Natural Policy Gradient as Doubly Smoothed Policy Iteration: A Bellman-Operator Framework

Phalguni Nanda
Edwardson School of Industrial Engineering
Purdue University
West Lafayette, IN 47906
nanda14@purdue.edu

Zaiwei Chen
Edwardson School of Industrial Engineering
Purdue University
West Lafayette, IN 47906
chen5252@purdue.edu

Abstract

In this work, we show that natural policy gradient (NPG), a core algorithm in reinforcement learning, admits an exact interpretation as a smoothed and averaged form of policy iteration. Specifically, we introduce *doubly smoothed policy iteration* (DSPI), a Bellman-operator framework in which each policy is obtained by applying a regularized greedy step to a weighted average of past Q-functions. DSPI includes policy iteration, dual-averaged policy iteration, NPG, and more general policy dual averaging methods as special cases.

Using only monotonicity and contraction of smoothed Bellman operators, we prove distribution-free global geometric convergence of DSPI. Consequently, standard NPG and policy dual averaging achieve an iteration complexity of $\mathcal{O}((1-\gamma)^{-1}\log((1-\gamma)^{-1}\epsilon^{-1}))$ for computing an ϵ-optimal policy, without modifying the MDP, adding regularization beyond the mirror map inherent in the update, or using adaptive, trajectory-dependent stepsizes. For the unregularized greedy case, we also prove finite termination of dual-averaged policy iteration. The same Bellman-operator framework extends to discounted MDPs with linear function approximation and to stochastic shortest path problems.

1 Introduction

In the last decade, reinforcement learning (RL) [39] has emerged as a principled framework for large-scale sequential decision making, with applications ranging from autonomous vehicles [49] and robotics [19] to large language models [11]. Mathematically, RL problems are typically modeled as Markov decision processes (MDPs) [32]. The seminal work [4] showed that solving an MDP reduces to finding a fixed point of the Bellman equation. By leveraging key properties of the Bellman operator, most notably contraction and monotonicity, a range of principled algorithms have been developed, including value iteration (VI) [4], policy iteration (PI) [20], and their variants [32, 5, 7].

In the modern era of sequential decision making, it was recognized that an MDP can be formulated as a continuous optimization problem over the policy space. From this perspective, the policy gradient (PG) method was introduced [40]. Building on this viewpoint and exploiting the geometry of the policy space, the natural policy gradient (NPG) method and more general approaches such as policy mirror ascent and policy dual averaging (PDA) were developed [24, 26, 21]. Over time, practical variants of these gradient-based methods, such as trust region policy optimization [35] and proximal policy optimization [36], have become dominant approaches in large-scale applications [19, 11].

The empirical success of NPG and its variants has motivated extensive theoretical work on their convergence properties. Existing approaches for studying NPG can be broadly categorized into two groups. One prominent line of work adopts a gradient-based optimization viewpoint, leveraging tools such as the policy gradient theorem, the performance difference lemma, the Polyak Łojasiewicz

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

condition [40, 1], as well as techniques from continuous optimization [26, 44]. A second line of work views NPG as an approximation of PI. Indeed, the original NPG paper [24, Theorem 2] observed that, as the stepsize approaches infinity, NPG recovers PI. Following this perspective, NPG has been analyzed in [15, 8] using properties of the Bellman operator, in a manner analogous to PI.

Despite substantial progress, existing analyses still leave open a conceptual gap. The defining structural property of discounted MDPs is the ℓ_∞ contraction of the Bellman operator, which underlies the design and analysis of classical methods such as VI and PI. This property is not naturally visible when NPG is analyzed as a continuous optimization method over the policy space. Consequently, many optimization-based analyses establish geometric convergence of NPG by introducing additional regularization of the MDP or the algorithm to induce curvature [13, 26, 37]. In contrast, approaches that view NPG as an approximation of PI [10, 15] exploit Bellman operator properties and can obtain distribution-free geometric convergence without additional regularization, but they require stepsizes that are chosen adaptively based on the algorithm trajectory. See Section 1.1 for further details.

More fundamentally, a unified perspective connecting classical Bellman iteration methods, such as PI, with modern gradient-based methods, such as NPG and PDA, remains lacking at both the algorithmic and analytical levels. In this paper, we address this gap by introducing *doubly smoothed policy iteration* (DSPI), a unified framework that contains PI, dual-averaged PI, NPG, and PDA as special cases. This perspective enables a unified analysis that yields strong guarantees for a broad class of algorithms. We summarize our contributions below.

- **Doubly Smoothed Policy Iteration.** We introduce a unified algorithmic framework, termed DSPI, in which the policy is updated via a regularized greedy step with respect to a weighted average of past Q-functions. This structure admits a compact Bellman operator formulation (cf. Algorithm 2, Line 4). We show that DSPI captures a broad class of existing algorithms, including PI, dual-averaged PI, and, most notably, NPG and PDA. Crucially, for NPG and PDA, interpreting these gradient-based methods as smoothed and averaged variants of PI allows the updates to be expressed via the Bellman operator, thereby enabling a principled analysis based on its properties.
- **A Unified Convergence Analysis.** We show that DSPI converges asymptotically whenever the stepsizes are non-summable. Moreover, under a constant stepsize, DSPI achieves distribution-free global geometric convergence. In particular,

$$\|V^{\pi_k} - V^*\|_\infty \leq (1-(1-\gamma)\beta)^{k-1}\left(\gamma\|V^* - V^{\pi_0}\|_\infty + 1\right), \quad \forall\, k \geq 0,$$

  where V^{π_0}, V^{π_k}, and V^* denote the value functions of the initial policy, the k-th iterate, and an optimal policy, respectively, γ is the discount factor, and β is the constant stepsize. See Theorem 4.1 for details. As byproducts,
  - this result implies an iteration complexity of $\mathcal{O}((1-\gamma)^{-1}\log((1-\gamma)^{-1}\epsilon^{-1}))$ for NPG and PDA to achieve $\|V^{\pi_k} - V^*\|_\infty \leq \epsilon$ (cf. Theorem 4.3), matching the state of the art without requiring additional regularization of the MDP or the algorithm, or the use of adaptive stepsizes;
  - it also enables us to establish finite termination of dual-averaged PI; in particular, an iteration complexity of $\mathcal{O}(mn(1-\gamma)^{-1}\log((1-\gamma)^{-1}))$ for finding an optimal policy (cf. Theorem 4.5), where n and m denote the sizes of the state and action spaces, respectively.

  Moreover, the algorithm and analysis extend to settings such as discounted MDPs with linear function approximation, where we obtain the same convergence rate up to a function approximation error, and stochastic shortest path problems, where we establish analogous geometric convergence. These results are deferred to Appendices C and D due to space limitations.

Scope. Our main results focus on the exact policy-improvement dynamics: at each iteration, the algorithm has access to the exact Q-function of the current policy. This isolates the policy-optimization mechanism from statistical policy-evaluation error. We therefore do not claim sample-complexity guarantees for model-free RL or actor–critic methods in the main results.

1.1 Related Literature

Since we focus on the MDP setting rather than the model-free RL setting, we restrict our discussion below to PI and NPG, rather than results that broadly analyze actor–critic algorithms.

Policy Iteration. The PI method was first proposed in [20]. The global geometric convergence of PI was shown in [32]. The proof has two main steps: (1) showing that the performance of the policies

is monotonically improving, and (2) leveraging the contraction property of the Bellman operator to establish a contractive recursion. This approach also inspires our analysis of DSPI. Beyond global geometric convergence, it was shown that PI can terminate in a finite number of iterations [34, 45]. The key property that enabled such strong results is that PI gradually eliminates sub-optimal actions. Such a property also plays a key role in showing the finite termination of dual-averaged PI in this work. Other variants of PI, including simplex PI, modified PI, and incremental PI, are analyzed in [34, 45, 7, 5].

An existing work that is closely related to ours is [43], which studies the impact of Kullback–Leibler (KL) regularization in approximate dynamic programming schemes. To that end, [43] introduces the Dual Averaging Modified Policy Iteration (DA-MPI). Although DA-MPI maintains averaged Q-values and applies a regularized greedy update to them, it differs from DSPI in the following aspects. First, the average is taken over Q-values that are computed through repeated application of KL-regularized Bellman operators, which makes these Q-values generally different from the unregularized action-value functions used by DSPI. Second, DA-MPI requires specific choices of stepsizes unlike DSPI, which allows any non-summable stepsize sequence. Lastly, DA-MPI crucially relies on entropy regularization and the corresponding KL-regularized greedy update. In contrast, DSPI permits any concave regularization function, including quadratic and zero regularization; the latter is essential for capturing dual-averaged PI. Therefore, DA-MPI does not exactly recover dual-averaged PI, NPG, or PDA, all of which are captured by DSPI. Moreover, [43] establishes an $\mathcal{O}(1/k)$ convergence rate, whereas our analysis establishes global geometric convergence.

Natural Policy Gradient. The PG method was first proposed in [40], which treats a weighted sum of the value function as a scalar objective and performs projected gradient ascent in the policy space. A comprehensive analysis of PG was carried out in [1] and subsequent works. NPG, proposed in [24], can be viewed as PG with a preconditioner. Alternatively, it can be interpreted as policy mirror ascent [26, 44] or PDA [21], with the divergence-generating function chosen as the Shannon entropy. The majority of existing analyses of NPG suffer from at least one of the following limitations: (1) lack of global geometric convergence [1, 37]; (2) dependence on the initial distribution or a concentrability coefficient [8, 44, 26, 3]; (3) additional regularization of the MDP or the algorithm [37, 13, 26, 30, 50, 44]; or (4) reliance on adaptive stepsizes that depend on the policy trajectory [8, 15]. See Table 1 from Appendix A for more details. The only exception is [23], which provides the state-of-the-art analysis of NPG from a continuous optimization perspective[1], but their choice of stepsize depends on the Bellman error (also called the advantage gap) of the initial policy, which is not required in our result.

2 Background

Consider an infinite horizon discounted MDP defined by the tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, p, \mathcal{R}, \gamma)$, where $\mathcal{S}$ and $\mathcal{A}$ denote the finite state and action spaces, respectively, $\{p(\cdot \mid s, a)\}_{(s,a)\in\mathcal{S}\times\mathcal{A}}$ denotes the transition kernel, $\mathcal{R} : \mathcal{S} \times \mathcal{A} \to [0, 1]$ is the reward function, and $\gamma \in (0, 1)$ is the discount factor. Throughout, we denote $n = |\mathcal{S}|$ and $m = |\mathcal{A}|$. Given a policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$, where $\Delta(\mathcal{A})$ denotes the set of probability distributions supported on $\mathcal{A}$, the Q-function $Q^\pi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is defined as

$$Q^\pi(s, a) = \mathbb{E}_\pi \left[\sum_{k=0}^{\infty} \gamma^k \mathcal{R}(S_k, A_k) \,\middle|\, S_0 = s, A_0 = a \right]$$

for all $(s, a) \in \mathcal{S} \times \mathcal{A}$, where the expectation $\mathbb{E}_\pi[\,\cdot\,]$ is taken with respect to the randomness in both the action selection $A_k \sim \pi(\cdot \mid S_k)$ and the state transitions $S_{k+1} \sim p(\cdot \mid S_k, A_k)$. With Q^π defined above, the value function $V^\pi : \mathcal{S} \to \mathbb{R}$ is defined as

$$V^\pi(s) = \sum_{a\in\mathcal{A}} \pi(a|s) Q^\pi(s, a)$$

for all $s \in \mathcal{S}$. Since the MDP has finitely many states and actions, Q^π and V^π can be equivalently viewed as vectors in $\mathbb{R}^{mn}$ and $\mathbb{R}^n$, respectively.

A policy π^* is *optimal* if $V^* := V^{\pi^*} \geq V^\pi$ for all π, or equivalently, $Q^* := Q^{\pi^*} \geq Q^\pi$ for all π. Throughout this paper, inequalities between vectors are understood componentwise. For notational

[1]The results in [23] were developed for general policy mirror ascent, which covers NPG as a special case.

simplicity, for any vector $x \in \mathbb{R}^{mn}$ (e.g., representing a policy or a Q-function), we denote by $x(s) \in \mathbb{R}^m$ the vector whose a-th entry is $x(s,a)$.

2.1 The Bellman Equation and Policy Iteration

The key to solving an MDP lies in the Bellman equation. We next present the Bellman equations for policy evaluation and policy optimization. To facilitate the connection between PI and NPG developed later in this paper, we express these equations in terms of Q-functions.

Given a policy π, its Q-function Q^π is the unique solution to the Bellman equation $Q = \mathcal{H}^\pi(Q)$, where $\mathcal{H}^\pi : \mathbb{R}^{mn} \to \mathbb{R}^{mn}$ is the Bellman operator associated with policy π, defined as

$$[\mathcal{H}^\pi(Q)](s,a) = \mathcal{R}(s,a) + \gamma \sum_{s' \in \mathcal{S}} p(s' \mid s,a) \sum_{a' \in \mathcal{A}} \pi(a' \mid s') Q(s',a')$$

for all $(s,a) \in \mathcal{S} \times \mathcal{A}$. Similarly, the optimal Q-function Q^* is the unique solution to the equation $Q = \mathcal{H}(Q)$, where $\mathcal{H} : \mathbb{R}^{mn} \to \mathbb{R}^{mn}$ is the Bellman optimality operator defined as

$$[\mathcal{H}(Q)](s,a) = \mathcal{R}(s,a) + \gamma \sum_{s' \in \mathcal{S}} p(s' \mid s,a) \max_{a' \in \mathcal{A}} Q(s',a')$$

for all $(s,a) \in \mathcal{S} \times \mathcal{A}$. Moreover, once Q^* is obtained, an optimal policy π^* can be computed by choosing actions greedily with respect to Q^*. The operators $\mathcal{H}^\pi$ and $\mathcal{H}$ enjoy several useful properties. In particular, both are contraction mappings with respect to the ℓ_∞ norm with contraction factor γ, and are monotone in the sense that $F(Q_1) \le F(Q_2)$ whenever $Q_1 \le Q_2$, where F denotes either $\mathcal{H}^\pi$ or $\mathcal{H}$ [32, 7]. These two properties enable the design and analysis of several classical algorithms, such as VI and PI.

When expressed in terms of Q-functions, the PI algorithm can be compactly written as

$$\mathcal{H}^{\pi_{k+1}}(Q^{\pi_k}) = \mathcal{H}(Q^{\pi_k}). \tag{1}$$

Note that (1) does not define a unique update. In particular, π_{k+1} can be any policy such that, when its associated Bellman operator $\mathcal{H}^{\pi_{k+1}}$ is applied to the previous Q-function Q^{π_k}, the result coincides with applying the Bellman optimality operator to Q^{π_k}. In view of the definition of the Bellman operators, one concrete way to implement the PI update in (1) is

$$\pi_{k+1}(s) \in \operatorname*{arg\,max}_{\mu \in \Delta(\mathcal{A})} \{\mu^\top Q^{\pi_k}(s)\}, \ \forall\, s \in \mathcal{S}, \tag{2}$$

where, according to our notation, $\pi_{k+1}(s)$ (respectively, $Q^{\pi_k}(s)$) denotes the m-dimensional vector whose a-th entry is $\pi_{k+1}(a \mid s)$ (respectively, $Q^{\pi_k}(s,a)$).

The global geometric convergence of PI is well established in the literature [32]. The analysis proceeds in two main steps: (1) using the monotonicity of the Bellman operators to show that the policies are monotonically improving, i.e., $Q^{\pi_k} \le Q^{\pi_{k+1}}$; (2) using the contraction property of the Bellman operators to show that the convergence error $\|Q^* - Q^{\pi_k}\|_\infty$ admits a contractive recursion. For completeness, we provide an analysis of PI in Appendix B.1.

2.2 Natural Policy Gradient and Policy Dual Averaging

Since an optimal policy maximizes the value function $V^\pi(s)$ uniformly over all states s, solving an MDP is equivalent to solving an optimization problem over the policy space with the scalar objective $V^\pi(\rho) := \sum_{s \in \mathcal{S}} \rho(s) V^\pi(s)$, where $\rho \in \Delta(\mathcal{S})$ can be interpreted as the initial state distribution. This viewpoint naturally motivates gradient-based algorithms. In particular, the vanilla PG method performs projected gradient ascent over the policy space [40], while the NPG method performs gradient ascent under a geometry induced by the Fisher information matrix [24].

There are many equivalent formulations of the NPG update. In this paper, to facilitate its connection with PI, we adopt the NPG update rule from [1][2], presented in Algorithm 1, where $h : \Delta(\mathcal{A}) \to \mathbb{R}$ is the entropy function defined by $h(\mu) = -\sum_{a \in \mathcal{A}} \mu(a) \log \mu(a)$. Note that NPG is a special case of PDA [21]. Specifically, replacing $h(\cdot)$ in Algorithm 1, Line 4, with the negative of any divergence-generating function yields PDA.

[2] It corresponds to Q-NPG in [1, Page 25] when using a tabular representation.

Algorithm 1 Natural Policy Gradient

1: **Input:** Initialize π_0 as the uniform policy and $\theta_0 = \mathbf{0} \in \mathbb{R}^{mn}$.
2: **for** $k = 0, 1, 2, 3, \cdots$ **do**
3: $\quad \theta_{k+1} = \theta_k + \alpha_k Q^{\pi_k}$
4: $\quad \pi_{k+1}(s) = \arg\max_{\mu \in \Delta(\mathcal{A})} \left\{ \mu^\top \theta_{k+1}(s) + h(\mu) \right\}$ for any $s \in \mathcal{S}$.
5: **end for**

3 Doubly Smoothed Policy Iteration

In this section, we introduce DSPI, a unified algorithmic framework that captures a broad class of algorithms and enables a unified convergence rate analysis.

To present the algorithm, we begin by introducing the concept of smoothed Bellman operators. Let $\nu : \Delta(\mathcal{A}) \to \mathbb{R}$ be a bounded, non-negative, and concave function; examples include the Shannon entropy, the Tsallis entropy, the shifted negative squared norm, and the zero function. Given a scalar $\eta \geq 0$, let the smoothed Bellman optimality operator $\mathcal{H}_\eta : \mathbb{R}^{mn} \to \mathbb{R}^{mn}$ be defined as

$$[\mathcal{H}_\eta(Q)](s,a) = \mathcal{R}(s,a) + \gamma \sum_{s' \in \mathcal{S}} p(s' \mid s,a) \max_{\mu \in \Delta(\mathcal{A})} (\mu^\top Q(s') + \eta \nu(\mu))$$

for all $(s,a) \in \mathcal{S} \times \mathcal{A}$. Given a policy π and a scalar $\eta \geq 0$, let $\mathcal{H}_\eta^\pi : \mathbb{R}^{mn} \to \mathbb{R}^{mn}$ be defined as

$$[\mathcal{H}_\eta^\pi(Q)](s,a) = \mathcal{R}(s,a) + \gamma \sum_{s' \in \mathcal{S}} p(s' \mid s,a)(\pi(s')^\top Q(s') + \eta \nu(\pi(s')))$$

for all $(s,a) \in \mathcal{S} \times \mathcal{A}$. Note that $\mathcal{H}_\eta$ (respectively, $\mathcal{H}_\eta^\pi$) reduces to $\mathcal{H}$ (respectively, $\mathcal{H}^\pi$) when $\eta = 0$. Moreover, the smoothed Bellman operators are also contraction mappings and are monotonic (cf. Appendix B.2).

With the smoothed Bellman operators in place, DSPI is presented in Algorithm 2.

Algorithm 2 Doubly Smoothed Policy Iteration

1: **Input:** Initialize $\pi_0(s) \in \arg\max_{\mu \in \Delta(\mathcal{A})} \nu(\mu)$ for all $s \in \mathcal{S}$ and $\bar{Q}_0 = \mathbf{0}$.
2: **for** $k = 0, 1, 2, 3, \cdots$ **do**
3: $\quad$ Maintain a running average of past Q-functions: $\bar{Q}_{k+1} = (1 - \beta_k)\bar{Q}_k + \beta_k Q^{\pi_k}$, where $\beta_k \in [0,1]$ is the stepsize.
4: $\quad$ Choose π_{k+1} such that $\mathcal{H}_{\eta_k}^{\pi_{k+1}}(\bar{Q}_{k+1}) = \mathcal{H}_{\eta_k}(\bar{Q}_{k+1})$, where $\eta_k = \tau \prod_{j=1}^k (1 - \beta_j)$ with $\tau \geq 0$ being a tunable parameter.
5: **end for**

Similar to PI, Line 4 of Algorithm 2 in general does not define a unique update. As long as π_{k+1} is chosen such that, when its associated smoothed Bellman operator $\mathcal{H}_{\eta_k}^{\pi_{k+1}}$ is applied to the running average $\bar{Q}_{k+1}$, the result coincides with that of applying the smoothed Bellman optimality operator $\mathcal{H}_{\eta_k}$ to $\bar{Q}_{k+1}$, it is a valid update. A concrete way to implement this step is given by

$$\pi_{k+1}(s) \in \arg\max_{\mu \in \Delta(\mathcal{A})} \left\{ \mu^\top \bar{Q}_{k+1}(s) + \tau \prod_{j=1}^k (1 - \beta_j) \nu(\mu) \right\}, \ \forall\, s \in \mathcal{S}. \tag{3}$$

While (3) is more explicit, the Bellman operator viewpoint in Line 4 of Algorithm 2 more naturally captures the fundamental nature of the update.

Compared with PI presented in (1) (or more explicitly, (2)), DSPI introduces two levels of *smoothing*. First, it updates the policy using a weighted average of all past Q-functions rather than relying solely on the most recent one. Second, it replaces the greedy policy update with a softmax-based update, where $\nu(\cdot)$ acts as a regularization term. These two smoothing mechanisms motivate the name *doubly smoothed policy iteration*. Note that the stepsize β_k simultaneously controls the weights in computing $\bar{Q}_{k+1}$ and the degree of smoothing through the scaling η_k. As β_k increases from zero to one, $\bar{Q}_{k+1}$ places more weight on the last iterate Q^{π_k} and the effect of smoothing fades away. In the extreme case where $\beta_k \equiv 1$, DSPI reduces to PI.

In addition to capturing PI, we next show that DSPI serves as a unified algorithmic framework that encompasses a broad class of algorithms, most importantly NPG and PDA.

Natural Policy Gradient. The following proposition shows that NPG is a special case of DSPI.

