POLICY GRADIENT METHODS FOR RL IN GENERAL SPACES

Contents

1. Introduction

Our aim is to learn about *policy gradient* methods for solving reinforcement learning (RL) problems modelled using the Markov decision problem (MDP) framework with general (possibly continuous, possibly infinite dimensional) state and action spaces. We will focus mainly on theoretical convergence of mirror descent with direct parametrisation and natural-gradient descent when employing log-linear parametrisation. For our purposes solving an RL problem means that we find a (nearly) optimal policy in a situation where the transition dynamics and costs are unknown but we can repeatedly interact with some system (or environment simulator).

There are two main approaches to solving RL problems: *action-value* methods which learn the state-action value function (the Q-function) and then select actions based on this. Their convergence is understood Watkins and Dayan [1992], [Sutton and Barto, 2018, Ch. 6] and will not be discussed here. Policy gradient methods directly update the policy by stepping in the direction of the gradient of the value function and have a long history for which the reader is referred to [Sutton and Barto, 2018, Ch. 13]. Their convergence is only understood in specific settings, as we will see below. The focus here is to cover generic (Polish) state and action spaces. We will touch upon the popular PPO algorithm Schulman et al. [2017] and explain the difficulties arising when trying to prove convergence of PPO.

Many related and interesting questions will not be covered upon: convergence of actor-critic methods, convergence in presence of Monte-Carlo errors, regret, off-policy gradient methods, near continuous time RL.

Large parts of what is presented here in particular on mirror descent and natural-gradient descent is from Kerimkulov et al. [2025a]. This work was itself inspired by the recent results of Agarwal et al. [2019], Mei et al. [2020], Lan [2023] and Cayci et al. [2021] that apply mainly to finite MDPs

Date: November 8, 2025.

though e.g. Agarwal et al. [2019], Lan [2023] already note that natural policy gradient (closely related to mirror descent) has dimension independent convergence rates.

1.1. Problem formulation and the Bellman principle. In this section, we formulate the entropy-regularised MDPs with continuous state and action spaces. Let S and A be Polish spaces,[1] $P \in \mathcal{P}(S|S \times A)$, $c \in B_b(S \times A)$ and $\gamma \in [0, 1)$. The five-tuple (S, A, P, c, γ) determines an infinite horizon Markov decision model, where S and A represent the state and action spaces, respectively, P represents the transition probability, c represents the cost function and γ represents the discount factor. Let $\Pi = \{\pi = \{\pi_n\}_{n\in\mathbb{N}_0} : \pi_n \in \mathcal{P}(A|H_n)\}$ denote the set of (possibly non-Markovian) stochastic policies, where for each $n \in \mathbb{N}_0$, $H_n := (S \times A)^n \times S$ is the space of admissible histories.

Let $(\Omega := (S\times A)^{\mathbb{N}_0}, \mathcal{F})$ denote the canonical sample space, where $\mathcal{F} = \mathcal{B}(\Omega)$ is the corresponding Borel sigma-algebra. Elements of Ω are of the form $(s_0, a_0, s_1, a_1, \ldots)$ with $s_n \in S$ and $a_n \in A$ denoting the projections and called the state and action variables, at time $n \in \mathbb{N}_0$, respectively. By [**?**, Proposition 7.28], for any given initial distribution $\rho \in \mathcal{P}(S)$ and policy $\pi \in \Pi$, there exists a unique product probability measure $\mathbb{P}^\pi_\rho$ on $(\Omega, \mathcal{F})$ with expectation denoted $\mathbb{E}^\pi_\rho$ such that for all $n \in \mathbb{N}_0$, $B \in \mathcal{B}(S)$ and $C \in \mathcal{B}(A)$, $\mathbb{P}^\pi_\rho(s_0 \in B) = \rho(B)$ and

$$\mathbb{P}^\pi_\rho(a_n \in C|h_n) = \pi_n(C|h_n), \quad \mathbb{P}^\pi_\rho(s_{n+1} \in B|h_n, a_n) = P(B|s_n, a_n)\,, \tag{1}$$

where $h_n = (s_0, a_0, \ldots, s_{n-1}, a_{n-1}, s_n) \in H_n$. In particular, if π is a Markov stochastic policy (i.e., $\pi_n \in \mathcal{P}(A|S)$ for all $n \in \mathbb{N}_0$), then $\{s_n\}_{n\in\mathbb{N}_0}$ is a Markov process with kernel $\{P_{\pi,n}\}_{n\in\mathbb{N}_0} \in \mathcal{P}(S|S)$ given by

$$P_{\pi,n}(ds'|s) = \int_A P(ds'|s, a)\pi_n(da|s), \quad \forall s \in S, n \in \mathbb{N}_0\,.$$

For $s \in S$, we denote $\mathbb{E}^\pi_s = \mathbb{E}^\pi_{\delta_s}$, where $\delta_s \in \mathcal{P}(S)$ denotes the Dirac measure at $s \in S$.

Let $\mu \in \mathcal{P}(A|S)$ denote a reference kernel and $\tau \in [0, \infty)$ denote a regularisation parameter. For each $\pi = \{\pi_n\}_{n\in\mathbb{N}_0} \in \Pi$ and $s \in S$, define the following regularised value function:

$$V^\pi_\tau(s) = \mathbb{E}^\pi_s\left[\sum_{n=0}^\infty \gamma^n\Big(c(s_n, a_n) + \tau\,\mathrm{KL}(\pi_n(\cdot|h_n)|\mu(\cdot|s))\Big)\right] \in \mathbb{R} \cup \{+\infty\}\,, \tag{2}$$

which may be infinite if $\pi_n \notin \mathcal{P}_\mu(A|S)$ for some $n \in \mathbb{N}_0$, or if $\mathbb{E}^\pi_s\left[\sum_{n=0}^\infty \gamma^n\,\mathrm{KL}(\pi_n(\cdot|h_n)|\mu(\cdot|s))\right]$ diverges. Since c is bounded and $H_n \times S \ni (h_n, s) \mapsto \mathrm{KL}(\pi_n(\cdot|h_n)|\mu(\cdot|s)) \in [0, \infty]$ is non-negative and measurable, $V^\pi_\tau : S \to \mathbb{R} \cup \{+\infty\}$ is a well-defined measurable function. We define the optimal value function $V^*_\tau : S \to \mathbb{R} \cup \{+\infty\}$ by

$$V^*_\tau(s) = \inf_{\pi\in\Pi} V^\pi_\tau(s), \quad \forall s \in S\,, \tag{3}$$

and refer to $\pi^* \in \Pi$ as an optimal policy if $V^{\pi^*}_\tau(s) = V^*_\tau(s)$, for all $s \in S$.

1.2. The unregularized case $\tau = 0$. By virtue of [Hernández-Lerma and Lasserre, 2012, Theorem 4.2.3], we have the following dynamic programming principle, as long as certain assumptions guaranteeing measurable selection hold.

Assumption 1.1.
(1) The kernel $P \in \mathcal{P}(S|S \times A)$ is strongly continuous, that is: for every $v \in B_b(S)$ (bounded and measurable) the function $w(s, a) = \int_S v(s')P(\mathrm{d}s'|s, a)$ is bounded and measurable as a function from $S \times A$ to $\mathbb{R}$.
(2) The cost function $c \in B_b(S \times A)$ is lower semi-continuous and inf-compact on $S \times A$ i.e. for any $s \in S$ and any $l \in \mathbb{R}$ the set $\{a \in A : c(s, a) \le l\}$ is compact.

[1] Complete metric spaces that have a countable dense subset.

Theorem 1.2 (Dynamic programming principle, $\tau = 0$)**.** *Let Assumption 1.1 hold. Then the optimal value function $V^* \in B_b(S)$ is the unique solution of the Bellman equation*

$$V^*(s) = \min_{a\in A}\left[c(s,a) + \gamma \int_S V^*(s')P(\mathrm{d}s'|s,a)\right]. \tag{4}$$

Moreover, writing $Q^*(s,a) = c(s,a) + \gamma\int_S V^*(s')P(\mathrm{d}s'|s,a)$, there exists a measurable function $f^* : S \to A$ called a selector such that $f^*(s) \in \operatorname{argmin}_{a\in A} Q^*(s,a)$ and the induced policy $\pi^* \in \mathcal{P}(A|S)$ defined by $\pi^*(\mathrm{d}a|s) = \delta_{f^*(s)}(\mathrm{d}a)$ for all $s \in S$ satisfies $V^* = V^{\pi^*}$.

Lemma 1.3. *Let $\pi \in \mathcal{P}(A|S)$. The value function V_0^π is the unique bounded solution of the on-policy Bellman equation:*

$$V_0^\pi(s) = \int_A \left(c(s,a) + \gamma \int_S V_0^\pi(s')P(ds'|s,a)\right)\pi(da|s), \quad \forall s \in S\,.$$

1.3. The entropy regularized case $\tau > 0$.

Theorem 1.4 (Dynamic programming principle, $\tau > 0$)**.** *Let $\tau > 0$. The optimal value function V_τ^* is the unique bounded solution of the following Bellman equation:*

$$V_\tau^*(s) = \inf_{m\in\mathcal{P}(A)} \int_A \left(c(s,a) + \tau \ln\frac{\mathrm{d}m}{\mathrm{d}\mu}(a) + \gamma\int_S V_\tau^*(s')P(ds'|s,a)\right) m(da), \quad \forall s \in S,.$$

Consequently, for all $s \in S$,

$$V_\tau^*(s) = -\tau \ln \int_A \exp\left(-\frac{1}{\tau}Q_\tau^*(s,a)\right)\mu(da|s),$$

where $Q^* \in B_b(S\times A)$ is defined by

$$Q_\tau^*(s,a) = c(s,a) + \gamma\int_S V_\tau^*(s')P(ds'|s,a)\,, \quad \forall (s,a) \in S\times A.$$

Moreover, there is an optimal policy $\pi_\tau^* \in \mathcal{P}_\mu(A|S)$ given by

$$\pi_\tau^*(da|s) = \exp\left(-(Q_\tau^*(s,a) - V_\tau^*(s))/\tau\right)\mu(da|s)\,, \quad \forall s \in S. \tag{5}$$

Definition 1.5. Let Π_μ denote the class of policies $\pi = \{\pi_n\}_{n\in\mathbb{N}_0} \in \Pi$ such that there exists $f \in B_b(S\times A)$ and we have $\pi_n(da|s) = \frac{\exp(f(s,a))}{\int_A \exp(f(s,a))\mu(da|s)}\mu(da|s)$ for all $s \in S$ and $n \in \mathbb{N}_0$. We identify Π_μ with the set $\{\boldsymbol{\pi}(f) \mid f \in B_b(S\times A)\} \subset \mathcal{P}_\mu(A|S)$, where $\boldsymbol{\pi} : B_b(S\times A) \to \mathcal{P}_\mu(A|S)$ is defined by

$$\boldsymbol{\pi}(f)(da|s) = \frac{e^{f(s,a)}}{\int_A e^{f(s,a')}\mu(da'|s)}\mu(da|s), \quad \forall f \in B_b(S\times A)\,. \tag{6}$$

For each $\pi \in \Pi_\mu$, we define the Q-function $Q_\tau^\pi \in B_b(S\times A)$ by

$$Q_\tau^\pi(s,a) = c(s,a) + \gamma\int_S V_\tau^\pi(s')P(ds'|s,a)\,. \tag{7}$$

Proposition 1.6. *Let $f \in B_b(S\times A)$ and $\pi \in \Pi_\mu$ be such that $\pi(da|s) = \frac{\exp(f(s,a))\mu(da)}{\int_A \exp(f(s,a'))\mu(da')}$ for all $s \in S$. Then*

$$\left\|\ln\frac{\mathrm{d}\pi}{\mathrm{d}\mu}\right\|_{B_b(S\times A)} \le 2\|f\|_{B_b(S\times A)}\,, \quad \|V_\tau^\pi\|_{B_b(S)} \le \frac{1}{1-\gamma}\left(\|c\|_{B_b(S\times A)} + 2\tau\|f\|_{B_b(S\times A)}\right),$$

$$\|Q_\tau^\pi\|_{B_b(S\times A)} \le \frac{1}{1-\gamma}\left(\|c\|_{B_b(S\times A)} + 2\tau\gamma\|f\|_{B_b(S\times A)}\right).$$

