Policy gradient methods for RL in general spaces

David Šiška

Tutorial for Bridging Stochastic Control And Reinforcement Learning

7th November 2025

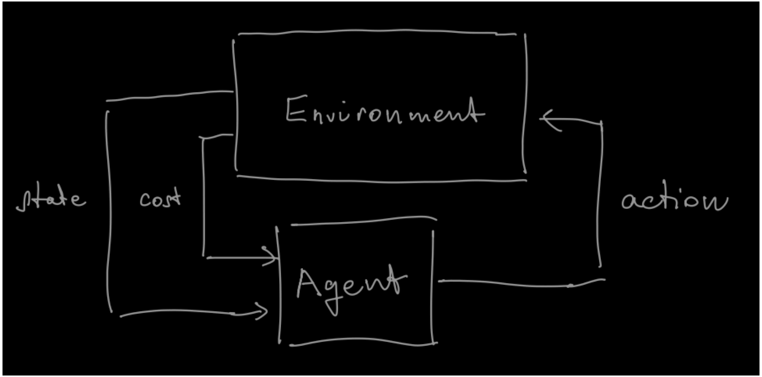
Slides

Notes

Outline

- Reinforcement Learning (RL) and its (entropy regularized) MDP formulation
  - Relative entropy and its key properties
  - Bellman principle for (entropy regularized) MDPs
- Classical Policy Gradient (PG)
  - Performance difference lemma and policy gradient theorem
  - Difficulty of convergence analysis due to lack of convexity
  - Polyak–Łojasiewicz (PL) gradient dominance condition
- Mirror descent
  - Role of Performance difference as convexity and L-smoothness
  - Convergence rate MDP case with inexact advantage
- PPO algorithm
  - Convergence of FR-PPO variant

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].

MDP

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 $s_{n+1} \sim P(\cdot|s_n, a_n)$, $s_0 = s$.

Relaxed and regularized formulation of the MDP

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

Kullback–Leibler divergence aka relative entropy

Relative entropy - definition and basics

If $\nu, \mu \in \mathcal{P}(A)$ and if $\mu(B) = 0 \implies \nu(B) = 0$ for every $B \in \mathcal{B}(A)$ then we say ν is absolutely continuous w.r.t. μ (notation $\nu << \mu$).

For $\mu \in \mathcal{P}(A)$ define

$$\mathcal{P}(A) \ni \nu \mapsto \mathrm{KL}(\nu|\mu) = \begin{cases} \int_A \ln \frac{\mathrm{d}\nu}{\mathrm{d}\mu}\, \nu(da) & \text{if } \nu << \mu\,, \\ +\infty & \text{otherwise}\,. \end{cases}$$

Note that

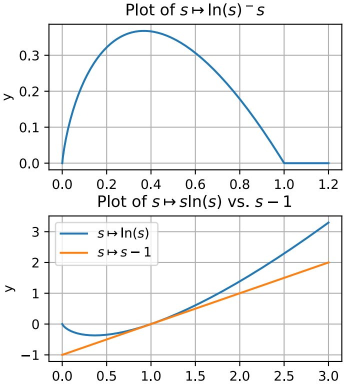
$$\int_A \left(\ln \tfrac{\mathrm{d}\nu}{\mathrm{d}\mu}\right)^- \nu(da) = \int_A \left(\ln \tfrac{\mathrm{d}\nu}{\mathrm{d}\mu}\right)^- \tfrac{\mathrm{d}\nu}{\mathrm{d}\mu}\, \mu(da)$$

and $s \mapsto (\ln s)^- s \geq 0$ is bounded for $s \geq 0$, so KL is well defined.

Moreover $s \ln s \geq s - 1$ for $s \geq 0$ (with equality only if $s = 1$) and so

$$\mathrm{KL}(\nu|\mu) = \int_A \left(\ln \tfrac{\mathrm{d}\nu}{\mathrm{d}\mu}\right) \tfrac{\mathrm{d}\nu}{\mathrm{d}\mu}\, \mu(da) \geq \int_A \left(\tfrac{\mathrm{d}\nu}{\mathrm{d}\mu} - 1\right) \mu(da) = 0\,,$$

with equality only if $\frac{\mathrm{d}\nu}{\mathrm{d}\mu} = 1$ i.e. if $\nu = \mu$.



Relative entropy - variational formula

Useful identity

$$\mathrm{KL}(\nu|\mu) - \mathrm{KL}(\nu'|\mu) = \mathrm{KL}(\nu|\nu') + \int_A \ln \tfrac{\mathrm{d}\nu'}{\mathrm{d}\mu}(a)(\nu - \nu')(da)\,. \tag{3}$$

which holds for any $\nu, \nu' \in \mathcal{P}(A)$ for which the quantities in the identity are finite.

