Fisher–Rao gradient flow for entropy regularised MDPs in Polish spaces

David Šiška[1]

Joint work with B. Kerimkulov,[2] J.-M. Leahy,[3] and L. Szpruch,[14] Y Zhang[3]

Séminaire FDD-FiME, Institut Henri Poincaré

17th November 2023

[1]University of Edinburgh
[2]Natwest, formerly University of Edinburgh
[3]Imperial College London
[4]The Alan Turing Institute

Overview

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 reward function, and γ

Minimise over Markov policies $S \ni s \mapsto \pi(da|s) \in \mathcal{P}(A)$ the objective:

$$V_\tau^\pi(s) = \mathbb{E}_s^\pi \left[\sum_{n=0}^{\infty} \gamma^n \Big(c(s_n, a_n) + \tau \, \mathrm{KL}(\pi(\cdot|s_n)|\mu) \Big) \right] \in \mathbb{R} \cup \{\infty\} \,. \tag{1}$$

Policy gradient (RL) / Fisher–Rao flow on the space of signed kernels:

$$\partial_t \pi_t(da|\cdot) = -\frac{\delta V_\tau^{\pi_t}}{\delta \pi}(\cdot, a) \, \pi_t(da|\cdot) \,, \quad t > 0 \,. \tag{2}$$

Mirror Descent on the space of bounded functions:

$$\partial_t Z_t(s, a) = -\frac{\delta V_\tau^{\pi_t}}{\delta \pi}(s, a) \,, \quad \pi_t(da|s) = \boldsymbol{\pi}(Z_t)(s, a) \propto e^{Z_t(s,a)} \,, \quad t > 0 \,. \tag{3}$$

Aim: Show (3) is well posed, $\pi_t = \boldsymbol{\pi}(Z_t)$ solves (2), $V^{\pi_t} \to V^*$ exponentially.

Outline

- Key properties of Kullback–Leibler divergence
- Entropy-regularised MDP for general state-action spaces
- Gradient flow derivation
- Short overview of existing results
- Well posedness and convergence
- Stability

Kullback–Leibler divergence aka relative entropy

Relative entropy - definition and basics

Recall that if A is Polish $\nu, \mu \in \mathcal{P}(A)$ 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.

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

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

Note that

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

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

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

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

Relative entropy - variational formula

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

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

Convexity:

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

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

Entropy-regularised MDPs on Polish spaces

Entropy-regularised MDP

- S, A Polish, $P \in \mathcal{P}(S|S \times A)$, $c \in B_b(S \times A)$ and $\gamma \in [0, 1)$.
- $H_n := (S \times A)^n \times S$ is the space of admissible histories.
- $\Pi = \{\pi = \{\pi_n\}_{n \in \mathbb{N}_0} : \pi_n \in \mathcal{P}(A|H_n)\}$ the set of randomised policies.
- $(\Omega := (S \times A)^{\mathbb{N}_0}, \mathcal{F})$ the canonical sample space, $\mathcal{F} = \mathcal{B}(\Omega)$.

By [Bertsekas and Shreve, 2004, Proposition 7.28], for any given initial distribution $\rho \in \mathcal{P}(S)$ and policy $\pi \in \Pi$, there exists a unique product probability measure $\mathbb{P}^\pi_\rho$ on $(\Omega, \mathcal{F})$ with expectation denoted $\mathbb{E}^\pi_\rho$ such that for all $n \in \mathbb{N}_0$, $B \in \mathcal{B}(S)$ and $C \in \mathcal{B}(A)$, $\mathbb{P}^\pi_\rho(s_0 \in B) = \rho(B)$ and

$$\mathbb{P}^\pi_\rho(a_n \in C|h_n) = \pi_n(C|h_n), \quad \mathbb{P}^\pi_\rho(s_{n+1} \in B|h_n, a_n) = P(B|s_n, a_n)\,, \tag{4}$$

where $h_n = (s_0, a_0, \ldots, s_{n-1}, a_{n-1}, s_n) \in H_n$.

- If π is a randomised Markov policy (i.e., $\pi_n \in \mathcal{P}(A|S)$ for all $n \in \mathbb{N}_0$), then $\{s_n\}_{n \in \mathbb{N}_0}$ is a Markov process with kernel $\{P_{\pi,n}\}_{n \in \mathbb{N}_0} \in \mathcal{P}(S|S)$ given by

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

Entropy-regularised MDP

- Fix $\tau > 0$ and $\mu \in \mathcal{P}(A)$.