Proposition 3.1. *In Algorithm 2, when choosing $\beta_k = \alpha_k / \sum_{i=0}^k \alpha_i$, $\tau = 1/\alpha_0$, and $\nu(\cdot) = h(\cdot)$, the sequence of policies generated by Algorithm 2 with Line 4 implemented via* (3) *is identical to that generated by Algorithm 1.*

Proof of Proposition 3.1. Since both algorithms initialize at the uniform policy, it suffices to show that their update equations generate identical policies thereafter. By Line 3 of Algorithm 1, we have $\theta_{k+1} = \sum_{i=0}^k \alpha_i Q^{\pi_i}$ for all $k \geq 0$. As a result, Line 4 of Algorithm 1 becomes

$$\pi_{k+1}(s) = \arg\max_{\mu \in \Delta(\mathcal{A})} \left\{ \mu^\top \left(\frac{\sum_{i=0}^k \alpha_i Q^{\pi_i}(s)}{\sum_{i=0}^k \alpha_i} \right) + \frac{\alpha_0 \tau}{\sum_{i=0}^k \alpha_i} h(\mu) \right\}, \ \forall\, s \in \mathcal{S},$$

where we used *(i)* the $\arg\max$ operator being indifferent to constant scaling and *(ii)* $\tau = 1/\alpha_0$. In view of the above equality and Line 4 of Algorithm 2 (which is equivalent to (3) since the Shannon entropy $h(\cdot)$ is strongly concave), it remains to show that $\bar{Q}_{k+1} = \sum_{i=0}^k \alpha_i Q^{\pi_i} / \sum_{i=0}^k \alpha_i$ and $\prod_{j=1}^k (1-\beta_j) = \alpha_0 / \sum_{i=0}^k \alpha_i$ for all $k \geq 0$, which we prove by induction.

For $k = 0$, since $\beta_0 = 1$, the base case clearly holds. Suppose that the desired equalities hold for some $k \geq 0$. Then, by the update equation in Line 3 of Algorithm 2, we have

$$\bar{Q}_{k+2} = (1-\beta_{k+1})\bar{Q}_{k+1} + \beta_{k+1} Q^{\pi_{k+1}} = \frac{(\sum_{i=0}^k \alpha_i)\bar{Q}_{k+1} + \alpha_{k+1} Q^{\pi_{k+1}}}{\sum_{i=0}^{k+1} \alpha_i} = \frac{\sum_{i=0}^{k+1} \alpha_i Q^{\pi_i}}{\sum_{i=0}^{k+1} \alpha_i},$$

where the last equality follows from the induction hypothesis. Similarly, we have

$$\prod_{j=1}^{k+1} (1-\beta_j) = \prod_{j=1}^{k} (1-\beta_j) \cdot (1-\beta_{k+1}) = \frac{\alpha_0}{\sum_{i=0}^k \alpha_i} \left(1 - \frac{\alpha_{k+1}}{\sum_{i=0}^{k+1} \alpha_i} \right) = \frac{\alpha_0}{\sum_{i=0}^{k+1} \alpha_i}.$$

The induction is complete. □

Proposition 3.1 is conceptually important, as it reveals that the widely used NPG algorithm, although originally developed from a gradient-based continuous optimization viewpoint, can be interpreted as a smoothed and averaged variant of classical PI. This perspective allows us to build on Bellman's foundational insights, such as the contraction and monotonicity of the Bellman operator, to establish strong convergence guarantees for NPG via the analysis of DSPI.

Policy Dual Averaging. Since NPG is a special case of PDA when the divergence-generating function is chosen as the negative entropy [23], by setting $\nu(\cdot)$ to a shifted negative divergence-generating function, i.e., $\nu(\mu) = C - \omega(\mu)$ for a bounded convex divergence-generating function ω, where $C := \max_\mu \omega(\mu)$ ensures that $\nu(\mu) \geq 0$, DSPI also captures PDA. The proof follows identically to that of Proposition 3.1 and is therefore omitted.

Dual-Averaged Policy Iteration. DSPI is more general than PDA, as it allows choosing $\nu(\cdot) \equiv 0$. In this case, Algorithm 2 updates the policy greedily based on a weighted average of historical Q-functions. We refer to this update as dual-averaged PI. One should not confuse dual-averaged PI with incremental PI [32], where the averaging is performed in the policy space, whereas in dual-averaged PI the averaging is performed in the Q-function space.

4 Analysis

This section presents our main theoretical results. We begin with the convergence rates of DSPI in Section 4.1. Due to its unified nature, this result immediately implies the distribution-free global geometric convergence of NPG and PDA in Section 4.2, and enables us to establish the finite termination of dual-averaged PI in Section 4.3.

4.1 Convergence Rates of Doubly Smoothed Policy Iteration

The following theorem presents the asymptotic convergence and convergence rates of the DSPI algorithm. The proof of Theorem 4.1 is presented in Section 5.

Theorem 4.1. *Consider $\{\pi_k\}$ generated by Algorithm 2.*

(1) When $\beta_0 = 1$, $\beta_k \in (0, 1]$, and $\sum_{k=0}^{\infty} \beta_k = \infty$, we have $\lim_{k\to\infty} \|V^* - V^{\pi_k}\|_\infty = 0$.
(2) When choosing $\beta_0 = 1$, $\beta_k \equiv \beta \in (0, 1]$ for all $k \geq 1$, we have

$$\|V^* - V^{\pi_k}\|_\infty \leq (1 - (1-\gamma)\beta)^{k-1} \left(\gamma \|V^* - V^{\pi_0}\|_\infty + \tau \nu_{\max}\right), \quad \forall\, k \geq 1,$$

where $\nu_{\max} := \max_{\mu \in \Delta(\mathcal{A})} \nu(\mu)$.

Theorem 4.1 (1) states that V^{π_k} asymptotically converges to the optimal value function as long as the stepsizes are non-summable, and Theorem 4.1 (2) characterizes global geometric convergence; see Appendix E for a numerical illustration. Moreover, Theorem 4.1 (2) implies the iteration complexity presented below.

Corollary 4.2. *For any $\epsilon > 0$, choose $\beta_0 = 1$, $\beta_k = 1/2$ for $k \geq 1$, and choose τ so that $\tau \nu_{\max} \leq 1$ in Algorithm 2, we have $\|V^* - V^{\pi_k}\|_\infty \leq \epsilon$ provided that*

$$k \geq 2(1-\gamma)^{-1} \log(\epsilon^{-1}(1-\gamma)^{-1}).$$

4.2 Convergence Rates of Natural Policy Gradient and Policy Dual Averaging

As Proposition 3.1 shows NPG as a special case of the DSPI algorithm, Theorem 4.1 allows us to establish the geometric convergence of NPG below.

Theorem 4.3. *Consider $\{\pi_k\}$ generated by Algorithm 1. When choosing $\alpha_0 = \log(m)$ and $\alpha_k = \beta\alpha_0/(1-\beta)^k$ for all $k \geq 1$ (where $\beta \in (0, 1)$ can be arbitrary), we have*

$$\|V^* - V^{\pi_k}\|_\infty \leq (1 - (1-\gamma)\beta)^{k-1} \left(\gamma \|V^* - V^{\pi_0}\|_\infty + 1\right), \quad \forall\, k \geq 1.$$

Remark. The geometrically increasing stepsize sequence $\{\alpha_k\}$ has also been used in prior works to establish geometric convergence of NPG [22, 15, 8]. Notably, Proposition 3.1 shows that the same policy sequence can be implemented via Algorithm 2 using the normalized recursion $\bar{Q}_{k+1} = (1-\beta)\bar{Q}_k + \beta Q^{\pi_k}$, together with the decaying smoothing parameter $\eta_k = \tau(1-\beta)^k$. Thus, the prescribed schedule does not require storing exponentially large variables.

The proof of Theorem 4.3 follows by combining Proposition 3.1 and Theorem 4.1, and is given in Appendix B.3. As a consequence, we obtain the following iteration complexity for NPG.

Corollary 4.4. *For any $\epsilon > 0$, by choosing $\alpha_0 = \log(m)$ and $\alpha_k = \log(m) \cdot 2^{k-1}$ for all $k \geq 1$ in Algorithm 1, we have $\|V^* - V^{\pi_k}\|_\infty \leq \epsilon$ provided that*

$$k \geq 2(1-\gamma)^{-1} \log(\epsilon^{-1}(1-\gamma)^{-1}).$$

Theorem 4.3 and Corollary 4.4 show that NPG enjoys global geometric convergence, measured by the optimality gap of the last iterate π_k. Moreover, the convergence bounds have several important features, which we discuss below.

Distribution-Free Convergence Rates. In the existing literature on NPG, convergence rates typically depend on the initial distribution ρ, either explicitly or through a distribution mismatch coefficient [31, 37, 44, 48, 9, 8, 29, 12]. In contrast, our convergence bound is expressed in terms of the ℓ_∞ norm of the value-function gap and is free of such dependencies.

No Additional Regularization. While our iteration complexity matches the state of the art, our result is obtained without modifying the MDP or the algorithm. Several existing analyses that establish geometric convergence for NPG either regularize the MDP [17, 13, 1, 26, 50, 12] or introduce an additional strongly convex regularizer [26, 30, 27]. In contrast, we do not rely on any such modifications; the only regularization arises from the entropy term inherent in the standard NPG formulation. Our DSPI viewpoint, together with its Bellman operator formulation (cf. Line 4 of Algorithm 2), allows us to directly exploit the contraction property of the Bellman operator, thereby eliminating the need for additional curvature-inducing regularization.

A Simple Prespecified Stepsize Schedule. Our results show that NPG converges at a geometric rate with a simple stepsize schedule $\alpha_k = \beta\alpha_0/(1-\beta)^k$, where $\beta \in (0,1)$ is arbitrary. In contrast, existing results that view NPG as an approximation of PI [8, 15] require adaptive stepsizes that depend on the value of $\pi_k(a_k^* \mid s)$, where $a_k^* \in \arg\max_{a\in\mathcal{A}} Q^{\pi_k}(s,a)$. Similarly, analyses that study NPG from an optimization viewpoint require stepsizes that depend on the Bellman error (also known as the advantage gap) of the initial policy [23, Theorems 3.4 and 3.5].

Policy Dual Averaging. Similarly, by choosing $\nu(\mu) = C - \omega(\mu)$, where ω is a bounded divergence-generating function and $C := \max_{\mu\in\Delta(\mathcal{A})} \omega(\mu)$, Algorithm 2 captures PDA. Theorem 4.1 therefore implies global geometric convergence for PDA. The results and proofs are identical to those of Theorem 4.3 and are therefore omitted.

4.3 Finite Termination of Dual-Averaged Policy Iteration

When setting $\nu(\cdot) \equiv 0$ in Algorithm 2, the DSPI algorithm reduces to the dual-averaged PI, which updates the policies greedily with respect to the averaged Q-functions. As a special case of DSPI, the dual-averaged PI also enjoys geometric convergence (cf. Theorem 4.1 with $\nu_{\max} \equiv 0$). Moreover, we show that the algorithm terminates in a finite number of iterations with an optimal policy. This is presented in our next result.

Theorem 4.5. *Let $\{\pi_k\}$ be generated by Algorithm 2 with $\nu(\cdot) \equiv 0$ and a deterministic π_0, where Line 4 selects a deterministic greedy policy according to a arbitrary tie-breaking rule. For a fixed $\beta \in (0,1)$, let $k^* = 1 + \lceil \beta^{-1}(1-\gamma)^{-1} \log\left((1-\gamma)^{-1}\right) \rceil$. When choosing the stepsize as*

$$\beta_k = \begin{cases} 1, \text{ if } k = rk^*, r \in \{0\} \cup \mathbb{N}, \\ \beta, \text{ otherwise}, \end{cases} \tag{4}$$

the algorithm terminates after at most $n(m-1)k^*$ iterations with an optimal policy, where $\lceil x \rceil$ denotes the smallest integer greater than or equal to x.

Remark. When γ is treated as fixed, this yields a strongly polynomial bound in the numbers of states and actions, in the standard sense used for discounted MDP policy-iteration analyses [45, 33]. More generally, the bound is polynomial in n, m, and the effective horizon $(1-\gamma)^{-1}$.

The proof of Theorem 4.5 is given in Appendix B.5. Following [33], the key idea is to show that dual-averaged PI eliminates at least one suboptimal action at some state every k^* iterations, where $k^* = 1 + \lceil \beta^{-1}(1-\gamma)^{-1} \log\left((1-\gamma)^{-1}\right) \rceil$. Since there are at most $n(m-1)$ suboptimal actions, dual-averaged PI finds an optimal policy in at most $n(m-1)k^*$ iterations.

4.4 Extensions

Although we have focused on discounted MDPs with a tabular representation, the generality of our framework allows both the algorithm and its analysis to extend naturally to other settings. We discuss these extensions below.

Natural Policy Gradient with Linear Function Approximation. To overcome the curse of dimensionality, algorithms often incorporate function approximation (e.g., approximate dynamic programming). In Appendix C, we show that NPG with linear function approximation also admits a DSPI formulation, and we establish distribution-free global geometric convergence up to a function approximation error.

Natural Policy Gradient for Stochastic Shortest Path Problems. Beyond the discounted setting, our framework also extends to stochastic shortest path problems, a special case of undiscounted MDPs. Classical results [7] show that, under mild assumptions, the Bellman operator is monotonic and is a contraction with respect to a weighted ℓ_∞ norm. Leveraging the DSPI formulation, we show that NPG enjoys distribution-free global geometric convergence. See Appendix D for more details.

One may also ask whether our framework extends to average-reward MDPs. This setting is more subtle because the Bellman operator is no longer norm-contracting; it is either seminorm-contracting or merely non-expansive, depending on the structural assumptions imposed on the problem, such as recurrent, unichain, or weakly communicating structures [32]. We believe that our formulation of NPG as DSPI still applies, but the resulting convergence rate, whether geometric or polynomial, will depend on these structural assumptions. We leave this as a direction for future work.

5 Proof of Theorem 4.1

Our proof proceeds in two steps: we first establish monotonic improvement of the Q-functions, and then derive and solve a contractive recursion for the convergence error. We begin by presenting the monotonic improvement lemma.

Lemma 5.1. *Algorithm 2 satisfies $Q^{\pi_k} \leq Q^{\pi_{k+1}}$ for all $k \geq 0$.*

Remark. By our convention, $Q^{\pi_k} \leq Q^{\pi_{k+1}}$ means that $Q^{\pi_k}(s,a) \leq Q^{\pi_{k+1}}(s,a)$ for all (s,a). This vector-version monotonic improvement lemma is stronger than those in the literature on NPG, where monotonicity is established only for the scalar objective $V^{\pi_k}(\rho)$ [1]. This stronger form is important for leveraging the contraction property of the Bellman operator in the next step to derive a contractive recursion for the convergence error.

Proof of Lemma 5.1. It suffices to show that $Q^{\pi_k} \leq \mathcal{H}^{\pi_{k+1}}(Q^{\pi_k})$ for all $k \geq 0$. Once this is established, repeatedly applying $\mathcal{H}^{\pi_{k+1}}$ to both sides and using its monotonicity yields $Q^{\pi_k} \leq [\mathcal{H}^{\pi_{k+1}}]^n(Q^{\pi_k})$ for all $n \geq 0$. Taking $n \to \infty$ gives $Q^{\pi_k} \leq Q^{\pi_{k+1}}$.

For simplicity of notation, define $f : \mathbb{R}^{mn} \to \mathbb{R}^{mn}$ by $[f(\pi)](s,a) = \gamma \sum_{s' \in \mathcal{S}} p(s' \mid s,a)\nu(\pi(s'))$ for all (s,a). Then, for any policy π and $Q \in \mathbb{R}^{mn}$, we have $\mathcal{H}^{\pi}_{\eta}(Q) = \mathcal{H}^{\pi}(Q) + \eta f(\pi)$. We now show that $Q^{\pi_k} \leq \mathcal{H}^{\pi_{k+1}}(Q^{\pi_k})$ for the two cases $k = 0$ and $k \geq 1$.

- For $k = 0$, since $\beta_0 = 1$, we have $\bar{Q}_1 = Q^{\pi_0}$. It follows that $Q^{\pi_0} = \mathcal{H}^{\pi_0}(\bar{Q}_1) = \mathcal{H}^{\pi_0}_{\eta_0}(\bar{Q}_1) - \eta_0 f(\pi_0)$. Since $\mathcal{H}^{\pi_0}_{\eta_0}(\bar{Q}_1) \leq \mathcal{H}_{\eta_0}(\bar{Q}_1) = \mathcal{H}^{\pi_1}_{\eta_0}(\bar{Q}_1)$ and $f(\pi_1) \leq f(\pi_0)$, where the latter follows from initializing $\pi_0(s) \in \arg\max_{\mu \in \Delta(\mathcal{A})} \nu(\mu)$ for all $s \in \mathcal{S}$, we further obtain
$$Q^{\pi_0} \leq \mathcal{H}^{\pi_1}_{\eta_0}(\bar{Q}_1) - \eta_0 f(\pi_1) = \mathcal{H}^{\pi_1}(\bar{Q}_1) = \mathcal{H}^{\pi_1}(Q^{\pi_0}).$$
- For $k \geq 1$, since $Q^{\pi_k} = \mathcal{H}^{\pi_k}(Q^{\pi_k})$ and $Q^{\pi_k} = (\bar{Q}_{k+1} - (1-\beta_k)\bar{Q}_k)/\beta_k$ (cf. Line 3 of Algorithm 2), we have
$$\begin{aligned}
Q^{\pi_k} &= \mathcal{H}^{\pi_k}_{\eta_{k-1}}\left(\frac{\bar{Q}_{k+1} - (1-\beta_k)\bar{Q}_k}{\beta_k}\right) - \eta_{k-1} f(\pi_k) \\
&= \frac{1}{\beta_k}\mathcal{H}^{\pi_k}_{\eta_{k-1}}(\bar{Q}_{k+1}) - \frac{(1-\beta_k)}{\beta_k}\mathcal{H}^{\pi_k}_{\eta_{k-1}}(\bar{Q}_k) - \eta_{k-1} f(\pi_k) \\
&= \frac{1}{\beta_k}\mathcal{H}^{\pi_k}_{\eta_k}(\bar{Q}_{k+1}) - \frac{(1-\beta_k)}{\beta_k}\mathcal{H}^{\pi_k}_{\eta_{k-1}}(\bar{Q}_k) + \left[\frac{(1-\beta_k)\eta_{k-1} - \eta_k}{\beta_k}\right] f(\pi_k) \\
&= \frac{1}{\beta_k}\mathcal{H}^{\pi_k}_{\eta_k}(\bar{Q}_{k+1}) - \frac{(1-\beta_k)}{\beta_k}\mathcal{H}^{\pi_k}_{\eta_{k-1}}(\bar{Q}_k),
\end{aligned}$$
where the second equality follows from $\mathcal{H}^{\pi_k}_{\eta_{k-1}}(\cdot)$ being an affine operator, and the last equality follows from $\eta_k = (1-\beta_k)\eta_{k-1}$.
To proceed, observe that by Line 4 of Algorithm 2, we have $\mathcal{H}^{\pi_k}_{\eta_k}(\bar{Q}_{k+1}) \leq \mathcal{H}^{\pi_{k+1}}_{\eta_k}(\bar{Q}_{k+1})$ and $\mathcal{H}^{\pi_k}_{\eta_{k-1}}(\bar{Q}_k) \geq \mathcal{H}^{\pi_{k+1}}_{\eta_{k-1}}(\bar{Q}_k)$. Therefore,
$$\begin{aligned}
Q^{\pi_k} &\leq \frac{1}{\beta_k}\mathcal{H}^{\pi_{k+1}}_{\eta_k}(\bar{Q}_{k+1}) - \frac{(1-\beta_k)}{\beta_k}\mathcal{H}^{\pi_{k+1}}_{\eta_{k-1}}(\bar{Q}_k) \\
&= \mathcal{H}^{\pi_{k+1}}\left(\frac{\bar{Q}_{k+1} - (1-\beta_k)\bar{Q}_k}{\beta_k}\right) + \left(\frac{\eta_k - (1-\beta_k)\eta_{k-1}}{\beta_k}\right) f(\pi_{k+1}) \\
&= \mathcal{H}^{\pi_{k+1}}(Q^{\pi_k}),
\end{aligned}$$
where the last equality follows from Algorithm 2, Line 3, and $\eta_k = (1-\beta_k)\eta_{k-1}$. □

With Lemma 5.1 in hand, we next derive a contractive recursion for $\bar{Q}_k$.

Lemma 5.2. *Algorithm 2 satisfies*
$$\|Q^* - \bar{Q}_{k+1}\|_\infty \leq (1 - (1-\gamma)\beta_k)\|Q^* - \bar{Q}_k\|_\infty + \gamma\beta_k\eta_{k-1}\nu_{\max}, \quad \forall\, k \geq 1.$$

Proof of Lemma 5.2. An immediate implication of Lemma 5.1 is that $\bar{Q}_k \leq Q^{\pi_k}$ for all $k \geq 0$, since $\bar{Q}_k$ is a weighted average of a monotonically increasing sequence.

Let $\mathbf{1}$ denote the all-ones vector. Using Algorithm 2, Line 3, we obtain for all $k \geq 1$ that

$$
\begin{aligned}
Q^* - \bar{Q}_{k+1} &= (1-\beta_k)(Q^* - \bar{Q}_k) + \beta_k (Q^* - Q^{\pi_k}) \\
&= (1-\beta_k)(Q^* - \bar{Q}_k) + \beta_k (\mathcal{H}(Q^*) - \mathcal{H}^{\pi_k}(Q^{\pi_k})) && \text{(5a)} \\
&\leq (1-\beta_k)(Q^* - \bar{Q}_k) + \beta_k (\mathcal{H}(Q^*) - \mathcal{H}^{\pi_k}(\bar{Q}_k)) && \text{(5b)} \\
&= (1-\beta_k)(Q^* - \bar{Q}_k) + \beta_k (\mathcal{H}(Q^*) - \mathcal{H}^{\pi_k}_{\eta_{k-1}}(\bar{Q}_k)) + \beta_k \eta_{k-1} f(\pi_k) \\
&\leq (1-\beta_k)(Q^* - \bar{Q}_k) + \beta_k (\mathcal{H}(Q^*) - \mathcal{H}(\bar{Q}_k)) + \beta_k \eta_{k-1} f(\pi_k) && \text{(5c)} \\
&\leq (1-\beta_k)\|Q^* - \bar{Q}_k\|_\infty \mathbf{1} + \beta_k \|\mathcal{H}(Q^*) - \mathcal{H}(\bar{Q}_k)\|_\infty \mathbf{1} + \gamma \beta_k \eta_{k-1} \nu_{\max} \mathbf{1} \\
&\leq \big(1 - (1-\gamma)\beta_k\big)\|Q^* - \bar{Q}_k\|_\infty \mathbf{1} + \gamma \beta_k \eta_{k-1} \nu_{\max} \mathbf{1}, && \text{(5d)}
\end{aligned}
$$

where (5a) follows from the Bellman equations, (5b) follows from $\bar{Q}_k \leq Q^{\pi_k}$ and $\mathcal{H}^{\pi_k}$ being a monotonic operator, (5c) follows from $\mathcal{H}(\bar{Q}_k) \leq \mathcal{H}_{\eta_{k-1}}(\bar{Q}_k) = \mathcal{H}^{\pi_k}_{\eta_{k-1}}(\bar{Q}_k)$, and (5d) follows from $\mathcal{H}$ being a contraction mapping with respect to $\|\cdot\|_\infty$. Since the vector $Q^* - \bar{Q}_{k+1}$ has non-negative entries, the desired inequality follows. □

The remainder of the proof proceeds by repeatedly applying Lemma 5.2 to obtain a bound on $\|Q^* - \bar{Q}_k\|_\infty$, translating this bound into one on $\|V^* - V^{\pi_k}\|_\infty$, and then solving the resulting recursion. This step involves only algebraic manipulations, and we defer the details to Appendix B.4.