Proof. As $\mu(A) = 1$, for all $g \in B_b(S \times A)$ and $s \in S$,

$$\ln \int_A \exp(g(s,a'))\mu(da') \le \ln\left(e^{\|g\|_{B_b(S\times A)}}\mu(A)\right) = \|g\|_{B_b(S\times A)}\,,$$
$$\ln \int_A \exp(g(s,a'))\mu(da') \ge \ln\left(e^{-\|g\|_{B_b(S\times A)}}\mu(A)\right) = -\|g\|_{B_b(S\times A)}\,.$$

Then, for all $(s,a) \in S \times A$, using $\ln \frac{\mathrm{d}\pi}{\mathrm{d}\mu}(a|s) = f(s,a) - \ln \int_A \exp(f(s,a'))\mu(da')$,

$$\left|\ln \frac{\mathrm{d}\pi}{\mathrm{d}\mu}(a|s)\right| \le |f(s,a)| + \left|\ln \int_A \exp(f(s,a'))\mu(da')\right| \le 2\|f\|_{B_b(S\times A)}\,,$$

which implies that

$$\left|\mathbb{E}_s^\pi\left[\sum_{t=0}^\infty \gamma^t\left(\tau \ln \frac{\mathrm{d}\pi}{\mathrm{d}\mu}(a_t|s_t)\right)\right]\right| \le 2\tau\|f\|_{B_b(S\times A)} \sum_{t=0}^\infty \gamma^t = \frac{2\tau\|f\|_{B_b(S\times A)}}{1-\gamma}\,.$$

By (2), for all $s \in S$,

$$\begin{aligned}|V_\tau^\pi(s)| &\le \mathbb{E}_s^\pi\left[\sum_{t=0}^\infty \gamma^t |c(s_t,a_t)|\right] + \left|\mathbb{E}_s^\pi\left[\sum_{t=0}^\infty \gamma^t\left(\tau \ln \frac{\mathrm{d}\pi}{\mathrm{d}\mu}(a_t|s_t)\right)\right]\right| \\ &\le \frac{1}{1-\gamma}\left(\|c\|_{B_b(S\times A)} + 2\tau\|f\|_{B_b(S\times A)}\right).\end{aligned}$$

Hence, for all $(s,a) \in S \times A$, by (7),

$$|Q_\tau^\pi(s,a)| \le \|c\|_{B_b(S\times A)} + \gamma\|V_\tau^\pi\|_{B_b(S)} \le \frac{1}{1-\gamma}\|c\|_{B_b(S\times A)} + \frac{2\tau\gamma}{1-\gamma}\|f\|_{B_b(S\times A)}\,.$$

This proves the desired bound of Q_τ^π. □

Lemma 1.7. *Let $\tau > 0$ and $\pi \in \Pi_\mu$. The value function V_τ^π is the unique bounded solution of the on-policy Bellman equation:*

$$V_\tau^\pi(s) = \int_A \left(c(s,a) + \tau \ln \frac{\mathrm{d}\pi}{\mathrm{d}\mu}(a|s) + \gamma \int_S V_\tau^\pi(s')P(ds'|s,a)\right)\pi(da|s), \quad \forall s \in S\,.$$

Note that from this and defn. of the Q-function (7) we have for all $\pi \in \Pi_\mu$ and $s \in S$ that

$$V_\tau^\pi(s') = \int_A \left(Q_\tau^\pi(s',a') + \tau \ln \frac{\mathrm{d}\pi}{\mathrm{d}\mu}(a'|s')\right)\pi(da'|s'), \quad \forall s \in S\,. \tag{8}$$

Using this in the defn. of the Q-function (7) we have the on policy Q-Bellman equation

$$Q_\tau^\pi(s,a) = c(s,a) + \gamma \int_S \int_A \left(Q_\tau^\pi(s',a') + \tau \ln \frac{\mathrm{d}\pi}{\mathrm{d}\mu}(a'|s')\right)\pi(da'|s')P(ds'|s,a)\,,\ \forall (s,a) \in S \times A\,. \tag{9}$$

2. Policy gradient

2.1. Performance difference. For each $\pi \in \mathcal{P}(A|S)$, we define the occupancy kernel $d^\pi \in \mathcal{P}(S|S)$ by

$$d^\pi(ds'|s) = (1-\gamma)\sum_{n=0}^\infty \gamma^n P_\pi^n(ds'|s)\,, \tag{10}$$

where P_π^n is the n-times product of the kernel P_π with $P_\pi^0(ds'|s) := \delta_s(ds')$ and the convergence is understood in $b\mathcal{M}(S|S)$. For a given initial distribution $\rho \in \mathcal{P}(S)$, we define

$$V_\tau^\pi(\rho) = \int_S V_\tau^\pi(s)\rho(ds) \quad \text{and} \quad d_\rho^\pi(ds) = \int_S d^\pi(ds|s')\rho(ds')\,. \tag{11}$$

The following lemma expresses the resolvent of the transition kernel using the occupancy kernel, which will be used in proving Lemma 2.2.

Lemma 2.1. *Let $\pi \in \mathcal{P}(A|S)$ and $f, g \in B_b(S)$ be such that for all $s \in S$,*

$$f(s) = \gamma \int_S \int_A f(s) P(ds'|s,a)\pi(da|s) + g(s)\,.$$

Then $f(s) = \frac{1}{1-\gamma}\int_S g(s')d^\pi(ds'|s)$ for all $s \in S$.

Proof. Recall that a kernel $k \in b\mathcal{M}(S|S)$ induces a linear operator $L_k \in \mathcal{L}(B_b(S))$ such that for all $h \in B_b(S)$, $L_k h(s) = \int_S h(s')k(ds'|s)$. Since $\|L_k h\|_{B_b(S)} \le \|h\|_{B_b(S)}\|k\|_{b\mathcal{M}(S|S)}$ for all $h \in B_b(S)$, $\|L_k\|_{\mathcal{L}(B_b(S))} \le \|k\|_{b\mathcal{M}(S|S)}$. Consider the kernel $\gamma P_\pi \in b\mathcal{M}(S|S)$ defined by $(\gamma P_\pi)(B) = \gamma \int_B \int_A P(ds'|s,a)\pi(da|s)$ for all $B \in \mathcal{B}(S)$. Then as $P_\pi \in \mathcal{P}(S|S)$ and $\|P_\pi\|_{b\mathcal{M}(S|S)} = 1$,

$$\|L_{\gamma P_\pi}\|_{\mathcal{L}(B_b(S))} \le \|\gamma P_\pi\|_{b\mathcal{M}(S|S)} = \gamma \|P_\pi\|_{b\mathcal{M}(S|S)} = \gamma\,.$$

The condition on f and g implies that $(\mathrm{id} - L_{\gamma P_\pi})f = g$, where id is the identity operator on $B_b(S)$. As $\|L_{\gamma P_\pi}\|_{\mathcal{L}(B_b(S))} \le \gamma < 1$, the operator $\mathrm{id} - L_{\gamma P_\pi} \in \mathcal{L}(B_b(S))$ is invertible, and the inverse operator is given by the Neumann series $(\mathrm{id} - L_{\gamma P_\pi})^{-1} = \sum_{n=0}^\infty L^n_{\gamma P_\pi}$. Thus, $f = \sum_{n=0}^\infty L^n_{\gamma P_\pi} g$. Observe that $L^n_{\gamma P_\pi} = L_{\gamma^n P^n_\pi}$ for all $n \in \mathbb{N}_0$, where P^n_π is the n-times product of the kernel P_π with $P^0_\pi(ds'|s) := \delta_s(ds')$. Then by the definition (10) of $d^\pi \in \mathcal{P}(S|S)$, $f = \sum_{n=0}^\infty L^n_{\gamma P_\pi} g = L_{(1-\gamma)^{-1}d^\pi} g$. This proves the desired identity. □

Lemma 2.2 (Performance difference)**.** *Let $\rho \in \mathcal{P}(S)$. Let $\tau \ge 0$. Let $\pi, \pi' \in \mathcal{P}(A|S)$ and if $\tau > 0$ assume further that $\pi, \pi' \in \Pi_\mu$. Then*

$$\begin{aligned} &V^\pi_\tau(\rho) - V^{\pi'}_\tau(\rho) \\ &\quad = \frac{1}{1-\gamma}\int_S \left[\int_A \left(Q^{\pi'}_\tau(s,a) + \tau \ln \frac{\mathrm{d}\pi'}{\mathrm{d}\mu}(a|s)\right)(\pi - \pi')(da|s) + \tau\,\mathrm{KL}(\pi(\cdot|s)|\pi'(\cdot|s))\right] d^\pi_\rho(ds)\,. \end{aligned}$$

Proof of Lemma 2.2. By (8), for all $s \in S$,

$$\begin{aligned} &V^\pi_\tau(s) - V^{\pi'}_\tau(s) \\ &\quad = \int_A \left(Q^\pi_\tau(a|s) + \tau \ln \frac{\mathrm{d}\pi}{\mathrm{d}\mu}(a|s)\right)\pi(da|s) - \int_A \left(Q^{\pi'}_\tau(s,a) + \tau \ln \frac{\mathrm{d}\pi'}{\mathrm{d}\mu}(a|s)\right)\pi'(da|s) \\ &\quad = \int_A \left(Q^{\pi'}_\tau(s,a) + \tau \ln \frac{\mathrm{d}\pi'}{\mathrm{d}\mu}(a|s)\right)(\pi - \pi')(da|s) \\ &\qquad + \int_A \left(Q^\pi_\tau(s,a) + \tau \ln \frac{\mathrm{d}\pi}{\mathrm{d}\mu}(a|s) - Q^{\pi'}_\tau(s,a) - \tau \ln \frac{\mathrm{d}\pi'}{\mathrm{d}\mu}(a|s)\right)\pi(da|s)\,. \end{aligned}$$

Hence for all $s \in S$ we have

$$\begin{aligned} V^\pi_\tau(s) - V^{\pi'}_\tau(s) = &\int_A \left(Q^{\pi'}_\tau(s,a) + \tau \ln \frac{\mathrm{d}\pi'}{\mathrm{d}\mu}(a|s)\right)(\pi - \pi')(da|s) \\ &+ \gamma \int_A \int_S \left(V^\pi_\tau(s') - V^{\pi'}_\tau(s')\right) P(ds'|s,a)\pi(da|s) + \tau\,\mathrm{KL}(\pi(\cdot|s)|\pi'(\cdot|s))\,, \end{aligned}$$

where the last equality used (7) and the fact that $\mathrm{KL}(\pi(\cdot|s)|\pi'(\cdot|s)) = \int_A \ln \frac{\mathrm{d}\pi}{\mathrm{d}\pi'}(a|s)\pi(da|s)$. Hence, by Fubini's theorem and Lemma 2.1, for all $s \in S$,

$$\begin{aligned} &V^\pi_\tau(s) - V^{\pi'}_\tau(s) \\ &\quad = \frac{1}{1-\gamma}\int_S \left[\int_A \left(Q^{\pi'}_\tau(s',a) + \tau \ln \frac{\mathrm{d}\pi'}{\mathrm{d}\mu}(a|s')\right)(\pi - \pi')(da|s') + \tau\,\mathrm{KL}(\pi(\cdot|s')|\pi'(\cdot|s'))\right] d^\pi(ds'|s). \end{aligned}$$

Integrating both sides with respect to ρ yields the desired identity. □

2.2. First variation, chain rule, policy gradient theorem.

Proposition 2.3. *Let $\pi, \pi' \in \Pi_\mu$ be such that $\pi(da|s) = \frac{\exp(f(s,a))\mu(da)}{\int_A \exp(f(s,a'))\mu(da')}$ for all $s \in S$. Then*

$$\|Q_\tau^{\pi'} - Q_\tau^\pi\|_{B_b(S\times A)} \le \frac{\gamma}{(1-\gamma)^2}\left(\|c\|_{B_b(S\times A)} + 2\tau\|f\|_{B_b(S\times A)}\right)\|\pi - \pi'\|_{b\mathcal{M}(A|S)} + \frac{\tau\gamma}{1-\gamma}\left\|\ln\frac{\mathrm{d}\pi'}{\mathrm{d}\pi}\right\|_{B_b(S\times A)}.$$

