Mirror descent actor-critic methods for entropy regularised MDPs in general spaces: stability and convergence[1]

David Šiška

7th July 2026

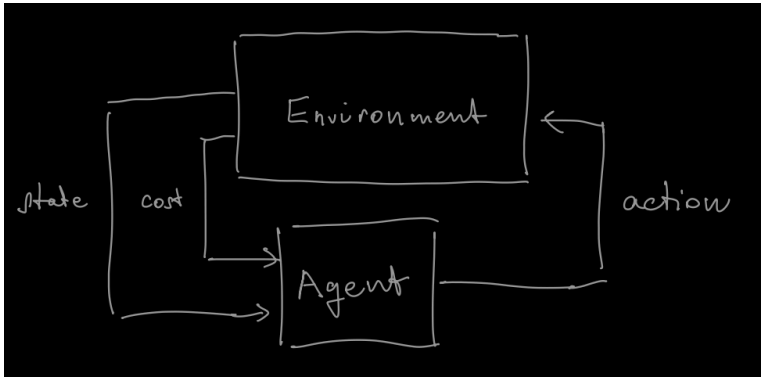
CIRM workshop on stochastic control and reinforcement learning

[1] Mostly based on joint work with D. Zorba and L. Szpruch [Zorba et al., 2026] https://arxiv.org/abs/2602.10838

Outline

- Reinforcement Learning (RL) and its (entropy regularized) MDP formulation
  - Performance difference lemma and policy gradient theorem
  - Difficulty of convergence analysis due to lack of convexity
- Mirror descent
  - Role of performance difference as convexity and L-smoothness
  - Convergence rate MDP case with inexact advantage
- Exact update Mirror descent actor critic [Zorba et al., 2026]
  - Critic TD learning
  - Stability
  - Convergence

Reinforcement Learning (RL)



RL Aim: learn to interact with an environment in an optimal (cost minimizing) way.

Data: $(s_t, a_t, c_t, s_{t+1}, a_{t+1}, \ldots)$.

Mathematical abstraction: MDP.

MDPs and relaxed, regularized MDPs

Key MDP / RL results

Overview of RL [Sutton and Barto, 2018] and results in discrete state-action spaces

- Classical policy gradient [Sutton et al., 1999].
- Natural policy gradient [Kakade, 2001].
- Actor-critic method [Haarnoja et al., 2018].
- Mirror descent method [Tomar et al., 2020].
- Convergence of classical PG in tabular setting [Mei et al., 2021].

Continuous state-action spaces: [Doya, 2000], [Van Hasselt, 2012], [Manna et al., 2022].

Entropy regularised: [Haarnoja et al., 2017, Geist et al., 2019].

Infinite-horizon Markov decision problem (S, A, P, c, γ):

- S is the state space, A is the action space
- $P \in \mathcal{P}(S|S \times A)$ is the transition probability kernel
- $c \in B_b(S \times A)$ is a cost function, and γ discount factor
- $H_n := (S \times A)^n \times S$ is the space of admissible histories

Aim: minimise the objective over policies $\alpha = (\alpha_n)_{n \in \mathbb{N}}$ s.t. $\alpha_n : H_n \to A$ measurable:

$$V^\alpha(s) = \mathbb{E}_s^\alpha \sum_{n=0}^{\infty} \gamma^n c(s_n, a_n)\,, \tag{1}$$

with $a_n := \alpha_n(h_n)$, $h_n = (s_0, a_0, \ldots, s_{n-1}, a_{n-1}, s_n)$ and $s_{n+1} \sim P(\cdot|s_n, a_n)$, $s_0 = s$.

Infinite-horizon Markov decision model (S, A, P, c, γ):

- S is the state space, A is the action space,
- $P \in \mathcal{P}(S|S \times A)$ is the transition probability kernel,
- $c \in B_b(S \times A)$ is a cost function, and γ a discount factor,
- $H_n := (S \times A)^n \times S$ is the space of admissible histories,
- $\tau \geq 0$ strenght of entropy regularizer,
- for $\mu', \mu \in \mathcal{P}(A)$ define $\mathrm{KL}(\mu'|\mu) = \int_A \ln \frac{\mathrm{d}\mu'}{\mathrm{d}\mu}(a)\mu'(da)$ if $\mu' \ll \mu$, and $+\infty$ otherwise.

Aim: minimise over relaxed policies $\pi = (\pi_n)_{n\in\mathbb{N}}$ s.t. $\pi_n : H_n \to P(A)$ measurable the objective:

$$V_\tau^\pi(s) = \mathbb{E}_s^\pi \left[\sum_{n=0}^{\infty} \gamma^n \Big(\int_A c(s_n, a)\, \pi_n(da) + \tau\, \mathrm{KL}(\pi_n|\mu) \Big) \right] \in \mathbb{R} \cup \{+\infty\}, \tag{2}$$

with $\pi_n := \pi_n(h_n)$, $h_n = (s_0, a_0, \ldots, s_{n-1}, a_{n-1}, s_n)$ and $s_{n+1} \sim \int_A P(\cdot|s_n, a)\, \pi_n(da)$, $s_0 = s$.

Relative entropy - variational formula

Variational formula: for $f \in B_b(A)$:

$$\inf_{\nu \in \mathcal{P}(A)} \left(\int_A f \, d\nu + \mathrm{KL}(\nu|\mu) \right) = -\ln \int_A e^{-f} \mu(da) \,,$$

and if

$$\frac{\mathrm{d}\nu^*}{\mathrm{d}\mu}(a) = \frac{e^{-f(a)}}{\int_A e^{-f(a')} \mu(da')}$$

then $\nu^* = \mathrm{argmin}_{\nu \in \mathcal{P}(A)} \left(\int_A f \, d\nu + \mathrm{KL}(\nu|\mu) \right)$.

