ACTOR-CRITIC IS IMPLICITLY BIASED TOWARDS HIGH ENTROPY OPTIMAL POLICIES

Anonymous authors
Paper under double-blind review

ABSTRACT

We show that the simplest actor-critic method — a linear softmax policy updated with TD through interaction with a linear MDP, but featuring no explicit regularization or exploration — does not merely find an optimal policy, but moreover prefers high entropy optimal policies. To demonstrate the strength of this bias, the algorithm not only has no regularization, no projections, and no exploration like ϵ-greedy, but is moreover trained on a single trajectory with no resets. The key consequence of the high entropy bias is that uniform mixing assumptions on the MDP, which exist in some form in all prior work, can be dropped: the implicit regularization of the high entropy bias is enough to ensure that all chains mix and an optimal policy is reached with high probability. As an auxiliary contribution, this work decouples concerns between the actor and critic by writing the actor update as an explicit mirror descent, and provides tools to analyze the critic's estimation in similar high entropy settings.

1 OVERVIEW

Reinforcement learning methods navigate an environment and seek to maximize their reward (Sutton & Barto, 2018). A key tension is the tradeoff between *exploration* and *exploitation*: does a learner (also called an *agent* or *policy*) *explore* for new high-reward states, or does it *exploit* the best states it has already found? This is a sensitive part of RL algorithm design, as it is easy for methods to become blind to parts of the state space; to combat this, many methods have an explicit exploration component, for instance the *ϵ-greedy method*, which forces exploration in all states with probability ϵ (Sutton & Barto, 2018; Tokic, 2010). Similarly, many methods must use projections and regularization to smooth their estimates (Williams & Peng, 1991; Mnih et al., 2016; Cen et al., 2020).

This work considers actor-critic methods, where a policy (or *actor*) is updated via the suggestions of a *critic*. In this setting, typically both the actor and the critic are regularized. For instance, in a recent work trained on a single trajectory just like the present work, there is an explicit *ϵ-greedy* component promoting exploration (Khodadadian et al., 2021). As further examples, TD methods in our linear MDP setting typically need projection steps (Bhandari et al., 2018), while neural actor-critic methods need many assumptions on mixing, the path the algorithm follows, the stability of all estimates, and so on (Wang et al., 2019).

Contribution. This work shows that a simple linear actor-critic (cf. Algorithm 1) in a linear MDP (cf. Assumption 1.3) with a finite but non-tabular state space (cf. Assumption 1.1) finds an ϵ-optimal policy in $\mathrm{poly}(1/\epsilon)$ steps with probability at least $1-\mathcal{O}(\epsilon)$, without any mixing assumptions on the underlying MDP (cf. Theorem 1.4). Instead, the only assumption is that there exists a maximum entropy optimal policy which mixes; the analysis shows that the actor-critic policy path never leaves a Kullback-Leibler (KL) divergence ball of the maximum entropy optimal policy, and this fact suffices to ensure mixing of all chains used in the actor and critic updates. In more detail:

1. **Actor analysis via mirror descent.** We write the actor update as an explicit mirror descent. While on the surface this does not change the method (e.g., in the tabular case, the method is identical to natural policy gradient (Agarwal et al., 2021)), it gives a clean optimization guarantee which carries a KL-based implicit bias consequence for free, and decouples concerns between the actor and critic.

2. **Critic analysis via projection-free sampling tools within KL balls.** The preceding mirror descent component guarantees that we stay within a small KL ball, *if* the statistical error of the critic is controlled. Concordantly, our sampling tools guarantee this statistical error is small, *if* we stay within a small KL ball. Concretely, we provide useful lemmas that every policy in a KL ball around the high entropy policy has uniformly upper bounded mixing times, and separately give a projection-free (implicitly regularized!) analysis of the standard *temporal-difference (TD)* update (Sutton, 1988), whereas prior work used projections (Bhandari et al., 2018). These mixing guarantees are in stark contrast to prior work, which either samples directly from stationary distributions (Wang et al., 2019), or makes uniform mixing assumptions (Khodadadian et al., 2021).

Our final proof has the above actor and critic components feeding off of each other: we use an induction to establish that since the actor stayed in a small KL ball in previous iterations, then the new critic update is accurate, and similarly the accuracy of the previous critic updates ensures the actor continues to stay in a small KL ball. We feel these tools will be useful in other work.

1.1 Setting and main results

We will now give the setting, main result, and algorithm in full. Further details on MDPs can be found in Section 1.3, but the actor-critic method appears below in Algorithm 1. To start, the environment and policies are as follows.

Assumption 1.1. The *Markov Decision Process (MDP)* has states $s \in \mathbb{R}^d$ and finitely many actions $a \in \mathcal{A} := \{\boldsymbol{e}_1, \ldots, \boldsymbol{e}_k\}$, and finite rewards $r \in [0, 1]$. States are observed in some feature encoding $s \in \mathbb{R}^d$, but the state space $\mathcal{S} \subseteq \{s \in \mathbb{R}^d : \|s\| \le 1\}$ is assumed finite: $|\mathcal{S}| < \infty$.

Policies are *linear softmax policies*: a policy is given by a weight matrix $W \in \mathbb{R}^{d\times k}$, and given a state $s \in \mathbb{R}^d$, uses a per-state softmax to sample a new action a:

$$a \sim \phi(s^\intercal W \cdot), \qquad \text{where } \phi(s^\intercal W a) = \frac{\exp(s^\intercal W a)}{\sum_{b\in\mathcal{A}} \exp(s^\intercal W b)}. \tag{1.1}$$

Let $\mathcal{A}_s$ denote the set of optimal actions. It is assumed that $\mathcal{A}_s$ is nonempty for every $s \in \mathcal{S}$; equivalently, for every state, there exists an optimal policy whose stationary distribution exists and places positive mass on that state. $\diamondsuit$

The choice of linear policies simplifies presentation, but the tools here should be applicable to many other settings. This choice also allows direct comparison to the widely-studied implicit bias of gradient descent in linear classification settings (Soudry et al., 2017; Ji & Telgarsky, 2018), as will be discussed further in Section 1.2. The choice of finite state space is to remove measure-theoretic concerns and to allow a simple characterization of the maximum entropy optimal policy.

Algorithm 1 Single-trajectory linear actor-critic.

Inputs: actor iterations t, step size θ; critic iterations N, step size η, minibatch size B.
Initialize: actor weights $W_0 = 0 \in \mathbb{R}^{d\times k}$, pre-softmax mapping $p_0(s, a) := s^\intercal W_0 a$, policy $\pi_0 := \phi(p_0)$ (cf. eq. (1.1)); sample initial state/action/reward triple $(s_{000}, a_{000}, r_{000})$.
for $i = 0, 1, 2, \ldots, t-1$ **do**
  Critic update: use π_i to interact with the MDP, obtaining state/action/reward triples $(s_{ijk}, a_{ijk}, r_{ijk})_{j<N,k\le B}$ with $(s_{ij0}, a_{ij0}, r_{ij0}) = (s_{i,j-1,B}, a_{i,j-1,B}, r_{i,j-1,B})$ (single trajectory with some notational convenience), and form *minibatch TD estimates*

$$U_{i,j+1} := U_{ij} - \frac{\eta}{B} \sum_{k<B} s_{ijk} \left(s_{ijk}^\intercal U_{ij} a_{ijk} - \gamma s_{ij,k+1}^\intercal U_{ij} a_{ij,k+1} - r_{ijk} \right) a_{ijk}^\intercal,$$

  with initial condition $U_{i,0} = 0$. Set $U_i := \frac{1}{N}\sum_{j<N} U_{i,j}$ and $\widehat{\mathcal{Q}}_i(s, a) := s^\intercal U_i a$, and also $(s_{i+1,0,0}, a_{i+1,0,0}, r_{i+1,0,0}) = (s_{i,N-1,B}, a_{i,N-1,B}, r_{i,N-1,B})$ to continue same trajectory.
  Actor update: Set $W_{i+1} := W_i + \theta U_i$ and $p_{i+1} := p_i + \theta \widehat{\mathcal{Q}}_i$ and $\pi_{i+1} := \phi(p_{i+1})$.
end for

Lemma 1.2 (simplification of Lemma A.1)**.** *Under Assumption 1.1, there exists a unique maximum entropy policy $\overline{\pi}$, which satisfies $\overline{\pi}(s,\cdot) = \mathrm{Uniform}(\mathcal{A}_s)$ for every state s.*

To round out this introductory presentation of the actor, the last component is the update: $p_{i+1} := p_i + \widehat{\mathcal{Q}}_i$ and $\pi_{i+1} := \phi(p_{i+1})$. This is explicitly a *mirror descent* or *dual averaging* representation of the policy, where we use a *mirror mapping* ϕ to obtain the policy π_{i+1} from pre-softmax values p_{i+1}. As mentioned before, this update appears in prior work in the tabular setting with natural policy gradient and actor-critic (Agarwal et al., 2021; Khodadadian et al., 2021). We will motivate this choice in our more general non-tabular setting in Section 2.

The final assumption and description of the critic are as follows. As will be discussed in Section 2, the policy becomes optimal if $\widehat{\mathcal{Q}}_i$ is an accurate estimate of the true Q function. We employ a standard TD update with no projections or constraints, albeit with minibatching to simplify the analysis. To guarantee that this linear model of $\mathcal{Q}_i$ is accurate, we make a standard *linear MDP assumption* (Bradtke & Barto, 1996; Melo & Ribeiro, 2007; Jin et al., 2020).

Assumption 1.3. In words, the *linear MDP assumption* is that the MDP rewards and transitions are modeled by linear functions. In more detail, for convenience first fix a canonical vector form for state/action pairs $(s,a) \in \mathbb{R}^{d\times k}$: let $x_{sa} \in \mathbb{R}^{dk}$ denote the vector obtained via unrolling the matrix $sa^\intercal$ row-wise (whereby vector inner products with x_{sa} match matrix inner products with $sa^\intercal$). The linear MDP assumption is then that there exists a fixed vector $y \in \mathbb{R}^{dk}$ and a fixed matrix $M \in \mathbb{R}^{d\times dk}$ so that for any state/action pair x_{sa} and any subsequent state $s' \in \mathbb{R}^d$,

$$\mathbb{E}[r|(s,a)] = x_{sa}^\intercal y, \qquad \text{and} \qquad \mathbb{E}[s'|(s,a)] = M x_{sa}. \qquad \diamondsuit$$

Though a strong assumption, it is not only common, but note also that since TD must continually interact with the MDP, then it would have little hope of accuracy if it can not model short-term MDP dynamics. Indeed, as is shown in Lemma C.1 (but appears in various forms throughout the literature), Assumption 1.3 implies that the fixed point of the TD update is the true Q function.

We now state our main result, which bounds not just the *value function* $\mathcal{V}$ (cf. Section 1.3) but also the KL divergence $K_{\mathfrak{v}^s_{\overline{\pi}}}(\overline{\pi},\pi_i)$, where $\mathfrak{v}^s_{\overline{\pi}}$ is the *visitation distribution* of the maximum entropy optimal policy $\overline{\pi}$ when run from state s (cf. Section 1.3).

Theorem 1.4. *Suppose Assumptions 1.1 and 1.3, which in particular implies the (unique) maximum entropy optimal policy is irreducible and aperiodic. Fix an iteration count t, and choose parameters*

$$\theta = \frac{1}{\sqrt{t}}, \qquad \eta = \frac{1-\gamma}{4(1+\gamma^2)}, \qquad N = t^3, \qquad B = \mathcal{O}(t^3),$$

where the constants hidden in B depend only on $\overline{\pi}$ and the MDP, and not on t. With these parameters in place, invoke Algorithm 1, and let $(\pi_i)_{i<t}$ be the resulting sequence of policies. Then with probability at least $1-\mathcal{O}(1/t)$ over the TD sampling, simultaneously for every state $s \in \mathbb{R}^d$ and every $i \le t$,

$$K_{\mathfrak{v}^s_{\overline{\pi}}}(\overline{\pi},\pi_i) + \theta(1-\gamma)\sum_{j<i}\big(\mathcal{V}_j(s) - \mathcal{V}_{\overline{\pi}}(s)\big) \le 1 + \ln k + \frac{1}{(1-\gamma)^2}.$$

Remark 1.5 (discussion of Theorem 1.4)**.**

1. **Implicit bias.** Since $\overline{\pi}$ is optimal, the second term can be deleted, and the bound implies

$$\max_{\substack{i\le t\\ s\in\mathcal{S}}} K_{\mathfrak{v}^s_{\overline{\pi}}}(\overline{\pi},\pi_i) \le 1 + \ln k + \frac{1}{(1-\gamma)^2};$$

   since this holds for all $i \le t$, it controls the optimization path. This term is a direct consequence of our mirror descent setup, and is used to control the TD errors at every iteration.