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

Relative entropy - dual formulation and convexity

Donsker–Varadhan variational formula

$$\mathrm{KL}(\nu|\mu) = \sup_{g\in C_b(A)} \left(\int_A g(a)\,\nu(da) - \ln \int_A e^{g(a)}\,\mu(da) \right)$$

and

$$\mathrm{KL}(\nu|\mu) = \sup_{\psi\in B_b(A)} \left(\int_A \psi(a)\,\nu(da) - \ln \int_A e^{\psi(a)}\,\mu(da) \right) .$$

N.B. for fixed g

$$(\nu,\mu) \mapsto \int_A g(a)\,\nu(da) - \ln \int_A e^{g(a)}\,\mu(da)$$

is convex. As a supremum over such g

- $\mathcal{P}(A) \times \mathcal{P}(A) \ni (\nu,\mu) \mapsto \mathrm{KL}(\nu|\mu)$ is convex, lower-semicontinuous.

Moreover

- For fixed $\mu \in \mathcal{P}(A)$ we have

$$\{\nu \in \mathcal{P}(A) : \mathrm{KL}(\nu|\mu) < \infty\} \ni \nu \mapsto \mathrm{KL}(\nu|\mu)$$

  strictly convex, from strict convexity of $[0,\infty) \ni s \mapsto s \ln s \in \mathbb{R}$.

All from [Dupuis and Ellis, 1997, Ch. 1, Sec. 4].

Bellman principle aka Dynamic programming principle (DPP)

DPP with $\tau > 0$

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

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

$$V_\tau^*(s) = \inf_\pi V_\tau^\pi(s), \quad \forall s \in S\,, \tag{4}$$

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

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

Let

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

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

$$Q_\tau^\pi(s,a) = c(s,a) + \gamma \int_S V_\tau^\pi(s') P(ds'|s,a)\,. \tag{6}$$

DPP consequences for $\tau > 0$

Proposition 1

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

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

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

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

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

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

which implies that

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

The rest follows as usual. □

DPP consequences for $\tau > 0$

Lemma 2

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

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

Note that from this and defn. of the Q-function (6) we have for all $\pi \in \Pi_\mu$ and $s \in S$ that

$$V^\pi_\tau(s') = \int_A \Big(Q^\pi_\tau(s',a') + \tau \ln \tfrac{\mathrm{d}\pi}{\mathrm{d}\mu}(a'|s') \Big)\, \pi(da'|s'), \quad \forall s \in S\,. \tag{7}$$

Using this in the defn. of the Q-function (6) we have the on policy Q-Bellman equation

$$Q^\pi_\tau(s,a) = c(s,a) + \gamma \int_S \int_A \Big(Q^\pi_\tau(s',a') + \tau \ln \tfrac{\mathrm{d}\pi}{\mathrm{d}\mu}(a'|s') \Big)\, \pi(da'|s')P(ds'|s,a)\,,\ \forall (s,a)\,. \tag{8}$$

Assumption 2

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.*

Theorem 3 (Dynamic programming principle, $\tau = 0$)

Let Assumption 2 hold. Then the optimal value function $V^* \in B_b(S)$ is the unique solution of the Bellman equation

$$V^*(s) = \min_{a \in A} \left[c(s,a) + \gamma \int_S V^*(s')P(\mathrm{d}s'|s,a) \right]. \tag{9}$$

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 $V^* = V^{\pi^*}$.

Proof. [Hernández-Lerma and Lasserre, 2012, Theorem 4.2.3].

Proposition 3

Let $\pi(da|s) \in \mathcal{P}(A|S)$. *Then*

$$\|V_0^\pi\|_{B_b(S)} \leq \frac{\|c\|_{B_b(S\times A)}}{1-\gamma}\,.$$

Proof. *Exercise, start with* (1) *which is definition of* V_0^π.

Lemma 4

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

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

What does “solving our RL problem” mean?

We will say we’ve “solved our RL problem” if we can find a near optimal policy for the MDP under the assumptions that:

- We do **not** have access to costs c and transitions $P \in \mathcal{P}(S|S \times A)$.
- We choose $\gamma > 0$, $\tau \geq 0$
- We have access to a simulator of the environment and we can repeatedly use it cost-free.
- The simulator will initialise at $s \sim \rho \in \mathcal{P}(S)$ of its choice and will run until termination or until we reset it.

Key meta-algorithms

Policy gradient (PG)

Initialize environment, parametrized policy π_θ,
for $n = 0, 1, \dots, N_{\text{episodes}}$ **do**
  Clear memory buffer
  for $t = 0, 1, \dots, N_{\text{steps in episode}}$ **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, parametrized state-action value Q_θ,
for $n = 0, 1, \dots, N_{\text{episodes}}$ **do**
  Make space in memory buffer
  for $t = 0, 1, \dots, N_{\text{steps in episode}}$ **do**
    Observe state s_t
    Take $a_t \in \operatorname{argmin}_{a \in A} Q_\theta(s_t | a)$
    Execute a_t in environment, accept cost c_t, new state s_{t+1}
    Store (s_t, a_t, c_t, s_{t+1}) in memory
    $t \leftarrow t + 1$, $s_t \leftarrow s_{t+1}$
  end for
  Sample $(s_j, a_j, c_j, s_{j+1})_{j=1}^N$
  For $j = 1, \dots, N$ set

$$v_j = c_j + \gamma \min_{a'} Q_{\theta_n}(s_{j+1}, a') \,.$$

  Let $L(\theta) := \sum_{j=1}^N |v_j - Q_\theta(s_j, a_j)|^2$ and update policy parameters

$$\theta_{n+1} = \theta_n - \eta \nabla_\theta L(\theta) \,.$$

end for

Classical policy gradient

Policy gradient (PG) methods

Recall we are minimizing

$$\Pi_\mu \ni \pi \mapsto V_\tau^\pi(\rho) \in \mathbb{R}\,.$$

A "gradient" update would be

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

But 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** $\mathcal{P}(A|S) \equiv \Delta(A)^{N_S}$.