Bellman principle aka Dynamic programming principle (DPP)

Recall $H_n := (S \times A)^n \times S$ is the space of admissible histories.

Let $V_\tau^* : S \to \mathbb{R}$ be

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

where infimum is over policies $\pi = (\pi_n)_{n\in\mathbb{N}}$ s.t. $\pi_n : H_n \to P(A)$ is measurable.

Theorem 1 (Dynamic programming principle)

Let $\tau > 0$. The optimal value function V_τ^* is the unique bounded solution of

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

DPP consequences for $\tau > 0$

For all $s \in S$,

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

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

$$\pi_\tau^*(da|s) = \exp\left(-(Q_\tau^*(s,a) - V_\tau^*(s))/\tau\right) \mu(da), \quad \forall s \in S. \tag{4}$$

Let

$$\Pi_\mu = \{\pi \in \mathcal{P}(A|S) : \ln \tfrac{d\pi}{d\mu} \in B_b(S \times A)\}.$$

Then

$$\inf_\pi V_\tau^\pi = V_\tau^*(s) = \inf_{\pi \in \Pi_\mu} V_\tau^\pi.$$

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

$$Q_\tau^\pi(s,a) = c(s,a) + \gamma \int_S V_\tau^\pi(s') P(ds'|s,a). \tag{5}$$

Lemma 2

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:

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

Note that from this and defn. of the Q-function (5) we have for all $\pi \in \Pi_\mu$ and $s \in S$ that

$$V_\tau^\pi(s') = \int_A \Big(Q_\tau^\pi(s',a') + \tau \ln \tfrac{\mathrm{d}\pi}{\mathrm{d}\mu}(a'|s') \Big) \pi(da'|s'), \quad \forall s \in S\,. \tag{6}$$

Using this in the defn. of the Q-function (5) we have the on policy Q-Bellman equation

$$Q_\tau^\pi(s,a) = c(s,a) + \gamma \int_S \int_A \Big(Q_\tau^\pi(s',a') + \tau \ln \tfrac{\mathrm{d}\pi}{\mathrm{d}\mu}(a'|s') \Big) \pi(da'|s') P(ds'|s,a)\,, \forall (s,a)\,. \tag{7}$$

What does "solving our RL problem" mean?

We will say we've "solved the RL problem" if we can find a near optimal policy by repeatedly interacting with an environment / simulator.

But under what "interaction" assumption?

Always accept we do **not** have access to costs c and transitions $P \in \mathcal{P}(S|S \times A)$.

Hardest

- The environment will initialise once at $s \sim \rho \in \mathcal{P}(S)$.
- We interact with the environment, incurring the costs.

Intermediate

- Initially and at resets the simulator will initialise at $s \sim \rho \in \mathcal{P}(S)$.
- We have access to a simulator of the environment and we can repeatedly use it.
- It will run until termination or until **we reset it**.

Easiest

- We can initialize the simulator at any $s \in S$ of **our choice**.
- We observe $s \in S$, can choose $a \in A$ and observe $s' \sim P(\cdot|s, a)$ and $c(s, a)$.

Key meta-algorithms

Policy gradient (PG)

Initialize environment, parametrized policy π_{θ_0},
for $n = 0, 1, \ldots, N_{\text{episodes}}$ **do**
  Clear memory buffer
  for $t = 0, 1, \ldots, N_{\text{steps in episode}}$ **do**
    Observe state s_t
    Sample action $a_t \sim \pi_{\theta_n}(a_t|s_t)$
    Execute a_t in environment, accept cost c_t, new state s_{t+1}
    Store $(s_t, a_t, c_t, \log \pi_{\theta_n}(a_t|s_t), V(s_t))$
    $t \leftarrow t + 1$, $s_t \leftarrow s_{t+1}$ in memory.
  end for
  Estimate $\nabla_\theta V^{\pi_\theta}$ from memory data
  Update policy parameters

$$\theta_{n+1} = \theta_n - \eta \nabla_\theta V^{\pi_\theta} .$$

end for

Q-learning

Initialize environment, schedule $(\eta_k)_k$, Q_{θ_0}, $a_0 \sim \mu$.
for $k = 0, 1, \ldots$ **do**
  Observe state s_k
  Execute a_k in environment, accept cost c_k, new state s_{k+1}
  Set $\hat{\theta}_k = \theta_k$, set $\pi_k(\cdot|s_k) \propto \exp(-\frac{1}{\tau} Q_{\hat{\theta}_k}(s_k|\cdot))\mu$
  Take $a_{k+1} \sim \pi_k(\cdot|s_{k+1})$.
  Calculate the one step TD error

$$\mathcal{E}_\theta := c_k + \gamma \left[Q_{\hat{\theta}_k}(s_{k+1}, a_{k+1}) + \tau \ln \frac{\mathrm{d}\pi_k}{\mathrm{d}\mu}(a_{k+1}|s_{k+1}) \right] - Q_\theta(s_k, a_k) .$$

  Update Q-function parameters:

$$\theta_{k+1} = \theta_k - \eta_k \nabla_\theta |\mathcal{E}_{\theta_k}|^2$$

  $s_k \leftarrow s_{k+1}$, $a_k \leftarrow a_{k+1}$
end for

Classical gradient descent

Policy gradient (PG) methods

Even if S and A are of finite cardinality and

$$\nabla_\pi V_\tau^\pi(\rho) := (\nabla_{\pi(s,a)} V_\tau^\pi(\rho))_{(s,a)\in S\times A}$$

with $(\nabla_{\pi(s,a)} V_\tau^\pi(\rho))_{(s,a)\in S\times A} \in \mathbb{R}^{N_S\times N_A}$ is a gradient in $\mathbb{R}^{N_S\times N_A}$ **not in** $\mathcal{P}(A|S) \equiv \Delta(A)^{N_S}$.