Proof. By Lemma 2.2, for all $s \in S$,

$$\begin{aligned}|V_\tau^\pi(s) - V_\tau^{\pi'}(s)| &\le \frac{1}{1-\gamma}\left|\int_S\int_A\left(Q_\tau^{\pi'}(s',a) + \tau\ln\frac{\mathrm{d}\pi'}{\mathrm{d}\mu}(a|s')\right)(\pi-\pi')(da|s')d_s^\pi(ds')\right| \\ &\quad + \frac{\tau}{1-\gamma}\int_S\int_A \ln\frac{\mathrm{d}\pi}{\mathrm{d}\pi'}(a|s')\pi(da|s')d_s^\pi(ds') \\ &\le \frac{1}{1-\gamma}\left\|Q_\tau^{\pi'} + \tau\ln\frac{\mathrm{d}\pi'}{\mathrm{d}\mu}\right\|_{B_b(S\times A)}\|\pi-\pi'\|_{b\mathcal{M}(A|S)} + \frac{\tau}{1-\gamma}\left\|\ln\frac{\mathrm{d}\pi}{\mathrm{d}\pi'}\right\|_{B_b(S\times A)}.\end{aligned}$$

Thus, by (7), for all $(s,a) \in S\times A$,

$$|Q_\tau^{\pi'}(s,a) - Q_\tau^\pi(s,a)| \le \gamma\int_S |V_\tau^{\pi'}(s') - V_\tau^\pi(s')|P(ds'|s,a) \le \gamma\|V_\tau^{\pi'} - V_\tau^\pi\|_{B_b(S)}.$$

By Proposition 1.6,

$$\begin{aligned}\left\|Q_\tau^{\pi'} + \tau\ln\frac{\mathrm{d}\pi'}{\mathrm{d}\mu}\right\|_{B_b(S\times A)} &\le \frac{1}{1-\gamma}\left(\|c\|_{B_b(S\times A)} + 2\tau\gamma\|f\|_{B_b(S\times A)}\right) + 2\tau\|f\|_{B_b(S\times A)} \\ &\le \frac{1}{1-\gamma}\left(\|c\|_{B_b(S\times A)} + 2\tau\|f\|_{B_b(S\times A)}\right).\end{aligned}$$

Combining the above inequalities yields the desired estimate. □

Proposition 2.4. *Let $\tau \ge 0$ and $\rho \in \mathcal{P}(S)$. For all $\pi, \pi' \in \Pi_\mu \subset \mathcal{P}(A|S)$ (cf. Definition 1.5),*

$$\begin{aligned}&\lim_{\varepsilon\searrow 0}\frac{V_\tau^{(1-\varepsilon)\pi+\varepsilon\pi'}(\rho) - V_\tau^\pi(\rho)}{\varepsilon} \\ &\quad = \frac{1}{1-\gamma}\int_S\int_A\left(Q_\tau^\pi(s,a) + \tau\ln\frac{\mathrm{d}\pi}{\mathrm{d}\mu}(a|s) - V_\tau^\pi(s)\right)(\pi'-\pi)(da|s)d_\rho^\pi(ds).\end{aligned} \tag{12}$$

Proof. Let $\pi^\varepsilon = (1-\varepsilon)\pi + \varepsilon\pi' = \pi + \varepsilon(\pi'-\pi)$ and note that $\pi - \pi^\varepsilon = -\varepsilon(\pi'-\pi) = \varepsilon(\pi-\pi')$. Then

$$\begin{aligned}\frac{1}{\varepsilon}(V_\tau^\pi(\rho) - V_\tau^{\pi^\varepsilon}(\rho)) &= \frac{1}{\varepsilon}\frac{1}{1-\gamma}\int_S\int_A\left(Q_\tau^{\pi^\varepsilon}(s,a) + \tau\ln\frac{\mathrm{d}\pi^\varepsilon}{\mathrm{d}\mu}(a|s)\right)(\pi-\pi^\varepsilon)(da|s)d_\rho^\pi(ds) \\ &\quad + \frac{1}{\varepsilon}\frac{\tau}{1-\gamma}\int_S \mathrm{KL}(\pi(\cdot|s)|\pi^\varepsilon(\cdot|s))d_\rho^\pi(ds) \\ &= \frac{1}{1-\gamma}\int_S\int_A\left(Q_\tau^{\pi^\varepsilon}(s,a) + \tau\ln\frac{\mathrm{d}\pi^\varepsilon}{\mathrm{d}\mu}(a|s)\right)(\pi-\pi')(da|s)d_\rho^\pi(ds) \\ &\quad + \frac{1}{\varepsilon}\frac{\tau}{1-\gamma}\int_S \mathrm{KL}(\pi(\cdot|s)|\pi^\varepsilon(\cdot|s))d_\rho^\pi(ds).\end{aligned}$$

We will now employ the identity which holds for any $m, m' \in \mathcal{P}(A)$ for which the quantities in the identity are finite:

$$\mathrm{KL}(m|\mu) - \mathrm{KL}(m'|\mu) = \mathrm{KL}(m|m') + \int_A \ln\tfrac{\mathrm{d}m'}{\mathrm{d}\mu}(a)(m-m')(da).$$

Hence

$$\frac{1}{\varepsilon}(V_\tau^\pi(\rho) - V_\tau^{\pi^\varepsilon}(\rho)) = \frac{1}{1-\gamma}\int_S\int_A Q_\tau^{\pi^\varepsilon}(s,a)(\pi-\pi')(da|s)d_\rho^\pi(ds) + \frac{1}{\varepsilon}\frac{\tau}{1-\gamma}\int_S \Big(\mathrm{KL}(\pi(\cdot|s)|\mu(\cdot|s)) - \mathrm{KL}(\pi^\varepsilon(\cdot|s)|\mu(\cdot|s))\Big)d_\rho^\pi(ds)\,.$$

Thus

$$\frac{1}{\varepsilon}(V_\tau^{\pi^\varepsilon}(\rho) - V_\tau^{\pi}(\rho)) = \frac{1}{1-\gamma}\int_S\int_A Q_\tau^{\pi^\varepsilon}(s,a)(\pi'-\pi)(da|s)d_\rho^\pi(ds) + \frac{\tau}{1-\gamma}\int_S \frac{1}{\varepsilon}\Big(\mathrm{KL}(\pi^\varepsilon(\cdot|s)|\mu(\cdot|s)) - \mathrm{KL}(\pi(\cdot|s)|\mu(\cdot|s))\Big)d_\rho^\pi(ds)\,.$$

The first integral on the right hand side converges to $\frac{1}{1-\gamma}\int_S\int_A Q_\tau^\pi(s,a)(\pi'-\pi)(da|s)d_\rho^\pi(ds)$ as $\varepsilon\to 0$ due to Proposition 2.3. Moreover, as $\pi,\pi'\in\Pi_\mu$, for all $s\in S$, by [Kerimkulov et al., 2025b, Lemma 3.8],

$$\lim_{\varepsilon\searrow 0}\frac{1}{\varepsilon}\Big(\mathrm{KL}(\pi^\varepsilon(\cdot|s)|\mu(\cdot|s)) - \mathrm{KL}(\pi(\cdot|s)|\mu(\cdot|s))\Big) = \int_A \ln\frac{\mathrm{d}\pi}{\mathrm{d}\mu}(a|s)(\pi'-\pi)(da|s)\,,$$

which along with Proposition 1.6 and the dominated yields the desired limit. □

For a fixed $\nu\in\mathcal{P}(S)$ define $\langle\cdot,\cdot\rangle_\nu : B_b(S\times A)\times b\mathcal{M}(A|S)\to\mathbb{R}$ by

$$\langle Z, m\rangle_\nu = \frac{1}{1-\gamma}\int_S\int_A Z(s,a)m(da|s)\nu(ds)\,,\quad (Z,m)\in B_b(S\times A)\times b\mathcal{M}(A|S).$$

As a consequence of Proposition 2.4, given $\nu\in\mathcal{P}(S)$ satisfying $d_\rho^\pi\ll\nu$,

$$\lim_{\varepsilon\searrow 0}\frac{V_\tau^{(1-\varepsilon)\pi+\varepsilon\pi'}(\rho) - V_\tau^\pi(\rho)}{\varepsilon} = \left\langle \frac{\delta V_\tau^\pi(\rho)}{\delta\pi}\Big|_\nu, \pi'-\pi\right\rangle_\nu,$$

with

$$\frac{\delta V_\tau^\pi(\rho)}{\delta\pi}\Big|_\nu(s,a) = \left(Q_\tau^\pi(s,a) + \tau\ln\frac{\mathrm{d}\pi}{\mathrm{d}\mu}(s,a) - V_\tau^\pi(s)\right)\frac{\mathrm{d}d_\rho^\pi}{\mathrm{d}\nu}(s)\,, \tag{13}$$

where $d_\rho^\pi\in\mathcal{P}(S)$ is the occupancy measure associated with π. The flat derivative (13) is consistent with the classical derivative in π when dealing with discrete action spaces (see, e.g., [Lan, 2023, Lemma 1]). It also generalises the notation of the flat derivative applied to probability measures to encompass probability transition kernels.

Let $(\mathbb{H},(\cdot,\cdot)_\mathbb{H})$ be a Hilbert space (we will either have $\mathbb{H}=\mathbb{R}^p$ or $\mathbb{H}=\ell_2$). If $\pi:\mathbb{H}\to\Pi_\mu$ i.e. we parametrize π in terms of $\theta\in\mathbb{H}$ and we can compute $\nabla_\theta\pi_\theta$ then a chain rule holds and can be proved similarly to [Kerimkulov et al., 2025a, Proposition 3.8].

Lemma 2.5 (Chain rule). *Let $\pi:\mathbb{H}\to\Pi_\mu$ be given. Then $\partial_{\theta_i}V_\tau^{\pi_\theta}(\rho) = \left\langle\frac{\delta V_\tau^\pi(\rho)}{\delta\pi}, \partial_{\theta_i}\pi_\theta\right\rangle_{d_\rho^\pi}$.*

Theorem 2.6 (Policy gradient theorem). *Let $\frac{\mathrm{d}\pi_\theta}{\mathrm{d}\mu}(a|s) := \frac{e^{g_\theta(s,a)}}{Z_\theta(s)}$, where $Z_\theta(s) := \int_A e^{g_\theta(s,a')}\mu(da')$. Then*

$$\nabla_\theta V_\tau^{\pi_\theta}(\rho) = \frac{1}{1-\gamma}\mathbb{E}^{s\sim d_\rho^{\pi_\theta}}_{a\sim\pi_\theta(\cdot|s)}\left[\frac{\delta V_\tau^{\pi_\theta}}{\delta\pi}(s,a)\nabla_\theta\ln\frac{\mathrm{d}\pi_\theta}{\mathrm{d}\mu}(a|s)\right].$$

Proof. From Lemma 2.5 (chain rule) we have:

$$\nabla_\theta V_\tau^{\pi_\theta}(\rho) = \frac{1}{1-\gamma}\int_S\int_A \frac{\delta V_\tau^{\pi_\theta}}{\delta\pi}(s,a)\nabla_\theta\frac{\mathrm{d}\pi_\theta}{\mathrm{d}\mu}(a|s)\mu(da)\,d_\rho^{\pi_\theta}(ds)\,.$$

Taking the gradient of the logarithm and re-arranging we see that

$$\nabla_\theta\frac{\mathrm{d}\pi_\theta}{\mathrm{d}\mu}(a|s) = \frac{\mathrm{d}\pi_\theta}{\mathrm{d}\mu}(a|s)\nabla_\theta\ln\frac{\mathrm{d}\pi_\theta}{\mathrm{d}\mu}(a|s)\,. \tag{14}$$