Policy gradient (PG) methods

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

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.
- Neural-net: $\frac{\mathrm{d}\pi_\theta}{\mathrm{d}\mu}(a|s) \propto e^{g_\theta(s,a)}$.

Then

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

classical gradient descent: [Cauchy, 1847][1] seems fine.

1. How to get $\nabla_\theta V_\tau^{\pi_\theta}(\rho)$ **from data?**
2. Convergence: e.g. is $\theta \mapsto V_\tau^{\pi_\theta}(\rho)$ convex?

[1]From [Lemaréchal, 2012] Cauchy and the gradient method, *Doc. Math. Extra*, 251–254.

Let

$$d^\pi(ds'|s) = (1-\gamma)\sum_{n=0}^{\infty}\gamma^n P_\pi^n(ds'|s) \quad \text{and} \quad d_\rho^\pi(ds) = \int_S d^\pi(ds|s')\rho(ds')\,. \tag{10}$$

We will refer to d^π as the occupancy kernel.

Lemma 5

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

$$f(s) = \gamma\int_A\int_S f(s')P(ds'|s,a)\pi(da|s) + g(s)\,. \tag{11}$$

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

Proof. A kernel $k \in b\mathcal{M}(S|S)$ induces a linear operator $L_k \in \mathcal{L}(B_b(S))$ by

$$B_b(S) \ni h \mapsto L_k h = \int_S h(s')k(ds'|\cdot)\,.$$

Since $\|L_k h\|_{B_b(S)} \leq \|h\|_{B_b(S)}\|k\|_{b\mathcal{M}(S|S)}$ for all $h \in B_b(S)$, $\|L_k\|_{\mathcal{L}(B_b(S))} \leq \|k\|_{b\mathcal{M}(S|S)}$.

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 $\|P_\pi\|_{b\mathcal{M}(S|S)} = 1$,

$$\|L_{\gamma P_\pi}\|_{\mathcal{L}(B_b(S))} \leq \|\gamma P_\pi\|_{b\mathcal{M}(S|S)} = \gamma \|P_\pi\|_{b\mathcal{M}(S|S)} = \gamma < 1\,.$$

The linear equation (11) that f satisfies g is equivalent to

$$(\mathrm{id} - L_{\gamma P_\pi})f = g\,.$$

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 $L^n_{\gamma P_\pi} = L_{\gamma^n P_\pi^n}$ 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 $d^\pi \in \mathcal{P}(S|S)$,

$$f = \sum_{n=0}^{\infty} L^n_{\gamma P_\pi} g = \frac{1}{1-\gamma} \int_S g(s') d^\pi(ds'|\cdot)$$

which is the desired identity. □

Performance difference

Lemma 6 (Performance difference[1])

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

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

[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]

Proof. By (7), for all $s \in S$,

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

Hence for all $s \in S$ we have

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

where the last equality used def. of Q. fn (6) and KL identity (3). Hence, by Fubini's theorem and Lemma 5, for all $s \in S$,

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

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

Proposition 4

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*

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

Proof. Start by getting the estimate for $\|V_\tau^{\pi'} - V_\tau^{\pi}\|_{B_b(S)}$ using Lemma 6 (performance difference).

Proposition 5

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

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

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

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

Proof of PG theorem for general state and action spaces

From the KL identity (3) we get

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

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 4. 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 and the dominated yields the desired limit. □

First variation and chain rule

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

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

with

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

Let $(\mathbb{H}, (\cdot,\cdot)_{\mathbb{H}})$ be a Hilbert space.

Lemma 7 (Chain rule)

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

Proof. Similar to [Kerimkulov et al., 2025a, Proposition 3.8].