Remark. In view of the proof of Theorem 4.1, although DSPI captures gradient-based policy optimization algorithms (e.g., NPG and PDA), its analysis differs fundamentally from approaches based on optimization techniques. It relies directly on Bellman's core principles: the monotonicity of the Bellman operator yields monotonic improvement of the policies (cf. Lemma 5.1), and its contraction property induces a contractive recursion for the convergence error (cf. Lemma 5.2).

6 Conclusion

We introduce DSPI, an algorithmic framework that bridges classical dynamic programming algorithms, such as PI and dual-averaged PI, with modern policy optimization methods, such as NPG and PDA. This perspective enables a unified analysis based on monotonicity and contraction properties of the Bellman operator, yielding distribution-free global geometric convergence for NPG and PDA, and establishing finite termination for dual-averaged PI. The framework also extends to discounted MDPs with linear function approximation and to stochastic shortest path problems.

Looking ahead, an important direction is to extend our results to the average-reward setting, as discussed in Section 4.4. In addition, our current analysis assumes access to the exact Q-function Q^π for each policy iterate, which isolates the policy-improvement dynamics. Extending DSPI to the model-free RL setting, where Q^π must be estimated from sampled trajectories, is therefore a natural next step. Algorithmically, this can be achieved through an actor–critic framework. Since the policy update can be implemented explicitly via (3), we only need to provide an estimate of Q^{π_k} in Line 3 of Algorithm 2. Such an estimate can be readily constructed by a critic using standard TD-learning [39]. We denote the resulting estimate by Q_k. The actor then replaces Q^{π_k} by Q_k in its policy update. Concretely, this yields a two-loop actor–critic method, where the actor, as the outer-loop iterate, applies DSPI using an estimate of Q^{π_k} constructed in the inner loop through a TD-learning. The asymptotic convergence and finite-sample convergence rates of TD-learning have been extensively studied in the literature [38, 10, 16], and these results can be combined with our analysis of DSPI to derive an overall sample-complexity guarantee for the resulting actor–critic algorithm, similar to existing analyses of actor–critic methods [25, 30]. Nonetheless, determining whether the actor–critic algorithm can achieve minimax-optimal dependence on all problem parameters, namely $\mathcal{O}(\epsilon^{-2}(1-\gamma)^{-3}mn)$ [18], is, to the best of our knowledge, still broadly open in the existing analysis of actor-critic methods, and is an important direction for future work.

References

[1] Agarwal, A., Kakade, S. M., Lee, J. D., and Mahajan, G. (2021). On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *Journal of Machine Learning Research*, 22(98):1–76.

[2] Alfano, C. and Rebeschini, P. (2022). Linear convergence for natural policy gradient with log-linear policy parametrization. *Preprint arXiv:2209.15382*.

[3] Alfano, C., Yuan, R., and Rebeschini, P. (2023). A novel framework for policy mirror descent with general parameterization and linear convergence. *Advances in Neural Information Processing Systems*, 36:30681–30725.

[4] Bellman, R. (1957). *Dynamic Programming*. Princeton University Press.

[5] Bertsekas, D. P. (1995). *Dynamic programming and optimal control*, volume 1. Athena scientific Belmont, MA.

[6] Bertsekas, D. P. and Tsitsiklis, J. N. (1991). An analysis of stochastic shortest path problems. *Mathematics of Operations Research*, 16(3):580–595.

[7] Bertsekas, D. P. and Tsitsiklis, J. N. (1996). *Neuro-dynamic programming*. Athena Scientific.

[8] Bhandari, J. and Russo, D. (2021). On the linear convergence of policy gradient methods for finite MDPs. In *International Conference on Artificial Intelligence and Statistics*, pages 2386–2394. PMLR.

[9] Bhandari, J. and Russo, D. (2024). Global optimality guarantees for policy gradient methods. *Operations Research*, 72(5):1906–1927.

[10] Bhandari, J., Russo, D., and Singal, R. (2018). A finite-time analysis of temporal difference learning with linear function approximation. In *Conference on learning theory*, pages 1691–1692. PMLR.

[11] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

[12] Cayci, S., He, N., and Srikant, R. (2024). Convergence of entropy-regularized natural policy gradient with linear function approximation. *SIAM Journal on Optimization*, 34(3):2729–2755.

[13] Cen, S., Cheng, C., Chen, Y., Wei, Y., and Chi, Y. (2022). Fast global convergence of natural policy gradient methods with entropy regularization. *Operations Research*, 70(4):2563–2578.

[14] Chen, L., Luo, H., and Rosenberg, A. (2022). Policy optimization for stochastic shortest path. In *Conference on Learning Theory*, pages 982–1046. PMLR.

[15] Chen, Z. and Maguluri, S. T. (2025). An approximate policy iteration viewpoint of actor–critic algorithms. *Automatica*, 179:112395.

[16] Chen, Z., Maguluri, S. T., Shakkottai, S., and Shanmugam, K. (2024). A Lyapunov theory for finite-sample guarantees of Markovian stochastic approximation. *Operations Research*, 72(4):1352–1367.

[17] Geist, M., Scherrer, B., and Pietquin, O. (2019). A theory of regularized Markov decision processes. In *International Conference on Machine Learning*, pages 2160–2169. PMLR.

[18] Gheshlaghi Azar, M., Munos, R., and Kappen, B. K. H. J. (2013). Minimax PAC bounds on the sample complexity of reinforcement learning with a generative model. *Machine Learning*, 91:325–349.

[19] Gu, S., Holly, E., Lillicrap, T., and Levine, S. (2017). Deep reinforcement learning for robotic manipulation with asynchronous off-policy updates. In *2017 IEEE international conference on robotics and automation (ICRA)*, pages 3389–3396. IEEE.

[20] Howard, R. (1964). *Dynamic Programming and Markov Processes*. Massachusetts Institute of Technology.

[21] Ju, C. and Lan, G. (2022). Policy optimization over general state and action spaces. *Preprint arXiv:2211.16715*.

[22] Ju, C. and Lan, G. (2024). Strongly-polynomial time and validation analysis of policy gradient methods. *Preprint arXiv:2409.19437*.

[23] Ju, C. and Lan, G. (2026). Strongly-polynomial time and validation analysis of policy gradient methods. *Mathematical Programming*, pages 1–45.

[24] Kakade, S. M. (2001). A natural policy gradient. *Advances in neural information processing systems*, 14.

[25] Khodadadian, S., Jhunjhunwala, P. R., Varma, S. M., and Maguluri, S. T. (2022). On linear and super-linear convergence of natural policy gradient algorithm. *Systems & Control Letters*, 164:105214.

[26] Lan, G. (2023). Policy mirror descent for reinforcement learning: Linear convergence, new sampling complexity, and generalized problem classes. *Mathematical programming*, 198(1):1059–1106.

[27] Lan, G., Li, Y., and Zhao, T. (2023). Block policy mirror descent. *SIAM Journal on Optimization*, 33(3):2341–2378.

[28] Lee, J. and Ryu, E. K. (2025). Why policy gradient algorithms work for undiscounted total-reward MDPs. *Preprint arXiv:2510.18340*.

[29] Li, Y., Lan, G., and Zhao, T. (2022). First-order policy optimization for robust Markov decision process. *Preprint arXiv:2209.10579*.

[30] Li, Y., Lan, G., and Zhao, T. (2024). Homotopic policy mirror descent: policy convergence, algorithmic regularization, and improved sample complexity. *Mathematical Programming*, 207(1):457–513.

[31] Mei, J., Xiao, C., Szepesvari, C., and Schuurmans, D. (2020). On the global convergence rates of softmax policy gradient methods. In *International Conference on Machine Learning*, pages 6820–6829. PMLR.

[32] Puterman, M. L. (2014). *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons.

[33] Scherrer, B. (2014). Approximate policy iteration schemes: a comparison. In *International Conference on Machine Learning*, pages 1314–1322. PMLR.

[34] Scherrer, B. (2016). Improved and generalized upper bounds on the complexity of policy iteration. *Mathematics of Operations Research*, 41(3):758–774.

[35] Schulman, J., Levine, S., Abbeel, P., Jordan, M., and Moritz, P. (2015). Trust region policy optimization. In *International conference on machine learning*, pages 1889–1897. PMLR.

[36] Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. (2017). Proximal policy optimization algorithms. *Preprint arXiv:1707.06347*.

[37] Shani, L., Efroni, Y., and Mannor, S. (2020). Adaptive trust region policy optimization: Global convergence and faster rates for regularized MDPs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 5668–5675.

[38] Srikant, R. and Ying, L. (2019). Finite-time error bounds for linear stochastic approximation and TD-learning. In *Conference on Learning Theory*, pages 2803–2830.

[39] Sutton, R. S. and Barto, A. G. (2018). *Reinforcement Learning: An Introduction*. MIT press.

[40] Sutton, R. S., McAllester, D., Singh, S., and Mansour, Y. (1999). Policy gradient methods for reinforcement learning with function approximation. In *Proceedings of the 12th International Conference on Neural Information Processing Systems*, pages 1057–1063.

[41] Tsitsiklis, J. N. (1994). Asynchronous stochastic approximation and Q-learning. *Machine learning*, 16(3):185–202.

[42] Tsitsiklis, J. N. and Van Roy, B. (1997). An analysis of temporal-difference learning with function approximation. *IEEE transactions on automatic control*, 42(5):674–690.

[43] Vieillard, N., Kozuno, T., Scherrer, B., Pietquin, O., Munos, R., and Geist, M. (2020). Leverage the average: an analysis of kl regularization in reinforcement learning. *Advances in Neural Information Processing Systems*, 33:12163–12174.

[44] Xiao, L. (2022). On the convergence rates of policy gradient methods. *Journal of Machine Learning Research*, 23(282):1–36.

[45] Ye, Y. (2011). The simplex and policy-iteration methods are strongly polynomial for the Markov decision problem with a fixed discount rate. *Mathematics of Operations Research*, 36(4):593–603.

[46] Yu, H. and Bertsekas, D. P. (2013a). On boundedness of Q-learning iterates for stochastic shortest path problems. *Mathematics of Operations Research*, 38(2):209–227.

[47] Yu, H. and Bertsekas, D. P. (2013b). Q-learning and policy iteration algorithms for stochastic shortest path problems. *Annals of Operations Research*, 208(1):95–132.

[48] Yuan, R., Du, S. S., Gower, R. M., Lazaric, A., and Xiao, L. (2023). Linear convergence of natural policy gradient methods with log-linear policies. In *The Eleventh International Conference on Learning Representations*.

[49] Yurtsever, E., Lambert, J., Carballo, A., and Takeda, K. (2020). A survey of autonomous driving: Common practices and emerging technologies. *IEEE Access*, 8:58443–58469.

[50] Zhan, W., Cen, S., Huang, B., Chen, Y., Lee, J. D., and Chi, Y. (2023). Policy mirror descent for regularized reinforcement learning: A generalized framework with linear convergence. *SIAM Journal on Optimization*, 33(2):1061–1091.

A Related Work on Natural Policy Gradient

Table 1: Summary of existing analyses of natural policy gradient.

	Global Geometric Convergence	Distribution Free	Free of Regularization	Free of Adaptive Stepsize
Agarwal et al. [1]	✗	✓	✓	✓
Shani et al. [37]	✗	✓	✗	✓
Bhandari and Russo [8][3]	✓	✗	✓	✗
Cen et al. [13]	✓	✓	✗	✓
Lan [26]	✓	✓	✗	✓
Li et al. [30]	✓	✓	✗	✓
Zhan et al. [50]	✓	✓	✗	✓
Xiao [44]	✓	✗	✗	✓
Khodadadian et al. [25]	✓	✓	✓	✗
Chen and Maguluri [15]	✓	✓	✓	✗
Ju and Lan [23]	✓	✓	✓	✓[4]
This Work	✓	✓	✓	✓

We compare our work (last row) with existing results in terms of (i) global geometric convergence of the last iterate, (ii) dependence of the guarantees on the initial distribution (explicitly or via concentrability coefficients), (iii) the use of regularization in the MDP or the algorithm, and (iv) the need for trajectory-dependent stepsizes.

B Proofs of All Technical Results

B.1 The Geometric Convergence of Policy Iteration

Theorem B.1. *PI* (1) *satisfies* $\|V^* - V^{\pi_k}\|_\infty \leq \gamma^k \|V^* - V^{\pi_0}\|_\infty$ *for all* $k \geq 0$.

Proof of Theorem B.1. Our proof follows the two-step approach outlined in the last paragraph of Section 2.1.

- **Step 1: Monotonic Improvement:** Using the Bellman equation, we have $Q^{\pi_k} = \mathcal{H}^{\pi_k}(Q^{\pi_k}) \leq \mathcal{H}(Q^{\pi_k}) = \mathcal{H}^{\pi_{k+1}}(Q^{\pi_k})$, where the last equality follows from the update rule (1). Repeatedly applying $\mathcal{H}^{\pi_{k+1}}$ to both sides of $Q^{\pi_k} \leq \mathcal{H}^{\pi_{k+1}}(Q^{\pi_k})$ and using its monotonicity yields $Q^{\pi_k} \leq [\mathcal{H}^{\pi_{k+1}}]^n(Q^{\pi_k})$ for all $n \geq 0$. Letting $n \to \infty$ gives $Q^{\pi_k} \leq Q^{\pi_{k+1}}$.
- **Step 2: A Contractive Recursion:** Since $Q^{\pi_k} \leq Q^{\pi_{k+1}}$ and $\mathcal{H}^{\pi_{k+1}}$ is monotonic, we have

  $$Q^* - Q^{\pi_{k+1}} = \mathcal{H}(Q^*) - \mathcal{H}^{\pi_{k+1}}(Q^{\pi_{k+1}}) \leq \mathcal{H}(Q^*) - \mathcal{H}^{\pi_{k+1}}(Q^{\pi_k}) = \mathcal{H}(Q^*) - \mathcal{H}(Q^{\pi_k}).$$

  Since both sides are non-negative vectors, we obtain

  $$\|Q^* - Q^{\pi_{k+1}}\|_\infty \leq \|\mathcal{H}(Q^*) - \mathcal{H}(Q^{\pi_k})\|_\infty \leq \gamma \|Q^* - Q^{\pi_k}\|_\infty, \ \forall\, k \geq 0.$$

  Solving this recursion yields $\|Q^* - Q^{\pi_k}\|_\infty \leq \gamma^k \|Q^* - Q^{\pi_0}\|_\infty$ for all $k \geq 0$.

To translate the Q-function gap $\|Q^{\pi_k} - Q^*\|_\infty$to the value function gap $\|V^{\pi_k} - V^*\|_\infty$, observe that on the one hand, we have

$$V^*(s) - V^{\pi_k}(s) = \max_{a \in \mathcal{A}} Q^*(s,a) - \pi_k(s)^\top Q^{\pi_k}(s)$$

[3] Global geometric convergence is established under two regimes: (i) using line search for stepsize selection where the convergence bounds depend explicitly on the initial distribution, and (ii) without such dependence using adaptive stepsizes that depend on policy updates.

[4] While [23] does not use trajectory-dependent adaptive stepsizes, its choice of stepsize depends on the Bellman error (also called the advantage gap) of the initial policy, which is not required in our result (cf. Theorem 4.3).

$$
\begin{aligned}
&\leq \max_{a\in\mathcal{A}} Q^*(s,a) - \pi_k(s)^\top Q^{\pi_{k-1}}(s) && (Q^{\pi_{k-1}} \leq Q^{\pi_k} \text{ for all } k \geq 1)\\
&= \max_{a\in\mathcal{A}} Q^*(s,a) - \max_{a\in\mathcal{A}} Q^{\pi_{k-1}}(s,a) && \text{(update rule (1))}\\
&\leq \|Q^* - Q^{\pi_{k-1}}\|_\infty.
\end{aligned}
$$

On the other hand, we have

$$
Q^*(s,a) - Q^{\pi_0}(s,a) = \gamma \sum_{s'\in\mathcal{S}} p(s' \mid s,a)\big(V^*(s') - V^{\pi_0}(s')\big) \leq \gamma \|V^* - V^{\pi_0}\|_\infty.
$$

Combining the above two inequalities with the convergence rate of $\|Q^* - Q^{\pi_k}\|_\infty$ established in Step 2 yields $\|V^* - V^{\pi_k}\|_\infty \leq \gamma^k \|V^* - V^{\pi_0}\|_\infty$ for all $k \geq 0$. □

B.2 Properties of the Smoothed Bellman Operators

Lemma B.2. *For any policy π and $\eta \geq 0$, let F denote either $\mathcal{H}_\eta$ or $\mathcal{H}_\eta^\pi$. Then, for any $Q_1, Q_2 \in \mathbb{R}^{mn}$, we have*

(1) $\|F(Q_1) - F(Q_2)\|_\infty \leq \gamma \|Q_1 - Q_2\|_\infty$;
(2) $F(Q_1) \leq F(Q_2)$ *whenever* $Q_1 \leq Q_2$.

Proof of Lemma B.2. We prove the result only for the smoothed Bellman optimality operator $\mathcal{H}_\eta$. The proof for $\mathcal{H}_\eta^\pi$ follows by a similar argument.

(1) For any (s,a), we have

$$
\begin{aligned}
&|[\mathcal{H}_\eta(Q_1)](s,a) - [\mathcal{H}_\eta(Q_2)](s,a)|\\
=&\gamma \left| \sum_{s'\in\mathcal{S}} p(s' \mid s,a) \left(\max_{\mu\in\Delta(\mathcal{A})} \{\mu^\top Q_1(s') + \eta\nu(\mu)\} - \max_{\mu\in\Delta(\mathcal{A})} \{\mu^\top Q_2(s') + \eta\nu(\mu)\} \right) \right|\\
\leq&\gamma \sum_{s'\in\mathcal{S}} p(s' \mid s,a) \max_{\mu\in\Delta(\mathcal{A})} \left|\mu^\top (Q_1(s') - Q_2(s'))\right|\\
\leq&\gamma \|Q_1 - Q_2\|_\infty.
\end{aligned}
$$

Since the above bound holds for all (s,a), we obtain

$$
\|\mathcal{H}_\eta(Q_1) - \mathcal{H}_\eta(Q_2)\|_\infty \leq \gamma \|Q_1 - Q_2\|_\infty.
$$

(2) For any (s,a), we have

$$
\begin{aligned}
[\mathcal{H}_\eta(Q_1)](s,a) =& \mathcal{R}(s,a) + \gamma \sum_{s'\in\mathcal{S}} p(s' \mid s,a) \max_{\mu\in\Delta(\mathcal{A})} \{\mu^\top Q_1(s') + \eta\nu(\mu)\}\\
\leq& \mathcal{R}(s,a) + \gamma \sum_{s'\in\mathcal{S}} p(s' \mid s,a) \max_{\mu\in\Delta(\mathcal{A})} \{\mu^\top Q_2(s') + \eta\nu(\mu)\}\\
=& [\mathcal{H}_\eta(Q_2)](s,a).
\end{aligned}
$$

□

B.3 Proof of Theorem 4.3

The proof follows by combining Proposition 3.1 with Theorem 4.1. By the definition of α_k, we have $\beta_0 = \alpha_0/\alpha_0 = 1$, and for every $k \geq 1$,

$$
\beta_k = \frac{\alpha_k}{\sum_{i=0}^k \alpha_i} = \frac{\beta\alpha_0/(1-\beta)^k}{\alpha_0 + \sum_{i=1}^k \beta\alpha_0/(1-\beta)^i} = \frac{\beta\alpha_0/(1-\beta)^k}{\alpha_0 + \alpha_0\left((1-\beta)^{-k} - 1\right)} = \beta.
$$

Further, we choose $\alpha_0 = \log(m)$, $\tau = 1/\alpha_0$, and $\nu(\cdot) = h(\cdot)$. Thus, by Theorem 4.1, for every $k \geq 1$,

$$
\begin{aligned}
\|V^* - V^{\pi_k}\|_\infty &\leq (1-(1-\gamma)\beta)^{k-1} \left(\gamma \|V^* - V^{\pi_0}\|_\infty + \frac{\max_{\mu\in\Delta(\mathcal{A})} h(\mu)}{\log(m)} \right)\\
&= (1-(1-\gamma)\beta)^{k-1} \left(\gamma \|V^* - V^{\pi_0}\|_\infty + 1 \right).
\end{aligned}
$$

B.4 Details in the Proof of Theorem 4.1

Repeatedly apply Lemma 5.2 yields

$$\|Q^* - \bar{Q}_{k+1}\|_\infty \leq \prod_{j=1}^{k}\big(1-(1-\gamma)\beta_j\big)\|Q^* - \bar{Q}_1\|_\infty + \gamma\nu_{\max}\sum_{i=1}^{k}\eta_{i-1}\beta_i\prod_{j=i+1}^{k}\big(1-(1-\gamma)\beta_j\big) \tag{6}$$

To translate this bound into a bound on the value function gap, we use the following lemma.

Lemma B.3. *Algorithm 2 satisfies:*

(1) $\|V^* - V^{\pi_{k+1}}\|_\infty \leq \|Q^* - \bar{Q}_{k+1}\|_\infty + \eta_k\nu_{\max}$ *for all* $k \geq 0$*;*
(2) $\|Q^* - \bar{Q}_1\|_\infty \leq \gamma\|V^* - V^{\pi_0}\|_\infty$.

Proof of Lemma B.3. The proof uses the relation between the value function and corresponding Q-function, and the update rule of Algorithm 2.