Hence

$$\nabla_\theta V_\tau^{\pi_\theta}(\rho) = \tfrac{1}{1-\gamma}\int_S\int_A \tfrac{\delta V_\tau^{\pi_\theta}}{\delta\pi}(s,a)\nabla_\theta \ln \tfrac{\mathrm{d}\pi_\theta}{\mathrm{d}\mu}(a|s)\pi_\theta(da|s)\, d_\rho^{\pi_\theta}(ds)\,.$$

We just need to rewrite this in terms of expectation to get the conclusion. □

We can take any $b \in B_b(S)$. Then

$$\int_A b(s)\nabla_\theta \ln \tfrac{\mathrm{d}\pi_\theta}{\mathrm{d}\mu}(a|s)\pi_\theta(da|s) = b(s)\int_A \nabla_\theta \ln \tfrac{\mathrm{d}\pi_\theta}{\mathrm{d}\mu}(a|s)\pi_\theta(da|s) = b(s)\int_A \nabla_\theta \tfrac{\mathrm{d}\pi_\theta}{\mathrm{d}\mu}(a|s)\mu(da)$$
$$= b(s)\nabla_\theta \int_A \tfrac{\mathrm{d}\pi_\theta}{\mathrm{d}\mu}(a|s)\mu(da) = b(s)\nabla_\theta 1 = 0\,.$$

Hence

$$\nabla_\theta V_\tau^{\pi_\theta}(\rho) = \tfrac{1}{1-\gamma}\mathbb{E}^{s\sim d_\rho^{\pi_\theta}}_{a\sim\pi_\theta(\cdot|s)}\left[\left(\tfrac{\delta V_\tau^{\pi_\theta}}{\delta\pi}(s,a) + b(s)\right)\nabla_\theta \ln \tfrac{\mathrm{d}\pi_\theta}{\mathrm{d}\mu}(a|s)\right].$$

We can see that to use this in an algorithm we (at least approximately) need $\frac{\delta V_\tau^{\pi_\theta}(\rho)}{\delta\pi}(s,a) = Q_\tau^{\pi_\theta}(s,a) + \tau \ln \frac{\mathrm{d}\pi_\theta}{\mathrm{d}\mu}(s,a) - V_\tau^{\pi_\theta}(s)$. Typically, we would have access to a stream of data

$$(s_0, a_0, \ln \tfrac{\mathrm{d}\pi_\theta}{\mathrm{d}\mu}(s_0,a_0), c_0, s_1, a_1, \ln \tfrac{\mathrm{d}\pi_\theta}{\mathrm{d}\mu}(s_1,a_1), c_1, \ldots, s_N, a_N, \ln \tfrac{\mathrm{d}\pi_\theta}{\mathrm{d}\mu}(s_N,a_N), c_N)\,,$$

produced by interacting with the environment using policy π_θ for an "episode" of length N. The most common approach is to use the generalised advantage estimation formula from Schulman et al. [2015]. Note that this relies on having access to a (separate) approximation of the value function. An alternative is to have a (separate) function approximation for e.g. the Q function updated from a Bellman error.

The following observation may be useful later.

Corollary 2.7 (to Policy Gradient Theorem)**.** *Let* $\frac{\mathrm{d}\pi_\theta}{\mathrm{d}\mu}(a|s) := \frac{e^{g_\theta(s,a)}}{Z_\theta(s)}$, $Z_\theta(s) := \int_A e^{g_\theta(s,a')}\mu(da')$. *Then*

$$\nabla_\theta V_\tau^{\pi_\theta}(\rho) = \frac{1}{1-\gamma}\mathbb{E}^{s\sim d_\rho^{\pi_\theta}}_{a\sim\pi_\theta(\cdot|s)}\left[\frac{\delta V_\tau^{\pi_\theta}}{\delta\pi}(s,a)\left(\nabla_\theta g_\theta(s,a) - \int_A (\nabla_\theta g_\theta)(s,a')\pi_\theta(da'|s)\right)\right].$$

Proof of Corollary 2.7. Noting that

$$\ln \tfrac{\mathrm{d}\pi_\theta}{\mathrm{d}\mu}(a|s) = g_\theta(s,a) - \ln Z_\theta(s)$$

and so

$$\nabla_\theta \ln \tfrac{\mathrm{d}\pi_\theta}{\mathrm{d}\mu}(s,a) = \nabla_\theta g_\theta(s,a) - \nabla_\theta Z_\theta(s)\tfrac{1}{Z_\theta(s)} = \nabla_\theta g_\theta(s,a) - \textstyle\int_A (\nabla_\theta g_\theta)(s,a')\frac{e^{g_\theta(s,a')}}{Z_\theta(s)}\mu(da')\,.$$

Hence we have an expression for gradient of the log-density:

$$\nabla_\theta \ln \tfrac{\mathrm{d}\pi_\theta}{\mathrm{d}\mu}(s,a) = \nabla_\theta g_\theta(s,a) - \textstyle\int_A (\nabla_\theta g_\theta)(s,a')\frac{\mathrm{d}\pi_\theta}{\mathrm{d}\mu}(a'|s)\mu(da') \tag{15}$$

which concludes the calculation. □

Remark 2.8. If the state and action spaces are finite and we take the direct (tabular) parametrizations so that $g_\theta(s,a) := \theta(s,a)$ then

$$\partial_{\theta_{\hat{s},\hat{a}}} g_\theta(s,a) - \sum_{a'} \partial_{\theta_{\hat{s},\hat{a}}} g_\theta(s,a')\pi_\theta(a'|s) = \delta_{\hat{s},\hat{a}}(s,a) - \sum_{a'}\delta_{\hat{s},\hat{a}}(s,a')\pi(a'|s) = \delta_{\hat{s},\hat{a}}(s,a) - \delta_{\hat{s}}(s)\pi(\hat{a}|s)$$
$$= \delta_{\hat{s}}(s)\big(\delta_{\hat{a}}(a) - \delta_{\hat{s}}(s)\pi(\hat{a}|s)\big)\,.$$

Hence

$$\partial_{\theta_{\hat{s},\hat{a}}} V_\tau^{\pi_\theta}(\rho) = \tfrac{1}{1-\gamma}\sum_{s,a}\tfrac{\delta V_\tau^{\pi_\theta}}{\delta\pi}(s,a)\delta_{\hat{s}}(s)\delta_{\hat{a}}(a)\pi_\theta(a|s)d_\rho^{\pi_\theta}(s) - \tfrac{1}{1-\gamma}\sum_{s,a}\tfrac{\delta V_\tau^{\pi_\theta}}{\delta\pi}(s,a)\delta_{\hat{s}}(s)\pi(\hat{a}|s)\pi_\theta(a|s)d_\rho^{\pi_\theta}(s)\,.$$

But

$$\sum_s \sum_a \tfrac{\delta V_\tau^{\pi_\theta}}{\delta \pi}(s,a)\delta_{\hat{s}}(s)\pi(\hat{a}|s)\pi_\theta(a|s)d_\rho^{\pi_\theta}(s) = \sum_s \delta_{\hat{s}}(s)\pi(\hat{a}|s)\sum_a \tfrac{\delta V_\tau^{\pi_\theta}}{\delta \pi}(s,a)\pi_\theta(a|s)d_\rho^{\pi_\theta}(s) = 0$$

and so

$$\partial_{\theta_{\hat{s},\hat{a}}} V_\tau^{\pi_\theta}(\rho) = \tfrac{1}{1-\gamma}\tfrac{\delta V_\tau^{\pi_\theta}}{\delta \pi}(\hat{s},\hat{a})\pi_\theta(\hat{a}|\hat{s})d_\rho^{\pi_\theta}(\hat{s})\,.$$

This is (for the $\tau = 0$ case) exactly Lemma C.1 in Agarwal et al. [2019].

If $g_\theta(s,a) = (\theta, \phi(s,a))_{\mathbb{H}}$ then $\partial_{\theta_i} g_\theta(s,a) = \theta_i(s,a)$ and so

$$\nabla_\theta V_\tau^{\pi_\theta}(\rho) = \tfrac{1}{1-\gamma}\mathbb{E}^{s\sim d_\rho^{\pi_\theta}}_{a\sim\pi_\theta(\cdot|s)}\Big[\tfrac{\delta V_\tau^{\pi_\theta}}{\delta \pi}(s,a)\Big(\phi(s,a) - \textstyle\int_A \phi(s,a')\pi_\theta(da'|s)\Big)\Big]\,. \tag{16}$$

3. Mirror descent

Mirror descent is now a classical approach to first order (gradient-based) methods for optimizing functions over convex sets, going back to Nemirovskij and Yudin [1983].

3.1. Motivation. There are at least three good reasons to study mirror descent in the context of RL. First of all, it allows one to consider gradient-like updates without introducing parametrization. Indeed, even in the finite action space setting an Euclidean gradient step in the convex space of policies provides no guarantee that after the update step is carried out we still have an element in the probability simplex.

Second, the lack of convexity of the map $\theta \mapsto V_\tau^{\pi_{\theta_n}}(\rho)$ makes convergence analysis of gradient descent challenging. The best results for direct parametrization in the finite action space setting are Mei et al. [2020] which first prove a non-local version of gradient dominance and then show that along the steps of the gradient descent the constant appearing is lower bounded thus obtaining convergence rate.

The third and final reason is algorithmic. The classical policy gradient methods update the policy parametrization using

$$\theta_{n+1} = \theta_n - \eta\nabla_\theta V_\tau^{\pi_{\theta_n}}(\rho)\,, n = 0, 1, \dots\,, \theta_0 \in \mathbb{R}^p \text{ given and } \eta > 0 \text{ a step size.}$$

Especially when $\frac{\mathrm{d}\pi_\theta}{\mathrm{d}\mu}(a|s) \propto e^{g_\theta(s,a)}$ with g_θ a neural network there is no guarantee that small step size η leads to a small update in the space of policies. We can only guarantee improvement with appropriate L-smoothness and the appropriately small step size or when asymptotically small step is taken. Indeed, taking $\eta \to 0$ the continuous time version of the above stepping is

$$\tfrac{d}{dt}\theta_t = -\nabla_\theta V_\tau^{\pi_{\theta_t}}(\rho)\,, t > 0\,, \theta_0 \in \mathbb{R}^p \text{ given.}$$

Then chair rule tells us that

$$\tfrac{d}{dt}V_\tau^{\pi_t}(\rho) = \tfrac{d}{dt}\theta_t \cdot \nabla_\theta V_\tau^{\pi_{\theta_t}}(\rho) = -|\nabla_\theta V_\tau^{\pi_{\theta_t}}(\rho)|^2 \le 0\,.$$

This makes choosing the "right" $\eta > 0$ difficult in practice: small values lead to slow convergence and already somewhat larger values can lead to instability.