Theorem 8 (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*

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

Proof. From Lemma 7 (chain rule) we have:

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

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

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

We just need to rewrite this in terms of expectation to get the conclusion. $\square$

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

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

Hence

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

Some remarks on PG

Remark on estimating the advantage function. First variation:

$$\frac{\delta V_\tau^{\pi_\theta}}{\delta \pi}(s,a) = \underbrace{Q_\tau^\pi(s,a) - V_\tau^\pi(s)}_{=:A_\tau^\pi(s,a) \text{ “advantage”}} + \tau \ln \frac{\mathrm{d}\pi}{\mathrm{d}\mu}(s,a)\,.$$

Advantage $A_\tau^\pi(s,a)$ can be estimated from data: $(s_t, a_t, c_t, s_{t+1}, a_{t+1}, \ldots)$.

$$\hat{A}_\tau^\pi := c_t + \gamma \hat{V}_\tau^\pi(s_{t+1}) - \hat{V}_\tau^\pi(s_t)\,,$$

where $\hat{V}_\tau \approx V_\tau^\pi$. N.B.

$$\mathbb{E}_{s_{t+1} \sim P(\cdot|s_t,a_t)} \hat{A}_\tau^\pi = c(s_t,a_t) + \gamma \int_S \hat{V}_\tau(s') P(ds'|s_t,a_t) - \hat{V}_\tau(s_t)$$

would be equal to $A_\tau^\pi(s_t,a_t)$ if $\hat{V}_\tau = V_\tau^\pi$ in which case it would be unbiased. Alternative

$$\hat{A}_\tau^\pi := \sum_{l=0}^\infty \gamma^{t+l} c_{t+l} - \hat{V}(s_t)\,.$$

Generalised advantage estimation (GAE) formula [Schulman et al., 2015] allows efficient variance vs bias tradeoffs.

Corollary 9 (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. Note that

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

and so

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

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

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

which concludes the calculation. □

Some remarks on PG

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

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

Hence

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

But

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

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].

Some remarks on PG

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

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

Summary of PG so far:

- We have expression for the gradient.
- It can be estimated from data.
- For some simple parametrizations it's nice and simple.

Next: what about convergence?

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

$$v^{(1-\varepsilon)\pi + \varepsilon\pi'} \leq (1-\varepsilon)v^\pi + \varepsilon v^{\pi'}$$

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].

Performance difference vs convexity

Definition (Convexity)

If for some $\tau \geq 0$ we have all $m, m' \in \mathcal{P}(A)$ that

$$F(m) - F(m') \geq \left\langle \frac{\delta F(m')}{\delta m}, m - m' \right\rangle + \tau \,\mathrm{KL}(m|m')\,,$$

then F is convex ($\tau = 0$) or strongly convex ($\tau > 0$).

Equiv.: $\frac{\delta F}{\delta m}(m,\cdot)$ exists and $F((1-\varepsilon)m + \varepsilon m') \leq (1-\varepsilon)F(m) + \varepsilon F(m')$ for all $m, m' \in \mathcal{P}(A)$, $\varepsilon \in [0,1]$.

Performance difference

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

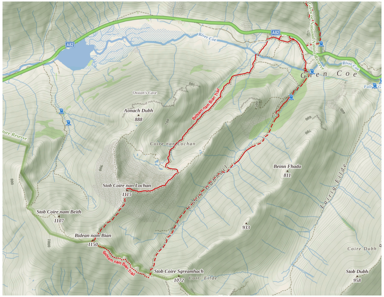
where

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

The map $\Pi_\mu \ni \pi \mapsto V_\tau^{\pi}(\rho) \in \mathbb{R} \cup \{+\infty\}$ 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

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



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

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

Grönwall:

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

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 discrete state-action setting best PL result is non-uniform [Mei et al., 2020] but shown lower bounded along PG and hence convergence.

Mirror descent

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].

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].

Deriving mirror descent

Aim: find

$$\pi^*(\cdot|s) = \arg\min_\pi V_\tau^\pi(s)\,.$$

Let's say we have π_{old}. Fix $\rho \in \mathcal{P}(S)$ and write $V_\tau^\pi = V_\tau^\pi(\rho)$. By perf. diff., Lemma 6,

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

Linearize and penalize with $\lambda \geq \tau$ to not move too far

$$L^\pi := V_\tau^{\pi_{\text{old}}} + \left\langle \frac{\delta V_\tau^{\pi_{\text{old}}}}{\delta\pi}, \pi - \pi_{\text{old}} \right\rangle_{\rho,\pi_{\text{old}}} + \tfrac{\lambda}{1-\gamma}\int_S \text{KL}(\pi|\pi_{\text{old}})(s) d_\rho^{\pi_{\text{old}}}(ds)\,.$$

Mirror descent optimizes $\pi \mapsto L^\pi(x)$ giving

$$\pi_{\text{new}}(da|s) = \arg\min_\pi \left(V_\tau^{\pi_{\text{old}}} + \int_A \tfrac{\delta V_\tau^{\pi_{\text{old}}}}{\delta\pi}(s,a)(\pi - \pi_{\text{old}})(da|s) + \lambda\,\text{KL}(\pi|\pi_{\text{old}})(s) \right).$$

Motivation for studying mirror descent

Policy gradient: introduce parametrized densities $\pi_\theta(da,s) \propto e^{g_\theta(a,s)}\mu(da)$. Step $\eta > 0$:

$$\theta_{\text{new}} = \theta_{\text{old}} + \eta \nabla_\theta V_\tau^{\pi_{\theta_{\text{old}}}}\,,$$

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

Problem: Even if θ_{new} and θ_{old} are close $\pi_{\theta_{\text{old}}}$ and $\pi_{\theta_{\text{new}}}$ may be *very different!*