~~$\pi_{n+1} = \pi_n - \eta \nabla_\pi V_\tau^\pi(\rho).$~~

Parametrize: Direct: $\frac{\mathrm{d}\pi_\theta}{\mathrm{d}\mu}(a|s) \propto e^{\theta(s,a)}$, Log-linear: $\frac{\mathrm{d}\pi_\theta}{\mathrm{d}\mu}(a|s) \propto e^{\langle\theta, g(s,a)\rangle}$ with $g : S\times A \to \mathbb{R}^p$ basis, ... Then

$$\theta_{n+1} = \theta_n - \eta \nabla_\theta V_\tau^{\pi_\theta}(\rho)$$

classical gradient descent: [Cauchy, 1847] seems fine.

Theorem 3 (PG for parametrization)

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*

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

Let

$$d^\pi(ds'|s) = (1-\gamma)\sum_{n=0}^{\infty}\gamma^n P_\pi^n(ds'|s) \quad \text{and} \quad d_\rho^\pi(ds) = \int_S d^\pi(ds|s')\rho(ds')\,. \tag{8}$$

Lemma 4

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

$$f(s) = \gamma \int_A \int_S f(s')P(ds'|s,a)\pi(da|s) + g(s)\,. \tag{9}$$

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

Lemma 5 (Performance difference[1])

For all $\rho \in \mathcal{P}(S)$ and $\pi, \pi' \in \Pi_\mu$,

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

[1] Tabular case [Howard, 1960, Ch. 7, p. 87], re-discovered in RL context [Kakade and Langford, 2002], Polish spaces + entropy [Kerimkulov et al., 2025a]

Proposition 1

Let $\tau \geq 0$ and $\rho \in \mathcal{P}(S)$. For all $\pi, \pi' \in \Pi_\mu \subset \mathcal{P}(A|S)$

$$\lim_{\varepsilon \searrow 0} \varepsilon^{-1} \Big(V_\tau^{(1-\varepsilon)\pi + \varepsilon\pi'}(\rho) - V_\tau^\pi(\rho) \Big) = \tfrac{1}{1-\gamma} \int_S \int_A \left(Q_\tau^\pi(s,a) + \tau \ln \tfrac{\mathrm{d}\pi}{\mathrm{d}\mu}(a|s) - V_\tau^\pi(s) \right) (\pi' - \pi)(da|s) d_\rho^\pi(ds) \,. \quad (10)$$

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 1, given $\nu \in \mathcal{P}(S)$ satisfying $d_\rho^\pi \ll \nu$,

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

with

$$\tfrac{\delta V_\tau^\pi(\rho)}{\delta \pi}(s,a) = Q_\tau^\pi(s,a) + \tau \ln \tfrac{\mathrm{d}\pi}{\mathrm{d}\mu}(s,a) - V_\tau^\pi(s) \,. \quad (11)$$

Lack of convexity in softmax parametrization

Consider *minimizing*, over $\pi \in \mathcal{P}(A)$ the objective

$$v^{\pi} := \int_A c(a)\pi(da)\,.$$

We trivially have, for $\pi, \pi' \in \mathcal{P}(A)$ that

$$v^{(1-\varepsilon)\pi+\varepsilon\pi'} \leq (1-\varepsilon)v^{\pi} + \varepsilon v^{\pi'}\,, \varepsilon \in [0,1]$$

so $\mathcal{P}(A) \ni \pi \mapsto v^{\pi} \in \mathbb{R}$ is convex.

Consider *minimizing*, over $\theta \in \mathbb{R}^{|A|}$ the objective

$$v^{\pi_\theta} := \int_A c(a)\pi_\theta(da),$$

with $\pi_\theta(a) = \frac{e^{\theta(a)}}{\sum_{a'} e^{\theta(a')}}$.

The map $\mathbb{R}^p \ni \theta \mapsto v^{\pi_\theta} \in \mathbb{R}$ is *not* convex [Mei et al., 2020, Propn. 1].

Recall

$$V_\tau^\pi(\rho) = \mathbb{E}_{s_0\sim\rho}^\pi\left[\sum_{n=0}^\infty \gamma^n\Big(\int_A c(s_n,a)\,\pi_n(da) + \tau\,\mathrm{KL}(\pi_n|\mu)\Big)\right] \in \mathbb{R}\cup\{+\infty\}\,,$$

Definition (Convexity)

If for any $\pi,\pi' \in \mathcal{P}(A|S)$ we have

$$V_\tau^{(1-\varepsilon)\pi+\varepsilon\pi'}(\rho) \le (1-\varepsilon)V_\tau^\pi(\rho) + \varepsilon V_\tau^{\pi'}(\rho)\,, \varepsilon\in[0,1]$$

then $\mathcal{P}(A|S)\ni\pi\mapsto V_\tau^\pi(\rho)\in\mathbb{R}\cup\{+\infty\}$ is convex.