How to overcome this? One can think of the classical policy gradient update as something that arises as follows: we do one step Taylor expansion in θ and then add a penalty term to ensure that we don't too large a step. Writing $V^{\pi_{\theta_n}} = V_\tau^{\pi_{\theta_n}}(\rho)$ we let

$$L_{\mathrm{PG}}(\theta) := V^{\pi_{\theta_n}} + \nabla_\theta V^{\pi_{\theta_n}} \cdot (\theta - \theta_n) + \frac{1}{2}\eta^{-1}|\theta|^2\,.$$

We can now take

$$\theta^{n+1} = \arg\min_\theta L_{\mathrm{PG}}(\theta)\,.$$

From the first order condition for optimizing $\theta \mapsto L_{\mathrm{PG}}(\theta)$ we get

$$0 = \nabla_\theta V^{\pi_{\theta_n}} + \eta^{-1}(\theta - \theta_n)$$

which is exactly the policy gradient step. Let us now penalize not in the parameter θ but in terms of policy. Let's say we have $\pi_n = \pi_{\theta_n}$. Fix $\rho \in \mathcal{P}(S)$ and write $V_\tau^\pi = V_\tau^\pi(\rho)$. By perf. diff. Lemma 2.2 we have

$$V_\tau^\pi = V_\tau^{\pi_n} + \left\langle \tfrac{\delta V_\tau^{\pi_n}}{\delta \pi}, \pi - \pi_n \right\rangle_{\rho,\pi} + \tfrac{\tau}{1-\gamma} \int_S \mathrm{KL}(\pi|\pi_n)(s) d_\rho^{\pi}(ds)\,.$$

This is almost a first order expansion except for the terms highlighted. We linearize and penalize with $\lambda \geq \tau$ to not move too far

$$V_\tau^\pi \approx V_\tau^{\pi_n} + \left\langle \tfrac{\delta V_\tau^{\pi_n}}{\delta \pi}, \pi - \pi_n \right\rangle_{\rho,\pi_n} + \tfrac{\lambda}{1-\gamma} \int_S \mathrm{KL}(\pi|\pi_n)(s) d_\rho^{\pi_n}(ds)\,.$$

We drop the terms that don't depend on π and the $(1-\gamma)^{-1}$ scaling as they won't matter if we're minimizing and define a surrogate objective

$$L_{\pi_n}^\pi := \left\langle \tfrac{\delta V_\tau^{\pi_n}}{\delta \pi}, \pi - \pi_n \right\rangle_{\rho,\pi_n} + \lambda \int_S \mathrm{KL}(\pi|\pi_n)(s) d_\rho^{\pi_n}(ds)$$

To see how this could be implemented we do a change of measure and write in terms of expectation

$$\begin{aligned} L_{\pi_n}^\pi &= \int_S \int_A \left(\tfrac{\delta V_\tau^{\pi_n}}{\delta \pi}(s,a) + \lambda \ln \tfrac{\mathrm{d}\pi}{\mathrm{d}\pi_n}(a|s) \right) \tfrac{\mathrm{d}\pi}{\mathrm{d}\pi_n}(a|s) \pi_n(da|s) d_\rho^{\pi_n}(ds) \\ &= \mathbb{E}_{a\sim\pi_{\theta_n}(\cdot|s)}^{s\sim d_\rho^{\pi_n}} \left[\left(\tfrac{\delta V_\tau^{\pi_n}}{\delta \pi}(s,a) + \lambda \ln \tfrac{\mathrm{d}\pi}{\mathrm{d}\pi_n}(a|s) \right) \tfrac{\mathrm{d}\pi}{\mathrm{d}\pi_n}(a|s) \right]. \end{aligned}$$

This is a quantity which can be estimated by collecting samples under the current policy π_n. The mirror descent update step is $\pi_{n+1} = \arg\min_\pi L_{\pi_n}^\pi$. If the policy is parametrized by θ then the proposed update in parameter space is: $\theta_{n+1} \in \arg\min_\theta L_{\pi_{\theta_n}}^{\pi_\theta}$.

3.2. Convergence of mirror descent with approximate advantage. We will first show that if we have access to the exact advantage function then the mirror descent updates guarantee improvement. Let

$$\pi^{n+1}(\cdot|s) = \arg\min_{m\in\mathcal{P}(A)} \int_A \frac{\delta V_\tau^{\pi_n}}{\delta \pi}(s,a)(m(da) - \pi^n(da|s)) + \lambda \,\mathrm{KL}(m|\pi^n(\cdot|s))\,. \tag{17}$$

For this exact scheme we have policy improvement.

Lemma 3.1 (Policy improvement)**.** *Let $V_\tau^n := V_\tau^{\pi^n}$ for $n \in \mathbb{N}$ and $\pi^n \in \Pi_\mu$ given by (17). If $\tau \leq \lambda$ then for any $\rho \in \mathcal{P}(S)$ we have $V_\tau^{n+1}(\rho) \leq V_\tau^n(\rho)$.*

Proof. From the performance difference lemma, see Lemma (2.2), we see that

$$\begin{aligned} (V_\tau^{n+1} - V_\tau^n)(\rho) &= \frac{1}{1-\gamma} \int_S \left(\int_A \frac{\delta V_\tau^n}{\delta \pi}(s,a)(\pi^{n+1} - \pi^n)(da|s) + \tau \,\mathrm{KL}(\pi^{n+1}|\pi^n)(s) \right) d_\rho^{\pi^{n+1}}(ds) \\ &\leq \frac{1}{1-\gamma} \int_S \left(\int_A \frac{\delta V_\tau^n}{\delta \pi}(s,a)(\pi^{n+1} - \pi^n)(da|s) + \lambda \,\mathrm{KL}(\pi^{n+1}|\pi^n)(s) \right) d_\rho^{\pi^{n+1}}(ds)\,. \end{aligned} \tag{18}$$

From the mirror descent update (17) we have, for all $\pi \in \Pi_\mu$ and $s \in S$ that

$$\begin{aligned} &\int_A \frac{\delta V_\tau^n}{\delta \pi}(s,a)(\pi - \pi^n)(da|s) + \lambda \,\mathrm{KL}(\pi|\pi^n)(s) \\ &\geq \int_A \frac{\delta V_\tau^n}{\delta \pi}(s,a)(\pi^{n+1} - \pi^n)(da|s) + \lambda \,\mathrm{KL}(\pi^{n+1}|\pi^n)(s)\,. \end{aligned}$$

This with $\pi = \pi^n$ allows us to conclude that for all $s \in S$ we have

$$\int_A \frac{\delta V_\tau^n}{\delta \pi}(s,a)(\pi^{n+1} - \pi^n)(da|s) + \lambda \,\mathrm{KL}(\pi^{n+1}|\pi^n)(s) \leq 0\,. \tag{19}$$

This with (18) concludes the proof. □

Now recall that $\frac{\delta V_\tau^{\pi_n}}{\delta \pi} = A_\tau^{\pi_n} + \tau \ln \frac{\mathrm{d}\pi^n}{\mathrm{d}\mu} = Q_\tau^{\pi_n} - V_\tau^{\pi_n} + \tau \ln \frac{\mathrm{d}\pi^n}{\mathrm{d}\mu}$. In practice updates can only be made with an approximation of $A_\tau^{\pi_n}$, say $\hat{A}_n(s,a) = A_\tau^{\pi_n}(s,a) + \mathcal{E}_n(s,a)$. We consider the scheme

$$\pi^{n+1}(da|s) = \operatorname*{arg\,min}_{m \in \mathcal{P}(A)} \int_A \Big(\hat{A}_n(s,a) + \tau \ln \tfrac{\mathrm{d}\pi^n}{\mathrm{d}\mu}(a|s)\Big)(m(da) - \pi^n(da|s)) + \lambda \,\mathrm{KL}(m|\pi^n(\cdot|s)\,. \tag{20}$$

What can we say about convergence of such a scheme, provided we can control the errors? We want to use the classical tools for analysis of mirror descent: 3-point lemma, convexity (substituted by performance difference) and L-smoothness (derived from performance difference).

Let $M_\mu = \{m \in \mathcal{P}(A) \,|\, \frac{\mathrm{d}m}{\mathrm{d}\mu} \text{ exists and } \ln \frac{\mathrm{d}m}{\mathrm{d}\mu} \in B_b(A)\}$ and notice this is a convex subset of $\mathcal{P}(A)$.

Lemma 3.2 (Three point lemma / Bregman proximal inequality)**.** *Let $G : M_\mu \to \mathbb{R}$ be convex. For all $m' \in M_\mu$ let*

$$m^* = \operatorname*{arg\,min}_{m \in M_\mu} \big\{G(m) + \mathrm{KL}(m|m')\big\}\,. \tag{21}$$

Then, for all $m \in M_\mu$, we have

$$G(m) + \mathrm{KL}(m|m') \geq G(m^*) + \mathrm{KL}(m|m^*) + \mathrm{KL}(m^*|m')\,. \tag{22}$$

The proof of Lemma 3.2 can be found e.g., in Aubin-Frankowski et al. [2022] noting that the flat derivative of KL is well defined on M_μ, see e.g. [Kerimkulov et al., 2025b, Lemma 3.8].

We will also need the following crucial observation with a trivial proof.

Lemma 3.3. *Let $F : S \to \mathbb{R}$ be such that $F \leq 0$. Then for any π and any $s \in S$*

$$\frac{1}{1-\gamma} \int_S F(s')\, d_s^\pi(ds') \leq F(s)\,. \tag{23}$$

Proof. From (10) and the fact that $P_\pi^0(ds'|s) = \delta_s(ds')$ we have for all $s \in S$ that

$$\begin{aligned} \frac{1}{1-\gamma} \int_S F(s')\, d_s^\pi(ds') &= \int_S F(s') P_\pi^0(ds'|s) + \sum_{k=1}^\infty \int_S \gamma^k F(s') P_\pi^k(ds'|s) \\ &\leq \int_S F(s') \delta_s(ds') = F(s)\,. \end{aligned} \tag{24}$$

This concludes the proof. □

Let π^n be generated by inductive application of the approximate mirror descent step (20). Let $V_\tau^n := V_\tau^{\pi^n}$ for $n \in \mathbb{N}$. We begin with an application of Bregman proximal inequality, see Lemma 3.2. Fix $s \in S$ and $\pi^n \in \Pi_\mu$ and define $G : M_\mu \to \mathbb{R}$ by

$$G(m) = \frac{1}{\lambda} \int_A \Big(\hat{A}_n(s,a) + \tau \ln \tfrac{\mathrm{d}\pi^n}{\mathrm{d}\mu}(a|s)\Big)(m(da) - \pi^n(da|s))\,.$$

It is linear and thus clearly convex and hence due to the mirror descent update (20) is equivalent to (21) and so we have, for all $\pi \in \Pi_\mu$, $s \in S$ and $n \in \mathbb{N}$ that

$$\begin{aligned} &\frac{1}{\lambda} \int_A \Big(\hat{A}_n(s,a) + \tau \ln \tfrac{\mathrm{d}\pi^n}{\mathrm{d}\mu}(a|s)\Big)(\pi - \pi^n)(da|s) + \mathrm{KL}(\pi|\pi^n)(s) \\ &\geq \frac{1}{\lambda} \int_A \Big(\hat{A}_n(s,a) + \tau \ln \tfrac{\mathrm{d}\pi^n}{\mathrm{d}\mu}(a|s)\Big)(\pi^{n+1} - \pi^n)(da|s) + \mathrm{KL}(\pi|\pi^{n+1})(s) + \mathrm{KL}(\pi^{n+1}|\pi^n)(s)\,. \end{aligned}$$

Re-arranging this leads to

$$\begin{aligned}&\mathrm{KL}(\pi|\pi^{n+1})(s)-\mathrm{KL}(\pi|\pi^n)(s)\\&\le\frac{1}{\lambda}\int_A\Big(\hat{A}_n(s,a)+\tau\ln\tfrac{\mathrm{d}\pi^n}{\mathrm{d}\mu}(a|s)\Big)(\pi-\pi^n)(da|s)\\&\quad-\frac{1}{\lambda}\int_A\Big(\hat{A}_n(s,a)+\tau\ln\tfrac{\mathrm{d}\pi^n}{\mathrm{d}\mu}(a|s)\Big)(\pi^{n+1}-\pi^n)(da|s)-\mathrm{KL}(\pi^{n+1}|\pi^n)(s)\,.\end{aligned}\tag{25}$$

From the performace difference, Lemma 2.2, we have

$$(V_\tau^{n+1}-V_\tau^n)(s)=\frac{1}{1-\gamma}\int_S\bigg(\int_A\Big(\hat{A}_n-\mathcal{E}_n+\tau\ln\tfrac{\mathrm{d}\pi^n}{\mathrm{d}\mu}\Big)(s,a)(\pi^{n+1}-\pi^n)(da|s)+\tau\,\mathrm{KL}(\pi^{n+1}|\pi^n)(s)\bigg)d_\rho^{\pi^{n+1}}(ds)\,.$$

Note that (20), together with $\lambda\ge\tau$ guarantees that

$$0\ge\int_A\Big(\hat{A}_n(s,a)+\tau\ln\tfrac{\mathrm{d}\pi^n}{\mathrm{d}\mu}(a|s)\Big)(\pi^{n+1}-\pi^n)(da|s)+\tau\,\mathrm{KL}(\pi^{n+1}|\pi^n)(s)=:F(s)$$

for all $s\in S$. Thus we may apply Lemma 3.3 and get

$$(V_\tau^{n+1}-V_\tau^n)(s)\le F(s)-\frac{1}{1-\gamma}\int_S\int_A\mathcal{E}_n(s,a)(\pi^{n+1}-\pi^n)(da|s)d_\rho^{\pi^{n+1}}(ds)\,.$$