Instead, re-write the mirror descent objective:

$$\begin{aligned}
L_{\text{MD}}(\theta) &= \left\langle \frac{\delta V_\tau^{\pi_{\theta_{\text{old}}}}}{\delta \pi}, \pi_\theta \right\rangle_{\rho,\pi_{\theta_{\text{old}}}} + \lambda \int_S \text{KL}(\pi_\theta|\pi_{\theta_{\text{old}}})(s)\, d_\rho^{\pi_{\theta_{\text{old}}}}(ds) \\
&= \mathbb{E}^{s\sim d_\rho^{\pi_{\theta_{\text{old}}}}}_{a\sim \pi_\theta(\cdot|s)}\left[\frac{\delta V_\tau^{\pi_{\theta_{\text{old}}}}}{\delta \pi}(s,a) + \lambda\, \text{KL}(\pi_\theta|\pi_{\theta_{\text{old}}})(s)\right] \\
&= \mathbb{E}^{s\sim d_\rho^{\pi_{\theta_{\text{old}}}}}_{a\sim \pi_{\theta_{\text{old}}}(\cdot|s)}\left[\frac{\delta V_\tau^{\pi_{\theta_{\text{old}}}}}{\delta \pi}(s,a)\frac{\mathrm{d}\pi_\theta}{\mathrm{d}\pi_{\theta_{\text{old}}}}(a|s) + \lambda\, \text{KL}(\pi_\theta|\pi_{\theta_{\text{old}}})(s)\right].
\end{aligned}$$

Step $\eta > 0$:

$$\theta_{\text{new}} = \theta_{\text{old}} + \eta \nabla_\theta L_{\text{MD}}(\theta_{\text{old}})\,.$$

Mirror descent policy improvement (with exact update)

Mirror descent update

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

From the performance difference lemma, see Lemma (6), we see that

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

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

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

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

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

From (18) we have

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

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

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

Towards L-smoothness

This is from [Lan, 2023].

Lemma 10

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

$$\frac{1}{1-\gamma}\int_S F(s')\, d_s^\pi(ds') \le F(s)\,. \tag{21}$$

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

$$\begin{aligned}\frac{1}{1-\gamma}\int_S F(s')\, d_s^\pi(ds') &= \int_S F(s') P_\pi^0(ds'|s) + \sum_{k=1}^\infty \int_S \gamma^k F(s') P_\pi^k(ds'|s)\\ &\le \int_S F(s')\delta_s(ds') = F(s)\,.\end{aligned} \tag{22}$$

This concludes the proof. □

Lemma 11 (L-smoothness for exact update)

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* (17) *satisfy this.*

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

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

Convergence of mirror descent with approximate advantage

Ingredients for convergence of mirror descent

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

Lemma 12 (Three point lemma / Bregman proximal inequality)

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

$$m^* = \operatorname*{argmin}_{m \in M_\mu} \left\{ G(m) + \mathrm{KL}(m|m') \right\} . \tag{23}$$

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

$$G(m) + \mathrm{KL}(m|m') \geq G(m^*) + \mathrm{KL}(m|m^*) + \mathrm{KL}(m^*|m') . \tag{24}$$

The proof of Lemma 12 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].

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 12. 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 (23) 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

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

From the performace difference, Lemma 6, we have

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

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

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

for all $s \in S$. Thus we may apply Lemma 10 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:

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

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

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

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

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

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

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

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 (27) 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 6, 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

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

Hence

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

and so

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

Convergence of mirror descent with approximate advantage

Theorem 13

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

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

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

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

Natural policy gradient is mirror descent

Natural policy gradient (NPG)

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 $g_\theta(s,a) = (\theta, \phi(s,a))_{\mathbb{H}}$.

Fisher information matrix

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

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

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

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).

NPG in RL is due to [Kakade, 2001].

Proposition 6

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*

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

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

Remark: Let[2]

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

where $A^{\pi_\theta}_\tau(s,a) = Q^{\pi_\theta}_\tau(s,a) - V^{\pi_\theta}_\tau(s)$. So NPG updates are:

$$\theta_{n+1} = \theta_n, -\tfrac{1}{\lambda}\left(\hat{w}(\theta_n) + \tau\theta_n\right)\,,$$

where $\hat{w}(\theta_n)$ is the minimizer for (29).

[2] We are not including the $\ln \frac{\mathrm{d}\pi_\theta}{\mathrm{d}\mu}$ term. 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 $y = \theta$.