The map $\mathcal{P}(A|S)\ni\pi\mapsto V_\tau^\pi(\rho)\in\mathbb{R}\cup\{+\infty\}$ is **not** convex, e.g. [Giegrich et al., 2024, Proposition 2.4] **even if underlying dynamics is linear and costs convex.**

Continuous time gradient flow

$$\frac{d}{ds}\theta_s = -\nabla f(\theta_s) \implies \frac{d}{ds}\big[f(\theta_s) - f(\theta^*)\big] = -|\nabla f(\theta_s)|^2 .$$

Non-uniform Polyak–Łojasiewicz: there is $\mu : \mathbb{R}^p \to (0, \infty)$ s.t. for all $\theta \in \mathbb{R}^p$

$$0 \le f(\theta) - f(\theta^*) \le \mu(\theta)|\nabla f(\theta)|^2 .$$

Hence

$$\frac{d}{ds}\big[f(\theta_s) - f(\theta^*)\big] = -|\nabla f(\theta_s)|^2 \le -\mu^{-1}(\theta_s)\big[f(\theta_s) - f(\theta^*)\big]$$

Grönwall:

$$0 \le f(\theta_s) - f(\theta^*) \le \big[f(\theta_0) - f(\theta^*)\big] \exp\Big(- \int_0^s \mu^{-1}(\theta_r)\, dr \Big) .$$

Q: Is $\inf_r \mu^{-1}(\theta_r) \ge \alpha > 0$?

Convergence of classical PG in MDPs

- Discrete time LQR: Polyak–Łojasiewicz (PL) / gradient dominance established and so PG has linear convergence [Fazel et al., 2018, Bu et al., 2019, Hu et al., 2023].
- In general MDPs with discrete state-action setting best PL result is non-uniform [Mei et al., 2020] but shown lower bounded along PG and hence convergence.
- In general MDPs with in general spaces under Q^π-realisibility best PL result is non-uniform [Chen et al., 2026] but *for suitable basis functions* shown lower bounded along PG and hence convergence.

Mirror descent

Static optimization mirror descent:

- Goes back to at least [Nemirovski, 1979].
- Modern proximal point form [Beck and Teboulle, 2003].
- For general probability measures [Aubin-Frankowski et al., 2022].

RL mirror descent update

$$\pi^{n+1}(\cdot|s) = \operatorname*{argmin}_{m\in\mathcal{P}(A)} \int_A \frac{\delta V_\tau^{\pi_n}}{\delta \pi}(s,a)(m(da) - \pi^n(da|s)) + \lambda\, \mathrm{KL}(m|\pi^n(\cdot|s))\,. \tag{12}$$

N.B.

$$\pi^{n+1}(\cdot|s) = \tfrac{1}{Z_n(s)} \exp\Big(-\tfrac{1}{\lambda}\tfrac{\delta V_\tau^{\pi_n}}{\delta \pi}(s,\cdot)\Big)\pi^n(\cdot|s)\,, \quad Z_n(s) = \int_A \exp\big(-\tfrac{1}{\lambda}\tfrac{\delta V_\tau^{\pi_n}}{\delta \pi}(s,a')\big)\pi^n(da'|s)\,.$$

Mirror descent for MDPs

Discrete space MDPs and constants **dependent** on $|S|$ and $|A|$:

- [Cen et al., 2022], entropy regularised, show linear convergence for disc. time. mirror descent
- [Cayci et al., 2021] same setting i.e natural policy gradient, log-linear policies i.e. mirror desc with func. approx.
- [Xiao, 2022] and [Khodadadian et al., 2022] achieved *linear convergence for unregularised MDPs* with inexact policy evaluation by employing geometrically increasing step sizes in NPG.

Discrete space MDPs and constants **independent of** of $|S|$ and $|A|$:

- [Lan, 2023] linear convergence of policy mirror descent with arbitrary convex regularisers and [Zhan et al., 2023] convergence rates independent of action space dimension.

MDPs with **general** S and A:

- Discrete step mirror descent and Fisher–Rao flow: Exponential convergence for entropy regularized MDPs in Polish state & action spaces [Kerimkulov et al., 2025a].

Mirror descent policy improvement (with exact update)

Recall the mirror descent update (12):

$$\pi^{n+1}(\cdot|s) = \operatorname*{argmin}_{m\in\mathcal{P}(A)} \int_A \frac{\delta V_\tau^{\pi^n}}{\delta \pi}(s,a)(m(da) - \pi^n(da|s)) + \lambda\, \mathrm{KL}(m|\pi^n(\cdot|s))\,.$$

From the performance difference lemma, see Lemma (5), we see that

$$\begin{aligned}(V_\tau^{n+1} - V_\tau^n)(\rho) &= \tfrac{1}{1-\gamma}\int_S \big(\int_A \tfrac{\delta V_\tau^n}{\delta \pi}(s,a)(\pi^{n+1} - \pi^n)(da|s) + \tau\, \mathrm{KL}(\pi^{n+1}|\pi^n)(s)\big)\, d_\rho^{\pi^{n+1}}(ds) \\ &\le \tfrac{1}{1-\gamma}\int_S \big(\int_A \tfrac{\delta V_\tau^n}{\delta \pi}(s,a)(\pi^{n+1} - \pi^n)(da|s) + \lambda\, \mathrm{KL}(\pi^{n+1}|\pi^n)(s)\big)\, d_\rho^{\pi^{n+1}}(ds)\,.\end{aligned} \tag{13}$$

From the mirror descent update (12) we have, for all $\pi \in \Pi_\mu$ and $s \in S$ that

$$\int_A \tfrac{\delta V_\tau^n}{\delta \pi}(s,a)(\pi - \pi^n)(da|s) + \lambda\, \mathrm{KL}(\pi|\pi^n)(s) \ge \int_A \tfrac{\delta V_\tau^n}{\delta \pi}(s,a)(\pi^{n+1} - \pi^n)(da|s) + \lambda\, \mathrm{KL}(\pi^{n+1}|\pi^n)(s)\,.$$

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

$$\int_A \tfrac{\delta V_\tau^n}{\delta \pi}(s,a)(\pi^{n+1} - \pi^n)(da|s) + \lambda\, \mathrm{KL}(\pi^{n+1}|\pi^n)(s) \le 0\,. \tag{14}$$

