Reinforcement Learning for Exponential Utility: Algorithms and Convergence in Discounted MDPs

Gugan Thoppe
Computer Science and Automation (CSA)
Indian Institute of Science (IISc)
Bengaluru 560012, India
gthoppe@iisc.ac.in

Prashanth L. A.
Computer Science and Engineering,
Indian Institute of Technology Madras
Chennai 600036, India
prashla@cse.iitm.ac.in

Ankur Naskar
Computer Science and Automation (CSA)
Indian Institute of Science (IISc)
Bengaluru 560012, India
ankurnaskar@iisc.ac.in

Sanjay P. Bhat
TCS Research,
IIIT-H Research Park,
Hyderabad 500032, India
sanjay.bhat@tcs.com

Abstract

Risk-sensitive reinforcement learning with exponential utility in discounted Markov decision processes remains unexplored from a model-free, value-based perspective. We address this gap in the risk-aversion setting. Building on a Bellman-type equation for exponential utility from [Porteus, 1975], we derive two Q-value-style extensions and show that the associated operators are contractions in the $\|\cdot\|_\infty$ and Thompson metrics, respectively. We characterize the fixed points of these operators and show that the induced greedy stationary policy minimizes an upper bound on the exponential-utility objective among stationary policies. Using these structural results, we develop two model-free algorithms: a two-timescale Q-learning variant and a one-timescale algorithm based on a sublinear power-law operator. We prove almost-sure convergence under Markovian sampling for both algorithms. While the two-timescale method is analyzed using timescale separation, the one-timescale variant needs a different approach based on local Lipschitzness, monotonicity, homogeneity, and Dini derivatives, as its operator is not globally contractive in a standard metric. We further derive finite-time convergence rates for the two-timescale method under i.i.d. sampling and for the one-timescale method in the scalar case. Our work opens up new possibilities for model-free learning of optimal policies for exponential-utility objectives, beyond the class of stationary policies.

1 Introduction

The traditional objective in reinforcement learning (RL) [Bertsekas and Tsitsiklis, 1996, Puterman, 1994, Sutton and Barto, 2018] is to find a policy that maximizes the expected cumulative reward. However, two policies with the same expected return may have very different return distributions, and hence different levels of risk. This is particularly important in applications such as finance, healthcare, and robotics. Risk-sensitive RL [Prashanth and Fu, 2022], which has garnered increased research attention over the last decade, aims to account for such differences. Broadly, risk-sensitive RL accounts for risk either through constraints or by incorporating risk directly into the objective. In the constrained setting, one maximizes the expected return subject to a constraint on a chosen risk measure. In the unconstrained setting, one optimizes an objective that directly accounts for risk. Risk measures studied in RL include variance [Mihatsch and Neuneier, 2002, Tamar et al., 2012] [Prashanth and Ghavamzadeh, 2016], quantile [Jiang and Powell, 2017], conditional value at

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

risk (CVaR) [Prashanth, 2014, Thoppe et al., 2024], distortion risk [Markowitz et al., 2023], and coherent risk [Tamar et al., 2016]. In this work, we focus on exponential utility as an unconstrained risk-sensitive objective [Bäuerle and Rieder, 2014].

Exponential-utility optimization has been studied extensively in average-reward MDPs [Borkar, 2010, Moharrami et al., 2022]. In discounted MDPs, however, the Bellman recursion is time-inhomogeneous: the optimal value at risk sensitivity λ depends on that at $\gamma\lambda$, which in turn depends on that at $\gamma^2\lambda$ and so on [Jaquette, 1976, Chung and Sobel, 1987]. Thus, computing the optimal value at λ requires determining the optimal values at infinitely many risk-sensitivity parameters. This contrasts with the risk-neutral setting, where the optimal value is characterized by a single fixed-point equation. An optimal policy for exponential utility, moreover, need not be stationary. While it can be chosen to be ultimately stationary [Jaquette, 1976], the stationarity horizon is generally unknown. Even if this horizon can be approximated, computing an optimal stationary tail remains challenging, as the optimal value over stationary policies generally satisfies only a time-inhomogeneous Bellman inequality. Existing model-based methods [Kumar et al., 2019, Hau et al., 2023] address this difficulty by using an optimal risk-neutral stationary policy as the tail and its value to initialize a finite backward Bellman recursion for computing the preceding values and decision rules. However, this approach ignores the residual risk sensitivity at the chosen horizon. Our work asks whether optimal risk-sensitive stationary policies can be learned using model-free RL.

A different formulation of discounted risk-sensitive control based on exponential utility was studied by Porteus [1975], who imposed an interstage consistency requirement. This requirement keeps the risk-sensitivity parameter fixed across the original Bellman recursion, leading to two time-homogeneous Bellman fixed-point equations, one for each stationary policy and another obtained by optimizing over actions. Porteus further showed that a stationary policy greedy with respect to the latter fixed point is optimal for the criterion defined by his Bellman recursion. However, how this policy performs under the direct exponential utility of the discounted return remains unclear. In this work, we study the Porteus formulation for $\lambda > 0$ and ask the following questions:

1. *Can we develop model-free RL algorithms to solve the Porteus Bellman optimality equation?*
2. *What relationship can we establish between the Porteus formulation and the direct exponential-utility objective?*

We address these questions through the following **key contributions**.

- **Bellman-type operators:** Building on Porteus [1975]'s state-value Bellman formulation, we study two Q-value extensions, covering both policy evaluation and optimality. We show that the associated operators are contractions in the $\|\cdot\|_\infty$ and sup-log/Thompson metrics, respectively. These results provide the foundation for our model-free RL algorithms.
- **Upper bound for direct exponential utility:** We show that, for any stationary policy, the fixed point of our policy-specific Q-value operator provides an upper bound on its direct exponential-utility Q-values. We further show that the greedy policy obtained from the optimal Q-value fixed point minimizes these upper bounds among all stationary policies.
- **Q-learning–style algorithms:** We develop the first tabular, model-free Q-learning–style algorithms for learning risk-sensitive stationary policies under the Porteus formulation in discounted MDPs. We propose two algorithms, one based on two-timescale stochastic approximation and the other on one-timescale stochastic approximation.
- **Two-timescale analysis:** For the proposed two-timescale algorithm, we establish almost-sure convergence under Markovian sampling and finite-time convergence rates under i.i.d. sampling, using timescale-separation arguments. These results extend classical Q-learning-style convergence guarantees to the exponential-utility setting.
- **One-timescale analysis:** For our one-timescale algorithm, whose dynamics are governed by a sublinear power-law operator, we prove almost-sure convergence under Markovian sampling using a novel analysis based on local Lipschitzness, monotonicity, homogeneity, and Dini derivatives. Under i.i.d. sampling, we also provide finite-time analysis of an analogous scalar recursion, that highlights the challenges due to the lack of a global contraction in standard metrics.

Related work. Exponential-utility optimization has been studied extensively in average-reward MDPs. Borkar [2002, 2001] developed tabular, model-free Q-learning and actor–critic algorithms,

while Basu et al. [2008] studied the linear function approximation case. More recently, Moharrami et al. [2022] developed a trajectory-based policy-gradient algorithm, and Murthy et al. [2023] established convergence guarantees for model-based modified policy iteration. See also Borkar [2010] for a survey. Of particular relevance to our work is the Q-learning algorithm of Borkar [2002], which uses exponential-coordinate updates similar to our one-timescale method.

For exponential utility with discounted returns, Jaquette [1976] showed that the Bellman recursion is time-inhomogeneous, as the risk parameter changes over time. Chung and Sobel [1987] showed that Markov policies still suffice for optimality, but need not be stationary. Jaquette [1976] further established the existence of optimal ultimately stationary policies, although the stationarity horizon depends on typically unknown problem-specific quantities. More recently, Bäuerle et al. [2026] revisited these results with simpler proofs and studied policy stability. Model-based approaches for approximating optimal policies were developed by Kumar et al. [2019], Hau et al. [2023]. For a chosen finite horizon, they use an optimal risk-neutral stationary policy as the tail and backward Bellman recursions to compute the preceding time-dependent decision rules. To approximate optimal stationary policies, Porteus [1975] introduced a time-homogeneous Bellman equation based on a fixed-point version of the exponential-utility Bellman recursion. Bäuerle and Glauner [2022] extended similar results to more general risk measures. For an overview, see Bäuerle and Jaśkiewicz [2024].

Model-free algorithms for exponential utility have also been developed for finite-horizon MDPs. Fei et al. [2020] proposed risk-sensitive value-iteration and Q-learning algorithms with regret guarantees, while Fei et al. [2021a] obtained improved guarantees using an exponential Bellman equation. Extensions to function approximation were developed by Fei et al. [2021b], and Liang and Luo [2024] established regret bounds for model-free and model-based distributional RL algorithms with entropic-risk objectives. These works focus on episodic problems.

More closely related to our work, Mortensen and Talebi [2026] established finite-sample guarantees for model-based Q-value iteration under recursive entropic risk in discounted MDPs. Their method estimates transition probabilities using a generative model, whereas ours learns the Bellman fixed point directly from sampled transitions.

Model-free policy-gradient and actor–critic methods for exponential utility have also been studied [Noorani et al., 2025, Granados and Pacheco, 2026], including applications to large language model alignment [Jiang et al., 2025]. Unlike these policy-based methods, our Q-learning–style algorithms for the Porteus formulation have almost-sure convergence and finite-time guarantees. We also connect the Porteus fixed points to the direct exponential utility of discounted returns.

2 Discounted reward MDP with exponential utility

We consider a discounted MDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{P}, r, \gamma)$. Here, $\mathcal{S}$ and $\mathcal{A}$ are finite state and action spaces with cardinalities $S = |\mathcal{S}|$ and $A = |\mathcal{A}|$, respectively. Further, $\mathcal{P} : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ is the transition kernel, where $\Delta(\mathcal{U})$ denotes the set of distributions over $\mathcal{U}$; in particular, $\mathcal{P}(s'|s, a)$ specifies the probability that the state of the MDP changes from s to s' under action a. Finally, $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is the reward function, and $\gamma \in (0, 1)$ the discount factor. Our goal is to develop model-free Q-learning–style algorithms for stationary policies under exponential utility.

Let Π denote the set of all policies, including randomized and history-dependent ones. For each $\pi \in \Pi$ and risk-sensitivity parameter $\lambda > 0$, define its direct exponential utility $X_\pi(\lambda) \in \mathbb{R}^{SA}_{++}$ by $X_\pi(s, a; \lambda) = \mathbb{E}^\pi_{s,a}\left[\exp\left(-\lambda \sum_{t=0}^\infty \gamma^t r(s_t, a_t)\right)\right]$, where $\mathbb{R}^{SA}_{++} := (0, \infty)^{SA}$ is the positive orthant in $\mathbb{R}^{SA}$ and $\mathbb{E}^\pi_{s,a}$ denotes expectation under the process initialized at $(s_0, a_0) = (s, a)$, with policy π followed from time 1 onward. The direct exponential-utility objective is to minimize $X_\pi(\lambda)$ over $\pi \in \Pi$. Following the arguments of [Jaquette, 1976] (see [Chung and Sobel, 1987, (8)]), the optimal value $X^*(s, a; \lambda) = \inf_{\pi\in\Pi} X_\pi(s, a; \lambda)$ satisfies the time-inhomogeneous Bellman recursion

$$X^*(s, a; \lambda) = e^{-\lambda r(s,a)} \sum_{s' \in \mathcal{S}} \mathcal{P}(s' \mid s, a) \min_{a' \in \mathcal{A}} X^*(s', a'; \gamma\lambda).$$

This recursion is time-inhomogeneous because the effective risk parameter changes over time. This implies the optimal decision rule may change even though the MDP is stationary. Indeed, Chung and Sobel [1987] showed that Markov policies suffice for optimality, but need not be stationary. Nevertheless, Jaquette [1976] established the existence of optimal ultimately stationary policies,

i.e., policies which become stationary after finitely many steps. However, this stationarity horizon depends on typically unknown problem-specific quantities.

To approximate such policies, Kumar et al. [2019], Hau et al. [2023] developed model-based methods that use a risk-neutral stationary policy beyond a chosen horizon k and compute preceding decision rules through backward Bellman recursion. However, the risk-neutral tail ignores the residual risk sensitivity $\lambda\gamma^k$. Motivated by this limitation, we seek a risk-sensitive policy that is stationary from the outset, corresponding to a stationarity horizon of $k = 0$.

Denote the set of all stationary Markovian policies (including randomized ones) for the MDP $\mathcal{M}$ by Π_{st}. Optimizing direct exponential utility over Π_{st} is challenging, as the time-inhomogeneous Bellman recursion does not directly characterize the optimal stationary policy. We therefore consider an alternative formulation originating from Porteus [1975], which we present in Q-value form. To motivate this formulation, define $Q_*^D(s,a;\lambda) = -\frac{1}{\lambda}\ln X^*(s,a;\lambda)$. The preceding Bellman recursion concerning X^* then becomes

$$Q_*^D(s,a;\lambda) = r(s,a) - \frac{1}{\lambda}\ln\left[\sum_{s'\in\mathcal{S}} \mathcal{P}(s' \mid s,a)\exp\left(-\lambda\gamma \max_{a'\in\mathcal{A}} Q_*^D(s',a';\gamma\lambda)\right)\right].$$

To avoid this time-inhomogeneity, Porteus's formulation leads to the fixed-point equation.

$$Q(s,a) = r(s,a) - \frac{1}{\lambda}\ln\left[\sum_{s'\in\mathcal{S}} \mathcal{P}(s' \mid s,a)\exp\left(-\lambda\gamma \max_{a'\in\mathcal{A}} Q(s',a')\right)\right].$$

We next study this recursive formulation, establish the existence of an optimal stationary policy for this formulation, and examine its relationship with direct exponential utility.

To analyze the Porteus formulation, we work in exponential coordinates $x(s,a) = e^{-\lambda Q(s,a)}$. Then, the preceding fixed-point equation becomes $x = F(x)$, where $F : \mathbb{R}_{++}^{SA} \to \mathbb{R}_{++}^{SA}$ is defined by

$$F(x)(s,a) = \exp\left(-\lambda r(s,a)\right)\sum_{s'\in\mathcal{S}} \mathcal{P}(s'|s,a)\left[\min_{a'} x(s',a')\right]^{\gamma}. \tag{1}$$

For any stationary policy $\pi \in \Pi_{\mathrm{st}}$, we also define the operator $F_\pi : \mathbb{R}_{++}^{SA} \to \mathbb{R}_{++}^{SA}$ by

$$F_\pi(x)(s,a) = \exp\left(-\lambda r(s,a)\right)\sum_{(s',a')\in\mathcal{S}\times\mathcal{A}} \mathcal{P}(s'|s,a)\pi(a'|s')\left[x(s',a')\right]^{\gamma}. \tag{2}$$

Consider the sup-log or Thompson metric $d(\cdot,\cdot) : \mathbb{R}_{++}^{SA} \times \mathbb{R}_{++}^{SA} \to [0,\infty)$ on $\mathbb{R}_{++}^{SA}$ defined by

$$d(x_1,x_2) := \|\ln x_1 - \ln x_2\|_\infty, \qquad x_1, x_2 \in \mathbb{R}_{++}^{SA}. \tag{3}$$

It is not difficult to see that $\mathbb{R}_{++}^{SA}$ is a complete metric space under the sup-log metric d. Our first result shows that, under this geometry, the operators F and F_π satisfy the structural properties that drive the rest of our analysis. For any $x_1, x_2 \in \mathbb{R}_{++}^{SA}$, let $x_1 \preceq x_2$ (equivalently, $x_2 \succeq x_1$) denote that $x_1(s,a) \le x_2(s,a)$ for every $(s,a) \in \mathcal{S}\times\mathcal{A}$.

Proposition 2.1 (***Structural properties of*** F ***and*** F_π)**.** *The following statements are true.*

1. *F is a γ-contraction in $d(\cdot,\cdot)$, that is, $\forall x_1, x_2 \in \mathbb{R}_{++}^{SA}$ we have $d(F(x_1),F(x_2)) \le \gamma d(x_1,x_2)$.*
2. *For each $\pi \in \Pi_{\mathrm{st}}$, F_π is a γ-contraction in $d(\cdot,\cdot)$.*
3. *Let $\pi \in \Pi_{\mathrm{st}}$. Then F and F_π are monotone, that is, for every $x_1, x_2 \in \mathbb{R}_{++}^{SA}$ such that $x_1 \preceq x_2$, we have $F(x_1) \preceq F(x_2)$ and $F_\pi(x_1) \preceq F_\pi(x_2)$.*
4. *For each $x \in \mathbb{R}_{++}^{SA}$, we have $F(x) = \min_{\pi\in\Pi_{\mathrm{st}}} F_\pi(x)$, where the minimum is taken element-wise.*
5. *F is homogeneous, that is, for each $c > 0$ and $x \in \mathbb{R}_{++}^{SA}$, $F(cx) = c^\gamma F(x)$.*

Proposition 2.1 identifies the geometry under which F and F_π, $\pi \in \Pi_{\mathrm{st}}$, are contractive. Since $\mathbb{R}_{++}^{SA}$ is complete under d, Banach's fixed-point theorem implies that F and each F_π admit unique fixed points, denoted by x^* and $\bar{x}_\pi$, respectively. Let $\pi^* \in \Pi_{\mathrm{st}}$ be a stationary deterministic policy that selects an action in $\arg\min_{a\in\mathcal{A}} x^*(s,a)$ at each state $s \in \mathcal{S}$. Our next main result establishes that π^* is optimal for the recursive formulation, in the sense that $\bar{x}_{\pi^*} = x^* \preceq \bar{x}_\pi$ for every $\pi \in \Pi_{\mathrm{st}}$. It also shows that $\bar{x}_\pi$ provides an upper bound on the direct exponential utility $X_\pi(\lambda)$.

Theorem 2.2. *Let x^* and π^* be as defined above. Then, for every $\pi \in \Pi_{\mathrm{st}}$, we have*

$$X_\pi(\lambda) \preceq \bar{x}_\pi \qquad \textit{and} \qquad x^* = \bar{x}_{\pi^*} \preceq \bar{x}_\pi. \tag{4}$$

Remark 2.3. *The Porteus-optimal stationary policy need not minimize direct exponential utility, even among stationary policies (Appendix A.3). However, for finite MDPs with nonnegative rewards and uniformly absorbing dynamics, the optimal values of the two criteria differ by at most $O(1-\gamma)$ as $\gamma \uparrow 1$ (Appendix A.4).*

Theorem 2.2 establishes the optimality of x^* for the Porteus formulation and shows that $\bar{x}_\pi$ upper-bounds the direct exponential utility $X_\pi(\lambda)$. We next develop two model-free algorithms for approximating x^*, one of which uses a Q-value operator associated with Porteus [1975]. For the subsequent analysis, we reparameterize $\lambda = \theta/\gamma$, where $\theta > 0$, and define $T : \mathbb{R}^{SA} \to \mathbb{R}^{SA}$ by

$$T(Q)(s,a) = r(s,a) - \frac{\gamma}{\theta} \ln \left[\sum_{s' \in \mathcal{S}} \mathcal{P}(s' \mid s,a) \exp\left(-\theta \max_{a' \in \mathcal{A}} Q(s',a') \right) \right]. \tag{5}$$

The operators F and T are related through

$$T(Q) = -\frac{\gamma}{\theta} \ln F\left(\exp\left(-\frac{\theta}{\gamma} Q \right) \right), \tag{6}$$

where the operations are applied element-wise. Consequently, T has the unique fixed point $Q^* = -\frac{\gamma}{\theta} \ln x^*$. Moreover, Proposition 2.1 implies that T is a γ-contraction in the sup norm.

2.1 Illustrative Example: Risk-averse behavior induced by T

The following example shows the qualitative difference between a policy obtained from the classical risk-neutral Bellman operator and one obtained from the risk-sensitive formulation represented by T.

Consider a discounted MDP with state space $\mathcal{S} = \{s, s_{\text{bad}}\}$, discount factor $\gamma = 0.9$, and risk parameter $\theta = 0.1$. The state s_{bad} is absorbing and yields a reward of -10, i.e., there is a single action a such that

$$r(s_{\text{bad}}, a) = -10, \qquad \mathcal{P}(s_{\text{bad}} \mid s_{\text{bad}}, a) = 1.$$

At state s, there are two available actions, safe and risk. The safe action yields zero reward and leaves the state unchanged:

$$r(s, \mathrm{safe}) = 0, \qquad \mathcal{P}(s \mid s, \mathrm{safe}) = 1.$$

In contrast, the risk action yields a reward of 1 and induces the following transition probabilities:

$$r(s, \mathrm{risk}) = 1, \qquad \mathcal{P}(s_{\text{bad}} \mid s, \mathrm{risk}) = 0.01, \qquad \mathcal{P}(s \mid s, \mathrm{risk}) = 0.99.$$

Solving the classical discounted risk-neutral Bellman equations shows that the optimal Q-values satisfy

$$Q^*_{\mathrm{rn}}(s_{\text{bad}}, a) = -100, \qquad Q^*_{\mathrm{rn}}(s, \mathrm{safe}) \approx 0.826, \qquad Q^*_{\mathrm{rn}}(s, \mathrm{risk}) \approx 0.917,$$

where rn denotes risk neutral. Thus, the optimal risk-neutral policy selects the risk action at state s.

In contrast, the fixed point Q^* of our operator T satisfies

$$Q^*(s_{\text{bad}}, a) = -100, \qquad Q^*(s, \mathrm{safe}) = 0, \qquad Q^*(s, \mathrm{risk}) \approx -47.59.$$

Thus, in this case, the optimal policy selects the safe action at state s.