Proof. 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

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

Have

$$F(\theta)^\dagger \nabla_\theta V_\tau^{\pi_\theta}(\rho) = \tfrac{1}{1-\gamma}\left(\hat{w}(\theta) + \tau\theta\right).$$

So the NPG stepping scheme (28) becomes

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

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

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

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

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

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

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

And so

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

Then

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

due to [Dupuis and Ellis, 1997], Lemma 1.4.3. $\square$

PPO and FR-PPO

Connection to PPO

Proximal policy optimization (PPO) [Schulman et al., 2017] optimizes:

$$J_{\text{PPO}}(\theta) = \mathbb{E}_{a \sim \pi_{\theta_{\text{old}}}(\cdot|s)}^{s \sim d_\rho^{\pi_{\theta_{\text{old}}}}} \Bigg(\min \Bigg[\frac{\delta V_\tau^{\pi_{\theta_{\text{old}}}}}{\delta \pi}(s,a) \frac{\mathrm{d}\pi_\theta}{\mathrm{d}\pi_{\theta_{\text{old}}}}(a|s), \\ \text{clip}\left(1-\varepsilon, 1+\varepsilon, \frac{\mathrm{d}\pi_\theta}{\mathrm{d}\pi_{\theta_{\text{old}}}}(a|s)\right) \frac{\delta V_\tau^{\pi_{\theta_{\text{old}}}}}{\delta \pi}(s,a) \Bigg] \Bigg).$$

Proximal policy optimization algorithms
J Schulman, F Wolski, P Dhariwal, A Radford… - arXiv p
… call **proximal policy optimization** (**PPO**), have some
Our experiments test **PPO** on a collection of benchmark
☆ Save 99 Cite Cited by 25072 Related articles

DeepSeek-V3 Technical Report

DeepSeek-AI

research@deepseek.com

$$\frac{1}{G}\sum_{i=1}^{G}\left(\min\left(\frac{\pi_\theta(o_i|q)}{\pi_{\theta_{old}}(o_i|q)}A_i, \text{clip}\left(\frac{\pi_\theta(o_i|q)}{\pi_{\theta_{old}}(o_i|q)}, 1-\varepsilon, 1+\varepsilon\right)A_i\right)\right)$$

TRPO

- Switch to maximizing in line with [Lascu et al., 2025].
- $A_\pi = Q_\pi - V_\pi = \frac{\delta V^\pi}{\delta \pi}$, (objective has $\tau = 0$ i.e. no KL term).

Theorem 14 ([Achiam et al., 2017])

For all $\rho \in \mathcal{P}(S)$ and any policies $\pi', \pi \in \mathcal{P}(A)$,

$$V^{\pi'}(\rho) - V^{\pi}(\rho) \geq \frac{1}{1-\gamma} \int_S \int_A \frac{\mathrm{d}\pi'}{\mathrm{d}\pi} A_\pi(s,a)\pi(\mathrm{d}a|s) d_\rho^\pi(\mathrm{d}s)$$
$$- \frac{4\gamma \|r\|_{B_b(S\times A)}}{(1-\gamma)^3} \int_S \mathrm{TV}(\pi'(\cdot|s), \pi(\cdot|s)) d_\rho^\pi(\mathrm{d}s) =: J(\pi').$$

N.B. $J(\pi) = 0$ so $\max_{\pi'} J(\pi') \geq 0$.

So

$$\pi' \in \operatorname*{argmax}_{\pi'} J(\pi')$$

guarantees improvement: $V^{\pi'} - V^\pi = \max_{\pi'} J(\pi') \geq 0$.

N.B.

$$\int_S \mathrm{TV}(\pi'(\cdot|s), \pi(\cdot|s)) d_\rho^\pi(\mathrm{d}s) = \frac{1}{2} \int_S \int_A \left| \frac{\mathrm{d}\pi'}{\mathrm{d}\pi}(a|s) - 1 \right| \pi(\mathrm{d}a|s) d_\rho^\pi(\mathrm{d}s),$$

we can write surrogate loss as

$$V^{\pi'}(\rho) - V^\pi(\rho) \geq \frac{1}{1-\gamma} \int_S \int_A \frac{\mathrm{d}\pi'}{\mathrm{d}\pi} A_\pi(s,a) \pi(\mathrm{d}a|s) d_\rho^\pi(\mathrm{d}s)$$
$$- \frac{2\gamma \|r\|_{B_b(S\times A)}}{(1-\gamma)^3} \int_S \int_A \left| \frac{\mathrm{d}\pi'}{\mathrm{d}\pi}(a|s) - 1 \right| \pi(\mathrm{d}a|s) = J(\pi').$$

This is a motivation for PPO's constraining the probability ratio $\left| \frac{\mathrm{d}\pi'}{\mathrm{d}\pi}(a|s) - 1 \right| \leq \varepsilon$, where $\varepsilon > 0$ is a hyperparameter.

PPO

$$J_{\mathrm{PPO}}(\pi') := \int_S \int_A \min \left\{ \frac{\mathrm{d}\pi'}{\mathrm{d}\pi}(a|s) A_\pi(s,a), \mathrm{clip}_{1-\varepsilon}^{1+\varepsilon} \left(\frac{\mathrm{d}\pi'}{\mathrm{d}\pi}(a|s) \right) A_\pi(s,a) \right\} \pi(\mathrm{d}a|s) d_\rho^\pi(\mathrm{d}s).$$

The challenge of TV geometry

$$J(\pi') = \frac{1}{1-\gamma}\int_S\int_A \tfrac{\mathrm{d}\pi'}{\mathrm{d}\pi} A_\pi(s,a)\pi(\mathrm{d}a|s)d_\rho^\pi(\mathrm{d}s) - \tfrac{2\gamma\|r\|_{B_b(S\times A)}}{(1-\gamma)^3}\int_S\int_A \left|\tfrac{\mathrm{d}\pi'}{\mathrm{d}\pi}(a|s)-1\right|\pi(\mathrm{d}a|s).$$

Seen how to get convergence if penalty satisfies 3-point property (Bregman proximal inequality).