From (13) we have

$$(V_\tau^{n+1} - V_\tau^n)(\rho) \le 0\,.$$

Lemma 6

Let $\pi, \pi' \in \Pi_\mu$ *satisfy* $\int_A \frac{\delta V_\tau^{\pi'}}{\delta \pi}(s,a)(\pi - \pi')(da|s) + \tau\,\mathrm{KL}(\pi|\pi')(s) \leq 0$ *for all* $s \in S$. *Then for all* $s \in S$,

$$(V_\tau^\pi - V_\tau^{\pi'})(s) \leq \int_A \frac{\delta V_\tau^{\pi'}}{\delta \pi}(s,a)(\pi - \pi')(da|s) + \tau\,\mathrm{KL}(\pi|\pi')(s)\,.$$

In particular with $\pi' = \pi_{old}$ *and* $\pi = \pi_{new}$ *given by the exact update* (12) *satisfy this.*

Proof. From perf. diff. lemma and Lan's trick:

$$\begin{aligned}(V_\tau^\pi - V_\tau^{\pi'})(s) &\leq \tfrac{1}{1-\gamma} \int_S \big(\int_A \tfrac{\delta V_\tau^{\pi'}}{\delta \pi}(s',a)(\pi - \pi')(da|s') + \tau\,\mathrm{KL}(\pi|\pi')(s') \big) d_s^\pi(ds') \\ &\leq \int_A \tfrac{\delta V_\tau^{\pi'}}{\delta \pi}(s,a)(\pi - \pi')(da|s) + \tau\,\mathrm{KL}(\pi|\pi')(s)\,,\end{aligned}$$

since, if $F \leq 0$ then

$$\tfrac{1}{1-\gamma} \int_S F(s')\, d_s^\pi(ds') = \int_S F(s') P_\pi^0(ds'|s) + \textstyle\sum_{k=1}^\infty \int_S \gamma^k F(s') P_\pi^k(ds'|s) \leq \int_S F(s') \delta_s(ds') = F(s)\,. \tag{15}$$

We can write performance difference as

$$(V_\tau^{\pi} - V_\tau^{\pi'})(\rho) = \left\langle \tfrac{\delta V_\tau^{\pi'}}{\delta \pi}, \pi - \pi' \right\rangle_{d_\rho^{\pi}} + \tfrac{\tau}{1-\gamma} \int_S \mathrm{KL}(\pi|\pi')(s) d_\rho^{\pi}(ds)\,, \tag{16}$$

where

$$\langle h, \hat{\pi} \rangle_{\rho,\pi} := \tfrac{1}{1-\gamma} \int_S \int_A h(s,a) \hat{\pi}(da|s)\, d_\rho^{\pi}(ds)$$

and

$$\tfrac{\delta V_\tau^{\pi}(\rho)}{\delta \pi}(s,a) = Q_\tau^{\pi}(s,a) + \tau \ln \tfrac{\mathrm{d}\pi}{\mathrm{d}\mu}(s,a) - V_\tau^{\pi}(s)\,.$$

Convergence of mirror descent with approximate advantage

Mirror descent with approximation

Recall that $\frac{\delta V_\tau^{\pi^n}}{\delta \pi} = A_\tau^{\pi^n} + \tau \ln \frac{\mathrm{d}\pi^n}{\mathrm{d}\mu} = Q_\tau^{\pi^n} - V_\tau^{\pi^n} + \tau \ln \frac{\mathrm{d}\pi^n}{\mathrm{d}\mu}$.

Updates can only be made with an approximation $\hat{A}_n(s,a) = A_\tau^{\pi^n}(s,a) + \mathcal{E}_n(s,a)$.

Consider the scheme

$$\pi^{n+1}(da|s) = \operatorname*{argmin}_{m \in \mathcal{P}(A)} \int_A \left(\hat{A}_n(s,a) + \tau \ln \frac{\mathrm{d}\pi^n}{\mathrm{d}\mu}(a|s) \right) (m(da) - \pi^n(da|s)) + \lambda \, \mathrm{KL}(m|\pi^n(\cdot|s)) \,. \tag{17}$$

Ingredients for convergence of mirror descent

- Convexity (strong for "linear" rate)
- L-smoothness
- Three point lemma

Lemma 7 (Three point lemma / Bregman proximal inequality)

Let $G : M_\mu \to \mathbb{R}$ be convex. For all $m' \in M_\mu$ let

$$m^* = \operatorname*{argmin}_{m \in M_\mu} \left\{ G(m) + \mathrm{KL}(m|m') \right\} . \tag{18}$$

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

$$G(m) + \mathrm{KL}(m|m') \geq G(m^*) + \mathrm{KL}(m|m^*) + \mathrm{KL}(m^*|m') . \tag{19}$$

The proof of Lemma 7 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].

Theorem 8

Given $\pi_0 \in \Pi_\mu$, let $(\pi_n)_{\mathbb{N}}$ be given by

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

where $\hat{A}_n(s,a) = A_\tau^{\pi_n}(s,a) + \mathcal{E}_n(s,a)$ and $\|\mathcal{E}\|_{B_b(S\times A)} = \delta_n < \infty$ for all $n \in \mathbb{N}$. Then

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

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

This is a small extension of results in [Kerimkulov et al., 2025a], [Lan, 2023].

Actor-Critic MD stepping

Function approximation and assumptions

Let $\phi_i : S \times A \to \mathbb{R}$, $i = 1, \dots, N$ be features. Write $\phi : S \times A \to \mathbb{R}^N$ and $|\phi(s,a)|_2^2 := \sum_{i=1}^N \phi_i(s,a)^2$.

Assumption 2 (Bounded features)

It holds that $\sup_{s \in S, a \in A} |\phi(s,a)|_2 \le 1$.