In summary, while the risk-neutral formulation favors the risk action due to its higher expected return, the operator T assigns it a substantially lower value by penalizing the possibility of rare but severe negative outcomes.

3 Model-free algorithms for optimizing exponential utility

The previous section motivated the fixed points of the T and F operators. In this section, we introduce two model-free algorithms for finding them.

Observe that, by treating s' as a random variable, the summations inside the definitions (5) and (1) of T and F, respectively, can be viewed as conditional expectations. The key difference is where this expectation appears. In the definition of T, it appears inside the $\log$ function; hence, unlike classical Q-learning, a single transition sample does not directly yield a stochastic estimate of the full operator value. In contrast, in the definition of F, the expectation appears linearly at the outermost level, so a single transition sample gives a natural stochastic estimate of $F(x)$. Keeping this distinction in mind, we now propose two model-free algorithms to find the fixed points of T and F.

Let $\{(s_n, a_n)\}_{n\geq 0}$ be a sequence of state-action pairs obtained using a (possibly time-varying) behavior policy. Specifically, let (s_0, a_0) be an arbitrary state-action pair and, for $n \geq 0$, suppose $s_{n+1} \sim \mathcal{P}(\cdot|s_n, a_n)$ and $a_{n+1} \sim \pi_n(\cdot|s_{n+1})$, where π_n is the behavior policy at time instance n.

3.1 Two-timescale algorithm for estimating Q^*

The first algorithm is motivated by classical Q-learning. Since Q^* is the unique fixed point of T, a natural approach is to perform a stochastic fixed-point iteration for T. The main difference from the risk-neutral case is that the conditional expectation in T lies inside a logarithm. To handle this, we introduce an auxiliary iterate g_n that tracks this inner expectation on a faster timescale. This leads to

$$\begin{aligned} Q_{n+1} &= Q_n + \alpha_n \left[-\frac{\gamma}{\theta} \ln g_n - Q_n \right], \\ g_{n+1} &= g_n + \beta_n e_{s_n,a_n} [\hat{G}_n - g_n(s_n, a_n)], \end{aligned} \tag{7}$$

where $\hat{G}_n = G(Q_n, s_n, a_n, s_{n+1})$ and

$$G(Q, s, a, s') = \exp\left[-\frac{\theta}{\gamma} r(s, a) - \theta \max_{a'} Q(s', a') \right]. \tag{8}$$

Here, $e_{s,a}$ denotes the standard basis vector in $\mathbb{R}^{SA}$ whose (s, a)-th entry is 1. The faster recursion for g_n estimates the conditional expectation inside $T(Q_n)$, while the slower recursion for Q_n uses $\log g_n$ as a plug-in estimate and moves toward Q^*.

3.2 One-timescale algorithm for estimating x^*

The second algorithm is motivated by the F-operator. Since x^* is the unique fixed point of F, and since the conditional expectation in F appears linearly at the outermost level, a single transition sample provides a natural stochastic estimate of $F(x_n)(s_n, a_n)$. This suggests the direct one-timescale recursion

$$x_{n+1} = x_n + \alpha_n e_{s_n,a_n} [\hat{F}_n - x_n(s_n, a_n)], \tag{9}$$

where

$$\hat{F}_n = \exp\left(-\frac{\theta}{\gamma} r(s_n, a_n) \right) \left[\min_{a'} x_n(s_{n+1}, a') \right]^{\gamma}. \tag{10}$$

This update should be contrasted with the sample-based multiplicative update suggested by the sup-log contraction of F:

$$x_{n+1} = x_n^{1-\alpha_n} \hat{F}_n^{\alpha_n},$$

where powers and products are understood componentwise. Although this update is natural from the sup-log viewpoint, it is not a stochastic approximation to the corresponding logarithmic fixed-point iteration. Indeed, in logarithmic coordinates it involves $\log \hat{F}_n$, whereas the desired drift contains $\log F(x_n)$; in general,

$$\mathbb{E}\Big[\log \hat{F}_n \mid s_n, a_n \Big] \neq \log F(x_n)(s_n, a_n).$$

Thus, exploiting the sup-log contraction of F in a model-free manner would again require estimating the conditional expectation inside F before taking the logarithm, leading naturally to an additional tracking recursion. The additive update (9) avoids this issue by working directly in the original x-coordinates, yielding a genuine one-timescale algorithm. Once x_n is close to x^*, an estimate of Q^* is obtained through the transformation $Q^* = -\frac{\gamma}{\theta} \ln x^*$. The price for this simplicity is analytical: convergence can no longer be obtained by directly invoking the sup-log contraction property of F.

Remark 3.1. *From a computational perspective, the one-timescale algorithm is preferable. Each iteration requires $\mathcal{O}(1)$ work, whereas the two-timescale method requires $\mathcal{O}(|\mathcal{S}||\mathcal{A}|)$ work because of the Q_n-update. The one-timescale method also has fewer hyperparameters: it uses a single step size sequence, while the two-timescale method requires separate step sizes for the Q_n- and g_n-updates. The per-iteration cost of the two-timescale method could be reduced to $\mathcal{O}(1)$ by updating only one coordinate of Q_n at each iteration. Beyond these computational considerations, the two-timescale method is a valuable addition to the literature since it is a direct algorithmic realization of the underlying Bellman equation $TQ = Q$ studied by Porteus [1975], Chung and Sobel [1987]. Moreover, T is monotone and a contraction in the $\|\cdot\|_\infty$ norm, closely paralleling the standard risk-neutral Bellman operator. This structural similarity makes the method particularly attractive for extending existing ideas from tabular RL to function approximation.*

4 Almost-sure Convergence and Finite-time Convergence Rates

We now discuss the convergence properties of the two proposed algorithms. The two-timescale recursion inherits a global contraction structure from the T-operator, enabling a relatively standard stochastic approximation analysis based on timescale separation. In contrast, the one-timescale recursion is driven by the power-law operator F. Although F is contractive in the log-sup metric, this contraction is not directly aligned with the additive update in the original x-coordinates. Consequently, standard global contraction arguments no longer apply, making both the asymptotic and finite-time analysis substantially more delicate.

4.1 Analysis of our two-timescale fixed-point method

Consider $(Q_n, g_n)_{n\geq 0}$ generated by (7). Let $C_{\max} := \frac{\|r\|_\infty}{1-\gamma}$, where $\|r\|_\infty := \max_{s,a} |r(s,a)|$.

Assumption 1 (***Controlled Markovian sampling***). *Fix $Q \in \mathbb{R}^{SA}$ with $\|Q\|_\infty \leq C_{\max}$ and let the iterates be frozen, i.e., $Q_n \equiv Q\ \forall n \geq 0$. Then the resulting $(s_n, a_n)_{n\geq 0}$ sequence is an irreducible Markov chain on $\mathcal{S} \times \mathcal{A}$ with transition kernel K_Q and unique stationary distribution π_Q. Also:*

1. *(**Uniform ergodicity**) There exist constants $C < \infty$ and $\rho \in (0,1)$ such that*

$$\sup_{y\in\mathcal{S}\times\mathcal{A}} \|K_Q^\ell(\cdot|y) - \pi_Q(\cdot)\|_{\mathrm{TV}} \leq C\rho^\ell, \qquad \forall \ell \geq 0,$$

   uniformly over all Q satisfying $\|Q\|_\infty \leq C_{\max}$.

2. *(**Continuity**) If $Q_n \to Q$ with $\|Q_n\|_\infty, \|Q\|_\infty \leq C_{\max}$, then*

$$K_{Q_n}(y, y') \to K_Q(y, y')$$

   uniformly over $(y, y') \in (\mathcal{S} \times \mathcal{A})^2$.

This is a standard assumption in the convergence analysis of stochastic approximation with controlled Markov noise, cf. Condition H2 (b) of Fort et al. [2016].

The next result establishes almost-sure convergence of the two-timescale iterates. The proof combines timescale separation with the global $\|\cdot\|_\infty$-contraction of the T-operator.

Theorem 4.1 (***Two-timescale asymptotic convergence***). *Let*

$$\max\left\{\|Q_0\|_\infty, \frac{\gamma}{\theta}\|\ln g_0\|_\infty\right\} \leq C_{\max},$$

and suppose

$$\sum_n \alpha_n = \sum_n \beta_n = \infty, \qquad \sum_n (\alpha_n^2 + \beta_n^2) < \infty, \qquad \frac{\alpha_n}{\beta_n} \to 0.$$

Further suppose Assumption 1 holds. Then

$$Q_n \xrightarrow{\text{a.s.}} Q^*, \qquad \left\| -\frac{\gamma}{\theta} \ln g_n - TQ_n \right\|_\infty \xrightarrow{\text{a.s.}} 0.$$

We next establish finite-time convergence rates in the generative-model setting, where (s_n, a_n) is sampled i.i.d. from a distribution ν satisfying $\lambda := \min_{(s,a)\in\mathcal{S}\times\mathcal{A}} \nu(s,a) > 0$. This allows us to focus on the impact of the nonlinear T-operator and the two-timescale coupling, while avoiding additional technicalities arising from Markovian noise, which can typically be handled using standard techniques [Srikant and Ying, 2019, Zhang et al., 2021].

Theorem 4.2 (***Two-timescale finite-time rates***)**.** *In the generative-model setting described above, let $\alpha_n = (n+1)^{-\alpha}$ and $\beta_n = (n+1)^{-\beta}$, where $\frac{1}{2} < \beta < \alpha < 1$. Then there exist constants $C_g, C_Q^{(1)}, C_Q^{(2)} > 0$ such that*

$$\mathbb{E}\left\|g_n - e^{-\frac{\theta}{\gamma}TQ_n}\right\|_\infty \le C_g\sqrt{\beta_n},$$

and

$$\mathbb{E}\|Q_n - Q^*\|_\infty \le C_Q^{(1)}\exp\left(-\frac{1-\gamma}{1-\alpha}n^{1-\alpha}\right) + C_Q^{(2)}\sqrt{\beta_n} = O(n^{-\beta/2}).$$

The precise constant definitions can be found in the appendix.

By choosing β close to 1, the rate can be made arbitrarily close to the optimal $O(n^{-1/2})$ rate.

4.2 Analysis of our one-timescale fixed-point method

We now state the convergence guarantee for (9). Let $\mathcal{K} := [C_\ell, C_u]^{SA}$, where

$$C_\ell := \exp\left(-\frac{\theta}{\gamma(1-\gamma)}\|r\|_\infty\right) \qquad \text{and} \qquad C_u := \exp\left(\frac{\theta}{\gamma(1-\gamma)}\|r\|_\infty\right). \tag{11}$$

Assumption 2 (***Markovian sampling***)**.** *For each frozen $x \in \mathcal{K}$, the process (s_n, a_n) is an irreducible Markov chain on $\mathcal{S}\times\mathcal{A}$ with kernel K_x and unique stationary distribution π_x. The family $\{K_x : x \in \mathcal{K}\}$ is uniformly ergodic, i.e., for some $C < \infty$ and $\rho \in (0,1)$,*

$$\sup_y \|K_x^\ell(\cdot \mid y) - \pi_x(\cdot)\|_{\mathrm{TV}} \le C\rho^\ell, \qquad \forall \ell \ge 0,\ x \in \mathcal{K}.$$

Moreover, $x \mapsto K_x$ is Lipschitz continuous on $\mathcal{K}$, uniformly over entries.

Theorem 4.3 (***One-timescale asymptotic convergence***)**.** *Let Assumption 2 hold. If $x_0 \in \mathcal{K}$ and $\alpha_n \in [0,1]$ satisfies $\sum_n \alpha_n = \infty$ and $\sum_n \alpha_n^2 < \infty$, then (x_n) generated by* (9) *satisfies*

$$x_n \to x^* \text{a.s.}$$

We also give a finite-time result for a scalar stochastic recursion associated with the x_n update in (9). This result is not a finite-time bound for the full Markovian recursion; rather, it shows that the power-law structure itself does not preclude the standard stochastic-approximation rate.

Proposition 4.4 (***Scalar rate for the one-timescale update***)**.** *Consider the scalar stochastic recursion associated with* (9)*, as defined in* (78)*, whose mean drift is induced by $F(x) = (x/x^*)^\gamma x^*$. Further suppose $x_0 \in [C_\ell, C_u]$. Define $\alpha_n = 1/(2C_1(n+1))$, where*

$$C_1 := \begin{cases} \dfrac{\underline{y} - \underline{y}^\gamma}{\underline{y} - 1}, & \underline{y} \ne 1, \\ 1-\gamma, & \underline{y} = 1, \end{cases} \qquad \text{and} \qquad \underline{y} := \frac{C_\ell}{x^*}.$$

Then, we have $\mathbb{E}\left|\dfrac{x_n}{x^*} - 1\right| = \tilde{O}(n^{-1/2})$, where $\tilde{O}$ hides logarithmic terms.

5 Outline of the Proofs

Here we discuss the main ideas underlying our structural and convergence analysis. We first establish the contraction, monotonicity, and homogeneity properties of the operators F and T, and use them to characterize an optimal stationary policy through their fixed points. We then discuss the convergence analysis of our proposed algorithms. Note that the two-timescale method leverages the global $\|\cdot\|_\infty$ contraction of T, enabling a relatively standard stochastic approximation analysis via timescale separation. In contrast, the one-timescale recursion is driven by the nonlinear power-law operator F, whose contraction holds only in the sup-log metric and is not directly compatible with additive updates in the original coordinates.

5.1 Structural properties and optimality characterization

The proof first establishes that the operators F and F_π are γ-contractions in the sup-log metric by converting multiplicative perturbations into additive logarithmic bounds. The key observation is that if two vectors are close in logarithmic scale, then their componentwise ratios are uniformly controlled, and this control is preserved under the operations defining the operators. Monotonicity and homogeneity follow directly from their definitions. Next, using Jensen's inequality, we show that $X_\pi(\lambda) \preceq F_\pi(X_\pi(\lambda))$. Repeated application of F_π, together with its monotonicity and contraction properties, then yields $X_\pi(\lambda) \preceq \bar{x}_\pi$. Finally, the greedy policy π^* satisfies $F_{\pi^*}(x^*) = F(x^*) = x^*$, implying $x^* = \bar{x}_{\pi^*}$. Since $F(x) \preceq F_\pi(x)$ for every stationary policy π, monotonicity and fixed-point convergence further give $x^* \preceq \bar{x}_\pi$, establishing the optimality of π^* for the recursive formulation.

5.2 Two-timescale analysis

The analysis of the two-timescale method is comparatively more standard due to the global $\|\cdot\|_\infty$ contraction of the operator T. The key idea is to exploit the separation between the slower (Q_n) recursion and the faster (g_n) recursion. Since

$$\frac{\alpha_n}{\beta_n} \to 0,$$

the slower iterate Q_n appears nearly static from the perspective of the faster recursion. Consequently, conditioned on a quasi-static Q_n, the faster iterate g_n tracks the equilibrium

$$g^*(Q_n) = \exp\left(-\frac{\theta}{\gamma} T Q_n\right).$$

To formalize this intuition, we define the tracking error

$$e_n := g_n - g^*(Q_n),$$

and analyze its dynamics on the faster timescale. Using the linear structure of the g-update together with standard martingale arguments, we show that

$$\|e_n\|_\infty \to 0 \qquad \text{a.s.}$$

Substituting this asymptotic relation into the slower recursion yields

$$Q_{n+1} = Q_n + \alpha_n[TQ_n - Q_n + \epsilon'_n],$$

where ϵ'_n is a logarithmic transformation of e_n. Since $\epsilon'_n \to 0$, the limiting ODE associated with (Q_n) is

$$\dot{Q}(t) = TQ(t) - Q(t).$$

Since T is a γ-contraction in the $\|\cdot\|_\infty$ norm, the above ODE admits a unique globally asymptotically stable equilibrium Q^*. Standard stochastic approximation arguments [Borkar, 2008] therefore imply

$$Q_n \to Q^* \qquad \text{a.s.}$$

For the finite-time analysis, we again exploit timescale separation. In particular, we first derive a tracking bound for the faster recursion of the form

$$\mathbb{E}\|g_n - g^*(Q_n)\|_\infty = O(\sqrt{\beta_n}),$$

and then propagate this error into the slower recursion. Combining the resulting perturbation bounds with the global contraction of T yields the finite-time convergence rates for (Q_n).

5.3 One-timescale analysis

The one-timescale recursion is significantly more delicate to analyze. Unlike the two-timescale method, the update is driven by the nonlinear power-law operator F, whose contraction holds only in the sup-log metric and is therefore not directly compatible with additive stochastic approximation updates in the original coordinates. Moreover, the map $x \mapsto x^\gamma$, $\gamma \in (0,1)$, is not globally Lipschitz near the boundary of $\mathbb{R}^{SA}_{++}$, preventing a direct application of standard global contraction arguments.

We first establish stability of the iterates by showing that the recursion remains within a compact subset

$$\mathcal{K} := [C_\ell, C_u]^{SA} \subset \mathbb{R}^{SA}_{++},$$

bounded away from the origin. This allows us to show that F is Lipschitz on a suitable neighborhood of $\mathcal{K}$ and that the limiting ODE

$$\dot{x}(t) = D_{\pi_{x(t)}}[F(x(t)) - x(t)]$$

is well posed. We then show that $\mathcal{K}$ is positively invariant under the above ODE.

Next, we establish global asymptotic stability of x^* for the above ODE. Since standard quadratic Lyapunov arguments are unavailable, we instead exploit the monotonicity and homogeneity properties of F from Proposition 2.1. In particular, homogeneity implies $F(cx) = c^\gamma F(x)$, while monotonicity gives

$$x \preceq y \implies F(x) \preceq F(y).$$

Using these properties, we compare the trajectory $x(t)$ with scaled versions of the fixed point x^*. Defining

$$m(t) := \min_i \frac{x_i(t)}{x_i^*}, \qquad \text{and} \qquad M(t) := \max_i \frac{x_i(t)}{x_i^*},$$

we obtain

$$m(t)x^* \preceq x(t) \preceq M(t)x^*.$$

Applying monotonicity and homogeneity to the left inequality yields

$$F(x(t)) \succeq F(m(t)x^*) = m(t)^\gamma F(x^*) = m(t)^\gamma x^*.$$

Using this relation, the ODE dynamics, and Danskin's theorem, we obtain the following bounds on the lower right Dini derivative of $m(t)$ and the upper right Dini derivative of $M(t)$:

$$D_+ m(t) \geq \begin{cases} \ell_{\mathcal{K}}[m(t)^\gamma - m(t)], & m(t) \leq 1, \\ u_{\mathcal{K}}[m(t)^\gamma - m(t)], & m(t) > 1, \end{cases}$$

and

$$D^+ M(t) \leq \begin{cases} u_{\mathcal{K}}[M(t)^\gamma - M(t)], & M(t) \leq 1, \\ \ell_{\mathcal{K}}[M(t)^\gamma - M(t)], & M(t) > 1, \end{cases}$$

where $0 < \ell_{\mathcal{K}} \leq u_{\mathcal{K}}$ are uniform bounds on the stationary probabilities of the state-action pairs.

We then compare these inequalities with the scalar dynamics $\dot{z}(t) = z(t)^\gamma - z(t)$, whose unique globally asymptotically stable equilibrium is 1. The comparison argument implies

$$1 \leq \liminf_{t\to\infty} m(t) \leq \limsup_{t\to\infty} M(t) \leq 1.$$

Consequently, both $m(t)$ and $M(t)$ converge to 1, which yields $x(t) \to x^*$. Standard stochastic approximation arguments [Borkar, 2008] now allow us to show $x_n \to x^*$ a.s.

We also provide a scalar finite-time analysis that isolates the effect of the power-law drift and highlights the main obstacle in extending standard stochastic approximation rate arguments to the full recursion. In particular, unlike the two-timescale method, the one-timescale recursion does not admit a directly usable global contraction inequality. Nevertheless, by carefully exploiting the structure of the scalar drift, we establish an $\tilde{O}(n^{-1/2})$ convergence rate.

6 Conclusions

Inspired by Porteus [1975], we studied a recursive exponential-utility formulation for discounted MDPs and developed two model-free, Q-learning–style algorithms for learning its optimal stationary policy. We established contraction properties of the associated Bellman operators and showed that their fixed points characterize an optimal stationary policy for the recursive formulation while providing upper bounds on direct exponential utility. We also demonstrated that the optimal stationary policies for the two criteria need not coincide, although their optimal values become close under additional assumptions. Our two algorithms operate on two timescales and one timescale, respectively, and both enjoy almost-sure convergence guarantees under Markovian sampling. We further established finite-time convergence rates for the two-timescale algorithm under i.i.d. sampling and for a scalar recursion associated with the one-timescale method. Extending the latter rate guarantees to general MDPs and developing model-free methods that directly optimize exponential utility beyond the stationary policy class remain interesting directions for future research.

Acknowledgments and Disclosure of Funding

We thank the anonymous reviewers for their valuable comments and suggestions, which helped improve the presentation of this work. Gugan Thoppe acknowledges support from the ANRF ARG Grant (ANRF/ARG/2025/011952/ENS), the Walmart Centre for Tech Excellence at IISc, the Indo-French Centre for the Promotion of Advanced Research (CEFIPRA; Project 7102-1), the Kotak-IISc AIML Centre (KIAC), and the Brain Computation and Data Science (BCD) Centre at IISc. L. A. Prashanth acknowledges support from the A. Raghunathan Center for Theoretical Computer Science and partial support from ANRF under Grant ANRF/ARG/2025/005386/ENS.