Regularised value function:

$$V_\tau^\pi(s) = \mathbb{E}_s^\pi \left[\sum_{n=0}^{\infty} \gamma^n \Big(c(s_n, a_n) + \tau \, \mathrm{KL}(\pi_n(\cdot|h_n)|\mu) \Big) \right] \in \mathbb{R} \cup \{\infty\}, \qquad (5)$$

where

$$\mathrm{KL}(\mu'|\mu) = \int_A \ln \frac{\mathrm{d}\mu'}{\mathrm{d}\mu}(a) \mu'(da)$$

if $\mu' \ll \mu$, and ∞ otherwise for $\mu' \in \mathcal{P}(A)$.

Optimal value function $V_\tau^* : S \to \mathbb{R} \cup \{\infty\}$ is

$$V_\tau^*(s) = \inf_{\pi \in \Pi} V_\tau^\pi(s), \quad \forall s \in S \, . \qquad (6)$$

Let

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

Lemma 1 (Policy Bellman equation)

Let $\pi \in \Pi_\mu$, $\tau > 0$. The value function V_τ^π is the unique bounded solution of

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

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

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

and show that it's a contraction on $B_b(S)$.

Theorem 2 (Dynamic programming principle)

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

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

Consequently, for all $s \in S$,

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

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

$$\pi_\tau^*(da|s) = \exp\left(-(Q_\tau^*(s,a) - V_\tau^*(s))/\tau\right) \mu(da)\,, \quad \forall s \in S. \tag{7}$$

Proof hint. The usual, no need for measurable selection as the infimum is attained [Dupuis and Ellis, 1997, Proposition 1.4.2].

Entropy-regularised MDP - Markov policies

DPP and the Bellman eq. imply we can restrict our policy class to

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

identified with $\{\boldsymbol{\pi}(f) \mid f \in B_b(S \times A)\} \subset \mathcal{P}_\mu(A|S)$, where $\boldsymbol{\pi} : B_b(S \times A) \to \mathcal{P}_\mu(A|S)$ is defined by

$$\boldsymbol{\pi}(f)(da|s) = \frac{e^{f(s,a)}}{\int_A e^{f(s,a')}\mu(da')}\mu(da), \quad \forall f \in B_b(S \times A)\,. \tag{8}$$

Entropy-regularised MDP - some useful objects

Q-function

$$Q_\tau^\pi(s,a) = c(s,a) + \gamma \int_S V_\tau^\pi(s') P(ds'|s,a)\,. \tag{9}$$

N.B. recall

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

so that

$$V_\tau^\pi(s) = \int_A \left(Q_\tau^\pi(s,a) + \tau \ln \frac{\mathrm{d}\pi}{\mathrm{d}\mu}(a|s) \right) \pi(da|s)\,. \tag{10}$$

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

$$d^\pi(ds'|s) = (1-\gamma) \sum_{n=0}^{\infty} \gamma^n P_\pi^n(ds'|s)\,. \tag{11}$$

Let

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

Gradient flow

Performance difference

Lemma 3 (Performance difference)

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

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

Lemma 4 (Feynmann–Kac formula)

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

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

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

Performance difference - proof
Proof of Lemma 3.

By (10), for all $s \in S$,

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

where the last equality used def. of Q fn. (9) and the fact that

$$
\int_A \ln \frac{\mathrm{d}\pi}{\mathrm{d}\mu}(a|s) - \ln \frac{\mathrm{d}\pi'}{\mathrm{d}\mu}(a|s)\pi(da|s) = \int_A \ln \frac{\frac{\mathrm{d}\pi}{\mathrm{d}\mu}}{\frac{\mathrm{d}\pi'}{\mathrm{d}\mu}}(a|s)\pi(da|s) = \mathrm{KL}(\pi(\cdot|s)|\pi'(\cdot|s)).
$$

Conclude by Fubini's theorem and Lemma 4. ■

From performance difference to a derivative

We have that for all $\rho \in \mathcal{P}(S)$ and $\pi, \pi' \in \Pi_\mu$,

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

Let $\pi^\varepsilon = \pi + \varepsilon(\pi' - \pi)$. Then $\frac{1}{\varepsilon}(V_\tau^{\pi^\varepsilon}(\rho) - V_\tau^\pi(\rho))$

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

Linear functional derivative / first variation

For $\pi, \pi' \in \Pi_\mu$ have

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

For any $\nu \in \mathcal{P}(S)$ let

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

First variation of $V_\tau^\cdot(\rho)$ relative to $\langle \cdot, \cdot \rangle_\nu$ must satisfy

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

and $\langle \frac{\delta V_\tau^\pi(\rho)}{\delta \pi}\big|_\nu, \pi \rangle_\nu = 0$.