Given $\theta \in \mathbb{R}^N$ let the *approximate Q-function* be

$$Q(s,a;\theta) := \langle \theta, \phi(s,a) \rangle = \sum_{i=1}^N \theta_i \phi_i(s,a)\,.$$

Further given $\pi \in \mathcal{P}(A|S)$ let the *approximate first variation* be

$$A(s,a;\theta,\pi) = Q(s,a;\theta) + \tau \ln \tfrac{d\pi}{d\mu}(s,a) - \textstyle\int_A \big(Q(s,a;\theta) + \tau \ln \tfrac{d\pi}{d\mu}(s,a)\big)\pi(da|s). \tag{20}$$

Assumption 3 (Feature-initialisation compatibility)

Let $\beta \in \mathcal{P}(S \times A)$ *be fixed. Then*

$$\lambda_\beta := \lambda_{\min}\big(\textstyle\int_{S\times A} \phi(s,a)\phi(s,a)^\top \beta(ds\,da)\big) > 0.$$

Assumption 4 (Q^π-realisability)

For all $\pi \in \Pi_\mu$ there exists $\theta_\pi \in \mathbb{R}^N$ such that $Q^\pi_\tau(s,a) = \langle \theta_\pi, \phi(s,a) \rangle$ for all $(s,a) \in S \times A$.

An MDPs is *linear* if there exists exists $\phi : S \times A \to \mathbb{R}^N$, $w \in \mathbb{R}^N$ and a sequence $\{\psi_i\}_{i=1}^N$ with $\psi_i \in \mathcal{M}(S)$ such that for all $(s,a) \in S \times A$,

$$c(s,a) = \langle w, \phi(s,a) \rangle, \qquad P(ds' \mid s,a) = \sum_{i=1}^N \phi_i(s,a)\psi_i(ds').$$

If the MDP is linear then by (5)

$$Q^\pi_\tau(s,a) = c(s,a) + \gamma \int_S V^\pi_\tau(s') P(ds'|s,a) = \langle w, \phi(s,a) \rangle + \gamma \int_S V^\pi_\tau(s') \sum_{i=1}^N \phi_i(s,a)\psi_i(ds').$$

So we have Q^π-realisability with $(\theta_\pi)_i = w_i + \gamma \int_S V^\pi(s')\psi_i(ds')$.

Critic updates with TD

Recall the on-policy Bellman equation for the Q-function:

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

Define the operator $\mathrm{T}_\tau^\pi : B_b(S \times A) \to B_b(S \times A)$ as

$$\mathrm{T}_\tau^\pi f(s,a) = c(s,a) + \gamma \int_{S\times A} f(s',a') P^\pi(ds',da'|s,a) + \tau\gamma \int_S \mathrm{KL}(\pi(\cdot|s')|\mu) P(ds'|s,a). \tag{21}$$

It can be shown this is γ-contraction.

The Mean Squared Bellman Error (MSBE) is defined as

$$\mathrm{MSBE}(\theta,\pi) := \frac{1}{2} \int_{S\times A} |Q(s,a;\theta) - \mathrm{T}_\tau^\pi Q(s,a;\theta)|^2 d_\beta^\pi(da,ds). \tag{22}$$

Here

$$d_\beta^\pi(ds',da') := \int_{S\times A} d^\pi(ds',da'|s,a)\beta(ds,da)\,, \quad d^\pi(ds',da'|s,a) := (1-\gamma)\sum_{n=0}^\infty \gamma^n (P^\pi)^n(ds',da'|s,a)$$

and $P^\pi \in \mathcal{P}(S\times A|S\times A)$ is the kernel which for all $f \in B_b(S\times A)$ satisfies

$$\textstyle\int_{S\times A} f(s',a') P^\pi(ds',da'|s,a) = \int_S \int_A f(s',a')\pi(da'|s')P(ds'|s,a)\,.$$

Let the approximate MD loss be

$$\widetilde{\mathrm{MD}}(\pi, \pi', \theta) = \int_S \left(\int_A A(s, a; \theta, \pi') \pi(da|s) + \tfrac{1}{\lambda} \mathrm{KL}(\pi|\pi')(s) \right) d\rho^{\pi'}(ds)\,. \tag{23}$$

Let the semi-gradient $g : \mathbb{R}^N \times \mathcal{P}(A|S) \to \mathbb{R}^N$ of the MSBE with respect to θ is given by

$$g(\theta, \pi) := \int_{S \times A} (Q(s, a; \theta) - \mathrm{T}_\tau^\pi Q(s, a; \theta)) \phi(s, a)\, d\beta^\pi(da, ds). \tag{24}$$

Mirror descent

Require: Actor step size $\lambda > 0$, $\theta_0 \in \mathbb{R}^N$, $\pi^0 = \Pi_\mu$.

for $n = 0, 1, \ldots, N_{\text{MD steps}}$ **do**
  $\theta^{n+1} \leftarrow \mathrm{argmin}_\theta \mathrm{MSBE}(\theta, \pi^n)$
  $\pi^{n+1} \leftarrow \underset{\pi \in \Pi_\mu}{\mathrm{argmin}}\, \widetilde{\mathrm{MD}}(\pi, \pi^n, \theta^{n+1})$
end for

Mirror descent actor-critic

Require: Critic step size $h > 0$, actor step size $\lambda > 0$, critic-step schedule $M : \mathbb{N} \to \mathbb{N}$, $\theta^0 \in \mathbb{R}^N$, $\pi^0 \in \Pi_\mu$

for $n = 0, 1, 2, \ldots$ **do**
  $\theta^{n,0} \leftarrow \theta^n$
  for $k = 1, \ldots, M(n)$ **do**
    $\theta^{n,k} \leftarrow \theta^{n,k-1} - h\, g(\theta^{n,k-1}, \pi^n)$
  end for
  $\theta^{n+1} \leftarrow \theta^{n,M(n)}$
  $\pi^{n+1} \leftarrow \underset{\pi \in \Pi_\mu}{\mathrm{argmin}}\, \widetilde{\mathrm{MD}}(\pi, \pi^n, \theta^{n+1})$
end for

Let $\Gamma := \lambda_\beta(1-\gamma)(1-\sqrt{\gamma})$.

