii) follows again by equation (6.17). ∎

To the best of our knowledge, our result gives the first convergence proof of SHB to global optima under weak gradient domination, with rates for almost sure convergence and convergence of expectations. The resulting convergence rate using the optimized step size are summarized in Table 6.1. In the strong gradient domination setting our rate in expectation gets arbitrarily close to the $O(\frac{1}{n})$ convergence obtained in Liang, Liu, and Xu [LLX23]. It is noteworthy that the utilization of SHB in our analysis does not yield a superior convergence rate compared to SGD. This arises from the proof technique and aligns with the findings in Liu and Yuan [LY22] and Sebbouh, Gower, and Defazio [SGD21] where the authors similarly achieve no acceleration. In general, for deterministic settings acceleration of gradient methods can achieve improvements of convergence rates [WMW19]. Although in the special case of gradient domination with $\beta = \frac{1}{2}$, Yue, Fang, and Lin [YFL23] showed that HB as well as Nesterov cannot accelerate in the deterministic setting.

6.4 NUMERICAL EXPERIMENT - TOY EXAMPLE

We have implemented the same toy example similar to Fatkhullin et al. [Fat+22] to test our theoretical findings. In our implementation, we consider both SGD and SHB applied to the objective function $f_p(x) = |x|^p$, where $x \in \mathbb{R}$, for various choices of $p \geq 2$. It is straightforward to verify that f_p satisfies the global gradient domination with parameter $\beta(p) = \frac{p-1}{p}$. It is noteworthy that for $p = 2$, the f_p obviously satisfies the PL condition with $\beta = \frac{1}{2}$, whereas for increasing $p \to \infty$, we move towards $\beta(p) \to 1$. We have used the step size schedule $\Theta(n^{-\frac{2\beta(p)}{4\beta(p)-1}})$ discussed in Table 6.1 and observed the almost sure convergence rates $n^{-\frac{1}{4\beta(p)-1}}$ as suggested by Theorem 6.2 and Theorem 6.3. Note that our derived rates are arbitrarily close to the sharp upper bound known in expectation [Fat+22].

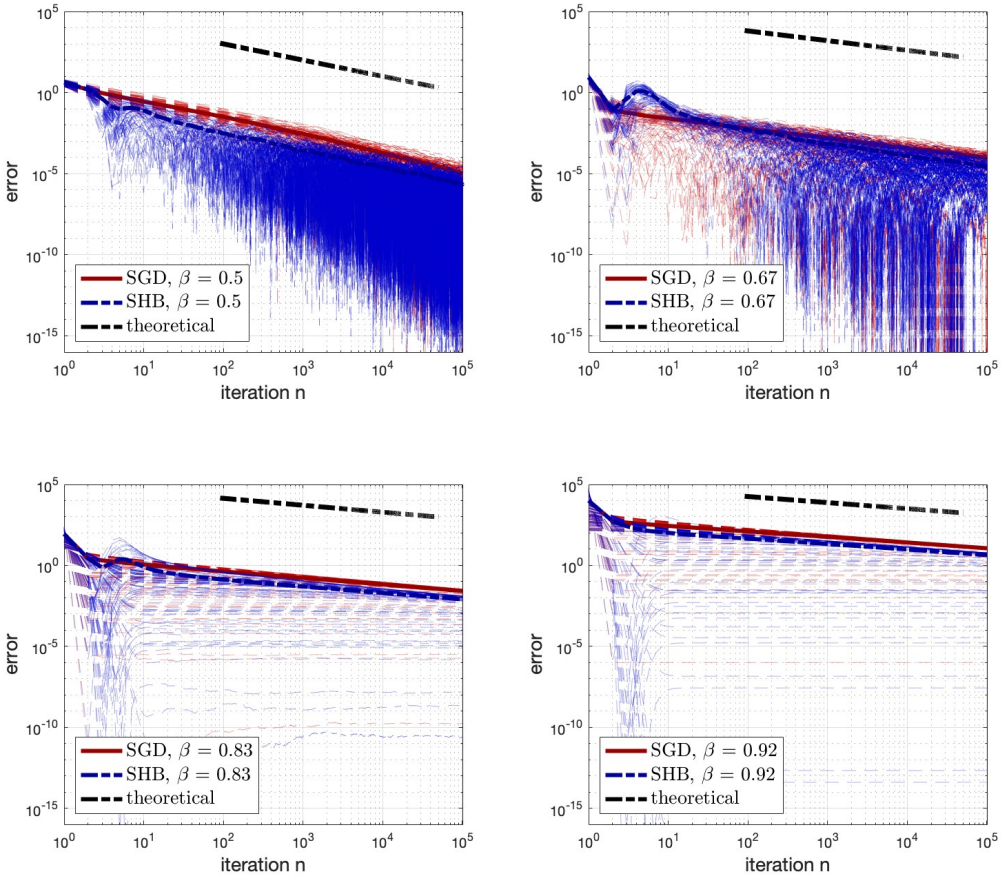


Figure 6.1: Pathwise error $(f_p(X_n))_{n=1,\dots,N}$ of SGD and SHB for various choices of β. For each setting we have simulated 100 runs of length , $N = 10^5$. The black dash-dotted line corresponds to the theoretical rate $n^{-\frac{1}{4\beta-1}}$.

Details of the implementation: Both algorithms have been implemented by hand using MATLAB. We have initialized both SGD and SHB with the initial state $X_1 \sim \frac{1}{2}\mathcal{U}([1.5, 2.5]) + \frac{1}{2}\mathcal{U}([-2.5, 1.5])$ to force initials which are not close to the actual minimum $x^* = 0$. The initial step sizes $\alpha_1(\beta)$ for both algorithms are chosen as

$$\alpha_1(0.5) = 0.2,\ \alpha_1(0.67) = 0.13,\ \alpha_1(0.83) = 0.004,\ \alpha_1(0.92) = 10^{-6}$$

through which we counteract the decreasing smoothness for $\beta \to 1$. The momentum parameter for SHB is fixed for all β as $\nu = 0.5$. The exact gradients ∇f_p are perturbed by independent additive noise following a standard normal distribution $\mathcal{N}(0, 1)$.

6.5 Almost Sure Convergence under Local Gradient Domination

6.5.1 *Local Smoothness and Local Gradient Domination*

From now on we will relax the assumption of smoothness and gradient domination to hold only locally.

Assumption 6.4. *The objective function $f : \mathbb{R}^d \to \mathbb{R}$ is differentiable and the gradient ∇f is locally L-Lipschitz continuous, i.e. for all $R > 0$ there exists $L(R) > 0$ such that $\|\nabla f(x) - \nabla f(y)\| \leq L(R)\|x - y\|$ for all $x, y \in \mathbb{R}^d$ with $|x|, |y| \leq R$.*

We collect the following types of local gradient domination properties.

Definition 6.5. Let $f : \mathbb{R}^d \to \mathbb{R}$ be continuously differentiable with $f^* = \inf_{x\in\mathbb{R}^d} f(x) > -\infty$.

(i) Let $x^* \in \mathbb{R}^d$ be a stationary point, i.e. $\nabla f(x^*) = 0$. We say that f satisfies a local gradient domination property in x^* with parameter $\beta_{x^*} \in [\frac{1}{2}, 1]$ if there exist a radius $r_{x^*} > 0$ and a constant $c_{x^*} > 0$ such that

$$\|\nabla f(x)\| \geq c_{x^*}|f(x) - f(x^*)|^{\beta_{x^*}}$$

for all $x \in \mathcal{B}_{r_{x^*}}(x^*) = \{y \in \mathbb{R}^d : \|x^* - y\| \leq r_{x^*}\}$.

(ii) We say that f satisfies a local gradient domination property in f^* with parameter $\beta \in [\frac{1}{2}, 1]$ if there exist a radius $r > 0$ and a constant $c > 0$ such that

$$\|\nabla f(x)\| \geq c(f(x) - f^*)^{\beta}$$

for all $x \in \mathcal{B}_r^* = \{y \in \mathbb{R}^d : f(y) - f^* \leq r\}$.

Remark 6.6. Moreover, note that for the local gradient domination property in x^* the parameters r and c may depend on x^*. Furthermore, we emphasize that for the definition of the local gradient domination in f^* we do not require the existence of $x^* \in \arg\min_{x\in\mathbb{R}^d} f(x)$.

In fact, there are many works analyzing (stochastic) first order methods under the weaker Łojasiewicz condition [Loj65] formulated in (local) areas around stationary points x^* and exponents $\beta \in [1/2, 1]$ [Lee+16]; [Fat+22]; [SMS22]; [WMW19]. Lojasiewicz [Loj65] has demonstrated that all analytic functions satisfy the local gradient domination property, emphasizing the particular significance of the local case. Further, Liu, Zhu, and Belkin [LZB22] proved

that all overparametrized neural networks fulfill the local gradient domination property. See also Madden, Dall'Anese, and Becker [MDB21], Dereich and Kassing [DK21], and Frei and Gu [FG21] and references therein for the application of (strong) gradient domination to (deep) neural networks.
Notably, we have seen in previous chapters that the tabular softmax parametrization in RL leads to a parametrized value function that satisfies the so-called "non-uniform" gradient domination property. In Section 6.7, we will show how this non-uniform gradient domination implies local gradient domination for f^*. This renders the local analysis of stochastic gradient methods specifically applicable in RL. As mentioned earlier, since every analytic function already satisfies local gradient domination, we expect that the local analysis can encompass further parametrizations for the policies, such as neural networks.
Recently, there has been a lot of effort to derive local convergence guarantees for stochastic first order methods (see Section 6.1). Especially, Mertikopoulos et al. [Mer+20] showed that, subject to certain assumptions on the objective function f, the SGD scheme converges almost surely towards a local minimum. By assuming strong convexity within a neighbourhood $\mathcal{U}$ of a local minimum x^*, they further established a local convergence rate for $\|X_n - x^*\|^2$ conditioning on the event of remaining in this neighbourhood. Additionally, they proved that the SGD scheme remains within $\mathcal{U}$ with high probability once it achieved sufficient closeness. In this section, we want to generalize this analysis under the weaker local gradient domination property for different cases of β. In this section, we extend the analysis in [Mer+20] under local strong convexity to the weaker local gradient domination property for different cases of β.
The contributions and differences of our results under less restricted assumptions are the following:

- We show that SGD still remains in the gradient dominated region with high probability by only assuming local gradient domination instead of local strong convexity. Especially in the case of a local minimum x^* this is a challenging task, as we have to ensure that the SGD scheme (X_n) remains close to x^* without exploiting convexity. We can guarantee this whenever x^* is a local minimum in a connected compact set of local minima and obtain Theorem 6.7.

- Additionally to convergence in expectation, we prove almost sure convergence conditioned on the "good event".

- Due to the weaker gradient domination assumption, one cannot expect the convergence of X_n to x^*, instead we focus on convergence of $f(X_n)$ to $f(x^*)$. Liu and Zhou [LZ23] delve into the rationale behind considering this as a more robust metric.

We consider the two cases of local gradient domination separately.

6.5.2 *Local gradient domination in stationary points*

THEOREM 6.7. *Fix some tolerance level $\delta > 0$ and let $\mathcal{X}^* \subset \mathbb{R}^d$ be an isolated compact connected set of local minima with level $l = f(x^*)$ for all $x^* \in \mathcal{X}^*$. Suppose that f satisfy the local gradient domination property in each $x^* \in \mathcal{X}^*$, f is locally G-Lipschitz continuous and satisfies Assumption 6.4. Moreover, suppose Assumption 2.12 hold true. Denote by $(X_n)_{n\in\mathbb{N}}$ the sequence generated by equation*

(SGD) *using a step size* $\alpha_n = \Theta(\frac{1}{n^\theta})$ *for* $\theta \in (\frac{1}{2}, 1)$ *and suppose that* $\alpha_n \leq \alpha_1$ *for* α_1 *small enough (dependent on* δ*). Then, the following holds:*

(*i*) *There exist subsets* $\mathcal{U}$ *and* $\mathcal{U}_1$ *of* $\mathbb{R}^d$ *such that, if* $X_1 \in \mathcal{U}_1$ *the event* $\Omega_{\mathcal{U}} = \{X_n \in \mathcal{U} \text{ for all } n = 1, 2, \dots\}$ *has probability at least* $1 - \delta$.

Moreover, there exists $\beta \in [\frac{1}{2}, 1]$ *such that for any*

$$\eta \in \begin{cases} \left(\max\{2 - 2\theta, \frac{\theta+2\beta-2}{2\beta-1}\}, 1\right) & : \beta \in (\frac{1}{2}, 1] \\ (2 - 2\theta, 1) & : \beta = \frac{1}{2} \end{cases}$$

it holds that

$$(ii)\ \ |f(X_n) - l|\mathbf{1}_{\Omega_{\mathcal{U}}} \in o\left(\frac{1}{n^{1-\eta}}\right),\ a.s., \quad \text{and} \quad (iii)\ \ \mathbb{E}[|f(X_n) - l|\mathbf{1}_{\Omega_{\mathcal{U}}}] \in o\left(\frac{1}{n^{1-\eta}}\right).$$

To prove the theorem a few preliminary results are necessary. Therefore, suppose that the assumptions of Theorem 6.7 hold throughout this section.
In contrast to the global gradient domination analysis we may assume w.l.o.g. the uniform second moment bounds, i.e. $A = B = 0$, instead of the more general (ABC) condition. Choosing $A, B > 0$ would imply the bounded variance assumption of the gradient estimator. Note therefore, that the first term $A(f(x) - f(x^*))$ and the second term $B\|\nabla f(x)\|^2$ are both locally bounded by the local Lipschitz assumptions on f and ∇f.
Note that every isolated local minimum $\{x^*\}$ is a special case of an isolated compact connected set of local minima. In this case it holds that $\beta = \beta_{x^*}$. If $\mathcal{X}^*$ contains more then one point, we can unify the gradient domination property in a neighbourhood of $\mathcal{X}^*$ due to compactness. The set $\mathcal{X}^*$ has to be connected to assure that all local minima are on the same level l.
The outline of the proof is structured as follows:

- First, we unify the gradient domination property around the set of local minima $\mathcal{X}^*$ and obtain a radius r such that the unified gradient domination property is fulfilled in all open balls with radius r around $x^* \in \mathcal{X}^*$ (Lemma 6.8).

- Based on this we construct sets $\mathcal{U}$, $\mathcal{U}_1 \subset \mathbb{R}^d$ and the events $\Omega_n \in \Omega$ (see equation (6.18), equation (6.19) and equation (6.20)), such that $\Omega_{\mathcal{U}} = \bigcap_n \Omega_n$ occurs with high probability. To be precise, $\mathcal{U}_1$ and $\mathcal{U}$ are neighborhoods of $\mathcal{X}^*$ constructed such that the gradient domination property holds within this region, and when starting in $\mathcal{U}_1$ the gradient trajectory does remain in $\mathcal{U}$ for all gradient steps with high probability. Then, Ω_n describes the event that $X_k \in \mathcal{U}$ for all $k \leq n$.

- All following Lemmata before the proof of Theorem 6.7 are devoted to show that $\mathbb{P}(\Omega_n) \geq 1 - \delta$ for all $n \in \mathbb{N}$. This then proves Claim (i) of the Theorem. Claim (ii) and (iii) will be shown directly in the proof of Theorem 6.7 at the end of this subsection.

- In order to show $\mathbb{P}(\Omega_n) \geq 1 - \delta$ we construct set C_n and E_n defined in equation (6.21) and equation (6.26) such that $E_n \cap C_n \subset \Omega_{n+1}$ (Lemma 6.13) while Lemma 6.12 is used to prove this claim.

- The sets E_n are such that $f(X_n)$ remains close to f^*. We exploit the unified gradient domination property to construct the sets E_n (Lemma 6.11) and derive a recursive inequality in Lemma 6.13 c) to prove that this event occurs with high probability (Lemma 6.14).
- The sets C_n are such that X_{n+1} remains close to X_n and we exploit the finite variance assumption to show that these events occur with high probability (Lemma 6.15).

We denote by

$$\widetilde{\mathcal{B}}_r(x) = \{y \in \mathbb{R}^d \,:\, ||x-y|| < r\}$$

the open ball with radius $r > 0$ around $x \in \mathbb{R}^d$ and by

$$\mathcal{B}_r(x) = \{y \in \mathbb{R}^d \,:\, ||x-y|| \leq r\}$$

the closed ball with radius $r > 0$ around $x \in \mathbb{R}^d$.
In the following Lemma we unify the gradient domination property around the set of local minima $\mathcal{X}^* \subset \mathbb{R}^d$.

LEMMA 6.8. . *There exists $r > 0$, $\beta \in [\frac{1}{2}, 1]$ and $c > 0$, such that for all $x \in \bigcup_{x^* \in \mathcal{X}^*} \widetilde{\mathcal{B}}_r(x^*)$ it holds that*

$$f(x) > l \text{ for } x \notin \mathcal{X}^* \quad \text{and} \quad ||\nabla f(x)|| \geq c(f(x)-l)^\beta .$$

Proof. By the local gradient domination property, for every $x^* \in \mathcal{X}^*$ there exist $r_{x^*} > 0$, $\beta_{x^*} \in [\frac{1}{2}, 1]$ and $c_{x^*} > 0$ such that

$$||\nabla f(x)|| \geq c_{x^*}|f(x)-l|^{\beta_{x^*}}, \quad \forall x \in \mathcal{B}_{r_{x^*}}(x^*).$$

Moreover, w.l.o.g we can assume that $f(x) > l$ for all $x \in \mathcal{B}_{r_{x^*}}(x^*) \setminus \mathcal{X}^*$, as $\mathcal{X}^*$ is an isolated compact connected set of local minima (otherwise choose r_{x^*} small enough).
By the compactness of $\mathcal{X}^*$ we can find a finite subset $\mathcal{Y}^* \subset \mathcal{X}^*$, such that

$$\widetilde{\mathcal{U}} := \bigcup_{y^* \in \mathcal{Y}^*} \widetilde{\mathcal{B}}_{r_{y^*}}(y^*) \supset \mathcal{X}^*.$$

Then, we define $\beta = \max_{y^* \in \mathcal{Y}^*} \beta_{y^*}$ and $c = \min_{y^* \in \mathcal{Y}^*} c_{y^*}$. For any $x \in \widetilde{\mathcal{U}}$ there exits $y^* \in \mathcal{Y}^*$ such that

$$||\nabla f(x)|| \geq c_{y^*}(f(x)-l)^{\beta_{y^*}} \geq c(f(x)-l)^\beta.$$

Thus, there exists an open neighbourhood $\widetilde{\mathcal{U}}$ of $\mathcal{X}^*$ and $\beta \in [\frac{1}{2}, 1]$, $c > 0$, such that for all $x \in \widetilde{\mathcal{U}}$ it holds that

$$f(x) > l \text{ for } x \notin \mathcal{X}^* \quad \text{and} \quad ||\nabla f(x)|| \geq c(f(x)-l)^\beta.$$

As $\widetilde{\mathcal{U}}$ is open by definition and $\mathcal{X}^* \subset \widetilde{\mathcal{U}}$, we can find a radius $r > 0$, such that $\bigcup_{x^* \in \mathcal{X}^*} \widetilde{\mathcal{B}}_r(x^*) \subset \widetilde{\mathcal{U}}$. This proves the claim. ∎

Remark 6.9. It is noteworthy that the unified gradient domination property obtained in the previous Lemma does not require an absolute value, as $f(x) \geq l$ for all $x \in \bigcup_{x^* \in \mathcal{X}^*} \widetilde{\mathcal{B}}_r(x^*)$. This is crucial to obtain the recursive inequalities in Lemma 6.11 and we will exploit this also in the proof of Theorem 6.7 to obtain the convergence rates.

In the following let $\mathbf{r} > 0$, $c > 0$ and $\beta \in [\frac{1}{2}, 1]$ chosen as in the previous Lemma, such that the unified gradient domination property holds for all $x \in \bigcup_{x^* \in \mathcal{X}^*} \widetilde{\mathcal{B}}_{\mathbf{r}}(x^*)$. Further define

$$s = \inf\left\{ f(x) - l \,:\, x \in \bigcup_{x^* \in \mathcal{X}^*} \mathcal{B}_{\frac{3\mathbf{r}}{4}}(x^*) \setminus \bigcup_{x^* \in \mathcal{X}^*} \widetilde{\mathcal{B}}_{\frac{\mathbf{r}}{2}}(x^*) \right\}.$$

LEMMA 6.10. *It holds that $s > 0$.*

Proof. If $s = 0$, then there exists a sequence $(x_n) \in \bigcup_{x^* \in \mathcal{X}^*} \mathcal{B}_{\frac{3\mathbf{r}}{4}}(x^*) \setminus \bigcup_{x^* \in \mathcal{X}^*} \widetilde{\mathcal{B}}_{\frac{\mathbf{r}}{2}}(x^*)$ with $f(x_n) \to l$ for $n \to \infty$. By definition of the set and compactness (boundedness) of $\mathcal{X}^*$, the sequence x_n is bounded:

$$||x_n|| \leq \frac{3\mathbf{r}}{4} + \sup_{x^* \in \mathcal{X}^*} ||x^*|| < \infty.$$

Hence, there is a convergent sub-sequence (x_{n_k}) with $x_{n_k} \to x$ for $k \to \infty$ and by continuity of f it holds that $f(x) = l$. Further, it holds for all $x^* \in \mathcal{X}^*$ that $||x_n - x^*|| \geq \frac{\mathbf{r}}{2}$ for all $n \in \mathbb{N}$ such that $\inf_{x^* \in \mathcal{X}^*} ||x - x^*|| \geq \frac{\mathbf{r}}{2}$.