It is is given by

$$\frac{\delta V_\tau^\pi(\rho)}{\delta \pi}\bigg|_\nu (s,a) = \left(Q_\tau^\pi(s,a) + \tau \ln \frac{\mathrm{d}\pi}{\mathrm{d}\mu}(s,a) - V_\tau^\pi(s)\right) \frac{d d_\rho^\pi}{d\nu}(s)\,. \tag{12}$$

Dissipation of energy

Consider:

$$\partial_t \pi_t(da|s) = -\frac{\delta V_\tau^{\pi_t}}{\delta \pi}(s,a)\, \pi_t(da|s)\,, \quad t > 0\,, \tag{13}$$

where

$$\frac{\delta V_\tau^\pi}{\delta \pi} = \frac{\delta V_\tau^\pi(\rho)}{\delta \pi}\bigg|_{d_\rho^\pi} = Q_\tau^\pi + \tau \ln \frac{\mathrm{d}\pi}{\mathrm{d}\mu} - V_\tau^\pi\,,$$

is the first variation rel. to state-dependent pairing $\langle \cdot, \cdot \rangle_{d_\rho^\pi}$.

Chain rule (formally):

$$\frac{d}{dt} V_\tau^{\pi_t}(\rho) = \left\langle \frac{\delta V_\tau^{\pi_t}}{\delta \pi}, \partial_t \pi_t \right\rangle_{d_\rho^{\pi_t}} = -\left\langle \left| \frac{\delta V_\tau^{\pi_t}}{\delta \pi} \right|^2, \pi_t \right\rangle_{d_\rho^{\pi_t}} \leq 0\,.$$

Policy mirror descent / Proximal policy optimization

Let $\lambda > 0$, $\pi_0 \in \Pi_\mu$ and for $n \in \mathbb{N}_0$,

$$\pi^{n+1} \in \operatorname*{arg\,min}_{\pi \in \mathcal{P}(A|S)} \left[\left\langle \frac{\delta V_\tau^{\pi^n}}{\delta \pi}, \pi - \pi^n \right\rangle_{d_\rho^{\pi^n}} + \frac{\lambda}{1-\gamma} \int_S \mathrm{KL}(\pi(\cdot|s)|\pi^n(\cdot|s)) d_\rho^{\pi^n}(ds) \right]$$

$$\in \operatorname*{arg\,min}_{\pi \in \mathcal{P}(A|S)} \int_S \left[\int_A \frac{\delta V_\tau^{\pi^n}}{\delta \pi}(da|s)(\pi - \pi^n)(da|s) + \lambda \,\mathrm{KL}(\pi(\cdot|s)|\pi^n(\cdot|s)) \right] \frac{1}{1-\gamma} d_\rho^{\pi^n}(ds)\,.$$

Spot the "adaptive Bregman divergence" associated with state int. neg. ent.

$$h_\nu(\pi) = \int_S \mathrm{KL}(m(s)|\mu)\nu(ds)\,. \tag{14}$$

Minimum is achieved by the pointwise in S optimization:

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

Mirror descent algo [Lan, 2022] (i.e., Algorithm 1) and [Xiao, 2022], natural policy gradient update of [Kakade, 2001].

Policy mirror descent as $\lambda \to 0$

Minimum at each step is

$$\frac{\mathrm{d}\pi^{n+1}}{\mathrm{d}\pi^n}(s,a) = \frac{\exp\left(-\frac{1}{\lambda}\frac{\delta V_\tau^{\pi^n}}{\delta \pi}(s,a)\right)}{\int_A \exp\left(-\frac{1}{\lambda}\frac{\delta V_\tau^{\pi^n}}{\delta \pi}(s,a')\right)\pi^n(da'|s)}\,.$$