References

J. Aubin, A. M. Bayen, and P. Saint-Pierre. *Viability theory: new directions*. Springer Science & Business Media, 2011.

A. Basu, T. Bhattacharyya, and V. S. Borkar. A learning algorithm for risk-sensitive cost. *Mathematics of Operations Research*, 33(4):880–898, 2008.

N. Bäuerle and A. Glauner. Markov decision processes with recursive risk measures. *European Journal of Operational Research*, 296(3):953–966, 2022.

N. Bäuerle and A. Jaśkiewicz. Markov decision processes with risk-sensitive criteria: an overview. *Mathematical Methods of Operations Research*, 99(1):141–178, 2024.

N. Bäuerle and U. Rieder. More risk-sensitive Markov decision processes. *Mathematics of Operations Research*, 39(1):105–120, 2014.

N. Bäuerle, M. Pitera, and Ł. Stettner. Policy stability and ultimate stationarity in discounted risk-sensitive stochastic control. *arXiv preprint arXiv:2601.06998*, 2026.

D. P. Bertsekas and J. N. Tsitsiklis. *Neuro-Dynamic Programming*. Athena Scientific, 1996.

V. S. Borkar. A sensitivity formula for risk-sensitive cost and the actor–critic algorithm. *Systems & Control Letters*, 44(5):339–346, 2001.

V. S. Borkar. Q-learning for risk-sensitive control. *Mathematics of operations research*, 27(2): 294–311, 2002.

V. S. Borkar. *Stochastic Approximation: A Dynamical Systems Viewpoint*. Cambridge University Press, 2008.

V. S. Borkar. Learning algorithms for risk-sensitive control. In *Proceedings of the 19th International Symposium on Mathematical Theory of Networks and Systems–MTNS*, volume 5, 2010.

K. Chung and M. Sobel. Discounted MDP's: Distribution functions and exponential utility maximization. *SIAM journal on control and optimization*, 25(1):49–62, 1987.

J. M. Danskin. The theory of max-min, with applications. *SIAM Journal on Applied Mathematics*, 14 (4):641–664, 1966.

A. Davydov. *Contraction Theory in Control, Learning, and Optimization*. PhD thesis, University of California Santa Barbara, 2025.

Y. Fei, Z. Yang, Y. Chen, Z. Wang, and Q. Xie. Risk-sensitive reinforcement learning: Near-optimal risk-sample tradeoff in regret. *Advances in Neural Information Processing Systems*, 33: 22384–22395, 2020.

Y. Fei, Z. Yang, Y. Chen, and Z. Wang. Exponential Bellman equation and improved regret bounds for risk-sensitive reinforcement learning. *Advances in neural information processing systems*, 34: 20436–20446, 2021a.

Y. Fei, Z. Yang, and Z. Wang. Risk-sensitive reinforcement learning with function approximation: A debiasing approach. In *International Conference on Machine Learning*, pages 3198–3207. PMLR, 2021b.

G. Fort, E. Moulines, A. Schreck, and M. Vihola. Convergence of markovian stochastic approximation with discontinuous dynamics. *SIAM Journal on Control and Optimization*, 54(2):866–893, 2016.

A. Granados and J. Pacheco. Risk-sensitive exponential actor critic. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 21343–21351, 2026.

Jia Lin Hau, Marek Petrik, and Mohammad Ghavamzadeh. Entropic risk optimization in discounted mdps. In *International Conference on Artificial Intelligence and Statistics*, pages 47–76. PMLR, 2023.

S. C. Jaquette. A utility criterion for Markov decision processes. *Management Science*, 23(1):43–49, 1976.

D. R. Jiang and W. B. Powell. Risk-averse approximate dynamic programming with quantile-based risk measures. *Mathematics of Operations Research*, 43(2):554–579, 2017.

Y. Jiang, J. Huang, Y. Yuan, X. Mao, Y. Yue, Q. Zhao, and L. Yan. Risk-sensitive RL for alleviating exploration dilemmas in large language models. *arXiv preprint arXiv:2509.24261*, 2025.

H. K. Khalil. *Nonlinear systems*, volume 3. Prentice Hall, 2002.

U. Kumar, S. P. Bhat, V. Kavitha, and N. Hemachandra. Ultimately stationary policies to approximate risk-sensitive discounted mdps. In *Proceedings of the 12th EAI International Conference on Performance Evaluation Methodologies and Tools*, pages 63–70, 2019.

H. Liang and Z. Luo. Bridging distributional and risk-sensitive reinforcement learning with provable regret bounds. *Journal of Machine Learning Research*, 25(221):1–56, 2024.

J. Markowitz, R. W. Gardner, A. Llorens, R. Arora, and I.J. Wang. A risk-sensitive approach to policy optimization. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(12):15019–15027, Jun. 2023.

O. Mihatsch and R. Neuneier. Risk-sensitive reinforcement learning. *Machine Learning*, 49(2): 267–290, 2002.

M. Moharrami, Y. Murthy, A. Roy, and R. Srikant. A policy gradient algorithm for the risk-sensitive exponential cost MDP, 2022.

O. Mortensen and M. S. Talebi. Recursive entropic risk optimization in discounted mdps: Sample complexity bounds with a generative model. *Transactions on Machine Learning Research*, 2026.

Y. Murthy, M. Moharrami, and R. Srikant. Modified policy iteration for exponential cost risk sensitive mdps. In Nikolai Matni, Manfred Morari, and George J. Pappas, editors, *Proceedings of The 5th Annual Learning for Dynamics and Control Conference*, volume 211 of *Proceedings of Machine Learning Research*, pages 395–406. PMLR, 15–16 Jun 2023.

E. Noorani, C. N. Mavridis, and J. S. Baras. Risk-sensitive reinforcement learning with exponential criteria. *IEEE Transactions on Cybernetics*, 55(8):3774–3787, 2025. doi: 10.1109/TCYB.2025.3575240.

E. L. Porteus. On the optimality of structured policies in countable stage decision processes. *Management Science*, 22(2):148–157, 1975.

L. A. Prashanth. Policy gradients for CVaR-constrained MDPs. In *Algorithmic Learning Theory (ALT)*, pages 155–169, 2014.

L. A. Prashanth and M. C. Fu. Risk-sensitive reinforcement learning via policy gradient search. *Foundations and Trends® in Machine Learning*, 15(5):537–693, 2022. ISSN 1935-8237.

L.A. Prashanth and Mohammad Ghavamzadeh. Variance-constrained actor-critic algorithms for discounted and average reward mdps. *Machine Learning*, 105:367–417, 2016.

M. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, 1994.

R. Srikant and L. Ying. Finite-Time Error Bounds For Linear Stochastic Approximation and TD Learning. In *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pages 2803–2830. PMLR, Jun 2019.

R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA, 2nd ed. edition, 2018.

A. Tamar, D. D. Castro, and S. Mannor. Policy gradients with variance related risk criteria. In *Proceedings of the Twenty-Ninth International Conference on Machine Learning*, pages 387–396, 2012.

A. Tamar, Y. Chow, M. Ghavamzadeh, and S. Mannor. Sequential decision making with coherent risk. *IEEE Transactions on Automatic Control*, 62(7):3323–3338, 2016.

G. Thoppe, Prashanth L. A., and S. P. Bhat. Risk estimation in a Markov cost process: lower and upper bounds. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pages 48124–48138, 2024.

S. Zhang, Z. Zhang, and S. T. Maguluri. Finite sample analysis of average-reward TD learning and Q-learning. *Advances in Neural Information Processing Systems*, 34:1230–1242, 2021.

A Structural Properties of F and F_π

A.1 Proof of Proposition 2.1

Proof. We first prove statements 1 and 2. Fix $x_1, x_2 \in \mathbb{R}^{SA}_{++}$ and let $\delta = d(x_1, x_2)$. By definition,

$$e^{-\delta} x_2(s,a) \le x_1(s,a) \le e^{\delta} x_2(s,a), \qquad \forall (s,a) \in \mathcal{S} \times \mathcal{A}.$$

Taking the minimum over actions, raising to the power γ, and using the definition of F gives

$$e^{-\gamma\delta} F(x_2) \preceq F(x_1) \preceq e^{\gamma\delta} F(x_2).$$

Taking logarithms and the sup norm yields

$$d(F(x_1), F(x_2)) \le \gamma\delta = \gamma d(x_1, x_2),$$

proving statement 1. Similarly, raising the coordinatewise bounds to the power γ and taking the weighted sums appearing in F_π gives

$$e^{-\gamma\delta} F_\pi(x_2) \preceq F_\pi(x_1) \preceq e^{\gamma\delta} F_\pi(x_2).$$

The same argument proves statement 2.

For statement 3, note that both F and F_π are monotone because their dependence on x involves only order-preserving operations.

To prove statement 4, fix $x \in \mathbb{R}^{SA}_{++}$. Since a minimum does not exceed a weighted average, $F(x) \preceq F_\pi(x)$ for every $\pi \in \Pi_{\mathrm{st}}$. Now choose a deterministic stationary policy π_x such that

$$\pi_x(s) \in \arg\min_{a \in \mathcal{A}} x(s,a), \qquad \forall s \in \mathcal{S}.$$

By construction, $F_{\pi_x}(x) = F(x)$, establishing statement 4.

Finally, statement 5 follows directly from the definition of F. For every $c > 0$ and $x \in \mathbb{R}^{SA}_{++}$,

$$F(cx) = c^\gamma F(x).$$

This completes the proof. □

A.2 Proof of Theorem 2.2

Proof. Fix $\pi \in \Pi_{\mathrm{st}}$ and $(s,a) \in \mathcal{S} \times \mathcal{A}$. Let

$$Z_1 = \exp\left(-\lambda \sum_{t=1}^{\infty} \gamma^{t-1} r(s_t, a_t)\right).$$

By the definition of direct exponential utility,

$$X_\pi(s,a;\lambda) = e^{-\lambda r(s,a)} \mathbb{E}^\pi_{s,a}[Z_1^\gamma].$$

Since $z \mapsto z^\gamma$ is concave, conditional Jensen's inequality gives

$$\begin{aligned} X_\pi(s,a;\lambda) &\le e^{-\lambda r(s,a)} \mathbb{E}^\pi_{s,a}\left[\left(\mathbb{E}^\pi_{s,a}[Z_1 \mid s_1, a_1]\right)^\gamma\right] \\ &= e^{-\lambda r(s,a)} \mathbb{E}^\pi_{s,a}\left[X_\pi(s_1, a_1; \lambda)^\gamma\right] \\ &= F_\pi(X_\pi(\lambda))(s,a), \end{aligned}$$

where the second equality follows from the Markov property and stationarity of π.

Since (s,a) was arbitrary, $X_\pi(\lambda) \preceq F_\pi(X_\pi(\lambda))$. By the monotonicity of F_π, this implies

$$X_\pi(\lambda) \preceq F^k_\pi(X_\pi(\lambda)), \qquad k \ge 1.$$

Proposition 2.1 guarantees that F_π is a contraction. Hence, $F^k_\pi(X_\pi(\lambda)) \to \bar{x}_\pi$ as $k \to \infty$, giving

$$X_\pi(\lambda) \preceq \bar{x}_\pi.$$

Next, by the definition of the greedy policy π^*,

$$F_{\pi^*}(x^*) = F(x^*) = x^*.$$

Since F_{π^*} has a unique fixed point, we obtain $x^* = \bar{x}_{\pi^*}$.

Finally, statement 4 of Proposition 2.1 gives

$$x^* = F(x^*) \preceq F_\pi(x^*).$$

Using the monotonicity of F_π again,

$$x^* \preceq F_\pi^k(x^*), \qquad k \geq 1.$$

Letting $k \to \infty$ and using the contraction property yields $x^* \preceq \bar{x}_\pi$. Therefore,

$$X_\pi(\lambda) \preceq \bar{x}_\pi, \qquad x^* = \bar{x}_{\pi^*} \preceq \bar{x}_\pi,$$

as required. □

A.3 Counterexample: A stationary policy minimizing $\bar{x}_{\pi,\gamma}$ need not minimize $X_{\pi,\gamma}$

We now show that optimizing the recursively defined criterion $\bar{x}_{\pi,\gamma}$ need not yield a stationary policy that optimizes the direct exponential criterion $X_{\pi,\gamma}$. This mismatch can occur even in a finite MDP for which every stationary policy induces an irreducible Markov chain.

We first state the main result. Its proof is divided into several lemmas, each addressing a separate part of the construction.

Theorem A.1 (Mismatch between the two exponential criteria)**.** *There exist a finite MDP, two deterministic stationary policies π_a and π_b, and a constant $\gamma_0 \in (0,1)$ such that the following statements hold.*

1. *For every $\gamma \in [0,1)$,*
$$\bar{x}_{\pi_b,\gamma} \preceq \bar{x}_{\pi,\gamma}, \qquad \forall \pi \in \Pi_{\mathrm{st}}.$$
Thus, π_b minimizes $\bar{x}_{\pi,\gamma}$ componentwise over the class of stationary policies.

2. *For every $\gamma \in (\gamma_0, 1)$,*
$$X_{\pi_a,\gamma} \prec X_{\pi_b,\gamma}.$$
In fact, for every admissible state-action pair $(\tilde{s}, \tilde{a})$,
$$\frac{X_{\pi_a,\gamma}(\tilde{s},\tilde{a})}{X_{\pi_b,\gamma}(\tilde{s},\tilde{a})} \longrightarrow 0 \qquad \text{as } \gamma \uparrow 1.$$
Thus, π_b becomes arbitrarily worse than π_a under the direct exponential criterion.

Consequently, a stationary policy minimizing $\bar{x}_{\pi,\gamma}$ need not minimize $X_{\pi,\gamma}$.

Construction of the MDP. Fix the risk parameter $\widehat{\theta} = 1$, and consider the state space

$$\mathcal{S} = \{s, s_h, s_\ell\},$$

where s is the decision state, s_h is the high-reward state, and s_ℓ is the low-reward state. At state s, the available actions are a and b. At s_h and s_ℓ, only action a is available.

The rewards at the decision state are

$$r(s,a) = 0, \qquad r(s,b) = \frac{1}{2},$$

while the rewards at the remaining states are

$$r(s_h,a) = 4, \qquad r(s_\ell,a) = 0.$$

From s_h and s_ℓ, the process returns deterministically to s:

$$P(s \mid s_h, a) = P(s \mid s_\ell, a) = 1.$$

At state s, action a induces the transition probabilities

$$P(s_h \mid s, a) = \frac{1}{2}, \qquad P(s_\ell \mid s, a) = \frac{1}{2},$$

whereas action b induces

$$P(s_h \mid s, b) = \frac{1}{4}, \qquad P(s_\ell \mid s, b) = \frac{3}{4}.$$

Let π_a denote the deterministic stationary policy that chooses action a at s, and let π_b denote the deterministic stationary policy that chooses action b at s. At s_h and s_ℓ, both policies choose the only available action a.

The set of admissible state-action pairs is

$$\mathcal{S}_{\mathrm{adm}} := \{(s, a), (s, b), (s_h, a), (s_\ell, a)\}.$$

We begin with a structural property of the example.

Lemma A.2 (Irreducibility)**.** *Every stationary policy in the preceding MDP induces an irreducible Markov chain.*

Proof. Under either action at s, both s_h and s_ℓ are reached with strictly positive probability. The same is therefore true under every randomized stationary policy. Moreover, both s_h and s_ℓ return deterministically to s. Hence, every state communicates with every other state under every stationary policy. □

We next determine the policy that minimizes the recursively defined criterion $\bar{x}_{\pi,\gamma}$.

Lemma A.3 (Optimality of π_b for the recursive criterion)**.** *For every $\gamma \in [0, 1)$, the policy π_b is greedy with respect to the unique fixed point x^*_γ of the optimality operator F_γ. Consequently,*

$$\bar{x}_{\pi_b,\gamma} = x^*_\gamma \preceq \bar{x}_{\pi,\gamma}, \qquad \forall \pi \in \Pi_{\mathrm{st}}.$$

Proof. Fix $\gamma \in [0, 1)$, and define

$$m^*_\gamma := \min\left\{x^*_\gamma(s, a), x^*_\gamma(s, b)\right\}.$$

Since action a is the only available action at s_h and s_ℓ, the fixed-point relation

$$F_\gamma(x^*_\gamma) = x^*_\gamma$$

gives

$$x^*_\gamma(s_h, a) = e^{-4}(m^*_\gamma)^\gamma$$

and

$$x^*_\gamma(s_\ell, a) = (m^*_\gamma)^\gamma.$$

It follows that

$$\begin{aligned} x^*_\gamma(s, a) &= \frac{1}{2}\left[x^*_\gamma(s_h, a)\right]^\gamma + \frac{1}{2}\left[x^*_\gamma(s_\ell, a)\right]^\gamma \\ &= \frac{1 + e^{-4\gamma}}{2}(m^*_\gamma)^{\gamma^2}. \end{aligned}$$

Similarly,

$$\begin{aligned} x^*_\gamma(s, b) &= e^{-1/2}\left(\frac{1}{4}\left[x^*_\gamma(s_h, a)\right]^\gamma + \frac{3}{4}\left[x^*_\gamma(s_\ell, a)\right]^\gamma\right) \\ &= e^{-1/2}\frac{3 + e^{-4\gamma}}{4}(m^*_\gamma)^{\gamma^2}. \end{aligned}$$

Therefore,

$$\frac{x^*_\gamma(s, b)}{x^*_\gamma(s, a)} = e^{-1/2}\frac{3 + e^{-4\gamma}}{2(1 + e^{-4\gamma})}.$$

For every $\gamma \in [0, 1)$,

$$\frac{3 + e^{-4\gamma}}{2(1 + e^{-4\gamma})} < \frac{3}{2} < e^{1/2}.$$

Hence,

$$x^*_\gamma(s, b) < x^*_\gamma(s, a).$$

Thus, action b is the unique greedy action at s, and π_b is greedy with respect to x^*_γ. It follows that

$$F_{\pi_b,\gamma}(x^*_\gamma) = F_\gamma(x^*_\gamma) = x^*_\gamma.$$

Since $F_{\pi_b,\gamma}$ has a unique fixed point,

$$\bar{x}_{\pi_b,\gamma} = x^*_\gamma.$$

Finally, Theorem 2.2 gives

$$x^*_\gamma \preceq \bar{x}_{\pi,\gamma}, \qquad \forall \pi \in \Pi_{\mathrm{st}},$$

which proves the result. □

Lemma A.3 establishes the first part of Theorem A.1. We now analyze the direct exponential criterion. For each $i \in \{a, b\}$, define

$$c_i := r(s, i),$$

so that

$$c_a = 0, \qquad c_b = \frac{1}{2}.$$

Let p_i denote the probability that action i at s moves the process to s_h. Thus,

$$p_a = \frac{1}{2}, \qquad p_b = \frac{1}{4}.$$

Let R_i denote the reward received at the next state-action pair after taking action i at state s. Since action a is the only available action at both s_h and s_ℓ, we have

$$R_i = \begin{cases} r(s_h, a) = 4, & \text{with probability } p_i, \\ r(s_\ell, a) = 0, & \text{with probability } 1 - p_i. \end{cases}$$

For $u \in [0, 1]$, define its Laplace transform by

$$M_i(u) := \mathbb{E}\left[e^{-uR_i}\right] = p_i e^{-4u} + (1 - p_i).$$

In particular,

$$M_a(u) = \frac{1 + e^{-4u}}{2}, \qquad M_b(u) = \frac{3 + e^{-4u}}{4}.$$

Recall that, for a stationary policy π and an admissible state-action pair $(\tilde{s}, \tilde{a})$,

$$X_{\pi,\gamma}(\tilde{s}, \tilde{a}) := \mathbb{E}\left[\exp\left(-\sum_{t=0}^{\infty} \gamma^t r(S_t, A_t)\right) \,\middle|\, S_0 = \tilde{s},\ A_0 = \tilde{a}\right].$$

The prescribed action $\tilde{a}$ is used at time zero, and the policy π is followed from time one onward.

The following lemma gives exact product representations for all components of $X_{\pi_i,\gamma}$.