(1) For any $k \geq 0$ and $s \in \mathcal{S}$, we have

$$\begin{aligned}
V^*(s) - V^{\pi_{k+1}}(s) &= \max_{\mu\in\Delta(\mathcal{A})}\mu^\top Q^*(s) - \pi_{k+1}(s)^\top Q^{\pi_{k+1}}(s) \\
&\leq \max_{\mu\in\Delta(\mathcal{A})}\mu^\top Q^*(s) - \pi_{k+1}(s)^\top \bar{Q}_{k+1}(s) && \text{(Lemma 5.1)} \\
&= \max_{\mu\in\Delta(\mathcal{A})}\mu^\top Q^*(s) - \max_{\mu\in\Delta(\mathcal{A})}\big\{\mu^\top\bar{Q}_{k+1}(s) + \eta_k\nu(\mu)\big\} + \eta_k\nu(\pi_{k+1}(s)) \\
& && \text{(Eq. (3))} \\
&\leq \max_{\mu\in\Delta(\mathcal{A})}\mu^\top(Q^*(s) - \bar{Q}_{k+1}(s)) + \eta_k\nu(\pi_{k+1}(s)) \\
&\leq \|Q^* - \bar{Q}_{k+1}\|_\infty + \eta_k\nu_{\max}.
\end{aligned}$$

Since the previous inequality holds for all $s \in \mathcal{S}$ and $V^* - V^{\pi_{k+1}} \geq 0$, we have

$$\|V^* - V^{\pi_{k+1}}\|_\infty \leq \|Q^* - \bar{Q}_{k+1}\|_\infty + \eta_k\nu_{\max},\ \forall\, k \geq 0.$$

(2) Since $\beta_0 = 1$, we have $\bar{Q}_1 = Q^{\pi_0}$. Now, for any (s,a), observe that

$$Q^*(s,a) - Q^{\pi_0}(s,a) = \gamma\sum_{s'\in\mathcal{S}}p(s' \mid s,a)(V^*(s') - V^{\pi_0}(s')) \leq \gamma\|V^* - V^{\pi_0}\|_\infty,$$

which implies $\|Q^* - Q^{\pi_0}\|_\infty \leq \gamma\|V^* - V^{\pi_0}\|_\infty$.

□

Combining Lemma B.3 with (6), we have for every $k \geq 0$ that

$$\|V^* - V^{\pi_{k+1}}\|_\infty \leq \prod_{j=1}^{k}(1-(1-\gamma)\beta_j)\gamma\|V^* - V^{\pi_0}\|_\infty + \gamma\nu_{\max}\sum_{i=1}^{k}\beta_i\eta_{i-1}\prod_{j=i+1}^{k}(1-(1-\gamma)\beta_j) + \eta_k\nu_{\max}. \tag{7}$$

We now use the preceding inequality to establish both asymptotic convergence and convergence rates for DSPI.

B.4.1 The Asymptotic Convergence

Since $\eta_k = \tau \prod_{j=1}^{k}(1-\beta_j)$ for all $k \geq 1$, we can write (7) as

$$\begin{aligned}
\|V^* - V^{\pi_{k+1}}\|_\infty &\leq \prod_{j=1}^{k}(1-(1-\gamma)\beta_j)\gamma\|V^* - V^{\pi_0}\|_\infty \\
&\quad + \gamma\nu_{\max}\tau\sum_{i=1}^{k}\beta_i\prod_{j=1}^{i-1}(1-\beta_j)\prod_{j=i+1}^{k}(1-(1-\gamma)\beta_j) + \nu_{\max}\tau\prod_{j=1}^{k}(1-\beta_j) \\
&\leq \prod_{j=1}^{k}(1-(1-\gamma)\beta_j)\gamma\|V^* - V^{\pi_0}\|_\infty \\
&\quad + \gamma\tau\nu_{\max}\prod_{j=1}^{k}(1-(1-\gamma)\beta_j)\sum_{i=1}^{k}\frac{\beta_i}{1-(1-\gamma)\beta_i} + \nu_{\max}\tau\prod_{j=1}^{k}(1-\beta_j) \\
&\leq \exp\left(-(1-\gamma)\sum_{i=1}^{k}\beta_i\right)\gamma\|V^* - V^{\pi_0}\|_\infty \\
&\quad + \gamma\tau\nu_{\max}\exp\left(-(1-\gamma)\sum_{i=1}^{k}\beta_i\right)\left(\sum_{i=1}^{k}\beta_i\right) + \nu_{\max}\tau\exp\left(-\sum_{i=1}^{k}\beta_i\right),
\end{aligned}$$

where the last inequality follows from $\beta_k \in (0,1]$ and $1+x \leq e^x$ for any $x \in \mathbb{R}$. Since $\lim_{x\to\infty} e^{-x} = 0$ and $\lim_{x\to\infty} xe^{-x} = 0$, as long as $\sum_{k=0}^{\infty}\beta_k = \infty$, we have $\lim_{k\to\infty}\|V^* - V^{\pi_k}\|_\infty = 0$.

B.4.2 Convergence Rates

We begin with Eq. (7) restated below:

$$\begin{aligned}
\|V^* - V^{\pi_{k+1}}\|_\infty \leq &\underbrace{\prod_{j=1}^{k}(1-(1-\gamma)\beta_j)\gamma\|V^* - V^{\pi_0}\|_\infty}_{:=E_1} \\
&\underbrace{+\gamma\nu_{\max}\sum_{i=1}^{k}\beta_i\eta_{i-1}\prod_{j=i+1}^{k}(1-(1-\gamma)\beta_j)}_{:=E_2} + \underbrace{\eta_k\nu_{\max}}_{:=E_3}.
\end{aligned}$$

When using $\beta_0 = 1$ and $\beta_k \equiv \beta \in (0,1]$ for all $k \geq 1$, we have

$$E_1 = (1-(1-\gamma)\beta)^k\gamma\|V^* - V^{\pi_0}\|_\infty, \quad \text{and} \quad E_3 = \tau(1-\beta)^k\nu_{\max}.$$

For the term E_2, we have

$$\begin{aligned}
E_2 &= \gamma\nu_{\max}\sum_{i=1}^{k}\beta_i\eta_{i-1}\prod_{j=i+1}^{k}(1-(1-\gamma)\beta_j) \\
&= \gamma\beta\nu_{\max}\tau\sum_{i=1}^{k}(1-\beta)^{i-1}(1-(1-\gamma)\beta)^{k-i} \\
&= \gamma\beta\nu_{\max}\tau(1-(1-\gamma)\beta)^{k-1}\sum_{i=1}^{k}\left(\frac{1-\beta}{1-(1-\gamma)\beta}\right)^{i-1} \\
&= \gamma\beta\nu_{\max}\tau(1-(1-\gamma)\beta)^{k-1}\frac{1-\left(\frac{1-\beta}{1-(1-\gamma)\beta}\right)^k}{1-\left(\frac{1-\beta}{1-(1-\gamma)\beta}\right)}
\end{aligned}$$

$$= \nu_{\max}\tau \left[(1-(1-\gamma)\beta)^k - (1-\beta)^k\right].$$

Altogether, we have

$$\|V^* - V^{\pi_{k+1}}\|_\infty \leq E_1 + E_2 + E_3 \leq (1-(1-\gamma)\beta)^k \left(\gamma\|V^* - V^{\pi_0}\|_\infty + \nu_{\max}\tau\right), \ \forall\, k \geq 0.$$

B.5 Proof of Theorem 4.5

As discussed, dual-averaged PI (DPI) is a particular case of DSPI corresponding to $\nu(\cdot) \equiv 0$; we present it below for clarity. Moreover, we restrict the iteration to deterministic policies in order to establish finite termination.

Algorithm 3 Dual-Averaged Policy Iteration

1: **Input:** Initialize π_0 at a deterministic policy and $\bar{Q}^0 = \mathbf{0}$
2: **for** $k = 0, 1, 2, \cdots$ **do**
3: Maintain a running average of past Q-functions: $\bar{Q}_{k+1} = (1-\beta_k)\bar{Q}_k + \beta_k Q^{\pi_k}$, where $\beta_k \in (0, 1]$ is the stepsize.
4: Choose a deterministic π_{k+1} such that $\mathcal{H}^{\pi_{k+1}}(\bar{Q}_{k+1}) = \mathcal{H}(\bar{Q}_{k+1})$.
5: **end for**

The key to showing that DPI terminates in a finite number of iterations is to identify a $k^* \in \mathbb{N}$ such that, every k^* iterations, at least one suboptimal action is eliminated in some state. To this end, we first construct a certificate that detects when two deterministic policies differ. For a given policy π, let $\mathcal{T}^\pi : \mathbb{R}^n \to \mathbb{R}^n$ denote its associated Bellman operator, defined by

$$[\mathcal{T}^\pi(V)](s) = \mathcal{R}^\pi(s) + \gamma \sum_{s' \in \mathcal{S}} P^\pi(s, s')V(s'), \quad s \in \mathcal{S},$$

with $\mathcal{R}^\pi$ and P^π given by

$$\mathcal{R}^\pi(s) = \sum_{a \in \mathcal{A}} \pi(a|s)\mathcal{R}(s, a), \quad \text{and} \quad P^\pi(s, s') = \sum_{a \in \mathcal{A}} \pi(a|s)p(s'|s, a), \quad s, s' \in \mathcal{S},$$

respectively. We define the certificate $G^\pi : \mathcal{S} \to \mathbb{R}$ as

$$G^\pi(s) \coloneqq V^*(s) - [\mathcal{T}^\pi(V^*)](s), \quad s \in \mathcal{S}. \tag{8}$$

Using the definition of $\mathcal{T}^\pi(\cdot)$ and the relation between Q^* and V^*, we can express the certificate G^π equivalently as

$$G^\pi(s) = V^*(s) - \pi(s)^\top Q^*(s), \quad s \in \mathcal{S}. \tag{9}$$

We will use this latter expression in our analysis. Furthermore, since $V^*(s) = \max_{a \in \mathcal{A}} Q^*(s, a)$, for every $s \in \mathcal{S}$, it follows that $G^\pi(s) \geq 0$ for any policy π and is zero if and only if the policy π is optimal. The following lemma shows how the certificate G^π detects when two deterministic policies differ.

Lemma B.4. *Let π_0 and π_1 be two deterministic policies. Let $s \in \mathcal{S}$ be such that $G^{\pi_0}(s) \neq G^{\pi_1}(s)$. Then $\pi_1(a_0|s) = 0$, where $a_0 \in \mathcal{A}$ is such that $\pi_0(a_0|s) = 1$.*

Proof of Lemma B.4. Let $a_1 \in \mathcal{A}$ be the action such that $\pi_1(a_1|s) = 1$. Then, the condition $G^{\pi_0}(s) \neq G^{\pi_1}(s)$, i.e., $V^*(s) - \pi_0(s)^\top Q^*(s) \neq V^*(s) - \pi_1(s)^\top Q^*(s)$, is equivalent to

$$\sum_{a \in \mathcal{A}} \pi_0(a|s)Q^*(s, a) \neq \sum_{a \in \mathcal{A}} \pi_1(a|s)Q^*(s, a) \quad \iff \quad Q^*(s, a_0) \neq Q^*(s, a_1).$$

Therefore, we must have $a_0 \neq a_1$. Since π_0 and π_1 are deterministic, we conclude that $\pi_1(a_0|s) = 0$ where $\pi_0(a_0|s) = 1$. □

Our next result gives a bound on the certificate corresponding to the policies generated by the DPI algorithm with stepsize chosen as

$$\beta_k = \begin{cases} 1, \text{ if } k = rk^*, r \in \{0\} \cup \mathbb{N}, \\ \beta, \text{ otherwise,} \end{cases} \tag{10}$$

Lemma B.5. *Let $\{\pi_k\}$ be the sequence of policies generated by Algorithm 3 with $\{\beta_k\}$ as defined in* (10)*. Then, for every $k \geq 1$ and $r \in \{0\} \cup \mathbb{N}$,*

$$\|G^{\pi_{rk^*+k}}\|_\infty \leq \frac{\gamma}{1-\gamma}\big(1-(1-\gamma)\beta\big)^{k-1}\|G^{\pi_{rk^*}}\|_\infty.$$

Proof of Lemma B.5. To begin with, we observe that $\bar{Q}_j \leq Q^*$, for every $j \geq 1$, since $\bar{Q}_j$ is the weighted average of $Q^{\pi_0}, \ldots, Q^{\pi_{j-1}}$ that are monotonically increasing. Then, for any $j \geq 1$, we have for every $s \in \mathcal{S}$ that

$$\begin{aligned}
G^{\pi_j}(s) &= V^*(s) - \pi_j(s)^\top Q^*(s) \\
&= \max_{a\in\mathcal{A}} Q^*(s,a) - \sum_{a\in\mathcal{A}} \pi_j(a|s) Q^*(s,a) \\
&\leq \max_{a\in\mathcal{A}} Q^*(s,a) - \sum_{a\in\mathcal{A}} \pi_j(a|s) \bar{Q}_j(s,a) && (\bar{Q}_j \leq Q^*) \\
&= \max_{a\in\mathcal{A}} Q^*(s,a) - \max_{a\in\mathcal{A}} \bar{Q}_j(s,a) && \text{(Line 4, Algorithm 3)} \\
&\leq \max_{a\in\mathcal{A}} \big\{Q^*(s,a) - \bar{Q}_j(s,a)\big\} \\
&\leq \|Q^* - \bar{Q}_j\|_\infty
\end{aligned}$$

Since $G^{\pi_j}(s) \geq 0$ for every $s \in \mathcal{S}$, it follows that

$$\|G^{\pi_j}\|_\infty \leq \|Q^* - \bar{Q}_j\|_\infty \tag{11}$$

As the DPI algorithm corresponds to $\nu(\cdot) \equiv 0$, the one-step contractive inequality from Lemma 5.2 reduces to

$$\|Q^* - \bar{Q}_{j+1}\|_\infty \leq (1-(1-\gamma)\beta_j)\|Q^* - \bar{Q}_j\|_\infty, \quad \forall\, j \geq 1.$$

By repeated application of the above inequality, we get for any $k \geq 1$ that

$$\|Q^* - \bar{Q}_{j+k}\|_\infty \leq \prod_{i=j}^{j+k-1} (1-(1-\gamma)\beta_i)\|Q^* - \bar{Q}_j\|_\infty, \quad \forall\, j \geq 1.$$

Combining this with (11) gives, for every $r \in \{0\} \cup \mathbb{N}$, that

$$\begin{aligned}
\|G^{\pi_{rk^*+k}}\|_\infty &\leq \prod_{i=rk^*+1}^{rk^*+k-1} \big(1-(1-\gamma)\beta_i\big)\|Q^* - \bar{Q}_{rk^*+1}\|_\infty \\
&\leq \big(1-(1-\gamma)\beta\big)^{k-1}\|Q^* - Q^{\pi_{rk^*}}\|_\infty
\end{aligned} \tag{12}$$

where we used $\bar{Q}_{rk^*+1} = Q^{\pi_{rk^*}}$ and $(1-(1-\gamma)\beta_{rk^*}) = (1-(1-\gamma)) < (1-(1-\gamma)\beta)$ that follow from $\beta_{rk^*} = 1$. To relate the right-hand side to the certificate $\|G^{\pi_{rk^*}}\|_\infty$, we proceed as follows.

$$\begin{aligned}
\|Q^* - Q^{\pi_{rk^*}}\|_\infty &= \|Q^* - \mathcal{H}^{\pi_{rk^*}}(Q^*) + \mathcal{H}^{\pi_{rk^*}}(Q^*) - Q^{\pi_{rk^*}}\|_\infty \\
&= \|Q^* - \mathcal{H}^{\pi_{rk^*}}(Q^*) + \mathcal{H}^{\pi_{rk^*}}(Q^*) - \mathcal{H}^{\pi_{rk^*}}(Q^{\pi_{rk^*}})\|_\infty \\
&\leq \|Q^* - \mathcal{H}^{\pi_{rk^*}}(Q^*)\|_\infty + \|\mathcal{H}^{\pi_{rk^*}}(Q^*) - \mathcal{H}^{\pi_{rk^*}}(Q^{\pi_{rk^*}})\|_\infty \\
&\leq \|Q^* - \mathcal{H}^{\pi_{rk^*}}(Q^*)\|_\infty + \gamma\|Q^* - Q^{\pi_{rk^*}}\|_\infty,
\end{aligned}$$

which implies

$$\|Q^* - Q^{\pi_{rk^*}}\|_\infty \leq \frac{1}{1-\gamma}\|Q^* - \mathcal{H}^{\pi_{rk^*}}(Q^*)\|_\infty. \tag{13}$$

The Bellman residual now can be bounded by the certificate $\|G^{\pi_{rk^*}}\|_\infty$ as follows. Since $Q^* = \mathcal{H}(Q^*)$, it follows that for every $(s,a) \in \mathcal{S}\times\mathcal{A}$

$$\begin{aligned}
Q^*(s,a) - [\mathcal{H}^{\pi_{rk^*}}(Q^*)](s,a) &= [\mathcal{H}(Q^*)](s,a) - [\mathcal{H}^{\pi_{rk^*}}(Q^*)](s,a) \\
&= \gamma \sum_{s'\in\mathcal{S}} p(s'|s,a)\left\{\max_{a'\in\mathcal{A}} Q^*(s',a') - \pi_{rk^*}(s')^\top Q^*(s')\right\}
\end{aligned}$$

$$= \gamma \sum_{s' \in \mathcal{S}} p(s'|s,a) G^{\pi_{rk^*}}(s')$$
$$\leq \gamma \|G^{\pi_{rk^*}}\|_\infty \sum_{s' \in \mathcal{S}} p(s'|s,a)$$
$$= \gamma \|G^{\pi_{rk^*}}\|_\infty$$

Again, since $Q^*(s,a) - [\mathcal{H}^{\pi_{rk^*}}(Q^*)](s,a) \geq 0$ for every $(s,a) \in \mathcal{S} \times \mathcal{A}$, we can conclude that

$$\|Q^* - \mathcal{H}^{\pi_{rk^*}}(Q^*)\|_\infty \leq \gamma \|G^{\pi_{rk^*}}\|_\infty$$

Combining this bound with (13), we get

$$\|Q^* - Q^{\pi_{rk^*}}\|_\infty \leq \frac{\gamma}{1-\gamma} \|G^{\pi_{rk^*}}\|_\infty$$

Applying this bound in (12) yields

$$\|G^{\pi_{rk^*+k}}\|_\infty \leq \frac{\gamma}{1-\gamma} \big(1 - (1-\gamma)\beta\big)^{k-1} \|G^{\pi_{rk^*}}\|_\infty$$

□

Corollary B.6. *Let $\{\pi_k\}$ be the sequence of policies generated by Algorithm 3 with $\{\beta_k\}$ as defined in* (10)*. Then, for every $r \in \{0\} \cup \mathbb{N}$ and every $k \geq (r+1)k^*$,*

$$\|G^{\pi_k}\|_\infty \leq \gamma \, \|G^{\pi_{rk^*}}\|_\infty.$$

In particular, if $\|G^{\pi_{rk^*}}\|_\infty > 0$, then $\|G^{\pi_k}\|_\infty < \|G^{\pi_{rk^*}}\|_\infty$ for every $k \geq (r+1)k^*$.

Proof of Corollary B.6. Fix $r \in \{0\} \cup \mathbb{N}$ and $k \geq (r+1)k^*$, and write $k = rk^* + j$ with $j := k - rk^* \geq k^*$. By the definition of k^*, we have

$$j - 1 \;\geq\; k^* - 1 \;=\; \left\lceil \frac{1}{\beta(1-\gamma)} \log\left(\frac{1}{1-\gamma}\right) \right\rceil \;\geq\; \frac{1}{\beta(1-\gamma)} \log\left(\frac{1}{1-\gamma}\right).$$

Since $1 + x \leq e^x$ for all $x \in \mathbb{R}$, it follows that

$$\big(1 - (1-\gamma)\beta\big)^{j-1} \;\leq\; e^{-(1-\gamma)\beta(j-1)} \;\leq\; \exp\left(-\log\left(\frac{1}{1-\gamma}\right)\right) \;=\; 1 - \gamma,$$

and therefore

$$\frac{\gamma}{1-\gamma} \big(1 - (1-\gamma)\beta\big)^{j-1} \;\leq\; \gamma. \tag{14}$$

Applying Lemma B.5 with the index rk^* and using (14), we obtain

$$\|G^{\pi_k}\|_\infty \;=\; \|G^{\pi_{rk^*+j}}\|_\infty \;\leq\; \frac{\gamma}{1-\gamma} \big(1 - (1-\gamma)\beta\big)^{j-1} \|G^{\pi_{rk^*}}\|_\infty \;\leq\; \gamma \, \|G^{\pi_{rk^*}}\|_\infty.$$

The second claim follows since $\gamma < 1$ and $\|G^{\pi_{rk^*}}\|_\infty > 0$. □

We are now ready to complete the proof. Fix $r \in \{0\} \cup \mathbb{N}$ and suppose that π_{rk^*} is not optimal. Then, $\|G^{\pi_{rk^*}}\|_\infty > 0$. Let $s_r \in \mathcal{S}$ satisfy

$$G^{\pi_{rk^*}}(s_r) = \|G^{\pi_{rk^*}}\|_\infty,$$

and let $a_r \in \mathcal{A}$ be the action such that $\pi_{rk^*}(a_r|s_r) = 1$. By Corollary B.6, for every $k \geq (r+1)k^*$, we have

$$G^{\pi_k}(s_r) \leq \|G^{\pi_k}\|_\infty < \|G^{\pi_{rk^*}}\|_\infty = G^{\pi_{rk^*}}(s_r).$$

Therefore, $G^{\pi_k}(s_r) \neq G^{\pi_{rk^*}}(s_r)$ for every $k \geq (r+1)k^*$. Applying Lemma B.4 with $\pi_0 = \pi_{rk^*}$ and $\pi_1 = \pi_k$, we conclude that

$$\pi_k(a_r|s_r) = 0, \quad \forall\, k \geq (r+1)k^*. \tag{15}$$

Moreover, $0 < G^{\pi_{rk^*}}(s_r) = V^*(s_r) - Q^*(s_r, a_r)$, which shows that a_r is a suboptimal action at state s_r, and is permanently eliminated after the next k^* iterations.

It remains to verify that the state-action pair eliminated in each block of k^* iterations is new. Let $0 \leq q < r$, and suppose that both π_{qk^*} and π_{rk^*} are not optimal. Since $rk^* \geq (q+1)k^*$, applying (15) to the pair (s_q, a_q) gives $\pi_{rk^*}(a_q|s_q) = 0$. On the other hand, the pair (s_r, a_r) is chosen so that $\pi_{rk^*}(a_r|s_r) = 1$. Consequently, $(s_r, a_r) \neq (s_q, a_q)$. Therefore, although the states s_r need not be distinct, every nonoptimal restart policy identifies a new suboptimal state-action pair that is permanently eliminated after the following k^* iterations.

At each state, there are at most $m-1$ suboptimal actions because at least one action is optimal. Hence, there are at most $n(m-1)$ suboptimal state-action pairs in total. Suppose, that π_{rk^*} is not optimal for every $r \in \{0, \ldots, n(m-1)\}$. The preceding argument would then produce $n(m-1)+1$ distinct suboptimal state-action pairs, which is a contradiction. It thus follows that DPI finds an optimal policy after at most $n(m-1)k^*$ iterations.