On the other hand, by construction we have that $x \in \overline{\bigcup_{x^* \in \mathcal{X}^*} \mathcal{B}_{\frac{3\mathbf{r}}{4}}(x^*) \setminus \bigcup_{x^* \in \mathcal{X}^*} \widetilde{\mathcal{B}}_{\frac{\mathbf{r}}{2}}(x^*)} \subset \overline{\bigcup_{x^* \in \mathcal{X}^*} \mathcal{B}_{\frac{3\mathbf{r}}{4}}(x^*)} \subset \bigcup_{x^* \in \mathcal{X}^*} \widetilde{\mathcal{B}}_{\mathbf{r}}(x^*)$. And as $f(y) > l$ for all $y \in \widetilde{\mathcal{B}}_{\mathbf{r}}(x^*) \setminus \mathcal{X}^*$ we deduce from $f(x) = l$ that $x \in \mathcal{X}^*$. This is a contradiction to $\inf_{x^* \in \mathcal{X}^*} ||x - x^*|| \geq \frac{\mathbf{r}}{2}$. ∎

We choose $\epsilon > 0$, such that $2\epsilon + \sqrt{\epsilon} < s$. We define the sets

$$\mathcal{U}_1 = \{x \in \mathbb{R}^d \,:\, \inf_{x^* \in \mathcal{X}^*} ||x - x^*|| < \frac{\mathbf{r}}{2}, f(x) - l \leq \frac{\epsilon}{2}\} \tag{6.18}$$

$$\mathcal{U} = \{x \in \mathbb{R}^d \,:\, \inf_{x^* \in \mathcal{X}^*} ||x - x^*|| < \frac{\mathbf{r}}{2}\} \tag{6.19}$$

which are subsets of $\mathbb{R}^d$ and the decreasing sequence of events

$$\Omega_n = \{X_k \in \mathcal{U} \text{ for all } k \leq n\} \tag{6.20}$$

$$C_n = \{||X_{k+1} - X_k|| \leq \frac{\mathbf{r}}{4} \text{ for all } k \leq n\}, \tag{6.21}$$

and $C_0 = \Omega$, which are measurable sets in $(\Omega, \mathcal{F}, \mathbb{P})$.
In order to prove Theorem 6.7 we will show that Ω_n has probability at least $1 - \delta$ for all $n \in \mathbb{N}$. To do this, we construct another sequence of events $(\widehat{E}_n)$ with $\widehat{E}_n \subset \Omega_n$ which occurs with probability at least $1 - \delta$ for any $n \in \mathbb{N}$.
Therefore, we fix the notation $D_n := f(X_n) - l$ and recall that $\mathbf{1}_{\mathcal{A}}$ denoted the indicator function for a measurable set $\mathcal{A}$ in $(\Omega, \mathcal{F}, \mathbb{P})$, i.e. $\mathbf{1}_{\mathcal{A}}(\omega) = 1$ if $\omega \in \mathcal{A}$ and $\mathbf{1}_{\mathcal{A}}(\omega) = 0$ if $\omega \notin \mathcal{A}$. We prove the following (recursive) inequalities.

LEMMA 6.11. *If $\beta = \frac{1}{2}$, then it holds that*

$$\begin{aligned} D_{n+1}\mathbf{1}_{\Omega_n} &\leq (1 - \alpha_n c^2) D_n \mathbf{1}_{\Omega_n} + \alpha_n \xi_{n+1} \mathbf{1}_{\Omega_n} + \frac{L\alpha_n^2}{2} \mathbf{1}_{\Omega_n} \|V_{n+1}(X_n)\|^2, \\ &\leq D_1 \prod_{k=1}^{n} (1 - \alpha_k c^2) \mathbf{1}_{\Omega_n} + \sum_{k=1}^{n} \left(\prod_{j=k}^{n} (1 - \alpha_j c^2) \right) \alpha_k \xi_{k+1} \mathbf{1}_{\Omega_n} \\ &\quad + \frac{L}{2} \sum_{k=1}^{n} \alpha_k^2 \|V_{k+1}(X_k)\|^2 \mathbf{1}_{\Omega_n}. \end{aligned} \tag{6.22}$$

If $\beta \in (\frac{1}{2}, 1]$, for any $1 \leq q < 2$, it holds that

$$\begin{aligned} D_{n+1}\mathbf{1}_{\Omega_n} &\leq (1-\alpha_n^q c^2) D_n \mathbf{1}_{\Omega_n} + (2\beta)^{-\frac{1}{2\beta-1}}(1-\frac{1}{2\beta})c^2 \alpha_n^{\frac{2\beta q-1}{2\beta-1}} + \alpha_n \xi_{n+1}\mathbf{1}_{\Omega_n} + \frac{L\alpha_n^2}{2}\|V_{n+1}(X_n)\|^2 \mathbf{1}_{\Omega_n} \\ &\leq \widetilde{D}_1 \prod_{k=1}^{n}(1-\alpha_k^q c^2) + \widetilde{c}\sum_{k=1}^{n}\alpha_k^{\frac{2\beta q-1}{2\beta-1}} + \sum_{k=1}^{n}\left(\prod_{j=k}^{n}(1-\alpha_j^q c^2)\right)\alpha_k \xi_{k+1}\mathbf{1}_{\Omega_n} \\ &\quad + \frac{L}{2}\sum_{k=1}^{n}\alpha_k^2 \|V_{k+1}(X_k)\|^2 \mathbf{1}_{\Omega_n}, \end{aligned} \tag{6.23}$$

for $\widetilde{c} = (2\beta)^{-\frac{1}{2\beta-1}}(1-\frac{1}{2\beta})c^2$.

Proof. From L-smoothness we can deduce that

$$\begin{aligned} D_{n+1} &\leq D_n - \alpha_n \langle \nabla f(X_n), V_{n+1}(X_n)\rangle + \frac{L\alpha_n^2}{2}\|V_{n+1}(X_n)\|^2 \\ &= D_n - \alpha_n \|\nabla f(X_n)\|^2 - \alpha_n \langle \nabla f(X_n), Z(X_n, \zeta_{n+1})\rangle + \frac{L\alpha_n^2}{2}\|V_{n+1}(X_n)\|^2 \\ &= D_n - \alpha_n \|\nabla f(X_n)\|^2 + \alpha_n \xi_{n+1} + \frac{L\alpha_n^2}{2}\|V_{n+1}(X_n)\|^2 \end{aligned}$$

for $Z(X_n, \zeta_{n+1})$ from Assumption 2.12 and $\xi_{n+1} = -\langle \nabla f(X_n), Z(X_n, \zeta_{n+1})\rangle$.
We separate the two cases of β:
$\underline{\beta = \frac{1}{2}}$: Iterating this inequality and using $\mathbf{1}_{\Omega_{n+1}} \leq \mathbf{1}_{\Omega_n}$ it follows that

$$\begin{aligned} D_{n+1}\mathbf{1}_{\Omega_n} &\leq D_n \mathbf{1}_{\Omega_n} - \alpha_n \mathbf{1}_{\Omega_n}\|\nabla f(X_n)\|^2 + \alpha_n \mathbf{1}_{\Omega_n}\xi_{n+1} + \frac{L\alpha_n^2}{2}\mathbf{1}_{\Omega_n}\|V_{n+1}(X_n)\|^2 \\ &\leq D_n \mathbf{1}_{\Omega_n} - \alpha_n c^2 (f(X_n) - l)\mathbf{1}_{\Omega_n} + \alpha_n \mathbf{1}_{\Omega_n}\xi_{n+1} + \frac{L\alpha_n^2}{2}\mathbf{1}_{\Omega_n}\|V_{n+1}(X_n)\|^2 \\ &= (1-\alpha_n c^2) D_n \mathbf{1}_{\Omega_n} + \alpha_n \xi_{n+1}\mathbf{1}_{\Omega_n} + \frac{L\alpha_n^2}{2}\mathbf{1}_{\Omega_n}\|V_{n+1}(X_n)\|^2, \\ &\leq D_1 \prod_{k=1}^{n}(1-\alpha_k c^2) + \sum_{k=1}^{n}\left(\prod_{j=k}^{n}(1-\alpha_j c^2)\right)\alpha_k \xi_{k+1}\mathbf{1}_{\Omega_n} \\ &\quad + \frac{L}{2}\sum_{k=1}^{n}\left(\prod_{j=k}^{n}(1-\alpha_j c^2)\right)\alpha_k^2 \|V_{k+1}(X_k)\|^2 \mathbf{1}_{\Omega_n} \\ &\leq D_1 \prod_{k=1}^{n}(1-\alpha_k c^2) + \sum_{k=1}^{n}\left(\prod_{j=k}^{n}(1-\alpha_j c^2)\right)\alpha_k \xi_{k+1}\mathbf{1}_{\Omega_n} \\ &\quad + \frac{L}{2}\sum_{k=1}^{n}\alpha_k^2 \|V_{k+1}(X_k)\|^2 \mathbf{1}_{\Omega_n}, \end{aligned} \tag{6.24}$$

where we used that the unified gradient domination property holds for all X_k, $k \leq n$ on the event Ω_n.

$\beta \in (\frac{1}{2}, 1]$: Similarly, the unified gradient domination property yields the claimed inequality for any $1 \leq q < 2$:

$$\begin{aligned}
D_{n+1}\mathbf{1}_{\Omega_n} &\leq D_n\mathbf{1}_{\Omega_n} - \alpha_n\mathbf{1}_{\Omega_n}\|\nabla f(X_n)\|^2 + \alpha_n\mathbf{1}_{\Omega_n}\xi_{n+1} + \frac{L\alpha_n^2}{2}\mathbf{1}_{\Omega_n}\|V_{n+1}(X_n)\|^2 \\
&\leq D_n\mathbf{1}_{\Omega_n} - \alpha_n c^2(f(X_n) - l)^{2\beta}\mathbf{1}_{\Omega_n} + \alpha_n\mathbf{1}_{\Omega_n}\xi_{n+1} + \frac{L\alpha_n^2}{2}\mathbf{1}_{\Omega_n}\|V_{n+1}(X_n)\|^2 \\
&= D_n\mathbf{1}_{\Omega_n} - \alpha_n c^2 D_n^{2\beta}\mathbf{1}_{\Omega_n} + \alpha_n\xi_{n+1}\mathbf{1}_{\Omega_n} + \frac{L\alpha_n^2}{2}\mathbf{1}_{\Omega_n}\|V_{n+1}(X_n)\|^2, \\
&= (1 - \alpha_n^q c^2)D_n\mathbf{1}_{\Omega_n} + \alpha_n c^2(\alpha_n^{1-q}D_n - D_n^{2\beta})\mathbf{1}_{\Omega_n} + \alpha_n\xi_{n+1}\mathbf{1}_{\Omega_n} + \frac{L\alpha_n^2}{2}\|V_{n+1}(X_n)\|^2\mathbf{1}_{\Omega_n} \\
&\leq (1 - \alpha_n^q c^2)D_n\mathbf{1}_{\Omega_n} + (2\beta)^{-\frac{1}{2\beta-1}}(1 - \frac{1}{2\beta})c^2\alpha_n^{\frac{2\beta q-1}{2\beta-1}} + \alpha_n\xi_{n+1}\mathbf{1}_{\Omega_n} + \frac{L\alpha_n^2}{2}\|V_{n+1}(X_n)\|^2\mathbf{1}_{\Omega_n} \\
&\leq D_1\prod_{k=1}^{n}(1 - \alpha_k^q c^2)\mathbf{1}_{\Omega_n} + \widetilde{c}\sum_{k=1}^{n}\left(\prod_{j=k}^{n}(1 - \alpha_j^q c^2)\right)\alpha_k^{\frac{2\beta q-1}{2\beta-1}} \\
&\quad + \sum_{k=1}^{n}\left(\prod_{j=k}^{n}(1 - \alpha_j^q c^2)\right)\alpha_k\xi_{k+1}\mathbf{1}_{\Omega_n} + \frac{L}{2}\sum_{k=1}^{n}\left(\prod_{j=k}^{n}(1 - \alpha_j^q c^2)\right)\alpha_k^2\|V_{k+1}(X_k)\|^2\mathbf{1}_{\Omega_n} \\
&\leq D_1\prod_{k=1}^{n}(1 - \alpha_k^q c^2)\mathbf{1}_{\Omega_n} + \widetilde{c}\sum_{k=1}^{n}\alpha_k^{\frac{2\beta q-1}{2\beta-1}} + \sum_{k=1}^{n}\left(\prod_{j=k}^{n}(1 - \alpha_j^q c^2)\right)\alpha_k\xi_{k+1}\mathbf{1}_{\Omega_n} \\
&\quad + \frac{L}{2}\sum_{k=1}^{n}\alpha_k^2\|V_{k+1}(X_k)\|^2\mathbf{1}_{\Omega_n},
\end{aligned} \tag{6.25}$$

for $\widetilde{c} = (2\beta)^{-\frac{1}{2\beta-1}}(1 - \frac{1}{2\beta})c^2$ from the function trick eq. (6.4) which we applied in the forth inequality. We also used that the unified gradient domination property holds for all X_k, $k \leq n$ on the event Ω_n . ■

For $\beta \in (\frac{1}{2}, 1]$ we know from the proof of Lemma 6.1 that we can choose the auxiliary parameter q from the previous lemma in such a way, that $\sum_{n=1}^{\infty} n^{1-\eta}\alpha_n^{\frac{2\beta q-1}{2\beta-1}}$ is convergent for all $\eta \in (\max\{2 - 2\theta, \frac{\theta+2\beta-2}{2\beta-1}\}, 1)$ (Condition (iii) to apply Lemma 2.15). As $\eta < 1$, it follows that $\sum_{n=1}^{\infty}\alpha_n^{\frac{2\beta q-1}{2\beta-1}} < \infty$ holds true for all these choices of q. Now define

$$M_n = \sum_{k=1}^{n}\left(\prod_{j=k}^{n}(1 - \alpha_j c^2)\right)\alpha_k\xi_{k+1}\mathbf{1}_{\Omega_k}, \quad M_n^{(q)} = \sum_{k=1}^{n}\left(\prod_{j=k}^{n}(1 - \alpha_j^q c^2)\right)\alpha_k\xi_{k+1}\mathbf{1}_{\Omega_k}$$
$$\text{and} \quad S_n = \frac{L}{2}\sum_{k=1}^{n}\alpha_k^2\|V_{k+1}(X_k)\|^2\mathbf{1}_{\Omega_k}.$$

Then, (M_n) and $(M_n^{(q)})$ are $(\mathcal{F}_{n+1})$-martingales with zero mean and (S_n) is a $(\mathcal{F}_{n+1})$-sub-martingale by Assumption 2.12. Note that by the choice of α_n we have that $\sum_n \alpha_n^2 < \infty$ and hence $\mathbb{E}[S_n] < \infty$ for all $n \in \mathbb{N}$.

Next, define $R_n = M_n^2 + S_n$ and $R_n = (M_n^{(q)})^2 + S_n$ respectively (with some abuse of notation), for every $n \in \mathbb{N}$. Moreover, let

$$E_n = \{R_k < \epsilon \text{ for all } k \leq n\}. \tag{6.26}$$

which is an $\mathcal{F}_{n+1}$-measurable event on $(\Omega, \mathcal{F}, \mathbb{P})$. We define $R_0 = 0$ such that $E_0 = \Omega$.
Now let $\widehat{E}_n = E_n \cap C_n$, then we will first show, that $\widehat{E}_n$ fulfills the property $\widehat{E}_n \subset \Omega_{n+1}$ for all $n \in \mathbb{N}$ in Lemma 6.13 and then that $\widehat{E}_n$ occurs with probability at least $1 - \delta$ in Lemma 6.15.
To prove that $\widehat{E}_n \subset \Omega_{n+1}$ we need one more auxiliary result.

LEMMA 6.12. *Suppose $x, y \in \mathbb{R}^d$ such that*

1. $\inf_{x^* \in \mathcal{X}^*} ||x - x^*|| < \frac{\mathbf{r}}{2}$,
2. $f(y) - l < s$,
3. $||x - y|| \leq \frac{\mathbf{r}}{4}$.

Then it holds that $\inf_{x^* \in \mathcal{X}^*} ||y - x^*|| < \frac{\mathbf{r}}{2}$.

Proof. By triangle inequality we have that $\inf_{x^* \in \mathcal{X}^*} ||y - x^*|| \leq \frac{3\mathbf{r}}{4}$, i.e there exists $x^* \in \mathcal{X}^*$ such that $||y - x^*|| \leq \frac{3\mathbf{r}}{4}$. Suppose now, that $\inf_{x^* \in \mathcal{X}^*} ||y - x^*|| \geq \frac{\mathbf{r}}{2}$, this means that $y \in \bigcup_{x^* \in \mathcal{X}^*} \mathcal{B}_{\frac{3r}{4}}(x^*) \setminus \bigcup_{x^* \in \mathcal{X}^*} \widetilde{\mathcal{B}}_{\frac{\mathbf{r}}{2}}(x^*)$. By the definition of $s = \inf\left\{f(z) - l \ : \ z \in \bigcup_{x^* \in \mathcal{X}^*} \mathcal{B}_{\frac{3\mathbf{r}}{4}}(x^*) \setminus \bigcup_{x^* \in \mathcal{X}^*} \widetilde{\mathcal{B}}_{\frac{\mathbf{r}}{2}}(x^*)\right\}$ this contradicts the second assumption $f(y) - l < s$. ∎

We deduce the following relations on the constructed sets:

LEMMA 6.13. *For $\beta \in (\frac{1}{2}, 1]$ let $\alpha_n \leq \alpha_1$ be sufficiently small such that $\sum_{n=1}^{\infty} \alpha_n^{\frac{2\beta q - 1}{2\beta - 1}} < \frac{\epsilon}{2\widetilde{c}}$, and for $\beta = \frac{1}{2}$ let $\alpha_1 > 0$ be arbitrary. Furthermore, assume that the initial $X_1 \in \mathcal{U}_1$ almost surely. Then,*

a) $E_{n+1} \subset E_n$, $\widehat{E}_{n+1} \subset \widehat{E}_n$ and $\Omega_{n+1} \subset \Omega_n$

b) $\widehat{E}_n \subset \Omega_{n+1}$

c) Define the events $\widetilde{E}_n = E_{n-1} \setminus E_n = E_{n-1} \cup \{R_n \geq \epsilon\}$. Then, for $\widetilde{R}_n = R_n \mathbf{1}_{E_{n-1}}$, there exists a $\widetilde{C} > 0$ such that

$$\mathbb{E}[\widetilde{R}_n] \leq \mathbb{E}[\widetilde{R}_{n-1}] + \gamma_n^2[G^2C^2 + G^2 + C] - \epsilon \mathbb{P}(\widetilde{E}_{n-1}).$$

Proof. a) Follows by definition of the events.
b) Note that $\widehat{E}_0 = \Omega = \Omega_1$ because

$$X_1 \in \mathcal{U}_1 = \{x : \inf_{x^* \in \mathcal{X}^*} ||x - x^*|| < \frac{\mathbf{r}}{2}, f(x) - l \leq \frac{\epsilon}{2}\} \subset \{x : \inf_{x^* \in \mathcal{X}^*} ||x - x^*|| < \frac{\mathbf{r}}{2}\} = \Omega_1$$

almost surely. We prove the assertion by induction. Let $\omega \in \widehat{E}_n$. Since $\widehat{E}_n \subset \widehat{E}_{n-1} \subset \Omega_n$ by induction assumption, we have $\omega \in \Omega_n$ and thus $\omega \in \Omega_k$ for all $k \leq n$. We will apply Lemma 6.12 with $x = X_n(\omega)$ and $y = X_{n+1}(\omega)$. By definition it holds that $\omega \in \widehat{E}_n$ implies condition 3. and $\omega \in \Omega_n$ implies condition 1. of Lemma 6.12. It remains to show condition 2., then it follows that $\inf_{x^* \in \mathcal{X}^*} ||X_{n+1}(\omega) - x^*|| < \frac{\mathbf{r}}{2}$, i.e. $X_{n+1}(\omega) \in \mathcal{U}$ and by $\omega \in \Omega_n$ we deduce $\omega \in \Omega_{n+1}$.
To Prove condition 2. we separate both cases for β:

$\underline{\beta = \frac{1}{2}}$: The inequality eq. (6.22) and the induction hypothesis yield

$$\begin{aligned}
D_{n+1}(\omega) &= D_{n+1}(\omega)\mathbf{1}_{\Omega_n}(\omega) \\
&\leq D_1(\omega)\prod_{k=1}^{n}(1-\alpha_k c^2) + \sum_{k=1}^{n}\left(\prod_{j=k}^{n}(1-\alpha_j c^2)\right)\alpha_k\xi_{k+1}(\omega)\mathbf{1}_{\Omega_n}(\omega) \\
&\quad + \frac{L}{2}\sum_{k=1}^{n}\alpha_k^2\|V_{k+1}(X_k(\omega))\|^2\mathbf{1}_{\Omega_n}(\omega) \\
&= D_1(\omega)\prod_{k=1}^{n}(1-\alpha_k c^2) + \sum_{k=1}^{n}\left(\prod_{j=k}^{n}(1-\alpha_j c^2)\right)\alpha_k\xi_{k+1}(\omega)\mathbf{1}_{\Omega_k}(\omega) \\
&\quad + \frac{L}{2}\sum_{k=1}^{n}\alpha_k^2\|V_{k+1}(X_k(\omega))\|^2\mathbf{1}_{\Omega_k}(\omega) \\
&\leq \frac{\epsilon}{2} + \sqrt{R_n(\omega)} + R_n(\omega) \\
&\leq 2\epsilon + \sqrt{\epsilon} < s,
\end{aligned}$$

where the equation in the third line is due to $\omega \in \Omega_k$ for all $k \leq n$ by induction.
$\underline{\beta \in (\frac{1}{2}, 1]}$: Similarly, we obtain from eq. (6.37)

$$\begin{aligned}
D_{n+1}(\omega) &= D_{n+1}(\omega)\mathbf{1}_{\Omega_n}(\omega) \\
&\leq D_1(\omega)\mathbf{1}_{\Omega_n}(\omega)\prod_{k=1}^{n}(1-\alpha_k^q c^2) + \widetilde{c}\sum_{k=1}^{n}\alpha_k^{\frac{2\beta q-1}{2\beta-1}} + \sum_{k=1}^{n}\left(\prod_{j=k}^{n}(1-\alpha_j^q c^2)\right)\alpha_k\xi_{k+1}(\omega)\mathbf{1}_{\Omega_n}(\omega) \\
&\quad + \frac{L}{2}\sum_{k=1}^{n}\alpha_k^2\|V_{k+1}(X_k(\omega))\|^2\mathbf{1}_{\Omega_n}(\omega) \\
&= D_1(\omega)\prod_{k=1}^{n}(1-\alpha_k^q c) + \sum_{k=1}^{n}\left(\prod_{j=k}^{n}(1-\alpha_j c^2)\right)\alpha_k\xi_{k+1}(\omega)\mathbf{1}_{\Omega_k}(\omega) \\
&\quad + \frac{L}{2}\sum_{k=1}^{n}\alpha_k^2\|V_{k+1}(X_k(\omega))\|^2\mathbf{1}_{\Omega_k}(\omega) \\
&\leq \frac{\epsilon}{2} + \frac{\epsilon}{2} + \sqrt{R_n(\omega)} + R_n(\omega) \\
&\leq 2\epsilon + \sqrt{\epsilon} < s.
\end{aligned}$$

We used in both cases that that $\prod_{k=1}^{n}(1-\alpha_k^q c) \leq 1$ and the choice of ϵ such that $2\epsilon + \sqrt{\epsilon} < s$. This proves that condition 2. in Lemma 6.12 is also satisfied which concludes the induction.
c) Without loss of generality we consider the case $\beta = 1/2$. The computations for $\beta \in (1/2, 1]$ follow in line by replacing M_n with $M_n^{(q)}$. By definition it holds that $E_n = E_{n-1} \setminus (E_{n-1} \setminus E_n) = E_{n-1} \setminus \widetilde{E}_n$. Then we have

$$\begin{aligned}
\widetilde{R}_n &= R_n\mathbf{1}_{E_{n-1}} \\
&= R_{n-1}\mathbf{1}_{E_{n-1}} + (R_n - R_{n-1})\mathbf{1}_{E_{n-1}}
\end{aligned}$$

$$= R_{n-1}\mathbf{1}_{E_{n-2}} - R_{n-1}\mathbf{1}_{\widetilde{E}_{n-1}} + (R_n - R_{n-1})\mathbf{1}_{E_{n-1}}$$
$$= \widetilde{R}_{n-1} - R_{n-1}\mathbf{1}_{\widetilde{E}_{n-1}} + (R_n - R_{n-1})\mathbf{1}_{E_{n-1}}$$

and for the last term

$$R_n - R_{n-1} = M_n^2 - M_{n-1}^2 + S_n - S_{n-1}$$
$$= \gamma_n^2(1-\gamma_n c^2)^2\xi_{n+1}^2\mathbf{1}_{\Omega_n} + 2\gamma_n(1-\gamma_n c)\xi_{n+1}\mathbf{1}_{\Omega_n}M_{n-1} + \gamma_n^2\frac{L}{2}\|V_{n+1}(X_n)\|^2\mathbf{1}_{\Omega_n}.$$

We treat each of the summands on the RHS seperately. It follows from the G-Lipschitz continuity and bounded variance assumption in Theorem 6.7, that

$$\begin{aligned}\mathbb{E}[\xi_{n+1}^2\mathbf{1}_{\Omega_n}] &= \mathbb{E}[\langle\nabla f(X_n), V_{n+1}(X_n) - \nabla f(X_n)\rangle^2\mathbf{1}_{\Omega_n}]\\ &\le \mathbb{E}[\|\nabla f(X_n)\|^2(\|V_{n+1}(X_n)\|^2+1)\mathbf{1}_{\Omega_n}] \le G^2(C^2+1),\\ \mathbb{E}[\xi_{n+1}(1-\gamma_n c)M_{n-1}\mathbf{1}_{\Omega_n}] &= \mathbb{E}[\mathbb{E}[\xi_{n+1}|\mathcal{F}_n]M_{n-1}\mathbf{1}_{\Omega_n}] = 0,\\ \mathbb{E}[\|V_{n+1}(X_n)\|^2\mathbf{1}_{\Omega_n}] &\le C.\end{aligned} \tag{6.27}$$

For the term $R_{n-1}\mathbf{1}_{\widetilde{E}_{n-1}}$ we have

$$\mathbb{E}[R_{n-1}\mathbf{1}_{\widetilde{E}_{n-1}}] \ge \epsilon\mathbb{P}(\widetilde{E}_{n-1}).$$

Using $(1-\gamma_n c) < 1$ and putting all together we obtain the claim

$$\mathbb{E}[\widetilde{R}_n] \le \mathbb{E}[\widetilde{R}_{n-1}] + \gamma_n^2[G^2C^2 + G^2 + C] - \epsilon\mathbb{P}(\widetilde{E}_{n-1}).$$

■

LEMMA 6.14. *Let $\delta > 0$ be a tolerance level and $\alpha_n \le \alpha_1$ be sufficiently small such that $\sum_{n=1}^{\infty}\alpha_n^2 < \frac{\delta\epsilon}{2(G^2C^2+G^2+C)}$ and the condition in Lemma 6.13 is fulfilled. Then, we have*

$$\mathbb{P}(E_n) \ge 1 - \frac{\delta}{2}.$$

Proof. The proof is along the lines of the proof of Proposition D2 in [Mer+20]. For completeness we repeat the arguments. First, observe that

$$\mathbb{P}(\widetilde{E}_{n-1}) = \mathbb{P}(E_{n-1}\setminus E_n) = \mathbb{P}(E_{n-1}\cap\{R_n \ge \epsilon\}) = \mathbb{E}[\mathbf{1}_{E_{n-1}}\mathbf{1}_{\{R_n>\epsilon\}}] \le \mathbb{E}[\mathbf{1}_{E_{n-1}}\frac{R_n}{\epsilon}] = \frac{\mathbb{E}[\widetilde{R}_n]}{\epsilon}.$$

On the other hand it follows from Lemma 6.18 that

$$\epsilon\mathbb{P}(\widetilde{E}_n) \le \mathbb{E}[\widetilde{R}_n] \le \mathbb{E}[\widetilde{R}_0] + [G^2C^2+G^2+C]\sum_{k=1}^{n}\alpha_k^2 - \epsilon\sum_{k=0}^{n}\mathbb{P}(\widetilde{E}_{k-1}). \tag{6.28}$$

Rearranging everything yields

$$\sum_{k=0}^{n}\mathbb{P}(\widetilde{E}_k) \le \frac{[G^2C^2+G^2+C]\Gamma}{\epsilon}$$

with $\Gamma = \sum_{n=1}^{\infty} \alpha_n^2$. By the assumption on the step size $\frac{[G^2C^2+G^2+C]\Gamma}{\epsilon} < \frac{\delta}{2}$ and moreover since the events $\widetilde{E}_n$ are disjoint we obtain

$$\mathbb{P}(\bigcup_{k=0}^{n} \widetilde{E}_k) = \sum_{k=0}^{n} \mathbb{P}(\widetilde{E}_k) \leq \frac{\delta}{2} \tag{6.29}$$

implying that

$$\mathbb{P}(E_n) = \mathbb{P}(\bigcap_{k=0}^{n} \widetilde{E}_k^c) \geq 1 - \frac{\delta}{2}\,. \tag{6.30}$$

■

LEMMA 6.15. *Let $\delta > 0$ be a tolerance level and $\alpha_n \leq \alpha_1$ be sufficiently small such that the condition in Lemma 6.13 and Lemma 6.14 are fulfilled. Moreover, we suppose α_1 small enough such that $\frac{4C}{\mathbf{r}^2} \sum_{k=1}^{n} \alpha_k^2 \leq \frac{\delta}{2}$. Then, we have*

$$\mathbb{P}(\widehat{E}_n) \geq 1 - \delta.$$

Proof. By Lemma 6.15, we have $\mathbb{P}(E_n) \geq 1 - \frac{\delta}{2}$. Moreover, by the additional step size assumption and Markov's inequality we deduce that

$$\begin{aligned}
\mathbb{P}(C_n) &= \mathbb{P}(\forall k \leq n \,:\, ||X_{k+1} - X_k|| \leq \frac{\mathbf{r}}{2}) \\
&\geq 1 - \sum_{k=1}^{n} \mathbb{P}(||X_{k+1} - X_k|| > \frac{\mathbf{r}}{2}) \\
&= 1 - \sum_{k=1}^{n} \mathbb{P}(||V_{k+1}(X_k)|| > \frac{\mathbf{r}}{2\alpha_k}) \\
&\geq 1 - \sum_{k=1}^{n} \mathbb{E}[||V_{k+1}(X_k)||^2] \frac{4\alpha_k^2}{\mathbf{r}^2} \\
&\geq 1 - \frac{4C}{\mathbf{r}^2} \sum_{k=1}^{n} \alpha_k^2 \\
&\geq 1 - \frac{\delta}{2}.
\end{aligned}$$

Together we obtain that $\mathbb{P}(\widehat{E}_n) = 1 - \mathbb{P}(\widehat{E}_n^c) \geq 1 - (\mathbb{P}(E_n^c) + \mathbb{P}(C_n^c)) \geq 1 - \delta$. ■

Finally, we are ready to prove the main result in the local setting for the set of local minima $\mathcal{X}^*$.

Proof of Theorem 6.7. (i): Recall the definitions of $\mathcal{U}_1$ and $\mathcal{U}$ above. Then it holds that

$$\Omega_{\mathcal{U}} = \bigcap_{n=1}^{\infty} \Omega_n.$$

Hence, using Lemma 6.15 we obtain

$$\mathbb{P}(\Omega_{\mathcal{U}}) = \inf_n \mathbb{P}(\Omega_n) \geq \inf_n \mathbb{P}(\widehat{E}_n) \geq 1 - \delta.$$

(ii): We define $\widetilde{D}_n := D_n \mathbf{1}_{\Omega_n}$ and prove that $\widetilde{D}_n \in o(1/n^{1-\eta})$. Then the claim follows since $\mathbf{1}_{\Omega_\mathcal{U}} \leq \mathbf{1}_{\Omega_n}$ almost surely.
From the proof of Lemma 6.11 we have

$$\widetilde{D}_{n+1} \leq \widetilde{D}_n - \alpha_n c \widetilde{D}_n^{2\beta} + \alpha_n \xi_n \mathbf{1}_{\Omega_n} + \frac{L\alpha_n^2}{2} \|V_n\|^2 \mathbf{1}_{\Omega_n}.$$

Hence, taking the conditional expectation gives

$$\begin{aligned}\mathbb{E}[\widetilde{D}_{n+1}|\mathcal{F}_n] &\leq \widetilde{D}_n - \alpha_n c \widetilde{D}_n^{2\beta} + \alpha_n \mathbb{E}[\xi_{n+1}|\mathcal{F}_n]\mathbf{1}_{\Omega_n} + \frac{L\alpha_n^2}{2}\mathbb{E}[\|V_{n+1}(X_n)\|^2|\mathcal{F}_n]\mathbf{1}_{\Omega_n} \\ &\leq \widetilde{D}_n - \alpha_n c \widetilde{D}_n^{2\beta} + LC\alpha_n^2,\end{aligned}$$

where we have used that D_n and $\mathbf{1}_{\Omega_n}$ are $\mathcal{F}_n$-measurable and $E[\|V_{n+1}(X_n)\|^2|\mathcal{F}_n] \leq C$ from equation (ABC) with $A = B = 0$. By our step size choice we can apply Lemma 6.1 to obtain Claim (ii).
(iii): In the following, we again separate between the two cases of β.
$\underline{\beta = \frac{1}{2}}$: We have from Lemma 6.11 equation (6.37) that

$$\widetilde{D}_{n+1} \leq (1 - \alpha_n c^2)\widetilde{D}_n + \alpha_n \xi_n \mathbf{1}_{\Omega_n} + \frac{L\alpha_n^2}{2}\|V_{n+1}(X_n)\|^2.$$

Taking expectations and multiplying by $(n+1)^{1-\eta}$ leads to

$$\begin{aligned}&\mathbb{E}[\widetilde{D}_{n+1}(n+1)^{1-\eta}] \\ &\leq (n+1)^{1-\eta}(1-\alpha_n c^2)\mathbb{E}[\widetilde{D}_n] + (n+1)^{1-\eta}\frac{LC\alpha_n^2}{2} \\ &\leq \left(n^{1-\eta} + (1-\eta)n^{-\eta}\right)(1-\alpha_n c^2)\mathbb{E}[\widetilde{D}_n] + (n+1)^{1-\eta}\frac{LC\alpha_n^2}{2} \\ &= \left(n^{1-\eta} + (1-\eta)n^{-\eta} - n^{1-\eta}\alpha_n c^2 - (1-\eta)n^{-\eta}\alpha_n c_{x^*}^2\right)\mathbb{E}[\widetilde{D}_n] + (n+1)^{1-\eta}\alpha_n^2\frac{LC}{2} \\ &= \left(1 + \frac{1-\eta}{n} - \alpha_n c^2 - \frac{(1-\eta)\alpha_n c^2}{n}\right)n^{1-\eta}\mathbb{E}[\widetilde{D}_n] + (n+1)^{1-\eta}\alpha_n^2\frac{LC}{2},\end{aligned}$$

where we used equation (6.27) in the first inequality. By our choice of α_n there exists $\widetilde{c} > 0$ and $N > 0$ such that $\alpha_n c^2 - \frac{1-\eta}{n} + \frac{(1-\eta)\alpha_n c^2}{n} \geq \widetilde{c}\alpha_n$ for all $n \geq N$. Thus, for all $n \geq N$

$$w_{n+1} \leq (1 - \widetilde{c}\alpha_n)w_n + (n+1)^{1-\eta}\alpha_n^2\frac{LC}{2},$$

where $w_n = \mathbb{E}[n^{1-\eta}\widetilde{D}_n]$. Define $a_n = \widetilde{c}\alpha_n$ and $b_n = (n+1)^{1-\eta}\alpha_n^2\frac{LC}{2}$. Since $\alpha_n = \Theta(\frac{1}{n^\theta})$, we have $\sum_n a_n = \widetilde{c}\sum_n \alpha_n = \infty$ and

$$\sum_n b_n = \frac{LC}{2}\sum_n (n+1)^{1-\eta}\alpha_n^2 < \infty,$$

by equation (6.8) in Lemma 6.1 Hence, we apply Lemma 2.16 to prove that $\lim_{n\to\infty} w_n = 0$. By the definition of w_n we have verified that $\mathbb{E}[(f(X_n) - l)\mathbf{1}_{\Omega_\mathcal{U}}] \leq \mathbb{E}[\widetilde{D}_n] \in o(\frac{1}{n^{1-\eta}})$

$\beta \in (\frac{1}{2}, 1]$: From Lemma 6.11 equation (6.37) we have

$$\widetilde{D}_{n+1} \leq (1 - \alpha_n^q c^2)\widetilde{D}_n + \widetilde{c}\alpha_n^{\frac{2\beta q-1}{2\beta-1}} + \alpha_n \xi_n \mathbf{1}_{\Omega_n} + \frac{L\alpha_n^2}{2}\|V_{n+1}(X_n)\|^2,$$

for $\widetilde{c} = (2\beta)^{-\frac{1}{2\beta-1}}(1 - \frac{1}{2\beta})c^2$. Next we multiply with $(n+1)^{1-\eta}$ and use equation (6.1) to obtain

$$\begin{aligned}
&\mathbb{E}[\widetilde{D}_{n+1}(n+1)^{1-\eta}] \\
&\leq (n+1)^{1-\eta}(1 - \alpha_n^q c^2)\mathbb{E}[\widetilde{D}_n] + (n+1)^{1-\eta}\widetilde{c}\alpha_n^{\frac{2\beta q-1}{2\beta-1}} + (n+1)^{1-\eta}LC\alpha_n^2 \\
&\leq \Big(n^{1-\eta} + (1-\eta)n^{-\eta}\Big)\big(1 - \alpha_n^q c^2\big)\mathbb{E}[\widetilde{D}_n] + c_1(n+1)^{1-\eta}(\alpha_n^{\frac{2\beta q-1}{2\beta-1}} + \alpha_n^2) \\
&= \Big(n^{1-\eta} + (1-\eta)n^{-\eta} - \alpha_n^q c^2 n^{1-\eta} - (1-\eta)\alpha_n^q c^2 n^{-\eta}\Big)\mathbb{E}[\widetilde{D}_n] \\
&\quad + c_1(n+1)^{1-\eta}(\alpha_n^{\frac{2\beta q-1}{2\beta-1}} + \alpha_n^2) \\
&= \mathbb{E}[\widetilde{D}_n n^{1-\eta}]\left(1 + \frac{1-\eta}{n} - \alpha_n^q c^2 - \frac{(1-\eta)\alpha_n^q c^2}{n}\right) + c_1(n+1)^{1-\eta}(\alpha_n^{\frac{2\beta q-1}{2\beta-1}} + \alpha_n^2),
\end{aligned}$$

for some $c_1 > 0$. By our choice of α_n and as $q \geq 1$, there exists a $c_2 > 0$ and $N > 0$ such that $\alpha_n^q c^2 - \frac{1-\eta}{n} + \frac{(1-\eta)\alpha_n^q c^2}{n} \geq c_2\alpha_n^q$ for all $n \geq N$. Thus, for $n \geq N$

$$\mathbb{E}[\widetilde{D}_{n+1}(n+1)^{1-\eta}] \leq \mathbb{E}[\widetilde{D}_n n^{1-\eta}]\big(1 - c_2\alpha_n^q\big) + c_1(n+1)^{1-\eta}(\alpha_n^{\frac{2\beta q-1}{2\beta-1}} + \alpha_n^2).$$

Define $w_n = \mathbb{E}[\widetilde{D}_n n^{1-\eta}]$, $a_n = c_2\alpha_n^q$ and $b_n = c_1(n+1)^{1-\eta}(\alpha_n^{\frac{2\beta q-1}{2\beta-1}} + \alpha_n^2)$. We will again apply Lemma 2.16. By the step size choice $\alpha_n = \Theta(\frac{1}{n^\theta})$ we have $\sum_n a_n = c_2 \sum_n \alpha_n^q = \infty$, because $q \leq \frac{1}{\theta}$. Further,

$$\sum_n b_n = c_1 \sum_n (n+1)^{1-\eta}(\alpha_n^{\frac{2\beta q-1}{2\beta-1}} + \alpha_n^2) < \infty,$$

because we choose the auxiliary parameter q as in the proof of Lemma 6.1 where we showed in equation (6.8) and equation (6.9) that

$$\sum_{n=N}^{\infty} n^{1-\eta-2\theta} < \infty \quad \text{and} \quad \sum_{n=N}^{\infty} n^{1-\eta-\frac{\theta(2\beta q-1)}{2\beta-1}} < \infty$$

All together we deduce that w_n vanishes at infinity. Again, by the definition of w_n we have that $\mathbb{E}[(f(X_n) - l)\mathbf{1}_{\Omega_l}] \leq \mathbb{E}[\widetilde{D}_n] \in o(\frac{1}{n^{1-\eta}})$ ∎

6.5.3 *Local gradient domination in f^**

The main result concerning local gradient domination in f^* is presented below and does not necessitate the existence of a local minimum or any stationary point. It is worth noting that the definition of local gradient domination in f^* guarantees the gradient domination property for any x with $f(x)$ close to f^*. Consequently, this definition ensures that functions satisfying this property cannot possess local minima or saddle points within this region.