Lemma A.4 (Exact representation of the direct criterion)**.** *For every $i, j \in \{a, b\}$,*

$$X_{\pi_i,\gamma}(s, j) = e^{-c_j} M_j(\gamma) \exp\left(-\frac{\gamma^2 c_i}{1 - \gamma^2}\right) \prod_{k=0}^{\infty} M_i\left(\gamma^{2k+3}\right). \tag{12}$$

Moreover,

$$X_{\pi_i,\gamma}(s_\ell, a) = \exp\left(-\frac{\gamma c_i}{1 - \gamma^2}\right) \prod_{k=0}^{\infty} M_i\left(\gamma^{2k+2}\right), \tag{13}$$

and

$$X_{\pi_i,\gamma}(s_h, a) = e^{-4} X_{\pi_i,\gamma}(s_\ell, a). \tag{14}$$

Proof. Fix $i, j \in \{a, b\}$. Starting from (s, j) and following π_i from time one onward, the reward sequence has the form

$$c_j,\ R^{(j)},\ c_i,\ R_0^{(i)},\ c_i,\ R_1^{(i)}, \ldots,$$

where $R^{(j)}$ has the same distribution as R_j, and $(R_k^{(i)})_{k\geq 0}$ are independent copies of R_i. Furthermore, $R^{(j)}$ is independent of $(R_k^{(i)})_{k\geq 0}$. Therefore,

$$\begin{aligned} X_{\pi_i,\gamma}(s, j) &= \mathbb{E}\left[\exp\left(-c_j - \gamma R^{(j)} - \sum_{k=0}^{\infty} \gamma^{2k+2} c_i - \sum_{k=0}^{\infty} \gamma^{2k+3} R_k^{(i)}\right)\right] \\ &= e^{-c_j} M_j(\gamma) \exp\left(-c_i \sum_{k=0}^{\infty} \gamma^{2k+2}\right) \prod_{k=0}^{\infty} M_i\left(\gamma^{2k+3}\right) \\ &= e^{-c_j} M_j(\gamma) \exp\left(-\frac{\gamma^2 c_i}{1-\gamma^2}\right) \prod_{k=0}^{\infty} M_i\left(\gamma^{2k+3}\right). \end{aligned}$$

Starting from (s_ℓ, a) and following π_i thereafter, the reward sequence is

$$0,\ c_i,\ R_0^{(i)},\ c_i,\ R_1^{(i)}, \ldots.$$

Consequently,

$$\begin{aligned} X_{\pi_i,\gamma}(s_\ell, a) &= \exp\left(-c_i \sum_{k=0}^{\infty} \gamma^{2k+1}\right) \prod_{k=0}^{\infty} M_i\left(\gamma^{2k+2}\right) \\ &= \exp\left(-\frac{\gamma c_i}{1-\gamma^2}\right) \prod_{k=0}^{\infty} M_i\left(\gamma^{2k+2}\right). \end{aligned}$$

Finally, starting from (s_h, a) produces the same continuation as starting from (s_ℓ, a), except that the initial reward is 4 instead of 0. Hence,

$$X_{\pi_i,\gamma}(s_h, a) = e^{-4} X_{\pi_i,\gamma}(s_\ell, a).$$

This completes the proof. □

The infinite products in Lemma A.4 can be evaluated asymptotically using the following shifted Riemann-sum result.

Lemma A.5 (Shifted Riemann-sum limit)**.** *Let $M : [0, 1] \to (0, \infty)$ be continuous and suppose that there exists a constant $C < \infty$ such that*

$$|\log M(u)| \leq Cu, \qquad u \in [0, 1].$$

Then, for every fixed positive integer κ,

$$\lim_{\gamma\uparrow 1}(1-\gamma) \sum_{k=0}^{\infty} \log M\left(\gamma^{2k+\kappa}\right) = \frac{1}{2}\int_0^1 \frac{\log M(u)}{u}\, du.$$

Proof. Define

$$g(t) := \log M(e^{-t}), \qquad t \geq 0.$$

The assumed bound implies

$$|g(t)| \leq Ce^{-t}, \qquad t \geq 0.$$

Thus, g is continuous and integrable on $[0, \infty)$.

Set

$$h := -\log \gamma.$$

As $\gamma \uparrow 1$, we have $h \downarrow 0$, and

$$\gamma^{2k+\kappa} = e^{-(2k+\kappa)h}.$$

Therefore,

$$2h\sum_{k=0}^{\infty}\log M\left(\gamma^{2k+\kappa}\right)=2h\sum_{k=0}^{\infty}g\left((2k+\kappa)h\right).$$

The expression on the right is a shifted Riemann sum with mesh size $2h$. Its evaluation points are

$$\kappa h,\ (\kappa+2)h,\ (\kappa+4)h,\ldots.$$

Since κ is fixed, the initial shift κh converges to zero as $h\downarrow 0$. The continuity of g on compact intervals, together with the bound

$$|g(t)|\le Ce^{-t},$$

for controlling the tail, gives

$$2h\sum_{k=0}^{\infty}g\left((2k+\kappa)h\right)\longrightarrow\int_0^{\infty}g(t)\,dt.$$

Using the change of variables $u=e^{-t}$, we obtain

$$\int_0^{\infty}g(t)\,dt=\int_0^1\frac{\log M(u)}{u}\,du.$$

Finally,

$$\frac{1-\gamma}{h}=\frac{1-\gamma}{-\log\gamma}\longrightarrow 1.$$

It follows that

$$\begin{aligned}&(1-\gamma)\sum_{k=0}^{\infty}\log M\left(\gamma^{2k+\kappa}\right)\\&\quad=\frac{1-\gamma}{2h}\left[2h\sum_{k=0}^{\infty}g\left((2k+\kappa)h\right)\right]\\&\quad\longrightarrow\frac{1}{2}\int_0^1\frac{\log M(u)}{u}\,du,\end{aligned}$$

as desired. □

We now apply Lemma A.5 to the functions M_a and M_b. For every $i\in\{a,b\}$ and $u\in[0,1]$,

$$e^{-4u}\le M_i(u)\le 1.$$

Consequently,

$$|\log M_i(u)|\le 4u,$$

and the conditions of Lemma A.5 are satisfied.

The next lemma shows that all components associated with a fixed policy have the same leading-order asymptotic rate.

Lemma A.6 (Common asymptotic rate)**.** *For $i\in\{a,b\}$, define*

$$\Lambda_i:=\frac{1}{2}\left[-c_i+\int_0^1\frac{\log M_i(u)}{u}\,du\right].$$

Then, for every admissible state-action pair $(\tilde{s},\tilde{a})\in\mathcal{S}_{\mathrm{adm}}$,

$$\lim_{\gamma\uparrow 1}(1-\gamma)\log X_{\pi_i,\gamma}(\tilde{s},\tilde{a})=\Lambda_i.$$

Proof. Consider first an initial pair (s,j), where $j\in\{a,b\}$. By (12),

$$(1-\gamma)\log X_{\pi_i,\gamma}(s,j)$$

$$= (1-\gamma)\left[-c_j + \log M_j(\gamma)\right] - (1-\gamma)\frac{\gamma^2 c_i}{1-\gamma^2} + (1-\gamma)\sum_{k=0}^{\infty} \log M_i\left(\gamma^{2k+3}\right).$$

The first term converges to zero because $-c_j + \log M_j(\gamma)$ remains bounded as $\gamma \uparrow 1$. Moreover,

$$(1-\gamma)\frac{\gamma^2}{1-\gamma^2} = \frac{\gamma^2}{1+\gamma} \longrightarrow \frac{1}{2}.$$

Applying Lemma A.5 with $\kappa = 3$ gives

$$\begin{aligned}\lim_{\gamma\uparrow 1}(1-\gamma)\log X_{\pi_i,\gamma}(s,j) &= -\frac{c_i}{2} + \frac{1}{2}\int_0^1 \frac{\log M_i(u)}{u}\,du \\ &= \Lambda_i.\end{aligned}$$

Similarly, by (13),

$$\begin{aligned}&(1-\gamma)\log X_{\pi_i,\gamma}(s_\ell, a) \\ &= -(1-\gamma)\frac{\gamma c_i}{1-\gamma^2} + (1-\gamma)\sum_{k=0}^{\infty}\log M_i\left(\gamma^{2k+2}\right).\end{aligned}$$

Since

$$(1-\gamma)\frac{\gamma}{1-\gamma^2} = \frac{\gamma}{1+\gamma} \longrightarrow \frac{1}{2},$$

Lemma A.5, applied with $\kappa = 2$, yields

$$\lim_{\gamma\uparrow 1}(1-\gamma)\log X_{\pi_i,\gamma}(s_\ell, a) = \Lambda_i.$$

Finally, by (14),

$$(1-\gamma)\log X_{\pi_i,\gamma}(s_h, a) = -4(1-\gamma) + (1-\gamma)\log X_{\pi_i,\gamma}(s_\ell, a).$$

The first term converges to zero, and hence the limit is again Λ_i. □

Lemma A.6 shows that differences arising from the initial state-action pair do not affect the leading-order behavior on the $(1-\gamma)\log$ scale. It remains to compare the rates Λ_a and Λ_b.

We first establish a pointwise estimate that will be used to bound the difference between these rates.

Lemma A.7 (A pointwise lower bound)**.** *Define*

$$y(u) := \frac{1-e^{-4u}}{3+e^{-4u}}, \qquad u \in [0,1].$$

Then

$$y(u) \geq \frac{u}{1+\frac{9}{4}u}, \qquad u \in [0,1].$$

Proof. The desired inequality is equivalent to

$$e^{-4u} \leq \frac{4-3u}{4+13u}.$$

To verify this inequality, define

$$\phi(u) := 4u + \log\left(\frac{4-3u}{4+13u}\right).$$

We have $\phi(0) = 0$, and a direct calculation gives

$$\phi'(u) = \frac{4u(40-39u)}{(4-3u)(4+13u)} \geq 0, \qquad u \in [0,1].$$

Therefore, $\phi(u) \geq 0$ for every $u \in [0,1]$. Equivalently,

$$4u + \log\left(\frac{4-3u}{4+13u}\right) \geq 0,$$

which gives

$$e^{-4u} \leq \frac{4-3u}{4+13u}.$$

This proves the claimed lower bound. □

We can now establish the strict comparison between the two asymptotic rates.

Lemma A.8 (Strict comparison of asymptotic rates)**.** *The rates in Lemma A.6 satisfy*

$$\Lambda_a < \Lambda_b.$$

Proof. By the definitions of Λ_a and Λ_b,

$$\begin{aligned}\Lambda_a - \Lambda_b &= \frac{1}{2}\left[c_b - c_a + \int_0^1 \frac{\log M_a(u) - \log M_b(u)}{u}\, du\right] \\ &= \frac{1}{2}\left[\frac{1}{2} + \int_0^1 \frac{1}{u}\log\left(\frac{2(1+e^{-4u})}{3+e^{-4u}}\right) du\right].\end{aligned}$$

Using the function y from Lemma A.7, we have

$$\frac{2(1+e^{-4u})}{3+e^{-4u}} = 1 - y(u).$$

Since $\log(1-z) \leq -z$ for every $z \in [0,1)$,

$$\int_0^1 \frac{1}{u}\log\left(\frac{2(1+e^{-4u})}{3+e^{-4u}}\right) du \leq -\int_0^1 \frac{y(u)}{u}\, du. \tag{15}$$

Lemma A.7 gives

$$\begin{aligned}\int_0^1 \frac{y(u)}{u}\, du &\geq \int_0^1 \frac{1}{1+\frac{9}{4}u}\, du \\ &= \frac{4}{9}\log\left(\frac{13}{4}\right) \\ &> \frac{1}{2}.\end{aligned}$$

Combining this inequality with (15) yields

$$\int_0^1 \frac{1}{u}\log\left(\frac{2(1+e^{-4u})}{3+e^{-4u}}\right) du < -\frac{1}{2}.$$

Consequently,

$$\Lambda_a - \Lambda_b < \frac{1}{2}\left[\frac{1}{2} - \frac{1}{2}\right] = 0.$$

Thus, $\Lambda_a < \Lambda_b$. □

We are now ready to prove the main result.

Proof of Theorem A.1. Lemma A.3 shows that, for every $\gamma \in [0,1)$,

$$\bar{x}_{\pi_b,\gamma} \preceq \bar{x}_{\pi,\gamma}, \qquad \forall \pi \in \Pi_{\mathrm{st}}.$$

Thus, π_b minimizes the recursively defined criterion $\bar{x}_{\pi,\gamma}$ componentwise over all stationary policies.

Now fix an admissible state-action pair $(\tilde{s}, \tilde{a}) \in S_{\mathrm{adm}}$. By Lemmas A.6 and A.8,

$$\lim_{\gamma \uparrow 1} (1-\gamma) \log \left(\frac{X_{\pi_a,\gamma}(\tilde{s},\tilde{a})}{X_{\pi_b,\gamma}(\tilde{s},\tilde{a})} \right) = \Lambda_a - \Lambda_b < 0.$$

Therefore, there exists $\gamma_{\tilde{s},\tilde{a}} \in (0,1)$ such that

$$X_{\pi_a,\gamma}(\tilde{s},\tilde{a}) < X_{\pi_b,\gamma}(\tilde{s},\tilde{a})$$

for every $\gamma \in (\gamma_{\tilde{s},\tilde{a}}, 1)$.

Since $\mathcal{S}_{\mathrm{adm}}$ is finite, define

$$\gamma_0 := \max_{(\tilde{s},\tilde{a}) \in \mathcal{S}_{\mathrm{adm}}} \gamma_{\tilde{s},\tilde{a}}.$$

Then $\gamma_0 < 1$, and, for every $\gamma \in (\gamma_0, 1)$,

$$X_{\pi_a,\gamma}(\tilde{s},\tilde{a}) < X_{\pi_b,\gamma}(\tilde{s},\tilde{a}), \qquad \forall (\tilde{s},\tilde{a}) \in \mathcal{S}_{\mathrm{adm}}.$$

Equivalently,

$$X_{\pi_a,\gamma} \prec X_{\pi_b,\gamma}.$$

Thus, although π_b minimizes $\bar{x}_{\pi,\gamma}$ componentwise over the class of stationary policies, it is strictly dominated by the stationary policy π_a under the direct exponential criterion whenever γ is sufficiently close to one. In particular, π_b does not minimize $X_{\pi,\gamma}$ over Π_{st}.

Finally, Lemma A.2 shows that this mismatch occurs even though every stationary policy induces an irreducible Markov chain. □

A.4 Asymptotic Agreement in Uniformly Absorbing MDPs

The preceding example shows that, in a general MDP, minimizing $\bar{x}_{\pi,\gamma}$ need not minimize $X_{\pi,\gamma}$, even when γ is arbitrarily close to one. We now show that, in a finite MDP with an absorbing set and nonnegative rewards, $\bar{x}_{\pi,\gamma}$ and $X_{\pi,\gamma}$ become uniformly close as $\gamma \uparrow 1$.

Let $\mathcal{S}_{\mathrm{abs}} \subseteq \mathcal{S}$ be a nonempty absorbing set. We assume that every action available at a state in $\mathcal{S}_{\mathrm{abs}}$ yields zero reward and keeps the process within $\mathcal{S}_{\mathrm{abs}}$. Thus, once the process enters $\mathcal{S}_{\mathrm{abs}}$, all subsequent rewards are zero.

Let

$$\mathcal{Z} := \{(s,a) : s \notin \mathcal{S}_{\mathrm{abs}},\ a \in \mathcal{A}(s)\}$$

denote the set of nonabsorbing state-action pairs. For $\pi \in \Pi_{\mathrm{st}}$, define the policy-induced state-action transition matrix

$$P_\pi\big((s,a),(s',a')\big) := P(s' \mid s,a)\pi(a' \mid s'),$$

and let Z_π denote the restriction of P_π to $\mathcal{Z} \times \mathcal{Z}$.

We impose the following standard assumption.

Assumption 3. *The rewards are nonnegative, and every stationary policy $\pi \in \Pi_{\mathrm{st}}$ is proper; that is, under π, the absorbing set $\mathcal{S}_{\mathrm{abs}}$ is reached almost surely from every initial state.*

The following lemma records the uniform-transience consequence of Assumption 3.

Lemma A.9. *Under Assumption 3, there exists a constant $K < \infty$ such that*

$$\sup_{\pi \in \Pi_{\mathrm{st}}} \left\| \sum_{n=1}^{\infty} Z_\pi^n \right\|_\infty \le K.$$

Proof. Fix $\pi \in \Pi_{\mathrm{st}}$. Since π is proper, every state-action pair in $\mathcal{Z}$ is transient under the state-action Markov chain induced by π. Hence,

$$\rho(Z_\pi) < 1,$$

and therefore

$$\sum_{n=1}^{\infty} Z_\pi^n = (I - Z_\pi)^{-1} - I.$$

The set Π_{st} of stationary policies is compact, and the map $\pi \mapsto Z_\pi$ is continuous. Moreover, $I - Z_\pi$ is nonsingular for every $\pi \in \Pi_{\mathrm{st}}$. Consequently, the map

$$\pi \longmapsto (I - Z_\pi)^{-1} - I$$

is continuous on the compact set Π_{st} and is therefore uniformly bounded. This proves the claim. □

For a fixed policy π, the vector

$$\sum_{n=1}^{\infty} Z_\pi^n \mathbb{1}$$

gives the expected number of visits, at positive time instances, to nonabsorbing state-action pairs before absorption. Thus, Lemma A.9 states that the expected time to absorption is uniformly bounded over stationary policies and initial nonabsorbing state-action pairs.

For $\gamma = 1$, define the undiscounted absorbing criterion by

$$X_{\pi,1}(s,a) := \mathbb{E}_{s,a}^{\pi}\left[\exp\left(-\widehat{\theta}\sum_{t=0}^{\tau-1} r(S_t, A_t)\right)\right],$$

where

$$\tau := \inf\{t \geq 0 : S_t \in \mathcal{S}_{\mathrm{abs}}\}$$

is the first hitting time of the absorbing set. Here, action a is applied at time zero and policy π is followed thereafter. By Assumption 3, $\tau < \infty$ almost surely; moreover, Lemma A.9 gives a uniform bound on its expectation.

For every state-action pair (s,a) with $s \in \mathcal{S}_{\mathrm{abs}}$, all subsequent rewards are zero, and hence

$$X_{\pi,1}(s,a) = X_{\pi,\gamma}(s,a) = \bar{x}_{\pi,\gamma}(s,a) = 1.$$

We begin with a simple ordering relation.

Lemma A.10. *Under Assumption 3, for every $\pi \in \Pi_{\mathrm{st}}$ and $\gamma \in (0,1)$,*

$$\mathbf{0} \prec X_{\pi,1} \preceq X_{\pi,\gamma} \preceq \bar{x}_{\pi,\gamma} \preceq \mathbb{1}.$$

Proof. Fix $\pi \in \Pi_{\mathrm{st}}$. By Assumption 3, the absorption time τ is finite almost surely. Since the MDP is finite, the reward function is bounded, and hence

$$\sum_{t=0}^{\tau-1} r(S_t, A_t)$$

is finite almost surely. Consequently, the random variable defining $X_{\pi,1}$ is strictly positive almost surely, which gives

$$\mathbf{0} \prec X_{\pi,1}.$$

Since the rewards are nonnegative and $\gamma \in (0,1)$, we have, on every sample path,

$$\sum_{t=0}^{\tau-1} \gamma^t r(S_t, A_t) \leq \sum_{t=0}^{\tau-1} r(S_t, A_t).$$

Because the map $u \mapsto e^{-\widehat{\theta}u}$ is decreasing, it follows that

$$X_{\pi,1} \preceq X_{\pi,\gamma}.$$

The inequality

$$X_{\pi,\gamma} \preceq \bar{x}_{\pi,\gamma}$$

follows from Theorem 2.2.

It remains to show that $\bar{x}_{\pi,\gamma} \preceq \mathbb{1}$. This follows from Proposition 2.1 since the rewards are nonnegative, which implies

$$F_{\pi,\gamma}(\mathbb{1}) \preceq \mathbb{1}.$$

Combining the preceding inequalities proves the result. □

The next lemma gives the matrix estimate needed in the proof.

Lemma A.11. *For every $\pi \in \Pi_{\mathrm{st}}$, define*

$$D := \operatorname{diag}\left(e^{-\widehat{\theta} r(s,a)}\right)_{(s,a)\in\mathcal{Z}}.$$

Then $I - DZ_\pi$ is invertible and

$$(I - DZ_\pi)^{-1} DZ_\pi = \sum_{n=1}^{\infty} (DZ_\pi)^n.$$

Moreover,

$$\mathbf{0} \preceq (I - DZ_\pi)^{-1} DZ_\pi \preceq \sum_{n=1}^{\infty} Z_\pi^n.$$

In particular,

$$\left\|(I - DZ_\pi)^{-1} DZ_\pi\right\|_\infty \le K,$$

where K is the constant from Lemma A.9.

Proof. Fix $\pi \in \Pi_{\mathrm{st}}$. Since the rewards are nonnegative,

$$0 < e^{-\widehat{\theta} r(s,a)} \le 1, \qquad (s,a) \in \mathcal{Z}.$$

Therefore,

$$\mathbf{0} \preceq D \preceq I.$$

Since Z_π is nonnegative, this implies

$$\mathbf{0} \preceq DZ_\pi \preceq Z_\pi.$$

By induction, it follows that

$$\mathbf{0} \preceq (DZ_\pi)^n \preceq Z_\pi^n, \qquad n \ge 1.$$

Indeed, if $(DZ_\pi)^n \preceq Z_\pi^n$, then

$$(DZ_\pi)^{n+1} = (DZ_\pi)^n DZ_\pi \preceq Z_\pi^n DZ_\pi \preceq Z_\pi^{n+1}.$$

By Lemma A.9, the series

$$\sum_{n=1}^{\infty} Z_\pi^n$$

converges. The preceding componentwise bound therefore implies that

$$\sum_{n=1}^{\infty} (DZ_\pi)^n$$

also converges. Hence $\rho(DZ_\pi) < 1$, so the Neumann-series identity gives

$$(I - DZ_\pi)^{-1} = \sum_{n=0}^{\infty} (DZ_\pi)^n.$$

Multiplying by DZ_π yields

$$(I - DZ_\pi)^{-1} DZ_\pi = \sum_{n=1}^{\infty} (DZ_\pi)^n.$$

Moreover,

$$\mathbf{0} \preceq \sum_{n=1}^{\infty} (DZ_\pi)^n \preceq \sum_{n=1}^{\infty} Z_\pi^n.$$

Since the matrices are nonnegative, the infinity norm preserves this ordering. Consequently,

$$\left\|(I - DZ_\pi)^{-1} DZ_\pi\right\|_\infty \le \left\|\sum_{n=1}^{\infty} Z_\pi^n\right\|_\infty \le K,$$

where the final inequality follows from Lemma A.9. □

We now obtain a uniform bound for each fixed policy.