C Natural Policy Gradient with Linear Function Approximation

In this section, we analyze the NPG algorithm with linear function approximation. In this case, the policy is parameterized as

$$\pi_\theta(a|s) = \frac{\exp(\phi_{s,a}^\top \theta)}{\sum_{a' \in \mathcal{A}} \exp(\phi_{s,a'}^\top \theta)}, \quad \forall\, (s,a) \in \mathcal{S} \times \mathcal{A},$$

where $\theta \in \mathbb{R}^d$ is the parameter vector ($d \leq mn$), and $\phi_{s,a} \in \mathbb{R}^d$ denotes the feature vector associated with the state-action pair (s,a). Let $\Phi \in \mathbb{R}^{mn \times d}$ be the matrix whose rows are the feature vectors. We assume without loss of generality that the columns of Φ are linearly independent.

We next present NPG with linear function approximation [1], where $d_\rho^\pi \in \Delta(\mathcal{S})$ denotes the discounted state visitation distribution with initial state distribution ρ.

Algorithm 4 Natural Policy Gradient with Linear Function Approximation

1: **Input:** Initialize $\theta_0 = \mathbf{0}$ and π_0 as the uniform policy.
2: **for** $k = 0, 1, 2, 3, \cdots$ **do**
3: $\quad w_k^* = \arg\min_{w \in \mathbb{R}^d} \mathbb{E}_{s \sim d_\rho^{\pi_k}, a \sim \pi_k(\cdot|s)} \left[Q^{\pi_k}(s,a) - \phi_{s,a}^\top w \right]^2$
4: $\quad \theta_{k+1} = \theta_k + \alpha_k w_k^*$
5: $\quad \pi_{k+1}(s) = \arg\max_{\mu \in \Delta(\mathcal{A})} \left\{ \mu^\top [\Phi \theta_{k+1}](s) + h(\mu) \right\}$ for any $s \in \mathcal{S}$
6: **end for**

C.1 Algorithm Reformulation

Similar to the tabular setting, we reformulate the algorithm as DSPI with linear function approximation, as given below.

Algorithm 5 Doubly Smoothed Policy Iteration with Linear Function Approximation

1: **Input:** Initialize $\theta_0 = \mathbf{0}$, $\bar{W}_0 = \mathbf{0}$, and $\pi_0(s) \in \arg\max_{\mu \in \Delta(\mathcal{A})} \nu(\mu)$ for any $s \in \mathcal{S}$.
2: **for** $k = 0, 1, 2, 3, \cdots$ **do**
3: $\quad w_k^* = \arg\min_w \mathbb{E}_{s \sim d_\rho^{\pi_k}, a \sim \pi_k(\cdot|s)} \left[Q^{\pi_k}(s,a) - \phi_{s,a}^\top w \right]^2$ and $W_k^* = \Phi w_k^*$
4: $\quad \bar{W}_{k+1} = (1-\beta_k)\bar{W}_k + \beta_k W_k^*$
5: $\quad \pi_{k+1}(s) \in \arg\max_{\mu \in \Delta} \left\{ \mu^\top \bar{W}_{k+1}(s) + \tau \prod_{j=1}^k (1-\beta_j) \nu(\mu) \right\}$ for any $s \in \mathcal{S}$
6: **end for**

The following result is analogous to Proposition 3.1

Proposition C.1. *When $\beta_k = \alpha_k / \sum_{i=0}^k \alpha_i$, $\tau = 1/\alpha_0$, and $\nu(\cdot) = h(\cdot)$, the sequence of policies generated by Algorithm 5 coincides with that generated by Algorithm 4.*

Proof of Proposition C.1. Denote $\Theta_k = \Phi\theta_k$. As stated in Line 3 of Algorithm 5, $W_k^* = \Phi w_k^*$. Then Line 4 of Algorithm 4 is equivalent to $\Theta_{k+1} = \Theta_k + \alpha_k W_k^*$, since Φ has linearly independent columns. Recursing yields $\Theta_{k+1} = \sum_{i=0}^k \alpha_i W_i^*$.

Following arguments similar to those in Proposition 3.1, we obtain

$$\begin{aligned}\pi_{k+1}(s) &= \operatorname*{arg\,max}_{\mu\in\Delta(\mathcal{A})} \left\{ \mu^\top \left(\sum_{i=0}^k \alpha_i W_i^*(s) \right) + h(\mu) \right\} \\ &= \operatorname*{arg\,max}_{\mu\in\Delta(\mathcal{A})} \left\{ \mu^\top \left(\frac{\sum_{i=0}^k \alpha_i W_i^*(s)}{\sum_{i=0}^k \alpha_i} \right) + \frac{\tau\alpha_0}{\sum_{i=0}^k \alpha_i} h(\mu) \right\}, \quad \forall\, s \in \mathcal{S}.\end{aligned}$$

For the above to be equivalent to Line 5 of Algorithm 5, it remains to show that $\bar{W}_{k+1} = \sum_{i=0}^k \alpha_i W_i^* / \sum_{i=0}^k \alpha_i$ and $\prod_{j=1}^k (1-\beta_j) = \alpha_0 / \sum_{i=0}^k \alpha_i$. The latter identity has already been proved in Proposition 3.1 and is omitted here.

We now prove the former by induction, following the same argument as in the proof of Proposition 3.1. The equality holds for $k = 0$, since $\beta_0 = 1$, which gives $\bar{W}_1 = W_0^*$. Assume that $\bar{W}_{n+1} = \sum_{i=0}^n \alpha_i W_i^* / \sum_{i=0}^n \alpha_i$ for some $n \in \mathbb{N}$. Then,

$$\begin{aligned}\bar{W}_{n+2} &= (1-\beta_{n+1})\bar{W}_{n+1} + \beta_{n+1} W_{n+1}^* \\ &= \frac{\sum_{i=0}^n \alpha_i}{\sum_{i=0}^{n+1} \alpha_i} \cdot \frac{\sum_{i=0}^n \alpha_i W_i^*}{\sum_{i=0}^n \alpha_i} + \frac{\alpha_{n+1}}{\sum_{i=0}^{n+1} \alpha_i} W_{n+1}^* \\ &= \frac{\sum_{i=0}^{n+1} \alpha_i W_i^*}{\sum_{i=0}^{n+1} \alpha_i}.\end{aligned}$$

This completes the induction and establishes the desired equality. □

C.2 Finite-Time Analysis of DSPI with Linear Function Approximation

Observe that, in contrast to the tabular setting where the natural gradient can be expressed in terms of the Q-function, Algorithm 5 uses only an approximation W_k^* to the Q-function at iteration k. We therefore write $W_k^* = Q^{\pi_k} + \epsilon_k$, where $\epsilon_k := W_k^* - Q^{\pi_k}$ captures the function approximation error. To separate approximation error from convergence, we assume there exists $\epsilon > 0$ such that $\sup_k \|\epsilon_k\|_\infty \le \epsilon$, where ϵ is analogous to the approximation error bound ϵ_{approx} defined in [1].

In model-free RL, the approximation W_k^* would need to be estimated via temporal-difference learning with linear function approximation [42, 38, 10]. This is beyond the scope of this work, as we focus on the policy optimization dynamics.

For simplicity, we consider only a constant stepsize in Algorithm 5. We now state the main convergence result, whose proof is presented in Appendix C.4.

Theorem C.2. *Consider $\{\pi_k\}$ generated by Algorithm 5. When choosing $\beta_0 = 1$ and $\beta_k \equiv \beta \in (0, 1]$ for all $k \ge 1$, we have*

$$\|V^* - V^{\pi_k}\|_\infty \le (1-\beta(1-\gamma))^{k-1} \left(\gamma \|V^* - V^{\pi_0}\|_\infty + \tau\nu_{\max}\right) + \frac{2\epsilon}{\beta(1-\gamma)^2},\ \forall\, k \ge 1.$$

Theorem C.2 shows that Algorithm 5 converges at a geometric rate up to a term that is proportional to the function approximation error ϵ.

C.3 Finite-Time Analysis of NPG with Linear Function Approximation

Combining Proposition C.1 with Theorem C.2, we have the following result for NPG with linear function approximation.

Theorem C.3. *Consider $\{\pi_k\}$ generated by Algorithm 4. When choosing $\alpha_0 = \log(m)$ and $\alpha_k = \beta\alpha_0/(1-\beta)^k$ for all $k \ge 1$ and for any $\beta \in (0, 1)$, we have*

$$\|V^* - V^{\pi_k}\|_\infty \le (1-\beta(1-\gamma))^{k-1} \left(\gamma \|V^* - V^{\pi_0}\|_\infty + 1\right) + \frac{2\epsilon}{\beta(1-\gamma)^2},\ \forall\, k \ge 1.$$

As a result, by choosing $\beta = 1/2$, for any $\xi > 0$, Algorithm 4 achieves $\|V^* - V^{\pi_k}\|_\infty \leq \xi + 4\epsilon/(1-\gamma)^2$ with iteration complexity $\mathcal{O}((1-\gamma)^{-1}\log(\xi^{-1}(1-\gamma)^{-1}))$.

Similar to the tabular setting, the proof of the above theorem follows from the equivalence of DSPI and NPG under a specific choice of stepsize and regularizer. Following the stepsize calculations in Appendix B.3, Proposition C.1 implies that Algorithm 4 is a special case of Algorithm 5. The convergence rate in Theorem C.3 then follows by taking $\nu(\cdot) = h(\cdot)$ and $\tau = 1/\log(m)$. Since the proof is similar to that of Theorem 4.3, we omit the details.

The NPG method with linear function approximation, also referred to as NPG with a log-linear policy class, has been extensively analyzed from a continuous optimization viewpoint [1, 48, 2, 3, 12], establishing convergence to the global optimum up to a function approximation error. In particular, [1] establish $\mathcal{O}(1/\sqrt{k})$ convergence for the best iterate (up to iteration k) using a decaying stepsize, while [48] and [2] establish geometric convergence of the last iterate with geometrically increasing stepsize sequences. Considering a more general framework of policy mirror descent, [3] achieve geometric convergence with geometrically increasing stepsize sequence. [12] show that geometric convergence of the last iterate can be achieved with a constant stepsize, but using entropy regularization of the MDP. The convergence rates in the aforementioned works depend on a concentrability or distribution mismatch coefficient; moreover, the results of [1, 48, 2] require a bounded relative condition number associated with the feature matrix. In contrast, Theorem C.3 establishes distribution-free global geometric convergence of the last iterate up to a function approximation error, without requiring additional assumptions on the feature matrix or regularization of the MDP.

C.4 Proof of Theorem C.2

The proof follows the same roadmap as in the tabular setting. Specifically, we establish an almost monotonic improvement and an almost contractive recursion, where the "almost" arises from the function approximation error.

The proof of the following lemma is given in Appendix C.5.1.

Lemma C.4 (Almost Monotonic Improvement)**.** *Algorithm 5 satisfies*

$$Q^{\pi_k} \leq Q^{\pi_{k+1}} + \frac{2\gamma\epsilon}{1-\gamma}\mathbf{1}, \quad \forall\, k \geq 0.$$

As a corollary of the above lemma, we have the following result on the relation between the averaged iterates $\bar{W}_k$ and the Q-functions Q^{π_k}, which is proved in Appendix C.5.2.

Corollary C.5. *Algorithm 5 satisfies*

$$\bar{W}_k \leq Q^{\pi_k} + \delta_k \mathbf{1}, \quad \forall\, k \geq 1,$$

where $\delta_k := \left(\frac{1+\gamma}{1-\gamma}\right)\epsilon\sum_{i=0}^{k-1}(1-\beta)^i$.

The above two results can now be used to obtain an almost one-step contraction inequality, which is the linear function approximation analogue of Lemma 5.2. We state the result below and prove it in Appendix C.5.3.

Lemma C.6 (An Almost One-step Contractive Recursion)**.** *Algorithm 5 satisfies*

$$\|Q^* - \bar{W}_{k+1}\|_\infty \leq (1-\beta(1-\gamma))\|Q^* - \bar{W}_k\|_\infty + \beta\gamma\nu_{\max}\eta_{k-1} + \beta\gamma\delta_k + \beta\epsilon, \quad \forall\, k \geq 1.$$

The following result translates the above bound on the Q-functions into a bound on the value functions. It is analogous to Lemma B.3 and is proved in Appendix C.5.4.

Lemma C.7. *Algorithm 5 satisfies*

(1) $\|V^* - V^{\pi_{k+1}}\|_\infty \leq \|Q^* - \bar{W}_{k+1}\|_\infty + \eta_k\nu_{\max} + \delta_{k+1}$, *for every* $k \geq 0$*;*
(2) $\|Q^* - \bar{W}_1\|_\infty \leq \gamma\|V^* - V^{\pi_0}\|_\infty + \epsilon$.

The next step toward the final bound is to apply Lemma C.7 (1) and recursively use the inequality in Lemma C.6. Specifically, we obtain

$$\|V^* - V^{\pi_{k+1}}\|_\infty \leq \|Q^* - \bar{W}_{k+1}\|_\infty + \eta_k\nu_{\max} + \delta_{k+1}$$

$$
\begin{aligned}
&\leq (1-\beta(1-\gamma))\|Q^* - \bar{W}_k\|_\infty + \beta\left(\gamma\eta_{k-1}\nu_{\max} + \gamma\delta_k + \epsilon\right) + \eta_k\nu_{\max} + \delta_{k+1} \\
&\leq (1-\beta(1-\gamma))^k\|Q^* - \bar{W}_1\|_\infty \\
&\quad + \sum_{j=1}^{k}(1-\beta(1-\gamma))^{k-j}\beta\left(\gamma\eta_{j-1}\nu_{\max} + \gamma\delta_j + \epsilon\right) + \eta_k\nu_{\max} + \delta_{k+1}.
\end{aligned}
$$

Applying Lemma C.7 (2) gives

$$
\begin{aligned}
\|V^* - V^{\pi_{k+1}}\|_\infty \leq \gamma(1-\beta(1-\gamma))^k\|V^* - V^{\pi_0}\|_\infty
&+ \epsilon\underbrace{\left((1-\beta(1-\gamma))^k + \sum_{j=1}^{k}(1-\beta(1-\gamma))^{k-j}\beta\right)}_{:=O_1} \\
&+ \nu_{\max}\underbrace{\left(\sum_{j=1}^{k}(1-\beta(1-\gamma))^{k-j}\beta\gamma\eta_{j-1} + \eta_k\right)}_{:=O_2} \\
&+ \underbrace{\sum_{j=1}^{k}(1-\beta(1-\gamma))^{k-j}\beta\gamma\delta_j + \delta_{k+1}}_{:=O_3}. \qquad (16)
\end{aligned}
$$

For the term O_1, we have

$$
O_1 = (1-\beta(1-\gamma))^k + \frac{1}{1-\gamma}\left[1 - (1-\beta(1-\gamma))^k\right] \leq \frac{1}{1-\gamma}.
$$

For the term O_2, since $\eta_k = \tau(1-\beta)^k$, we have

$$
\begin{aligned}
O_2 &= \tau\beta\gamma\sum_{j=1}^{k}(1-\beta(1-\gamma))^{k-j}(1-\beta)^{j-1} + \tau(1-\beta)^k \\
&= \tau\beta\gamma(1-\beta(1-\gamma))^{k-1}\sum_{j=1}^{k}\left(\frac{1-\beta}{1-\beta(1-\gamma)}\right)^{j-1} + \tau(1-\beta)^k \\
&= \tau\beta\gamma(1-\beta(1-\gamma))^{k-1}\frac{1-\left(\frac{1-\beta}{1-\beta(1-\gamma)}\right)^k}{1-\left(\frac{1-\beta}{1-\beta(1-\gamma)}\right)} + \tau(1-\beta)^k \\
&= \tau\left((1-\beta(1-\gamma))^k - (1-\beta)^k\right) + \tau(1-\beta)^k \\
&= \tau(1-\beta(1-\gamma))^k.
\end{aligned}
$$

For the term O_3, we have

$$
\begin{aligned}
&\sum_{j=1}^{k}(1-\beta(1-\gamma))^{k-j}\beta\gamma\delta_j + \delta_{k+1} \\
=&\gamma\left(\frac{1+\gamma}{1-\gamma}\right)\epsilon\sum_{j=1}^{k}(1-\beta(1-\gamma))^{k-j}(1-(1-\beta)^j) + \frac{1}{\beta}\left(\frac{1+\gamma}{1-\gamma}\right)(1-(1-\beta)^{k+1})\epsilon \\
\leq&\frac{1+\gamma}{\beta(1-\gamma)^2}\epsilon.
\end{aligned}
$$

Substituting the bounds obtained for O_1, O_2, and O_3 into (16), we have

$$
\|V^* - V^{\pi_{k+1}}\|_\infty \leq (1-\beta(1-\gamma))^k\left(\gamma\|V^* - V^{\pi_0}\|_\infty + \tau\nu_{\max}\right) + \frac{2\epsilon}{\beta(1-\gamma)^2}.
$$

C.5 Proof of All Technical Lemmas

C.5.1 Proof of Lemma C.4

It is enough to show, for any $k \geq 0$, that $Q^{\pi_k} \leq \mathcal{H}^{\pi_{k+1}}(Q^{\pi_k}) + 2\gamma\epsilon\mathbf{1}$. Once this is established, repeatedly applying $\mathcal{H}^{\pi_{k+1}}$ on both sides gives us $Q^{\pi_k} \leq [\mathcal{H}^{\pi_{k+1}}]^n(Q^{\pi_k}) + 2\epsilon\mathbf{1}\sum_{i=1}^{n}\gamma^i$. Letting $n \to \infty$ gives us the desired result.

We next prove that $Q^{\pi_k} \leq \mathcal{H}^{\pi_{k+1}}(Q^{\pi_k}) + 2\gamma\epsilon\mathbf{1}$ for all $k \geq 0$. Similar to the proof of Lemma 5.1, we consider the two cases $k = 0$ and $k \geq 1$ separately.

(1) For $k = 0$, since $\beta_0 = 1$, we have $\bar{W}_1 = W_0^*$. It follows that

$$\begin{aligned}
Q^{\pi_0} &= \mathcal{H}^{\pi_0}(Q^{\pi_0}) \\
&= \mathcal{H}^{\pi_0}(W_0^* - \epsilon_0) \\
&\leq \mathcal{H}^{\pi_0}(W_0^* + \epsilon\mathbf{1}) \\
&= \mathcal{H}^{\pi_0}(W_0^*) + \gamma\epsilon\mathbf{1} && \text{(Translation invariance)} \\
&= \mathcal{H}^{\pi_0}(\bar{W}_1) + \gamma\epsilon\mathbf{1} \\
&= \mathcal{H}^{\pi_0}_{\eta_0}(\bar{W}_1) - \eta_0 f(\pi_0) + \gamma\epsilon\mathbf{1} \\
&\leq \mathcal{H}^{\pi_1}_{\eta_0}(\bar{W}_1) - \eta_0 f(\pi_0) + \gamma\epsilon\mathbf{1} && \text{(Algorithm 5, Line 5)} \\
&= \mathcal{H}^{\pi_1}(\bar{W}_1) + \eta_0[f(\pi_1) - f(\pi_0)] + \gamma\epsilon\mathbf{1} \\
&\leq \mathcal{H}^{\pi_1}(Q^{\pi_0} + \epsilon_0) + \gamma\epsilon\mathbf{1} && \text{(Algorithm 5, Line 1)} \\
&\leq \mathcal{H}^{\pi_1}(Q^{\pi_0} + \epsilon\mathbf{1}) + \gamma\epsilon\mathbf{1} \\
&\leq \mathcal{H}^{\pi_1}(Q^{\pi_0}) + 2\gamma\epsilon\mathbf{1}. && \text{(Translation invariance)}
\end{aligned}$$

(2) For $k \geq 1$, observe that

$$\begin{aligned}
Q^{\pi_k} &= \mathcal{H}^{\pi_k}(Q^{\pi_k}) \\
&= \mathcal{H}^{\pi_k}(W_k^* - \epsilon_k) \\
&\leq \mathcal{H}^{\pi_k}(W_k^* + \epsilon\mathbf{1}) \\
&\leq \mathcal{H}^{\pi_k}(W_k^*) + \gamma\epsilon\mathbf{1} && \text{(Translation invariance)} \\
&= \mathcal{H}^{\pi_k}\left(\frac{\bar{W}_{k+1} - (1-\beta_k)\bar{W}_k}{\beta_k}\right) + \gamma\epsilon\mathbf{1} && \text{(Algorithm 5, Line 4)} \\
&= \mathcal{H}^{\pi_k}_{\eta_{k-1}}\left(\frac{\bar{W}_{k+1} - (1-\beta_k)\bar{W}_k}{\beta_k}\right) - \eta_{k-1} f(\pi_k) + \gamma\epsilon\mathbf{1} \\
&= \frac{1}{\beta_k}\mathcal{H}^{\pi_k}_{\eta_{k-1}}(\bar{W}_{k+1}) - \frac{1-\beta_k}{\beta_k}\mathcal{H}^{\pi_k}_{\eta_{k-1}}(\bar{W}_k) - \eta_{k-1} f(\pi_k) + \gamma\epsilon\mathbf{1} && (\mathcal{H}^{\pi_k}_{\eta_{k-1}} \text{ being affine}) \\
&= \frac{1}{\beta_k}\mathcal{H}^{\pi_k}_{\eta_k}(\bar{W}_{k+1}) - \frac{1-\beta_k}{\beta_k}\mathcal{H}^{\pi_k}_{\eta_{k-1}}(\bar{W}_k) - \frac{\eta_k - (1-\beta_k)\eta_{k-1}}{\beta_k} f(\pi_k) + \gamma\epsilon\mathbf{1} \\
&= \frac{1}{\beta_k}\mathcal{H}^{\pi_k}_{\eta_k}(\bar{W}_{k+1}) - \frac{1-\beta_k}{\beta_k}\mathcal{H}^{\pi_k}_{\eta_{k-1}}(\bar{W}_k) + \gamma\epsilon\mathbf{1},
\end{aligned}$$

where the last step follows from $\eta_k = (1-\beta_k)\eta_{k-1}$.
Since Line 5 of Algorithm 5 implies $\mathcal{H}^{\pi_{k+1}}_{\eta_k}(\bar{W}_{k+1}) = \mathcal{H}_{\eta_k}(\bar{W}_{k+1})$ for all $k \geq 0$, we have

$$\begin{aligned}
\mathcal{H}^{\pi_k}_{\eta_k}(\bar{W}_{k+1}) \leq \mathcal{H}_{\eta_k}(\bar{W}_{k+1}) = \mathcal{H}^{\pi_{k+1}}_{\eta_k}(\bar{W}_{k+1}) \\
\mathcal{H}^{\pi_k}_{\eta_{k-1}}(\bar{W}_k) = \mathcal{H}_{\eta_{k-1}}(\bar{W}_k) \geq \mathcal{H}^{\pi_{k+1}}_{\eta_{k-1}}(\bar{W}_k).
\end{aligned}$$

Combining the previous two inequalities, we obtain

$$\begin{aligned}
Q^{\pi_k} &\leq \frac{1}{\beta_k}\mathcal{H}^{\pi_{k+1}}_{\eta_k}(\bar{W}_{k+1}) - \frac{1-\beta_k}{\beta_k}\mathcal{H}^{\pi_{k+1}}_{\eta_{k-1}}(\bar{W}_k) + \gamma\epsilon\mathbf{1} \\
&= \frac{1}{\beta_k}\mathcal{H}^{\pi_{k+1}}(\bar{W}_{k+1}) - \frac{1-\beta_k}{\beta_k}\mathcal{H}^{\pi_{k+1}}(\bar{W}_k) + \frac{\eta_k - (1-\beta_k)\eta_{k-1}}{\beta_k} f(\pi_{k+1}) + \gamma\epsilon\mathbf{1} \\
&= \frac{1}{\beta_k}\mathcal{H}^{\pi_{k+1}}(\bar{W}_{k+1}) - \frac{1-\beta_k}{\beta_k}\mathcal{H}^{\pi_{k+1}}(\bar{W}_k) + \gamma\epsilon\mathbf{1} && (\eta_k = (1-\beta_k)\eta_{k-1})
\end{aligned}$$

$$
\begin{aligned}
&= \mathcal{H}^{\pi_{k+1}}\left(\frac{\bar{W}_{k+1} - (1-\beta_k)\bar{W}_k}{\beta_k}\right) + \gamma\epsilon\mathbf{1} && (\mathcal{H}^{\pi_{k+1}} \text{ being affine})\\
&= \mathcal{H}^{\pi_{k+1}}(W_k^*) + \gamma\epsilon\mathbf{1} && (\text{Algorithm 5, Line 4})\\
&= \mathcal{H}^{\pi_{k+1}}(Q^{\pi_k} + \epsilon_k) + \gamma\epsilon\mathbf{1}\\
&\leq \mathcal{H}^{\pi_{k+1}}(Q^{\pi_k} + \epsilon\mathbf{1}) + \gamma\epsilon\mathbf{1}\\
&\leq \mathcal{H}^{\pi_{k+1}}(Q^{\pi_k}) + 2\gamma\epsilon\mathbf{1} && (\text{Translation invariance})
\end{aligned}
$$

C.5.2 Proof of Corollary C.5

The proof is by induction. Since $\bar{W}_0 = 0$ and $\beta_0 = 1$, we have by Lemma C.4 that

$$
\bar{W}_1 = W_0^* = Q^{\pi_0} + \epsilon_0 \leq Q^{\pi_0} + \epsilon\mathbf{1} \leq Q^{\pi_1} + \frac{2\gamma\epsilon}{1-\gamma}\mathbf{1} + \epsilon\mathbf{1} \leq Q^{\pi_1} + \frac{\epsilon(1+\gamma)}{1-\gamma}\mathbf{1}.
$$

Thus, the base case holds.