Assume that $\|\mathcal{E}\|_{B_b(S\times A)}=\delta_n<\infty$. Then we have the following approximate L-smoothness:

$$(V_\tau^{n+1}-V_\tau^n)(s)\le F(s)+\frac{2\delta_n}{1-\gamma}\,,\quad s\in S.$$

Applying this in (25) and taking we thus have, for all $s\in S$, that

$$\begin{aligned}\mathrm{KL}(\pi_\tau^*|\pi^{n+1})(s)-\mathrm{KL}(\pi_\tau^*|\pi^n)(s)\le\frac{1}{\lambda}\int_A\Big(\hat{A}_n(s,a)+\tau\ln\tfrac{\mathrm{d}\pi^n}{\mathrm{d}\mu}(a|s)\Big)(\pi_\tau^*-\pi^n)(da|s)\\-\frac{1}{\lambda}(V_\tau^{n+1}-V_\tau^n)(s)+\frac{2\delta_n}{(1-\gamma)\lambda}\,.\end{aligned}\tag{26}$$

Summing up over $n=0,1,\dots,N-1$ we see (spotting the telescoping sums) that for all $s\in S$,

$$\begin{aligned}\mathrm{KL}(\pi_\tau^*|\pi^N)(s)-\mathrm{KL}(\pi_\tau^*|\pi^0)(s)\le\sum_{n=0}^{N-1}\frac{1}{\lambda}\int_A\Big(\hat{A}_n(s,a)+\tau\ln\tfrac{\mathrm{d}\pi^n}{\mathrm{d}\mu}(a|s)\Big)(\pi_\tau^*-\pi^n)(da|s)\\-\frac{1}{\lambda}(V_\tau^N-V_\tau^0)(s)+\frac{2}{(1-\gamma)\lambda}\sum_{n=0}^{N-1}\delta_n\,.\end{aligned}$$

We wish to apply performance difference in due course and so we observe that the above is equivalent to

$$\begin{aligned}\mathrm{KL}(\pi_\tau^*|\pi^N)(s)-\mathrm{KL}(\pi_\tau^*|\pi^0)(s)\le\sum_{n=0}^{N-1}\frac{1}{\lambda}\int_A\Big(A_\tau^{\pi^n}(s,a)+\tau\ln\tfrac{\mathrm{d}\pi^n}{\mathrm{d}\mu}(a|s)\Big)(\pi_\tau^*-\pi^n)(da|s)\\+\sum_{n=0}^{N-1}\frac{1}{\lambda}\int_A\mathcal{E}_n(s,a)(\pi_\tau^*-\pi^n)(da|s)-\frac{1}{\lambda}(V_\tau^N-V_\tau^0)(s)+\frac{2}{(1-\gamma)\lambda}\sum_{n=0}^{N-1}\delta_n\,.\end{aligned}\tag{27}$$

Notice that $V_\tau^N(s)\ge V_\tau^*(s)$ and so $(V_\tau^N-V_\tau^0)(s)\ge(V_\tau^*-V_\tau^0)(s)$ for all $N\in\mathbb{N}$. Let

$$y^n:=\int_S\mathrm{KL}(\pi_\tau^*|\pi^n)(s)d_\rho^{\pi_\tau^*}(ds)\ \text{ and }\ \alpha:=-\int_S(V_\tau^*-V^0)(s)d_\rho^{\pi_\tau^*}(ds)$$

so that, after integrating (27) over $d_\rho^{\pi_\tau^*}$ and using $\|\mathcal{E}\|_{B_b(S\times A)} = \delta_n < \infty$ we have

$$y^N - y^0 \le \sum_{n=0}^{N-1} \frac{1}{\lambda} \int_S \int_A \frac{\delta V_\tau^n}{\delta \pi}(s,a)(\pi_\tau^* - \pi^n)(da|s) d_\rho^{\pi_\tau^*}(ds) + \frac{2}{\lambda}\sum_{n=0}^{N-1}\delta_n + \frac{\alpha}{\lambda} + \frac{2}{(1-\gamma)\lambda}\sum_{n=0}^{N-1}\delta_n \,.$$

Using the performance difference lemma, see Lemma 2.2, and upper bounding the approximation error terms we get

$$y^N - y^0 \le \sum_{n=0}^{N-1} \left[\frac{1-\gamma}{\lambda}(V^{\pi_\tau^*} - V^{\pi^n})(\rho) - \frac{\tau}{\lambda}\int_S \mathrm{KL}(\pi_\tau^*|\pi^n)(s) d_\rho^{\pi_\tau^*}(ds) \right] + \frac{\alpha}{\lambda} + \frac{4}{(1-\gamma)\lambda}\sum_{n=0}^{N-1}\delta_n \,.$$

Since since $\mathrm{KL}(\cdot|\cdot) \ge 0$ we get that

$$y^N - y^0 \le N\frac{1-\gamma}{\lambda}\Big(V_\tau^{\pi_\tau^*}(\rho) - \min_{n=0,1,\dots,N-1} V_\tau^{\pi^N}(\rho)\Big) + \frac{\alpha}{\lambda} + \frac{4}{(1-\gamma)\lambda}\sum_{n=0}^{N-1}\delta_n \,.$$

Hence

$$N\frac{1-\gamma}{\lambda}\Big(\min_{n=0,1,\dots,N-1} V_\tau^{\pi^N}(\rho) - V_\tau^{\pi_\tau^*}(\rho)\Big) \le \frac{\alpha}{\lambda} + y^0 + \frac{4}{(1-\gamma)\lambda}\sum_{n=0}^{N-1}\delta_n \,.$$

and so

$$0 \le \min_{n=0,1,\dots,N-1} V_\tau^{\pi^N}(\rho) - V_\tau^{\pi_\tau^*}(\rho) \le \frac{1}{N}\frac{\alpha + \lambda y^0}{1-\gamma} + \frac{1}{N}\frac{4}{(1-\gamma)^2}\sum_{n=0}^{N-1}\delta_n \,.$$

3.3. Natural policy gradient is mirror descent. Natural policy gradient (NPG) leads to the same updates as mirror descent and we'll show this for log-linear policies. NPG in RL is due to Kakade [2001] but the argument connecting to mirror descent updates is closer to Agarwal et al. [2019].

Let $\frac{\mathrm{d}\pi_\theta}{\mathrm{d}\mu}(a|s) := \frac{e^{g_\theta(s,a)}}{Z_\theta(s)}$, $Z_\theta(s) := \int_A e^{g_\theta(s,a')}\mu(da')$ with $g_\theta(s,a) = (\theta, \phi(s,a))_{\mathbb{H}}$. Let us defined the Fisher information matrix

$$F(\theta) := \int_S \int_A \nabla_\theta \ln \tfrac{\mathrm{d}\pi_\theta}{\mathrm{d}\mu} \otimes \nabla_\theta \ln \tfrac{\mathrm{d}\pi_\theta}{\mathrm{d}\mu}(a|s)\pi_\theta(da|s) d_\rho^{\pi_\theta}(da|s)\,,$$

where for $\theta, \theta' \in \mathbb{H}$ we have $(\theta \otimes \theta')_{jk} = \theta_j \theta'_k$. Let

$$\phi_{\pi_\theta} := \phi(s,a) - \int_A \phi(s,a')\pi_\theta(da'|s)\,.$$

Recalling (15) we have that $\nabla_\theta \ln \frac{\mathrm{d}\pi_\theta}{\mathrm{d}\mu}(a|s) = \nabla_\theta g_\theta(s,a) - \int_A (\nabla_\theta g_\theta)(s,a')\frac{\mathrm{d}\pi_\theta}{\mathrm{d}\mu}(a'|s)\mu(da') = \phi_{\pi_\theta}(s,a)$. Hence

$$F(\theta) = \int_S \int_A \phi_{\pi_\theta} \otimes \phi_{\pi_\theta}(s,a)\pi_\theta(da|s) d_\rho^{\pi_\theta}(da|s)\,.$$

Natural policy gradient (NPG) updates are

$$\theta_{n+1} = \theta_n - \eta F(\theta)^\dagger \nabla_\theta V_\tau^{\pi_{\theta^n}}(\rho)\,, \quad n = 0, 1, \dots\,, \quad \theta^0 \in \mathbb{H} \text{ given.} \tag{28}$$

Here, for $M \in \mathcal{L}(\mathbb{H}, \mathbb{H})$ we use $M^\dagger$ to denote the Moore–Penrose pseudo-inverse (which coincides with M^{-1} for invertible M).

Proposition 3.4. *If given $\theta \in \mathbb{H}$ we take $\ln \frac{\mathrm{d}\pi_{\pi_\theta}}{\mathrm{d}\mu}(a|s) = (\theta, \phi_\theta)_{\mathbb{H}}$ and thus obtain π_{θ_n} corresponding to θ_n then $\pi_{\theta_{n+1}}$ with θ_{n+1} given by the NPG update (28) is equal to π^{n+1} given by*

$$\pi_{\theta_{n+1}}(\cdot|s) = \operatorname*{arg\,min}_{m\in\mathcal{P}(A)} \int_A \big(\hat{w}(\theta_n) + \tau\theta_n, \phi_{\pi_{\theta_n}}(s,a)\big)_{\mathbb{H}}(m(da) - \pi_{\theta_n}(da|s)) + \lambda\,\mathrm{KL}(m|\pi_{\theta_n}(\cdot|s))$$

which is the mirror descent update (17) *where the flat derivative is replaced by its approximation* $\hat{A}_n = (\hat{w}(\theta) + \tau\theta, \phi_{\pi_\theta})_{\mathbb{H}}$.

Proof. To see the connection between (28) and the mirror descent updates, let[2]

$$L^{\pi_\theta}(w) := \frac{1}{2}\int_S\int_A |A_\tau^{\pi_\theta}(s,a) - (w, \phi_{\pi_\theta}(s,a))_{\mathbb{H}}|^2 \pi_\theta(da|s) d_\rho^{\pi_\theta}(ds)\,, \tag{29}$$

where $A_\tau^{\pi_\theta}(s,a) = Q_\tau^{\pi_\theta}(s,a) - V_\tau^{\pi_\theta}(s)$. Notice that

$$\nabla_w L^{\pi_\theta}(w) = \int_S\int_A (A_\tau^{\pi_\theta}(s,a) - (w, \phi_{\pi_\theta}(s,a))_{\mathbb{H}})\phi_{\pi_\theta}(s,a)\pi_\theta(da|s)d_\rho^{\pi_\theta}(ds)$$

and so the first order condition for any minimizer $\hat{w}$ of (29) is

$$\int_S\int_A (\hat{w}, \phi_{\pi_\theta}(s,a))_{\mathbb{H}}\phi_{\pi_\theta}(s,a)\pi_\theta(da|s)d_\rho^{\pi_\theta}(ds) = \int_S\int_A A_\tau^{\pi_\theta}(s,a)\phi_{\pi_\theta}(s,a)\pi_\theta(da|s)d_\rho^{\pi_\theta}(ds)\,.$$

Moreover, for any $w \in \mathbb{H}$ we have $F(\theta)w = \int_S\int_A (w, \phi_{\pi_\theta}(s,a))_{\mathbb{H}}\phi_{\pi_\theta}(s,a)\pi_\theta(da|s)d_\rho^{\pi_\theta}(ds)$. Noting also that the minimizer above depends on θ we have

$$F(\theta)\hat{w}(\theta) = \int_S\int_A A_\tau^{\pi_\theta}(s,a)\phi_{\pi_\theta}(s,a)\pi_\theta(da|s)d_\rho^{\pi_\theta}(ds)\,.$$

Note that the Moore–Penrose pseudo-inverse provides the smallest norm solution to this i.e.

$$\hat{w}(\theta) = F(\theta)^\dagger \int_S\int_A A_\tau^{\pi_\theta}(s,a)\phi_{\pi_\theta}(s,a)\pi_\theta(da|s)d_\rho^{\pi_\theta}(ds)\,.$$