Theorem 9 (Mirror descent actor-critic near-improvement)

Let Assumptions 2, 3 and 4 hold. Let $0 < h < \min\left\{\frac{\Gamma}{2(1+\gamma)}, \frac{1}{\Gamma}\right\}$. *Then for all* $n \in \mathbb{N}$ *and* $s \in S$ *it holds that*

$$V_\tau^{\pi^*}(s) \leq V_\tau^{\pi^{n+1}}(s) \leq V_\tau^{\pi^n}(s) + \frac{2e^{-\frac{M(n)h\Gamma}{2}}}{1-\gamma}|\theta^n - \theta_{\pi^n}|_2. \tag{25}$$

Theorem 10 (Mirror descent actor-critic stability)

Let Assumptions 2, 3 and 4 hold. Let $0 < h < \min\left\{\frac{\Gamma}{2(1+\gamma)}, \frac{1}{\Gamma}\right\}$ *and* $0 < \tau\lambda < 1$. *Then there exists* $c_0 > 0$ *such that, for any* $c \geq c_0$, *if the number of inner temporal difference steps at mirror descent step* $n \in \mathbb{N}$ *satisfies*

$$M(n) \geq \frac{4}{h\Gamma}\log\left(c(n+1)\right), \tag{26}$$

then there exists $R \geq 0$ *such that for all* $s \in S$ *and* $n \in \mathbb{N}$ *it holds that*

$$\mathrm{KL}(\pi^n(\cdot|s)|\mu) \leq R. \tag{27}$$

Theorem 11 (Best iterate sublinear convergence)

Let Assumptions 2, 3 and 4 hold and let $0 < h < \frac{\Gamma}{2(1+\gamma)}$. *Moreover for each* $n \in \mathbb{N}$, *let the number of inner critic steps satisfy* $M(n) \geq \frac{4}{h\Gamma} \log\left(c(n+1)\right)$ *with* $c \geq c_0$, *where* c_0 *is the stability constant from Theorem 10. Then, for any* $\rho \in \mathcal{P}(S)$, *there exists a* ≥ 0 *such that for all* $n \in \mathbb{N}$ *it holds that*

$$\min_{0 \leq r \leq n-1} V_\tau^{\pi^r}(\rho) - V_\tau^{\pi^*}(\rho) \leq \frac{a}{n}. \tag{28}$$

References I

[Aubin-Frankowski et al., 2022] Aubin-Frankowski, P.-C., Korba, A., and Léger, F. (2022). Mirror descent with relative smoothness in measure spaces, with application to Sinkhorn and EM. *Advances in Neural Information Processing Systems*, 35:17263–17275.

[Beck and Teboulle, 2003] Beck, A. and Teboulle, M. (2003). Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3):167–175.

[Bu et al., 2019] Bu, J., Mesbahi, A., Fazel, M., and Mesbahi, M. (2019). LQR through the lens of first order methods: Discrete-time case. *arXiv preprint arXiv:1907.08921*.

[Cayci et al., 2021] Cayci, S., He, N., and Srikant, R. (2021). Linear convergence of entropy-regularized natural policy gradient with linear function approximation. *arXiv preprint arXiv:2106.04096*.

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

[Chen et al., 2026] Chen, Z., Šiška, D., and Szpruch, L. (2026). Global linear convergence of entropy-regularized softmax policy gradient beyond tabular mdps. *arXiv preprint arXiv:2605.24939*.

[Doya, 2000] Doya, K. (2000). Reinforcement learning in continuous time and space. *Neural computation*, 12(1):219–245.

[Fazel et al., 2018] Fazel, M., Ge, R., Kakade, S., and Mesbahi, M. (2018). Global convergence of policy gradient methods for the linear quadratic regulator. In *International Conference on Machine Learning*, pages 1467–1476. PMLR.

References II

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

[Giegrich et al., 2024] Giegrich, M., Reisinger, C., and Zhang, Y. (2024). Convergence of policy gradient methods for finite-horizon exploratory linear-quadratic control problems. *SIAM Journal on Control and Optimization*, 62(2):1060–1092.

[Haarnoja et al., 2017] Haarnoja, T., Tang, H., Abbeel, P., and Levine, S. (2017). Reinforcement learning with deep energy-based policies. In *International Conference on Machine Learning*, pages 1352–1361. PMLR.

[Haarnoja et al., 2018] Haarnoja, T., Zhou, A., Abbeel, P., and Levine, S. (2018). Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International Conference on Machine Learning*, pages 1861–1870. PMLR.

[Howard, 1960] Howard, R. A. (1960). *Dynamic programming and markov processes*. John Wiley.

[Hu et al., 2023] Hu, B., Zhang, K., Li, N., Mesbahi, M., Fazel, M., and Başar, T. (2023). Toward a theoretical foundation of policy optimization for learning control policies. *Annual Review of Control, Robotics, and Autonomous Systems*, 6:123–158.

[Kakade and Langford, 2002] Kakade, S. and Langford, J. (2002). Approximately optimal approximate reinforcement learning. In *Proceedings of the Nineteenth International Conference on Machine Learning*, pages 267–274.