THEOREM 6.16. *Fix some tolerance level $\delta > 0$. Suppose f satisfies the local gradient domination property in f^* from Definition 6.5 with $\beta \in [\frac{1}{2}, 1]$ and $\mathcal{B}_r^* \subset \mathbb{R}^d$. Moreover, suppose within $\mathcal{B}_r^*$ f is G-Lipschitz continuous, Assumption 2.1 and Assumption 2.12 hold true. Denote by $(X_n)_{n\in\mathbb{N}}$ the sequence generated by equation* (SGD) *using a step size $\alpha_n = \Theta(\frac{1}{n^\theta})$ for $\theta \in (\frac{1}{2}, 1)$ and suppose that $\alpha_n \leq \alpha_1$ for α_1 small enough (dependent on δ). Then, the following holds:*

(i) There exist subsets $\mathcal{U}$ and $\mathcal{U}_1$ of $\mathbb{R}^d$ such that, if $X_1 \in \mathcal{U}_1$ the event $\Omega_{\mathcal{U}} = \{X_n \in \mathcal{U}$ for all $n = 1, 2, \dots\}$ has probability at least $1 - \delta$.

Moreover, for any

$$\eta \in \begin{cases} \left(\max\{2 - 2\theta, \frac{\theta+2\beta-2}{2\beta-1}\}, 1\right) & : \beta \in (\frac{1}{2}, 1] \\ (2 - 2\theta, 1) & : \beta = \frac{1}{2} \end{cases}$$

it holds that

$$(ii)\ \ (f(X_n) - f^*)\mathbf{1}_{\Omega_{\mathcal{U}}} \in o\left(\frac{1}{n^{1-\eta}}\right),\ a.s., \quad and \quad (iii)\ \ \mathbb{E}[(f(X_n) - f^*)\mathbf{1}_{\Omega_{\mathcal{U}}}] \in o\left(\frac{1}{n^{1-\eta}}\right).$$

Suppose throughout this section that the assumptions in Theorem 6.16 are satisfied.
The proof will be similar to the previous section. Instead of assuring that (X_n) remains close to the set where we could guarantee the unified gradient domination property, it is now sufficient that $f(X_n)$ remains close to f^* by the different definition of gradient domination definition in f^*. This will simplify the proof. Moreover, we may again assume w.l.o.g. the uniform second moment bounds, i.e. $A = B = 0$, instead of the more general (ABC) condition by the same argument as above but on the level sets.
Recall the notation

$$\mathcal{B}_r^* = \{x \in \mathbb{R}^d \,:\, f(x) - f^* \leq r\}.$$

and let $r > 0$ be the radius of the gradient domination property in f^*, then there exists $\epsilon > 0$, such that $2\epsilon + \sqrt{\epsilon} < r$, i.e

$$\mathcal{U} := \mathcal{B}^*_{2\epsilon+\sqrt{\epsilon}} \subset \mathcal{B}_r^*. \tag{6.31}$$

Moreover, we define the set

$$\mathcal{U}_1 := \mathcal{B}^*_{\frac{\epsilon}{2}} \tag{6.32}$$

and the measurable subsets

$$\Omega_n = \{X_k \in \mathcal{U},\ \text{for all } k \leq n\}$$

in $(\Omega, \mathcal{F}, \mathbb{P})$.
The proof of Theorem 6.16 is again based on a series of auxiliary lemmas. The goal of these is to prove that with high probability we do not leave the gradient dominated region, i.e. Claim (i) in Theorem 6.16.
In the following, we fix the notation $D_n := f(X_n) - f^*$ and obtain the parallel result to Lemma 6.11.

LEMMA 6.17. *If $\beta = \frac{1}{2}$, it holds that*

$$D_{n+1}\mathbf{1}_{\Omega_n} \tag{6.33}$$

$$\leq (1 - \alpha_n c^2) D_n \mathbf{1}_{\Omega_n} + \alpha_n \xi_n \mathbf{1}_{\Omega_n} + \frac{L\alpha_n^2}{2}\mathbf{1}_{\Omega_n}\|V_{n+1}(X_n)\|^2, \tag{6.34}$$

$$\leq D_1 \mathbf{1}_{\Omega_n} \prod_{k=1}^{n}(1-\alpha_k c^2) + \sum_{k=1}^{n}\prod_{j=k}^{n}(1-\alpha_k c)\alpha_k \xi_k \mathbf{1}_{\Omega_n} + \frac{L}{2}\sum_{k=1}^{n}\alpha_k^2 \|V_{k+1}(X_k)\|^2 \mathbf{1}_{\Omega_n}. \tag{6.35}$$

If $\beta \in (\frac{1}{2}, 1]$, for any $1 \leq q \leq 2$, it holds that

$$D_{n+1}\mathbf{1}_{\Omega_n} \tag{6.36}$$

$$\leq (1-\alpha_n^q c^2) D_n \mathbf{1}_{\Omega_n} + (2\beta)^{-\frac{1}{2\beta-1}}(1-\frac{1}{2\beta})c^2 \alpha_n^{\frac{2\beta q-1}{2\beta-1}} + \alpha_n \xi_n \mathbf{1}_{\Omega_n} + \frac{L\alpha_n^2}{2}\|V_{n+1}(X_n)\|^2 \mathbf{1}_{\Omega_n} \tag{6.37}$$

$$\leq D_1 \mathbf{1}_{\Omega_n} \prod_{k=1}^{n}(1-\alpha_k^q c^2) + \widetilde{c}\sum_{k=1}^{n}\prod_{j=k}^{n}(1-\alpha_j^q c^2)\alpha_k^{\frac{2\beta q-1}{2\beta-1}} \tag{6.38}$$

$$+ \sum_{k=1}^{n}\prod_{j=k}^{n}(1-\alpha_j^q c^2)\alpha_k \xi_k \mathbf{1}_{\Omega_n} + \frac{L}{2}\sum_{k=1}^{n}\alpha_k^2 \|V_{k+1}(X_k)\|^2 \mathbf{1}_{\Omega_n}, \tag{6.39}$$

for $\widetilde{c} = (2\beta)^{-\frac{1}{2\beta-1}}(1-\frac{1}{2\beta})c^2$.

Proof. The proof follows line for line as in Lemma 6.11 by replacing l with f^* and taking the different definition of Ω_n into account. ∎

We continue as in the previous section:
For $\beta > \frac{1}{2}$ we know from the proof of Lemma 6.1 that we can choose the auxiliary parameter q from the previous lemma in such a way, that $\sum_{n=1}^{\infty} n^{1-\eta}\alpha_n^{\frac{2\beta q-1}{2\beta-1}}$ is convergent for all $\eta \in (\max\{2-2\theta, \frac{\theta+2\beta-2}{2\beta-1}\}, 1)$ (Condition (iii) to apply Lemma 2.15). As $\eta < 1$, it follows that $\sum_{n=1}^{\infty}\alpha_n^{\frac{2\beta q-1}{2\beta-1}} < \infty$ holds true for all these choices of q. Now define

$$M_n = \sum_{k=1}^{n}\prod_{j=k}^{n}(1-\alpha_j c^2)\alpha_k \boldsymbol{\xi}_{k+1}\mathbf{1}_{\Omega_k}, \quad M_n^{(q)} = \sum_{k=1}^{n}\prod_{j=k}^{n}(1-\alpha_j^q c^2)\alpha_k \boldsymbol{\xi}_{k+1}\mathbf{1}_{\Omega_k}$$

and

$$S_n = \frac{L}{2}\sum_{k=1}^{n}\alpha_k^2\|V_{k+1}(X_k)\|^2\mathbf{1}_{\Omega_k}.$$

Then, $(M_n)_{n\in\mathbb{N}}$ and $(M_n^{(q)})_{n\in\mathbb{N}}$ is a $(\mathcal{F}_{n+1})$-martingale with zero mean and $(S_n)_{n\in\mathbb{N}}$ is a $(\mathcal{F}_{n+1})$-sub-martingale by Assumption 2.12. Note that by the choice of α_n we have that $\sum_n \alpha_n^2 < \infty$ and hence $\mathbb{E}[S_n] < \infty$ for all $n \in \mathbb{N}$.
Next, define $R_n = M_n^2 + S_n$ and $R_n = (M_n^{(q)})^2 + S_n$ (with sight abuse of notation) for every $n \in \mathbb{N}$. Moreover, let

$$E_n = \{R_k < \epsilon \text{ for all } k \leq n\}.$$

which is an $\mathcal{F}_{n+1}$-measurable event on $(\Omega, \mathcal{F}, \mathbb{P})$. We define $R_0 = 0$ such that $E_0 = \Omega$.
With these definitions we can directly prove a parallel result to Lemma 6.13 without the auxiliary result in Lemma 6.12.

LEMMA 6.18. *For $\beta \in (\frac{1}{2}, 1]$ let $\alpha_n \leq \alpha_1$ be sufficiently small such that $\sum_{n=1}^{\infty}\alpha_n^{\frac{2\beta q-1}{2\beta-1}} < \frac{\epsilon}{2\widetilde{c}}$, and for $\beta = \frac{1}{2}$ let $\alpha_1 > 0$ be arbitrary. Furthermore, assume that the initial $X_1 \in \mathcal{U}_1 = \{x : f(x) - f(x^*) \leq \frac{\epsilon}{2}\}$ almost surely. Then,*

a) $E_{n+1} \subset E_n$ *and* $\Omega_{n+1} \subset \Omega_n$

b) $E_n \subset \Omega_{n+1}$

c) Define the events $\widetilde{E}_n = E_{n-1} \setminus E_n = E_{n-1} \cup \{R_n \geq \epsilon\}$. *Then, for* $\widetilde{R}_n = R_n \mathbf{1}_{E_{n-1}}$, *there exists a* $\widetilde{C} > 0$ *such that*

$$\mathbb{E}[\widetilde{R}_n] \leq \mathbb{E}[\widetilde{R}_{n-1}] + \alpha_n^2[G^2C^2 + G^2 + C] - \epsilon \mathbb{P}(\widetilde{E}_{n-1}).$$

Proof. a) Follows by definition.
b) Note that $E_0 = \Omega = \Omega_1$ because $X_1 \in \mathcal{U}_1$ almost surely by assumption. We prove the assertion by induction. Let $\omega \in E_n$. Since $E_n \subset E_{n-1} \subset \Omega_n$ by induction assumption, we have $\omega \in \Omega_n$ and thus $\omega \in \Omega_k$ for all $k \leq n$. It remains to show that $X_{n+1}(\omega) \in \mathcal{U}$ to prove that $\omega \in \Omega_{n+1}$. We separate both cases for β:
$\underline{\beta = \frac{1}{2}}$: The inequality (6.34) and the induction hypothesis yield

$$\begin{aligned}
&D_{n+1}(\omega) \\
&\leq D_1(\omega) \prod_{k=1}^{n}(1 - \alpha_k c) + \sum_{k=1}^{n} \prod_{j=k}^{n}(1 - \alpha_k c)\alpha_k \xi_{k+1}(\omega)\mathbf{1}_{\Omega_k}(\omega) + \frac{L}{2}\sum_{k=1}^{n} \alpha_k^2 \|V_{k+1}(X_k(\omega))\|^2 \mathbf{1}_{\Omega_k} \\
&\leq \frac{\epsilon}{2} + \sqrt{R_n(\omega)} + R_n(\omega) \\
&\leq 2\epsilon + \sqrt{\epsilon}.
\end{aligned}$$

Hence, $X_{n+1}(\omega) \in \mathcal{U}$ by definition of $\mathcal{U}$.
$\underline{\beta \in (\frac{1}{2}, 1]}$: Similarly, we obtain from equation (6.37)

$$\begin{aligned}
&D_{n+1}(\omega) \\
&\leq D_1(\omega) \prod_{k=1}^{n}(1 - \alpha_k^q c) + \widetilde{c} \sum_{k=1}^{n} \alpha_k^{\frac{2\beta q - 1}{2\beta - 1}} \\
&+ \sum_{k=1}^{n} \prod_{j=k}^{n}(1 - \alpha_k^q c)\alpha_k \xi_{k+1}(\omega)\mathbf{1}_{\Omega_k}(\omega) + \frac{L}{2}\sum_{k=0}^{n} \alpha_k^2 \|V_{k+1}(X_k(\omega))\|^2 \mathbf{1}_{\Omega_k} \\
&\leq \frac{\epsilon}{2} + \frac{\epsilon}{2} + \sqrt{R_n(\omega)} + R_n(\omega) \\
&\leq 2\epsilon + \sqrt{\epsilon},
\end{aligned}$$

where we used that $\prod_{k=1}^{n}(1 - \alpha_k^q c^*) < 1$. Hence, it holds again that $X_{n+1}(\omega) \in \mathcal{U}$.
This prove that $\omega \in \Omega_{n+1}$ and closes the induction.
c) Follows line by line as in Lemma 6.13 part c). ∎

LEMMA 6.19. *Let* $\delta > 0$ *be a tolerance level and* $\alpha_n \leq \alpha_1$ *be sufficiently small such that* $\sum_{n=1}^{\infty} \alpha_n^2 < \frac{\delta\epsilon}{2(G^2C^2+G^2+C)}$ *and the condition in Lemma 6.18 is fulfilled. Then, we have*

$$\mathbb{P}(E_n) \geq 1 - \delta.$$

Proof. Line by line as in Lemma 6.14. ∎

Finally, we are ready to prove the main result in the local setting for f^*.

Proof of Theorem 6.16. (i): Recall the definition of $\mathfrak{U}_1$ and $\mathfrak{U}$ above. Then it holds that

$$\Omega_{\mathfrak{U}} = \bigcap_{n=1}^{\infty} \Omega_n.$$

Hence, using Lemma 6.19 we obtain

$$\mathbb{P}(\Omega_{\mathfrak{U}}) = \inf_n \mathbb{P}(\Omega_n) \geq \inf_n \mathbb{P}(E_n) \geq 1 - \delta.$$

The proof of Claim (ii) and (iii) follows line by line as in the proof of Theorem 6.7 by replacing l with f^* and taking the different definitions of D_n and Ω_n into account. ∎

6.6 Application in the training of neural networks

In supervised learning one aims to approximate an unknown model $\varphi : \mathbb{R}^{d_z} \to \mathbb{R}^{d_y}$ by a parametrized function $g_w : \mathbb{R}^{d_z} \to \mathbb{R}^{d_y}$ with parameter $w \in \mathbb{R}^{d_w}$. Given a family of training data $((Z^{(m)}, Y^{(m)}))_{m\in\mathbb{N}}$ generated as i.i.d. samples from an unknown distribution $\mu_{(Z,Y)}$ one usually chooses the parameter $w \in \mathbb{R}^{d_w}$ by solving

$$\min_{w\in\mathbb{R}^{d_w}} \mathbb{E}_{\mu_{(Z,Y)}}[\Phi(g_w(Z), Y)],$$

where $\Phi : \mathbb{R}^{d_y} \times \mathbb{R}^{d_y} \to \mathbb{R}_+$ is a user specific data discrepancy. One popular choice of parametrizations are DNNs. We define a neural network of depth $L \in \mathbb{N}$ by the recursion

$$z_0 := z, \quad z_\ell = \sigma^{\otimes d_\ell}(A_\ell z_{\ell-1} + b_\ell),\ \ell = 1, \dots, L-1, \quad g_w(z) := A_L z_{L-1} + b_L.$$

The weights $((A_\ell, b_\ell))_{\ell=1}^L$ of the DNN are collected in $w \in \mathcal{W} := \times_{\ell=1}^L (\mathbb{R}^{d_\ell \times d_{\ell-1}} \times \mathbb{R}^{d_\ell}) \simeq \mathbb{R}^{d_w}$, and $\sigma^{\otimes d} : \mathbb{R}^d \to \mathbb{R}^d$ describes the component-wise application of the activation function $\sigma : \mathbb{R} \to \mathbb{R}$. Provided that σ and Φ are analytic, and (Z, Y) are compactly supported $\mathbb{R}^{d_z} \times \mathbb{R}^{d_y}$-valued random variables, then $f^{\mathrm{DNN}} : \mathbb{R}^{d_w} \to \mathbb{R}_+$ defined by $w \mapsto \mathbb{E}_{\mu_{(Z,Y)}}[\Phi(g_w(Z), Y)]$ is analytic [DK21, Thm. 5.2] and therefore satisfies local gradient domination in any stationary point w_* [Loj65].
In our notation, the stochastic first order oracle in Section 2.2.1 takes the form

$$V(w, (Z, Y)) = \nabla_w f^{\mathrm{DNN}}(w) + (\nabla_w \Phi(g_w(Z), Y) - \nabla_w f^{\mathrm{DNN}}(w)),$$

where we denote $\zeta = (Z, Y)$ and the iterative SGD then reads as

$$W_{n+1} = W_n - \alpha_n \nabla_w \Phi(g_{W_n}(Z_{n+1}), Y_{n+1})$$

with $\zeta_n = (Z_n, Y_n)$ independent and identical distributed. The iterative scheme of SHB can be written similarly. Note that this scenario also includes the empirical risk minimization of $\frac{1}{M}\sum_{m=1}^M \Phi(g_w(z^{(m)}), y^{(m)})$ when $\zeta = (Z, Y) \sim \frac{1}{M}\sum_{m=1}^M \delta_{(z^{(m)}, y^{(m)})}$, see Example 2.13 for more details. The following local convergence is a direct consequence of Theorem 6.7.

COROLLARY 6.20. *Let $\delta > 0$. Denote by $(W_n)_{n\in\mathbb{N}}$ the sequence generated by SGD with $w \mapsto \nabla_w f^{\mathrm{DNN}}(w)$ as objective function, step size $\alpha_n \in \Theta(n^{-\theta})$ for $\theta \in (\frac{1}{2}, 1)$, and assume that f^{DNN} is analytic. Let $\mathcal{W}^*$ be an isolated compact set of local minima with level $l = f^{\mathrm{DNN}}(w^*)$ for all $w^* \in \mathcal{W}^*$ and suppose Assumption 2.12 is satisfied within $\mathcal{W}^*$. Suppose that $\alpha_n \leq \alpha_1$ for sufficiently small α_1 (depending on δ), then there exist two subsets $\mathcal{U}, \mathcal{U}_1$ of $\mathbb{R}^{d_w}$ such that $W_1 \in \mathcal{U}_1$ implies that the event $\Omega_{\mathcal{U}} = \{W_n \in \mathcal{U}, \text{ for all } n \geq 1\}$ has probability at least $1-\delta$. Moreover, there exists $\beta \in [\frac{1}{2}, 1]$ such that for any*

$$\eta \in \begin{cases} \left(\max\{2-2\theta, \frac{\theta+2\beta-2}{2\beta-1}\}, 1\right) & : \beta \in (\frac{1}{2}, 1] \\ (2-2\theta, 1) & : \beta = \frac{1}{2} \end{cases}$$

it holds that $|f^{\mathrm{DNN}}(W_n) - l|\mathbf{1}_\Omega \in o\big(n^{\eta-1}\big)$ almost surely and in expectation.

In words: If the iterates of SGD reach a certain area around a local minimum, they are likely to become trapped in that region with high probability, provided that the step size is sufficiently small. This results shows that, under very general conditions, SGD converges to local minima and furthermore quantifies the convergence speed.

Remark 6.21. One may similarly apply Theorem 6.16 in the training of DNNs to derive convergence towards a global minimum with high probability provided that the initial loss $f^{\mathrm{DNN}}(X_1)$ and initial step size α_1 are sufficiently small.

6.7 APPLICATION IN POLICY GRADIENT

We will see in this section how the non-uniform gradient domination property for tabular softmax PG implies a local gradient domination property around the global optimum. We consider the case of infinite-horizon discounted MDPs with and without entropy regularization and in the case of finite-time MDPs.

Infinite-time horizon discounted MDPs. Let $(\mathcal{S}, \mathcal{A}, \gamma, r, p)$ be a discounted MDP with finite state and action space and discount factor $\gamma \in [0, 1)$. Further, we assume that the rewards are bounded in $[0, 1]$. Consider the stationary tabular softmax policy for parameter $w \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$[1], i.e.

$$\pi_w(a|s) = \frac{\exp(w(s,a))}{\sum_{a'\in\mathcal{A}_s}\exp(w(s,a'))}, \quad \forall s \in \mathcal{S},\ a \in \mathcal{A}.$$

Then, for an initial state distribution μ, recall that the value function under this parametrization is given by

$$V^{\pi_w}(\mu) = \mathbb{E}_\mu^{\pi_w}\Big[\sum_{t=0}^{\infty}\gamma^t r(S_t, A_t)\Big]$$

and the optimal value function is denoted by $V^*(\mu)$. In order to maximize the objective we use stochastic gradient ascent as in Algorithm 2, with unbiased gradient estimator in equation (3.11). Note that this stochastic first order oracle meets the conditions required in Assumption 2.12 with $A = B = 0$ [Zha+20, see]. Moreover, the following non-uniform weak gradient domination property holds for this optimization problem [Lemma 3.19]: For every $w \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$ it holds that

$$\|\nabla_w V^{\pi_w}(\mu)\|_2 \geq c(w)(V^*(\mu) - V^{\pi_w}(\mu))),$$

[1] We use parameter w instead of θ in this section, as θ is already used in the step size schedule.

with non-constant $c(w) = \frac{\min_{s\in\mathcal{S}} \pi_w(a^*(s)|s)}{\sqrt{|\mathcal{S}|}(1-\gamma)} \left\| \frac{d_\mu^{\pi^*}}{\mu} \right\|_\infty^{-1}$, where $a^*(s)$ denotes the (w.l.o.g. unique) best possible action in state s.
We prove that this non-uniform weak gradient domination property implies local gradient domination with $\beta = 1$.