Lemma A.12. *Under Assumption 3, for every $\pi \in \Pi_{\mathrm{st}}$ and $\gamma \in (0,1)$,*

$$\mathbf{0} \preceq \bar{x}_{\pi,\gamma} - X_{\pi,\gamma} \preceq K(1-\gamma)\mathbb{1},$$

where K is the constant from Lemma A.9.

Proof. Fix $\pi \in \Pi_{\mathrm{st}}$. Throughout the proof, all vectors and matrices are restricted to the set $\mathcal{Z}$ of nonabsorbing state-action pairs. Let

$$x := \bar{x}_{\pi,\gamma}, \qquad v := X_{\pi,1},$$

and define

$$q_\pi := \mathbb{1} - Z_\pi \mathbb{1}.$$

For each $(s,a) \in \mathcal{Z}$, the component $q_\pi(s,a)$ is the probability that the process enters the absorbing set after one transition from (s,a).

Recall that

$$D := \operatorname{diag}\left(e^{-\widehat{\theta} r(s,a)}\right)_{(s,a)\in\mathcal{Z}}.$$

Since the value at every absorbing state-action pair equals one, the fixed-point equation for $\bar{x}_{\pi,\gamma}$, restricted to $\mathcal{Z}$, can be written as

$$x = D\left(q_\pi + Z_\pi x^\gamma\right),$$

where the power is applied componentwise. Similarly, the undiscounted absorbing criterion satisfies

$$v = D\left(q_\pi + Z_\pi v\right).$$

Subtracting these equations gives

$$\begin{aligned} x - v &= DZ_\pi(x^\gamma - v) \\ &= DZ_\pi(x^\gamma - x) + DZ_\pi(x - v). \end{aligned}$$

Therefore,

$$x - v = (I - DZ_\pi)^{-1} DZ_\pi (x^\gamma - x).$$

By Lemma A.10,

$$\mathbf{0} \prec x \preceq \mathbb{1}.$$

For every scalar $y \in [0,1]$, we have

$$y \le y^\gamma.$$

Moreover, since $y \mapsto y^\gamma$ is concave, its tangent at $y = 1$ gives

$$y^\gamma \le \gamma y + (1-\gamma).$$

Consequently,

$$0 \le y^\gamma - y \le (1-\gamma)(1-y) \le 1 - \gamma.$$

Applying this componentwise yields

$$\mathbf{0} \preceq x^\gamma - x \preceq (1-\gamma)\mathbb{1}.$$

Let

$$R_\pi := (I - DZ_\pi)^{-1} DZ_\pi.$$

By Lemma A.11, R_π is entrywise nonnegative and

$$\|R_\pi\|_\infty \le K.$$

It follows that

$$\begin{aligned} \mathbf{0} \preceq x - v &= R_\pi(x^\gamma - x) \\ &\preceq (1-\gamma) R_\pi \mathbb{1} \\ &\preceq (1-\gamma)\|R_\pi\|_\infty \mathbb{1} \\ &\preceq K(1-\gamma)\mathbb{1}. \end{aligned}$$

Finally, Lemma A.10 gives

$$X_{\pi,1} \preceq X_{\pi,\gamma} \preceq \bar{x}_{\pi,\gamma}.$$

Consequently,

$$\mathbf{0} \preceq \bar{x}_{\pi,\gamma} - X_{\pi,\gamma} \preceq \bar{x}_{\pi,\gamma} - X_{\pi,1} \preceq K(1-\gamma)\mathbb{1}.$$

At state-action pairs belonging to the absorbing set, all three quantities equal one, so the same conclusion holds on the entire state-action space. □

We can now compare the optimal values of the two criteria. Since a single stationary policy need not simultaneously attain the minimum of $X_{\pi,\gamma}(s,a)$ at every state-action pair, we define the componentwise optimal-value envelopes

$$X^\star_\gamma(s,a) := \inf_{\pi\in\Pi_{\mathrm{st}}} X_{\pi,\gamma}(s,a)$$

and

$$\bar{x}^\star_\gamma(s,a) := \inf_{\pi\in\Pi_{\mathrm{st}}} \bar{x}_{\pi,\gamma}(s,a).$$

Theorem A.13 (Asymptotic agreement of the optimal values)**.** *Under Assumption 3, for every* $\gamma \in (0,1)$*,*

$$\mathbf{0} \preceq \bar{x}^\star_\gamma - X^\star_\gamma \preceq K(1-\gamma)\mathbb{1},$$

where K *is the constant from Lemma A.9. Consequently,*

$$\left\|\bar{x}^\star_\gamma - X^\star_\gamma\right\|_\infty \le K(1-\gamma) \longrightarrow 0 \qquad \text{as } \gamma \uparrow 1.$$

Proof. Fix an admissible state-action pair (s,a). By Lemma A.12, for every $\pi \in \Pi_{\mathrm{st}}$,

$$X_{\pi,\gamma}(s,a) \le \bar{x}_{\pi,\gamma}(s,a) \le X_{\pi,\gamma}(s,a) + K(1-\gamma).$$

Since the sets

$$\{X_{\pi,\gamma}(s,a) : \pi \in \Pi_{\mathrm{st}}\} \qquad \text{and} \qquad \{\bar{x}_{\pi,\gamma}(s,a) : \pi \in \Pi_{\mathrm{st}}\}$$

are nonempty and bounded below by zero, their infima are well defined.

Taking the infimum over $\pi \in \Pi_{\mathrm{st}}$ in the first inequality gives

$$X^\star_\gamma(s,a) \le \bar{x}^\star_\gamma(s,a).$$

Similarly, taking the infimum over $\pi \in \Pi_{\mathrm{st}}$ in the second inequality gives

$$\begin{aligned}\bar{x}^\star_\gamma(s,a) &\le \inf_{\pi\in\Pi_{\mathrm{st}}} [X_{\pi,\gamma}(s,a) + K(1-\gamma)] \\ &= \inf_{\pi\in\Pi_{\mathrm{st}}} X_{\pi,\gamma}(s,a) + K(1-\gamma) \\ &= X^\star_\gamma(s,a) + K(1-\gamma),\end{aligned}$$

where the second equality follows because $K(1-\gamma)$ does not depend on π. Thus,

$$0 \le \bar{x}^\star_\gamma(s,a) - X^\star_\gamma(s,a) \le K(1-\gamma).$$

Since (s,a) was arbitrary, we conclude that

$$\mathbf{0} \preceq \bar{x}^\star_\gamma - X^\star_\gamma \preceq K(1-\gamma)\mathbb{1}.$$

Taking the infinity norm yields

$$\left\|\bar{x}^\star_\gamma - X^\star_\gamma\right\|_\infty \le K(1-\gamma).$$

Since the right-hand side converges to zero as $\gamma \uparrow 1$, the desired result follows. □

B Analysis for the Two-Timescale Fixed Point Method

Here we discuss the stability, asymptotic convergence, and convergence rates of the two-timescale fixed-point method in (7).

B.1 Stability

We first prove the almost sure boundedness of the sequences (Q_n) and (g_n).

Lemma B.1. *Suppose* $Q_0 \in \mathbb{R}^{SA}$ *and* $g_0 \in \mathbb{R}^{SA}_{++}$ *be such that* $\max\{\|Q_0\|_\infty, \frac{\gamma}{\theta}\|\ln g_0\|_\infty\} \le C_{\max}$*, where* $C_{\max} := \frac{\|r\|_\infty}{1-\gamma}$*. Also, suppose* $\alpha_n, \beta_n \in [0,1]$ *for all* $n \ge 0$*. Then, for each* $n \ge 0$*, we have*

$$\max\left\{\|Q_n\|_\infty, \frac{\gamma}{\theta}\|\ln g_n\|_\infty\right\} \le C_{\max}. \tag{16}$$

Proof. We prove using induction. For $n = 0$, the claim holds trivially. Now assume that

$$\|Q_n\|_\infty \le C_{\max} \qquad \text{and} \qquad \frac{\gamma}{\theta}\|\ln g_n\|_\infty \le C_{\max}$$

for some $n \ge 0$. The second inequality implies, for all (s,a),

$$\exp\left[-\frac{\theta}{\gamma}C_{\max}\right] \le g_n(s,a) \le \exp\left[\frac{\theta}{\gamma}C_{\max}\right]. \tag{17}$$

The rest of the proof proceeds via three steps.

Step 1: Bound $\frac{\gamma}{\theta}\|\ln g_{n+1}\|_\infty$. For $(s,a) \neq (s_n, a_n)$, we have $g_{n+1}(s,a) = g_n(s,a)$; hence, for every $(s,a) \neq (s_n, a_n)$, we have

$$\frac{\gamma}{\theta}|\ln g_{n+1}(s,a)| = \frac{\gamma}{\theta}|\ln g_n(s,a)| \leq C_{\max}.$$

For the updated coordinate, we have

$$g_{n+1}(s_n, a_n) = (1-\beta_n) g_n(s_n, a_n) + \beta_n \hat{G}_n. \tag{18}$$

Also, $\hat{G}_n = \exp\Big[-\frac{\theta}{\gamma}\Big(r(s_n,a_n) + \gamma \max_{a'} Q_n(s_{n+1}, a')\Big)\Big]$, and

$$\big|r(s_n,a_n) + \gamma \max_{a'} Q_n(s_{n+1},a')\big| \leq \|r\|_\infty + \gamma C_{\max} \leq C_{\max},$$

where the first inequality follows since $|r(s_n,a_n)| \leq \|r\|_\infty$ and $\|Q_n\|_\infty \leq C_{\max}$, while the last inequality uses $C_{\max} \geq \|r\|_\infty/(1-\gamma)$. Therefore,

$$\exp\left[-\frac{\theta}{\gamma}C_{\max}\right] \leq \hat{G}_n \leq \exp\left[\frac{\theta}{\gamma}C_{\max}\right]. \tag{19}$$

Since $\beta_n \in [0,1]$, it follows from (17), (19), and (18) that

$$\exp\left[-\frac{\theta}{\gamma}C_{\max}\right] \leq g_{n+1}(s_n,a_n) \leq \exp\left[\frac{\theta}{\gamma}C_{\max}\right].$$

Hence,

$$\frac{\gamma}{\theta}\|\ln g_{n+1}\|_\infty \leq C_{\max}. \tag{20}$$

Step 2: Bound $\|Q_{n+1}\|_\infty$. For all (s,a), our hypothesis implies that

$$\begin{aligned}
|Q_{n+1}(s,a)| &\leq (1-\alpha_n)|Q_n(s,a)| + \alpha_n \frac{\gamma}{\theta}|\ln g_n(s,a)| \\
&\leq (1-\alpha_n)C_{\max} + \alpha_n C_{\max} \\
&= C_{\max}.
\end{aligned}$$

Combining Steps 1 and 2 proves

$$\max\Big\{\|Q_{n+1}\|_\infty,\ \frac{\gamma}{\theta}\|\ln g_{n+1}\|_\infty\Big\} \leq C_{\max},$$

closing the induction. The desired result now follows. □

B.2 Asymptotic Convergence

Here we establish (Q_n) and (g_n)'s asymptotic convergence. Let $(\mathcal{F}_n)_{n\geq 0}$ be the filtration given by

$$\mathcal{F}_n = \sigma(Q_0, g_0, s_0, a_0, \ldots, s_n, a_n). \tag{21}$$

Proof of Theorem 4.1. . We first analyze (g_n)'s asymptotic behavior. For that, we consider the update rules in (7) from the (β_n)-stepsize perspective. That is, we rewrite (7) as

$$z_{n+1} = z_n + \beta_n[\phi(z_n, s_n, a_n, s_{n+1}) + \epsilon_n], \quad n \geq 0,$$

where $z_n = \begin{bmatrix} Q_n \\ g_n \end{bmatrix}$, $\phi : \mathbb{R}^{2SA} \times \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}^{2SA}$ is given by

$$\begin{aligned}
\phi(z,s,a,s') &= \begin{bmatrix} 0 \\ e_{s,a}[G(Q,s,a,s') - g(s,a)] \end{bmatrix} \\
&= \begin{bmatrix} 0 \\ e_{s,a}\left[\exp\left(-\frac{\theta}{\gamma}r(s,a) - \theta \max_{a'} Q(s',a')\right) - g(s,a)\right] \end{bmatrix}
\end{aligned} \tag{22}$$

for any $z = \begin{bmatrix} Q \\ g \end{bmatrix}$, and

$$\epsilon_n = \begin{bmatrix} \frac{\alpha_n}{\beta_n} e_{s_n,a_n}[-\frac{\gamma}{\theta} \ln g_n(s_n,a_n) - Q_n(s_n,a_n)] \\ 0 \end{bmatrix},$$

which again is a vector in $\mathbb{R}^{2SA}$. Clearly, z_n is $\mathcal{F}_n$-measurable for each $n \geq 0$. With this in mind, the above recursion can be equivalently expressed in the standard stochastic approximation form

$$z_{n+1} = z_n + \beta_n[\bar{\phi}(z_n, s_n, a_n) + \epsilon_n + M_{n+1}], \quad n \geq 0.$$

Here, $\bar{\phi}(z_n, s_n, a_n) := \mathbb{E}[\phi(z_n, s_n, a_n, s_{n+1}) \mid \mathcal{F}_n] = \mathbb{E}[\phi(z_n, s_n, a_n, s_{n+1}) \mid z_n, s_n, a_n]$. More explicitly, for $z \in \mathbb{R}^{2SA}$ and $(s,a) \in \mathcal{S} \times \mathcal{A}$, we have

$$\begin{aligned} \bar{\phi}(z,s,a) &= \mathbb{E}[\phi(z_n, s_n, a_n, s_{n+1}) \mid z_n = z, s_n = s, a_n = a] \\ &= \begin{bmatrix} 0 \\ e_{s,a}\Big[\sum_{s' \in \mathcal{S}} \mathcal{P}(s'|s,a) \exp\Big(-\frac{\theta}{\gamma} r(s,a) - \theta \max_{a'} Q(s',a')\Big) - g(s,a)\Big] \end{bmatrix} \\ &= \begin{bmatrix} 0 \\ e_{s,a}\Big(\exp\Big[-\frac{\theta}{\gamma} TQ(s,a)\Big] - g(s,a)\Big) \end{bmatrix}, \end{aligned} \tag{23}$$

where the last relation follows from (5). The noise term

$$M_{n+1} = \phi(z_n, s_n, a_n, s_{n+1}) - \bar{\phi}(z_n, s_n, a_n), \tag{24}$$

and it satisfies $\mathbb{E}[M_{n+1}|\mathcal{F}_n] = 0$ a.s. Hence, $(M_n)_{n \geq 1}$ forms a Martingale-difference sequence with respect to the filtration $(\mathcal{F}_n)$. Also, since $z_n \in \mathcal{K}$ from Lemma B.1, we have from (22) and (23) that

$$\|M_{n+1}\|_\infty \leq 2 \exp\left[\frac{\theta}{\gamma} C_{\max}\right] \text{ a.s.}$$

This relation also shows that (M_n) is square-integrable and satisfies

$$\mathbb{E}\big[\|M_{n+1}\|_\infty^2 \mid \mathcal{F}_n\big] \leq 4 \exp\left[\frac{2\theta}{\gamma} C_{\max}\right] \leq 4 \exp\left[\frac{2\theta}{\gamma} C_{\max}\right] \big(1 + \|z_n\|_\infty^2\big) \quad \text{a.s.}$$

Finally, since $z_n \in \mathcal{K}$ from Lemma B.1 for each $n \geq 0$ and since $\alpha_n/\beta_n \to 0$, we have $\epsilon_n \xrightarrow{\text{a.s.}} 0$.

Now define the averaged drift $\tilde{\phi} : \mathbb{R}^{2SA} \to \mathbb{R}^{2SA}$ by

$$\tilde{\phi}(z) := \sum_{(s,a) \in \mathcal{S} \times \mathcal{A}} \pi_Q(s,a)\, \bar{\phi}(z,s,a) = \begin{bmatrix} 0 \\ D_{\pi_Q}\Big[\exp\Big(-\frac{\theta}{\gamma} TQ\Big) - g\Big] \end{bmatrix},$$

where π_Q, recall, denotes the stationary distribution of the Markov chain with transition kernel K_Q, and $D_{\pi_Q} \in \mathbb{R}^{SA \times SA}$ is the diagonal matrix whose (s,a)-th diagonal entry equals $\pi_Q(s,a)$. It then follows from [Borkar, 2008, Corollary 8.1 and Remark 3, Section 2.2] that the sequence (z_n) converges almost surely to a connected, internally chain transitive invariant set of the ODE

$$\dot{z}(t) = \tilde{\phi}\big(z(t)\big).$$

The above ODE admits

$$\mathcal{A} := \left\{ \begin{bmatrix} Q \\ \exp\Big[-\frac{\theta}{\gamma} TQ\Big] \end{bmatrix} : \; Q \in \mathbb{R}^{SA} \right\}$$

as its unique global attractor. Indeed, since $\dot{Q}(t) = 0$, the Q-component remains constant along trajectories. For a fixed Q, the dynamics of the g-component reduces to

$$\dot{g}(t) = D_{\pi_Q}\Big(\exp\Big[-\frac{\theta}{\gamma} TQ\Big] - g(t)\Big),$$

which is a linear ODE with $\exp\Big[-\frac{\theta}{\gamma} TQ\Big]$ as its globally asymptotically stable equilibrium. Therefore, we conclude that $z_n \xrightarrow{\text{a.s.}} \mathcal{A}$, or equivalently,

$$\Big\| g_n - \exp\Big[-\frac{\theta}{\gamma} TQ_n\Big] \Big\| \xrightarrow{\text{a.s.}} 0, \tag{25}$$

as claimed.

We now discuss (Q_n)'s behavior from the (α_n)-stepsize perspective. The update rule for Q_n can be rewritten as

$$Q_{n+1} = Q_n + \alpha_n[TQ_n - Q_n + \epsilon'_n],$$

where

$$\epsilon'_n = \left[-\frac{\gamma}{\theta}\ln g_n - TQ_n\right].$$

Now, from (25), we have $\epsilon'_n \xrightarrow{\text{a.s.}} 0$. Hence, from [Borkar, 2008, Corollary 8.1 and Remark 3 of Section 2.2], we have that (Q_n) converges to a connected, internally chain transitive invariant set of the ODE

$$\dot{Q}(t) = TQ(t) - Q(t). \tag{26}$$

To complete the proof, it remains to show that Q^* is the unique globally asymptotically stable equilibrium of the above ODE.

For $t \geq 0$, define $z(t) := Q(t) - Q^*$ and $u(t) := \|z(t)\|_\infty$. Since $TQ^* = Q^*$, we have

$$\dot{z}_i(t) = (TQ(t))_i - (TQ^*)_i - z_i(t), \quad i \in \mathcal{S} \times \mathcal{A}. \tag{27}$$

Since $Q(\cdot)$ is absolutely continuous and $\|\cdot\|_\infty$ is Lipchitz continuous, it follows that $u(\cdot)$ is absolutely continuous and hence differential for a.e. $t \geq 0$. Fix such a t, and let i_t be any index satisfying

$$|z_{i_t}(t)| = \|z(t)\|_\infty = u(t).$$

Then,

$$\frac{\mathrm{d}}{\mathrm{d}t}u(t) = \text{sign}(z_{i_t}(t))\dot{z}_{i_t}(t).$$

Substituting (27) with $i = i_t$, and using the γ-contraction property of T, we obtain

$$\frac{\mathrm{d}}{\mathrm{d}t}u(t) = \text{sign}(z_{i_t}(t))\Big((TQ(t))_{i_t} - (TQ^*)_{i_t}\Big) - u(t) \leq -(1-\gamma)u(t). \tag{28}$$

Since this is true for a.e. $t \geq 0$, applying the differential form of Gronwall's inequality yields

$$\|Q(t) - Q^*\|_\infty \leq e^{-(1-\gamma)t}\|Q(0) - Q^*\|_\infty, \quad t \geq 0.$$

Thus, every solution trajectory of (26) converges exponentially fast to Q^*. In particular, Q^* is the unique globally asymptotically stable equilibrium of (26), as desired. □

B.3 Convergence Rates

Throughout this section, we assume that at each iteration $n \geq 0$, the state-action pair (s_n, a_n) is sampled from a stationary distribution ν, while the next state s'_n is sampled from $\mathcal{P}(\cdot|s_n, a_n)$. Also let $D^\nu := \text{diag}(\nu(s,a) : (s,a) \in \mathcal{S} \times \mathcal{A})$ and $\lambda := \min_{s,a}\nu(s,a) > 0$.

Before we proceed, we define some relevant notation. Define for $0 \leq n_1 < n_2$,

$$\Gamma^{(\beta)}_{n_1:n_2} := \prod_{n=n_1}^{n_2-1}\left(1 - 2\lambda\beta_n + 3\beta_n^2\right) \quad \text{and} \quad \Gamma^{(\alpha)}_{n_1:n_2} := \prod_{n=n_1}^{n_2-1}\left(1 - (1-\gamma)\alpha_n\right).$$

Also, define the constants

$$C_g := e^{\frac{2\theta}{\gamma}C_{\max}}\left[4e^{\frac{6\beta}{2\beta-1}}e^{\frac{2}{1-\beta}}\left(\frac{\beta}{2e}\right)^{\frac{\beta}{1-\beta}} + 2C'(1 + 6\theta^2 C_{\max}^2)\right]^{1/2},$$

$$C_Q^{(1)} := 2C_{\max}e^{\frac{(1-\gamma)}{(1-\alpha)}}, \quad C_Q^{(2)} := \frac{\gamma}{\theta}C_g C'' e^{\frac{\theta}{\gamma}C_{\max}},$$

where

$$\begin{aligned} C' &:= e^{\frac{4\beta}{2\beta-1}} \cdot \max\left\{2^{-\beta}, \left[\frac{4\beta}{2^\beta e(1-\beta)}\right]^{\frac{\beta}{1-\beta}}\right\}, \\ C'' &:= \frac{2}{1-\gamma} \cdot \max\left\{2^{-\frac{\beta}{2}}, \left[\frac{2\beta}{2^\alpha e(1-\gamma)(1-\alpha)}\right]^{\frac{\beta}{2(1-\alpha)}}\right\}. \end{aligned} \tag{29}$$

Finally, we present the following technical lemma, which we prove later.