Now suppose that, for some $n \geq 1$,

$$
\bar{W}_n \leq Q^{\pi_n} + \left(\frac{1+\gamma}{1-\gamma}\right)\epsilon\sum_{i=0}^{n-1}(1-\beta)^i\mathbf{1}.
$$

Then,

$$
\begin{aligned}
\bar{W}_{n+1} &= (1-\beta)\bar{W}_n + \beta W_n^*\\
&\leq (1-\beta)Q^{\pi_n} + (1-\beta)\left(\frac{1+\gamma}{1-\gamma}\right)\epsilon\sum_{i=0}^{n-1}(1-\beta)^i\mathbf{1} + \beta(Q^{\pi_n} + \epsilon_n)\\
&\qquad (\text{The induction hypothesis})\\
&\leq Q^{\pi_n} + \left(\frac{1+\gamma}{1-\gamma}\right)\epsilon\sum_{i=1}^{n}(1-\beta)^i\mathbf{1} + \beta\epsilon\mathbf{1}\\
&\leq Q^{\pi_{n+1}} + \frac{2\gamma\epsilon}{1-\gamma}\mathbf{1} + \left(\frac{1+\gamma}{1-\gamma}\right)\epsilon\sum_{i=1}^{n}(1-\beta)^i\mathbf{1} + \beta\epsilon\mathbf{1} && (\text{Lemma C.4})\\
&\leq Q^{\pi_{n+1}} + \left(\frac{1+\gamma}{1-\gamma}\right)\epsilon\sum_{i=0}^{n}(1-\beta)^i\mathbf{1}.
\end{aligned}
$$

This completes the induction.

C.5.3 Proof of Lemma C.6

By Line 4 of Algorithm 5, we have

$$
\begin{aligned}
\|Q^* - \bar{W}_{k+1}\|_\infty &= \|(1-\beta)(Q^* - \bar{W}_k) + \beta(Q^* - W_k^*)\|_\infty\\
&\leq (1-\beta)\|Q^* - \bar{W}_k\|_\infty + \beta\|Q^* - W_k^*\|_\infty\\
&= (1-\beta)\|Q^* - \bar{W}_k\|_\infty + \beta\|Q^* - Q^{\pi_k} - \epsilon_k\|_\infty\\
&\leq (1-\beta)\|Q^* - \bar{W}_k\|_\infty + \beta\|Q^* - Q^{\pi_k}\|_\infty + \beta\epsilon. \qquad (17)
\end{aligned}
$$

We next bound $\|Q^* - Q^{\pi_k}\|_\infty$ in terms of $\|Q^* - \bar{W}_k\|_\infty$. Observe that

$$
\begin{aligned}
Q^* - Q^{\pi_k} &= \mathcal{H}(Q^*) - \mathcal{H}^{\pi_k}(Q^{\pi_k})\\
&\leq \mathcal{H}(Q^*) - \mathcal{H}^{\pi_k}(\bar{W}_k - \delta_k\mathbf{1}) && (\text{Corollary C.5})\\
&= \mathcal{H}(Q^*) - \mathcal{H}^{\pi_k}(\bar{W}_k) + \gamma\delta_k\mathbf{1} && (\text{Translation invariance})\\
&= \mathcal{H}(Q^*) - \mathcal{H}^{\pi_k}_{\eta_{k-1}}(\bar{W}_k) + \eta_{k-1}f(\pi_k) + \gamma\delta_k\mathbf{1}\\
&= \mathcal{H}(Q^*) - \mathcal{H}_{\eta_{k-1}}(\bar{W}_k) + \eta_{k-1}f(\pi_k) + \gamma\delta_k\mathbf{1} && (\text{Algorithm 5, Line 5})\\
&\leq \mathcal{H}(Q^*) - \mathcal{H}(\bar{W}_k) + \eta_{k-1}f(\pi_k) + \gamma\delta_k\mathbf{1} && (\nu(\cdot) \geq 0)
\end{aligned}
$$

$$\begin{aligned}
&\leq \|\mathcal{H}(Q^*) - \mathcal{H}(\bar{W}_k)\|_\infty \mathbf{1} + \eta_{k-1}\gamma\nu_{\max}\mathbf{1} + \gamma\delta_k\mathbf{1} \\
&\leq \gamma\|Q^* - \bar{W}_k\|_\infty \mathbf{1} + \eta_{k-1}\gamma\nu_{\max}\mathbf{1} + \gamma\delta_k\mathbf{1}.
\end{aligned}$$

Since $Q^* - Q^{\pi_k} \geq 0$, we obtain

$$\|Q^* - Q^{\pi_k}\|_\infty \leq \gamma\|Q^* - \bar{W}_k\|_\infty + \gamma\nu_{\max}\eta_{k-1} + \gamma\delta_k.$$

Substituting the above bound into (17) yields

$$\begin{aligned}
\|Q^* - \bar{W}_{k+1}\|_\infty &\leq (1-\beta)\|Q^* - \bar{W}_k\|_\infty + \beta\gamma\|Q^* - \bar{W}_k\|_\infty + \beta\gamma\nu_{\max}\eta_{k-1} + \beta\gamma\delta_k + \beta\epsilon \\
&= (1-\beta(1-\gamma))\|Q^* - \bar{W}_k\|_\infty + \beta\gamma\nu_{\max}\eta_{k-1} + \beta\gamma\delta_k + \beta\epsilon.
\end{aligned}$$

C.5.4 Proof of Lemma C.7

(1) For any $k \geq 0$ and $s \in \mathcal{S}$,

$$\begin{aligned}
V^*(s) - V^{\pi_{k+1}}(s) &= \max_{a\in\mathcal{A}} Q^*(s,a) - \pi_{k+1}(s)^\top Q^{\pi_{k+1}}(s) \\
&\leq \max_{a\in\mathcal{A}} Q^*(s,a) - \pi_{k+1}(s)^\top (\bar{W}_{k+1}(s) - \delta_{k+1}\mathbf{1}) && \text{(Corollary C.5)} \\
&= \max_{a\in\mathcal{A}} Q^*(s,a) - \pi_{k+1}(s)^\top \bar{W}_{k+1}(s) + \delta_{k+1} \\
&\leq \max_{a\in\mathcal{A}} Q^*(s,a) - \left\{\pi_{k+1}(s)^\top \bar{W}_{k+1}(s) + \eta_k\nu(\pi_{k+1}(s))\right\} + \eta_k\nu_{\max} + \delta_{k+1} \\
&= \max_{\mu_{\Delta(\mathcal{A})}} \mu^\top Q^*(s) - \max_{\mu\in\Delta(\mathcal{A})} \left\{\mu^\top \bar{W}_{k+1}(s) + \eta_k\nu(\mu)\right\} + \eta_k\nu_{\max} + \delta_{k+1} \\
& && \text{(Algorithm 5, Line 5)} \\
&\leq \max_{\mu_{\Delta(\mathcal{A})}} \mu^\top Q^*(s) - \max_{\mu\in\Delta(\mathcal{A})} \mu^\top \bar{W}_{k+1}(s) + \eta_k\nu(\pi_{k+1}(s)) + \delta_{k+1} \\
& && (\nu(\cdot) \geq 0) \\
&\leq \max_{\mu_{\Delta(\mathcal{A})}} \left\{\mu^\top (Q^*(s) - \bar{W}_{k+1}(s))\right\} + \eta_k\nu_{\max} + \delta_{k+1} \\
&\leq \|Q^* - \bar{W}_{k+1}\|_\infty + \nu_{\max}\eta_k + \delta_{k+1}
\end{aligned}$$

Since $V^*(s) - V^{\pi_{k+1}}(s) \geq 0$ and the above inequality holds for every $s \in \mathcal{S}$, we have

$$\|V^* - V^{\pi_{k+1}}\|_\infty \leq \|Q^* - \bar{W}_{k+1}\|_\infty + \eta_k\nu_{\max} + \delta_{k+1}.$$

(2) Since $\beta_0 = 1$ and $\bar{W}_1 = W_0^*$, we have

$$\|Q^* - \bar{W}_1\|_\infty = \|Q^* - W_0^*\|_\infty = \|Q^* - Q^{\pi_0} - \epsilon_0\|_\infty \leq \|Q^* - Q^{\pi_0}\|_\infty + \epsilon.$$

Using the Bellman equation for the Q-function, we have

$$\begin{aligned}
\|Q^* - Q^{\pi_0}\|_\infty &= \|\mathcal{H}(Q^*) - \mathcal{H}^{\pi_0}(Q^{\pi_0})\|_\infty \\
&= \gamma \max_{(s,a)} \left| \sum_{s'\in\mathcal{S}} p(s'|s,a) \left(\max_{a'\in\mathcal{A}} Q^*(s',a') - \sum_{a'\in\mathcal{A}} \pi_0(a'|s') Q^{\pi_0}(s',a') \right) \right| \\
&= \gamma \max_{(s,a)} \left| \sum_{s'\in\mathcal{S}} p(s'|s,a) \left(V^*(s') - V^{\pi_0}(s') \right) \right| \\
&\leq \gamma \|V^* - V^{\pi_0}\|_\infty.
\end{aligned}$$

Combining the previous two inequalities yields

$$\|Q^* - \bar{W}_1\|_\infty \leq \gamma\|V^* - V^{\pi_0}\|_\infty + \epsilon.$$

D Natural Policy Gradient for Stochastic Shortest Path Problems

Consider an infinite-horizon undiscounted MDP $\mathcal{M} = (\mathcal{S}_0, \mathcal{A}, p, \mathcal{R})$, where $\mathcal{S}_0$ is a finite state space that includes a terminal state, $\mathcal{A}$ is a finite action space, $\mathcal{R} : \mathcal{S}_0 \times \mathcal{A} \to [-1, 0]$ is the reward function, and p is the transition kernel. We denote the terminal state by s_{tm} and define $\mathcal{S} = \mathcal{S}_0 \setminus \{s_{\text{tm}}\}$. Once

the system reaches s_{tm}, it remains there and receives no further rewards, i.e., $p(s_{\text{tm}} \mid s_{\text{tm}}, a) = 1$ and $\mathcal{R}(s_{\text{tm}}, a) = 0$ for all $a \in \mathcal{A}$. As in the discounted setting, we denote $n = |\mathcal{S}|$ and $m = |\mathcal{A}|$.

Given a stationary policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$, the expected total reward starting from a state $s \in \mathcal{S}$ is defined as

$$V^{\pi}(s) = \limsup_{K \to \infty} \mathbb{E}_{\pi} \left[\sum_{k=0}^{K} \mathcal{R}(S_k, A_k) \;\middle|\; S_0 = s \right], \tag{18}$$

where $A_k \sim \pi(\cdot \mid S_k)$ and $S_{k+1} \sim p(\cdot \mid S_k, A_k)$. A stationary policy π^* is said to be optimal if $V^*(s) \coloneqq V^{\pi^*}(s) \geq V^{\pi}(s)$ for all stationary policies π.

To ensure termination under the optimal policy, the notion of a proper policy is commonly adopted in the SSP literature [46, 47, 6, 14]. The formal definition of a proper policy is stated below.

Definition D.1. A policy π is said to be proper if, for any initial state $s \in \mathcal{S}$, there is a positive probability of reaching the terminal state within at most n stages, i.e., $\mathbb{P}\{S_n \neq s_{\text{tm}} \mid S_0 = s\} < 1$ for all $s \in \mathcal{S}$.

The notion of a proper policy was standardized in the SSP literature following [6], where it was used to relax the prevailing assumption that the reward function be either nonnegative or nonpositive. In particular, [6] required that there exists at least one proper policy, and for every improper policy π, there exists a state $s \in \mathcal{S}$ such that $V^{\pi}(s) = -\infty$, and established asymptotic convergence of VI and PI. Under the same assumption, subsequent works have also established asymptotic convergence of Q-learning [41, 46] and of variants of PI and Q-learning [47]. However, this assumption may not be sufficient for establishing convergence rates, which typically relies on a contraction property of the Bellman optimality operator[7, 47]. In SSP setting, the Bellman optimality operator is only non-expansive with respect to the ℓ_∞ norm. Nonetheless, it becomes a contraction with respect to a weighted ℓ_∞ norm (cf. Lemma D.9) when all stationary policies are proper–a standard condition in the literature [6, 7, 47] when establishing contraction properties of the Bellman operator. We therefore adopt the same assumption, formally stated below.

Assumption D.2. All stationary policies are proper.

For a proper policy, since the MDP terminates in finite time, the limit in (18) exists. Under Assumption D.2, it has been established in [6] that there exists an optimal proper policy π^* for which $V^*(s)$ is finite for every $s \in \mathcal{S}$. Furthermore, V^* satisfies the Bellman equation

$$V^*(s) = \max_{a \in \mathcal{A}} \left\{ \mathcal{R}(s, a) + \sum_{s' \in \mathcal{S}_0} p(s' \mid s, a)\, V^*(s') \right\}.$$

For a proper policy π, the state-action value function Q^{π} is analogously defined as

$$Q^{\pi}(s, a) = \lim_{K \to \infty} \mathbb{E} \left[\sum_{k=0}^{K} \mathcal{R}(S_k, A_k) \;\middle|\; S_0 = s, A_0 = a \right].$$

Under Assumption D.2, there exists a unique optimal state-action value function Q^*, which is related to the optimal value function via $V^*(s) = \max_{a \in \mathcal{A}} Q^*(s, a)$ and uniquely solves the Bellman equation $Q = \widetilde{\mathcal{H}}(Q)$ [6], where $\widetilde{\mathcal{H}} : \mathbb{R}^{mn} \to \mathbb{R}^{mn}$ is the Bellman optimality operator defined as

$$[\widetilde{\mathcal{H}}(Q)](s, a) = \mathcal{R}(s, a) + \sum_{s' \in \mathcal{S}} p(s' \mid s, a) \max_{a' \in \mathcal{A}} Q(s', a'), \quad \forall\, (s, a) \in \mathcal{S} \times \mathcal{A}.$$

Similarly, for a proper stationary policy π, Q^{π} uniquely solves the Bellman equation $Q = \widetilde{\mathcal{H}}^{\pi}(Q)$, where $\widetilde{\mathcal{H}}^{\pi} : \mathbb{R}^{mn} \to \mathbb{R}^{mn}$ is the Bellman operator associated with π defined as

$$[\widetilde{\mathcal{H}}^{\pi}(Q)](s, a) = \mathcal{R}(s, a) + \sum_{s' \in \mathcal{S}} p(s' \mid s, a) \sum_{a' \in \mathcal{A}} \pi(a' \mid s') Q(s', a'), \quad \forall\, (s, a) \in \mathcal{S} \times \mathcal{A}.$$

Observe that the sums in the above expressions are over $\mathcal{S} = \mathcal{S}_0 \setminus \{s_{\text{tm}}\}$. Further, due to absence of a discount factor, the Bellman optimality operator $\widetilde{\mathcal{H}}$ and the Bellman operator $\widetilde{\mathcal{H}}^{\pi}$ for arbitrary stationary policy π are only non-expansive with respect to the ℓ_∞-norm. Nonetheless, both $\widetilde{\mathcal{H}}$ and $\widetilde{\mathcal{H}}^{\pi}$ are monotonic: for $Q_1 \leq Q_2$, we have $\widetilde{\mathcal{H}}(Q_1) \leq \widetilde{\mathcal{H}}(Q_2)$ and $\widetilde{\mathcal{H}}^{\pi}(Q_1) \leq \widetilde{\mathcal{H}}^{\pi}(Q_2)$.

D.1 Natural Policy Gradient for Stochastic Shortest Path Problems

In comparison to the discounted setting, policy gradient methods in the stochastic shortest path (SSP) setting are relatively underexplored. Policy optimization algorithms for SSP problems have been analyzed by [14] in terms of regret bounds. Concerning last-iterate convergence, policy gradient methods for undiscounted total reward MDPs were studied by [28] only very recently. The NPG update for the SSP problem admits a similar formulation as that for the discounted problem [28]; we present it below in Algorithm 6.

Algorithm 6 Natural Policy Gradient for Stochastic Shortest Path Problems

1: **Input:** Initialize π_0 as the uniform policy and $\theta_0 = \mathbf{0} \in \mathbb{R}^{mn}$.
2: **for** $k = 0, 1, 2, 3, \cdots$ **do**
3: $\quad \theta_{k+1} = \theta_k + \alpha_k Q^{\pi_k}$
4: $\quad \pi_{k+1}(s) = \arg\max_{\mu \in \Delta(\mathcal{A})} \left\{\mu^\top \theta_{k+1}(s) + h(\mu)\right\}$ for any $s \in \mathcal{S}$.
5: **end for**

D.2 Reformulation as Doubly Smoothed Policy Iteration

The development of the DSPI algorithm for SSP problems follows the same roadmap as that for the discounted problems. To that end, we first introduce the smoothed versions of the Bellman operators for the SSP setting. Given a non-negative, bounded, and concave function $\nu : \Delta(\mathcal{A}) \to \mathbb{R}$ and $\eta \geq 0$, the smoothed Bellman optimality operator $\widetilde{\mathcal{H}}_\eta : \mathbb{R}^{mn} \to \mathbb{R}^{mn}$ is defined as

$$[\widetilde{\mathcal{H}}_\eta(Q)](s,a) = \mathcal{R}(s,a) + \sum_{s' \in \mathcal{S}} p(s' \mid s,a) \max_{\mu \in \Delta(\mathcal{A})} \left\{\mu^\top Q(s') + \eta \nu(\mu)\right\}, \quad \forall\, (s,a) \in \mathcal{S} \times \mathcal{A}.$$

Similarly, given policy π, the smoothed Bellman operator $\widetilde{\mathcal{H}}_\eta^\pi : \mathbb{R}^{mn} \to \mathbb{R}^{mn}$ associated with π is defined as

$$[\widetilde{\mathcal{H}}_\eta^\pi(Q)](s,a) = \mathcal{R}(s,a) + \sum_{s' \in \mathcal{S}} p(s' \mid s,a) \left\{\pi(s')^\top Q(s') + \eta \nu(\pi(s'))\right\}, \quad \forall\, (s,a) \in \mathcal{S} \times \mathcal{A}.$$

Similar to the discounted setting, the smoothed Bellman operator $\widetilde{\mathcal{H}}_\eta^\pi$ is related to $\widetilde{\mathcal{H}}^\pi$ as $\widetilde{\mathcal{H}}_\eta^\pi(Q) = \widetilde{\mathcal{H}}^\pi(Q) + \eta g(\pi)$, where the operator $g : \mathbb{R}^{mn} \to \mathbb{R}^{mn}$ is defined as

$$[g(\pi)](s,a) = \sum_{s' \in \mathcal{S}} p(s' \mid s,a)\, \nu(\pi(s'))$$

for all $(s,a) \in \mathcal{S} \times \mathcal{A}$. Moreover, the operators $\widetilde{\mathcal{H}}_\eta(\cdot)$ and $\widetilde{\mathcal{H}}_\eta^\pi(\cdot)$ for arbitrary stationary policy π and for $\eta \geq 0$ are non-expansive with respect to the ℓ_∞ norm and are monotonic. These properties can be established using arguments similar to those in Appendix B.2 and hence are omitted. With the smoothed versions of the Bellman operators defined, we can now formulate the DSPI algorithm for the SSP problem, which differs from that in the discounted setting only with respect to the Bellman operators. We state the algorithm below.