LEMMA 6.22. *There exists $r, c > 0$ such that for all $w \in \mathcal{B}_r^* = \{w : V^*(\mu) - V^{\pi_w}(\mu) \leq r\}$ it holds that $c(w) \geq c$.*

Proof. Define the optimal reward gap in every state $s \in \mathcal{S}$ by

$$\Delta^*(s) = Q^*(s, a^*(s)) - \max_{a \neq a^*(s)} Q^*(s, a) > 0,$$

where $a^*(s)$ denotes the best possible action in state s and $Q^* : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ denotes the optimal Q-function defined by $Q^*(s,a) = \mathbb{E}_\mu^{\pi^*}[\sum_{t=0}^\infty \gamma^{-t} r(S_t, A_t)|A_0 = a]$. W.l.o.g. we assume that $a^*(s)$ is unique. Similarly let $Q^{\pi_w}(s,a) = \mathbb{E}_\mu^{\pi_w}[\sum_{t=0}^\infty \gamma^{-t} r(S_t, A_t)|A_0 = a]$ be the Q-function for policy π_w. For any $0 < \alpha < 1$ choose $r = \min_{s\in\mathcal{S}} \mu(s) \min_{s\in\mathcal{S}} \Delta^*(s)(1-\alpha)$ and assume that $w \in \mathcal{B}_r^*$, i.e. $V^*(\mu) - V^{\pi_w}(\mu) \leq r$. Then, we have for every $s \in \mathcal{S}$ that

$$V^*(\delta_s) - V^{\pi_w}(\delta_s) \leq \frac{r}{\min_{s\in\mathcal{S}} \mu(s)}.$$

It follows for every $s \in \mathcal{S}$ that

$$\begin{aligned}
\frac{r}{\min_{s\in\mathcal{S}} \mu(s)} &\geq V^*(\delta_s) - V^{\pi_w}(\delta_s) \\
&= Q^*(s, a^*(s)) - \sum_{a\in\mathcal{A}_s} \pi_w(a|s) Q^{\pi_w}(s,a) \\
&\geq \sum_{a\in\mathcal{A}_s} \pi_w(a|s)(Q^*(s,a^*(s)) - Q^*(s,a)) \\
&= \sum_{a\neq a^*(s)} \pi_w(a|s)(Q^*(s,a^*(s)) - Q^*(s,a)) \\
&\geq (1 - \pi_w(a^*(s)|s) \min_s \Delta^*(s).
\end{aligned}$$

Rearranging results in

$$\pi_w(a^*(s)|s) \geq 1 - \frac{r}{\min_{s\in\mathcal{S}} \mu(s) \min_{s\in\mathcal{S}} \Delta^*(s)} = \alpha.$$

Hence, for all $w \in \mathcal{B}_r^*$ we can bound $c(w)$ by

$$c(w) \geq \frac{\alpha}{\sqrt{|\mathcal{S}|}(1-\gamma)} \left\| \frac{d_\mu^{\pi^*}}{\mu} \right\|_\infty^{-1} > 0.$$

Thus, setting $c = \frac{\alpha}{\sqrt{|\mathcal{S}|}(1-\gamma)} \left\| \frac{d_\mu^{\pi^*}}{\mu} \right\|_\infty^{-1}$ proves the claim. ∎

As the objective function $w \mapsto V^{\pi_w}(\mu)$ is smooth and moreover Lipschitz on $\mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ [YGL22, Lem. E.1], all assumptions in Theorem 6.16 are satisfied and we obtain the following result.

COROLLARY 6.23. *Let $\delta > 0$. Denote by (w_n) the sequence generated by SGD with $w \mapsto -V^{\pi_w}(\mu)$ as objective function, step size $\alpha_n \in \Theta(n^{-\theta})$ for $\theta \in (\frac{1}{2}, 1)$ and suppose $\alpha_n \leq \alpha_0$ for sufficiently small α_0 (depending on δ). Then, there exist two subsets $\mathcal{U}, \mathcal{U}_1$ of $\mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ such that $w_0 \in \mathcal{U}_1$ implies that the event $\Omega_{\mathcal{U}} = \{w_n \in \mathcal{U}, \text{ for all } n \geq 0\}$ has probability at least $1 - \delta$. Moreover, for any $\eta \in (\max\{2 - 2\theta, \theta\}, 1)$ it holds that $(V^*(\mu) - V^{\pi_{w_n}}(\mu))\mathbf{1}_\Omega \in o\big(n^{\eta-1}\big)$ almost surely and in expectation.*

To the best of our knowledge, Corollary 6.23 present the first local analysis for stochastic policy gradient without regularization.
In words: If the stochastic policy gradient algorithm is started close enough to the optimum a nearly $o(n^{-\frac{1}{3}})$ almost sure rate of convergence can be obtained by choosing $\theta = \frac{2}{3}$ (in contrast to $o(n^{-1})$ in policy gradient with access to exact gradients).

Infinite-time horizon discounted MDPs with entropy regularization. In the following, we apply Theorem 6.16 also to the entropy regularized setting, with regularization parameter $\lambda > 0$. We will see that one can also achieve convergence arbitrarily close to $o(\frac{1}{n})$ without the need for an increasing batch size. Moreover, the local convergence occurs almost surely on an event with high probability. Consider the regularized objective

$$V_\lambda^{\pi_w}(\mu) = V^{\pi_w}(\mu) - \lambda \mathbb{E}_\mu^{\pi_w}\Big[\sum_{t=0}^{\infty} \gamma^t \log(\pi_w(A_t|S_t)\Big] \tag{6.40}$$

and denote by $V_\lambda^*(\mu)$ the global optimum. Due to regularization there exists a continuum of optimal parameters w^*, such that $V_\lambda^{\pi_{w^*}}(\mu) = V_\lambda^*(\mu)$. We will write $\pi^* = \pi_{w^*}$.
Ding, Zhang, and Lavaei [DZL23] present in Equation (4) a stochastic gradient estimator that satisfies Assumption 2.12 with $A = B = 0$. Moreover, the following stronger non-uniform PL-inequality holds [Mei+20, Lem. 15]: For every $w \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$ it holds that

$$\|\nabla_w V_\lambda^{\pi_w}(\mu)\|_2 \geq c(w)^{\frac{1}{2}}\big[V_\lambda^*(\mu) - V_\lambda^{\pi_w}(\mu)\big]^{\frac{1}{2}},$$

with $c(w) = \frac{2\lambda}{|\mathcal{S}|(1-\gamma)} \min_s \mu(s) \min_{s,a} \pi_w(a|s)^2 \left\|\frac{d_\mu^{\pi^*}}{\mu}\right\|_\infty^{-1}$. We prove that this implies a local strong gradient domination property with $\beta = \frac{1}{2}$ in the following sense.

LEMMA 6.24. *There exists $r, c > 0$ such that for all $w \in \mathcal{B}_{r,\lambda}^* = \{w : V_\lambda^*(\mu) - V_\lambda^{\pi_w}(\mu) \leq r\}$ it holds that $c(w) \geq c$.*

Proof. For any $\alpha \in (0, 1)$ choose $r = \alpha^2 \exp\Big(\frac{-1}{(1-\gamma)\lambda}\Big)^2 \frac{\lambda \min_s \mu(s)}{2\ln 2}$ and assume that $w \in \mathcal{B}_{r,\lambda}^*$. By Ding, Zhang, and Lavaei [DZL23, Lem. 12] we have

$$\begin{aligned}
|\pi_w(a|s) - \pi^*(a|s)| &\leq \sqrt{\frac{2(V_\lambda^*(\mu) - V_\lambda^{\pi_w}(\mu))\ln 2}{\lambda \min_s \mu(s)}} \\
&\leq \sqrt{\frac{2r\ln 2}{\lambda \min_s \mu(s)}} = \alpha \exp\left(\frac{-1}{(1-\gamma)\lambda}\right) \\
&\leq \alpha \min_{s,a} \pi^*(a|s).
\end{aligned}$$

where the last inequality is due to Nachum et al. [Nac+17, Thm. 1]. It follows directly that

$$\min_{s,a} \pi_w(s,a) \geq (1-\alpha) \min_{s,a} \pi^*(s,a) > 0.$$

Hence, we can bound $c(w)$ uniformly for all $w \in \mathcal{B}^*_{r,\lambda}$ by

$$c(w) \geq \frac{2\lambda}{|\mathcal{S}|(1-\gamma)} \min_s \mu(s)(1-\alpha)^2 \min_{s,a} \pi^*(a|s)^2 \left\| \frac{d_\mu^{\pi^*}}{\mu} \right\|_\infty^{-1}. \tag{6.41}$$

Thus, setting $c = \frac{2\lambda}{|\mathcal{S}|(1-\gamma)} \min_s \mu(s)(1-\alpha)^2 \min_{s,a} \pi^*(a|s)^2 \left\| \frac{d_\mu^{\pi^*}}{\mu} \right\|_\infty^{-1}$ proves the claim. ∎

As $w \mapsto V_\lambda^{\pi_w}$ is also smooth and globally Lipschitz [DZL23] we deduce the convergence corollary.

Corollary 6.25. *Let $\delta > 0$. Denote by (w_n) the sequence generated by SGD with $w \mapsto -V_\lambda^{\pi_w}(\mu)$ as objective function, step size $\alpha_n \in \Theta(n^{-\theta})$ for $\theta \in (\frac{1}{2}, 1)$ and suppose $\alpha_n \leq \alpha_1$ for sufficiently small α_1 (depending on δ). Then, there exist two subsets $\mathcal{U}, \mathcal{U}_1$ of $\mathbb{R}^{H|\mathcal{S}||\mathcal{A}|}$ such that $w_1 \in \mathcal{U}_1$ implies that the event $\Omega_{\mathcal{U}} = \{w_n \in \mathcal{U}, \text{ for all } n \geq 0\}$ has probability at least $1-\delta$. Moreover, for any $\eta \in (2-2\theta, 1)$ it holds that $(V_\lambda^*(\mu) - V_\lambda^{\pi_{w_n}}(\mu))\mathbf{1}_\Omega \in o\big(n^{\eta-1}\big)$ almost surely and in expectation.*

In words: If the regularized stochastic policy gradient algorithm is started close enough to the optimum a nearly $o(n^{-1})$ almost sure rate of convergence can be obtained by choosing θ close to 1 (in contrast to linear convergence known in regularized policy gradient with access to exact gradients).

Finite-time horizon MDPs. A similar result holds also for finite-time horizon MDPs, $(\mathcal{H}, \mathcal{S}, \mathcal{A}, p, r)$, with $\mathcal{S}$, $\mathcal{A}$, p, r as before, but finite decision epochs $\mathcal{H} = \{0, \dots, H-1\}$ and no discounting ($\gamma = 1$). We consider the dynamic policy gradient algorithm in finite time (FT-DynPG) from Chapter 4. Then, recall that the value function is given by

$$V_0^{\pi^{\mathcal{W}}}(\mu) = \mathbb{E}_\mu^{\pi^{\mathcal{W}}}\Big[\sum_{h=0}^{H-1} r(S_h, A_h)\Big],$$

where $\pi^{\mathcal{W}} = (\pi^{w_h})_{h=0}^{H-1}$ the non-stationary tabular softmax parametrization, i.e. $w_h \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$

$$\pi^{w_h}(a|s) = \frac{\exp(w_h(s,a))}{\sum_{a' \in \mathcal{A}_{s_h}} \exp(w_h(s,a'))}, \quad \forall s \in \mathcal{S},\ a \in \mathcal{A}.$$

The objective functions

$$J_h(w_h, \pi_{(h+1)}, \mu_h) = \mathbb{E}_{\mu_h}^{(\pi^{w_h}, \pi_{(h+1)})}[\sum_{l=h}^{H-1} r(S_l, A_l)]$$

are optimized backwards in time, given the already optimized fixed future policy $\pi_{(h+1)} = (\pi_h)_{h=h+1}^{H-1}$. Let $J_h^*(\pi_{(h+1)}, \mu_h)$ be the optimal value function given that the policy after h is fixed

by $\pi_{(h+1)}$. Then, the following non-uniform weak gradient domination holds for this optimization problem (see Lemma 4.19): For all $w_h \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$ holds that

$$\begin{aligned}&\|\nabla_{w_h} J_h(w_h, \pi_{(h+1)}, \mu_h)\|_2 \geq \\ &\min_{s\in\mathcal{S}} \pi^{w_h}(a^*(s)|s)(J_h^*(\pi_{(h+1)}, \mu_h) - J_h(w_h, \pi_{(h+1)}, \mu_h)),\end{aligned}$$

where $a^*(s)$ denotes the best possible action in state s and is dependent on the fixed future policy $\pi_{(h+1)}$. Again, w.l.o.g. $a^*(s)$ is unique and the following Lemma shows how $c(w_h) = \min_{s\in\mathcal{S}} \pi^{\theta_h}(a^*|s)$ can be bounded away from zero from below around the global optimum.

LEMMA 6.26. *Fix $h \in \{0, \dots, H-1\}$. There exists $c, r > 0$ such that for all $w_h \in \mathcal{B}_{r,h}^* := \{w_h : J_h^*(\pi_{(h+1)}, \mu_h) - J_h(w_h, \pi_{(h+1)}, \mu_h) \leq r\}$ it holds that $c(w_h) \geq c$.*

Proof. Define the optimal reward gap in every state $s \in \mathcal{S}$ of epoch h (given the fixed future policies) by

$$\Delta_h^*(s) = Q_h^{\pi_{(h+1)}}(s, a^*(s)) - \max_{a\neq a^*(s)} Q_h^{\pi_{(h+1)}}(s, a) > 0,$$

where $a^*(s)$ denotes the best possible action in state s (given the fixed future policy). W.l.o.g. we assume that this action is unique.
For any $0 < \alpha < 1$, choose $r = \min_{s\in\mathcal{S}} \mu_h(s) \min_{s\in\mathcal{S}} \Delta_h^*(s)(1-\alpha)$. Then, for $w_h \in \mathcal{B}_{r,h}^*$ with $J_h^*(\pi_{(h+1)}, \mu_h) - J_h(w_h, \pi_{(h+1)}, \mu_h) \leq r$ we have for every $s \in \mathcal{S}$ that

$$J_h^*(\pi_{(h+1)}, \delta_s) - J_h(\theta_h, \pi_{(h+1)}, \delta_s) \leq \frac{r}{\mu_h(s)} \leq \frac{r}{\min_{s\in\mathcal{S}} \mu_h(s)}.$$

It follows for every $s \in \mathcal{S}$ that

$$\begin{aligned}\frac{r}{\min_{s\in\mathcal{S}} \mu_h(s)} &\geq J_h^*(\pi_{(h+1)}, \delta_s) - J_h(\theta_h, \pi_{(h+1)}, \delta_s) \\ &= \left(Q_h^{\pi_{(h+1)}}(s, a^*(s)) - \sum_{a\in\mathcal{A}_s} \pi^{\theta_h}(a|s) Q_h^{\pi_{(h+1)}}(s, a)\right) \\ &= \sum_{a\in\mathcal{A}_s} \pi^{\theta_h}(a|s)\left(Q_h^{\pi_{(h+1)}}(s, a^*(s)) - Q_h^{\pi_{(h+1)}}(s, a)\right) \\ &= \sum_{a\neq a^*(s)} \pi^{\theta_h}(a|s)\left(Q_h^{\pi_{(h+1)}}(s, a^*(s)) - Q_h^{\pi_{(h+1)}}(s, a)\right) \\ &\geq \sum_{a\neq a^*(s)} \pi^{\theta_h}(a|s)\Delta_h^*(s) \\ &= \left(1 - \pi^{\theta_h}(a^*(s)|s)\right)\Delta_h^*(s) \\ &\geq \left(1 - \pi^{\theta_h}(a^*(s)|s)\right) \min_{s\in\mathcal{S}} \Delta_h^*(s).\end{aligned}$$

Rearranging results in

$$\pi^{\theta_h}(a^*(s)|s) \geq 1 - \frac{r}{\min_{s\in\mathcal{S}} \mu_h(s) \min_{s\in\mathcal{S}} \Delta_h^*(s)} = \alpha, \quad \forall s \in \mathcal{S}.$$

Hence, for all $w_h \in \mathcal{B}^*_{r,h}$ we can bound $c(w_h)$ by

$$c(w_h) \geq \alpha > 0.$$

Thus, setting $c = \alpha$ proves the claim. ■

We deduce the following convergence result.

COROLLARY 6.27. *Let $\delta > 0$. Denote by $(w_h^{(n)})$ the sequence generated by SGD with $w_h \mapsto -J_h(w, \pi_{(h+1)}, \mu_h)$ as objective function, step size $\alpha_h^{(n)} \in \Theta(n^{-\theta})$ for $\theta \in (\frac{1}{2}, 1)$ and suppose $\alpha_h^{(n)} < \alpha_h^{(1)}$ for sufficiently small $\alpha_h^{(1)}$ (depending on δ). Then, there exist two sets $\mathcal{U}, \mathcal{U}_1 \in R^{|\mathcal{S}||\mathcal{A}|}$ such that $w_h^{(1)} \in \mathcal{U}_1$ implies that $\Omega_{\mathcal{U}} = \{w_h^{(n)} \in \mathcal{U}, \text{ for all } n \geq 0\}$ occurs with probability at least $1 - \delta$. Moreover, for any $\eta \in (\max\{2 - 2\theta, \theta\}, 1)$ it holds that $(J_h(w_h^{(n)}, \pi_{(h+1)}, \mu_h) - J_h^*(\pi_{(h+1)}, \mu_h))\mathbf{1}_{\Omega_n} \in o\big(n^{\eta-1}\big)$ almost surely and in expectation.*

In words: In every optimization loop of stochastic FT-DynPG, we converge almost surely to the global optimum with rate $o\big(n^{-\frac{1}{3}+\epsilon}\big)$, when the initialization condition is fulfilled and the step size schedule is chosen with $\theta = \frac{2}{3}$.

Take aways. In all three cases we obtain almost sure convergence with high probability, without the need of a large batch size.
Note also, that in all cases we obtain from the proofs of Lemma 6.22, Lemma 6.24 and Lemma 6.26 that r and c can be explicitly chosen (depending on α). Hence, one can choose the neighbourhoods $\mathcal{U}$ and $\mathcal{U}_1$ w.r.t. r in Lemma 6.18 as in equation (6.31) and equation (6.32). We obtain an explicit characterization of the neighbourhood $\mathcal{U}_1$ as condition for initialization.

Conclusion and Future Work

7

We close this thesis by summarizing the discussed methods and giving a brief overview on interesting future work.

Chapter 4

In Chapter 4, we have presented two PG methods for finite-time horizon MDPs and derived a convergence analysis of under the tabular softmax parametrization. Assuming exact gradients we have obtained an $\mathcal{O}(1/n)$-convergence rate for both approaches where the behavior regarding the time horizon and the model-dependent constant c is better in the dynamic approach than in the simultaneous approach. In the model-free setting with estimated gradients, we have derived complexity bounds to approximate the error to global optima with high probability. It would be desirable to derive tighter bounds in the stochastic setting, using for example adaptive step sizes or variance reduction methods that lead to more realistic batch sizes and step sizes. To partly answer this question, we refer to Section 6.7, where we presented a local convergence analysis without the need of a batch size.
Similar to many recent results, the presented analysis relied on the tabular parametrization. However, the heuristic intuition from the policy gradient theorem does not, and the dynamic programming perspective suggests that parameters should be trained backwards in time. It would be interesting future work to see how this theoretical insight can be implemented in lower dimensional parametrizations using for instance neural networks.

Chapter 5

We continued the thesis in Chapter 5 by transferring the observed results for finite-time MDPs in the previous chapter to infinite-time horizon MDPs. As in this scenario a stationary optimal policy is sufficient, it is not straight forwards to see if a dynamic approach can improve the convergence behaviour of vanilla PG. We introduced DynPG, carried out a convergence analysis and derived a sample complexity result under tabular softmax parametrization. It became clear that indeed the model-dependent constant of vanilla PG can be omitted. We also discussed the challenges when applying DynPG in practise and introduced DynAC and DynNPG as possible modifications.
A natural extension of this work would be to examine the practical performance of DynPG or its modifications. As in the finite-time horizon setting, it would be interesting to investigate how DynPG works under different, more complex parametrizations and also in non-tabular MDPs.

Chapter 6

In Chapter 6, we zoomed out and considered the stochastic gradient descent method independent of RL. We derived almost sure convergence rates under the global and local gradient domination assumption. Finally, we concluded by applying the local results to supervised learning with neural networks and to the previously analyzed PG methods.

Bibliography

[Aga+21] A. Agarwal, S. M. Kakade, J. D. Lee, and G. Mahajan. "On the Theory of Policy Gradient Methods: Optimality, Approximation, and Distribution Shift". In: *Journal of Machine Learning Research* 22.98 (2021), pp. 1–76.

[AGK12] M. G. Azar, V. Gómez, and H. J. Kappen. "Dynamic policy programming". In: *The Journal of Machine Learning Research* 13.1 (2012), pp. 3207–3245.

[AL24] J. An and J. Lu. "Convergence of stochastic gradient descent under a local Lojasiewicz condition for deep neural networks". In: *arXiv Preprint* arXiv:2304.09221 (2024).

[AR23] C. Alfano and P. Rebeschini. "Linear Convergence for Natural Policy Gradient with Log-linear Policy Parametrization". In: *arXiv Preprint* arXiv:2209.15382 (2023).