Lemma B.2. *For every $0 \le k \le N$, we have*

1. $\Gamma^{(\beta)}_{k:N} \le C_{\Gamma,\beta} \exp\left(-2\lambda \sum_{n=k}^{N-1} \beta_n\right)$ *and* $\Gamma^{(\alpha)}_{k:N} \le \exp\left(-(1-\gamma)\sum_{n=k}^{N-1} \alpha_n\right)$;
2. $\sum_{n=k}^{N-1} \beta_n f_n \, \Gamma^{(\beta)}_{n+1:N} \le C^{a,b}_{\Gamma,\beta} \, f_N$ *and* $\sum_{n=k}^{N-1} \alpha_n f_n \, \Gamma^{(\alpha)}_{n+1:N} \le C^{a,b}_{\Gamma,\alpha} \, f_N$,

where $f_n := \alpha_n^a \beta_n^b$, for some $a, b \in \mathbb{Z} \cup \{1/2\}$, and the constants are defined as

$$C_{\Gamma,\beta} := \exp\left(\frac{6\beta}{2\beta-1}\right), \quad C^{a,b}_{\Gamma,\beta} := C_{\Gamma,\beta} \cdot \max\left\{2^{-(a\alpha+b\beta)}, \left[\frac{a\alpha+b\beta}{\lambda e(1-\beta)2^\beta}\right]^{\frac{a\alpha+b\beta}{1-\beta}}\right\},$$

and

$$C^{a,b}_{\Gamma,\alpha} := \frac{2}{1-\gamma} \max\left\{2^{-(a\alpha+b\beta)}, \left[\frac{4(a\alpha+b\beta)}{2^\alpha e(1-\gamma)(1-\alpha)}\right]^{\left(\frac{a\alpha+b\beta}{1-\alpha}\right)}\right\}.$$

Proof of Theorem 4.2. We first discuss (g_n)'s behavior from the (β_n)-stepsize perspective. Let $y_n = g_n - e^{-\theta/\gamma TQ_n}$. Then, we have for all $n \ge 0$,

$$\begin{aligned}
y_{n+1} &= g_{n+1} - e^{-\frac{\theta}{\gamma}TQ_{n+1}} \\
&\overset{(a)}{=} y_n + [e^{-\frac{\theta}{\gamma}TQ_n} - e^{-\frac{\theta}{\gamma}TQ_{n+1}}] + \beta_n e_{s_n,a_n}\left[\hat{G}_n - g_n(s_n,a_n)\right] \\
&\overset{(b)}{=} y_n + \epsilon_n + \beta_n e_{s_n,a_n}\left[e^{-\frac{\theta}{\gamma}TQ_n(a_n,s_n)} - g_n(s_n,a_n) + \hat{G}_n - e^{-\frac{\theta}{\gamma}TQ_n(a_n,s_n)}\right] \\
&\overset{(c)}{=} y_n - \beta_n e_{s_n,a_n} y_n(s_n,a_n) + \beta_n e_{s_n,a_n}\left[\hat{G}_n - e^{-\frac{\theta}{\gamma}TQ_n(a_n,s_n)}\right] + \epsilon_n, \qquad (30)
\end{aligned}$$

where (a) follows from the update rule of (g_n), (b) follows from defining $\epsilon_n := e^{-(\theta/\gamma)TQ_n} - e^{-(\theta/\gamma)TQ_{n+1}}$, and (c) uses the definition of y_n.

For $n \ge 0$, define $M_{n+1} := e_{s_n,a_n}\left[\hat{G}_n - e^{-\frac{\theta}{\gamma}TQ_n(a_n,s_n)}\right]$ and let $\mathbb{E}_n[\cdot]$ denote conditional expectation w.r.t the σ-algebra

$$\mathcal{F}'_n := \sigma(Q_0, g_0, s_0, a_0, s'_0, \ldots, s_{n-1}, a_{n-1}, s'_{n-1}). \qquad (31)$$

Then, squaring both sides of (30) and taking expectation gives us

$$\begin{aligned}
&\mathbb{E}\|y_{n+1}\|_2^2 \\
&\overset{(a)}{=} \mathbb{E}\|\left(I - \beta_n e_{s_n,a_n} e_{s_n,a_n}^\top\right) y_n\|_2^2 + \mathbb{E}\|\epsilon_n\|_2^2 + \beta_n^2 \mathbb{E}\|M_{n+1}\|_2^2 \\
&\quad + 2\beta_n \mathbb{E}\left\langle \mathbb{E}_n[\epsilon_n], \mathbb{E}\left[I - \beta_n e_{s_n,a_n} e_{s_n,a_n}^\top\right] y_n\right\rangle \\
&\quad + 2\beta_n \mathbb{E}\left\langle \mathbb{E}_n[M_{n+1}], \mathbb{E}\left[I - \beta_n e_{s_n,a_n} e_{s_n,a_n}^\top\right] y_n\right\rangle + 2\beta_n \mathbb{E}\left[\langle M_{n+1}, \epsilon_n\rangle\right] \\
&\overset{(b)}{\le} \mathbb{E}\|\left(I - \beta_n e_{s_n,a_n} e_{s_n,a_n}^\top\right) y_n\|_2^2 + \mathbb{E}\|\epsilon_n\|_2^2 + \beta_n^2 \mathbb{E}\|M_{n+1}\|_2^2 \\
&\quad + \beta_n\left(\beta_n \mathbb{E}\|(I - \beta_n D^\nu) y_n\|_2^2 + \frac{1}{\beta_n}\mathbb{E}\|\mathbb{E}_n[\epsilon_n]\|_2^2\right) + \beta_n\left(\beta_n \mathbb{E}\|M_{n+1}\|_2^2 + \frac{1}{\beta_n}\mathbb{E}\|\epsilon_n\|_2^2\right) \\
&\overset{(c)}{\le} \mathbb{E}\|\left(I - \beta_n e_{s_n,a_n} e_{s_n,a_n}^\top\right) y_n\|_2^2 + 2\beta_n^2 \mathbb{E}\|M_{n+1}\|_2^2 + \beta_n^2 \mathbb{E}\|(I - \beta_n D^\nu) y_n\|_2^2 + 3\mathbb{E}\|\epsilon_n\|_2^2 \\
&\overset{(d)}{\le} \mathbb{E}\|\left(I - \beta_n e_{s_n,a_n} e_{s_n,a_n}^\top\right) y_n\|_2^2 + 2e^{2\theta C_{\max}}\beta_n^2 + \beta_n^2 \mathbb{E}\|(I - \beta_n D^\nu) y_n\|_2^2 + 3\mathbb{E}\|\epsilon_n\|_2^2, \qquad (32)
\end{aligned}$$

where (a) uses the fact that e_{s_n,a_n} is independent of $\mathcal{F}'_n$ while y_n is $\mathcal{F}'_n$-measurable, whereas (b) follows by applying the Young's inequality, taking $(s_n, a_n) \sim \nu$, and using the fact that $\mathbb{E}_n[M_{n+1}] = 0$ since

$$\mathbb{E}_n[\hat{G}_n] = \mathbb{E}_n[G(Q_n, s_n, a_n, s_{n+1})] = \sum_{s'} P(s'|s_n,a_n) G(Q_n, s_n, a_n, s') = e^{-\frac{\theta}{\gamma}TQ_n}.$$

Further, (c) uses Jensen's inequality for conditional expectation, whereas (d) follows since $r(s,a) \geq 0$ and $Q_n(s,a) \geq -C_{\max}$, giving

$$\|M_{n+1}\|_2 \leq 2|\hat{G}_n| \leq \exp\left(-\frac{\theta}{\gamma} r(s_n, a_n) - \theta \max_{a'} Q_n(s_{n+1}, a')\right) \leq e^{\theta C_{\max}}. \tag{33}$$

Finally, note that

$$\begin{aligned}
|\epsilon_n(s,a)| &\overset{(a)}{\leq} \sum_{s'} P(s'|s,a) \left| e^{-\theta \max_{a'} Q_n(s',a')} - e^{-\theta \max_{a'} Q_{n+1}(s',a')} \right| \\
&\overset{(b)}{\leq} \theta e^{\theta C_{\max}} \sum_{s'} P(s'|s,a) \left| \max_{a'} Q_n(s',a') - \max_{a'} Q_{n+1}(s',a') \right| \\
&\leq \theta e^{\theta C_{\max}} \|Q_n - Q_{n+1}\|_\infty \\
&\overset{(c)}{\leq} \theta e^{\theta C_{\max}} \left(\frac{\gamma}{\theta} \|\ln g_n\|_\infty + \|Q_n\|_\infty \right) \alpha_n \overset{(d)}{\leq} 2C_{\max} \theta e^{\theta C_{\max}} \alpha_n,
\end{aligned} \tag{34}$$

where (a) uses the definition of T, (b) follows from Lemma B.1 since $u \mapsto \exp(-\theta u)$ is $\theta \exp(\theta C_{\max})$-Lipschitz on $[-C_{\max}, C_{\max}]$, (c) uses the update rule for (Q_n), and (d) uses Lemma B.1.

Now, since y_n is $\mathcal{F}'_n$-measurable, we have

$$\begin{aligned}
\mathbb{E}_n \| \left(I - \beta_n e_{s_n,a_n} e_{s_n,a_n}^\top\right) y_n \|_2^2 &= y_n^\top \mathbb{E}_n \left[\left(I - \beta_n e_{s_n,a_n} e_{s_n,a_n}^\top\right)^\top \left(I - \beta_n e_{s_n,a_n} e_{s_n,a_n}^\top\right) \right] y_n \\
&\overset{(a)}{=} y_n^\top \mathbb{E} \left[\left(I - \beta_n e_{s_n,a_n} e_{s_n,a_n}^\top\right)^\top \left(I - \beta_n e_{s_n,a_n} e_{s_n,a_n}^\top\right) \right] y_n \\
&= y_n^\top \mathbb{E} \left[\left(I - \beta_n e_{s_n,a_n} e_{s_n,a_n}^\top\right) \left(I - \beta_n e_{s_n,a_n} e_{s_n,a_n}^\top\right) \right] y_n \\
&\overset{(b)}{=} y_n^\top \mathbb{E} \left[I - 2\beta_n e_{s_n,a_n} e_{s_n,a_n}^\top + \beta_n^2 e_{s_n,a_n} e_{s_n,a_n}^\top \right] y_n \\
&\overset{(c)}{=} y_n^\top \mathbb{E} \left[I - 2\beta_n D^\nu + \beta_n^2 D^\nu \right] y_n \\
&\overset{(d)}{\leq} \left(1 - 2\lambda\beta_n + \beta_n^2\right) \|y_n\|_2^2,
\end{aligned}$$

where (a) follows since $(s_n, a_n) \sim \nu \perp \mathcal{F}'_n$, (b) follows since $e_{s_n,a_n}^\top e_{s_n,a_n} = 1$, (c) follows from the fact that $(s_n, a_n) \sim \nu$, and (d) follows from the definition of λ, and using $y^\top D^\nu y \leq y^\top y$.

Furthermore, (d) implies

$$\|(I - \beta_n D^\nu) y_n\|_2^2 \leq \left(1 - 2\lambda\beta_n + \beta_n^2\right) \|y_n\|_2^2.$$

Plugging the above bounds into (32) gives us

$$\begin{aligned}
\mathbb{E}\|y_{n+1}\|_2^2 &\leq \left(1 + \beta_n^2\right) \left(1 - 2\lambda\beta_n + \beta_n^2\right) \mathbb{E}\|y_n\|_2^2 + 2e^{2\theta C_{\max}} \beta_n + 12\theta^2 C_{\max}^2 e^{2\theta C_{\max}} \alpha_n^2, \\
&\overset{(a)}{\leq} (1 - 2\lambda\beta_n + 3\beta_n^2) \mathbb{E}\|y_n\|_2^2 + 2e^{2\theta C_{\max}} (1 + 6\theta^2 C_{\max}^2) \beta_n^2,
\end{aligned} \tag{35}$$

where (a) follows as we choose β_n such that $(1 - 2\lambda\beta_n + \beta) \leq (1 - 2\lambda\beta_n + 3\beta) \leq 1$. Then, recursively applying (35) gives

$$\mathbb{E}\|y_N\|^2 \leq \Gamma_{0:N}^{(\beta)} \left\| g_0 - e^{-\frac{\theta}{\gamma} T Q_0} \right\|_2^2 + 2e^{2\theta C_{\max}} (1 + 6\theta^2 C_{\max}^2) \sum_{n=0}^{N-1} \beta_n^2 \Gamma_{n+1:N}^{(\beta)}. \tag{36}$$

Applying Lemma B.2 to (36) gives us

$$\begin{aligned}
&\mathbb{E}\|y_N\|_2^2 \\
&\overset{(a)}{\leq} C_{\Gamma,\beta} \left\| g_0 - e^{-\frac{\theta}{\gamma} T Q_0} \right\|_2^2 e^{-2\lambda \sum_{n=0}^{N-1} \beta_n} + 2C_{\Gamma,\beta}^{0,1} e^{2\theta C_{\max}} (1 + 6\theta^2 C_{\max}^2) \beta_N \\
&\overset{(b)}{\leq} C_{\Gamma,\beta} \left\| g_0 - e^{-\frac{\theta}{\gamma} T Q_0} \right\|_2^2 e^{\frac{2}{1-\beta}} \cdot e^{-\frac{2\lambda}{1-\beta} N^{1-\beta}} + 2C_{\Gamma,\beta}^{0,1} e^{2\theta C_{\max}} (1 + 6\theta^2 C_{\max}^2) \beta_N
\end{aligned}$$

$$\overset{(c)}{\leq} 4C_{\Gamma,\beta}e^{\left(\frac{2}{1-\beta}+\frac{2\theta C_{\max}}{\gamma}\right)}e^{-\frac{2\lambda}{1-\beta}(N+1)^{1-\beta}} + 2C^{0,1}_{\Gamma,\beta}e^{2\theta C_{\max}}(1+6\theta^2C^2_{\max})\beta_N, \tag{37}$$

where in (a), we additionally use the fact that $\alpha_n^2 = \beta_n(\alpha_n^2/\beta_n)$ and α_n^2/β_n is decreasing in n. Further, (b) follows since $\sum_{n=0}^{N-1}(n+1)^{-\beta} \geq [(N+1)^{1-\beta}-1]/(1-\beta)$, and (c) uses bounds from Lemma B.1.

At last, we use the fact that $x \mapsto x^\beta \cdot e^{-\frac{2\lambda}{1-\beta}x^{1-\beta}}$ attains maximum at $x = (\beta/2\lambda)^{1/(1-\beta)}$ to conclude

$$e^{-\frac{2\lambda}{1-\beta}(N+1)^{1-\beta}} \leq \left[(N+1)^\beta \cdot e^{-\frac{2\lambda}{1-\beta}(N+1)^{1-\beta}}\right]\beta_N \leq \left(\frac{\beta}{2\lambda e}\right)^{\frac{\beta}{1-\beta}}\beta_N. \tag{38}$$

Hence, for all $N \geq 0$,

$$\mathbb{E}\left\|g_N - e^{\frac{-\theta}{\gamma}TQ_N}\right\|_2^2 \leq C_g^2\beta_N, \tag{39}$$

where we define

$$C_g := e^{\frac{2\theta}{\gamma}C_{\max}}\left[4C_{\Gamma,\beta}e^{\frac{2}{1-\beta}}\left(\frac{\beta}{2\lambda e}\right)^{\frac{\beta}{1-\beta}} + 2C^{0,1}_{\Gamma,\beta}(1+6\theta^2C^2_{\max})\right]^{1/2}.$$

This completes the finite-time analysis for (g_n).

Now, we discuss (Q_n)'s behavior from the (α_n)-stepsize perspective. The update rule for Q_n can be rewritten as

$$Q_{n+1} = Q_n + \beta_n[TQ_n - Q_n] + \alpha_n\epsilon_n',$$

where

$$\epsilon_n' := \left[-\frac{\gamma}{\theta}\ln g_n - TQ_n\right].$$

To proceed, we need to bound ϵ_n' appropriately. To do so, note that

$$\begin{aligned}
\|\epsilon_n'\|_\infty &= \left\|\frac{\gamma}{\theta}\ln g_n + TQ_n\right\|_\infty \\
&\leq \left\|\frac{\gamma}{\theta}\ln g_n + TQ_n\right\|_2 = \frac{\gamma}{\theta}\left\|\ln g_n - \ln e^{-\frac{\theta}{\gamma}TQ_n}\right\|_2 \\
&\overset{(a)}{\leq} \frac{\gamma}{\theta}e^{\frac{\theta C_{\max}}{\gamma}}\left\|g_n - e^{-\frac{\theta}{\gamma}TQ_n}\right\|_2 = \frac{\gamma}{\theta}e^{\frac{\theta C_{\max}}{\gamma}}\|y_n\|_2,
\end{aligned} \tag{40}$$

where (a) follows since $g_n, e^{-\frac{\theta}{\gamma}TQ_n} \in [\exp(-\theta C_{\max}/\gamma), \exp(\theta C_{\max}/\gamma)]^{SA}$, and on any interval $[c,d]$ with $c > 0$, the function $u \mapsto \ln(u)$ is $(1/c)$-Lipschitz.

Then, the update rule for (Q_n) gives us:

$$\begin{aligned}
\Delta_{n+1} &= \Delta_n + \alpha_n[TQ_n - Q_n] + \alpha_n\epsilon_n' \\
&\overset{(a)}{=} (1-\alpha_n)\Delta_n + \alpha_n[TQ_n - \Delta_n - Q_n] + \alpha_n\epsilon_n' \\
&\overset{(b)}{=} (1-\alpha_n)\Delta_n + \alpha_n[TQ_n - Q^*] + \alpha_n\epsilon_n',
\end{aligned}$$

where (a) is obtained by adding and subtracting $\alpha_n\Delta_n$, and (b) follows since $Q^* = TQ^*$. Now, taking $\|\cdot\|_\infty$ followed by expectation on both sides

$$\begin{aligned}
\mathbb{E}\|\Delta_{n+1}\|_\infty &\leq (I-\alpha_n)\mathbb{E}\|\Delta_n\|_\infty + \alpha_n\mathbb{E}\|TQ_n - TQ^*)\|_\infty + \alpha_n\mathbb{E}\|\epsilon_n'\|_\infty \\
&\overset{(a)}{\leq} (1-\alpha_n)\mathbb{E}\|\Delta_n\|_\infty + \gamma\alpha_n\mathbb{E}\|\Delta_n\|_\infty + \alpha_n\mathbb{E}\|\epsilon_n'\|_\infty \\
&\overset{(b)}{\leq} (1-\alpha_n)\|\Delta_n\|_\infty + \gamma\alpha_n\mathbb{E}\|\Delta_n\|_\infty + \frac{\gamma}{\theta}e^{\frac{\theta C_{\max}}{\gamma}}\alpha_n\mathbb{E}\|y_n\|_2 \\
&= (1-(1-\gamma)\alpha_n)\|\Delta_n\|_\infty + \frac{\gamma}{\theta}e^{\frac{\theta C_{\max}}{\gamma}}\alpha_n\sqrt{\mathbb{E}\|y_n\|_2^2} \\
&\overset{(c)}{\leq} (1-(1-\gamma)\alpha_n)\|\Delta_n\|_\infty + \frac{\gamma}{\theta}e^{\frac{\theta C_{\max}}{\gamma}}C_g\alpha_n\sqrt{\beta_n},
\end{aligned} \tag{41}$$

where (a) follows since T is a γ-contraction w.r.t $\|\cdot\|_\infty$, (b) follows from (40), and (c) uses (39). Iterating the above inequality gives

$$\begin{aligned}\mathbb{E}\|\Delta_N\|_\infty &\le \Gamma^{(\alpha)}_{0:N}\|\Delta_0\|_\infty + \frac{\gamma}{\theta}e^{\frac{\theta C_{\max}}{\gamma}}C_g\sum_{n=0}^{N-1}\Gamma^{(\alpha)}_{n+1:N}\alpha_n\sqrt{\beta_n}\\ &\overset{(a)}{\le} \|\Delta_0\|_\infty e^{-(1-\gamma)\sum_{n=0}^{N-1}\alpha_n} + \frac{\gamma}{\theta}e^{\frac{\theta C_{\max}}{\gamma}}C_g C^{0,1/2}_{\Gamma,\alpha}\sqrt{\beta_N}\\ &\overset{(b)}{\le} 2C_{\max}e^{\frac{(1-\gamma)}{(1-\alpha)}}e^{-\frac{(1-\gamma)}{(1-\alpha)}(N+1)^{1-\alpha}} + \frac{\gamma}{\theta}e^{\frac{\theta C_{\max}}{\gamma}}C_g C^{0,1/2}_{\Gamma,\alpha}\sqrt{\beta_N},\end{aligned}$$

where (a) follows from Lemma B.2, (b) follows since $\sum_{n=0}^{N-1}\alpha_n \ge [(N+1)^{1-\alpha}-1]/(1-\alpha)$. Therefore, for all $N \ge 0$,

$$\mathbb{E}\|\Delta_N\|_\infty \le C^{(1)}_Q e^{-\frac{1-\gamma}{1-\alpha}N^{1-\alpha}} + C^{(2)}_Q\sqrt{\beta_n},$$

where

$$C^{(1)}_Q := 2C_{\max}e^{\frac{(1-\gamma)}{(1-\alpha)}} \quad \text{and} \quad C^{(2)}_Q := \frac{\gamma}{\theta}e^{\frac{\theta C_{\max}}{\gamma}}C_g C^{0,1/2}_{\Gamma,\alpha}.$$

This completes the proof of Theorem 4.2. □

Proof of Lemma B.2. We prove the three statements one by one.