Rearranging:

$$\lambda\left(\ln\frac{\mathrm{d}\pi^{n+1}}{\mathrm{d}\mu}(s,a) - \ln\frac{\mathrm{d}\pi^n}{\mathrm{d}\mu}(s,a)\right) = -\frac{\delta V_\tau^{\pi^n}}{\delta\pi}(s,a) - \lambda \ln\int_A e^{-\frac{1}{\lambda}\frac{\delta V_\tau^{\pi^n}}{\delta\pi}(s,a')}\pi^n(da'|s)\,.$$

Interpolating in the time variable and letting $\lambda \to \infty$, we obtain

$$\partial_t \ln\frac{\mathrm{d}\pi_t}{\mathrm{d}\mu}(s,a) = -\left(\frac{\delta V_\tau^{\pi_t}}{\delta\pi}(s,a) - \int_A \frac{\delta V_\tau^{\pi_t}}{\delta\pi}(s,a')\pi_t(da')\right) = -\frac{\delta V_\tau^{\pi_t}}{\delta\pi}(s,a)\,,$$

where in the second equality, we used the fact that $\int_A \frac{\delta V_\tau^{\pi_t}}{\delta\pi}(s,a')\pi_t(da') = 0$ (see Lemma 1).

Dual and primal (Fisher–Rao) flows

Defining $Z_t = \ln \frac{\mathrm{d}\pi_t}{\mathrm{d}\mu}$, we find

$$\partial_t Z_t(s,a) = -\frac{\delta V_\tau^{\pi_t}}{\delta \pi}(s,a)\,, \quad \pi_t(da|s) = \boldsymbol{\pi}(Z_t)(s,a)\ , \ \ t>0\,, \tag{15}$$

where the map $\boldsymbol{\pi} : B_b(S\times A) \to \Pi_\mu$ is defined by

$$\boldsymbol{\pi}(Z)(da|s) := \frac{e^{Z(s,a)}}{\int_A e^{Z(s,a')}\mu(da')}\,\mu(da)\,. \tag{16}$$

Chain rule:

$$\frac{1}{\frac{\mathrm{d}\pi_t}{\mathrm{d}\mu}(s,a)}\partial_t \frac{\mathrm{d}\pi_t}{\mathrm{d}\mu}(s,a) = -\frac{\delta V_\tau^{\pi_t}}{\delta \pi}(s,a)\,, \ \ t>0\,,$$

which yields (13) i.e.

$$\partial_t \pi_t(da|s) = -\frac{\delta V_\tau^{\pi_t}}{\delta \pi}(s,a)\,\pi_t(da|s)\,, \quad t>0\,. \tag{17}$$

Log-linear parametrisation

Hilbert space $(\mathbb{H}, \|\cdot\|_{\mathbb{H}})$ with the inner product $\langle\cdot,\cdot\rangle_{\mathbb{H}}$.

- $\mathbb{H} = \mathbb{R}^N$ or $\mathbb{H} = \ell^2$
- $g \in B_b(S \times A; \mathbb{H})$ be a fixed feature basis.

$$\pi_\theta = \boldsymbol{\pi}(\langle\theta, g(\cdot)\rangle_{\mathbb{H}}) = \boldsymbol{\pi}(\langle\theta, g_{\pi_\theta}(\cdot)\rangle_{\mathbb{H}}) \tag{18}$$

where

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

Chain rule:

$$\nabla_\theta V_\tau^{\pi_\theta}(\rho) = \frac{1}{1-\gamma}\int_S\int_A \left(Q_\tau^{\pi_\theta}(s,a) + \tau \ln\frac{\mathrm{d}\pi_\theta}{\mathrm{d}\mu}(a|s)\right) g_{\pi_\theta}(s,a)\pi_\theta(da|s)d_\rho^{\pi_\theta}(ds)\,. \tag{19}$$

Natural policy gradient (NPG):

$$\partial_t\theta_t = -\mathscr{F}(\theta_t)^{-1}\nabla_\theta V_\tau^{\pi_{\theta_t}}(\rho)\,, \quad t > 0\,. \tag{20}$$

Fisher information operator defined by

$$\mathscr{F}(\theta) := \int_S\int_A \big(g_{\pi_\theta}(s,a)\otimes g_{\pi_\theta}(s,a)\big)\pi_\theta(da|s)d_\rho^{\pi_\theta}(ds)\,.$$

Log-linear parametrisation: NPG

Compatible function approximation to approximate a centered Q-function.