2. **Mixing time constants.** The minibatch parameter B hides mixing time constants; these mixing time constants depend only on the KL bound $1 + \ln k + (1-\gamma)^{-2}$, and in particular there is no hidden growth in these terms with t. That is to say, mixing times are uniformly controlled over a fixed KL ball that does not depend on t; prior work by contrast makes strong mixing assumptions (Wang et al., 2019; Khodadadian et al., 2021).

3. **High probability guarantee.** Though prior work focuses on bounds in expectation, we chose a high probability guarantee to highlight the power of the analysis technique, and to highlight that quantities are truly not blowing up.

4. **Single trajectory.** A single trajectory through the MDP is used to remove the option of the algorithm escaping from poor choices with resets; only the implicit bias can save it.

5. **Rate.** Since the step size is $1/\sqrt{t}$, to obtain error $\epsilon > 0$, then $t = 1/\epsilon^2$ iterations are needed, and the total number of samples is $tNB = 1/\epsilon^{14}$. This is a slow rate, but the goal here is to remove assumptions and provide clean analysis tools. Note that the closest related work, due to Khodadadian et al. (2021), which also considers a simple algorithm and a single trajectory, in fact achieves a better rate, but the comparison is hard to interpret as that work makes uniform mixing assumptions (cf. Khodadadian et al. (2021, Lemma C.1)), and is in the tabular setting; as minor comparison points, that work makes use of ϵ-greedy, and gives bounds in expectation, whereas the high probability bounds here directly worsened the rate. $\diamondsuit$

The organization of the remainder of this work is as follows. The rest of this introduction gives further related work, and some notation and MDP background. Section 2 presents and discusses the mirror descent framework which provides optimization and implicit bias guarantees on the actor for free. Section 3 presents the sampling lemmas we use to control the TD error. Section 4 concludes with some discussion and open problems, and the appendices contain the full proofs.

1.2 Further related work

For the standard background in reinforcement learning, see Sutton & Barto (2018).

Natural and regular policy gradient (PG & NPG). As mentioned before, the actor update here agrees with the *natural* policy gradient update in the tabular setting (Kakade, 2001); see also (Agarwal et al., 2021) for a well-known analysis of natural and regular policy gradient methods. These methods are widespread in theory and practice (Williams, 1992; Sutton et al., 2000; Bagnell & Schneider, 2003; Liu et al., 2020; Fazel et al., 2018).

Natural and regular actor-critic (AC & NAC). The study of regular and natural actor-critic methods started with Konda & Tsitsiklis (2000) and Peters & Schaal (2008) respectively. These methods are very common both in theory and practice, and there are many variants and improvements to both the actor component and the critic component (Xu et al., 2020a;b; Wu et al., 2020; Bhatnagar et al., 2009). Temporal difference (TD) updates are common in this literature, and originated with (Sutton, 1988).

Regularization and constraints. It is standard with neural policies to explicitly maintain a constraint on the network weights (Wang et al., 2019; Cai et al., 2019). Relatedly, many works both in theory and practice use explicit entropy regularization to prevent small probabilities (Williams & Peng, 1991; Mnih et al., 2016; Abdolmaleki et al., 2018), and which can seem to yield convergence rate improvements (Cen et al., 2020).

NPG and mirror descent. The original and recent analyses of NPG had a mirror descent flavor, though mirror descent and its analysis were not explicitly invoked as a black box (Kakade, 2001; Agarwal et al., 2021). Further connections to mirror descent have appeared many times (Geist et al., 2019; Shani et al., 2020), though with a focus on the design of new algorithms, and not for any implicit regularization effect or proof.

Implicit regularization in supervised learning. A pervasive topic in supervised learning is the *implicit regularization* effect of common descent methods; concretely, standard descent methods prefer low or even minimum norm solutions, which can be converted into generalization bounds. This effect has been observed in coordinate descent with linear predictors (Zhang & Yu, 2005; Telgarsky, 2013), gradient descent with linear predictors (Soudry et al., 2017; Ji & Telgarsky, 2018), and deep learning in various settings (Chizat & Bach, 2020; Lyu & Li, 2019), just to name a few.

1.3 NOTATION

This brief notation section summarizes various concepts and notation used throughout; modulo a few inventions, the presentation mostly matches standard ones in RL (Sutton & Barto, 2018) and policy gradient (Agarwal et al., 2021). A policy $\pi : \mathbb{R}^d \times \mathbb{R}^k \to \mathbb{R}$ maps state-action pairs to reals, and $\pi(s, \cdot)$ will always be a probability distribution. Given a state, the agent samples an action from $a \sim \pi(s, \cdot)$, the environment returns some random reward (which has a fixed distribution conditioned on the observed (s, a) pair), and then uses a *transition kernel* to choose a new state given (s, a).

Taking τ to denote a random trajectory followed by a policy π interacting with the MDP from an arbitrary initial state distribution μ, the *value* $\mathcal{V}$ and *Q functions* are respectively

$$\mathcal{V}_\pi(\mu) := \mathop{\mathbb{E}}_{\substack{s_0 \sim \mu \\ \tau = s_0, a_0, r_0, s_1, \cdots}} \sum_{t \geq 0} \gamma^t r_t,$$

$$\mathcal{Q}_\pi(s, a) := \mathop{\mathbb{E}}_{\substack{s_0 = s, a_0 = a \\ \tau = s_0, a_0, r_0, s_1, \cdots}} \sum_{t \geq 0} \gamma^t r_t = \mathbb{E}_{r_0 \sim (s,a)} \left(r_0 + \gamma \mathbb{E}_{s_1 \sim (s,a)} \mathcal{V}_\pi(s_1) \right),$$

where the simplified notation $\mathcal{V}_\pi(s) = \mathcal{V}_\pi(\delta_s)$ for Dirac distribution δ_s on state s will often be used, as well as the shorthand $\mathcal{V}_i = \mathcal{V}_{\pi_i}$ and $\mathcal{Q}_i = \mathcal{Q}_{\pi_i}$. Additionally, let $\mathcal{A}_\pi(s, a) := \mathcal{Q}_\pi(s, a) - \mathcal{V}_\pi(s)$ denote the *advantage function*; note that the natural policy gradient update could interchangeably use $\mathcal{A}_i$ or $\mathcal{Q}_i$ since they only differ by an action-independent constant, namely $\mathcal{V}_\pi(s)$, which the softmax normalizes out. As in Assumption 1.1, the state space $\mathcal{S}$ is finite but a subset of $\mathbb{R}^d$, specifically $\mathcal{S} \subseteq \{s \in \mathbb{R}^d : \|s\| \leq 1\}$, and the action space $\mathcal{A}$ is just the k standard basis vectors $\{\boldsymbol{e}_1, \ldots, \boldsymbol{e}_k\}$. The other MDP assumption, namely of a *linear MDP* (cf. Assumption 1.3), will be used whenever TD guarantees are needed. Lastly, the *discount factor* $\gamma \in (0, 1)$ has not been highlighted, but is standard in the RL literature, and will be treated as given and fixed throughout the present work.

A common tool in RL is the *performance difference lemma* (Kakade & Langford, 2002): letting $\mathfrak{v}_\pi^\mu$ denote the *visitation distribution* corresponding to policy π starting from μ, meaning

$$\mathfrak{v}_\pi^\mu := \frac{1}{1-\gamma} \mathbb{E}_{s' \sim \mu} \sum_{t \geq 0} \gamma^t \mathrm{Pr}[s_t = s | s_0 = s'],$$

the performance difference lemma can be written as

$$\mathcal{V}_\pi(\mu) - \mathcal{V}_{\pi'}(\mu) = \frac{1}{1-\gamma} \mathbb{E}_{s \sim \mathfrak{v}_{\pi'}^\mu} \sum_a \mathcal{Q}_\pi(s, a)(\pi(s, a) - \pi'(s, a)) = \frac{1}{1-\gamma} \left\langle \mathcal{Q}_\pi, \pi - \pi' \right\rangle_{\mathfrak{v}_{\pi'}^\mu}, \tag{1.2}$$

where the final notation will often be employed for convenience.

As mentioned above, $\mathfrak{s}_\pi$ will denote the stationary distribution of a policy π whenever it exists. The only relevant assumption we make here is that the maximum entropy optimal policy $\overline{\pi}$ is aperiodic and irreducible, which implies it has a stationary distribution with positive mass on every state (Levin et al., 2006, Chapter 1). Via Lemma 3.1, it follows that all policies in a KL ball around $\overline{\pi}$ also have stationary distributions with positive mass on every state.

The max entropy optimal policy $\overline{\pi}$ is complemented by a (unique) optimal Q function $\overline{\mathcal{Q}}$ and optimal advantage function $\overline{\mathcal{A}}$. The optimal Q function $\overline{\mathcal{Q}}$ dominates all other Q functions, meaning $\overline{\mathcal{Q}}(s, a) \geq \mathcal{Q}_\pi(s, a)$ for any policy π; cf. Lemma A.1.

In a few places, we need the Markov chain *on states*, P_π, which is induced by a policy π: that is, the chain where given a state s, we sample $a \sim \pi(s, \cdot)$, and then transition to $s' \sim (s, a)$, where the latter sampling is via the MDP's transition kernel.

We use $\|\mu - \nu\|_{\mathrm{TV}} = \sup_{U \subseteq \mathcal{S}} |\mu(U) - \nu(U)|$ to denote the *total variation distance*. This distance is pervasive in mixing time analyses (Levin et al., 2006).

2 MIRROR DESCENT TOOLS

To see how nicely mirror descent and its guarantees fit with the NPG/NAC setup, first recall our updates: $p_{i+1} := p_i + \widehat{\mathcal{Q}}_i$, and $\pi_{i+1} := \phi(p_{i+1})$ (e.g., matching NPG in the tabular case (Kakade,

2001; Agarwal et al., 2021)). In the online learning literature (Shalev-Shwartz, 2011; Lattimore & Szepesvári, 2020), the basic mirror ascent (or *dual averaging*) guarantee is of the form

$$\sum_{i<t} \left\langle \widehat{\mathcal{Q}}_i, \pi - \pi_i \right\rangle \leq \mathcal{O}(\sqrt{t}),$$

where notably $\widehat{\mathcal{Q}}_i$ does not need to mean anything, it can just be an element of a vector space. The most common results are stated when $\widehat{\mathcal{Q}}_i$ is the gradient of some convex function, but here instead we can use the performance difference lemma: recalling the inner product and visitation distribution notation from Section 1.3,

$$\begin{aligned}\left\langle \widehat{\mathcal{Q}}_i, \pi_i - \pi \right\rangle_{\mathfrak{v}_\pi^\mu} &= \langle \mathcal{Q}_i, \pi_i - \pi \rangle_{\mathfrak{v}_\pi^\mu} + \left\langle \widehat{\mathcal{Q}}_i - \mathcal{Q}_i, \pi_i - \pi \right\rangle_{\mathfrak{v}_\pi^\mu} \\ &= (1-\gamma)\left(\mathcal{V}_i(\mu) - \mathcal{V}_\pi(\mu)\right) + \left\langle \widehat{\mathcal{Q}}_i - \mathcal{Q}_i, \pi_i - \pi \right\rangle_{\mathfrak{v}_\pi^\mu}.\end{aligned}$$

The term $\widehat{\mathcal{Q}}_i - \mathcal{Q}_i$ is exactly what we will control with the TD analysis, and thus the mirror descent approach has neatly decoupled concerns into an actor term, and a critic term.

In order to apply the mirror descent framework, we need to choose a *mirror mapping*. Rather than using ϕ, we use a choice which bakes the measure $\mathfrak{v}_\pi^\mu$ from the above inner product above into the dual object q! This may seem strange, but it does not change the induced policy, and thus is a degree of freedom, and allows us to state guarantees for *all* possible starting distributions for free.

Our full mirror descent setup is detailed in Appendix B, but culminates in the following guarantee.