This, together with (16) leads to

$$\begin{aligned} F(\theta)^\dagger \nabla_\theta V_\tau^{\pi_\theta}(\rho) &= \tfrac{1}{1-\gamma} F(\theta)^\dagger \mathbb{E}_{a\sim\pi_\theta(\cdot|s)}^{s\sim d_\rho^{\pi_\theta}}\Big[\Big(A_\tau^{\pi_\theta}(s,a) + \tau \ln \tfrac{\mathrm{d}\pi_\theta}{\mathrm{d}\mu}(a|s)\Big)\phi_{\pi_\theta}(s,a)\Big] \\ &= \tfrac{1}{1-\gamma}\big(\hat{w}(\theta) + \tau\theta\big)\,. \end{aligned}$$

If $F(\theta)$ is invertible then $F(\theta)^{-1}\nabla_\theta V_\tau^{\pi_\theta}(\rho) = \frac{1}{1-\gamma}\big(\hat{w}(\theta) + \tau\theta\big)$. So the NPG stepping scheme (28) becomes

$$\theta_{n+1} = \theta_n - \tfrac{\eta}{1-\gamma}\big(\hat{w}(\theta_n) + \tau\theta_n\big)\,,\quad n = 0, 1, \dots\,,\quad \theta_0 \in \mathbb{H} \text{ given.}$$

Letting $\lambda = \eta(1-\gamma)^{-1}$ we have

$$(\theta_{n+1}, \phi)_{\mathbb{H}} = (\theta_n, \phi)_{\mathbb{H}} - \tfrac{1}{\lambda}\big(\hat{w}(\theta_n) + \tau\theta_n, \phi(s,a)\big)_{\mathbb{H}}\,.$$

Since $\ln \frac{\mathrm{d}\pi_{\pi_{\theta_n}}}{\mathrm{d}\mu}(a|s) = (\theta_n, \phi)_{\mathbb{H}} - \Big(\theta_n, \int_A \phi(\cdot, a')\pi_{\theta_n}(da'|\cdot)\Big)_{\mathbb{H}}$ and collecting all the terms constant in a in some $b = b(s)$ we then have

$$\ln \tfrac{\mathrm{d}\pi_{\theta_{n+1}}}{\mathrm{d}\mu}(a|s) = \ln \tfrac{\mathrm{d}\pi_{\pi_{\theta_n}}}{\mathrm{d}\mu}(a|s) - \tfrac{1}{\lambda}\big(\hat{w}(\theta_n) + \tau\theta_n, \phi_{\pi_{\theta_n}}(s,a)\big)_{\mathbb{H}} + b(s)\,,$$

with b chosen such that $\pi_{\theta_{n+1}} \in \mathcal{P}(A|S)$. Hence

$$\ln \tfrac{\mathrm{d}\pi_{\theta_{n+1}}}{\mathrm{d}\pi_{\pi_{\theta_n}}}(a|s) = -\tfrac{1}{\lambda}\big(\hat{w}(\theta_n) + \tau\theta_n, \phi_{\pi_{\theta_n}}(s,a)\big)_{\mathbb{H}} + b(s)\,.$$

And so

$$\tfrac{\mathrm{d}\pi_{\theta_{n+1}}}{\mathrm{d}\pi_{\pi_{\theta_n}}}(a|s) = \exp\Big(-\tfrac{1}{\lambda}\big(\hat{w}(\theta_n) + \tau\theta_n, \phi_{\pi_{\theta_n}}(s,a)\big)_{\mathbb{H}} + b(s)\Big)\,.$$

Due to Dupuis and Ellis [1997], Lemma 1.4.3 we know that

$$\pi_{\theta_{n+1}}(\cdot|s) = \operatorname*{arg\,min}_{m\in\mathcal{P}(A)} \int_A \big(\hat{w}(\theta_n) + \tau\theta_n, \phi_{\pi_{\theta_n}}(s,a)\big)_{\mathbb{H}}(m(da) - \pi_{\theta_n}(da|s)) + \lambda\,\mathrm{KL}(m|\pi_{\theta_n}(\cdot|s))\,.$$

[2]As you see we are not including the $\ln \frac{\mathrm{d}\pi_\theta}{\mathrm{d}\mu}$ term. The reason is that as it's just an additive term we can trivially see that $|\ln \frac{\mathrm{d}\pi_\theta}{\mathrm{d}\mu} - (y, \phi_{\pi_\theta})_{\mathbb{H}}|^2$ is minimized by $y = \theta$.

This is the mirror descent update (17) where the flat derivative is replaced by its approximation $\hat{A}_n = (\hat{w}(\theta) + \tau\theta, \phi_{\pi_\theta})_{\mathbb{H}}$. □

Appendix A. Proofs for results in Section 1

Proof of Lemma 1.3. For $u \in B_b(S)$ let

$$(Lu)(s) := \int_A \left(c(s,a) + \gamma \int_S u(s') P(ds'|s,a) \right) \pi(da|s), \quad \forall s \in S\,.$$

Then

$$|(Lu)(s)| \le \int_A |c(s,a)| + \gamma \int_S |u(s')| P(ds'|s,a)\pi(da|s) \le \|c\|_{B_b(S)} + \gamma \|u\|_{B_b(S)}\,.$$

Hence $L : B_b(S) \to B_b(s)$ is well defined. Moreover for $u, v \in B_b(s)$ we have

$$|(Lu - Lv)(s)| = \left| \int_A \gamma \int_S (u(s') - v(s')) P(ds'|s,a)\pi(da|s) \right| \le \gamma \|u - v\|_{B_b(S)}\,.$$

Hence $\|Lu - Lv\|_{B_b(S)} \le \gamma \|u - v\|_{B_b(S)}$ and so $L : B_b(S) \to B_b(s)$ is a contraction on the Banach space of bounded functions and there is a unique solution $\bar{V} \in B_b(S)$ to the equation

$$\bar{V}(s) = \int_A \left(c(s,a) + \gamma \int_S \bar{V}(s') P(ds'|s,a) \right) \pi(da|s), \quad \forall s \in S\,. \tag{30}$$

It remains to show that $\bar{V} = V_0^\pi$. Iterating (30) and using (1), we get that for all $N \in \mathbb{N}$,

$$\bar{V}(s) = \mathbb{E}_s^\pi \sum_{n=0}^{N-1} \gamma^n c(s_n, a_n) + \gamma^N \int_A P^{(N)} \bar{V}(s,a)\pi(da|s),$$

where $P^{(N)} \in \mathcal{L}(B_b(S), B_b(S \times A))$ is the operator induced by the N-step transition kernel. Since $P^{(N)}$ has an operator norm less than one, we have $\int_A P^{(N)} \bar{V}(s,a)\pi(da|s) \le \|\bar{V}\|_{B_b(S)}$, and hence by Lebesgue's dominated convergence theorem, for all $s \in S$,

$$\bar{V}(s) = \mathbb{E}_s^\pi \sum_{n=0}^{\infty} \gamma^n c(s_n, a_n) = V_0^\pi(s),$$

where the last identity used the definition of V_0^π in (2). This proves the desired identity. □

Proof of Theorem 1.4. This proof can mostly be seen as a special case of the proof of the DPP for generic Borel state and action spaces (e.g., [Hernández-Lerma and Lasserre, 2012, Theorem 4.2.3]) once one enriches the action space to $\mathcal{P}(A)$ and understands the entropy/KL as an additional cost. Here, we present a self-contained proof for the reader's convenience.

Let $\tau > 0$ be fixed. For each $u \in B_b(S)$ and each $s \in S$, define

$$\begin{aligned} T_\tau u(s) &= \inf_{m \in \mathcal{P}(A)} \int_A \left[c(s,a) + \tau \ln \frac{\mathrm{d}m}{\mathrm{d}\mu}(a|s) + \gamma \int_S u(s') P(ds'|s,a) \right] m(da) \\ &= \tau \inf_{m \in \mathcal{P}(A)} \left[\tau^{-1} \int_A Q_u(s,a) m(da) + \mathrm{KL}(m|\mu(\cdot|s)) \right], \end{aligned}$$

where $Q_u(s,a) := c(s,a) + \gamma \int_S u(s') P(ds'|s,a)$. Since $\|Q_u\|_{B_b(S\times A)} \le \|c\|_{B_b(S\times A)} + \gamma \|u\|_{B_b(S)}$, by [Dupuis and Ellis, 1997, Proposition 1.4.2], for each $s \in S$, we have

$$T_\tau u(s) = -\tau \ln \int_A \exp\left(-\tau^{-1} Q_u(s,a)\right) \mu(da|s)\,,$$

where the infimum is uniquely attained at $\pi_u \in \mathcal{P}_\mu(A|S)$ given by

$$\pi_u(da|s) = \frac{\exp\left(-\tau^{-1} Q_u(s,a)\right)}{\int_A \exp\left(-\tau^{-1} Q_u(s,a')\right) \mu(da'|s)} \mu(da|s).$$

It is clear that $T_\tau u : S \to \mathbb{R}$ is measurable by Fubini's theorem. Moreover, since the natural logarithm is increasing, for all $s \in S$, we have

$$|T_\tau u(s)| \le \tau \left| \ln \int_A \exp\left(\tau^{-1} \|Q_u\|_{B_b(S\times A)}\right) \mu(da|s) \right| \le \|c\|_{B_b(S\times A)} + \gamma \|u\|_{B_b(S)} .$$

Thus, the Bellman operator $T_\tau : B_b(S) \to B_b(S)$ is well defined.

We will now show that T_τ is a contraction on the Banach space $B_b(S)$, following the proof in Haarnoja et al. [2017]. Let $u, v \in B_b(S)$ be fixed. Note that for all $(s,a) \in S \times A$, we have

$$Q_v(s,a) - Q_u(s,a) = \gamma \int_S \big(v(s') - u(s')\big) P(ds'|s,a) \le \gamma \|u - v\|_{B_b(S)} .$$

Using that the natural logarithm is increasing, for all $s \in S$, we get

$$\begin{aligned} -T_\tau u(s) &= \tau \ln \int_A \exp(\tau^{-1} Q_v(s,a) - \tau^{-1} Q_u(s,a) - \tau^{-1} Q_v(s,a)) \mu(da|s) \\ &\le \tau \ln \left(\exp\left(\frac{\gamma}{\tau} \|u - v\|_{B_b(S)}\right) \int_A \exp\left(-\tau^{-1} Q_v(s,a)\right) \mu(da|s) \right) \\ &= \gamma \|u - v\|_{B_b(S)} - T_\tau v(s) , \end{aligned}$$

and hence $T_\tau v(s) - T_\tau u(s) \le \gamma \|u - v\|_{B_b(S)}$. Swapping the roles of u and v in the above, we find $T_\tau u - T_\tau v \le \gamma \|u - v\|_{B_b(S)}$, and thus

$$\|T_\tau u - T_\tau v\|_{B_b(S)} \le \gamma \|u - v\|_{B_b(S)} .$$

Since $\gamma \in [0,1)$, $T_\tau : B_b(S) \to B_b(S)$ is a contraction, and there is a unique fixed point $\bar{V} \in B_b(S)$ such that $T_\tau \bar{V} = \bar{V}$. In particular, for all $s \in S$,

$$\bar{V}(s) = \inf_{m \in \mathcal{P}(A)} \int_A \left[c(s,a) + \tau \ln \frac{\mathrm{d}m}{\mathrm{d}\mu}(a|s) + \gamma \int_S \bar{V}(s') P(ds'|s,a) \right] m(da) \tag{31}$$

$$= \int_A \left[c(s,a) + \tau \ln \frac{\mathrm{d}\bar{\pi}}{\mathrm{d}\mu}(a|s) + \gamma \int_S \bar{V}(s') P(ds'|s,a) \right] \bar{\pi}(da|s), \tag{32}$$

where the the unique infimum is attained at $\bar{\pi} \in \mathcal{P}_\mu(A|S)$ given by

$$\bar{\pi}(da|s) = \frac{\exp\left(-\tau^{-1} Q_{\bar{V}}(s,a)\right)}{\int_A \exp\left(-\tau^{-1} Q_{\bar{V}}(s,a')\right) \mu(da'|s)} \mu(da|s) .$$

Thus, we have proved that $\bar{V}$ is the unique bounded solution of the Bellman equation (32).