For any $\theta \in \mathbb{H}$ the quadratic loss $\ell^{\pi_\theta} : \mathbb{H} \to \mathbb{R}$ is

$$\ell^{\pi_\theta}(w) = \frac{1}{2}\int_S \int_A |A_\tau^{\pi_\theta}(s,a) - \langle w, g_{\pi_\theta}(s,a)\rangle_{\mathbb{H}}|^2 \pi_\theta(da|s) d_\rho^{\pi_\theta}(ds)\,, \quad \text{where} \quad A_\tau^{\pi_\theta}(s,a) := Q_\tau^{\pi_\theta}(s,a) - \int_A Q_\tau^{\pi_\theta}(s,a')\pi_\theta(da'|s)\,. \tag{21}$$

If $w^*(\theta) \in \arg\min_{w\in\mathbb{H}} \ell^{\pi_\theta}(w)$, then the first-order condition of (21) produces

$$\mathscr{F}(\theta) w^*(\theta) = \int_S \int_A Q_\tau^{\pi_\theta}(s,a) g_{\pi_\theta}(s,a) \pi_\theta(da|s) d_\rho^{\pi_\theta}(ds)\,. \tag{22}$$

Substituting (22) into (19) and using $\pi_\theta = \langle \theta, g(\cdot)\rangle_{\mathbb{H}}$, we obtain

$$\nabla_\theta V^{\pi_\theta}(\rho) = \frac{1}{1-\gamma}\mathscr{F}(\theta)\left(w^*(\theta) + \tau\theta\right)\,.$$

Hence the NPG (20) is

$$\partial_t \theta_t = -(w^*(\theta_t) + \tau\theta_t)\,, \quad t > 0\,.$$

Log-linear parametrisation: NPG to Fisher–Rao

If $(\theta_t)_{t\geq 0}$ satisfies (25) i.e.

$$\partial_t \theta_t = -(w^*(\theta_t) + \tau\theta_t)\,, \quad t > 0$$

then the chain rule shows that $(\pi_{\theta_t})_{t\geq 0}$ satisfies

$$\begin{aligned}
&\partial_t \pi_t(da|s) \\
&= -\left(Q_t(s,a) + \tau \ln \frac{\mathrm{d}\pi_t}{\mathrm{d}\mu}(s,a) - \int_A \left(Q_t(s,a') + \tau \ln \frac{\mathrm{d}\pi_t}{\mathrm{d}\mu}(a'|s)\right)\pi_t(da'|s)\right)\pi_t(da|s) \\
&= -\left(\frac{\delta V_\tau^{\pi_t}}{\delta \pi}(s,a) + \mathcal{E}_t(s,a) - \int_A \mathcal{E}_t(s,a')\pi_t(da'|s)\right)\pi_t(da|s)\,,
\end{aligned}$$

where

$$\begin{aligned}
Q_t(s,a) &= \langle w^*(\theta_t), g(s,a)\rangle_{\mathbb{H}}, \\
\mathcal{E}_t(s,a) &= \langle w^*(\theta_t), g(s,a)\rangle_{\mathbb{H}} - A_\tau^{\pi_{\theta_t}}(s,a)\,.
\end{aligned}$$

Gradient flows: summary

We have abstract but exact:

i) Primal:

$$\partial_t \pi_t(da|s) = -\frac{\delta V_\tau^{\pi_t}}{\delta \pi}(s,a)\,\pi_t(da|s)\,, \quad t>0\,. \tag{23}$$

ii) Dual:

$$\partial_t Z_t(s,a) = -\frac{\delta V_\tau^{\boldsymbol{\pi}(Z_t)}}{\delta \pi}(s,a)\,, \quad t>0\,. \tag{24}$$

Connected via the mirror map

$$\pi_t(da|s) = \boldsymbol{\pi}(Z_t)(s,a) = \frac{e^{Z(s,a)}}{\int_A e^{Z(s,a')}\mu(da')}\,\mu(da)\,.$$

Practical but approximate:

i) Dual, clearly natural policy gradient:

$$\partial_t \theta_t = -\mathscr{F}(\theta_t)^{-1}\nabla_\theta V_\tau^{\pi_{\theta_t}}(\rho)\,, \quad t>0\,.$$

ii) Dual:

$$\partial_t \theta_t = -(w^*(\theta_t) + \tau\theta_t)\,, \quad t>0\,. \tag{25}$$

Existing results

Key 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].