Lemma 2.1. *Consider step size $\theta > 0$, any reference policy π, and two treatments of the error $\widehat{\mathcal{Q}}_i - \mathcal{Q}_i$.*

1. *Define $\hat{\epsilon}_i := \sup_{s,a} |\widehat{\mathcal{Q}}_i(s,a) - \mathcal{Q}_i(s,a)|$, and also assume $\widehat{\mathcal{Q}}_i \in [0, 1/(1-\gamma)]$. For any starting measure μ,*

$$\begin{aligned}K_{\mathfrak{v}_\pi^\mu}(\pi, \pi_t) + \theta(1-\gamma)\sum_{i<t}\left(\mathcal{V}_\pi(\mu) - \mathcal{V}_i(\mu)\right) \leq K_{\mathfrak{v}_\pi^\mu}(\pi, \pi_0) + \frac{1}{1-\gamma} \\ + \sum_{i<t}\left(\frac{2\gamma\hat{\epsilon}_i}{1-\gamma} + \hat{\epsilon}_i + \hat{\epsilon}_{i+1}\right),\end{aligned}$$

   and additionally $\mathcal{V}_i$ and $\widehat{\mathcal{Q}}_i$ are approximately monotone: for any starting state s and any action a,

$$\mathcal{V}_{i+1}(s) \geq \mathcal{V}_i(s) - \frac{2\hat{\epsilon}_i}{1-\gamma} \qquad \text{and} \qquad \mathcal{Q}_{i+1}(s,a) \geq \mathcal{Q}_i(s,a) - \frac{2\gamma\hat{\epsilon}_i}{1-\gamma} - \hat{\epsilon}_i - \hat{\epsilon}_{i+1}.$$

2. *For any starting measure μ,*

$$\begin{aligned}K_{\mathfrak{v}_\pi^\mu}(\pi, \pi_t) + \theta(1-\gamma)\sum_{i<t}\left(\mathcal{V}_\pi(\mu) - \mathcal{V}_i(\mu)\right) \leq K_{\mathfrak{v}_\pi^\mu}(\pi, \pi_0) + \frac{t\theta^2}{(1-\gamma)^2} \\ + \theta\sum_{i<t}\left\langle \mathcal{Q}_i - \widehat{\mathcal{Q}}_i, \pi_i - \pi \right\rangle_{\mathfrak{v}_\pi^\mu}.\end{aligned}$$

Remark 2.2 (Regarding the mirror descent setup, Lemma 2.1)**.**

1. **Two rates.** For the first bound, it is most natural to set $\theta = 1$, which requires $\mathcal{O}(1/\epsilon)$ iterations to reach accuracy $\epsilon > 0$; by contrast, the second guarantee requires $\mathcal{O}(1/\epsilon^2)$ iterations for the same $\epsilon > 0$. We used the second form to prove Theorem 1.4, since the error term provided here is less stringent. Indeed, the TD analysis we provide in Section 3 will not be able to give the uniform control needed for the first bound.

2. **Comparison to standard rates.** Comparing the first bound to the standard NPG rate in the literature (Agarwal et al., 2021), the rate is exactly recovered; as such, this mirror descent setup at the very least has not paid a price in rates.

3. **Implicit regularization term.** A conspicuous difference between these bounds and both the standard NPG bounds (cf. (Agarwal et al., 2021, Theorem 5.3)), but also many mirror descent treatments, is the term $K_{\mathfrak{v}_\pi^\mu}(\pi, \pi_t)$; one could argue that this term is nonnegative and moreover we care more about the value function, so why not drop it, as is usual? It is precisely this term that gives our implicit regularization effect: instead, we can drop *the value function term* and uniformly upper bound the right hand side to get $K_{\mathfrak{v}_{\overline{\pi}}^\mu}(\overline{\pi}, \pi_t) \leq 1 + \ln k + (1-\gamma)^{-2}$ in Theorem 1.4. $\diamondsuit$

3 SAMPLING TOOLS

Via Lemma 2.1 above, our mirror descent black box analysis gives us a KL bound and a value function bound: what remains, and is the job of this section, is to control the Q function error terms involving $\mathcal{Q}_i - \widehat{\mathcal{Q}}_i$.

As a starting point, first we show that a KL bound is enough: that is to say, if we are in a KL ball around some well-behaved policy, then we are reasonably well-behaved as well.

Lemma 3.1. *Let policy $\tilde{\pi}$ be given, and suppose the induced transition kernel on states $P_{\tilde{\pi}}$ is irreducible and aperiodic (Levin et al., 2006, Section 1.3). Then $\tilde{\pi}$ has a stationary distribution $\mathfrak{s}_{\tilde{\pi}}$, and moreover for any $c > 0$ and any measure ν which is positive on all states and a corresponding set of policies*

$$\mathcal{P}_c := \left\{\pi : K_\nu(\tilde{\pi}, \pi) \leq c\right\},$$

there exist constants C, m_1, m_2 so that mixing is uniform over $\mathcal{P}_c$, meaning for any t, and any $\pi \in \mathcal{P}_c$ with induced kernel P_π,

$$\sup_s \|P_\pi^t(s) - \mathfrak{s}_\pi(s)\|_{\mathrm{TV}} \leq m_1 e^{-m_2 t},$$

and for any state s and any $\pi \in \mathcal{P}_c$, and any action a with $\tilde{\pi}(s,a) > 0$,

$$\frac{1}{C} \leq \frac{\tilde{\pi}(s,a)}{\pi(s,a)} \leq C \qquad \text{and} \qquad \frac{1}{C} \leq \frac{\mathfrak{s}_{\tilde{\pi}}(s)}{\mathfrak{s}_\pi(s)} \leq C.$$

Remark 3.2 (Implicit vs explicit exploration)**.** On the surface, Lemma 3.1 might seem quite nice. Worrying about it a little more, and especially after inspecting the proof, it is clear that the constants C, m_1, and m_2 can be quite bad. On the one hand, one may argue that this is inherent to implicit exploration, and something like ϵ-greedy is preferable, as it arguably gives an explicit control on all these quantities.

Some aspects of this situation are unavoidable, however. Consider a *combination lock* MDP, where a precise, hard to find sequence of actions must be followed to arrive at some good reward. Suppose this sequence has length n and we have a reference policy $\tilde{\pi}$ which takes each of these good actions with probability $1 - 1/n$, whereby the probability of the sequence is $(1 - 1/n)^n \approx 1/e$; a policy $\pi \in \mathcal{P}_c$ with $\pi(s,a)/\tilde{\pi}(s,a) \leq 1/2$ can drop the probability of this good sequence of actions all the way down to $1/2^n$! $\diamondsuit$

The last tool we present is the TD guarantee. The version here is specialized to Algorithm 1; a general form is presented in Appendix C.

Lemma 3.3 (See also Lemma C.2)**.** *Suppose the MDP and linear MDP assumptions (cf. Assumptions 1.1 and 1.3). Consider a policy π_i in some iteration i of Algorithm 1, and suppose there exist mixing constants $m \geq 1$ and $c > 0$ so that the induced transition kernel P_{π_i} on $\mathcal{S}$ satisfies*

$$\sup_s \|P_{\pi_i}^t(s) - \mathfrak{s}_{\pi_i}(s)\|_{\mathrm{TV}} \leq m e^{-ct}.$$

Then for any step size $\eta \leq \frac{1-\gamma}{4(1+\gamma^2)}$, any number of minibatches $N \geq 1$, and any minibatch size $B \geq \frac{192 m N \eta}{(1-e^{-c})}$, letting $\mathbb{E}_{\widehat{\mathcal{Q}}_i}$ denote the randomness of the trajectory sampled by π_i to construct $\widehat{\mathcal{Q}}_i$,

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\widehat{\mathcal{Q}}_i} \mathop{\mathbb{E}}_{(s,a)\sim(\mathfrak{s}_{\pi_i},\pi_i)} \left(\widehat{\mathcal{Q}}_i(s,a) - \mathcal{Q}_i(s,a)\right)^2 \\
&\leq \frac{1}{N\eta(1-\gamma)} \left[\frac{1}{(1-\gamma)^2} + \mathbb{E}_{\widehat{\mathcal{Q}}_i} \sup_{s,a} \mathcal{Q}_i(s,a)^2 - \mathbb{E}_{\widehat{\mathcal{Q}}_i} \sup_{s,a} \left(s^\intercal U_{i,N-1} a - \mathcal{Q}_i(s,a)\right)^2\right].
\end{aligned}$$

Remark 3.4 (Implicit bias *of TD*)**.** The right hand side of the final bound in Lemma 3.3 may seem odd, but the negated expectation term is analogous to the term $K_{\mathfrak{v}_\pi^\mu}(\pi, \pi_t)$ in Lemma B.1 which was used to control the entropy of the policy path. Here, this term is used to prevent $U_{i,N-1}$ from becoming too large, and is how the analysis here can omit a projection step which was needed in prior work (Bhandari et al., 2018). $\diamondsuit$

All the pieces are now in place to sketch the proof of Theorem 1.4, which is presented in full in Appendix D. To start, instantiate Lemma 3.1 with KL divergence upper bound $1 + \ln k + (1-\gamma)^{-2}$, which gives the various mixing constants used throughout the proof (which we need to instantiate now, before seeing the sequence of policies, to avoid any dependence). With that out of the way, consider some iteration i, and suppose that for all iterations $j < i$, we have a handle both on the TD error, and also a guarantee that we are in a small KL ball around $\overline{\pi}$ (specifically, of radius $1 + \ln k + (1-\gamma)^{-2}$ as in Theorem 1.4). The right hand side of the second mirror descent bound in Lemma 2.1 only needs a control on all previous TD errors, therefore it implies both a bound on $\sum_{j<i} \mathcal{V}_j(s)$ and on $K_{\mathfrak{v}_{\overline{\pi}}^s}(\overline{\pi}, \pi_i)$. But this KL control on π_i means that the mixing and other constants we assumed at the start will hold for π_i, and thus we can invoke Lemma 3.3 to bound the error on $\widehat{\mathcal{Q}}_i - \mathcal{Q}_i$, which we will use in the next loop of the induction. In this way, the actor and critic analysis complement each other and work together in each step of the induction.

There was one issue overlooked in the preceding paragraph. Notice that Lemma C.2 only grants an error control on average over (s, a) pairs sampled from the stationary distribution of π_i (which we mix towards thanks to Lemma 3.1). To control the error in Lemma 2.1, superficially we need something closer to a uniform error over (s, a) pairs; this ends up not being a problem though, as the only pairs we can't control are when $a \notin \mathcal{A}_s$, which ends up not mattering due to the exact form of the error term.

4 Discussion and open problems

This work, in contrast to prior work in natural actor-critic and natural policy gradient methods, dropped many assumptions from the analysis, and components of the algorithms. The analysis was meant to be fairly general purpose and unoptimized. As such, there are many open problems.

Faster rates. How much can this analysis be squeezed? Moreover, does it suggest any algorithmic improvements?

Implicit vs explicit regularization/exploration. What are some situations where one is better than the other, and vice versa? The analysis here only says you can get away with doing everything implicitly, but not necessarily that this is the best option.

More general settings. The paper here is for linear MDPs, linear softmax policies, finite state and action spaces. How much does the implicit bias phenomenon (and this analysis) help in more general settings?

Tightening the TD and MD coupling. The proof of Theorem 1.4 here relied on a very tight coupling of the actor (mirror descent) and the critic (temporal difference). But perhaps the coupling can be made even tighter, both in the algorithm and the analysis?

References

Abbas Abdolmaleki, Jost Tobias Springenberg, Yuval Tassa, Remi Munos, Nicolas Heess, and Martin Riedmiller. Maximum a posteriori policy optimisation. *arXiv preprint arXiv:1806.06920*, 2018.

Alekh Agarwal, Sham M Kakade, Jason D Lee, and Gaurav Mahajan. On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *Journal of Machine Learning Research*, 22(98):1–76, 2021.

J Andrew Bagnell and Jeff Schneider. Covariant policy search. 2003.

Jalaj Bhandari, Daniel Russo, and Raghav Singal. A finite time analysis of temporal difference learning with linear function approximation. *arXiv preprint arXiv:1806.02450*, 2018.

Shalabh Bhatnagar, Richard S Sutton, Mohammad Ghavamzadeh, and Mark Lee. Natural actor–critic algorithms. *Automatica*, 45(11):2471–2482, 2009.

Steven J Bradtke and Andrew G Barto. Linear least-squares algorithms for temporal difference learning. *Machine learning*, 22(1):33–57, 1996.

Sébastien Bubeck. Convex optimization: Algorithms and complexity. *Foundations and Trends in Machine Learning*, 2015.