Statement 1: Recall that for $0 \le k < N$,

$$\Gamma^{(\beta)}_{k:N} := \prod_{n=k}^{N-1}\left(1-2\lambda\beta_n+3\beta_n^2\right), \quad \text{and} \quad \Gamma^{(\alpha)}_{k:N} := \prod_{n=k}^{N-1}(1-(1-\gamma)\alpha_n)$$

Using the fact that $1-x \le e^{-x}$, we have

$$\Gamma^{(\beta)}_{k:N} \le \prod_{n=k}^{N-1} e^{-2\lambda\beta_n+2\beta_n^2} = \left(e^{-2\lambda\sum_{n=k}^{N-1}\beta_n}\right)\cdot\left(e^{3\sum_{n=k}^{N-1}\beta_n^2}\right) \tag{42}$$

and

$$\Gamma^{(\alpha)}_{k:N} := \prod_{n=k}^{N-1}\exp\left(-(1-\gamma)\alpha_n\right) = \exp\left(-(1-\gamma)\sum_{n=k}^{N-1}\alpha_n\right). \tag{43}$$

Taking $\beta_n = (n+1)^{-\beta}$, with $\beta \in (1/2, 1)$ gives

$$\begin{aligned}\sum_{n=k}^{N-1}\beta_n^2 \le \sum_{n=0}^{\infty}\frac{1}{(n+1)^{2\beta}} &= 1+\sum_{n=1}^{\infty}\frac{1}{(n+1)^{2\beta}}\\ &\le 1+\sum_{n=1}^{\infty}\int_n^{n+1}\frac{1}{x^{2\beta}}dx = 1+\int_1^{\infty}\frac{1}{x^{2\beta}}dx = 1+\frac{1}{2\beta-1} = \frac{2\beta}{2\beta-1},\end{aligned} \tag{44}$$

Hence, setting $C_{\Gamma,\beta} := e^{\left(\frac{6\beta}{2\beta-1}\right)}$ gives

$$\Gamma^{(\beta)}_{k:N} \le C_{\Gamma,\beta}e^{-2\sum_{n=k}^{N-1}\beta_n} \quad \text{and} \quad \Gamma^{(\alpha)}_{k:N} \le e^{-(1-\gamma)\sum_{n=k}^{N-1}\alpha_n}. \tag{45}$$

This completes the proof of **Statement 1**. Now, we focus on the next set of bounds.

Statement 2: Note that, for $k \le n < N$,

$$\sum_{n=k}^{N-1}\beta_n f_n \cdot e^{-2\lambda\sum_{m=n+1}^{N-1}\beta_m} \le \left[\max_{k\le n<N} f_n\, e^{-\lambda\sum_{m=n+1}^{N-1}\beta_m}\right]\left(\sum_{n=k}^{N-1}\beta_n\, e^{-\lambda\sum_{m=n+1}^{N-1}\beta_m}\right).' \tag{46}$$

We bound these factors one by one. First, define $t_{-1} := 0$ and $t_n := \sum_{m=0}^{n} \beta_m$. Then, for $0 \le k < N$,

$$\sum_{n=k}^{N-1} \beta_n e^{-\lambda \sum_{m=n+1}^{N-1} \beta_m} \le \sum_{n=0}^{N-1} e^{-\lambda(t_{N-1}-t_n)}(t_n - t_{n-1}). \tag{47}$$

Since $0 \le \beta_n \le \beta_0$, for every $t \in [t_{n-1}, t_n]$ we have $t_n \le t + \beta_0 \le 1$. Thus, for $\lambda > 0$,

$$\begin{aligned}\sum_{n=0}^{N-1} e^{-\lambda(t_{N-1}-t_n)}(t_n - t_{n-1}) &\le e^{\lambda\beta_0} \sum_{n=0}^{N-1} \int_{t_{n-1}}^{t_n} e^{\lambda(t-t_{N-1})}\, dt \\ &= e^{\lambda\beta_0} \int_0^{t_{N-1}} e^{\lambda(t-t_{N-1})}\, dt \\ &= \frac{e^{\lambda\beta_0}}{\lambda}\left(1 - e^{-\lambda t_{N-1}}\right) \le \frac{e^{\lambda}}{\lambda}. \end{aligned} \tag{48}$$

Recall that $f_n = \alpha_n^a\, \beta_n^b = (n+1)^{-(a\alpha+b\beta)}$ and define for each $k \le n < N$:

$$h_n := f_n\, e^{-\lambda \sum_{m=n+1}^{N-1} \beta_m}. \tag{49}$$

Now, consider the following two cases:

Case 1: When $n \ge \lceil N/2 \rceil$. We have $n+1 \ge N/2 + 1 > (N+1)/2$. Thus,

$$\frac{h_n}{f_N} = \left(\frac{f_n}{f_N}\right) \cdot e^{-\lambda \sum_{m=n+1}^{N-1} \beta_m} \overset{(a)}{\le} \frac{f_n}{f_N} = \left(\frac{N+1}{n+1}\right)^{a\alpha+b\beta} < 2^{-(a\alpha+b\beta)}, \tag{50}$$

where (a) follows since $e^{-\lambda \sum_{m=n+1}^{N-1} \beta_m} \le 1$.

Case 2: When $n \le \lfloor N/2 \rfloor$. Here, we have

$$\sum_{m=n+1}^{N-1} \beta_m \ge \sum_{m=\lceil N/2 \rceil}^{N-1} \beta_m = \sum_{m=\lceil N/2 \rceil}^{N-1} \frac{1}{(m+1)^{\beta}} \ge \sum_{m=\lceil N/2 \rceil}^{N-1} \frac{1}{N^{\beta}} > \frac{N}{2N^{\beta}}. \tag{51}$$

Using the fact that $f_n \le 1$, we have

$$\begin{aligned}\frac{h_n}{f_N} = \frac{f_n}{f_N} \cdot e^{-\lambda \sum_{m=n+1}^{N-1} \beta_m} &\overset{(a)}{\le} \frac{1}{f_N} \cdot e^{-\lambda \sum_{m=n+1}^{N-1} \beta_m} = (N+1)^{a\alpha+b\beta} \cdot e^{-\lambda \sum_{m=n+1}^{N-1} \beta_m} \\ &\overset{(b)}{\le} (N+1)^{a\alpha+b\beta} \cdot e^{-\frac{\lambda}{2} N^{1-\beta}} \overset{(c)}{\le} (2N)^{a\alpha+b\beta} \cdot e^{-\frac{\lambda}{2} N^{1-\beta}}, \end{aligned} \tag{52}$$

where (a) follows since $f_n \le 1$, (b) follows from (51), and (c) follows since $N \ge 1$. Finally, we can show using calculus that for $p > 0$, the function $x \mapsto (2x)^p \cdot \exp(-\frac{\lambda}{2} x^{1-\beta})$ attains maximum at $x = \left(\frac{2p}{\lambda(1-\beta)}\right)^{1/(1-\beta)}$ to get

$$\begin{aligned}\frac{h_n}{f_N} &\le \sup_{x \in \mathbb{R}} \left[(2x)^{a\alpha+b\beta} \cdot e^{-\frac{\lambda}{2} x^{1-\beta}}\right] \\ &= (2x)^{a\alpha+b\beta} \cdot e^{-\frac{\lambda}{2} x^{1-\beta}}\Big|_{x=\left(\frac{2p}{\lambda(1-\beta)}\right)^{\frac{1}{1-\beta}}} = \left[\frac{4(a\alpha+b\beta)}{2^{\beta}\lambda e(1-\beta)}\right]^{\frac{a\alpha+b\beta}{1-\beta}}. \end{aligned} \tag{53}$$

Hence, for all $k \le n < N$, we have

$$h_n \le C_{a,b} f_N, \quad \text{where} \quad C_{a,b} := \max\left\{2^{-(a\alpha+b\beta)}, \left[\frac{4(a\alpha+b\beta)}{2^{\beta}\lambda e(1-\beta)}\right]^{\frac{a\alpha+b\beta}{1-\beta}}\right\}. \tag{54}$$

Defining $C^{a,b}_{\Gamma,\beta} := C_{a,b}C_{\Gamma,\beta}(e^{\lambda}/\lambda)$, we conclude that

$$\begin{aligned}\sum_{n=k}^{N-1}\beta_n f_n\,\Gamma^{(\beta)}_{n+1:N} &\overset{(a)}{\leq} C_{\Gamma,\beta}\left[\max_{k\leq n<N} f_n\, e^{-\lambda\sum_{m=n+1}^{N-1}\beta_m}\right]\left(\sum_{n=k}^{N-1}\beta_n\, e^{-\lambda\sum_{m=n+1}^{N-1}\beta_m}\right)\\ &\overset{(b)}{\leq} C_{\Gamma,\beta}\left[\max_{k\leq n<N} f_n\, e^{-\lambda\sum_{m=n+1}^{N-1}\beta_m}\right]\left(\frac{e^{\lambda}}{\lambda}\right)\\ &\overset{(c)}{\leq} C_{\Gamma,\beta}\cdot\left[\max_{k\leq n<N} C_{a,b}f_N\right]\cdot\left(\frac{e^{\lambda}}{\lambda}\right) = C^{a,b}_{\Gamma,\beta}f_N, \end{aligned} \tag{55}$$

where (a) is obtained by combining from (45) and (46), (b) follows by combining (47) and (48), and (c) follows from (54).

To finish the proof of **Statement 2**, we write

$$\begin{aligned}\sum_{n=k}^{N-1}\alpha_n f_n\cdot e^{-(1-\gamma)\sum_{m=n+1}^{N-1}\alpha_m} \leq \left[\max_{k\leq n<N} f_n\, e^{-\frac{(1-\gamma)}{2}\sum_{m=n+1}^{N-1}\alpha_m}\right]\\ \times\left(\sum_{n=k}^{N-1}\alpha_n\, e^{-\frac{(1-\gamma)}{2}\sum_{m=n+1}^{N-1}\alpha_m}\right).\end{aligned} \tag{56}$$

Treating the second factor in the above expression gives

$$\begin{aligned}\sum_{n=k}^{N-1}\alpha_n\, e^{-\frac{1-\gamma}{2}\sum_{m=n+1}^{N-1}\alpha_m} &= \sum_{n=k}^{N-1} e^{-\frac{1-\gamma}{2}(t_{N-1}-t_n)}(t_n-t_{n-1})\\ &\leq e^{\frac{1-\gamma}{2}\alpha_0}\int_{t_{k-1}}^{t_{N-1}} e^{\frac{1-\gamma}{2}(t-t_{N-1})}\,dt\\ &= \frac{2e^{\frac{1-\gamma}{2}}}{1-\gamma}\left(1-e^{-\frac{1-\gamma}{2}(t_{N-1}-t_{k-1})}\right)\leq\frac{2e^{\frac{1-\gamma}{2}}}{1-\gamma}.\end{aligned} \tag{57}$$

where we let $t_n := \sum_{m=0}^{n}\alpha_m$ and $t_{-1}=0$.

Now, for the second factor, we define $\bar{h}_n := f_n\, e^{-\frac{(1-\gamma)}{2}\sum_{m=n+1}^{N-1}\alpha_m}$, and claim that there exists a constant $C^{a,b}_{\Gamma,\alpha}>0$, such that for all $k\leq n<N$, $\bar{h}_n \leq C^{a,b}_{\Gamma,\alpha}\, f_N$.

Using a similar analysis as earlier, we can show that

$$\frac{\bar{h}_n}{f_N}\leq\begin{cases}\sup_{x\in\mathbb{R}}\left[(2x)^{a\alpha+b\beta}\cdot e^{\frac{(1-\gamma)}{2}x^{1-\alpha}}\right] & ,\text{when } n\leq\lfloor N/2\rfloor\\ 2^{-(a\alpha+b\beta)} & ,\text{when } n\geq\lceil N/2\rceil\end{cases}. \tag{58}$$

Since the function $x\mapsto(2x)^{a\alpha+b\beta}\cdot e^{\frac{(1-\gamma)}{2}x^{1-\alpha}}$ attains global maximum at $x=\left[\frac{2(a\alpha+b\beta)}{(1-\gamma)(1-\alpha)}\right]^{\frac{1}{1-\alpha}}$, we conclude that

$$\max_{k\leq n\leq N}\bar{h}_n\leq f_N\cdot\max\left\{2^{-(a\alpha+b\beta)},\left[\frac{4(a\alpha+b\beta)}{2^{\alpha}e(1-\gamma)(1-\alpha)}\right]^{\left(\frac{a\alpha+b\beta}{1-\alpha}\right)}\right\}. \tag{59}$$

Thus, combining (45), (56), (57), and (59) gives

$$\sum_{n=k}^{N-1}\alpha_n f_n\,\Gamma^{(\alpha)}_{n+1:N}\leq C^{a,b}_{\Gamma,\alpha}\, f_N, \tag{60}$$

where

$$C^{a,b}_{\Gamma,\alpha} := \frac{2e^{\frac{1-\gamma}{2}}}{1-\gamma}\cdot\max\left\{2^{-(a\alpha+b\beta)},\left[\frac{4(a\alpha+b\beta)}{2^{\alpha}e(1-\gamma)(1-\alpha)}\right]^{\left(\frac{a\alpha+b\beta}{1-\alpha}\right)}\right\}.$$

This completes the proof. □

C Analysis for the One-Timescale Fixed-Point Method

We discuss the stability, asymptotic convergence, and the convergence rates of the one-timescale fixed-point method in (9).

C.1 Stability

We first show that the sequence (x_n) remains within the compact set

$$\mathcal{K} := [C_\ell, C_u]^{SA}, \tag{61}$$

where C_ℓ and C_u are as in (11). Since $C_\ell > 0$, $\mathcal{K}$ lies in the interior of the positive orthant $\mathbb{R}^{SA}_{++}$.

Lemma C.1. *Let $x_0 \in \mathcal{K}$. Then, $x_n \in \mathcal{K}$ and $\hat{F}_n \in [C_\ell, C_u]$ for all $n \geq 0$.*

Proof. We use induction to first prove that $x_n(s,a) \geq C_\ell$ for all s, a and all $n \geq 0$. It trivially follows that $x_0(s,a) \geq C_\ell$ for all s, a. Now, for some $n \geq 0$, suppose that $x_n(s,a) \geq C_\ell$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$. Then, for any s, a, and s',

$$\begin{aligned}\exp\left(-\frac{\theta}{\gamma} r(s,a) + \gamma \min_{a'} \ln x_n(s',a')\right) &\overset{\text{(a)}}{\geq} \exp\left(-\frac{\theta}{\gamma}\|r\|_\infty + \gamma \ln C_\ell\right) \\ &\overset{\text{(b)}}{\geq} C_\ell,\end{aligned} \tag{62}$$

where (a) follows since $r(s,a) \leq \|r\|_\infty$ and $x_n(s,a) \geq C_\ell$ for all s, a, while (b) holds since the definition of C_ℓ implies $(1-\gamma)\ln C_\ell = -\frac{\theta}{\gamma}\|r\|_\infty$. Using this last inequality, it follows from the update rule in (9) that $x_{n+1}(s,a) \geq C_\ell$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$, completing the induction proof. The same argument also show that $\hat{F}_n \geq C_\ell$ for all $n \geq 0$.

A similar induction argument shows that $x_n(s,a) \leq C_u$ for all s, a and $\hat{F}_n \leq C_u$ for all $n \geq 0$. The overall claim now follows. □

Building on the above result, it is easy to see that F maps $\mathcal{K}$ to itself.

Corollary C.2. *Let K be defined as in* (61)*. Then, $x \in \mathcal{K} \implies F(x) \in \mathcal{K}$.*

Proof. The definition of F given in (1) shows that

$$F(x)(s,a) = \mathbb{E}_{s' \in \mathcal{P}(\cdot|s,a)}\left[\exp\left(-\frac{\theta}{\gamma} r(s,a) + \gamma \min_{a'} \ln x_n(s',a')\right)\right].$$

Hence, by arguing as in the proof of Lemma C.1, the desired result is easy to see. □

C.2 Asymptotic Convergence

We now discuss (9)'s asymptotic behavior. Let $(\mathcal{F}_n)_{n\geq 0}$ be the filtration of σ-fields, where

$$\mathcal{F}_n := \sigma(x_0, s_0, a_0, \ldots, s_n, a_n).$$

The update in (9) can then be rewritten as

$$x_{n+1} = x_n + \alpha_n \left[e_{s_n,a_n}\Big((F(x_n))_{s_n,a_n} - x_n(s_n,a_n)\Big) + M_{n+1}\right], \tag{63}$$

where

$$M_{n+1} = e_{s_n,a_n}\Big(\hat{F}_n - (F(x_n))_{s_n,a_n}\Big). \tag{64}$$

Then, it is straightforward to see that the associated limiting ODE is given by

$$\dot{x}(t) = D_{\pi_{x(t)}}[F(x(t)) - x(t)], \tag{65}$$

where D_{π_x} is the diagonal matrix made up the stationary distribution π_x (defined below (11)). Let

$$\ell_{\mathcal{K}} := \min_{x\in\mathcal{K}} \min_{i \in \mathcal{S}\times\mathcal{A}} D_{\pi_x}(i,i) \quad \text{and} \quad u_{\mathcal{K}} := \max_{x\in\mathcal{K}} \max_{i \in \mathcal{S}\times\mathcal{A}} D_{\pi_x}(i,i). \tag{66}$$

Since $\mathcal{K}$ is compact and bounded away from 0, it follows that

$$0 < \ell_{\mathcal{K}} \leq u_{\mathcal{K}} < \infty. \tag{67}$$

For (65) to be well-posed—i.e., to ensure existence and uniqueness of solutions on $[0, \infty)$, along with continuous dependence on initial conditions—standard results typically require either global Lipschitz continuity of the driving function or, more generally, local Lipschitz continuity together with the property that solution trajectories remain within a compact set.

In our setting, F involves terms of the form $[\min_{s,a} x(s', a')]^\gamma$ with $\gamma \in [0, 1)$, and hence is not globally Lipschitz; moreover, it fails to be locally Lipschitz near the boundary of $\mathbb{R}^{SA}_{++}$. We therefore adopt the second approach to show (65)'s well-posedness: we show that F is locally Lipschitz on a neighbourhood of $\mathcal{K}$ in $\mathbb{R}^{SA}_{++}$ and that $\mathcal{K}$ is positively invariant under the induced dynamics.

Lemma C.3. *The function F is Lipschitz continuous on an open neighbourhood of $\mathcal{K}$ in $\mathbb{R}^{SA}_{++}$.*

Proof. Let $\delta := C_\ell/2 > 0$; the inequality holds since $C_\ell > 0$. Then,

$$\mathcal{K} \subseteq \mathcal{K}^\delta := (C_\ell - \delta,\, C_u + \delta)^{SA} \subset \mathbb{R}^{SA}_{++}.$$

It suffices to show that F is Lipschitz continuous on $\mathcal{K}^\delta$.

Let $h : \mathbb{R} \to \mathbb{R}$ be given by $h(z) = z^\gamma$. Then, for $z_1, z_2 \geq C_\ell/2 > 0$, there exists some c between z_1 and z_2 such that

$$\begin{aligned} |h(z_1) - h(z_2)| &\overset{\text{(a)}}{=} |h'(c)||z_1 - z_2|, \\ &\overset{\text{(b)}}{=} \frac{\gamma}{c^{1-\gamma}}|z_1 - z_2| \\ &\leq \frac{2^{1-\gamma}\gamma}{C_\ell^{1-\gamma}}|z_1 - z_2|, \end{aligned} \tag{68}$$

where (a) follows from the mean value theorem, while (b) follows since $c \geq \min\{z_1, z_2\} \geq C_\ell/2$.

Therefore, for any $(s, a) \in \mathcal{S} \times \mathcal{A}$,

$$\begin{aligned} &|(F(x_1)(s,a) - F(x_2)(s,a)| \\ &\quad\overset{\text{(a)}}{\leq} \sum_{s' \in \mathcal{S}} \mathcal{P}(s'|s,a) \exp\left(-\frac{\theta}{\gamma} r(s,a)\right) \left| \Big[\min_{a'} x_1(s', a')\Big]^\gamma - \Big[\min_{a'} x_2(s', a')\Big]^\gamma \right| && (69) \\ &\quad\overset{\text{(b)}}{\leq} \frac{2^{1-\gamma}\gamma}{C_\ell^{1-\gamma}} \sum_{s' \in \mathcal{S}} \mathcal{P}(s'|s,a) \exp\left(-\frac{\theta}{\gamma} r(s,a)\right) \left| \min_{a'} x_1(s', a') - \min_{a'} x_2(s', a') \right| && (70) \\ &\quad\overset{\text{(c)}}{\leq} \frac{2^{1-\gamma}\gamma}{C_\ell^{1-\gamma}} \exp\left(\frac{\theta}{\gamma}\|r\|_\infty\right) \|x_1 - x_2\|_\infty, && (71) \end{aligned}$$

where (a) follows from (1), (b) follows from (68), while (c) holds from since $r(s, a) \geq -\|r\|_\infty$ and since a standard argument shows

$$\left| \min_{a'} x_1(s', a') - \min_{a'} x_2(s', a') \right| \leq \|x_1 - x_2\|_\infty$$

Taking the supremum over (s, a) yields

$$\|F(x_1) - F(x_2)\|_\infty \leq \frac{\gamma}{C_\ell^{1-\gamma}} \exp\left(\frac{\theta}{\gamma}\|r\|_\infty\right) \|x_1 - x_2\|_\infty, \tag{72}$$

which proves that F is Lipschitz continuous on $\mathcal{K}^\delta$, as desired. □

Our next result establishes that $\mathcal{K}$ is positively invariant for the limiting ODE in (65) and that x^* is globally asymptotically stable relative to $\mathcal{K}$.