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

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

NPG / Mirror descent in discrete state and action spaces

- [Cen et al., 2022], entropy regularised, show linear convergence for disc. time. versions of (13) and (15).
- [Cayci et al., 2021] same setting i.e natural policy gradient but for log-linear policies i.e. discrete-time analog of (25).
- [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.
- [Lan, 2023] linear convergence of policy mirror descent with arbitrary convex regularisers and [Zhan et al., 2023] convergence rates independent of action space dimension.
- [Yuan et al., 2022] and [Xiao, 2022] log-linear parameterised policies with compatible Q-function approximations, stability estimates based on L^∞ and L^2 approx errors respectively.

Except [Zhan et al., 2023]) all depend explicitly on the action space cardinality.

NPG / Mirror descent in continuous state and action spaces

- Discrete time LQR: Polyak–Łojasiewicz (PL) / gradient dominance condition is used to show Policy gradient has linear convergence [Fazel et al., 2018, Bu et al., 2019, Hu et al., 2023].
- General MDPs with finite-dimensional parameterised policies *assuming uniform PL condition* [Bhandari and Russo, 2019, Zhang et al., 2022, Fatkhullin et al., 2023] obtain corresponding rate.

In discrete state-action setting best PL result is non-uniform [Mei et al., 2021].

- Mean-field softmax policies with Wasserstein gradient flow:
  - [Agazzi and Lu, 2020] - if flow converges then to optimal.
  - [Leahy et al., 2022] - if you regularise enough, flow converges and error estimates to unreg.
  - [Zhang et al., 2021] - two timescale actor/critic NPG combined with Wasserstein, error of $O(1/T)$ but "proof" is formal only.

Well posedness and convergence

Existence of primal equiv. to existence of dual

Primal

$$\partial_t \pi_t(da|s) = -\left(Q_\tau^{\pi_t}(s,a) + \tau \ln \frac{\mathrm{d}\pi_t}{\mathrm{d}\mu}(a|s) - V_\tau^{\pi_t}(s) \right) \pi_t(da|s)\,. \tag{26}$$

Dual

$$\partial_t Z_t(s,a) = -\left(Q_\tau^{\pi_t}(s,a) + \tau Z_t(s,a) - V_\tau^{\pi_t}(s)\right), \quad \pi_t(da|s) = \boldsymbol{\pi}(Z_t)(da|s)\,. \tag{27}$$

Lemma 5

Let $T > 0$.

(1) *Let $Z \in C^1((0,T); B_b(S \times A))$ be such that (27) holds and define $\pi_t = \boldsymbol{\pi}(Z_t)$ for all $t \in (0,T)$. Then $\pi \in C^1((0,T); \Pi_\mu)$[5] satisfies (26).*

(2) *Let $\pi \in C^1((0,T); \Pi_\mu)$ be such that (26) holds and define $Z_t = \ln \frac{\mathrm{d}\pi_t}{\mathrm{d}\mu}$ for all $t \in (0,T)$. Then $Z \in C^1((0,T); B_b(S \times A))$ satisfies (27).*

[5] For any $I \subset [0,\infty]$, we write $\pi \in C^1(I; \Pi_\mu)$ if $\pi \in C^1(I; b\mathcal{M}(A|S))$ and $\pi_t \in \Pi_\mu$ for all $t \in I$.

Existence of and dissipation of energy in dual flow

Proposition 1

Let $Z_0 \in B_b(S \times A)$ and $T > 0$. If $Z \in C^1([0,T); B_b(S \times A))$ satisfies (27) and $\pi_t = \boldsymbol{\pi}(Z_t)$ for all $t \in [0,T)$, then $t \mapsto V_\tau^{\pi t}$ is differentiable, and $\partial_t V_\tau^{\pi t}(s) \leq 0$ for all $s \in S$ and $t \in [0,T)$.