It remains to show that $\bar{V}(s) = V_\tau^*(s)$ for all $s \in S$. We will first show $\bar{V}(s) \ge V_\tau^*(s)$ for all $s \in S$. Iterating (32) and using (1), we get that for all $N \in \mathbb{N}$,

$$\bar{V}(s) = \mathbb{E}_s^{\bar{\pi}} \sum_{n=0}^{N-1} \gamma^n \left(c(s_n, a_n) + \tau \ln \frac{\mathrm{d}\bar{\pi}}{\mathrm{d}\mu}(a_n|s_n) \right) + \gamma^N \int_A P^{(N)} \bar{V}(s,a) \bar{\pi}(da|s),$$

where $P^{(N)} \in \mathcal{L}(B_b(S), B_b(S \times A))$ is the operator induced by the N-step transition kernel. Since $P^{(N)}$ has operator norm less than one, we have $\int_A P^{(N)} \bar{V}(s,a) \bar{\pi}(da|s) \le \|\bar{V}\|_{B_b(S)}$, and hence by Lebesgue's dominated convergence theorem, for all $s \in S$,

$$\bar{V}(s) = \mathbb{E}_s^{\bar{\pi}} \sum_{n=0}^{\infty} \gamma^n \left(c(s_n, a_n) + \tau \ln \frac{\mathrm{d}\bar{\pi}}{\mathrm{d}\mu}(a_n|s_n) \right) \ge V_\tau^*(s).$$

We will now show that $\bar{V}(s) \le V_\tau^\pi(s)$ for all $\pi \in \Pi$ and $s \in S$, which then implies that $\bar{V}(s) \le V^*(s)$ for all $s \in S$. Let $\pi = \{\pi_n\}_{n \in \mathbb{N}_0} \in \Pi$, so that for each $n \in \mathbb{N}_0$, $\pi_n \in \mathcal{P}(A|H_n)$. Let $s \in S$ denote an

arbitrary fixed initial state. Without loss of generality, we assume $\pi_n \in \mathcal{P}_\mu(A|H_n)$ for all $n \in \mathbb{N}_0$, since otherwise $V^\pi(s) = \infty$. For each $n \in \mathbb{N}$, applying (1) and adding and subtracting, we find

$$\begin{aligned}\gamma^{n+1}\mathbb{E}_s^\pi\left[\bar{V}(s_{n+1})|h_n, a_n\right] &= \gamma^{n+1}\int_S \bar{V}(s')P(ds'|s_n, a_n)\\ &= \gamma^n\left[c(s_n, a_n) + \tau\ln\frac{\mathrm{d}\pi}{\mathrm{d}\mu}(a_n|h_n) + \gamma\int_S \bar{V}(s')P(ds'|s_n, a_n)\right] - \gamma^n\left(c(s_n, a_n) + \tau\ln\frac{\mathrm{d}\pi}{\mathrm{d}\mu}(a_n|h_n)\right).\end{aligned}$$

By the tower property of conditional expectations,

$$\begin{aligned}\gamma^{n+1}\mathbb{E}_s^\pi\left[\bar{V}(s_{n+1})|h_n\right] &= \gamma^{n+1}\mathbb{E}_s^\pi\left[\mathbb{E}_s^\pi\left[\bar{V}(s_{n+1})|h_n, a_n\right]|h_n\right]\\ &= \gamma^n\mathbb{E}_s^\pi\left[c(s_n, a_n) + \tau\ln\frac{\mathrm{d}\pi}{\mathrm{d}\mu}(a_n|h_n) + \gamma\int_S \bar{V}(s')P(ds'|s_n, a_n)\middle|h_n\right]\\ &\quad - \gamma^n\mathbb{E}_s^\pi\left[c(s_n, a_n) + \tau\ln\frac{\mathrm{d}\pi}{\mathrm{d}\mu}(a_n|h_n)\middle|h_n\right]\\ &= \gamma^n\mathbb{E}_s^\pi\left[\int_A\left(c(s_n, a) + \tau\ln\frac{\mathrm{d}\pi}{\mathrm{d}\mu}(a|h_n) + \gamma\int_S \bar{V}(s')P(ds'|s_n, a)\right)\pi_n(da|h_n)\middle|h_n\right]\\ &\quad - \gamma^n\mathbb{E}_s^\pi\left[c(s_n, a_n) + \tau\ln\frac{\mathrm{d}\pi}{\mathrm{d}\mu}(a_n|h_n)\middle|h_n\right],\end{aligned}$$

where we have used (1) in the last identity.

Applying (31) (with $m = \pi_n(da|h_n)$),

$$\int_A\left(c(s_n, a) + \tau\ln\frac{\mathrm{d}\pi}{\mathrm{d}\mu}(a|h_n) + \gamma\int_S \bar{V}(s')P(ds'|s_n, a)\right)\pi_n(da|h_n) \geq \bar{V}(s_n),$$

and hence

$$\gamma^{n+1}\mathbb{E}_s^\pi\left[\bar{V}(s_{n+1})|h_n\right] \geq \gamma^n\mathbb{E}_s^\pi[\bar{V}(s_n)|h_n] - \gamma^n\mathbb{E}_s^\pi\left[c(s_n, a_n) + \tau\ln\frac{\mathrm{d}\pi}{\mathrm{d}\mu}(a_n|h_n)\middle|h_n\right].$$

Rearranging the inequality, applying the expectation operator $\mathbb{E}^\pi$, and using a telescoping sum argument, we get

$$\mathbb{E}_s^\pi\left[\sum_{n=0}^{N-1}\gamma^n\left(c(s_n, a_n) + \tau\ln\frac{\mathrm{d}\pi}{\mathrm{d}\mu}(a_n|s_n)\right)\right] \geq \bar{V}(s) - \gamma^N\mathbb{E}_s^\pi\left[\bar{V}(s_N)\right].$$

Letting $N \to \infty$ and using that $\bar{V} \in B_b(S)$, we find $V^\pi(s) \geq \bar{V}(s)$ for all $s \in S$, which gives $\bar{V}(s) \leq V_\tau^*(s)$ for all $s \in S$, and finally $\bar{V} \equiv V_\tau^*$. This completes the proof. □

Proof of Lemma 1.7. For each $u \in B_b(S)$, $\pi \in \Pi_\mu$, and $s \in S$, define

$$L_\tau^\pi u(s) = \int_A\left(c(s, a) + \tau\ln\frac{\mathrm{d}\pi}{\mathrm{d}\mu}(a|s) + \gamma\int_S u(s')P(ds'|s, a)\right)\pi(da|s),$$

which is well-defined as $\pi \in \Pi_\mu$ and $\left\|\int_S u(s')P(ds'|s, \cdot)\right\|_{B_b(A)} \leq \|u\|_{B_b(S)}$. Recalling that $\pi = \boldsymbol{\pi}(f)$ for some $f \in B_b(S \times A)$, and thus by Proposition 1.6,

$$\left\|c + \tau\ln\frac{\mathrm{d}\pi}{\mathrm{d}\mu}\right\|_{B_b(S\times A)} \leq \|c\|_{B_b(S\times A)} + \tau\left\|\ln\frac{\mathrm{d}\pi}{\mathrm{d}\mu}\right\|_{B_b(S\times A)} \leq \|c\|_{B_b(S\times A)} + 2\tau\|f\|_{B_b(S\times A)}.$$

Thus for all $u \in B_b(S)$, $L_\tau u \in B_b(S)$ and

$$\|L_\tau^\pi u\|_{B_b(S)} \leq \|c\|_{B_b(S\times A)} + 2\tau\|f\|_{B_b(S\times A)} + \gamma\|u\|_{B_b(S)}.$$

Moreover, for all $u, v \in B_b(S)$, we have

$$\|L_\tau^\pi u - L_\tau^\pi v\|_{B_b(S)} = \gamma\left\|\int_A\int_S (u(s') - v(s'))P(ds'|\cdot, a)\pi(da|\cdot)\right\|_{B_b(s)} \leq \gamma\|u - v\|_{B_b(S)}.$$

Since $\gamma \in [0,1)$, the map $L_\tau : B_b(S) \to B_b(S)$ is a contraction, and thus there is a unique $V \in B_b(S)$ such that for all $s \in S$,

$$V(s) = \int_A \left(c(s,a) + \tau \ln \frac{\mathrm{d}\pi}{\mathrm{d}\mu}(a|s) + \gamma \int_S V(s') P(ds'|s,a) \right) \pi(da|s)\,. \tag{33}$$

To verify $V = V_\tau^\pi$, iterating (33) and using (1), we get that for all $N \in \mathbb{N}$,

$$V(s) = \mathbb{E}_s^\pi \sum_{n=0}^{N-1} \gamma^n \left(c(s_n,a_n) + \tau \ln \frac{\mathrm{d}\pi}{\mathrm{d}\mu}(a_n|s_n) \right) + \gamma^N \int_A P^{(N)} V(s,a) \pi(da|s),$$

where $P^{(N)} \in \mathcal{L}(B_b(S), B_b(S \times A))$ is the operator induced by the N-step transition kernel. Since $P^{(N)}$ has an operator norm less than one, we have $\int_A P^{(N)} V(s,a)\pi(da|s) \le \|V\|_{B_b(S)}$, and hence by Lebesgue's dominated convergence theorem, for all $s \in S$,

$$V(s) = \mathbb{E}_s^\pi \sum_{n=0}^{\infty} \gamma^n \left(c(s_n,a_n) + \tau \ln \frac{\mathrm{d}\pi}{\mathrm{d}\mu}(a_n|s_n) \right) = V_\tau^\pi(s),$$

where the last identity used the definition of V_τ^π in (2). This proves the desired identity. □

References

Alekh Agarwal, Sham M Kakade, Jason D Lee, and Gaurav Mahajan. Optimality and approximation with policy gradient methods in Markov decision processes. *arXiv preprint arXiv:1908.00261*, 2019.

Pierre-Cyril Aubin-Frankowski, Anna Korba, and Flavien Léger. Mirror descent with relative smoothness in measure spaces, with application to Sinkhorn and EM. *Advances in Neural Information Processing Systems*, 35:17263–17275, 2022.

Semih Cayci, Niao He, and R Srikant. Linear convergence of entropy-regularized natural policy gradient with linear function approximation. *arXiv preprint arXiv:2106.04096*, 2021.

Paul Dupuis and Richard S. Ellis. *A weak convergence approach to the theory of large deviations*. John Wiley & Sons, Inc., New York, 1997.

Tuomas Haarnoja, Haoran Tang, Pieter Abbeel, and Sergey Levine. Reinforcement learning with deep energy-based policies. In *International Conference on Machine Learning*, pages 1352–1361. PMLR, 2017.

Onésimo Hernández-Lerma and Jean B Lasserre. *Discrete-time Markov control processes: basic optimality criteria*, volume 30. Springer Science & Business Media, 2012.

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

Bekzhan Kerimkulov, James-Michael Leahy, David Šiška, Lukasz Szpruch, and Yufei Zhang. A Fisher–Rao gradient flow for entropy-regularised Markov decision processes in Polish spaces. *Foundations of Computational Mathematics*, 2025a.

Bekzhan Kerimkulov, David Šiška, Łukasz Szpruch, and Yufei Zhang. Mirror descent for stochastic control problems with measure-valued controls. *Stochastic Processes and their Applications*, page 104765, 2025b.

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

Jincheng Mei, Chenjun Xiao, Csaba Szepesvari, and Dale Schuurmans. On the global convergence rates of softmax policy gradient methods. In *International Conference on Machine Learning*, pages 6820–6829. PMLR, 2020.

Arkadij Semenovič Nemirovskij and David Borisovich Yudin. *Problem complexity and method efficiency in optimization*. Wiley-Interscience, 1983.

John Schulman, Philipp Moritz, Sergey Levine, Michael Jordan, and Pieter Abbeel. High-dimensional continuous control using generalized advantage estimation. *arXiv preprint arXiv:1506.02438*, 2015.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

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

Christopher JCH Watkins and Peter Dayan. Q-learning. *Machine learning*, 8(3):279–292, 1992.