Qi Cai, Zhuoran Yang, Jason D Lee, and Zhaoran Wang. Neural temporal-difference and q-learning provably converge to global optima. *arXiv preprint arXiv:1905.10027*, 2019.

Shicong Cen, Chen Cheng, Yuxin Chen, Yuting Wei, and Yuejie Chi. Fast global convergence of natural policy gradient methods with entropy regularization. *arXiv preprint arXiv:2007.06558*, 2020.

Lenaic Chizat and Francis Bach. Implicit bias of gradient descent for wide two-layer neural networks trained with the logistic loss. *arXiv preprint arXiv:2002.04486*, 2020.

Maryam Fazel, Rong Ge, Sham Kakade, and Mehran Mesbahi. Global convergence of policy gradient methods for the linear quadratic regulator. In *International Conference on Machine Learning*, pp. 1467–1476. PMLR, 2018.

Matthieu Geist, Bruno Scherrer, and Olivier Pietquin. A theory of regularized markov decision processes. In *International Conference on Machine Learning*, pp. 2160–2169. PMLR, 2019.

Mark Jerrum and Alistair Sinclair. Conductance and the rapid mixing property for markov chains: The approximation of permanent resolved. In *STOC*, pp. 235–244, 1988.

Ziwei Ji and Matus Telgarsky. Risk and parameter convergence of logistic regression. `arXiv:1803.07300v3 [cs.LG]`, 2018.

Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I Jordan. Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory*, pp. 2137–2143. PMLR, 2020.

Sham M Kakade. A natural policy gradient. *Advances in neural information processing systems*, 14, 2001.

Sham M. Kakade and John Langford. Approximately optimal approximate reinforcement learning. In *ICML*, 2002.

Sajad Khodadadian, Thinh T Doan, Siva Theja Maguluri, and Justin Romberg. Finite sample analysis of two-time-scale natural actor-critic algorithm. *arXiv preprint arXiv:2101.10506*, 2021.

Vijay R Konda and John N Tsitsiklis. Actor-critic algorithms. In *Advances in neural information processing systems*, pp. 1008–1014, 2000.

Tor Lattimore and Csaba Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2020. doi: 10.1017/9781108571401.

David A. Levin, Yuval Peres, and Elizabeth L. Wilmer. *Markov chains and mixing times*. American Mathematical Society, 2006.

Yanli Liu, Kaiqing Zhang, Tamer Basar, and Wotao Yin. An improved analysis of (variance-reduced) policy gradient and natural policy gradient methods. In *NeurIPS*, 2020.

Kaifeng Lyu and Jian Li. Gradient descent maximizes the margin of homogeneous neural networks. *arXiv preprint arXiv:1906.05890*, 2019.

Francisco S Melo and M Isabel Ribeiro. Q-learning with linear function approximation. In *International Conference on Computational Learning Theory*, pp. 308–322. Springer, 2007.

Volodymyr Mnih, Adria Puigdomenech Badia, Mehdi Mirza, Alex Graves, Timothy Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. Asynchronous methods for deep reinforcement learning. In *International conference on machine learning*, pp. 1928–1937. PMLR, 2016.

Jan Peters and Stefan Schaal. Natural actor-critic. *Neurocomputing*, 71(7-9):1180–1190, 2008.

Shai Shalev-Shwartz. Online learning and online convex optimization. *Foundations and trends in Machine Learning*, 4(2):107–194, 2011.

Lior Shani, Yonathan Efroni, and Shie Mannor. Adaptive trust region policy optimization: Global convergence and faster rates for regularized mdps. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 5668–5675, 2020.

Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *arXiv preprint arXiv:1710.10345*, 2017.

Richard S Sutton. Learning to predict by the methods of temporal differences. *Machine learning*, 3 (1):9–44, 1988.

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

Richard S Sutton, David A McAllester, Satinder P Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In *Advances in neural information processing systems*, pp. 1057–1063, 2000.

Matus Telgarsky. Margins, shrinkage, and boosting. In *ICML*, 2013.

Michel Tokic. Adaptive ε-greedy exploration in reinforcement learning based on value differences. In *Annual Conference on Artificial Intelligence*, pp. 203–210. Springer, 2010.

Lingxiao Wang, Qi Cai, Zhuoran Yang, and Zhaoran Wang. Neural policy gradient methods: Global optimality and rates of convergence. *arXiv preprint arXiv:1909.01150*, 2019.

Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3):229–256, 1992.

Ronald J Williams and Jing Peng. Function optimization using connectionist reinforcement learning algorithms. *Connection Science*, 3(3):241–268, 1991.

Yue Wu, Weitong Zhang, Pan Xu, and Quanquan Gu. A finite time analysis of two time-scale actor critic methods. *arXiv preprint arXiv:2005.01350*, 2020.

Tengyu Xu, Zhe Wang, and Yingbin Liang. Improving sample complexity bounds for actor-critic algorithms. *arXiv preprint arXiv:2004.12956*, 2020a.

Tengyu Xu, Zhe Wang, and Yingbin Liang. Non-asymptotic convergence analysis of two time-scale (natural) actor-critic algorithms. *arXiv preprint arXiv:2005.03557*, 2020b.

Tong Zhang and Bin Yu. Boosting with early stopping: Convergence and consistency. *The Annals of Statistics*, 33:1538–1579, 2005.

A BACKGROUND PROOF: EXISTENCE OF $\overline{\pi}$

The only thing in this section is the expanded version of Lemma 1.2, namely giving the unique maximum entropy optimal policy, and some key properties.

Lemma A.1. *There exists a unique maximum margin optimal policy $\overline{\pi}$ and corresponding $\overline{\mathcal{Q}}$ and $\overline{\mathcal{A}}$ which satisfy the following properties.*

1. *For any state s, let $\mathcal{A}_s$ denote the set of actions taken by optimal policies; Define $\overline{\pi}(s,\cdot) := \mathrm{Uniform}(\mathcal{A}_s)$, which is unique; then $\overline{\pi}$ is also an optimal policy, and let $\overline{\mathcal{A}}$ and $\overline{\mathcal{Q}}$ denote its advantage and Q functions.*

2. *For every state s and every action a, then $\overline{\mathcal{Q}}(s,a) = \max_{Q\in\mathcal{Q}} Q(s,a)$, where Q is the set of all Q functions (that is, the set of Q functions corresponding to the set of all policies). Moreover, $\max_{a\in\mathcal{A}_s} \overline{\mathcal{Q}}(s,a) > \max_{a\notin\mathcal{A}_s} \overline{\mathcal{Q}}(s,a)$.*

3. *$\overline{\pi} = \lim_{r\to\infty} \phi(r\overline{\mathcal{A}})$.*

Proof of Lemmas 1.2 and A.1. 1. We provide an iterative construction of $\overline{\pi}$. Start with $\overline{\pi}_0$ equal to any optimal deterministic policy (which must exist as usual for MDPs), and consider any enumeration $(s_1,\ldots,s_{|\mathcal{S}|})$ of the set of states. The construction produces $\overline{\pi}_{i+1}$ from $\overline{\pi}_i$, and will assume by induction that $\overline{\pi}_i$ is an optimal policy which for every state s_j with $j \le i$ satisfies $\overline{\pi}_i(s_j,\cdot) = \mathrm{Uniform}(\mathcal{A}_{s_j})$. The base case was handled directly, thus consider constructing $\overline{\pi}_{i+1}$. Since Markov chains have no memory, the behavior in state s_{i+1} is independent of the behavior in all prior and subsequent states; therefore we can safely define $\overline{\pi}_{i+1}(s,\cdot) = \overline{\pi}_i(s,\cdot)$ for $s \neq s_{i+1}$ and $\overline{\pi}_{i+1}(s_{i+1},\cdot) = \mathrm{Uniform}(\mathcal{A}_{s_{i+1}})$, and $\overline{\pi}_{i+1}$ is still an optimal policy. To complete the construction, set $\overline{\pi} := \overline{\pi}_{|\mathcal{S}|}$, and let $\overline{\mathcal{Q}}$ and $\overline{\mathcal{A}}$ correspond to it.

2. For any other $Q \in \mathcal{Q}$ with corresponding V, and any (s,a), then
$$\begin{aligned}\overline{\mathcal{Q}}(s,a) - Q(s,a) &= \mathbb{E}_{r,s'\sim(s,a)}\left[r(s,a) + \gamma\overline{\mathcal{V}}(s') - r(s,a) - \gamma V(s')\right]\\ &= \gamma\mathbb{E}_{s'\sim(s,a)}\left[\overline{\mathcal{V}}(s') - V(s')\right]\\ &\ge 0.\end{aligned}$$
It follows that $\overline{\mathcal{Q}}(s,a) \ge \sup_{Q\in\mathcal{Q}} Q(s,a)$, and since $\overline{\mathcal{Q}} \in \mathcal{Q}$, then the inequality is an equality and the supremum is attained.

3. By the previous point, for any state s and any $a \in \mathcal{A}_s$, then $\overline{\mathcal{A}}(s,a) = \overline{\mathcal{Q}}(s,a) - \overline{\mathcal{V}}(s) = 0$ whereas for any $b \notin \mathcal{A}_s$, then $\overline{\mathcal{A}}(s,b) = \overline{\mathcal{Q}}(s,b) - \overline{\mathcal{V}}(s) \le \max_{b\notin\mathcal{A}_s} \overline{\mathcal{Q}}(s,b) - \min_{a\in\mathcal{A}_s} \overline{\mathcal{Q}}(s,a) < 0$. It follows that
$$\lim_{r\to\infty} \phi(r\mathcal{A}(s,\cdot)) = \mathrm{Uniform}(\mathcal{A}_s) = \overline{\pi}(s,\cdot).$$

□

B FULL MIRROR DESCENT SETUP AND PROOFS

This section first gives a basic mirror descent / dual averaging setup. This characterization is mostly standard (Bubeck, 2015), though the main guarantees are given with some flexibility to allow for various natural policy setups.

First, here is the basic notation (which, unlike the paper body, will allow for step size θ to differ between iterations):

$$\begin{aligned}
p_{i+1} &:= p_i - \theta_i g_i, && \theta_i > 0,\\
q_i &:= \nabla\psi(p_i), && \text{closed proper convex } \psi,\\
\langle\!\langle\cdot,\cdot\rangle\!\rangle & && \text{bilinear pairing},\\
D(p,p_i) &:= \psi(p) - \left[\psi(p_i) + \langle\!\langle q_i, p - p_i\rangle\!\rangle\right] && \text{primal Bregman},\\
D_*(q,q_i) &:= \psi^*(q) - \left[\psi^*(q_i) + \langle\!\langle p_i, q - q_i\rangle\!\rangle\right], && \text{dual Bregman}.
\end{aligned}$$

One nonstandard choice here is that the Bregman divergence bakes in a conjugate element, rather than using $\nabla\psi$ and $\nabla\psi^*$; this gives an easy way to handle certain settings (like the boundary of the simplex) which run into non-uniqueness issues. Secondly, $\langle\!\langle \cdot, \cdot \rangle\!\rangle$ is just some bilinear form and need not be interpreted as a standard inner product.

The standard Bregman identities used in mirror descent proofs are as follows:

$$D_*(q, q_i) - D_*(q, q_{i+1}) - D_*(q_{i+1}, q_i) = \langle\!\langle p_i - p_{i+1}, q_{i+1} - q \rangle\!\rangle, \tag{B.1}$$
$$D_*(q_{i+1}, q_i) = D(p_i, p_{i+1}), \tag{B.2}$$
$$D_*(q_{i+1}, q_i) + D_*(q_i, q_{i+1}) = \langle\!\langle p_i - p_{i+1}, q_i - q_{i+1} \rangle\!\rangle, \tag{B.3}$$
$$D_*(q, q_i) = \psi^*(q) + \psi(p_i) - \langle\!\langle p_i, q \rangle\!\rangle \geq 0. \tag{B.4}$$

With these in hand, the core mirror descent guarantee is as follows. The bound is written with equalities to allow for careful handling of error terms. Note that this version of mirror descent does not interpret the "gradient" g_i in any way, and treats it as a vector and no more.