Another approach is to go via first order optimality condition. But ...

$$2\frac{\delta\,\mathrm{TV}(\pi',\pi)(s)}{\delta\pi'}(a) = \frac{\delta}{\delta\pi'}\left[\int_S\int_A \left|\tfrac{\mathrm{d}\pi'}{\mathrm{d}\pi}(a|s)-1\right|\pi(\mathrm{d}a|s)\right] = \mathrm{sign}\left(\tfrac{\mathrm{d}\pi'}{\mathrm{d}\pi}(a|s)-1\right)\tfrac{\mathrm{d}\pi}{\mathrm{d}\lambda}(a|s).$$

Therefore, the first-order condition for the update rule gives the scheme

$$A_\pi(s,a) - \tfrac{1}{\tau}\,\mathrm{sign}\left(\tfrac{\mathrm{d}\pi'}{\mathrm{d}\pi}(a|s)-1\right)\tfrac{\mathrm{d}\pi}{\mathrm{d}\lambda}(a|s) = \text{const in a}.$$

- Two solutions for $\frac{\mathrm{d}\pi'}{\mathrm{d}\pi}$ depending on whether $\frac{\mathrm{d}\pi'}{\mathrm{d}\pi}(a|s) - 1 \geq 0$ or not.

Improved lower bound

Theorem 15 (Lower bound on performance difference [Lascu et al., 2025])

For any $\pi', \pi \in \mathcal{P}_\lambda(A|S)$, it holds that

$$
\begin{aligned}
V^{\pi'}(\rho) - V^{\pi}(\rho) \geq \frac{1}{1-\gamma} \int_S \int_A \frac{\mathrm{d}\pi'}{\mathrm{d}\pi}(a|s) A_\pi(s,a) \pi(\mathrm{d}a|s) d_\rho^\pi(\mathrm{d}s) \\
- \frac{8\gamma \|r\|_{B_b(S\times A)}}{(1-\gamma)^3} \int_S \mathrm{TV}^2(\pi'(\cdot|s), \pi(\cdot|s)) d_\rho^\pi(\mathrm{d}s).
\end{aligned}
$$

Towards Fisher-Rao PPO

Fix $\lambda \in \mathcal{M}_+(A)$. Then the squared Fisher-Rao (FR) distance between measures μ and μ' is defined by

$$\mathrm{FR}^2(\mu,\mu') = 4\int_A \left|\sqrt{\frac{\mathrm{d}\mu}{\mathrm{d}\lambda}} - \sqrt{\frac{\mathrm{d}\mu'}{\mathrm{d}\lambda}}\right|^2 \mathrm{d}\lambda\,,$$

if $\mu, \mu' \ll \lambda$, and ∞ otherwise.

Cauchy-Schwarz inequality gives

$$\int_S \mathrm{TV}^2(\pi'(\cdot|s),\pi(\cdot|s))d_\rho^\pi(\mathrm{d}s) \leq \frac{1}{16}\int_S \mathrm{FR}^2(\pi'(\cdot|s)^2,\pi(\cdot|s)^2)d_\rho^\pi(\mathrm{d}s),$$

Corollary 16

For any $\pi',\pi \in \mathcal{P}_\lambda(A|S)$, it holds that

$$\begin{aligned} V^{\pi'}(\rho) - V^\pi(\rho) \geq{}& \frac{1}{1-\gamma}\int_S\int_A \frac{\mathrm{d}\pi'}{\mathrm{d}\pi}(a|s)A_\pi(s,a)\pi(\mathrm{d}a|s)d_\rho^\pi(\mathrm{d}s) \\ &- \frac{\gamma\|r\|_{B_b(S\times A)}}{2(1-\gamma)^3}\int_S \mathrm{FR}^2(\pi'(\cdot|s)^2,\pi(\cdot|s)^2)d_\rho^\pi(\mathrm{d}s), \end{aligned}$$

where $\pi(\cdot|s)^2 := \left(\frac{\mathrm{d}\pi}{\mathrm{d}\lambda}\right)^2(\cdot|s)$, for all $s \in S$.