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

References III

[Kerimkulov et al., 2025a] Kerimkulov, B., Leahy, J.-M., Šiška, D., Szpruch, L., and Zhang, Y. (2025a). A Fisher–Rao gradient flow for entropy-regularised Markov decision processes in Polish spaces. *Foundations of Computational Mathematics*.

[Kerimkulov et al., 2025b] Kerimkulov, B., Šiška, D., Szpruch, Ł., and Zhang, Y. (2025b). Mirror descent for stochastic control problems with measure-valued controls. *Stochastic Processes and their Applications*, page 104765.

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

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

[Manna et al., 2022] Manna, S., Loeffler, T. D., Batra, R., Banik, S., Chan, H., Varughese, B., Sasikumar, K., Sternberg, M., Peterka, T., Cherukara, M. J., et al. (2022). Learning in continuous action space for developing high dimensional potential energy models. *Nature communications*, 13(1):368.

[Mei et al., 2021] Mei, J., Gao, Y., Dai, B., Szepesvari, C., and Schuurmans, D. (2021). Leveraging non-uniformity in first-order non-convex optimization. In *International Conference on Machine Learning*, pages 7555–7564. PMLR.

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

[Nemirovski, 1979] Nemirovski, A. (1979). Efficient methods. *Ekonomika i Mat. Metody*, 15.

References IV

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

[Sutton et al., 1999] Sutton, R. S., McAllester, D., Singh, S., and Mansour, Y. (1999). Policy gradient methods for reinforcement learning with function approximation. *Advances in neural information processing systems*, 12.

[Tomar et al., 2020] Tomar, M., Shani, L., Efroni, Y., and Ghavamzadeh, M. (2020). Mirror descent policy optimization. *arXiv preprint arXiv:2005.09814*.

[Van Hasselt, 2012] Van Hasselt, H. (2012). Reinforcement learning in continuous state and action spaces. In *Reinforcement Learning: State-of-the-Art*, pages 207–251. Springer.

[Xiao, 2022] Xiao, L. (2022). On the convergence rates of policy gradient methods. *arXiv preprint arXiv:2201.07443*.

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

[Zorba et al., 2026] Zorba, D., Šiška, D., and Szpruch, L. (2026). Mirror descent actor-critic methods for entropy regularised mdps in general spaces: stability and convergence. *arXiv preprint arXiv:2602.10838*.

Proof of convergence rate for mirror descent with approximate advantage

Let π^n be generated by inductive application of the approximate mirror descent step (17). Let $V_\tau^n := V_\tau^{\pi^n}$ for $n \in \mathbb{N}$. We begin with an application of Bregman proximal inequality, see Lemma 7. 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 (17) is equivalent to (18) 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

$$\begin{aligned}
&\mathrm{KL}(\pi|\pi^{n+1})(s) - \mathrm{KL}(\pi|\pi^n)(s) \\
&\leq \frac{1}{\lambda} \int_A \Big(\hat{A}_n(s,a) + \tau \ln \tfrac{\mathrm{d}\pi^n}{\mathrm{d}\mu}(a|s) \Big) (\pi - \pi^n)(da|s) \\
&\quad - \frac{1}{\lambda} \int_A \Big(\hat{A}_n(s,a) + \tau \ln \tfrac{\mathrm{d}\pi^n}{\mathrm{d}\mu}(a|s) \Big) (\pi^{n+1} - \pi^n)(da|s) - \mathrm{KL}(\pi^{n+1}|\pi^n)(s) \,.
\end{aligned} \tag{29}$$

From the performance difference, Lemma 5, we have

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

Note that (17), together with $\lambda \geq \tau$ guarantees that

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

for all $s \in S$. Thus we may "forget the future" and get

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

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

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

Applying this in (29) and taking we thus have, for all $s \in S$, that

$$
\begin{aligned}
\mathrm{KL}(\pi_\tau^*|\pi^{n+1})(s) - \mathrm{KL}(\pi_\tau^*|\pi^n)(s) \leq \frac{1}{\lambda} \int_A \big(\hat{A}_n(s,a) + \tau \ln \tfrac{\mathrm{d}\pi^n}{\mathrm{d}\mu}(a|s)\big)(\pi_\tau^* - \pi^n)(da|s) \\
- \tfrac{1}{\lambda}(V_\tau^{n+1} - V_\tau^n)(s) + \tfrac{2\delta_n}{(1-\gamma)\lambda}\,.
\end{aligned} \tag{30}
$$

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

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

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

$$\mathrm{KL}(\pi_\tau^*|\pi^N)(s) - \mathrm{KL}(\pi_\tau^*|\pi^0)(s) \leq \sum_{n=0}^{N-1} \frac{1}{\lambda} \int_A \Big(A_\tau^{\pi^n}(s,a) + \tau \ln \tfrac{\mathrm{d}\pi^n}{\mathrm{d}\mu}(a|s) \Big) (\pi_\tau^* - \pi^n)(da|s)$$
$$+ \sum_{n=0}^{N-1} \frac{1}{\lambda} \int_A \mathcal{E}_n(s,a)(\pi_\tau^* - \pi^n)(da|s) - \frac{1}{\lambda}(V_\tau^N - V_\tau^0)(s) + \frac{2}{(1-\gamma)\lambda} \sum_{n=0}^{N-1} \delta_n \,. \tag{31}$$

Notice that $V_\tau^N(s) \geq V_\tau^*(s)$ and so $(V_\tau^N - V_\tau^0)(s) \geq (V_\tau^* - V_\tau^0)(s)$ 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 (31) over $d_\rho^{\pi_\tau^*}$ and using $\|\mathcal{E}\|_{B_b(S\times A)} = \delta_n < \infty$ we have

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

Using the performance difference lemma, see Lemma 5, and upper bounding the approximation error terms we get

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

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

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

Hence

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

and so the rate follows. □