Lemma B.1. *Suppose $\theta_i > 0$. For any t and q where $q \in \mathrm{dom}(\psi^*)$,*

$$\begin{aligned}
\sum_{i<t} \theta_i \langle\!\langle g_i, q_i - q \rangle\!\rangle &= D_*(q, q_0) - D_*(q, q_t) + \sum_{i<t} D(p_{i+1}, p_i) \\
&= D_*(q, q_0) - D_*(q, q_t) + \sum_{i<t} \left[\langle\!\langle p_i - p_{i+1}, q_i - q_{i+1} \rangle\!\rangle - D_*(q_{i+1}, q_i) \right] \\
&= D_*(q, q_0) + \theta_0 \langle\!\langle g_0, q_0 \rangle\!\rangle - D_*(q, q_t) - \theta_{t-1} \langle\!\langle g_t, q_t \rangle\!\rangle - \sum_{i<t} D_*(q_{i+1}, q_i) \\
&\quad + \sum_{i=1}^{t-1} \langle\!\langle g_i, q_i \rangle\!\rangle (\theta_i - \theta_{i-1}) + \sum_{i<t} \theta_i \langle\!\langle g_{i+1} - g_i, q_{i+1} \rangle\!\rangle .
\end{aligned}$$

Moreover, for any i, $\langle\!\langle g_i, q_{i+1} \rangle\!\rangle \leq \langle\!\langle g_i, q_i \rangle\!\rangle$.

Proof. For any fixed iterate $i < t$, by eqs. (B.1) to (B.3),

$$\begin{aligned}
\theta_i \langle\!\langle g_i, q_i - q \rangle\!\rangle &= \langle\!\langle p_i - p_{i+1}, q_i - q \rangle\!\rangle \\
&= \langle\!\langle p_i - p_{i+1}, q_i - q_{i+1} \rangle\!\rangle + \langle\!\langle p_i - p_{i+1}, q_{i+1} - q \rangle\!\rangle \\
&= \langle\!\langle p_i - p_{i+1}, q_i - q_{i+1} \rangle\!\rangle + D_*(q, q_i) - D_*(q, q_{i+1}) - D_*(q_{i+1}, q_i) && \because \text{eq. (B.1)} \\
&= D_*(q, q_i) - D_*(q, q_{i+1}) + D_*(q_i, q_{i+1}) && \because \text{eq. (B.3)} \\
&= -D_*(q, q_i) + D_*(q, q_{i+1}) + D(p_{i+1}, p_i), && \because \text{eq. (B.2)}
\end{aligned}$$

and additionally note

$$\frac{1}{\theta_i} \langle\!\langle p_i - p_{i+1}, q_i - q_{i+1} \rangle\!\rangle = \langle\!\langle g_i, q_i - q_{i+1} \rangle\!\rangle = \langle\!\langle g_{i+1} - g_i, q_{i+1} \rangle\!\rangle - \langle\!\langle g_{i+1}, q_{i+1} \rangle\!\rangle + \langle\!\langle g_i, q_i \rangle\!\rangle .$$

The first equalities now follow by applying $\sum_{i<t}$ to both sides and telescoping.

For the second part, for any i, by convexity of ψ,

$$0 \leq \left\langle\!\!\left\langle p_i - p_{i+1}, \nabla\psi(p_i) - \nabla\psi(p_{i+1}) \right\rangle\!\!\right\rangle = \theta_i \langle\!\langle g_i, q_i - q_{i+1} \rangle\!\rangle ,$$

which rearranges to give $\langle\!\langle g_i, q_{i+1} \rangle\!\rangle \leq \langle\!\langle g_i, q_i \rangle\!\rangle$ since $\theta_i > 0$. □

All that remains is to instantiate the various mirror descent objects to match Algorithm 1, and control the various resulting terms. This culminates in Lemma 2.1; its proof is as follows.

Proof of Lemma 2.1. The core of both parts of the proof is to apply the mirror descent guarantees from Lemma B.1, using the following choices:

$$
\begin{aligned}
g_i &:= -\widehat{\mathcal{Q}}_i, \\
p_{i+1} &:= p_i - \theta g_i = p_i + \theta \widehat{\mathcal{Q}}_i, \\
\langle\!\langle p, q \rangle\!\rangle &:= \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} p(s,a) q(s,a), \\
\psi(p) &:= \mathbb{E}_{s\sim\mathfrak{v}_\pi^\mu} \ln \sum_{a\in\mathcal{A}} \exp(p(s,a)) = \sum_{s\in\mathcal{S}} \mathfrak{v}_\pi^\mu(s) \ln \sum_{a\in\mathcal{A}} \exp(p(s,a)), \\
[\nabla\psi(p)](s,a) &= \frac{\mathfrak{v}_\pi^\mu(s)\exp(p(s,a))}{\sum_{b\in\mathcal{A}} \exp(p(s,b))}, \\
q_i &:= \nabla\psi(p_i), \\
q_\pi(s,a) &:= \mathfrak{v}_\pi^\mu(s)\pi(s,a), \\
D(p,p_i) &:= \psi(p) - \left[\psi(p_i) - \langle\!\langle q_i, p - p_i \rangle\!\rangle\right], \\
\psi^*(q) &= \begin{cases} \langle\!\langle q, \ln q \rangle\!\rangle - \sum_s \mathfrak{v}_\pi^\mu(s) \ln \mathfrak{v}_\pi^\mu(s), & q \in \Delta_{\mathcal{S}\times\mathcal{A}}, \\ \infty & \text{o.w.}, \end{cases} \\
D_*(q,q_i) &:= \psi^*(q) - \left[\psi^*(q_i) - \langle\!\langle p_i, q - q_i \rangle\!\rangle\right] \\
&= \langle\!\langle q, \ln q \rangle\!\rangle - \langle\!\langle q_i, \ln q_i \rangle\!\rangle \\
&\qquad - \sum_{s,a} \left(\ln(q_i(s,a)) + \ln \mathfrak{v}_\pi^\mu(s) + \ln \sum_b \exp(p_i(s,b) \right) \big(q(s,a) - q_i(s,a)\big) \\
&= \left\langle\!\!\left\langle q, \ln \frac{q}{q_i} \right\rangle\!\!\right\rangle = K_{\mathfrak{v}_\pi^\mu}(\pi, \pi_i).
\end{aligned}
$$

A key consequence of these constructions is that q_i, treated for any fixed s as an unnormalized policy, agrees with $\pi_i := \phi(p_i)$ after normalization; that is to say, it gives the same policy, and the choice of $\mathfrak{v}_\pi^\mu$ baked into the definition is not needed by the algorithm, is only used in the analysis. The "gradient" $\widehat{\mathcal{Q}}_i$ makes no use of it.

Plugging this notation in to Lemma B.1 but making use of two of its equalities, and the performance difference lemma eq. (1.2), then for any μ

$$
\begin{aligned}
\theta(1-\gamma) \sum_{i<t} \big(\mathcal{V}_i(\mu) - \mathcal{V}_\pi(\mu)\big) &= \sum_{i<t} \theta \left\langle \mathcal{Q}_i, \pi_i - \pi \right\rangle_{\mathfrak{v}_\pi^\mu} \\
&= \sum_{i<t} \theta \left\langle \mathcal{Q}_i - \widehat{\mathcal{Q}}_i, \pi_i - \pi \right\rangle_{\mathfrak{v}_\pi^\mu} + \sum_{i<t} \left\langle \theta \widehat{\mathcal{Q}}_i, \pi_i - \pi \right\rangle_{\mathfrak{v}_\pi^\mu} \\
&= \sum_{i<t} \theta \left\langle \mathcal{Q}_i - \widehat{\mathcal{Q}}_i, \pi_i - \pi \right\rangle_{\mathfrak{v}_\pi^\mu} - \sum_{i<t} \langle\!\langle \theta_i g_i, q_i - q_\pi \rangle\!\rangle. \qquad \text{(B.5)}
\end{aligned}
$$

The proof now splits into the two different settings.

1. Consider the case of ℓ_∞ norm errors $\hat{\epsilon}_i$. By Lemma B.1 with the above choices and any measure $\mathfrak{v}_\pi^\mu$, then

$$
\left\langle \widehat{\mathcal{Q}}_i, \pi_{i+1} \right\rangle_{\mathfrak{v}_\pi^\mu} = -\langle\!\langle g_i, q_{i+1} \rangle\!\rangle \geq -\langle\!\langle g_i, q_i \rangle\!\rangle = \left\langle \widehat{\mathcal{Q}}_i, \pi_i \right\rangle_{\mathfrak{v}_\pi^\mu}, \qquad \text{(B.6)}
$$

and also

$$
\begin{aligned}
-\sum_{i<t} \langle\!\langle \theta_i g_i, q_i - q_\pi \rangle\!\rangle &= -D_*(q,q_0) - \theta \langle\!\langle g_0, q_0 \rangle\!\rangle + D_*(q,q_t) + \theta \langle\!\langle g_t, q_t \rangle\!\rangle + \sum_{i<t} D_*(q_{i+1}, q_i) \\
&\qquad - \sum_{i<t} \theta \langle\!\langle g_{i+1} - g_i, q_{i+1} \rangle\!\rangle \\
&\geq K_{\mathfrak{v}_\pi^\mu}(\pi, \pi_t) - K_{\mathfrak{v}_\pi^\mu}(\pi, \pi_0) - \frac{\theta}{1-\gamma} + \sum_{i<t} \theta \left\langle \widehat{\mathcal{Q}}_{i+1} - \widehat{\mathcal{Q}}_i, \pi_{i+1} \right\rangle_{\mathfrak{v}_\pi^\mu}.
\end{aligned}
$$

To simplify these further, first note by eq. (B.6) with measure $\mathfrak{v}_{\pi_{i+1}}^{\delta_s}$ for any state s and the performance difference lemma that

$$\begin{aligned}
0 &\leq \left\langle \widehat{\mathcal{Q}}_i, \pi_{i+1} - \pi_i \right\rangle_{\mathfrak{v}_{\pi_{i+1}}^{\delta_s}} \\
&= \left\langle \widehat{\mathcal{Q}}_i - \mathcal{Q}_i, \pi_{i+1} - \pi_i \right\rangle_{\mathfrak{v}_{\pi_{i+1}}^{\delta_s}} + \left\langle \mathcal{Q}_i, \pi_{i+1} - \pi_i \right\rangle_{\mathfrak{v}_{\pi_{i+1}}^{\delta_s}} \\
&\leq 2\hat{\epsilon}_i + (1-\gamma)\left(V_{i+1}(\delta_s) - V_i(\delta_s)\right)
\end{aligned}$$

which rearranges to give

$$V_{i+1}(\delta_s) \geq V_i(\delta_s) - \frac{2\hat{\epsilon}_i}{1-\gamma} \qquad \forall s,$$

which itself in turn implies

$$\begin{aligned}
\left\langle \widehat{\mathcal{Q}}_{i+1} - \widehat{\mathcal{Q}}_i, \pi_{i+1} \right\rangle_{\mathfrak{v}_\pi^\mu} &\geq \left\langle \mathcal{Q}_{i+1} - \mathcal{Q}_i, \pi_{i+1} \right\rangle_{\mathfrak{v}_\pi^\mu} - \hat{\epsilon}_i - \hat{\epsilon}_{i+1} \\
&= \gamma \mathbb{E}_{\substack{s \sim \mathfrak{v}_\pi^\mu \\ a \sim \pi_{i+1}(s,\cdot) \\ s' \sim (s,a)}} \left(\mathcal{V}_{i+1}(\delta_{s'}) - \mathcal{V}_i(\delta_{s'})\right) - \hat{\epsilon}_i - \hat{\epsilon}_{i+1} \\
&\geq -\frac{2\gamma\hat{\epsilon}_i}{1-\gamma} - \hat{\epsilon}_i - \hat{\epsilon}_{i+1}.
\end{aligned}$$

(The preceding derivation works if $\mathfrak{v}_\pi^\mu$ is replaced with any measure on states.) Plugging this all back in to eq. (B.5) gives

$$\begin{aligned}
\mathcal{V}_{t-1}(\mu) - \mathcal{V}_\pi(\mu) &\geq \frac{1}{t}\sum_{i<t}\left(\mathcal{V}_i(\mu) - \mathcal{V}_\pi(\mu)\right) - \sum_{i<t}\frac{2(1+i)\epsilon_i}{t(1-\gamma)} \\
&\geq \frac{K_{\mathfrak{v}_\pi^\mu}(\pi, \pi_t) - K_{\mathfrak{v}_\pi^\mu}(\pi, \pi_0)}{t\theta(1-\gamma)} - \frac{1}{t(1-\gamma)^2} - \frac{1}{t(1-\gamma)}\sum_{i<t}\left(\frac{2\gamma\epsilon_i}{1-\gamma} + \epsilon_i + \epsilon_{i+1}\right) \\
&\quad - \sum_{i<t}\frac{2(1+i)\epsilon_i}{t(1-\gamma)},
\end{aligned}$$

where the last term may be omitted for the summation version, and these expressions rearrange to give the final bounds.