[Att+10] H. Attouch, J. Bolte, P. Redont, and A. Soubeyran. "Proximal Alternating Minimization and Projection Methods for Nonconvex Problems: An Approach Based on the Kurdyka-Łojasiewicz Inequality". In: *Mathematics of Operations Research* 35.2 (2010), pp. 438–457.

[Bag+03] J. Bagnell, S. M. Kakade, J. Schneider, and A. Ng. "Policy Search by Dynamic Programming". In: *Advances in Neural Information Processing Systems*. Ed. by S. Thrun, L. Saul, and B. Schölkopf. Vol. 16. MIT Press, 2003.

[Ban22] S. Banach. "Sur les opérations dans les ensembles abstraits et leur application aux équations intégrales". In: *Fundamenta mathematicae* 3.1 (1922), pp. 133–181.

[BBM18] R. Bassily, M. Belkin, and S. Ma. "On exponential convergence of SGD in non-convex over-parametrizedlearning". In: *arXiv Preprint* arXiv:1811.02564 (2018).

[BCN18] L. Bottou, F. E. Curtis, and J. Nocedal. "Optimization Methods for Large-Scale Machine Learning". In: *SIAM Review* 60.2 (2018), pp. 223–311.

[Bec17] A. Beck. *First-Order Methods in Optimization*. Philadelphia, PA: Society for Industrial and Applied Mathematics, 2017.

[Ber01] D. P. Bertsekas. *Dynamic Programming and Optimal Control. 2*. eng. 2. ed. Belmont, Mass: Athena Scientific, 2001.

[BR21] J. Bhandari and D. Russo. "On the Linear Convergence of Policy Gradient Methods for Finite MDPs". In: *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*. Vol. 130. Proceedings of Machine Learning Research. PMLR, Apr. 2021, pp. 2386–2394.

[BR22] J. Bhandari and D. Russo. "Global Optimality Guarantees For Policy Gradient Methods". In: *arXiv Preprint* arXiv:1906.01786 (2022).

[BST14] J. Bolte, S. Sabach, and M. Teboulle. "Proximal alternating linearized minimization for nonconvex and nonsmooth problems". In: *Mathematical Programming* 146.1 (2014), pp. 459–494.

[BT96a] D. Bertsekas and J. N. Tsitsiklis. *Neuro-Dynamic Programming*. Athena Scientific, 1996.

[BT96b] D. P. Bertsekas and J. N. Tsitsiklis. *Neuro-Dynamic Programming*. 1st. Athena Scientific, 1996.

[Cen+22] S. Cen, C. Cheng, Y. Chen, Y. Wei, and Y. Chi. "Fast Global Convergence of Natural Policy Gradient Methods with Entropy Regularization". In: *Operations Research* 70.4 (2022), pp. 2563–2578.

[CFR23] E. Chouzenoux, J.-B. Fest, and A. Repetti. "A Kurdyka-Lojasiewicz property for stochastic optimization algorithms in a non-convex setting". In: *arXiv Preprint* arXiv:2302.06447 (2023).

[Cha22] S. Chatterjee. *Convergence of gradient descent for deep neural networks*. 2022.

[Cla+18] I. Clavera, J. Rothfuss, J. Schulman, Y. Fujita, T. Asfour, and P. Abbeel. "Model-Based Reinforcement Learning via Meta-Policy Optimization". In: *Proceedings of The 2nd Conference on Robot Learning*. Vol. 87. Proceedings of Machine Learning Research. PMLR, Oct. 2018, pp. 617–629.

[DK21] S. Dereich and S. Kassing. "Convergence of stochastic gradient descent schemes for Lojasiewicz-landscapes". In: *arXiv Preprint* arXiv:2102.09385 (2021).

[DZL22] Y. Ding, J. Zhang, and J. Lavaei. "Beyond Exact Gradients: Convergence of Stochastic Soft-Max Policy Gradient Methods with Entropy Regularization". In: *arXiv Preprint* arXiv:2110.10117 (2022).

[DZL23] Y. Ding, J. Zhang, and J. Lavaei. "Local analysis of entropy-regularized stochastic soft-max policy gradient methods". In: *2023 European Control Conference (ECC)*. IEEE. 2023, pp. 1–8.

[Fat+22] I. Fatkhullin, J. Etesami, N. He, and N. Kiyavash. "Sharp Analysis of Stochastic Optimization under Global Kurdyka-Lojasiewicz Inequality". In: *Advances in Neural Information Processing Systems*. Vol. 35. Curran Associates, Inc., 2022, pp. 15836–15848.

[Fat+23] I. Fatkhullin, A. Barakat, A. Kireeva, and N. He. "Stochastic Policy Gradient Methods: Improved Sample Complexity for Fisher-non-degenerate Policies". In: *Proceedings of the 40th International Conference on Machine Learning*. Vol. 202. Proceedings of Machine Learning Research. PMLR, July 2023, pp. 9827–9869.

[Faz+18] M. Fazel, R. Ge, S. Kakade, and M. Mesbahi. "Global Convergence of Policy Gradient Methods for the Linear Quadratic Regulator". In: *Proceedings of the 35th International Conference on Machine Learning*. Vol. 80. Proceedings of Machine Learning Research. PMLR, July 2018, pp. 1467–1476.

[FBD21] X. Fontaine, V. D. Bortoli, and A. Durmus. "Convergence rates and approximation results for SGD and its continuous-time counterpart". In: *Proceedings of Thirty Fourth Conference on Learning Theory*. Vol. 134. Proceedings of Machine Learning Research. PMLR, Aug. 2021, pp. 1965–2058.

[FG21] S. Frei and Q. Gu. “Proxy Convexity: A Unified Framework for the Analysis of Neural Networks Trained by Gradient Descent”. In: *Advances in Neural Information Processing Systems*. Vol. 34. Curran Associates, Inc., 2021, pp. 7937–7949.

[FGJ20] B. Fehrman, B. Gess, and A. Jentzen. “Convergence Rates for the Stochastic Gradient Descent Method for Non-Convex Objective Functions”. In: *Journal of Machine Learning Research* 21.136 (2020), pp. 1–48.

[FHM18] S. Fujimoto, H. van Hoof, and D. Meger. “Addressing Function Approximation Error in Actor-Critic Methods”. In: *Proceedings of the 35th International Conference on Machine Learning*. Ed. by J. Dy and A. Krause. Vol. 80. Proceedings of Machine Learning Research. PMLR, July 2018, pp. 1587–1596.

[GB23] S. Guin and S. Bhatnagar. “A Policy Gradient Approach for Finite Horizon Constrained Markov Decision Processes”. In: *2023 62nd IEEE Conference on Decision and Control (CDC)*. 2023, pp. 3353–3359.

[GG24] G. Garrigos and R. M. Gower. “Handbook of Convergence Theorems for (Stochastic) Gradient Methods”. In: *arXiv Preprint* arXiv:2301.11235 (2024).

[GHZ22] X. Guo, A. Hu, and J. Zhang. “Theoretical Guarantees of Fictitious Discount Algorithms for Episodic Reinforcement Learning and Global Convergence of Policy Gradient Methods”. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 36.6 (June 2022), pp. 6774–6782.

[GK23] B. Gess and S. Kassing. “Convergence rates for momentum stochastic gradient descent with noise of machine learning type”. In: *arXiv Preprint* arXiv:2302.03550 (2023).

[GL13] S. Ghadimi and G. Lan. “Stochastic First- and Zeroth-Order Methods for Nonconvex Stochastic Programming”. In: *SIAM Journal on Optimization* 23.4 (2013), pp. 2341–2368.

[GPS18] S. Gadat, F. Panloup, and S. Saadane. “Stochastic heavy ball”. In: *Electronic Journal of Statistics* 12.1 (2018), pp. 461 –529.

[HGH22] F. Huang, S. Gao, and H. Huang. “Bregman Gradient Policy Optimization”. In: *International Conference on Learning Representations*. 2022.

[Hua+20] F. Huang, S. Gao, J. Pei, and H. Huang. “Momentum-Based Policy Gradient Methods”. In: *Proceedings of the 37th International Conference on Machine Learning*. Vol. 119. Proceedings of Machine Learning Research. PMLR, July 2020, pp. 4422–4433.

[HXY21] B. Hambly, R. Xu, and H. Yang. “Policy Gradient Methods for the Noisy Linear Quadratic Regulator over a Finite Horizon”. In: *SIAM Journal on Control and Optimization* 59.5 (2021), pp. 3359–3391.

[HXY23] B. Hambly, R. Xu, and H. Yang. “Policy Gradient Methods Find the Nash Equilibrium in N-player General-sum Linear-quadratic Games”. In: *Journal of Machine Learning Research* 24.139 (2023), pp. 1–56.

[JPBR23] E. Johnson, C. Pike-Burke, and P. Rebeschini. “Optimal Convergence Rate for Exact Policy Mirror Descent in Discounted Markov Decision Processes”. In: *Advances in Neural Information Processing Systems*. Vol. 36. Curran Associates, Inc., 2023, pp. 76496–76524.

[Kak01] S. M. Kakade. “A Natural Policy Gradient”. In: *Advances in Neural Information Processing Systems*. Vol. 14. MIT Press, 2001.

[Kak03] S. M. Kakade. “On the Sample Complexity of Reinforcement Learning”. PhD Thesis. University College London, 2003.

[KL02] S. Kakade and J. Langford. “Approximately Optimal Approximate Reinforcement Learning”. In: *Proceedings of the Nineteenth International Conference on Machine Learning*. Morgan Kaufmann Publishers Inc., 2002, 267–274.

[Kle+24] S. Klein, X. Zhang, T. Başar, S. Weissmann, and L. Döring. “Structure Matters: Dynamic Policy Gradient”. In: *arXiv Preprint* arXiv:2411.04913 (2024).

[KNS16] H. Karimi, J. Nutini, and M. Schmidt. “Linear Convergence of Gradient and Proximal-Gradient Methods Under the Polyak-Łojasiewicz Condition”. In: *Machine Learning and Knowledge Discovery in Databases*. Cham: Springer International Publishing, 2016, pp. 795–811.

[KR23] A. Khaled and P. Richtárik. “Better Theory for SGD in the Nonconvex World”. In: *Transactions on Machine Learning Research* (2023). Survey Certification.

[KT99] V. Konda and J. Tsitsiklis. “Actor-Critic Algorithms”. In: *Advances in Neural Information Processing Systems*. Vol. 12. MIT Press, 1999.

[KWD24] S. Klein, S. Weissmann, and L. Döring. “Beyond Stationarity: Convergence Analysis of Stochastic Softmax Policy Gradient Methods”. In: *The Twelfth International Conference on Learning Representations*. 2024.

[Lee+16] J. D. Lee, M. Simchowitz, M. I. Jordan, and B. Recht. “Gradient Descent Only Converges to Minimizers”. In: *29th Annual Conference on Learning Theory*. Vol. 49. Proceedings of Machine Learning Research. PMLR, June 2016, pp. 1246–1257.

[Lei+20] Y. Lei, T. Hu, G. Li, and K. Tang. “Stochastic Gradient Descent for Nonconvex Learning Without Bounded Gradient Assumptions”. In: *IEEE Transactions on Neural Networks and Learning Systems* 31.10 (2020), pp. 4394–4400.

[Li+23a] G. Li, Y. Wei, Y. Chi, and Y. Chen. “Softmax policy gradient methods can take exponential time to converge”. In: *Mathematical Programming* 201.1 (2023), pp. 707–802.

[Li+23b] H. Li, J. Qian, Y. Tian, A. Rakhlin, and A. Jadbabaie. “Convex and Non-convex Optimization Under Generalized Smoothness”. In: *Thirty-seventh Conference on Neural Information Processing Systems*. 2023.

[Liu+23] Z. Liu, T. D. Nguyen, T. H. Nguyen, A. Ene, and H. Nguyen. “High Probability Convergence of Stochastic Gradient Methods”. In: *Proceedings of the 40th International Conference on Machine Learning*. Vol. 202. Proceedings of Machine Learning Research. PMLR, July 2023, pp. 21884–21914.

[LLX23] Y. Liang, J. Liu, and D. Xu. “Stochastic momentum methods for non-convex learning without bounded assumptions”. In: *Neural Networks* 165 (2023), pp. 830–845.

[Loj65] S. Lojasiewicz. “Ensembles semi-analytiques”. In: *Lectures Notes IHES (Bures-sur-Yvette)* (1965).

[LY22] J. Liu and Y. Yuan. “On Almost Sure Convergence Rates of Stochastic Gradient Methods”. In: *Proceedings of Thirty Fifth Conference on Learning Theory*. Vol. 178. Proceedings of Machine Learning Research. PMLR, July 2022, pp. 2963–2983.

[LZ23] Z. Liu and Z. Zhou. “Revisiting the Last-Iterate Convergence of Stochastic Gradient Methods”. In: *arXiv Preprint* arXiv:2312.08531 (2023).

[LZB22] C. Liu, L. Zhu, and M. Belkin. “Loss landscapes and optimization in over-parameterized non-linear systems and neural networks”. In: *Applied and Computational Harmonic Analysis* 59 (2022). Special Issue on Harmonic Analysis and Machine Learning, pp. 85–116.

[MB11] E. Moulines and F. Bach. “Non-Asymptotic Analysis of Stochastic Approximation Algorithms for Machine Learning”. In: *Advances in Neural Information Processing Systems*. Vol. 24. Curran Associates, Inc., 2011.

[MDB21] L. Madden, E. Dall’Anese, and S. Becker. “High-probability Convergence Bounds for Non-convex Stochastic Gradient Descent”. In: *arXiv Preprint* arXiv:2006.05610 (2021).

[Mei+20] J. Mei, C. Xiao, C. Szepesvari, and D. Schuurmans. “On the Global Convergence Rates of Softmax Policy Gradient Methods”. In: *Proceedings of the 37th International Conference on Machine Learning*. Vol. 119. Proceedings of Machine Learning Research. PMLR, July 2020, pp. 6820–6829.

[Mei+21] J. Mei, Y. Gao, B. Dai, C. Szepesvari, and D. Schuurmans. “Leveraging Non-uniformity in First-order Non-convex Optimization”. In: *Proceedings of the 38th International Conference on Machine Learning*. Vol. 139. Proceedings of Machine Learning Research. PMLR, July 2021, pp. 7555–7564.

[Mer+20] P. Mertikopoulos, N. Hallak, A. Kavis, and V. Cevher. “On the Almost Sure Convergence of Stochastic Gradient Descent in Non-Convex Problems”. In: *Advances in Neural Information Processing Systems*. Vol. 33. Curran Associates, Inc., 2020, pp. 1117–1128.

[MJ20] V. Mai and M. Johansson. “Convergence of a Stochastic Gradient Method with Momentum for Non-Smooth Non-Convex Optimization”. In: *Proceedings of the 37th International Conference on Machine Learning*. Vol. 119. Proceedings of Machine Learning Research. PMLR, July 2020, pp. 6630–6639.

[Nac+17] O. Nachum, M. Norouzi, K. Xu, and D. Schuurmans. “Bridging the Gap Between Value and Policy Based Reinforcement Learning”. In: *Advances in Neural Information Processing Systems*. Vol. 30. Curran Associates, Inc., 2017.

[Nes13] Y. Nesterov. *Introductory Lectures on Convex Optimization*. Springer New York, NY, 2013.

[Nes83] Y. E. Nesterov. “A method of solving a convex programming problem with convergence rate $O(k^2)$”. In: *Doklady Akademii Nauk*. Vol. 269. Russian Academy of Sciences. 1983, pp. 543–547.

[Ngu+18] L. M. Nguyen, P. H. Nguyen, P. Richtárik, K. Scheinberg, M. Takác, and M. van Dijk. “New Convergence Aspects of Stochastic Gradient Algorithms”. In: *Journal of Machine Learning Research* 20 (2018), 176:1–176:49.

[Ngu+23] T. D. Nguyen, T. H. Nguyen, A. Ene, and H. L. Nguyen. “High Probability Convergence of Clipped-SGD Under Heavy-tailed Noise”. In: *arXiv Preprint* arXiv:2302.05437 (2023).

[OKL20] A. Orvieto, J. Kohler, and A. Lucchi. “The Role of Memory in Stochastic Optimization”. In: *Proceedings of The 35th Uncertainty in Artificial Intelligence Conference*. Vol. 115. Proceedings of Machine Learning Research. PMLR, July 2020, pp. 356–366.

[Pap+18] M. Papini, D. Binaghi, G. Canonaco, M. Pirotta, and M. Restelli. “Stochastic Variance-Reduced Policy Gradient”. In: *Proceedings of the 35th International Conference on Machine Learning*. Vol. 80. Proceedings of Machine Learning Research. PMLR, July 2018, pp. 4026–4035.

[Pol63] B. Polyak. “Gradient methods for the minimisation of functionals”. In: *USSR Computational Mathematics and Mathematical Physics* 3.4 (1963), pp. 864–878.

[Pol64] B. T. Polyak. “Some methods of speeding up the convergence of iteration methods”. In: *Ussr computational mathematics and mathematical physics* 4.5 (1964), pp. 1–17.

[Pol87] B. T. Polyak. *Introduction to Optimization*. Translations series in mathematics and engineering. Optimization Software, Publications Division, 1987.

[PRB13] M. Pirotta, M. Restelli, and L. Bascetta. “Adaptive Step-Size for Policy Gradient Methods”. In: *Advances in Neural Information Processing Systems*. Vol. 26. Curran Associates, Inc., 2013.

[Put05] M. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, 2005.

[QMM24] J. Qiu, B. Ma, and A. Milzarek. “Convergence of SGD with momentum in the nonconvex case: A time window-based analysis”. In: *arXiv Preprint* arxiv:2405.16954 (2024).

[RM51] H. Robbins and S. Monro. “A Stochastic Approximation Method”. In: *The Annals of Mathematical Statistics* 22.3 (1951), pp. 400–407.

[RS71] H. Robbins and D. Siegmund. “A CONVERGENCE THEOREM FOR NON NEGATIVE ALMOST SUPERMARTINGALES AND SOME APPLICATIONS”. In: *Optimizing Methods in Statistics*. Academic Press, 1971, pp. 233–257.

[SB18] R. S. Sutton and A. G. Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

[Sch14] B. Scherrer. “Approximate Policy Iteration Schemes: A Comparison”. In: *Proceedings of the 31st International Conference on Machine Learning*. Ed. by E. P. Xing and T. Jebara. Vol. 32. Proceedings of Machine Learning Research. Bejing, China: PMLR, June 2014, pp. 1314–1322.

[Sch+15a] B. Scherrer, M. Ghavamzadeh, V. Gabillon, B. Lesner, and M. Geist. “Approximate modified policy iteration and its application to the game of Tetris.” In: *J. Mach. Learn. Res.* 16.49 (2015), pp. 1629–1676.

[Sch+15b] J. Schulman, S. Levine, P. Abbeel, M. Jordan, and P. Moritz. “Trust Region Policy Optimization”. In: *Proceedings of the 32nd International Conference on Machine Learning*. Vol. 37. Proceedings of Machine Learning Research. Lille, France: PMLR, July 2015, pp. 1889–1897.

[Sch+17] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov. “Proximal policy optimization algorithms”. In: *arXiv preprint arXiv:1707.06347* (2017).

[SGD21] O. Sebbouh, R. M. Gower, and A. Defazio. “Almost sure convergence rates for Stochastic Gradient Descent and Stochastic Heavy Ball”. In: *Proceedings of Thirty Fourth Conference on Learning Theory*. Vol. 134. Proceedings of Machine Learning Research. PMLR, Aug. 2021, pp. 3935–3971.

[She+19] Z. Shen, A. Ribeiro, H. Hassani, H. Qian, and C. Mi. “Hessian Aided Policy Gradient”. In: *Proceedings of the 36th International Conference on Machine Learning*. Vol. 97. Proceedings of Machine Learning Research. PMLR, June 2019, pp. 5729–5738.

[SMS08] R. S. Sutton, H. Maei, and C. Szepesvári. “A Convergent O(n) Temporal-difference Algorithm for Off-policy Learning with Linear Function Approximation”. In: *Advances in Neural Information Processing Systems*. Ed. by D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou. Vol. 21. Curran Associates, Inc., 2008.

[SMS22] K. Scaman, C. Malherbe, and L. D. Santos. “Convergence Rates of Non-Convex Stochastic Gradient Descent Under a Generic Lojasiewicz Condition and Local Smoothness”. In: *Proceedings of the 39th International Conference on Machine Learning*. Vol. 162. Proceedings of Machine Learning Research. PMLR, July 2022, pp. 19310–19327.

[Sut+09] R. S. Sutton, H. R. Maei, D. Precup, S. Bhatnagar, D. Silver, C. Szepesvári, and E. Wiewiora. “Fast gradient-descent methods for temporal-difference learning with linear function approximation”. In: *Proceedings of the 26th annual international conference on machine learning*. 2009, pp. 993–1000.

[Sut+99] R. S. Sutton, D. McAllester, S. Singh, and Y. Mansour. “Policy Gradient Methods for Reinforcement Learning with Function Approximation”. In: *Advances in Neural Information Processing Systems*. Vol. 12. MIT Press, 1999.

[Wan+21] H. Wang, M. Gurbuzbalaban, L. Zhu, U. Simsekli, and M. A. Erdogdu. “Convergence Rates of Stochastic Gradient Descent under Infinite Noise Variance”. In: *Advances in Neural Information Processing Systems*. Vol. 34. Curran Associates, Inc., 2021, pp. 18866–18877.