Theorem C.4. *The set $\mathcal{K}$ defined in* (61) *contains the fixed point x^* of F and is positively invariant under the ODE in* (65)*. Moreover, x^* is a globally asymptotically stable equilibrium of* (65) *relative to $\mathcal{K}$; that is, for every initial condition $x(0) \in \mathcal{K}$, the corresponding trajectory satisfies*

$$x(t) \to x^* \quad \textit{as} \quad t \to \infty.$$

Proof. We first show that $x^* \in \mathcal{K}$. Clearly, $\mathcal{K}$ is a non-empty convex compact subset of $\mathbb{R}^{SA}_{++}$. Moreover, F is continuous and Lemma C.2 implies that $F(\mathcal{K}) \subseteq \mathcal{K}$. It thus follows from Brouwer's fixed-point theorem that F has a fixed point in $\mathcal{K}$. Since x^* is the unique fixed point of F on $\mathbb{R}^{SA}_{++}$, it follows that $x^* \in \mathcal{K}$, as desired.

Next, we show that $\mathcal{K}$ is positively invariant under the ODE in (65). Under Assumption 2, the stationary distribution π_x depends Lipschitz continuously on $x \in \mathcal{K}$, since the state-action space is finite and the kernels K_x are irreducible. Together with Lemma C.3, this implies that the vector field

$$f(x) := D_{\pi_x}[F(x) - x]$$

is Lipschitz continuous on $\mathcal{K}$ and admits a locally Lipschitz extension to a neighborhood of $\mathcal{K}$. Hence, the associated ODE has unique solutions.

Now, let $x \in \mathcal{K}$. If $x_i = C_\ell$, then

$$f_i(x) = D_{\pi_x}(i,i)[F_i(x) - C_\ell] \geq 0,$$

since $F(x) \in \mathcal{K}$. Similarly, if $x_i = C_u$, then $f_i(x) \leq 0$. These inequalities show that $f(x)$ belongs to the tangent cone of $\mathcal{K}$ at every $x \in \mathcal{K}$. Therefore, by Nagumo's invariance theorem [Aubin et al., 2011, Theorem 11.2.3], together with uniqueness of solutions, $\mathcal{K}$ is positively invariant under the ODE in (65). Since $\mathcal{K}$ is compact, it then follows that all solutions starting in $\mathcal{K}$ exist for all $t \geq 0$.

By Proposition 2.1, the operator F is monotone and homogeneous of degree γ, i.e., for all $x, y \in \mathbb{R}^{SA}_{++}$ and $c > 0$,

$$x \leq y \implies F(x) \leq F(y), \qquad F(cx) = c^\gamma F(x).$$

We use these properties to establish the global asymptotic stability of x^*.

Let x be any solution trajectory of (65), starting in $\mathcal{K}$. Since $x^* \in \mathcal{K}$, the quantities

$$m(t) := \min_{i \in \mathcal{S}\times\mathcal{A}} \frac{x_i(t)}{x_i^*} \text{ and } M(t) := \max_{i \in \mathcal{S}\times\mathcal{A}} \frac{x_i(t)}{x_i^*}$$

are well defined for any $t \geq 0$. These definitions imply

$$m(t)x^* \leq x(t) \leq M(t)x^*. \tag{73}$$

To show $\lim_{t\to\infty} x(t) = x^*$, it then suffices to show that

$$1 \leq \liminf_{t\to\infty} m(t) \leq \limsup_{t\to\infty} M(t) \leq 1. \tag{74}$$

We first show that $\liminf_{t\to\infty} m(t) \geq 1$. By the homogeneity and monotonicity of F,

$$m(t)^\gamma F(x^*) = F(m(t)x^*) \leq F(x(t)). \tag{75}$$

Let $D_+m(t)$ denote the lower right Dini derivative of $m(t)$, defined by

$$D_+m(t) := \liminf_{h\downarrow 0} \frac{m(t+h) - m(t)}{h},$$

and let

$$I(t) := \arg\min_{i \in \mathcal{S}\times\mathcal{A}} \frac{x_i(t)}{x_i^*}.$$

Then, by Danskin's theorem [Danskin, 1966, Theorem 1] (cf. [Davydov, 2025, Lemma 2.2.6]) and the ODE dynamics,

$$\begin{aligned}
D_+m(t) &= \min_{i\in I(t)} \frac{\dot{x}_i(t)}{x_i^*} \\
&= \min_{i\in I(t)} D_{\pi_{x(t)}}(i,i) \frac{F_i(x(t)) - x_i(t)}{x_i^*} \\
&\geq \min_{i\in I(t)} D_{\pi_{x(t)}}(i,i) \frac{m(t)^\gamma F_i(x^*) - x_i(t)}{x_i^*} \\
&= \min_{i\in I(t)} D_{\pi_{x(t)}}(i,i)[m(t)^\gamma - m(t)],
\end{aligned}$$

where the inequality follows from (75), and the last equality uses $F(x^*) = x^*$ and $x_i(t) = m(t)x_i^*$ for every $i \in I(t)$.

Since $0 < \ell_{\mathcal{K}} \le D_{\pi_{x(t)}}(i,i) \le u_{\mathcal{K}}$, we distinguish two cases. If $m(t) \le 1$, then $m(t)^\gamma - m(t) \ge 0$, and hence

$$D_+ m(t) \ge \ell_{\mathcal{K}}[m(t)^\gamma - m(t)].$$

If $m(t) > 1$, then $m(t)^\gamma - m(t) < 0$, so

$$D_+ m(t) \ge u_{\mathcal{K}}[m(t)^\gamma - m(t)].$$

Combining the two cases gives

$$D_+ m(t) \ge \begin{cases} \ell_{\mathcal{K}}[m(t)^\gamma - m(t)], & m(t) \le 1, \\ u_{\mathcal{K}}[m(t)^\gamma - m(t)], & m(t) > 1. \end{cases} \tag{76}$$

Next, let $z : [0,\infty) \to \mathbb{R}_{++}$ solve

$$\dot{z}(t) = \begin{cases} \ell_{\mathcal{K}}[z(t)^\gamma - z(t)], & z(t) \le 1, \\ u_{\mathcal{K}}[z(t)^\gamma - z(t)], & z(t) > 1, \end{cases}$$

with $z(0) = m(0) > 0$. Since $\gamma \in (0,1)$ and $0 < \ell_{\mathcal{K}} \le u_{\mathcal{K}}$, the right-hand side is locally Lipschitz on $\mathbb{R}_{++}$, positive for $z < 1$, and negative for $z > 1$. Consequently, 1 is the unique globally asymptotically stable equilibrium on $\mathbb{R}_{++}$. Applying the Dini comparison lemma[1] to (76) yields $m(t) \ge z(t)$ for all $t \ge 0$. Since $z(t) \to 1$, we conclude that

$$\liminf_{t\to\infty} m(t) \ge 1.$$

A symmetrical argument using the upper right Dini derivative of $M(t)$ shows $\limsup M(t) \le 1$. Separately, $m(t) \le M(t)$ implies $\liminf_{t\to\infty} m(t) \le \limsup_{t\to\infty} M(t)$. This completes the proof of (74) and, hence, of the claim that $\lim_{t\to\infty} x(t) = x^*$.

It remains to establish Lyapunov stability of x^*. The preceding scalar comparison arguments also imply that

$$m(t) \ge \min\{m(0), 1\}, \qquad M(t) \le \max\{M(0), 1\}, \quad t \ge 0.$$

Consequently, letting

$$\delta_0 := \max_i \left| \frac{x_i(0)}{x_i^*} - 1 \right|,$$

we obtain

$$(1 - \delta_0)x^* \le x(t) \le (1 + \delta_0)x^*, \qquad t \ge 0.$$

It follows that

$$\|x(t) - x^*\|_\infty \le \frac{\|x^*\|_\infty}{\min_i x_i^*} \|x(0) - x^*\|_\infty, \qquad t \ge 0.$$

Since x^* has strictly positive coordinates, this establishes Lyapunov stability. Combined with the global attractivity established above, it follows that x^* is globally asymptotically stable relative to $\mathcal{K}$.

Since the solution trajectory x was arbitrary, it follows that x^* is a globally asymptotically stable equilibrium of (65) relative to $\mathcal{K}$, as desired. □

Since F is only locally Lipschitz (Lemma 6), we now show how the ODE method from [Borkar, 2008] can be adapted to prove Theorem 4.3.

Proof of Theorem 4.3. We verify the standard conditions required for the ODE method.

Lemma C.3 shows that F is Lipschitz continuous on an open neighbourhood of $\mathcal{K}$. In particular, this ensures that the limiting ODE (65) is well-posed. Further, by Lemma C.1 and Theorem C.4, both the iterate sequence (x_n) and the solution trajectories of (65) remain confined to $\mathcal{K}$. The stepsize sequence (α_n) satisfies the Robbins–Monro conditions.

[1]The result in Khalil [2002, Lemma 3.4] is stated for the upper right Dini derivative. The corresponding result for the lower right Dini derivative follows by negating the functions.

It remains to verify the properties of the noise sequence (M_n) defined in (64). By construction,

$$\mathbb{E}[M_{n+1} \mid \mathcal{F}_n] \overset{a.s.}{=} 0, \quad n \geq 0,$$

so (M_n) is a martingale difference sequence with respect to the filtration $(\mathcal{F}_n)$. Moreover, since $\hat{F}_n, F(x_n) \in \mathcal{K}$ (see Lemma C.1 and Corollary C.2), we have

$$\|M_{n+1}\|_2 \leq C_u - C_\ell,$$

where C_ℓ and C_u are defined in (11). Consequently,

$$\mathbb{E}\big[\|M_{n+1}\|_2^2 \mid \mathcal{F}_n\big] \leq (C_u - C_\ell)^2.$$

Thus, all the conditions of the ODE method are satisfied. It follows that (x_n) tracks the solution trajectories of the limiting ODE (65). Since x^* is the unique globally asymptotically stable equilibrium of (65) relative to $\mathcal{K}$, we conclude that $x_n \to x^*$ almost surely. □

C.3 Convergence Rates: Scalar Case

Here, we analyze the convergence rate of a scalar stochastic approximation model associated with the one-timescale update in (9). The aim is to isolate the challenges posed by the power-law structure.

We begin by considering the one-state, one-action reduction of the function F defined in (1). Let the corresponding reward be denoted by $r \in \mathbb{R}$. Then, $F : \mathbb{R}_{++} \to \mathbb{R}_{++}$ takes the form

$$F(x) = \exp\left(-\frac{\theta}{\gamma} r\right) x^\gamma.$$

Solving $F(x^*) = x^*$ yields

$$x^* = \exp\left(-\frac{\theta}{\gamma(1-\gamma)} r\right),$$

so that

$$F(x) = \left(\frac{x}{x^*}\right)^\gamma x^*. \tag{77}$$

In the literal single-state, single-action model with deterministic rewards, the update in (63) becomes deterministic. To study the effect of stochastic approximation noise while preserving the same power-law drift, we consider the following scalar recursion related to 9:

$$x_{n+1} = x_n + \alpha_n \left[F(x_n) - x_n + \zeta_{n+1}\right], \tag{78}$$

where $(\zeta_n)_{n\geq 1}$ is a martingale-difference sequence with respect to the filtration $(\mathcal{F}_n)_{n\geq 0}$, defined by

$$\mathcal{F}_n := \sigma(x_0, \zeta_1, \ldots, \zeta_n). \tag{79}$$

For this new scalar update, we presume that there exists constants C_ℓ and C_u, similar to (11) such that $0 < C_\ell \leq C_u$ and $x^* \in \mathcal{K} := [C_\ell, C_u]$. As in Lemma C.1, we assume that $x_0 \in \mathcal{K}$ implies $x_n \in \mathcal{K}$ for all $n \geq 0$. We also assume that the martingale differences satisfy

$$|\zeta_n| \leq C_u - C_\ell, \qquad n \geq 0. \tag{80}$$

For the convergence and convergence-rate analysis of (7), the global contraction of T plays a crucial role. In particular, it allows us to use $\|Q(t) - Q^*\|_\infty$ as a global Lyapunov function in (28) for the limiting ODE associated with the Q_n-update in (7). At the discrete-time level, the same contraction property yields the strict shrinkage factor in (41), which provides a negative drift toward Q^* and thereby enables the convergence-rate analysis.

In contrast, the power-law form of F in (77) implies that $F'(x) \to \infty$ as $x \downarrow 0$. Thus, F is contractive only on the subset of $\mathbb{R}_{++}$ that is bounded away from the origin. Indeed, for any $\epsilon > 0$ such that $\gamma + \epsilon < 1$, there exists $\delta > 0$ such that

$$|F(x) - F(x')| \leq (\gamma + \epsilon)|x - x'|, \qquad x, x' \geq x^* - \delta.$$

This lack of global contractivity is the main obstacle in obtaining convergence rates for our one-timescale method. We now explain how this can be overcome in the scalar case.

Theorem C.5. *Let*

$$\underline{y} := \frac{C_\ell}{x^*}, \qquad \overline{y} := \frac{C_u}{x^*}.$$

Define

$$C_1 := \begin{cases} \dfrac{\underline{y} - \underline{y}^\gamma}{\underline{y} - 1}, & \underline{y} \neq 1, \\ 1 - \gamma, & \underline{y} = 1. \end{cases}$$

Let

$$\alpha_n = \frac{1}{2C_1(n+1)}.$$

Then

$$\mathbb{E}\left|\frac{x_n}{x^*} - 1\right| \leq \sqrt{\frac{\widetilde{C}_2(1+\ln n)}{n}} = \tilde{O}(n^{-1/2}),$$

where

$$\widetilde{C}_2 := \frac{1}{4C_1^2}\left[\max\left\{\left(\frac{C_u}{x^*}\right)^{2\gamma}, \left(\frac{C_u}{x^*}\right)^2\right\} + \left(\frac{C_u}{x^*}\right)^2\right],$$

and $\tilde{O}$ hides logarithmic terms.

Proof. Let $y_n := \frac{x_n}{x^*}$. Since $0 < C_\ell \leq x^* \leq C_u < \infty$, we have

$$0 < \underline{y} \leq y_n \leq \overline{y} < \infty, \qquad n \geq 0. \tag{81}$$

Dividing (78) by x^* and using (77), we obtain

$$\begin{aligned} y_{n+1} &= y_n + \alpha_n\left[\frac{F(x_n)}{x^*} - y_n + \frac{\zeta_{n+1}}{x^*}\right] \\ &= y_n + \alpha_n\left[y_n^\gamma - y_n + \frac{\zeta_{n+1}}{x^*}\right]. \end{aligned} \tag{82}$$

Let

$$z_n := y_n - 1.$$

Then

$$z_{n+1} = z_n + \alpha_n(y_n^\gamma - y_n) + \alpha_n\frac{\zeta_{n+1}}{x^*}.$$

Squaring both sides, taking conditional expectation with respect to $\mathcal{F}_n$, and using $\mathbb{E}[\zeta_{n+1} \mid \mathcal{F}_n] = 0$, we get

$$\mathbb{E}[z_{n+1}^2 \mid \mathcal{F}_n] = z_n^2 + 2\alpha_n z_n(y_n^\gamma - y_n) + \alpha_n^2(y_n^\gamma - y_n)^2 + \alpha_n^2\mathbb{E}\left[\left(\frac{\zeta_{n+1}}{x^*}\right)^2\middle|\mathcal{F}_n\right]. \tag{83}$$

From (80), we have

$$\left(\frac{\zeta_{n+1}}{x^*}\right)^2 \leq \left(\frac{C_u - C_\ell}{x^*}\right)^2 \leq \left(\frac{C_u}{x^*}\right)^2.$$

Also, using (81),

$$|y_n^\gamma - y_n| \leq \max\{y_n^\gamma, y_n\} \leq \max\left\{\left(\frac{C_u}{x^*}\right)^\gamma, \frac{C_u}{x^*}\right\}.$$

Therefore, with

$$C_2 := \max\left\{\left(\frac{C_u}{x^*}\right)^{2\gamma}, \left(\frac{C_u}{x^*}\right)^2\right\} + \left(\frac{C_u}{x^*}\right)^2,$$

we obtain

$$\mathbb{E}[z_{n+1}^2 \mid \mathcal{F}_n] \leq z_n^2 + 2\alpha_n z_n(y_n^\gamma - y_n) + C_2\alpha_n^2. \tag{84}$$

We now establish a uniform drift bound. Define

$$h(y) := \frac{y - y^\gamma}{y - 1}, \qquad y > 0,\ y \neq 1,$$

with the continuous extension $h(1) := 1 - \gamma$. We first show that h is nondecreasing on $\mathbb{R}_{++}$. For $y > 0$ and $y \neq 1$,

$$(y-1)^2 h'(y) = (1-\gamma)y^\gamma + \gamma y^{\gamma-1} - 1.$$

Multiplying by $y^{1-\gamma} > 0$, define

$$g(y) := y^{1-\gamma}(y-1)^2 h'(y) = (1-\gamma)y + \gamma - y^{1-\gamma}.$$

Then

$$g'(y) = (1-\gamma) - (1-\gamma)y^{-\gamma}, \qquad g''(y) = \gamma(1-\gamma)y^{-(1+\gamma)} \geq 0.$$

Thus g is convex on $(0, \infty)$. Since

$$g(1) = 0, \qquad g'(1) = 0,$$

it follows that g attains its global minimum at 1, and hence $g(y) \geq 0$ for all $y > 0$. Therefore, $h'(y) \geq 0$ for all $y > 0$, $y \neq 1$, and so h is nondecreasing on $(0, \infty)$.

Since h is nondecreasing and $y_n \geq \underline{y}$, we have $h(y_n) \geq C_1$. Moreover,

$$y_n - y_n^\gamma = h(y_n)(y_n - 1).$$

Hence

$$(y_n - 1)(y_n^\gamma - y_n) = -h(y_n)(y_n - 1)^2 \leq -C_1(y_n - 1)^2 = -C_1 z_n^2. \tag{85}$$

Substituting (85) into (84) gives

$$\mathbb{E}[z_{n+1}^2 \mid \mathcal{F}_n] \leq z_n^2 - 2C_1\alpha_n z_n^2 + C_2\alpha_n^2.$$

Taking expectation on both sides yields

$$\mathbb{E}z_{n+1}^2 \leq (1 - 2C_1\alpha_n)\mathbb{E}z_n^2 + C_2\alpha_n^2.$$

Recalling that

$$\alpha_n = \frac{1}{2C_1(n+1)},$$

we get

$$\mathbb{E}z_{n+1}^2 \leq \frac{n}{n+1}\mathbb{E}z_n^2 + \frac{\widetilde{C}_2}{(n+1)^2}. \tag{86}$$

Multiplying both sides of (86) by $n + 1$ gives

$$(n+1)\mathbb{E}z_{n+1}^2 \leq n\mathbb{E}z_n^2 + \frac{\widetilde{C}_2}{n+1}.$$

Iterating the above inequality from 0 to $n - 1$, and using that the initial term is killed by the factor $1 - 2C_1\alpha_0 = 0$, gives

$$n\mathbb{E}z_n^2 \leq \widetilde{C}_2 \sum_{k=1}^{n} \frac{1}{k}.$$

Therefore,

$$\mathbb{E}z_n^2 \leq \widetilde{C}_2 \frac{1 + \ln n}{n}.$$

Finally, by Jensen's inequality,

$$\mathbb{E}\left|\frac{x_n}{x^*} - 1\right| = \mathbb{E}|z_n| \leq \sqrt{\mathbb{E}z_n^2} \leq \sqrt{\widetilde{C}_2 \frac{1 + \ln n}{n}}.$$

This completes the proof. □

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: All the results derived are in line with our abstract and introduction.

   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: In Section 6, we mention that the rate results require a generative model and the single timescale rate is derived for the scalar case.

   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: Yes, we detail the assumptions made either before stating the theorem or within its statement, and provide complete proofs in the appendix.

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: [N/A]

   Justification: The contributions in this paper are purely theoretical.

   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: [N/A]

Justification: This paper does not include experiments.

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: [N/A]

   Justification: This paper does not include experiments.

   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: [N/A]

   Justification: This paper does not include experiments.

   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: [N/A]

   Justification: This paper does not include experiments.

   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 contributions are primarily theoretical and do not involve experiments, data, or human subjects and, hence, aligns 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: The contributions are primarily theoretical. We do not foresee a strong societal impact that is worth pondering about.

    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: Being theoretical, the work presented in this paper does not pose such risks.

    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: This paper reports theoretical results and hence does not use existing assets.

    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: This paper does not release new assets.

    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 paper does not involve crowdsourcing nor research with human subjects.

    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 paper does not involve crowdsourcing nor research with human subjects.

    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: The work presented in this paper does not involve LLMs as an important or non-trivial component.

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.