2. Now consider the case of ℓ_2 bounded errors. Then by the above definitions and the first equality in Lemma B.1,

$$-\sum_{i<t}\left\langle\!\left\langle \theta_i g_i, q_i - q_\pi \right\rangle\!\right\rangle = D_*(q, q_t) - D_*(q, q_0) - \sum_{i<t} D(p_{i+1}, p_i),$$

where the last term may be bounded in a way common to the online learning literature: since $e^z \leq 1 + z + z^2$ when $z \leq 1$, setting $Z(s,a) := \widehat{\mathcal{Q}}_i(s,a) - 1/(1-\gamma)$ for convenience (whereby

$Z(s,a) \leq 0$ as needed by the preceding inequality),

$$
\begin{aligned}
&D(p_{i+1}, p_i) \\
&= \mathbb{E}_{s\sim\mathfrak{v}_\pi^\mu}\left(\ln\sum_a \exp(p_{i+1}(s,a)) - \ln\sum_a \exp(p_i(s,a)) - \sum_a \pi_i(s,a)(p_{i+1}(s,a) - p_i(s,a))\right) \\
&= \mathbb{E}_{s\sim\mathfrak{v}_\pi^\mu}\left(\ln\left(\sum_a \pi_i(s,a)\exp\left(\theta\widehat{\mathcal{Q}}_i(s,a)\right)\right) - \theta\sum_a \pi_i(s,a)\widehat{\mathcal{Q}}_i(s,a)\right) \\
&= \mathbb{E}_{s\sim\mathfrak{v}_\pi^\mu}\left(\ln\left(\sum_a \pi_i(s,a)\exp\left(\theta Z(s,a)\right)\right) - \theta\sum_a \pi_i(s,a)Z(s,a)\right) \\
&\leq \mathbb{E}_{s\sim\mathfrak{v}_\pi^\mu}\left(\ln\left(\sum_a \pi_i(s,a)(1+\theta Z(s,a)+\theta^2 Z(s,a)^2)\right) - \theta\sum_a \pi_i(s,a)Z(s,a)\right) \\
&\leq \mathbb{E}_{s\sim\mathfrak{v}_\pi^\mu}\left(\sum_a \pi_i(s,a)(1+\theta Z(s,a)+\theta^2 Z(s,a)^2) - 1 - \theta\sum_a \pi_i(s,a)Z(s,a)\right) \\
&= \mathbb{E}_{s\sim\mathfrak{v}_\pi^\mu}\sum_a \pi_i(s,a)\theta^2 Z(s,a)^2 \leq \frac{\theta^2}{(1-\gamma)^2}.
\end{aligned}
$$

which together with the preceding as well as eq. (B.5) gives

$$
\begin{aligned}
&\theta(1-\gamma)\sum_{i<t}\left(\mathcal{V}_i(\mu) - \mathcal{V}_\pi(\mu)\right) \\
&= \sum_{i<t}\theta\left\langle \mathcal{Q}_i - \widehat{\mathcal{Q}}_i, \pi_i - \pi\right\rangle_{\mathfrak{v}_\pi^\mu} - \sum_{i<t}\left\langle\!\left\langle \theta_i g_i, q_i - q_\pi\right\rangle\!\right\rangle. \\
&\geq \sum_{i<t}\theta\left\langle \mathcal{Q}_i - \widehat{\mathcal{Q}}_i, \pi_i - \pi\right\rangle_{\mathfrak{v}_\pi^\mu} + K_{\mathfrak{v}_\pi^\mu}(\pi,\pi_t) - K_{\mathfrak{v}_\pi^\mu}(\pi,\pi_0) - \frac{t\theta^2}{(1-\gamma)^2},
\end{aligned}
$$

which after dividing both sides by $t\theta(1-\gamma)$ gives the bound.

□

C SAMPLING PROOFS

To start, here is the proof controlling various probability ratios and mixing times uniformly over KL balls.

Proof of Lemma 3.1. Since $\tilde{\pi}$ is irreducible and aperiodic, it has a stationary distribution, and moreover this stationary distribution is positive on all states (Levin et al., 2006, Proposition 1.19), and define $\tilde{\pi}_{\min} := \min_{s\in\mathcal{S}}\tilde{\pi}(s) > 0$. By definition of $K_{\mathfrak{s}_{\tilde{\pi}}}$ and $\mathcal{P}_c$, for any $\pi \in \mathcal{P}_c$, letting (s', a') denote any pair which maximizes $\tilde{\pi}(s,a)/\pi(s,a)$, and since $\sum_a v_a \ln v_a \geq -\ln k$ for any $v \in \Delta_k$,

$$
\begin{aligned}
c \geq K_\nu(\tilde{\pi},\pi) &= \sum_{s\in\mathcal{S}}\nu(s)\sum_a \tilde{\pi}(s,a)\ln\frac{\tilde{\pi}(s,a)}{\pi(s,a)} \\
&\geq \nu(s')\tilde{\pi}(s',a')\ln\frac{\tilde{\pi}(s',a')}{\pi(s',a')} + \sum_{(s,a)\neq(s',a')}\nu(s)\sum_a \tilde{\pi}(s,a)\ln\frac{\tilde{\pi}(s,a)}{\pi(s,a)} \\
&\geq \nu(s')\tilde{\pi}(s',a')\ln\frac{\tilde{\pi}(s',a')}{\pi(s',a')} + \nu(s')\sum_{a\neq a'}\tilde{\pi}(s',a)\ln\frac{\tilde{\pi}(s',a)}{\pi(s',a)} \\
&\geq \nu(s')\tilde{\pi}(s',a')\ln\frac{\tilde{\pi}(s',a')}{\pi(s',a')} - \nu(s')\ln k,
\end{aligned}
$$

then

$$\frac{\tilde{\pi}(s',a')}{\pi(s',a')} \leq \exp\left(\frac{c}{\nu(s')\tilde{\pi}(s',a')} + \frac{\ln k}{\tilde{\pi}(s',a')}\right),$$

where the right hand side does not depend on π; it follows that $\tilde{\pi}(s,a)/\pi(s,a)$ is uniformly upper bounded over $\mathcal{P}_c$. On the other hand, $\tilde{\pi}(s,a)/\pi(s,a) \geq \tilde{\pi}(s,a)$, so the ratio uniformly bounded in both directions over $\mathcal{P}_c$, and let C_0 denote this ratio. This in turn implies that the ratio of stationary distributions is also uniformly bounded; for this, first recall that the stationary distribution can be characterized in terms of hitting times (Levin et al., 2006, Proposition 1.19), meaning

$$\mathfrak{s}_{\tilde{\pi}}(s) = \frac{1}{\mathbb{E}[\min\{t > 0 : s_t = s\}|s_0 = s]}.$$

As discussed in proofs that the denominator is finite (Levin et al., 2006, proof of Lemma 1.13), letting P_π and $P_{\tilde{\pi}}$ corresponding to the state transitions induced by taking a step by π and $\tilde{\pi}$ respectively and then using the MDP dynamics to get to a state, there exist $r > 0$ and $\epsilon > 0$ so that $P^j_{\tilde{\pi}}(s,s') > \epsilon$ for any states s, s' and any $j \geq r$, therefore $P^j_\pi(s,s') \geq \epsilon C_0^{-j}$. In fact, for a fixed s, letting $j_s \geq 1$ denote the smallest exponent so that $P^{j_s}_{\tilde{\pi}}(s,s) > 0$, then $\mathbb{E}_{\tilde{\pi}}[\min\{t > 0 : s_t = s\}|s_0 = s] \geq j_s$, meanwhile, as in Levin et al. (2006, proof of Lemma 1.13), defining $\tau^+_\pi := \min t > 0 : s_t = s$,

$$\begin{aligned}
\mathbb{E}[\tau^+_\pi|s_0 = s] &= \sum_{t\geq 0} \mathrm{Pr}[\tau^+_\pi > t|s_0 = s] \\
&\leq \sum_{k\geq 0} j_s \mathrm{Pr}[\tau^+_\pi > kj_s|s_0 = s] \\
&\leq j_s \sum_{k\geq 0} (1 - C_0^{-j_s} P^{j_s}_{\tilde{\pi}}(s,s))^k \\
&\leq \frac{j_s}{1 - C_0^{-j_s} P^{j_s}_{\tilde{\pi}}(s,s)} \\
&\leq \frac{\mathbb{E}[\tau^+_{\tilde{\pi}}|s_0 = s]}{1 - C_0^{-j_s} P^{j_s}_{\tilde{\pi}}(s,s)},
\end{aligned}$$

where the denominator does not depend on π and thus the ratio is uniformly bounded over $\mathcal{P}_c$. (We can produce a bound in the reverse direction trivially, by using $\mathfrak{s}_\pi(s)/\mathfrak{s}_{\tilde{\pi}}(s) \leq 1/\mathfrak{s}_{\tilde{\pi}}(s)$.) Let C denote the maximum of this ratio and C_0.

The proof of uniform mixing times is now a consequence of these bounded probability ratios. Recall the definition of *conductance*:

$$\Phi^*_\pi := \min_{S\subseteq|\mathcal{S}|:\mathfrak{s}_\pi(S)\leq 1/2} \Phi_\pi(S), \qquad \text{where } \Phi_\pi(S) := \frac{\sum_{s\in S}\sum_{s'\notin S}\mathfrak{s}_\pi(s)P_\pi(s,s')}{\mathfrak{s}_\pi(S)};$$

as established by Jerrum & Sinclair (1988) (see also Levin et al. (2006, eq. (7.7), Theorem 12.4, Theorem 13.10)), uniform mixing constants m_1, m_2 can be chosen purely as a function of (a lower bound on) Φ^*_π, therefore it suffices to bound $\inf\{\Phi^*_\pi/\Phi^*_{\tilde{\pi}} : \pi \in \mathcal{P}_c\}$. But in that case we are already done: using the definition of C and of Φ^*, we immediately obtain $\Phi^*_\pi \geq \Phi^*_{\tilde{\pi}}/C^4$. □

Turning to the TD guarantees, the first step is to characterize the fixed points of the TD update, which in turn motivates the linear MDP assumption (cf. Assumption 1.3).

Lemma C.1. *Suppose x is distributed according to the stationary distribution of π, and let x' denote a subsequent state, and suppose $\mathbb{E}xx^\intercal$ is invertible. Then*

$$\bar{u} := \sum_{t\geq 0}\left(\gamma\left[\mathbb{E}xx^\intercal\right]^{-1}\mathbb{E}x(x')^\intercal\right)^t\left[\mathbb{E}xx^\intercal\right]^{-1}\mathbb{E}xr$$

is a fixed point of the expected TD update, meaning

$$\mathbb{E}\bar{u} = \mathbb{E}\left(\bar{u} - \eta x\left(\langle x - \gamma x', \bar{u}\rangle - r\right)\right) = \bar{u}.$$

Moreover, under the linear MDP assumption (cf. Assumption 1.3), then $x^\intercal_{sa}\bar{u} = \mathcal{Q}_\pi(s,a)$.