Algorithm 7 Doubly Smoothed Policy Iteration for Stochastic Shortest Path Problems

1: **Input:** Initialize $\pi_0(s) \in \arg\max_{\mu \in \Delta(\mathcal{A})} \nu(\mu)$ for all $s \in \mathcal{S}$ and $\bar{Q}_0 = \mathbf{0}$.
2: **for** $k = 0, 1, 2, 3, \cdots$ **do**
3: $\quad$ Maintain a running average of past Q-functions: $\bar{Q}_{k+1} = (1 - \beta_k)\bar{Q}_k + \beta_k Q^{\pi_k}$, where $\beta_k \in (0, 1]$ is the stepsize.
4: $\quad$ Choose π_{k+1} such that $\widetilde{\mathcal{H}}_{\eta_k}^{\pi_{k+1}}(\bar{Q}_{k+1}) = \widetilde{\mathcal{H}}_{\eta_k}(\bar{Q}_{k+1})$, where $\eta_k = \tau \prod_{j=1}^{k}(1 - \beta_j)$ with $\tau \geq 0$ being a tunable parameter.
5: **end for**

Observe that the policy update is now expressed in terms of the SSP analogue of the smoothed Bellman operators. A concrete way to implement the update step remains the same as that in the

discounted setting:

$$\pi_{k+1}(s) \in \arg\max_{\mu \in \Delta(\mathcal{A})} \Big\{ \mu^\top \bar{Q}_{k+1}(s) + \tau \prod_{j=1}^{k} (1-\beta_j) \nu(\mu) \Big\}, \ \forall\, s \in \mathcal{S}. \tag{19}$$

However, as discussed in Section 3, the Bellman operator viewpoint facilitates a simplified analysis and leads to improved convergence rates, leveraging the properties of the Bellman operators.

The DSPI algorithm includes the NPG algorithm as a special case, which is given by the following result.

Proposition D.3. *In Algorithm 7, when choosing $\beta_k = \alpha_k / \sum_{i=0}^{k} \alpha_i$, $\tau = 1/\alpha_0$, and $\nu(\cdot) = h(\cdot)$, the sequence of policies generated by Algorithm 7 and Algorithm 6 are identical.*

Since the NPG and DSPI updates for the SSP problems remain the same as those for the discounted setting, the proof of the above result remains the same as that of Proposition 3.1, and hence we omit it here.

D.3 Finite-Time Analysis of DSPI for Stochastic Shortest Path Problems

Following is the main result showing convergence of Algorithm 7. For simplicity, we state it only for using a constant stepsize.

Theorem D.4. *Consider the sequence $\{\pi_k\}$ generated by Algorithm 7. Let the stepsizes be chosen as $\beta_0 = 1$ and $\beta_k \equiv \beta \in (0, 1]$ for all $k \geq 1$. Then, under Assumption D.2, there exists a $\kappa \in (0, 1)$ such that*

$$\|V^* - V^{\pi_k}\|_\infty \leq (1 - (1-\kappa)\beta)^{k-1} \left[\frac{\|V^* - V^{\pi_0}\|_\infty + 2\tau\nu_{\max}}{1-\kappa} \right], \quad \forall\, k \geq 1.$$

The above result shows that Algorithm 7 achieves global geometric convergence. The convergence rate is similar to that in discounted tabular setting, except for the factor $1/(1-\kappa)$ which arises due to the Bellman optimality operator not being a contraction mapping with respect to the ℓ_∞ norm. Moreover, it leads to the following iteration complexity.

Corollary D.5. *Let $\beta_0 = 1$, $\beta_k = 1/2$ for $k \geq 1$, and $\tau = 1/\nu_{\max}$ in Algorithm 7. Then for any $\epsilon > 0$, to achieve $\|V^* - V^{\pi_k}\|_\infty \leq \epsilon$, the iteration complexity is $\mathcal{O}((1-\kappa)^{-1} \log(\epsilon^{-1}(1-\kappa)^{-1})$.*

D.4 Finite-Time Analysis of NPG for Stochastic Shortest Path Problems

As a special case of DSPI, NPG also enjoys global geometric convergence, as stated in the following theorem.

Theorem D.6. *Consider $\{\pi_k\}$ generated by Algorithm 6. When choosing $\alpha_0 = \log(m)$ and $\alpha_k = \beta\alpha_0/(1-\beta)^k$ for all $k \geq 1$ and for any $\beta \in (0, 1)$, we have*

$$\|V^* - V^{\pi_k}\|_\infty \leq (1 - (1-\kappa)\beta)^{k-1} \left[\frac{\|V^* - V^{\pi_0}\|_\infty + 2}{1-\kappa} \right], \quad \forall\, k \geq 1.$$

The proof of Theorem D.6 follows from combining Proposition D.3 and Theorem D.4. To apply Proposition D.3, we first verify that $\{\alpha_k\}$ as chosen in Theorem D.6 corresponds to $\{\beta_k\}$ as chosen in Theorem D.4. We then substitute the specific choices $\nu(\cdot) = h(\cdot)$ and $\tau = 1/\log(m)$. These steps follow the arguments in Appendix B.3 and hence are omitted. We now state the iteration complexity of NPG.

Corollary D.7. *Let $\alpha_0 = \log(m)$ and $\alpha_k = \log(m) \cdot 2^{k-1}$ for all $k \geq 1$ in Algorithm 6. Then for any $\epsilon > 0$, to achieve $\|V^* - V^{\pi_k}\|_\infty \leq \epsilon$ the iteration complexity is $\mathcal{O}((1-\kappa)^{-1} \log(\epsilon^{-1}(1-\kappa)^{-1})$.*

As discussed, the analysis of NPG for SSP problems has received relatively limited attention compared to the discounted setting. To the best of our knowledge, the only existing analysis of NPG in the SSP setting is due to [28], where it is studied as part of a broader framework of policy gradient methods for undiscounted total-reward MDPs. For NPG, [28] establish an $\mathcal{O}(1/k)$ convergence rate with a constant stepsize and a geometric convergence rate with adaptive stepsizes. Their analysis

is based on the mirror descent framework of [44], resulting in convergence rates that depend on a distribution mismatch coefficient; moreover, the adaptive stepsizes depend on the optimal policy. In our setting, Theorem D.6 establishes distribution-free geometric convergence using a simple geometrically increasing stepsize. As mentioned earlier, policy gradient methods for SSP problems have also been studied by [14], but in an episodic setting and from a regret perspective; therefore, their results are not directly comparable to ours.

D.5 Proof of Theorem D.4

The proof follows the same line of argument as before, establishing monotonic policy improvement and a contractive recursion. The following lemma formalizes the monotonicity of the DSPI policy updates and is proved in Appendix D.6.1.

Lemma D.8. *Under Assumption D.2, the policies π_k generated by Algorithm 7 satisfy $Q^{\pi_{k+1}} \geq Q^{\pi_k}$ for all $k \geq 0$.*

Remark. Although we are assuming that all stationary policies are proper, monotonic policy improvement can be established under the assumption that (1) π_0 is proper, and (2) for every improper policy π, there exists $(s,a) \in \mathcal{S} \times \mathcal{A}$ such that $Q^{\pi}(s,a) = -\infty$. We present the proof in Appendix D.7.4. Such an assumption is similar to the assumptions formulated in [6], where the authors require (1) the existence of at least one proper policy, and (2) that every improper policy must lead to $Q^{\pi}(s,a) = -\infty$ for some (s,a). However, as discussed, existing analyses require all policies being proper in order to establish convergence rate. The condition that π_0 be proper is in view of the initialization in Algorithm 7.

Following our approach thus far, the next step is to establish a contractive recursion, which relies on the contraction property of the Bellman optimality operator. However, unlike the discounted setting, the Bellman optimality operator $\widetilde{\mathcal{H}}(\cdot)$ in the SSP setting is not a ℓ_∞ norm contraction in general; rather, it is a contraction with respect to a weighted ℓ_∞ norm [7]. We formally state this property in the following lemma.

Lemma D.9 (Proposition 2.2 of [7])**.** *Under Assumption D.2, there exists $\xi \in \mathbb{R}^{mn}$ such that $\xi > \mathbf{1}$ and $\widetilde{\mathcal{H}}$ is a κ-contraction with respect to $\|\cdot\|_\xi$, where*

$$\kappa = \max_{1 \leq i \leq mn} \frac{\xi(i) - 1}{\xi(i)} \in (0,1) \quad \textit{and} \quad \|Q\|_\xi \coloneqq \max_{1 \leq i \leq mn} \frac{|Q(i)|}{\xi(i)}, \ \ \forall Q \in \mathbb{R}^{mn}$$

Remark. As detailed in the proof of Proposition 2.2 in [7], the weight vector $\xi \in \mathbb{R}^{mn}$ is defined as the optimal value function of an auxiliary SSP problem that has the same transition kernel as the original problem but assigns a one-stage reward of 1 at every nonterminal state and 0 at the terminal state. Thus, $\xi(i)$ represents the largest expected time to termination, over all policies, starting from state i. In other words, states with potentially longer transient behavior receive larger weights. The contraction factor $\kappa = \max_{1 \leq i \leq mn} (1 - 1/\xi(i))$ thus captures the worst-case fraction of the remaining transient behavior that can persist after one transition. Therefore, faster termination yields a smaller κ and stronger contraction, whereas longer expected termination times lead to weaker contraction.

The following result shows the equivalence between the weighted ℓ_∞ norm $\|\cdot\|_\xi$ and the unweighted ℓ_∞ norm $\|\cdot\|_\infty$, which we prove in Appendix D.7.

Corollary D.10. *It holds for any $Q \in \mathbb{R}^{mn}$ that $(1-\kappa)\|Q\|_\infty \leq \|Q\|_\xi \leq \|Q\|_\infty$.*

For the remaining part of the proof, we denote $\kappa' = \min\left\{1, \frac{\kappa}{1-\kappa}\right\}$. Now, we present below the contractive recursion with respect to the weighted ℓ_∞ norm $\|\cdot\|_\xi$.

Lemma D.11. *Under Assumption D.2, Algorithm 7 satisfies*

$$\|Q^* - \bar{Q}_{k+1}\|_\xi \leq (1 - (1-\kappa)\beta_k)\|Q^* - \bar{Q}_k\|_\xi + \kappa'\beta_k\eta_{k-1}\nu_{\max}, \quad \forall\, k \geq 1.$$

Recursively applying the inequality in Lemma D.11 gives

$$\|Q^* - \bar{Q}_{k+1}\|_\xi \leq \prod_{j=1}^{k}(1 - (1-\kappa)\beta_j)\|Q^* - \bar{Q}_1\|_\xi$$

$$+ \kappa' \nu_{\max} \sum_{i=1}^{k} \beta_i \eta_{i-1} \prod_{j=i+1}^{k} (1 - (1-\kappa)\beta_j) \tag{20}$$

To translate the above bound to one in terms of the value function, we apply the following lemma which is proved in Appendix D.7.3.

Lemma D.12. *Let $\tilde{\xi}(s) = \max_{a \in \mathcal{A}} \xi(s,a)$. Then, under Assumption D.2, Algorithm 7 satisfies*

(1) $\|V^* - V^{\pi_k}\|_{\tilde{\xi}} \leq \|Q^* - \bar{Q}_k\|_\infty + \eta_{k-1}\nu_{\max}$ *for all* $k \geq 1$

(2) $\|Q^* - \bar{Q}_1\|_\infty \leq \|V^* - V^{\pi_0}\|_\infty$

For $k \geq 1$, combining Lemma D.12 with the bound in (20) yields

$$\|V^* - V^{\pi_k}\|_{\tilde{\xi}} \leq \underbrace{\prod_{j=1}^{k-1}(1-(1-\kappa)\beta_j)\|V^* - V^{\pi_0}\|_\infty}_{=:T_1} + \underbrace{\kappa' \nu_{\max} \sum_{i=1}^{k-1} \beta_i \eta_{i-1} \prod_{j=i+1}^{k-1} (1-(1-\kappa)\beta_j)}_{=:T_2} + \underbrace{\eta_{k-1}\nu_{\max}}_{=:T_3} \tag{21}$$

Recall that $\eta_{k-1} \coloneqq \tau \prod_{j=1}^{k-1}(1-\beta_j)$. Thus, when choosing $\beta_k \equiv \beta$, we get

$$T_1 = (1-(1-\kappa)\beta)^{k-1}\|V^* - V^{\pi_0}\|_\infty \quad \text{and} \quad T_3 = \tau(1-\beta)^{k-1}\nu_{\max}$$

The term T_2 can be simplified as follows

$$\begin{aligned}
&\kappa' \tau \nu_{\max} \beta \sum_{i=1}^{k-1} (1-\beta)^{i-1}(1-(1-\kappa)\beta)^{k-i-1} \\
&= \kappa' \tau \nu_{\max} \beta (1-(1-\kappa)\beta)^{k-2} \sum_{i=1}^{k-1} \frac{(1-\beta)^{i-1}}{(1-(1-\kappa)\beta)^{i-1}} \\
&= \kappa' \tau \nu_{\max} \beta (1-(1-\kappa)\beta)^{k-2} \frac{1 - \left(\frac{(1-\beta)}{(1-(1-\kappa)\beta)}\right)^{k-1}}{1 - \frac{(1-\beta)}{(1-(1-\kappa)\beta)}} \\
&= \kappa' \tau \nu_{\max} \beta \frac{\left[(1-(1-\kappa)\beta)^{k-1} - (1-\beta)^{k-1}\right]}{(1-\beta+\beta\kappa) - (1-\beta)} \\
&= \frac{\kappa'}{\kappa} \tau \nu_{\max} \left[(1-(1-\kappa)\beta)^{k-1} - (1-\beta)^{k-1}\right]
\end{aligned}$$

Now, observe that $\frac{\kappa'}{\kappa} = \min\left\{1, \frac{\kappa}{1-\kappa}\right\}\frac{1}{\kappa} = \min\left\{\frac{1}{\kappa}, \frac{1}{1-\kappa}\right\} \in (1,2)$ for $\kappa \in (0,1)$. Consequently, the bound on the value function gap simplifies as

$$\begin{aligned}
\|V^* - V^{\pi_k}\|_{\tilde{\xi}} &\leq (1-(1-\kappa)\beta)^{k-1}\|V^* - V^{\pi_0}\|_\infty \\
&\quad + \frac{\kappa'}{\kappa}\tau\nu_{\max}\left[(1-(1-\kappa)\beta)^{k-1} - (1-\beta)^{k-1}\right] + \tau\nu_{\max}(1-\beta)^{k-1} \\
&= (1-(1-\kappa)\beta)^{k-1}\|V^* - V^{\pi_0}\|_\infty + \frac{\kappa'}{\kappa}\tau\nu_{\max}(1-(1-\kappa)\beta)^{k-1} \\
&\quad + \nu_{\max}\left(1 - \frac{\kappa'}{\kappa}\right)(1-\beta)^{k-1} \\
&\leq (1-(1-\kappa)\beta)^{k-1}\left[\|V^* - V^{\pi_0}\|_\infty + 2\tau\nu_{\max}\right]
\end{aligned}$$

where the last inequality is due to $\kappa'/\kappa \in (1,2)$. Finally, by the equivalence of norms $\|\cdot\|_{\tilde{\xi}}$ and $\|\cdot\|_\infty$, we get

$$\|V^* - V^{\pi_k}\|_\infty \leq (1-(1-\kappa)\beta)^{k-1}\left[\frac{\|V^* - V^{\pi_0}\|_\infty + 2\tau\nu_{\max}}{1-\kappa}\right]$$

D.6 Proof of Technical Lemmas in Section D.5

D.6.1 Proof of Lemma D.8

We take the same approach as in our previous results showing monotonic policy improvement. It suffices to show that $Q^{\pi_k} \leq \widetilde{\mathcal{H}}^{\pi_{k+1}}(Q^{\pi_k})$ for every $k \geq 0$. Then, by monotonicity of $\widetilde{\mathcal{H}}^{\pi_{k+1}}$, we will have $Q^{\pi_k} \leq \widetilde{\mathcal{H}}^{\pi_{k+1}}(Q^{\pi_k}) \leq [\widetilde{\mathcal{H}}^{\pi_{k+1}}]^2(Q^{\pi_k}) \leq \ldots \leq [\widetilde{\mathcal{H}}^{\pi_{k+1}}]^n(Q^{\pi_k})$ for every $n \in \mathbb{N}$. Under Assumption D.2, π_{k+1} is a proper policy and hence letting $n \to \infty$ yields $Q^{\pi_{k+1}} = \lim_{n\to\infty}[\widetilde{\mathcal{H}}^{\pi_{k+1}}]^n(Q^{\pi_k})$ which implies $Q^{\pi_k} \leq Q^{\pi_{k+1}}$. Again, we prove the claim in two steps: $k = 0$ and $k \geq 1$.

(1) For $k = 0$, $\beta_0 = 1$ and $\bar{Q}_1 = Q^{\pi_0}$. By properness of π_0, we have

$$\begin{aligned}
Q^{\pi_0} &= \widetilde{\mathcal{H}}^{\pi_0}(Q^{\pi_0}) \\
&= \widetilde{\mathcal{H}}^{\pi_0}(\bar{Q}_1) \\
&= \widetilde{\mathcal{H}}^{\pi_0}_{\eta_0}(\bar{Q}_1) - \eta_0 g(\pi_0) \\
&\leq \widetilde{\mathcal{H}}_{\eta_0}(\bar{Q}_1) - \eta_0 g(\pi_0) && (\widetilde{\mathcal{H}}^{\pi}_{\eta}(\cdot) \leq \widetilde{\mathcal{H}}_{\eta}(\cdot)) \\
&= \widetilde{\mathcal{H}}^{\pi_1}_{\eta_0}(\bar{Q}_1) - \eta_0 g(\pi_0) && \text{(Algorithm 7, Line 4)} \\
&= \widetilde{\mathcal{H}}^{\pi_1}(\bar{Q}_1) + \eta_0 g(\pi_1) - \eta_0 g(\pi_0) \\
&\leq \widetilde{\mathcal{H}}^{\pi_1}(\bar{Q}_1)
\end{aligned}$$

where the last inequality follows from the initialization $\pi_0(s) \in \arg\max_{\mu\in\Delta(\mathcal{A})} \nu(\mu)$.

(2) For $k \geq 1$, since π_k is proper, we proceed as follows:

$$\begin{aligned}
Q^{\pi_k} &= \widetilde{\mathcal{H}}^{\pi_k}(Q^{\pi_k}) \\
&= \widetilde{\mathcal{H}}^{\pi_k}\left(\frac{\bar{Q}_{k+1} - (1-\beta_k)\bar{Q}_k}{\beta_k}\right) && \text{(Algorithm 7, Line 3)} \\
&= \widetilde{\mathcal{H}}^{\pi_k}_{\eta_{k-1}}\left(\frac{\bar{Q}_{k+1} - (1-\beta_k)\bar{Q}_k}{\beta_k}\right) - \eta_{k-1} g(\pi_k) \\
&= \frac{1}{\beta_k}\widetilde{\mathcal{H}}^{\pi_k}_{\eta_{k-1}}(\bar{Q}_{k+1}) - \left(\frac{1}{\beta_k} - 1\right)\widetilde{\mathcal{H}}^{\pi_k}_{\eta_{k-1}}(\bar{Q}_k) - \eta_{k-1} g(\pi_k) && (\widetilde{\mathcal{H}}^{\pi_k}_{\eta_{k-1}}(\cdot) \text{ being affine}) \\
&= \frac{1}{\beta_k}\widetilde{\mathcal{H}}^{\pi_k}_{\eta_k}(\bar{Q}_{k+1}) - \left(\frac{1}{\beta_k} - 1\right)\widetilde{\mathcal{H}}^{\pi_k}_{\eta_{k-1}}(\bar{Q}_k) + \left[\frac{(1-\beta_k)\eta_{k-1} - \eta_k}{\beta_k}\right] g(\pi_k) \\
&= \frac{1}{\beta_k}\widetilde{\mathcal{H}}^{\pi_k}_{\eta_k}(\bar{Q}_{k+1}) - \left(\frac{1}{\beta_k} - 1\right)\widetilde{\mathcal{H}}^{\pi_k}_{\eta_{k-1}}(\bar{Q}_k)
\end{aligned}$$

where the last step is due to $\eta_k = (1-\beta_k)\eta_{k-1}$. As per the update rule in Line 4 of Algorithm 7, $\widetilde{\mathcal{H}}^{\pi_{k+1}}_{\eta_k}(\bar{Q}_{k+1}) = \widetilde{\mathcal{H}}_{\eta_k}(\bar{Q}_{k+1})$ for $k \geq 0$, and we thus get

$$\begin{aligned}
\widetilde{\mathcal{H}}^{\pi_k}_{\eta_k}(\bar{Q}_{k+1}) &\leq \widetilde{\mathcal{H}}_{\eta_k}(\bar{Q}_{k+1}) = \widetilde{\mathcal{H}}^{\pi_{k+1}}_{\eta_k}(\bar{Q}_{k+1}) \\
\widetilde{\mathcal{H}}^{\pi_k}_{\eta_{k-1}}(\bar{Q}_k) &= \widetilde{\mathcal{H}}_{\eta_{k-1}}(\bar{Q}_k) \geq \widetilde{\mathcal{H}}^{\pi_{k+1}}_{\eta_{k-1}}(\bar{Q}_k)
\end{aligned}$$

Applying these in the bound on Q^{π_k} yields

$$\begin{aligned}
Q^{\pi_k} &\leq \frac{1}{\beta_k}\widetilde{\mathcal{H}}^{\pi_{k+1}}_{\eta_k}(\bar{Q}_{k+1}) - \left(\frac{1}{\beta_k} - 1\right)\widetilde{\mathcal{H}}^{\pi_{k+1}}_{\eta_{k-1}}(\bar{Q}_k) \\
&= \frac{1}{\beta_k}\widetilde{\mathcal{H}}^{\pi_{k+1}}(\bar{Q}_{k+1}) - \left(\frac{1}{\beta_k} - 1\right)\widetilde{\mathcal{H}}^{\pi_{k+1}}(\bar{Q}_k) + \left(\frac{\eta_k - (1-\beta_k)\eta_{k-1}}{\beta_k}\right) g(\pi_{k+1}) \\
&= \frac{1}{\beta_k}\widetilde{\mathcal{H}}^{\pi_{k+1}}(\bar{Q}_{k+1}) - \left(\frac{1}{\beta_k} - 1\right)\widetilde{\mathcal{H}}^{\pi_{k+1}}(\bar{Q}_k) && (\eta_k = (1-\beta_k)\eta_{k-1}) \\
&= \widetilde{\mathcal{H}}^{\pi_{k+1}}(Q^{\pi_k})
\end{aligned}$$

where the last equality is due to $\widetilde{\mathcal{H}}^{\pi_{k+1}}(\cdot)$ being affine and Line 3 of Algorithm 7.