Proof. The map on RHS of

$$\partial_t Z_t(s,a) = -\left(Q_\tau^{\pi t}(s,a) + \tau Z_t(s,a) - V_\tau^{\pi t}(s)\right), \quad \pi_t(da|s) = \boldsymbol{\pi}(Z_t)(da|s)\,. \tag{28}$$

is only local-Lipschitz; however $Z \mapsto V_\tau^{\boldsymbol{\pi}(Z)}(\rho)$ is a Lyapunov function.

i) Dissipation of energy $\partial_t V_\tau^{\boldsymbol{\pi}(Z_t)} \leq 0$ means $V_\tau^{\pi 0} \geq V_\tau^{\pi t} \geq V_\tau^*$ means

$$\sup_t \|V_\tau^{\pi t}\| \leq \max(\|V_\tau^{\pi 0}\|, \|V_\tau^*\|)\,,$$

$$\sup_t \|Q_\tau^{\pi t}\| \leq \|c\| + \gamma \max(\|V_\tau^{\pi 0}\|, \|V_\tau^*\|)\,.$$

ii) If $(Z_t)_{t\geq 0}$ solves (28) then $\forall t \geq 0$

$$\|Z_t\| \leq e^{-\tau t}\|Z_0\| + \tfrac{1-e^{-\tau t}}{\tau}\left(\|c\| + (1+\gamma)(\|V_\tau^{\pi 0}\|, \|V_\tau^*\|)\right)$$

iii) Show a Lipschitz continuous truncation of (28) has unique soln. and Z_t need never exceed the truncation. □

Theorem 6 (Linear convergence)

Let $Z \in C^1(\mathbb{R}_+; B_b(S \times A))$ satisfy (27). Then for all $\rho \in \mathcal{P}(S)$ and $t > 0$,

$$V_\tau^{\pi_t}(\rho) - V_\tau^{\pi_\tau^*}(\rho) \leq \frac{\tau}{(1-\gamma)(e^{\tau t} - 1)} \int_S \mathrm{KL}(\pi_\tau^*(\cdot|s)|\pi_0(\cdot|s)) d_\rho^{\pi_\tau^*}(ds)\,,$$

and

$$\int_S \|\pi_t(\cdot|s) - \pi_\tau^*(\cdot|s)\|_{\mathcal{M}(A)}^2 d_\rho^{\pi_\tau^*}(ds) \leq 2e^{-\tau t} \int_S \mathrm{KL}(\pi_\tau^*(\cdot|s)|\pi_0(\cdot|s)) d_\rho^{\pi_\tau^*}(ds)\,,$$

where $\pi_t = \boldsymbol{\pi}(Z_t)$ for all $t \geq 0$ and π_τ^* is the optimal policy defined in (7).

Proof. introduce the Bregman divergence $D_\nu : B_b(S \times A) \times B_b(S \times A) \to \mathbb{R}$ defined for all $f, g \in B_b(S \times A)$ by

$$D_\nu(f, g) = \int_S \left(\Phi(f)(s) - \Phi(g)(s) - \int_A (f(s,a) - g(s,a)) \boldsymbol{\pi}(g)(da|s) \right) \nu(ds)\,, \tag{29}$$

where $\boldsymbol{\pi} : B_b(S \times A) \to \mathcal{P}_\mu(A|S)$ is defined by (8) and $\Phi : B_b(S \times A) \to B_b(S)$ is defined for all $f \in B_b(S \times A)$ by

$$\Phi(f)(s) = \ln \left(\int_A e^{f(s,a)} \mu(da) \right), \quad s \in S\,. \tag{30}$$

Lemma 7

For all $\rho \in \mathcal{P}(S)$ and $f, g \in B_b(S \times A)$,

$$D_{d_\rho^{\boldsymbol{\pi}(g)}}(f, g) = \int_S \mathrm{KL}(\boldsymbol{\pi}(g)(\cdot|s)|\boldsymbol{\pi}(f)(\cdot|s)) d_\rho^{\boldsymbol{\pi}(g)}(ds)\,.$$