Proof. For the fixed point claim, it suffices to show

$$\mathbb{E}x\left(\langle x-\gamma x', \bar{u}\rangle - r\right) = 0,$$

but this is direct from the definition of $\bar{u}$, since

$$\begin{aligned}
&\mathbb{E}x\left(\langle x-\gamma x, \bar{u}\rangle - r\right)\\
&= \left(\left[\mathbb{E}xx^\intercal\right] - \gamma\left[\mathbb{E}x(x')^\intercal\right]\right)\bar{u} - \mathbb{E}xr\\
&= \left(\left[\mathbb{E}xx^\intercal\right] - \gamma\left[\mathbb{E}x(x')^\intercal\right]\right)\sum_{t\geq 0}\left(\gamma\left[\mathbb{E}xx^\intercal\right]^{-1}\mathbb{E}x(x')^\intercal\right)^t\left[\mathbb{E}xx^\intercal\right]^{-1}\mathbb{E}xr - \mathbb{E}xr\\
&= \left[\mathbb{E}xx^\intercal\right]\sum_{t\geq 0}\left(\gamma\left[\mathbb{E}xx^\intercal\right]^{-1}\mathbb{E}x(x')^\intercal\right)^t\left[\mathbb{E}xx^\intercal\right]^{-1}\mathbb{E}xr\\
&\quad - \left[\mathbb{E}xx^\intercal\right]\sum_{t\geq 1}\left(\gamma\left[\mathbb{E}xx^\intercal\right]^{-1}\mathbb{E}x(x')^\intercal\right)^t\left[\mathbb{E}xx^\intercal\right]^{-1}\mathbb{E}xr - \mathbb{E}xr\\
&= \left[\mathbb{E}xx^\intercal\right]\left(\gamma\left[\mathbb{E}xx^\intercal\right]^{-1}\mathbb{E}x(x')^\intercal\right)^0\left[\mathbb{E}xx^\intercal\right]^{-1}\mathbb{E}xr - \mathbb{E}xr = 0.
\end{aligned}$$

For the claims under the linear MDP assumption, using the notation from Assumption 1.3, letting X_π denote the transition mapping induced by π, note by the tower property that

$$\mathbb{E}\left(x(x')^\intercal\right) = \mathbb{E}\left(x\mathbb{E}\left[(x')^\intercal|x\right]\right) = \mathbb{E}\left(x\mathbb{E}\left[(X_\pi v')^\intercal|x\right]\right) = \mathbb{E}\left(x(X_\pi M x)^\intercal\right) = \mathbb{E}xx^\intercal M^\intercal X_\pi^\intercal,$$

and therefore

$$\begin{aligned}
x_{sa}^\intercal\bar{u} &= x_{sa}^\intercal\sum_{t\geq 0}\left(\gamma\left[\mathbb{E}xx^\intercal\right]^{-1}\mathbb{E}x(x')^\intercal\right)^t\left[\mathbb{E}xx^\intercal\right]^{-1}\mathbb{E}xr\\
&= x_{sa}^\intercal\sum_{t\geq 0}\left(\gamma\left[\mathbb{E}xx^\intercal\right]^{-1}\mathbb{E}xx^\intercal M^\intercal X_\pi\right)^t\left[\mathbb{E}xx^\intercal\right]^{-1}\mathbb{E}xx^\intercal y\\
&= x_{sa}^\intercal\sum_{t\geq 0}\left(\gamma M^\intercal X_\pi^\intercal\right)^t y\\
&= \mathcal{Q}_\pi(s,a).
\end{aligned}$$

□

Now comes the core TD guarantee. This guarantee is most similar to one in (Bhandari et al., 2018), however that work employs a projection step to control the norm of the Q function estimate. Here, instead, another *implicit regularization* property, now internal to TD, is used to control the norm without projection.

Lemma C.2. *Fix a policy π, and let P_π denote the transition function for the corresponding induced Markov chain on the state space of the MDP, and let $((x_{i,0}, x_{i,1}, \ldots, x_{i,B}))_{i<N}$ evolve according to P_π with $x_{i,0} = x_{i-1,B}$, with associated rewards $(r_{i,j})_{j\leq B, i<N}$, and assume the following mixing time for P_π:*

$$\sup_s \|P_\pi^t(s) - \mathfrak{s}_\pi(s)\|_{\mathrm{TV}} \leq me^{-ct}.$$

Consider the minibatch TD updates

$$u_{i+1} := u_i - \frac{\eta}{B}\sum_{j<B} x_{i,j}\left(\langle x_{i,j} - \gamma x_{i,j+1}, u_i\rangle - r_{i,j}\right),$$

with the initial condition $\|u_0 - \bar{u}\| \leq 1/(1-\gamma)$. and any step size $\eta \leq \frac{1-\gamma}{4(1+\gamma^2)}$. Let $\bar{u}$ denote a fixed point of the Bellman operator as in Lemma C.1. If $B \geq \frac{192mN\eta}{(1-e^{-c})}$, then

$$\begin{aligned}
&\mathop{\mathbb{E}}_{(x_{i,j})_{j\leq B, i<N}} \mathop{\mathbb{E}}_{x\sim(\mathfrak{s}_\pi,\pi)} \left\langle x, \frac{1}{N}\sum_{i<N} u_i - \bar{u} \right\rangle^2 \\
&\leq \frac{1}{N}\sum_{i<N} \mathop{\mathbb{E}}_{(x_{i,j})_{j\leq B, i<N}} \mathop{\mathbb{E}}_{x\sim(\mathfrak{s}_\pi,\pi)} \langle x, u_i - \bar{u}\rangle^2 \\
&\leq \frac{1}{N\eta(1-\gamma)}\left[\frac{1}{(1-\gamma)^2} + \|u_0 - \bar{u}\|^2 - \mathbb{E}_{((x_{i,j}))}\|u_N - \bar{u}\|^2\right].
\end{aligned}$$

Proof. Since the policy π is fixed throughout, we will use the shorthand $\mathfrak{s} = \mathfrak{s}_\pi$. The proof proceeds by induction, and will maintain $\mathbb{E}\|u - \bar{u}\|^2 \leq 2/(1-\gamma)^2$ throughout.

Let $\widehat{\mathcal{T}}_i$ denote the affine function corresponding to the minibatch Bellman update at time i, meaning $u_{i+1} := u_i - \eta\widehat{\mathcal{T}}_i(u_i)$, and let $\mathcal{T}_\mathfrak{s}$ denote the expected Bellman update if the change is run from the stationary distribution $\mathfrak{s}$, whereby $\bar{u} = \bar{u} - \eta\mathcal{T}_\mathfrak{s}(\bar{u})$.

The first part of the proof controls the difference between $\widehat{\mathcal{T}}_i$ and $\mathcal{T}_\mathfrak{s}$; completing the proof from there follows a usual scheme from stochastic optimization. To this end, for each $x_{i,j}$, there exist (coupled) random variables $y_{i,j}$ and $z_{i,j}$ and an event $E_{i,j}$ where $y_{i,j}$ is distributed according to $\mathfrak{s}$, and when $E_{i,j}$ holds, then $x_{i,j} = y_{i,j}$, and otherwise $\neg E_{i,j}$ corresponds to $x_{i,j} = z_{i,j}$. By an equivalent characterization of total variation in terms of couplings (Levin et al., 2006, Proposition 4.7),

$$\Pr[\neg E_{i,j}] = \|x_{i,j} - \mathfrak{s}\|_{\text{TV}} \leq me^{-cj}.$$

Letting $\widehat{\mathcal{T}}(u; x_{i,j})$ denote the operator using $(x_{i,j}, r_{i,j}, x_{i,j+1})$, whereby $\mathbb{E}\widehat{\mathcal{T}}(\cdot; y_{i,j}) = \mathcal{T}_\mathfrak{s}$ and $\widehat{\mathcal{T}}_i = B^{-1}\sum_j \widehat{\mathcal{T}}(\cdot; x_{i,j})$, and noting $\|\bar{u}\| \leq 1/(1-\gamma)$ and $\|\widehat{\mathcal{T}}(u; z_{i,j})\| \leq (1+\gamma)\|u\| + 1$, and similarly for $\|\mathcal{T}_\mathfrak{s}(u)\|$, using the inductively assumed bound $\mathbb{E}\|u_i - \bar{u}\|^2 \leq 2/(1-\gamma)^2$,

$$\begin{aligned}
\mathbb{E}\left\langle u_i - \bar{u}, \widehat{\mathcal{T}}_i(u_i) - \mathcal{T}_\mathfrak{s}(u_i)\right\rangle &= \frac{1}{B}\mathbb{E}\sum_{j<B}\left\langle u_i - \bar{u}, \widehat{\mathcal{T}}(u_i; x_{i,j}) - \mathcal{T}_\mathfrak{s}(u_i)\right\rangle \\
&= \frac{1}{B}\mathbb{E}\sum_{j<B}\left\langle u_i - \bar{u}, \widehat{\mathcal{T}}(u_i; y_{i,j}) - \mathcal{T}_\mathfrak{s}(u_i)\right\rangle \Pr[E_{i,j}] \\
&\quad + \frac{1}{B}\mathbb{E}\sum_{j<B}\left\langle u_i - \bar{u}, \widehat{\mathcal{T}}(u_i; z_{i,j}) - \mathcal{T}_\mathfrak{s}(u_i)\right\rangle \Pr[\neg E_{i,j}] \\
&= \frac{1}{B}\mathbb{E}\sum_{j<B}\left\langle u_i - \bar{u}, \widehat{\mathcal{T}}(u_i; z_{i,j}) - \mathcal{T}_\mathfrak{s}(u_i)\right\rangle \|x_{i,j} - \mathfrak{s}\|_{\text{TV}} \\
&\leq \frac{1}{B}\mathbb{E}\sum_{j<B}\|u_i - \bar{u}\| \cdot \|\widehat{\mathcal{T}}(u_i; z_{i,j}) - \mathcal{T}_\mathfrak{s}(u_i)\| me^{-cj} \\
&\leq \frac{1}{2B}\mathbb{E}\sum_{j<B}\left(\|u_i - \bar{u}\|^2 + \|\widehat{\mathcal{T}}(u_i; z_{i,j}) - \mathcal{T}_\mathfrak{s}(u_i)\|^2\right) me^{-cj} \\
&\leq \frac{32m}{B(1-\gamma)^2}\left(\frac{1}{1-e^{-c}}\right)
\end{aligned}$$

and by similar reasoning

$$\begin{aligned}
\mathbb{E}\left\|\widehat{\mathcal{T}}_i(u_i) - \mathcal{T}_\mathfrak{s}(u_i)\right\|^2 &= \frac{2}{B^2}\sum_{j=0}^{B-1}\sum_{k=j}^{B-1}\left\langle \widehat{\mathcal{T}}(u_i; x_{i,j}) - \mathcal{T}_\mathfrak{s}(u_i), \widehat{\mathcal{T}}(u_i; x_{i,k}) - \mathcal{T}_\mathfrak{s}(u_i)\right\rangle \\
&\leq \frac{64}{B^2(1-\gamma)^2}\sum_{j=0}^{B-1}\sum_{k=j}^{B-1} me^{-c(k-j)} \\
&\leq \frac{64m}{B(1-\gamma)^2}\left(\frac{1}{1-e^{-c}}\right).
\end{aligned}$$

(Note that u_i and $x_{i,j}$ are *not* independent, but conditioning on u_i and using a worst case mixing time circumvents this.)

Now that $\widehat{\mathcal{T}}_i$ has been relaxed to $\mathcal{T}_{\mathfrak{s}}$, the next technical point is to control the gap between $\mathcal{T}_{\mathfrak{s}}(u_i)$ and $\mathcal{T}_{\mathfrak{s}}(\bar{u})$. First note that

$$\mathcal{T}_{\mathfrak{s}}(u_i) - \mathcal{T}_{\mathfrak{s}}(\bar{u}) = \mathbb{E}\left(x\left\langle x - \gamma x', u_i\right\rangle - xr\right) - \mathbb{E}\left(x\left\langle x - \gamma x', \bar{u}\right\rangle - xr\right) = \mathbb{E}x\left\langle x - \gamma x', u_i - \bar{u}\right\rangle,$$

and thus

$$\begin{aligned}\left\langle u_i - \bar{u}, \mathcal{T}_{\mathfrak{s}}(u_i) - \mathcal{T}_{\mathfrak{s}}(\bar{u})\right\rangle &= \mathbb{E}\left\langle u_i - \bar{u}, x\left\langle x - \gamma x', u_i - \bar{u}\right\rangle\right\rangle \\ &= \mathbb{E}\left\langle u_i - \bar{u}, x\right\rangle^2 - \gamma\mathbb{E}\left\langle u_i - \bar{u}, x\right\rangle\left\langle x', u_i - \bar{u}\right\rangle \\ &\geq \mathbb{E}\left\langle u_i - \bar{u}, x\right\rangle^2 - \gamma\mathbb{E}\left\langle u_i - \bar{u}, x\right\rangle^2 \\ &= (1-\gamma)\mathbb{E}\left\langle u_i - \bar{u}, x\right\rangle^2,\end{aligned}$$

and

$$\begin{aligned}&\left\|\mathcal{T}_{\mathfrak{s}}(u_i) - \mathcal{T}_{\mathfrak{s}}(\bar{u})\right\|^2 \\ &= \left\langle \mathbb{E}x\left\langle x - \gamma x', u_i - \bar{u}\right\rangle, \mathbb{E}x\left\langle x - \gamma x', u_i - \bar{u}\right\rangle\right\rangle \\ &= \mathbb{E}_{x,x',z,z'}\left\langle x, z\right\rangle\left\langle x - \gamma x', u_i - \bar{u}\right\rangle\left\langle z - \gamma z', u_i - \bar{u}\right\rangle \\ &= \mathbb{E}_{x,x',z,z'}\left\langle x, z\right\rangle\left(\left\langle x, u_i - \bar{u}\right\rangle\left\langle z, u_i - \bar{u}\right\rangle - 2\gamma\left\langle x', u_i - \bar{u}\right\rangle\left\langle z, u_i - \bar{u}\right\rangle + \gamma^2\left\langle x', u_i - \bar{u}\right\rangle\left\langle z', u_i - \bar{u}\right\rangle\right) \\ &\leq 2(1+\gamma^2)\mathbb{E}\left\langle x, u_i - \bar{u}\right\rangle^2,\end{aligned}$$

and together, using $\eta \leq \frac{1-\gamma}{4(1+\gamma^2)}$,

$$\begin{aligned}-\left\langle u_i - \bar{u}, \mathcal{T}_{\mathfrak{s}}(u_i) - \mathcal{T}_{\mathfrak{s}}(\bar{u})\right\rangle + \eta\left\|\mathcal{T}_{\mathfrak{s}}(u_i) - \mathcal{T}_{\mathfrak{s}}(\bar{u})\right\|^2 &\leq -(1-\gamma)\mathbb{E}\left\langle u_i - \bar{u}, x\right\rangle^2 + 2\eta(1+\gamma^2)\mathbb{E}\left\langle x, u_i - \bar{u}\right\rangle^2 \\ &= -\frac{1-\gamma}{2}\mathbb{E}\left\langle x, u_i - \bar{u}\right\rangle^2.\end{aligned}$$