[Wei+24] S. Weissmann, S. Klein, W. Azizian, and L. Döring. “Almost sure convergence rates of stochastic gradient methods under gradient domination”. In: *arXiv preprint arXiv:2405.13592* (2024).

[Wil92] R. J. Williams. “Simple statistical gradient-following algorithms for connectionist reinforcement learning”. In: *Machine Learning* 8.3 (1992), pp. 229–256.

[WMW19] A. C. Wilson, L. Mackey, and A. Wibisono. “Accelerating Rescaled Gradient Descent: Fast Optimization of Smooth Functions”. In: *Advances in Neural Information Processing Systems*. Vol. 32. Curran Associates, Inc., 2019.

[Woj23] S. Wojtowytsch. “Stochastic Gradient Descent with Noise of Machine Learning Type Part I: Discrete Time Analysis”. In: *Journal of Nonlinear Science* 33.3 (2023), p. 45.

[WWZ22] S. Weissmann, A. Wilson, and J. Zech. “Multilevel Optimization for Inverse Problems”. In: *Proceedings of Thirty Fifth Conference on Learning Theory*. Ed. by P.-L. Loh and M. Raginsky. Vol. 178. Proceedings of Machine Learning Research. PMLR, July 2022, pp. 5489–5524.

[XGG20a] P. Xu, F. Gao, and Q. Gu. “An Improved Convergence Analysis of Stochastic Variance-Reduced Policy Gradient”. In: *Proceedings of The 35th Uncertainty in Artificial Intelligence Conference*. Vol. 115. Proceedings of Machine Learning Research. PMLR, July 2020, pp. 541–551.

[XGG20b] P. Xu, F. Gao, and Q. Gu. “Sample Efficient Policy Gradient Methods with Recursive Variance Reduction”. In: *International Conference on Learning Representations*. 2020.

[Xia22] L. Xiao. “On the Convergence Rates of Policy Gradient Methods”. In: *Journal of Machine Learning Research* 23.282 (2022), pp. 1–36.

[Yan+18] Y. Yan, T. Yang, Z. Li, Q. Lin, and Y. Yang. “A Unified Analysis of Stochastic Momentum Methods for Deep Learning”. In: *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*. International Joint Conferences on Artificial Intelligence Organization, July 2018, pp. 2955–2961.

[YFL23] P. Yue, C. Fang, and Z. Lin. “On the Lower Bound of Minimizing Polyak-Łojasiewicz functions”. In: *Proceedings of Thirty Sixth Conference on Learning Theory*. Ed. by G. Neu and L. Rosasco. Vol. 195. Proceedings of Machine Learning Research. PMLR, July 2023, pp. 2948–2968.

[YGL22] R. Yuan, R. M. Gower, and A. Lazaric. “A general sample complexity analysis of vanilla policy gradient”. In: *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*. Ed. by G. Camps-Valls, F. J. R. Ruiz, and I. Valera. Vol. 151. Proceedings of Machine Learning Research. PMLR, Mar. 2022, pp. 3332–3380.

[YLL16] T. Yang, Q. Lin, and Z. Li. “Unified convergence analysis of stochastic momentum methods for convex and non-convex optimization”. In: *arXiv Preprint* arXiv:1604.03257 (2016).

[ZB23] X. Zhang and T. Başar. “Revisiting LQR Control From the Perspective of Receding-Horizon Policy Gradient”. In: *IEEE Control Systems Letters* 7 (2023), pp. 1664–1669.

[Zha+20] K. Zhang, A. Koppel, H. Zhu, and T. Başar. "Global Convergence of Policy Gradient Methods to (Almost) Locally Optimal Policies". In: *SIAM Journal on Control and Optimization* 58.6 (2020), pp. 3586–3612.

[Zha+23] X. Zhang, S. Mowlavi, M. Benosman, and T. Başar. "Global Convergence of Receding-Horizon Policy Search in Learning Estimator Designs". In: *arXiv Preprint* arXiv:2309.04831 (2023).

[ZHB23] X. Zhang, B. Hu, and T. Başar. "Learning the Kalman Filter with Fine-Grained Sample Complexity". In: *2023 American Control Conference (ACC)*. 2023, pp. 4549–4554.

[Zho+20] B. Zhou, J. Liu, W. Sun, R. Chen, C. Tomlin, and Y. Yuan. "pbSGD: Powered Stochastic Gradient Descent Methods for Accelerated Non-Convex Optimization". In: *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*. International Joint Conferences on Artificial Intelligence Organization, July 2020, pp. 3258–3266.

[ZWL18] Y. Zhou, Z. Wang, and Y. Liang. "Convergence of Cubic Regularization for Nonconvex Optimization under KL Property". In: *Advances in Neural Information Processing Systems*. Vol. 31. Curran Associates, Inc., 2018.

Asymptotic convergence of FT-SimPG

A

We prove asymptotic convergence of FT-SimPG under softmax parametrization towards the global optimum. This result is used in the proof of Lemma 4.14.[1]
We use the extended notation of the state value, state-action value and advantage function introduced in Remark 4.8. For the rest of the section, we will write θ instead of Θ, $\theta_n = \Theta^{(n)}$ and further $J(\theta)$ or J^* instead of $J(\theta, \mu)$ or $J^*(\mu)$ to save notation.

Theorem A.1. *Let μ be a probability measure such that $\mu(s) > 0$ for all $s \in \mathcal{S}$ and let $0 < \eta \leq \frac{1}{5H^2R^*}$. Consider the sequence $(\theta^{(n)})$ generated by Algorithm 4 for arbitrary $\theta^{(0)} \in \mathbb{R}^{\sum_h d_h}$. Then, for all $s \in \mathcal{S}^{[\mathcal{H}]}$ we have $V^{\pi^{\theta^{(n)}}}(s) \to V^*(s)$ as $n \to \infty$. Especially we have $V_0^{\pi^{\theta^{(n)}}}(s) \to V_0^*(s)$ as $n \to \infty$ for all $s \in \mathcal{S}_0$.*

Before we can proof this result we have to prove a row of lemmata. The outline follows the proof of Agarwal et al. [Aga+21, Theorem 5].

Lemma A.2 (Monotonicity). *If the learning rate satisfies $0 < \eta \leq \frac{1}{H^2R^*5} \leq \frac{1}{H^2R^*\left(2-\frac{1}{|\mathcal{A}|}\right)} = \frac{1}{\beta}$ then $V^{\pi^{\theta_{n+1}}}(s) \geq V^{\pi^{\theta_n}}(s)$ and $Q^{\pi^{\theta_{n+1}}}(s,a) \geq Q^{\pi^{\theta_n}}(s,a)$ for all $s \in \mathcal{S}^{[\mathcal{H}]}$ and all $a \in \mathcal{A}$. Furthermore, there exist limits $V^\infty(s)$ and $Q^\infty(s,a)$ such that*

$$\lim_{n\to\infty} V^{\pi^{\theta_n}}(s) = V^\infty(s) < \infty.$$
$$\lim_{n\to\infty} Q^{\pi^{\theta_n}}(s,a) = Q^\infty(s,a) < \infty.$$

Proof. We will show that $V_h^{\pi^{\theta_n}}(s) \leq V_h^{\pi^{\theta_{n+1}}}(s)$ for each state $s \in \mathcal{S}$ (in the not enlarged state space) and each epoch h. Then by the bounded reward assumption there exists $V_h^\infty(s)$ such that $V_h^{\pi^{\theta_n}}(s) \to V_h^\infty$ for $n \to \infty$. If this holds true we see the mononicity and convergence of the Q-functions from the relation

$$Q_h^{\pi^\theta}(s,a) = r(s,a) + \sum_{s' \in \mathcal{S}} p(s'|s,a) V_{h+1}^{\pi^\theta}(s'),$$

with $V_H \equiv 0$.
In order to show the claim we first see from the performance difference lemma, that

$$\begin{aligned} V_h^{\pi^{\theta_{n+1}}}(s) - V_h^{\pi^{\theta_n}}(s) &= \mathbb{E}_{S_h=s}^{\pi^{\theta_{n+1}}}\left[\sum_{t=h}^{H-1} A_t^{\pi^{\theta_n}}(S_t, A_t)\right] \\ &= \sum_{s_l \in \mathcal{S}^{[\mathcal{H}]}} \widetilde{\rho}_{s,h}^{\pi^{\theta_{n+1}}}(s_l) \sum_{a \in \mathcal{A}} \pi^{\theta_{n+1}}(a|s_l) A^{\pi^{\theta_n}}(s_l, a), \end{aligned}$$

where $\widetilde{\rho}_{s,h}^{\pi^{\theta_{n+1}}}(s_l) := \sum_{t=h}^{H-1} \mathbb{P}_{S_h=s}^{\pi^{\theta_{n+1}}}(S_t = s_l)$ the state visitation measure from epoch h to $H-1$ on the enlarged state space $\mathcal{S}^{[\mathcal{H}]}$. Note that $\widetilde{\rho}_{s,h}^{\pi^{\theta_{n+1}}}(s_l) = 0$ for $l < h$, as we cannot visit states from previous epochs.

[1] This chapter contains the results in [KWD24, App. C].

We will prove that $\sum_{a\in\mathcal{A}} \pi^{\theta_{n+1}}(a|s)A^{\pi^{\theta_n}}(s,a) \geq \sum_{a\in\mathcal{A}} \pi^{\theta_n}(a|s)A^{\pi^{\theta_n}}(s,a)$, for any $s \in \mathcal{S}^{[\mathcal{H}]}$. Then the fact that $\sum_{a\in\mathcal{A}} \pi^{\theta_n}(a|s)A^{\pi^{\theta_n}}(s,a) = 0$ leads to the desired result.
Therefore, we consider the function

$$F_s(\theta^s) := \sum_{a\in\mathcal{A}} \pi^{\theta^s}(a|s)c(s,a) \quad s \in \mathcal{S}^{[\mathcal{H}]},$$

for $\theta^s = (\theta(s,a))_{a\in\mathcal{A}} \in \mathbb{R}^{|\mathcal{A}|}$. We will set $c(s,a) = A_h^{\pi^{\theta_n}}(s,a)$, but for θ_n fix, i.e. the following derivatives with respect to θ^s of F are independent of $A^{\pi^{\theta_n}}$. From [Aga+21] Lemma C.2 we know that

$$\frac{\partial F_s(\theta^s)}{\partial\theta(s,a)}\Big|_{\theta_n^s} = \pi^{\theta_s^{(n)}}(a|s)A_h^{\pi^{\theta_n}}(s,a). \tag{A.1}$$

Furthermore, $F_s(\theta_s)$ is $5HR^*$-smooth for every s by Lemma D.1 in [Aga+21] and the bounded reward assumption. Considering our gradient ascent updates from simultaneous training we get

$$\begin{aligned}\theta_{n+1}(s,a) &= \theta_n(s,a) + \eta\frac{\partial V^{\pi^{\theta_n}}(\mu)}{\partial\theta_n(s,a)} &\text{(A.2)}\\ &= \theta_n(s,a) + \eta\widetilde{\rho}_\mu^{\pi^{\theta_n}}(s)\pi^{\theta_n}(a|s)A^{\pi^{\theta_n}}(s,a) &\text{(A.3)}\\ &= \theta_n(s,a) + \eta\widetilde{\rho}_\mu^{\pi^{\theta_n}}(s)\frac{\partial F_s(\theta_s)}{\partial\theta_n(s,a)}\Big|_{\theta_n^s}. &\text{(A.4)}\end{aligned}$$

As $\eta\widetilde{\rho}_\mu^{\pi^{\theta_n}}(s) = \eta H d_\mu^{\pi^{\theta_n}}(s)$ and $d_\mu^{\pi^{\theta_n}}(s)$ a probability measure we see that $\eta\widetilde{\rho}_\mu^{\pi^{\theta_n}}(s) \leq \frac{1}{5HR^*}$ by our choice of $\eta \leq \frac{1}{H^2R^*5}$. Then the descent lemma for the $5HR^*$-smooth function F_s gives the desired inequality

$$\sum_{a\in\mathcal{A}} \pi^{\theta_{n+1}}(a|s)A^{\pi^{\theta_n}}(s,a) \geq \sum_{a\in\mathcal{A}} \pi^{\theta_n}(a|s)A^{\pi^{\theta_n}}(s,a).$$

■

Remark A.3. We want to point out that the proof of Lemma A.2 is crucial for the choice of the step size in the convergence analysis of the simultaneous PG algorithm. As we can only use the descent lemma for a step size $0 < \eta \leq \frac{1}{5H^2R^*}$, we can only achieve asymptotic convergence towards global minima under this assumption. Hence, we also need this step size requirement in the convergence analysis.

We introduce the following definitions:

$$\Delta = \min_{\{(s,a)\in\mathcal{S}^{[\mathcal{H}]}\times\mathcal{A} : A^\infty(s,a)\neq 0\}} |A^\infty(s,a)|$$

where $A^\infty(s,a) = Q^\infty(s,a) - V^\infty(s)$.
We define the sets for each $s \in \mathcal{S}^{[\mathcal{H}]}$:

$$\begin{aligned}I_0^s &= \{a \in \mathcal{A} \mid Q^\infty(s,a) = V^\infty(s)\},\\ I_+^s &= \{a \in \mathcal{A} \mid Q^\infty(s,a) > V^\infty(s)\},\\ I_-^s &= \{a \in \mathcal{A} \mid Q^\infty(s,a) < V^\infty(s)\}.\end{aligned}$$

We aim to prove that I_+^s is an empty set, then $V^\infty(s) = V^*(s)$ the optimal value function (epoch wise true).

LEMMA A.4. *There exists a time $N_1 > 0$ such that for all $n > N_1$, and $s \in \mathcal{S}^{[\mathcal{H}]}$, we have*

$$A^{\theta_n}(s,a) < -\frac{\Delta}{4} \text{ for } a \in I_-^s; \quad A^{\theta_n}(s,a) > \frac{\Delta}{4} \text{ for } a \in I_+^s.$$

Proof. Fix $s \in \mathcal{S}^{[\mathcal{H}]}$ arbitrarily. As $V^{\pi^{\theta_n}}(s) \to V^{\infty}(s)$ for $n \to \infty$ and $\mathcal{S}$ is finite, we have that there exists $N_1 > 0$ such that for all $n > N_1$ and $s \in \mathcal{S}^{[\mathcal{H}]}$,

$$V^{\pi^{\theta_n}}(s) > V^{\infty}(s) - \frac{\Delta}{4}.$$

It follows for all $n > N_1$, $s \in \mathcal{S}^{[\mathcal{H}]}$ and $a \in I_-^s$ by the definition of Δ:

$$A^{\theta_n}(s,a) = Q^{\theta_n}(s,a) - V^{\pi^{\theta_n}}(s) \leq Q^{\infty}(s,a) - V^{\infty}(s) + \frac{\Delta}{4} \leq -\Delta + \frac{\Delta}{4} < -\frac{\Delta}{4}.$$

Similarly, for all $n > N_1$, $s \in \mathcal{S}^{[\mathcal{H}]}$ and $a \in I_+^s$ we obtain from monotonicity Lemma A.2 and the definition of Δ,

$$A_h^{\theta_n}(s,a) = Q^{\theta_n}(s,a) - V^{\pi^{\theta_n}}(s) \geq Q^{\infty}(s,a) - \frac{\Delta}{4} - V^{\infty}(s) \geq \Delta - \frac{\Delta}{4} > \frac{\Delta}{4}.$$

■

LEMMA A.5. *It holds that $\frac{\partial J(\theta_n)}{\partial \theta_n(s,a)} \to 0$ as $n \to \infty$ for all $s \in \mathcal{S}^{[\mathcal{H}]}$, $a \in \mathcal{A}_s$. This implies that for $a \in I_+^s \cup I_-^s$, $\pi^{\theta_n}(a|s) \to 0$ and that $\sum_{a \in I_0^s} \pi^{\theta_n}(a|s) \to 1$ for $n \to \infty$.*

Proof. From [Bec17, Theorem 10.15] we deduce for any β-smooth function $f : \mathbb{R}^d \to \mathbb{R}$, that $\|\nabla f(x^k)\| \to 0$ for $k \to \infty$, if $x^{k+1} = x^k - \eta \nabla f(x^k)$, when $\eta < \frac{1}{\beta}$. By Lemma 4.10 $J(\cdot)$ is $H^2R^*(2 - \frac{1}{|\mathcal{A}|})$-smooth. It follows by our choice of $\eta < \frac{1}{5H^2R^*}$ that $\frac{\partial J(\theta_n)}{\partial \theta_n(s,a)} \to 0$ as $n \to \infty$ for all $s \in \mathcal{S}^{[\mathcal{H}]}$, $a \in \mathcal{A}_s$. Now remember the derivative of the softmax parametrization in the stationary case

$$\frac{\partial J(\theta_n)}{\partial \theta_n(s,a)} = \widetilde{\rho}_{\mu}^{\pi^{\theta_n}}(s) \pi^{\theta_n}(a|s) A^{\theta_n}(s,a),$$

and by Lemma A.4 $|A^{\theta_n}(s,a)| > \frac{\Delta}{4}$ for all $n > N_1$ and $a \in I_+^s \cup I_-^s$. As $\widetilde{\rho}_{\mu}^{\pi^{\theta_n}}(s) > 0$ by assumption on μ and the positivity of the softmax parametrization. It follows that $\pi^{\theta_n}(a|s) \to 0$ for $n \to \infty$ for all $a \in I_+^s \cup I_-^s$ from $\frac{\partial J(\theta_n)}{\partial \theta_n(s,a)} \to 0$ as $n \to \infty$.
The last claim, $\sum_{a \in I_0^s} \pi^{\theta_n}(a|s) \to 1$ for $n \to \infty$, follows immediately from $\sum_{a \in \mathcal{A}_s} \pi^{\theta_n}(a|s) = 1$ by:

$$\begin{aligned}
\lim_{n \to \infty} \sum_{a \in I_0^s} \pi^{\theta_n}(a|s) &= \lim_{n \to \infty} \Big(\sum_{a \in \mathcal{A}} \pi^{\theta_n}(a|s) - \sum_{a \in I_+^s \cup I_-^s} \pi^{\theta_n}(a|s) \Big) \\
&= 1 - \sum_{a \in I_+^s \cup I_-^s} \lim_{n \to \infty} \pi^{\theta_n}(a|s) \\
&= 1.
\end{aligned}$$

■

LEMMA A.6. *For $a \in I_+^s$, the sequence $(\theta_n(s,a))_{n\geq 0}$ is strictly increasing for $n > N_1$ and for $a \in I_-^s$, the sequence $(\theta_n(s,a))_{n\geq 0}$ is strictly decreasing for $n > N_1$.*

Proof. With Lemma A.4 we know that for $n > N_1$

$$A_h^{\theta_n}(s,a) > 0 \text{ for } a \in I_+^s; \quad A_h^{\theta_n}(s,a) < 0 \text{ for } a \in I_-^s,$$

and by the derivative of the value function

$$\frac{\partial J(\theta_n)}{\partial \theta_n(s,a)} = \widetilde{\rho}_\mu^{\pi^{\theta_n}}(s)\pi^{\theta_n}(a|s)A_h^{\theta_n}(s,a).$$

As $\widetilde{\rho}_\mu^{\pi^{\theta_n}}(s) > 0$ by the assumption $\mu(s) > 0$ and the positivity of the softmax parametrization, we have for all $n > N_1$

$$\frac{\partial J(\theta_n)}{\partial \theta_n(s,a)} > 0 \text{ for } a \in I_+^s; \quad \frac{\partial J(\theta_n)}{\partial \theta_n(s,a)} < 0 \text{ for } a \in I_-^s.$$

This implies for $a \in I_+^s$,

$$\theta_{n+1}(s,a) - \theta_n(s,a) = \eta \frac{\partial J(\theta_n)}{\partial \theta(s,a)} > 0,$$

i.e. $(\theta_n(s,a))_{n\geq 0}$ is strictly increasing for $n > N_1$ and similar for $a \in I_-^s$,

$$\theta_{n+1}(s,a) - \theta_n(s,a) = \eta \frac{\partial J(\theta_n)}{\partial \theta_n(s,a)} < 0,$$

i.e. $(\theta_n(s,a))_{n\geq 0}$ is strictly decreasing for $n > N_1$. ∎

LEMMA A.7. *For all $s \in \mathcal{S}^{[\mathcal{H}]}$ where $I_+^s \neq \emptyset$, we have that*

$$\max_{a\in I_0^s} \theta_n(s,a) \to \infty \quad \textit{and} \quad \min_{a\in\mathcal{A}} \theta_n(s,a) \to -\infty \quad \textit{for } n \to \infty.$$

Proof. By assumption $I_+^s \neq \emptyset$ there exists an $a_+ \in I_+^s$ and by Lemma A.5 we have $\pi^{\theta_n}(a_+|s) \to 0$, as $n \to \infty$. Hence, by softmax parametrization this is equivalent to

$$\frac{\exp(\theta_n(s,a_+))}{\sum_{a\in\mathcal{A}} \exp(\theta_n(s,a))} \to 0, \text{ for } n \to \infty.$$

Using Lemma A.6, i.e. $\theta_n(s,a_+)$ is strictly increasing for $n > N_1$, we imply that $\exp(\theta_n(s,a_+))$ is strictly increasing for $n > N_1$. This implies that

$$\sum_{a\in\mathcal{A}} \exp(\theta_n(s,a)) \to \infty, \text{ for } n \to \infty.$$