Consider

$$\pi^{n+1}(\cdot|s) = \operatorname*{argmax}_{m\in\mathcal{P}} \left[\int_A A_{\pi^n}(s,a) \frac{\mathrm{d}m}{\mathrm{d}\pi^n}(a|s)\pi^n(da|s) - \frac{1}{2\tau} \mathrm{FR}^2(m^2, \pi^n(\cdot|s)^2) \right].$$

Theorem 17 (Policy improvement [Lascu et al., 2025])

Let $V^n := V^{\pi^n}$. If $\frac{1}{\tau} \geq \frac{\|r\|_{B_b(S\times A)}}{(1-\gamma)^2}$, then for any $\rho \in \mathcal{P}(S)$ we have

$$V^{n+1}(\rho) \geq V^n(\rho) \text{ for all } n > 0\,.$$

Theorem 18 (Sublinear convergence [Lascu et al., 2025])

Let $V^n := V^{\pi^n}$. For all $\rho \in \mathcal{P}(S)$, $\frac{1}{\tau} \geq \frac{\gamma\|r\|_{B_b(S\times A)}}{(1-\gamma)^2}$, for any $\pi \in \mathcal{P}(A|S)$ and $n \in \mathbb{N}_0$,

$$(V^\pi - V^n)(\rho) \leq \frac{1}{n(1-\gamma)} \left(\frac{1}{\tau} \int_S \mathrm{FR}^2((\pi(\cdot|s)^2, \pi_0(\cdot|s)^2) d^\pi_\rho(ds) + \int_S (V^0 - V^\pi)(s) d^\pi_\rho(ds) \right).$$

Summary

General

- MDPs can be formulated in very general (Polish) state and action spaces.
- Relaxed MDP with entropy regularization has good mathematical properties (existence and uniqueness of optimal policy) under minimal assumptions (boundedness and measurability).

Policy gradient

- We have PG theorem but convergence analysis is hard due to non-convexity (best known results need non-local Lojasiewicz are in finite state & action spaces or structural assumptions - LQR).
- Mirror descent (NPG) is much more ameaneable to convergence analysis (with direct parametrization of linear basis functions for log gensities), key ingredients
  - Three point lemma
  - L-smoothness (from perf. diff. and Lan's trick)
  - The "replacement-for-convexity-property" (that perf. diff provides)
- Convergence of popular PG schemes e.g. PPO still not understood.
- Replacing linear parametrizations (basis functions, NTK, mean-field) with deep networks is . . . not understood even for supervised learning.

Thank you!

References I

[Achiam et al., 2017] Achiam, J., Held, D., Tamar, A., and Abbeel, P. (2017). Constrained policy optimization. In *International conference on machine learning*, pages 22–31. PMLR.

[Agarwal et al., 2019] Agarwal, A., Kakade, S. M., Lee, J. D., and Mahajan, G. (2019). Optimality and approximation with policy gradient methods in Markov decision processes. *arXiv preprint arXiv:1908.00261*.

[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.

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

[Dupuis and Ellis, 1997] Dupuis, P. and Ellis, R. S. (1997). *A weak convergence approach to the theory of large deviations*. John Wiley & Sons, Inc., New York.

[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.

[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.

[Hernández-Lerma and Lasserre, 2012] Hernández-Lerma, O. and Lasserre, J. B. (2012). *Discrete-time Markov control processes: basic optimality criteria*, volume 30. Springer Science & Business Media.

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

References III

[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.

[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.

[Lascu et al., 2025] Lascu, R.-A., Šiška, D., and Szpruch, Ł. (2025). PPO in the Fisher–Rao geometry. *arXiv preprint arXiv:2506.03757*.

References IV

[Lemaréchal, 2012] Lemaréchal, C. (2012). Cauchy and the gradient method. *Doc Math Extra*, 251(254):10.

[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.

[Schulman et al., 2015] Schulman, J., Moritz, P., Levine, S., Jordan, M., and Abbeel, P. (2015). High-dimensional continuous control using generalized advantage estimation. *arXiv preprint arXiv:1506.02438*.

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

[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.

Q-learning

DPP equation for Q_τ^*:

$$Q_\tau^*(s,a) = c(s,a) + \gamma \int_S \inf_{m \in \mathcal{P}(A)} \left(Q_\tau^*(s',a') + \tau \tfrac{\mathrm{d}m}{\mathrm{d}\mu}(a') \right) m(da') \Big) P(ds'|s,a)\,.$$

Re-write:

$$0 = \mathbb{E}^{s' \sim P(\cdot|s,a)}_{a' \sim \pi_\tau^*(\cdot|s')} \left[c(s,a) + \gamma \Big(Q_\tau^*(s',a') + \tau \tfrac{\mathrm{d}\pi_\tau^*}{\mathrm{d}\mu}(a'|s') \Big) \Big) - Q_\tau^*(s,a) \right].$$

Q-learning:

$$Q_{k+1}(s_k,a_k) = Q_k(s_k,a_k) + \delta_k \left[c(s_k,a_k) + \gamma \Big(Q_k(s_{k+1},a_{k+1}) + \tau \tfrac{\mathrm{d}\pi_k}{\mathrm{d}\mu}(a_{k+1}|s_{k+1}) \Big) - Q_k(s_k,a_k) \right],$$

where $s_{k+1} \sim P(\cdot|s_k,a_k)$, $\pi_k(da'|s_k) \propto \exp(-\tau^{-1} Q_k(s_k,a'))\mu(da')$, $a_{k+1} \sim \pi_k(\cdot|s_k)$.

Convergence: like value iteration + stochastic approximation (Robbins–Monro).