Using these facts about $\widehat{\mathcal{T}}_i$ and $\mathcal{T}_{\mathfrak{s}}$,

$$\begin{aligned}\mathbb{E}\|u_{i+1} - \bar{u}\|^2 &= \mathbb{E}\left\|u_i - \bar{u} - \eta\left(\widehat{\mathcal{T}}_i(u_i) - \mathcal{T}_{\mathfrak{s}}(\bar{u})\right)\right\|^2 \\ &= \mathbb{E}\|u_i - \bar{u}\|^2 - 2\eta\mathbb{E}\left\langle u_i - \bar{u}, \widehat{\mathcal{T}}_i(u_i) - \mathcal{T}_{\mathfrak{s}}(\bar{u})\right\rangle + \eta^2\mathbb{E}\left\|\widehat{\mathcal{T}}_i(u_i) - \mathcal{T}_{\mathfrak{s}}(\bar{u})\right\|^2 \\ &= \mathbb{E}\|u_i - \bar{u}\|^2 - 2\eta\mathbb{E}\left\langle u_i - \bar{u}, \widehat{\mathcal{T}}_i(u_i) - \mathcal{T}_{\mathfrak{s}}(u_i)\right\rangle - 2\eta\mathbb{E}\left\langle u_i - \bar{u}, \mathcal{T}_{\mathfrak{s}}(u_i) - \mathcal{T}_{\mathfrak{s}}(\bar{u})\right\rangle \\ &\quad + 2\eta^2\mathbb{E}\left\|\widehat{\mathcal{T}}_i(u_i) - \mathcal{T}_{\mathfrak{s}}(u_i)\right\|^2 + 2\eta^2\mathbb{E}\left\|\mathcal{T}_{\mathfrak{s}}(u_i) - \mathcal{T}_{\mathfrak{s}}(\bar{u})\right\|^2 \\ &\leq \mathbb{E}\|u_i - \bar{u}\|^2 - \eta(1-\gamma)\mathbb{E}\left\langle x, u_i - \bar{u}\right\rangle^2 + \frac{64m\eta(1+2\eta)}{B(1-\gamma)^2}\left(\frac{1}{1-e^{-c}}\right),\end{aligned}$$

which after rearranging and telescoping and using $\eta \leq 1$ and $B \geq \frac{192mN\eta}{(1-e^{-c})}$ gives

$$\begin{aligned}\mathbb{E}\sum_{j<i+1}\mathbb{E}_{x\sim\mathfrak{s}}\left\langle x, u_j - \bar{u}\right\rangle^2 &\leq \frac{1}{\eta(1-\gamma)}\left[\|u_0 - \bar{u}\|^2 - \mathbb{E}\|u_{i+1} - \bar{u}\|^2\right] + \frac{192m(i+1)}{B(1-\gamma)^3}\left(\frac{1}{1-e^{-c}}\right) \\ &\leq \frac{1}{\eta(1-\gamma)}\left[\frac{1}{(1-\gamma)^2} + \mathbb{E}\|u_0 - \bar{u}\|^2 - \mathbb{E}\|u_{i+1} - \bar{u}\|^2\right],\end{aligned}$$

which implies $\mathbb{E}\|u_{i+1} - \bar{u}\|^2 \leq 2/(1-\gamma)^2$ since $\|u_0 - \bar{u}\|^2 \leq 1/(1-\gamma)^2$.

The proof is completed by here by instantiating the above inequality for $i + 1 = N$ and using Jensen's inequality to obtain a guarantee on the averaged iterate. □

Lastly, the proof of Lemma 3.3 in the body is a quick consequence of Lemma C.2 and Lemma C.1.

Proof of Lemma 3.3. The proof is via Lemma C.2, which gives a more general statement, but also substitutes in $\mathcal{Q}_i$ via Lemma C.1. □

D PROOF OF THEOREM 1.4

Combining the mirror descent and sampling tools, we can finally prove Theorem 1.4.

Proof of Theorem 1.4. Define the following KL-bounded subset of policy space:

$$\mathcal{P} := \bigcap_{s\in\mathcal{S}} \mathcal{P}_s, \qquad \text{where } \mathcal{P}_s := \left\{\pi : K_{\mathfrak{v}^s_{\overline{\pi}}}(\overline{\pi},\pi) \le 1 + \ln k + \frac{1}{(1-\gamma)^2}\right\}.$$

By $|\mathcal{S}|$ applications of Lemma 3.1 (one for each $\mathcal{P}_s$) and taking maxima/minima of the resulting constants, since $\mathfrak{s}_{\overline{\pi}}$ is positive for every state, then there exist constants $p_{\min} > 0$, $C_1 > 0$, and $C_2 \ge 1$ so that, for any $\pi \in \mathcal{P}$ and any state s and any optimal action $a \in \mathcal{A}_s$,

$$p_{\min} \le \mathfrak{s}_\pi(s), \qquad \pi(s,a) \ge \frac{\overline{\pi}(s,a)}{C_2},$$

and letting P_π denote the transition matrix on the induced chain on $\mathcal{S}$, for any time t,

$$\max_{s\in\mathcal{S}} \|P^t_\pi(s) - \mathfrak{s}_\pi(s)\|_{\mathrm{TV}} \le C_2 \exp(-C_1 t). \tag{D.1}$$

Lastly, the fully specified constants N and B are

$$N := t^3 \qquad \text{and} \qquad B := \frac{192 C_2 N \eta}{1-\exp(-C_1)},$$

where $B = \mathcal{O}(\eta N) = \mathcal{O}(t^3)$ as in the statement.

The proof establishes the following inequalities inductively:

$$K_{\mathfrak{v}^s_{\overline{\pi}}}(\overline{\pi},\pi_i) + \theta(1-\gamma)\sum_{j<i}\left(\mathcal{V}_j(s) - \mathcal{V}_{\overline{\pi}}(s)\right) \le 1 + \ln k + \frac{1}{(1-\gamma)^2}, \qquad \forall s,$$

$$\left\langle \mathcal{Q}_i - \widehat{\mathcal{Q}}_i, \pi_i - \pi\right\rangle_{\delta_s} \le \theta, \qquad \forall s,$$

where the first implies $\pi_i \in \mathcal{P}$ directly, and instantiating these inequalities with $i = t$ and rearranging gives the desired statement.

To establish the inductive claim, consider some i, and suppose the inductive hypothesis holds for $j < i$. The first inequality comes more easily: by the inductive hypothesis on $\mathcal{Q}_i - \widehat{\mathcal{Q}}_i$ and since $K_\nu(\overline{\pi},\pi_0) \le \ln k$ for any measure ν, then the second part of Lemma 2.1 instantiated for the first i iterations for any s gives

$$\begin{aligned} K_{\mathfrak{v}^s_{\overline{\pi}}}(\overline{\pi},\pi_i) + \theta(1-\gamma)\sum_{j<i}\left(\mathcal{V}_{\overline{\pi}}(s) - \mathcal{V}_j(s)\right) &\le K_{\mathfrak{v}^s_{\overline{\pi}}}(\overline{\pi},\pi_0) + \theta\sum_{j<i}\left\langle \mathcal{Q}_j - \widehat{\mathcal{Q}}_j, \pi_j - \overline{\pi}\right\rangle_{\mathfrak{v}^s_{\overline{\pi}}} + \frac{i\theta^2}{(1-\gamma)^2} \\ &\le \ln k + 1 + \frac{1}{(1-\gamma)^2}, \end{aligned}$$

as desired. For the second inequality, since $\pi \in \mathcal{P}$, the mixing constants in eq. (D.1) hold, which after plugging the above choice of N and B into Lemma C.2 (and using Lemma C.1 and Assumption 1.3 to guarantee the output of Lemma C.2 approximates the true $\mathcal{Q}_i$ and is not merely a TD fixed point) gives

$$\mathbb{E}_{\widehat{\mathcal{Q}}_i}\mathbb{E}_{(s,a)\sim(\mathfrak{s}_i,\pi_i)}(\widehat{\mathcal{Q}}_i(s,a) - \mathcal{Q}_i(s,a))^2 \le \frac{2}{N\eta(1-\gamma)^3} \le \frac{2}{t^3\eta(1-\gamma)^3}.$$

By Markov's inequality,

$$\begin{aligned} \Pr_{\widehat{\mathcal{Q}}_i}\left[\mathbb{E}_{(s,a)}(\widehat{\mathcal{Q}}_i(s,a) - \mathcal{Q}_i(s,a))^2 > \frac{p^2_{\min}}{4C_2^2 t}\right] &\le \frac{4C_2^2 t \mathbb{E}_{\widehat{\mathcal{Q}}_i}\mathbb{E}_{(s,a)}(\widehat{\mathcal{Q}}_i(s,a) - \mathcal{Q}_i(s,a))^2}{p^2_{\min}} \\ &= \mathcal{O}\left(\frac{1}{t^2}\right), \end{aligned}$$

and henceforth discard this failure event (the whole proof unions and discards all t such events, together giving the $\mathcal{O}(1/t)$ failure probability in the statement). Using this bound, and noting that $\overline{\pi}(s,b) = 0$ for $b \notin \mathcal{A}_s$,

$$\begin{aligned}
\max_s \left\langle \mathcal{Q}_i - \widehat{\mathcal{Q}}_i, \pi_i - \pi \right\rangle_{\delta_s} &\leq \frac{1}{p_{\min}} \left\langle \mathcal{Q}_i - \widehat{\mathcal{Q}}_i, \pi_i - \pi \right\rangle_{\mathfrak{s}_{\pi_i}} \\
&= \frac{1}{p_{\min}} \mathop{\mathbb{E}}_{(s,a)\sim(\mathfrak{s}_i,\pi_i)} (\mathcal{Q}_i(s,a) - \widehat{\mathcal{Q}}_i(s,a)) \\
&\quad + \frac{1}{p_{\min}} \mathop{\mathbb{E}}_{s\sim\mathfrak{s}_i} \sum_{a\in\mathcal{A}_s} \overline{\pi}(s,a)(\widehat{\mathcal{Q}}_i(s,a) - \mathcal{Q}_i(s,a)) \\
&\leq \frac{2C_2}{p_{\min}} \mathop{\mathbb{E}}_{(s,a)\sim(\mathfrak{s}_i,\pi_i)} |\mathcal{Q}_i(s,a) - \widehat{\mathcal{Q}}_i(s,a)| \\
&\leq \frac{2C_2}{p_{\min}} \sqrt{\mathop{\mathbb{E}}_{(s,a)\sim(\mathfrak{s}_i,\pi_i)} (\mathcal{Q}_i(s,a) - \widehat{\mathcal{Q}}_i(s,a))^2} \\
&\leq \frac{2C_2}{p_{\min}} \sqrt{\frac{p_{\min}^2}{4C^2 t}} \\
&= \theta,
\end{aligned}$$

as desired. □