D.7 Proof of Corollary D.10

Observe that since $\xi(s,a) > 1$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$,

$$\|Q^* - Q^{\pi_0}\|_\xi = \max_{s,a} \frac{|Q^*(s,a) - Q^{\pi_0}(s,a)|}{\xi(s,a)} < \max_{s,a} |Q^*(s,a) - Q^{\pi_0}(s,a)| = \|Q^* - Q^{\pi_0}\|_\infty$$

Further, from the proof of existence of ξ and κ ([7, Proposition 2.2]), we have

$$\xi(s,a) - 1 \leq \kappa \xi(s,a), \ \ \forall\, (s,a) \in \mathcal{S} \times \mathcal{A}$$

which implies

$$(1-\kappa)\xi(s,a) \leq 1, \text{ or equivalently, } \xi(s,a) \leq \frac{1}{1-\kappa}, \ \ \forall\, (s,a) \in \mathcal{S} \times \mathcal{A}$$

It then follows that

$$\|Q\|_\xi = \max_{s,a} \frac{|Q(s,a)|}{\xi(s,a)} \geq (1-\kappa) \max_{s,a} |Q(s,a)| = (1-\kappa)\|Q\|_\infty.$$

D.7.1 Proof of Lemma D.11

For any $k \geq 0$, we have by Line 3 of Algorithm 7 that

$$\begin{aligned}
Q^* - \bar{Q}_{k+1} &= Q^* - [(1-\beta_k)\bar{Q}_k + \beta_k Q^{\pi_k}] \\
&= Q^* - \bar{Q}_k + \beta_k \bar{Q}_k - \beta_k Q^{\pi_k} \\
&= Q^* - \bar{Q}_k + \beta_k \bar{Q}_k - \beta_k Q^* + \beta_k Q^* - \beta_k Q^{\pi_k} \\
&= (1-\beta_k)(Q^* - \bar{Q}_k) + \beta_k (Q^* - Q^{\pi_k}).
\end{aligned}$$

As argued in the proof of Lemma 5.2, since $\bar{Q}_k$ is a weighted average of $\{Q^{\pi_j}\}_{0 \leq j \leq k-1}$ and $Q^{\pi_j} \leq Q^{\pi_k}$ for all $j \leq k$, we have $\bar{Q}_k \leq Q^{\pi_k}$ for all $k \geq 0$. From this, we can deduce that

$$\begin{aligned}
Q^* - Q^{\pi_k} &= \widetilde{\mathcal{H}}(Q^*) - \widetilde{\mathcal{H}}^{\pi_k}(Q^{\pi_k}) \\
&\leq \widetilde{\mathcal{H}}(Q^*) - \widetilde{\mathcal{H}}^{\pi_k}(\bar{Q}_k) && (\widetilde{\mathcal{H}}^{\pi_k} \text{ being monotonic}) \\
&= \widetilde{\mathcal{H}}(Q^*) - \widetilde{\mathcal{H}}^{\pi_k}_{\eta_{k-1}}(\bar{Q}_k) + \eta_{k-1} g(\pi_k) \\
&= \widetilde{\mathcal{H}}(Q^*) - \widetilde{\mathcal{H}}_{\eta_{k-1}}(\bar{Q}_k) + \eta_{k-1} g(\pi_k) && (\text{Algorithm 7, Line 4}) \\
&\leq \widetilde{\mathcal{H}}(Q^*) - \widetilde{\mathcal{H}}(\bar{Q}_k) + \eta_{k-1} g(\pi_k) && (\widetilde{\mathcal{H}}(Q)(\cdot) \leq \widetilde{\mathcal{H}}_\eta(\cdot))
\end{aligned}$$

Under Assumption D.2, we have that $g(\pi) \leq \kappa' \nu_{\max} \mathbf{1}$ for any stationary policy π, which we show in Appendix D.7.2. With this, we proceed as follows. For any $(s,a) \in \mathcal{S} \times \mathcal{A}$, we have

$$\begin{aligned}
\frac{1}{\xi(s,a)} |Q^*(s,a) - Q^{\pi_k}(s,a)| &= \frac{1}{\xi(s,a)} [Q^*(s,a) - Q^{\pi_k}(s,a)] \\
&\leq \frac{1}{\xi(s,a)} \left\{ [\widetilde{\mathcal{H}}(Q^*)](s,a) - [\widetilde{\mathcal{H}}(\bar{Q}_k)](s,a) \right\} + \frac{\eta_{k-1}}{\xi(s,a)} [g(\pi_k)](s,a) \\
&\leq \frac{1}{\xi(s,a)} \left| [\widetilde{\mathcal{H}}(Q^*)](s,a) - [\widetilde{\mathcal{H}}(\bar{Q}_k)](s,a) \right| + \kappa' \eta_{k-1} \nu_{\max} \\
& && (\xi \geq \mathbf{1}) \\
&= \|\widetilde{\mathcal{H}}(Q^*) - \widetilde{\mathcal{H}}(\bar{Q}_k)\|_\xi + \kappa' \eta_{k-1} \nu_{\max} \\
&\leq \kappa \|Q^* - \bar{Q}_k\|_\xi + \kappa' \eta_{k-1} \nu_{\max},
\end{aligned}$$

where the last inequality follows due to Lemma D.9. Since the above holds for every $(s,a) \in \mathcal{S} \times \mathcal{A}$, we obtain

$$\|Q^* - Q^{\pi_k}\|_\xi \leq \kappa \|Q^* - \bar{Q}_k\|_\xi + \kappa' \eta_{k-1} \nu_{\max}$$

Now, we can bound $\|Q^* - \bar{Q}_{k+1}\|_\xi$ as follows.

$$\begin{aligned}
\|Q^* - \bar{Q}_{k+1}\|_\xi &\leq (1-\beta_k)\|Q^* - \bar{Q}_k\|_\xi + \beta_k \|Q^* - Q^{\pi_k}\|_\xi \\
&\leq (1 - (1-\kappa)\beta_k)\|Q^* - \bar{Q}_k\|_\xi + \kappa' \beta_k \eta_{k-1} \nu_{\max}
\end{aligned}$$

D.7.2 Bound on g

By definition of the operator g, we have for every $(s,a) \in \mathcal{S} \times \mathcal{A}$ that

$$[g(\pi)](s,a) = \sum_{s' \in \mathcal{S}} p(s' \mid s,a)\nu(\pi(s')), \quad s \in \mathcal{S},\ a \in \mathcal{A}$$

which satisfies $\widetilde{\mathcal{H}}_\eta^\pi(Q) = \widetilde{\mathcal{H}}^\pi(Q) + \eta g(\pi)$ for any $\eta \geq 0$ and any policy π. Now, by definition of $\widetilde{\mathcal{H}}(\cdot)$, we see that

$$[\widetilde{\mathcal{H}}(\mathbf{0})](s,a) = \mathcal{R}(s,a) \text{ and } [\widetilde{\mathcal{H}}(\mathbf{1})](s,a) = \mathcal{R}(s,a) + \sum_{s' \in \mathcal{S}} p(s'|s,a)$$

According to Lemma D.9, $\widetilde{\mathcal{H}}(\cdot)$ is a contraction mapping with respect to $\|\cdot\|_\xi$ and we thus have

$$\|\widetilde{\mathcal{H}}(\mathbf{1}) - \widetilde{\mathcal{H}}(\mathbf{0})\|_\xi = \max_{s,a} \frac{\sum_{s' \in \mathcal{S}} p(s'|s,a)}{\xi(s,a)} \leq \kappa \|\mathbf{1}\|_\xi \leq \kappa \|\mathbf{1}\|_\infty = \kappa$$

Since $1/\xi(s,a) \geq 1/[\max_{s,a} \xi(s,a)]$ for every (s,a), we get

$$\frac{1}{\max_{s,a} \xi(s,a)} \max_{s,a} \sum_{s' \in \mathcal{S}} p(s'|s,a) \leq \max_{s,a} \frac{\sum_{s' \in \mathcal{S}} p(s'|s,a)}{\xi(s,a)} \leq \kappa$$

$$\Rightarrow \max_{s,a} \sum_{s' \in \mathcal{S}} p(s'|s,a) \leq \kappa \max_{s,a} \xi(s,a) \leq \frac{\kappa}{1-\kappa}$$

Note that $\kappa/(1-\kappa) > 1$ only if $\kappa > 1/2$. Since $p(s'|s,a) \in [0,1)$, $\sum_{s' \in \mathcal{S}} p(s'|s,a) \leq 1$, and we get

$$[g(\pi)](s,a) = \sum_{s' \in \mathcal{S}} p(s' \mid s,a)\nu(\pi(s')) \leq \nu_{\max} \sum_{s' \in \mathcal{S}} p(s' \mid s,a) \leq \nu_{\max} \min\left\{1, \frac{\kappa}{1-\kappa}\right\} = \nu_{\max}\kappa'$$

for every $(s,a) \in \mathcal{S} \times \mathcal{A}$, and hence $g(\pi) \leq \kappa' \nu_{\max} \mathbf{1}$.

D.7.3 Proof of Lemma D.12

(1) For any $s \in \mathcal{S}$, we have that

$$\begin{aligned}
&\frac{1}{\tilde{\xi}(s)}[V^*(s) - V^{\pi_k}(s)] \\
&\leq \frac{1}{\tilde{\xi}(s)} \left[\max_{a \in \mathcal{A}} Q^*(s,a) - \pi_k(s)^\top \bar{Q}_k(s)\right] && (Q^{\pi_k} \geq \bar{Q}_k) \\
&= \frac{1}{\tilde{\xi}(s)} \left[\max_{a \in \mathcal{A}} Q^*(s,a) - \max_{\mu \in \Delta(\mathcal{A})} \left\{\mu^\top \bar{Q}_k(s) + \eta_{k-1}\nu(\mu)\right\}\right] + \frac{1}{\tilde{\xi}(s)} \eta_{k-1}\nu(\pi_k(s)) \\
& && \text{(Algorithm 7, Line 4)} \\
&\leq \frac{1}{\tilde{\xi}(s)} \left[\max_{a \in \mathcal{A}} Q^*(s,a) - \max_{a \in \mathcal{A}} \bar{Q}_k(s,a)\right] + \frac{1}{\tilde{\xi}(s)} \eta_{k-1}\nu(\pi_k(s)) \\
&\leq \frac{1}{\tilde{\xi}(s)} \max_{a \in \mathcal{A}} \left|Q^*(s,a) - \bar{Q}_k(s,a)\right| + \frac{1}{\tilde{\xi}(s)} \eta_{k-1}\nu(\pi_k(s)) \\
&\leq \max_{a \in \mathcal{A}} \frac{\left|Q^*(s,a) - \bar{Q}_k(s,a)\right|}{\xi(s,a)} + \frac{1}{\tilde{\xi}(s)} \eta_{k-1}\nu(\pi_k(s)) && (\tilde{\xi}(s) \geq \xi(s,a), \forall a \in \mathcal{A}) \\
&\leq \|Q^* - \bar{Q}_k\|_\xi + \frac{1}{\tilde{\xi}(s)} \eta_{k-1}\nu(\pi_k(s)) \\
&\leq \|Q^* - \bar{Q}_k\|_\infty + \eta_{k-1}\nu_{\max}
\end{aligned}$$

where the last inequality follows from the equivalence of norms $\|\cdot\|_\xi$ and $\|\cdot\|_\infty$, and $\xi \geq \mathbf{1}$.

(2) Since $\beta_0 = 1$, we have $\bar{Q}_1 = Q^{\pi_0}$. Now, for any (s,a), observe that

$$Q^*(s,a) - Q^{\pi_0}(s,a) = \sum_{s' \in \mathcal{S}} p(s' \mid s,a)(V^*(s') - V^{\pi_0}(s')) \leq \|V^* - V^{\pi_0}\|_\infty,$$

which implies $\|Q^* - Q^{\pi_0}\|_\infty \leq \|V^* - V^{\pi_0}\|_\infty$.

D.7.4 Monotonic Improvement Under A Weaker Assumption

We first recall the assumptions from Remark D.5.

Assumption D.13. (1) The policy π_0 is proper.
(2) For every improper policy π, there exists $(s,a) \in \mathcal{S} \times \mathcal{A}$ such that $Q^\pi(s,a) = -\infty$.

Under this assumption, we first establish the following result, whose proof follows arguments similar to those of [6, Proposition 2.4].

Lemma D.14. *Let Assumption D.13 hold. Let π be a proper policy and π' be any policy such that $Q^\pi \leq \widetilde{\mathcal{H}}^{\pi'}(Q^\pi)$. Then π' is also a proper policy.*

Proof. We prove by contradiction. Suppose π' is not proper. First, observe that by monotonicity of $\widetilde{\mathcal{H}}^{\pi'}$, $Q^\pi \leq \widetilde{\mathcal{H}}^{\pi'}(Q^\pi) \leq [\widetilde{\mathcal{H}}^{\pi'}]^2(Q^\pi) \leq \cdots \leq [\widetilde{\mathcal{H}}^{\pi'}]^n(Q^\pi)$ for every $n \in \mathbb{N}$. Now, for any $Q \in \mathbb{R}^{mn}$, $\widetilde{\mathcal{H}}^{\pi'}(Q) = \mathcal{R} + \mathcal{P}^{\pi'} Q$, where $\mathcal{P}^{\pi'}$ is a $mn \times mn$ matrix with $\mathcal{P}^{\pi'}((s,a),(s',a')) = p(s'|s,a)\pi'(a'|s')$ for $(s,a),(s',a') \in \mathcal{S} \times \mathcal{A}$. Consequently, $[\widetilde{\mathcal{H}}^{\pi'}]^n$ can be written as

$$\begin{aligned}
[\widetilde{\mathcal{H}}^{\pi'}]^n(Q) &= \widetilde{\mathcal{H}}^{\pi'}[\widetilde{\mathcal{H}}^{\pi'}]^{n-1}(Q) \\
&= \mathcal{R} + \mathcal{P}^{\pi'}[\widetilde{\mathcal{H}}^{\pi'}]^{n-1}(Q) \\
&= \mathcal{R} + \mathcal{P}^{\pi'}[\mathcal{R} + \mathcal{P}^{\pi'}[\widetilde{\mathcal{H}}^{\pi'}]^{n-2}(Q)] \\
&= \mathcal{R} + \mathcal{P}^{\pi'}\mathcal{R} + (\mathcal{P}^{\pi'})^2[\widetilde{\mathcal{H}}^{\pi'}]^{n-2}(Q).
\end{aligned}$$

Continuing in the above way thus leads to

$$[\widetilde{\mathcal{H}}^{\pi'}]^n(Q) = \sum_{i=0}^{n-1} (\mathcal{P}^{\pi'})^i \mathcal{R} + (\mathcal{P}^{\pi'})^n Q$$

Also notice that for any $i \in \{1,\ldots,n-1\}$, $(\mathcal{P}^{\pi'})^i((s,a),(s',a'))$ is the probability of going from (s,a) to (s',a') in i-transitions. Thus, we can write

$$\sum_{i=0}^{n-1} [(\mathcal{P}^{\pi'})^i \mathcal{R}](s,a) = \sum_{i=0}^{n-1} \mathbb{E}[\mathcal{R}(S_i,A_i) \mid S_0 = s, A_0 = a].$$

Since, π' is improper, by Assumption D.13(2), there exists $(s,a) \in \mathcal{S} \times \mathcal{A}$ such that $\limsup_n \sum_{i=0}^{n-1} [(\mathcal{P}^{\pi'})^i \mathcal{R}](s,a) = -\infty$ and hence $\limsup_n [(\widetilde{\mathcal{H}}^{\pi'})^n(Q)](s,a) = -\infty$. This implies that $Q^\pi(s,a) = -\infty$, which is a contradiction given π is proper. Hence, π' cannot be improper. □

With the above lemma in hand, it is now enough to show that π_k's are proper and $Q^{\pi_k} \leq \widetilde{\mathcal{H}}^{\pi_{k+1}}(Q^{\pi_k})$ for every $k \geq 0$. Indeed, this implies that π_{k+1} is proper by Lemma D.14, and $Q^{\pi_{k+1}}$ will be the unique solution to the Bellman equation $\widetilde{\mathcal{H}}^{\pi_{k+1}}(Q) = Q$. Following the same arguments as before, we can apply the monotonicity of $\widetilde{\mathcal{H}}^{\pi_{k+1}}(\cdot)$ to get $Q^{\pi_k} \leq \widetilde{\mathcal{H}}^{\pi_{k+1}}(Q^{\pi_k}) \leq [\widetilde{\mathcal{H}}^{\pi_{k+1}}]^2(Q^{\pi_k}) \leq \ldots \leq [\widetilde{\mathcal{H}}^{\pi_{k+1}}]^n(Q^{\pi_k})$ for every $n \in \mathbb{N}$. Since $Q^{\pi_{k+1}} = \lim_{n\to\infty} [\widetilde{\mathcal{H}}^{\pi_{k+1}}]^n(Q^{\pi_k})$, we will obtain $Q^{\pi_k} \leq Q^{\pi_{k+1}}$.

To prove the claim, we proceed as follows. Starting from π_0, the arguments in Appendix D.6.1 (1) imply that $Q^{\pi_0} \leq \widetilde{\mathcal{H}}^{\pi_1}(Q^{\pi_0})$. Since π_0 is proper, it follows from Lemma D.14 that π_1 is also proper. Next, the arguments in Appendix D.6.1 (2) ensure that $Q^{\pi_1} \leq \widetilde{\mathcal{H}}^{\pi_2}(Q^{\pi_1})$ and hence Lemma D.14 implies that π_2 is proper. Continuing inductively, applying the arguments in Appendix D.6.1 (2) and Lemma D.14, we can conclude that π_k's are proper and $Q^{\pi_k} \leq \widetilde{\mathcal{H}}^{\pi_{k+1}}(Q^{\pi_k})$ for every $k \geq 0$.

E Numerical Experiments

We consider randomly generated MDPs with state space $\mathcal{S} = \{1,2,\ldots,20\}$, action space $\mathcal{A} = \{1,2\}$, and discount factor $\gamma = 0.99$. Thus, $n = 20$ and $m = 2$. For each $(s,a) \in \mathcal{S} \times \mathcal{A}$, the transition probabilities $p(\cdot \mid s,a)$ are drawn uniformly at random from the probability simplex $\Delta(\mathcal{S})$, i.e.,

independently according to a Dirichlet distribution with parameter vector $\mathbf{1} \in \mathbb{R}^n$, and the reward $R(s,a)$ is drawn uniformly at random from $[0,1]$, independently across state-action pairs. We generate 30 such MDPs independently, and for each of them we compute the Q-function Q^{π_k} of every iterate exactly by solving the linear Bellman equation $Q = \mathcal{H}^{\pi_k}(Q)$, and the optimal value function V^* by policy iteration, which terminates in finitely many iterations.

In Figure 1, we plot the value function gap $\|V^* - V^{\pi_k}\|_\infty$ of the last iterate π_k against the number of iterations k. In particular, we show the convergence of the error for three key special cases of DSPI:

- PI: $\nu \equiv 0$ and $\beta_k \equiv 1$
- dual-averaged PI: $\nu \equiv 0$, $\beta_0 = 1$, and $\beta_k = 1/2$ for $k \geq 1$
- NPG: $\nu = h$, $\tau = 1/\log m$, $\beta_0 = 1$, and $\beta_k = 1/2$ for $k \geq 1$

For PI and dual-averaged PI, the initial policy π_0 is deterministic, while for NPG, π_0 is the uniform policy.

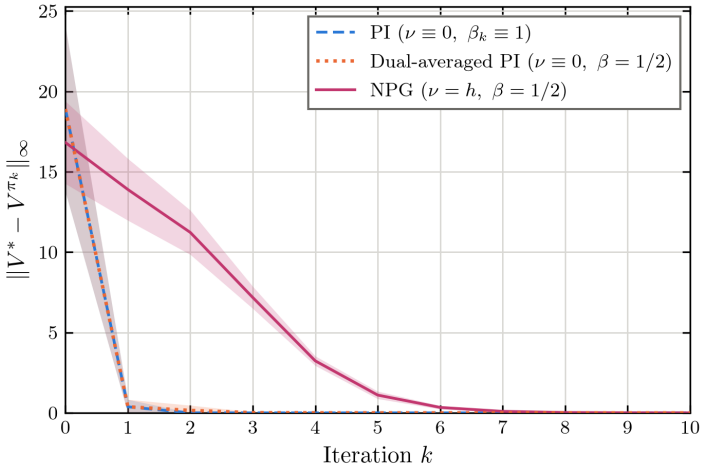

Figure 1: Value function gap $\|V^* - V^{\pi_k}\|_\infty$ of the last iterate versus the number of iterations k for three special cases of DSPI: PI, dual-averaged PI, and NPG. The curves show the mean over 30 randomly generated MDPs with $n = 20$, $m = 2$, and $\gamma = 0.99$, and the shaded regions indicate one standard deviation.

In all three cases, the value function gap decreases monotonically and converges to zero, in agreement with Theorem 4.1. Recall that for $\beta_0 = 1$ and $\beta_k \equiv \beta$ for $k \geq 1$, Theorem 4.1 guarantees that

$$\|V^* - V^{\pi_k}\|_\infty \leq \big(1 - (1-\gamma)\beta\big)^{k-1}\big(\gamma\|V^* - V^{\pi_0}\|_\infty + \tau\nu_{\max}\big), \qquad \forall\, k \geq 1.$$

The ordering of the three curves is consistent with this bound. PI, which corresponds to $\beta = 1$ and $\nu \equiv 0$, enjoys the fastest guaranteed rate γ and converges fastest, reaching an optimal policy within three iterations. Dual-averaged PI and NPG, which correspond to $\beta = 1/2$, have the slower guaranteed rate $(1 - (1-\gamma)/2)$. Although not clearly visible in the plot, dual-averaged PI lags slightly behind PI because the running average $\bar{Q}_k$ retains the Q-functions of earlier policies. NPG decreases more slowly during the first few iterations, because of the regularization term until the multiplier $\eta_k = \tau(1-\beta)^k$ has decayed, after which the error drops rapidly.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims in the abstract and introduction accurately reflect the paper's contributions and scope.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The limitations are discussed in the conclusion section.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: The theorem statements and assumptions are presented in the main body of the paper, while the complete proofs are deferred to the appendix due to space limitations.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.

- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The experimental set-up is described in detail in Appendix E.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: The numerical experiments in this work use randomly generated Markov decision processes and do not require any dataset. Although the paper does not provide open access to the code, the experimental set-up and implementation details are described in Appendix E.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.

- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The details of the experimental setting are included in Appendix E.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The numerical experiments in this work use randomly generated Markov decision processes. Figure 1 shows the mean value-function gap over 30 independently generated Markov decision processes, where the shaded regions indicate one standard deviation computed across these realizations.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [No]

Justification: The numerical experiments are based on simulated Markov decision processes of modest size that can be simulated on a standard computer without specialized hardware or substantial computational resources.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

Answer: [Yes]

Justification: The research conducted in the paper conforms, in every respect, with the NeurIPS Code of Ethics.

Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [N/A]

Justification: This is a theoretical work, and no societal impact of the work is identified.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer:[N/A]

    Justification: This is a theoretical work, and no data are involved.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [N/A]

    Justification: The numerical experiments are based on a simulated set-up and do not rely on any existing datasets, pretrained models, or externally sourced code.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not release any new datasets, code, or models.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

Justification: This is a theoretical work, and no human subjects are involved in the research.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: This is a theoretical work, and no human subjects are involved in the research.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: We only used LLMs for polishing the language.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.