Again by Lemma A.5 we know that

$$\sum_{a\in I_0^s} \pi^{\theta_n}(a|s) \to 1, \text{ for } n \to \infty,$$

i.e. by definition

$$\sum_{a\in I_0^s} \frac{\exp(\theta_n(s,a))}{\sum\limits_{a'\in\mathcal{A}} \exp(\theta_n(s,a'))} \to 1, \text{ for } n\to\infty.$$

As $\sum\limits_{a'\in\mathcal{A}} \exp(\theta_n(s,a')) \to \infty$ it follows that

$$\sum_{a\in I_0^s} \exp(\theta_n(s,a)) \to \infty, \text{ for } n\to\infty$$

implying

$$\max_{a\in I_0^s} \theta_n(s,a) \to \infty, \text{ for } n\to\infty.$$

For the second claim it holds that

$$\begin{aligned}
\sum_{a\in\mathcal{A}} \frac{\partial J(\theta_n)}{\partial\theta_n(s,a)} &= \widetilde{\rho}_\mu^{\pi^{\theta_n}}(s) \sum_{a\in\mathcal{A}} \pi^{\theta_n}(a|s)(Q_h^{\pi^{\theta_n}}(s,a) - V_h^{\pi^{\theta_n}}(s)) \\
&= \widetilde{\rho}_\mu^{\pi^{\theta_n}}(s)(\mathbb{E}_{S_h=s}^{\pi^{\theta_n}}[Q_h^{\pi^{\theta_n}}(s,a)] - V_h^{\pi^{\theta_n}}(s)) \\
&= \widetilde{\rho}_\mu^{\pi^{\theta_n}}(s)(V_h^{\pi^{\theta_n}}(s) - V_h^{\pi^{\theta_n}}(s)) \\
&= 0.
\end{aligned}$$

By induction, we obtain $\sum_{a\in\mathcal{A}} \theta_n(s,a) = \sum_{a\in\mathcal{A}} \theta_0(s,a) := c$ for every $n > 0$ and hence

$$\min_{a\in\mathcal{A}} \theta_n(s,a) < \sum_{a\in\mathcal{A}} \theta_n(s,a) - \max_{a\in\mathcal{A}} \theta_n(s,a) = -\max_{a\in\mathcal{A}} \theta_n(s,a) + c.$$

Since $\max_{a\in\mathcal{A}} \theta_n(s,a) \to \infty$, because $\max_{a\in I_0^s} \theta_n(s,a) \to \infty$, we conclude $\min_{a\in\mathcal{A}} \theta_n(s,a) \to -\infty$ for $n\to\infty$. ∎

LEMMA A.8. *Suppose $a_+ \in I_+^s$. If there exists $a \in I_0^s$ such that for some $n > N_1$, $\pi^{\theta_n}(a|s) \le \pi^{\theta_n}(a_+|s)$, then for all $m > n$ it holds that $\pi^{\theta_m}(a|s) \le \pi^{\theta_m}(a_+|s)$.*

Proof. Suppose there exists $a \in I_0^s$ such that for an $n > 0$, $\pi^{\theta_n}(a|s) \le \pi^{\theta_n}(a_+|s)$. We show that $\pi^{\theta_{n+1}}(a|s) \le \pi^{\theta_{n+1}}(a_+|s)$, then the claim follows by induction. We have

$$\begin{aligned}
\frac{\partial J_h(\theta_n)}{\partial\theta_n(s,a)} &= \widetilde{\rho}_\mu^{\pi^{\theta_n}}(s)\pi^{\theta_n}(a|s)(Q_h^{\pi^{\theta_n}}(s,a) - V_h^{\pi^{\theta_n}}(s)) \\
&\le \widetilde{\rho}_\mu^{\pi^{\theta_n}}(s)\pi^{\theta_n}(a_+|s)(Q_h^{\pi^{\theta_n}}(s,a_+) - V_h^{\pi^{\theta_n}}(s)) \\
&= \frac{\partial J(\theta_n)}{\partial\theta_n(s,a_+)},
\end{aligned}$$

where the inequality follows with

$$\begin{aligned}
Q_h^{\pi^{\theta_n}}(s,a_+) &\ge Q_h^{\infty}(s,a_+) - \frac{\Delta}{4} \\
&\ge Q_h^{\infty}(s,a) + \Delta - \frac{\Delta}{4}
\end{aligned}$$

$$> Q_h^{\pi^{\theta_n}}(s,a).$$

The first inequaility is due to Lemma A.4 and the second by the definition of Δ and $a \in I_0^s$. Now by assumption we have $\pi^{\theta_n}(a|s) \leq \pi^{\theta_n}(a_+|s)$ and thus $\theta_n(s,a) \leq \theta_n(s,a_+)$. It follows

$$\theta_{n+1}(s,a) = \theta_n(s,a) + \eta \frac{\partial J(\theta_n)}{\partial \theta_n(s,a)} \leq \theta(s,a_+) + \eta \frac{\partial J(\theta_n)}{\partial \theta_n(s,a_+)} = \theta_{n+1}(s,a_+).$$

■

Now define for every $a_+ \in I_+^s$ the set

$$B_0^s(a_+) = \{a \in I_0^s | \pi^{\theta_n}(a_+|s) \leq \pi^{\theta_n}(a|s) \text{ for all } l > 0\}$$

and denote its complement in I_0^s as $\bar{B}_0^s(a_+) = I_0^s \setminus B_0^s(a_+)$.

LEMMA A.9. *Suppose $I_+^s \neq \emptyset$. For all $a_+ \in I_+^s$, we have that $B_0^s(a_+) \neq \emptyset$ and*

$$\sum_{a \in B_0^s(a_+)} \pi^{\theta_n}(a|s) \to 1, \text{ as } n \to \infty.$$

This implies:

$$\max_{a \in B_0^s(a_+)} \theta_n(s,a) \to \infty, \text{ for } n \to \infty.$$

Proof. Let $a_+ \in I_+^s$ and consider $a \in \bar{B}_0^s(a_+)$. Then by definition of $\bar{B}_0^s(a_+)$ there exists $n' > N_1$ such that $\pi^{\theta_{n'}}(a_+|s) \geq \pi^{\theta_{n'}}(a|s)$. Hence, by Lemma A.8 for all $n \geq n'$ we have $\pi^{\theta_n}(a_+|s) \geq \pi^{\theta_n}(a|s)$. As $\pi^{\theta_n}(a_+|s) \to 0$ for $n \to \infty$. We obtain $\pi^{\theta_n}(a|s) \to 0$ for $n \to \infty$, for all $a \in \bar{B}_0^s(a_+)$. Since by Lemma A.5 $\sum_{a \in I_0^s} \pi^{\theta_n}(a|s) \to 1$ for $n \to \infty$, we have that $B_0^s(a_+) \neq \emptyset$ and that $\sum_{a \in B_0^s(a_+)} \pi^{\theta_n}(a|s) \to 1$, as $n \to \infty$. The second claim follows from this as in Lemma A.7. ■

LEMMA A.10. *Consider $s \in \mathcal{S} \times \mathcal{H}$ such that $I_+^s \neq \emptyset$. Then, for any $a_+ \in I_+^s$, there exists an N_{a_+} such that for all $n > N_{a_+}$ we have*

$$\pi^{\theta_n}(a_+|s) > \pi^{\theta_n}(a|s) \text{ for all } a \in \bar{B}_0^s(a_+).$$

Proof. For every $a \in \bar{B}_0^s(a_+)$ exists time n_a such that

$$\pi^{\theta_n}(a_+|s) > \pi^{\theta_n}(a|s) \text{ for all } a \in \bar{B}_0^s(a_+)$$

for all $n > n_a$ by definition. Set $N_{a_+} = \max_{a \in \bar{B}_0^s(a_+)} n_a$ and the proof is completed. ■

LEMMA A.11. *Assume again $I_+^s \neq \emptyset$. For all actions $a \in I_+^s$, we have that $\theta_n(s,a)$ is bounded from below as $n \to \infty$. And for all $a \in I_-^s$, we have that $\theta_n(s,a) \to -\infty$ as $n \to \infty$.*

Proof. The first claim follows directly with Lemma A.6 as $\theta_n(s,a)$ is strictly increasing for all $a \in I_+^s$, $n > N_1$, and thus for all $n > N_1$ we have $\theta_n(s,a) \geq \theta_{N_1}(s,a)$. Now suppose $a \in I_-^s$, then by Lemma A.6 we have that $\theta_n(s,a)$ is strictly decreasing for $n > N_1$. Assume there exists b such that $\lim_{n \to \infty} \theta_n(s,a) = b$, then $\theta_n(s,a) > b$ for all $n > N_1$. By Lemma A.7 there exists an action

$a' \in \mathcal{A}$ such that $\theta_n(s,a') \to -\infty$ for $n \to \infty$. Consider $\delta > 0$ such that $\theta_{N_1}(s,a') \geq b - \delta$. Define for all $n > N_1$

$$\tau(n) = \max\{k \in (N_1, n] : \theta_k(s,a') \geq b - \delta\}.$$

Define also

$$\mathcal{T}^{(n)} = \left\{\tau(n) < n' < n : \frac{\partial J(\theta_{n'})}{\partial \theta_{n'}(s,a')} \leq 0\right\},$$

as the set of all indices n' in $(\tau(n), n)$, where $\theta_{n'}(s,a')$ is decreasing. Next we define $Z_n := \sum_{n' \in \mathcal{T}^{(n)}} \frac{\partial J(\theta_{n'})}{\partial \theta_{n'}(s,a')}$, then it holds that

$$\begin{aligned} Z_n &= \sum_{n' \in \mathcal{T}^{(n)}} \frac{\partial J(\theta_{n'})}{\partial \theta_{n'}(s,a')} \\ &\leq \sum_{n'=\tau(n)+1}^{n-1} \frac{\partial J(\theta_{n'})}{\partial \theta_{n'}(s,a')} \\ &\leq \sum_{n'=\tau(n)}^{n-1} \frac{\partial J(\theta_{n'})}{\partial \theta_{n'}(s,a')} + \left|\frac{\partial J(\theta_{\tau(n)})}{\partial \theta_{\tau(n)}(s,a')}\right|. \end{aligned}$$

By Lemma 4.9 and the bounded reward assumption we have

$$\left|\frac{\partial J(\theta_{\tau(n)})}{\partial \theta_{\tau(n)}(s,a')}\right| = \widetilde{\rho}_\mu^{\pi^{\theta_{\tau(n)}}}(s)\pi^{\theta_{\tau(n)}}(a'|s)|A_h^{\theta_{\tau(n)}}(s,a')| \leq H^2 R^*.$$

Hence,

$$\begin{aligned} Z_n &\leq \sum_{n'=\tau(n)}^{n-1} \frac{\partial J(\theta_{n'})}{\partial \theta_{n'}(s,a')} + H^2 R^* \\ &= \frac{1}{\eta}(\theta_n(s,a') - \theta_{\tau(n)}(s,a')) + H^2 R^* \\ &\leq \frac{1}{\eta}(\theta_n(s,a') - b + \delta) + H^2 R^*. \end{aligned}$$

Then $\theta_n(s,a') \to -\infty$ for $n \to \infty$ implies that $Z_n \to -\infty$ for $n \to \infty$. As we chose $a \in I_-^s$ it holds that $|A_h^{\theta_n}(s,a)| \geq \frac{\Delta}{4}$ for $n > N_1$ with Lemma A.4 and so for all $n' \in \mathcal{T}^{(n)}$:

$$\begin{aligned} \left|\frac{\frac{\partial J(\theta_{n'})}{\partial \theta_{n'}(s,a)}}{\frac{\partial J(\theta_{n'})}{\partial \theta_{n'}(s,a')}}\right| &= \left|\frac{\pi^{\theta_{n'}}(a|s)A_h^{\theta_{n'}}(s,a)}{\pi^{\theta_{n'}}(a'|s)A_h^{\theta_{n'}}(s,a')}\right| \\ &\geq \frac{\pi^{\theta_{n'}}(a|s)}{\pi^{\theta_{n'}}(a'|s)}\frac{\Delta}{4HR^*} \\ &= \exp(\theta_{n'}(s,a) - \theta_{n'}(s,a'))\frac{\Delta}{4HR^*} \end{aligned}$$

$$\geq \exp(b-(b-\delta))\frac{\Delta}{4HR^*}$$
$$= \exp(\delta)\frac{\Delta}{4HR^*},$$

where we used in the last inequality that $\theta_{n'}(s,a') \leq b-\delta$ for all $n' > \tau(n)$ and $\theta_{n'}(s,a) > b$ for all $n' > N_1$. By the definition of $\mathcal{T}^{(n)}$ these inequalities holds especially for all $n' \in \mathcal{T}^{(n)}$. Using this we can imply that for all $n > N_1$ with $\mathcal{T}^{(n)} \neq \emptyset$,

$$\begin{aligned}\frac{1}{\eta}\Big(\theta_{N_1}(s,a)-\theta_n(s,a)\Big) &= \sum_{n'=N_1+1}^{n-1}\frac{\partial J(\theta_{n'})}{\partial\theta_{n'}(s,a)}\\ &\leq \sum_{n'\in\mathcal{T}^{(n)}}\frac{\partial J(\theta_{n'})}{\partial\theta_{n'}(s,a)}\\ &\leq \exp(\delta)\frac{\Delta}{4HR^*}\sum_{n'\in\mathcal{T}^{(n)}}\frac{\partial J(\theta_{n'})}{\partial\theta_{n'}(s,a')}\\ &= \exp(\delta)\frac{\Delta}{4HR^*}Z_n,\end{aligned}$$

where the first inequality holds because $\theta_{n'}(s,a)$ is strictly decreasing for $n' > N_1$, i.e. $\frac{\partial J(\theta_{n'})}{\partial\theta_{n'}(s,a)} < 0$ for all $n' \in \{N_1+1,\dots,n-1\}$. In the second inequality we used

$$\left|\frac{\frac{\partial J(\theta_{n'})}{\partial\theta_{n'}(s,a)}}{\frac{\partial J(\theta_{n'})}{\partial\theta_{n'}(s,a')}}\right| \geq \exp(\delta)\frac{\Delta}{4HR^*}.$$

Note that $\frac{\partial J(\theta_{n'})}{\partial\theta_{n'}(s,a)} < 0$ and $\frac{\partial J(\theta_{n'})}{\partial\theta_{n'}(s,a')} < 0$ for $n' \in \mathcal{T}^{(n)}$ so that the sign of the inequality reverses. Finally, we deduce from $Z_n \to -\infty$ that $\theta_n(s,a) \to \infty$ for $n \to \infty$, which is a contradiction to $\theta_n(s,a)$ strictly decreasing for all $n > N_1$. ∎

LEMMA A.12. *Consider $s \in \mathcal{S}^{[\mathcal{H}]}$ such that $I_+^s \neq \emptyset$. Then for any $a_+ \in I_+^s$ it holds that*

$$\sum_{a\in B_0^s(a_+)}\theta_n(s,a) \to \infty, \quad \text{for } n \to \infty.$$

Proof. Let $a_+ \in I_+^s$ and $a \in B_0^s(a_+)$. Then by definition of $B_0^s(a_+)$ we have

$$\pi^{\theta_n}(a_+|s) \leq \pi^{\theta_n}(a|s)$$

for all $n > 0$ and hence by softmax parametrization $\theta_n(s,a_+) \leq \theta_n(s,a)$ for all $n > 0$. By Lemma A.11 we have that $\theta_n(s,a_+)$ and thus also $\theta_n(s,a)$ is bounded from below for $n \to \infty$. Together with

$$\max_{\{a\in B_0^s(a_+)\}}\theta_n(s,a) \to \infty, \quad \text{for } n \to \infty$$

by Lemma A.9 we deduce the claim. ∎

Finally, we are ready to prove the asymptotic convergence of simultaneous PG with tabular softmax parametrization.

Proof of Theorem A.1. We have to show that $I_+^s = \emptyset$ for all $s \in \mathcal{S}^{[\mathcal{H}]}$. So assume there exists $s \in \mathcal{S}^{[\mathcal{H}]}$ such that $I_+^s \neq \emptyset$ and let $a_+ \in I_+^s$. Then by Lemma A.12 we have

$$\sum_{a \in B_0^s(a_+)} \theta_n(s,a) \to \infty, \quad \text{for } n \to \infty. \tag{A.5}$$

For any $a \in I_-^s$ we have by Lemma A.11 that

$$\frac{\pi^{\theta_n}(a|s)}{\pi^{\theta_n}(a_+|s)} = \exp(\underbrace{\theta_n(s,a)}_{\to -\infty} - \underbrace{\theta_n(s,a_+)}_{\text{bounded from below}}) \to 0, \quad n \to \infty.$$

Hence, there exists $N_2 > N_1$ such that for all $n > N_2$

$$\frac{\pi^{\theta_n}(a|s)}{\pi^{\theta_n}(a_+|s)} < \frac{\Delta}{16|\mathcal{A}|HR^*},$$

which leads for $n > N_2$ to

$$-HR^* \sum_{a \in I_-^s} \pi^{\theta_n}(a|s) > -\frac{\Delta}{16}\pi^{\theta_n}(a_+|s). \tag{A.6}$$

Note that if $I_-^s = \emptyset$ we can just ignore this sum later on.
Next consider $a \in \bar{B}_0^s(a_+) \subset I_0^s$. By the definition of I_0^s we have that $A_h^{\theta_n}(s,a) \to A_h^{\infty}(s,a) = 0$ for $n \to \infty$. By Lemma A.10 we have for $n \geq N_{a_+}$

$$1 < \frac{\pi^{\theta_n}(a_+|s)}{\pi^{\theta_n}(a|s)}.$$

Thus, there exists $N_3 > \max\{N_2, N_{a_+}\}$ such that for all $n \geq N_3$

$$|A_h^{\theta_n}(s,a)| < \frac{\pi^{\theta_n}(a_+|s)}{\pi^{\theta_n}(a|s)} \frac{\Delta}{16|\mathcal{A}|}.$$

This implies

$$\sum_{a \in \bar{B}_0^s(a_+)} \pi^{\theta_n}(a|s)|A_h^{\theta_n}(s,a)| < \pi^{\theta_n}(a_+|s)\frac{\Delta}{16}$$

and so

$$-\pi^{\theta_n}(a_+|s)\frac{\Delta}{16} < \sum_{a \in \bar{B}_0^s(a_+)} \pi^{\theta_n}(a|s)A_h^{\theta_n}(s,a) < \pi^{\theta_n}(a_+|s)\frac{\Delta}{16}, \tag{A.7}$$

for all $n > N_3$. We can conclude again for $n > N_3$,

$$0 = \sum_{a \in \mathcal{A}} \pi^{\theta_n}(a|s)A_h^{\theta_n}(s,a)$$

$$
\begin{aligned}
&= \sum_{a\in B_0^s(a_+)} \pi^{\theta_n}(a|s) A_h^{\theta_n}(s,a) + \sum_{a\in \bar{B}_0^s(a_+)} \pi^{\theta_n}(a|s) A_h^{\theta_n}(s,a) \\
&\quad + \sum_{a\in I_+^s} \pi^{\theta_n}(a|s) A_h^{\theta_n}(s,a) + \sum_{a\in I_-^s} \pi^{\theta_n}(a|s) A_h^{\theta_n}(s,a) \\
&> \sum_{a\in B_0^s(a_+)} \pi^{\theta_n}(a|s) A_h^{\theta_n}(s,a) - \pi^{\theta_n}(a_+|s)\frac{\Delta}{16} + \pi^{\theta_n}(a_+|s)\frac{\Delta}{4} - HR^* \sum_{a\in I_-^s} \pi^{\theta_n}(a|s) \\
&\geq \sum_{a\in B_0^s(a_+)} \pi^{\theta_n}(a|s) A_h^{\theta_n}(s,a) - \pi^{\theta_n}(a_+|s)\frac{\Delta}{16} + \pi^{\theta_n}(a_+|s)\frac{\Delta}{4} - \frac{\Delta}{16}\pi^{\theta_n}(a_+|s) \\
&> \sum_{a\in B_0^s(a_+)} \pi^{\theta_n}(a|s) A_h^{\theta_n}(s,a),
\end{aligned}
$$

where we used (A.7) and Lemma A.4 in the first inequality and (A.6) in the second inequality. Finally, by our assumption and (A.5) for $n > N_3$,

$$
\begin{aligned}
\infty \overset{n\to\infty}{\longleftarrow} &\sum_{a\in B_0^s(a_+)} (\theta_n(s,a) - \theta_{N_3}(s,a)) \\
&= \eta \sum_{n'=N_3}^{n} \sum_{a\in B_0^s(a_+)} \frac{\partial J(\theta_{n'})}{\partial \theta_{n'}(s,a)} \\
&= \eta \sum_{n'=N_3}^{n} \widetilde{\rho}_\mu^{\pi^{\theta_{n'}}}(s) \sum_{a\in B_0^s(a_+)} \pi^{\theta_{n'}}(a|s) A_h^{\theta_{n'}}(s,a),
\end{aligned}
$$

which contradicts $\sum_{a\in B_0^s(a_+)} \pi^{\theta_n}(a|s) A_h^{\theta_n}(s,a) < 0$ for all $n > N_3$. ∎