Proof of linear convergence

By the chain rule and the definition of Φ given in (30), for all $t > 0$,

$$
\begin{aligned}
\partial_t D_{d_\rho^{\pi^*_\tau}}(Z_t, Z^*) &= \partial_t \left(\int_S \left(\Phi(Z_t)(s) - \Phi(Z^*)(s) - \int_A (Z_t(s,a) - Z^*(s,a)) \pi^*_\tau(da|s) \right) d_\rho^{\pi^*_\tau}(ds) \right) \\
&= \partial_t \left(\int_S \ln \left(\int_A e^{Z_t(s,a)} \mu(da) \right) d_\rho^{\pi^*_\tau}(ds) \right) - \int_S \int_A \partial_t Z_t(s,a) \pi^*_\tau(da|s) d_\rho^{\pi^*_\tau}(ds) \\
&= \int_S \int_A \partial_t Z_t(s,a) \pi_t(da|s) d_\rho^{\pi^*_\tau}(ds) - \int_S \int_A \partial_t Z_t(s,a) \pi^*_\tau(da|s) d_\rho^{\pi^*_\tau}(ds) \\
&= \int_S \int_A \partial_t Z_t(s,a) (\pi_t(da|s) - \pi^*_\tau(da|s)) d_\rho^{\pi^*_\tau}(ds) \\
&= -\int_S \int_A (Q_\tau^{\pi_t}(s,a) + \tau Z_t(s,a) - V_\tau^{\pi_t}(s)) (\pi_t(da|s) - \pi^*_\tau(da|s)) d_\rho^{\pi^*_\tau}(ds) \\
&= -\int_S \int_A \left(Q_\tau^{\pi_t}(s,a) + \tau \ln \frac{\mathrm{d}\pi_t}{\mathrm{d}\mu}(a|s) \right) (\pi_t(da|s) - \pi^*_\tau(da|s)) d_\rho^{\pi^*_\tau}(ds) ,
\end{aligned}
\tag{31}
$$

where the second to last line used the fact that $Z \in C^1(\mathbb{R}_+; B_b(S \times A))$ satisfies (27), and the last line used the facts that $Z_t(s,a) = \ln \frac{\mathrm{d}\pi_t}{\mathrm{d}\mu}(a|s) + \ln \left(\int_A e^{Z_t(s,a')} \mu(da') \right)$ for all $(s,a) \in S \times A$ and

$$
\int_A g(s) (\pi_t(da|s) - \pi^*_\tau(da|s)) = g(s) (\pi_t(A|s) - \pi^*_\tau(A|s)) = 0
$$

for all $s \in S$ and $g \in B_b(S)$.

Proof of linear convergence

Note that by Lemma 3 (performance difference lemma), for all $t > 0$,

$$\begin{aligned}\int_S \int_A \left(Q_\tau^{\pi_t}(s,a) + \tau \ln \frac{\mathrm{d}\pi_t}{\mathrm{d}\mu}(s,a) \right) (\pi_t - \pi_\tau^*)(da|s) d_\rho^{\pi_\tau^*}(ds) \\ = (1-\gamma)(V_\tau^{\pi_t}(\rho) - V_\tau^{\pi_\tau^*}(\rho)) + \tau \int_S \mathrm{KL}(\pi_\tau^*(\cdot|s)|\pi_t(\cdot|s)) d_\rho^{\pi_\tau^*}(ds)\,.\end{aligned} \tag{32}$$

Substituting this identity into (31) yields

$$\begin{aligned}\partial_t D_{d_\rho^{\pi_\tau^*}}(Z_t, Z^*) &= -\left((1-\gamma)(V_\tau^{\pi_t}(\rho) - V_\tau^{\pi_\tau^*}(\rho)) + \tau \int_S \mathrm{KL}(\pi_\tau^*(\cdot|s)|\pi_t(\cdot|s)) d_\rho^{\pi_\tau^*}(ds) \right) \\ &= -(1-\gamma)(V_\tau^{\pi_t}(\rho) - V_\tau^{\pi_\tau^*}(\rho)) - \tau D_{d_\rho^{\pi_\tau^*}}(Z_t, Z^*)\,,\end{aligned}$$

where the last line follows from Lemma 7 (with $f = Z_t$ and $g = Z^*$). Then for all $t > 0$, by Duhamel's principle and the fact that $V_\tau^{\pi_t}(\rho) \le V_\tau^{\pi_{t'}}(\rho)$ for all $0 \le t' \le t$ (see Proposition 1),

$$\begin{aligned}D_{d_\rho^{\pi_\tau^*}}(Z_t, Z^*) &= e^{-\tau t} D_{d_\rho^{\pi_\tau^*}}(Z_0, Z^*) - (1-\gamma) \int_0^t e^{-\tau(t-t')} (V_\tau^{\pi_{t'}}(\rho) - V_\tau^{\pi_\tau^*}(\rho)) dt' \\ &\le e^{-\tau t} D_{d_\rho^{\pi_\tau^*}}(Z_0, Z^*) - (1-\gamma) \int_0^t e^{-\tau(t-t')} dt' (V_\tau^{\pi_t}(\rho) - V_\tau^{\pi_\tau^*}(\rho))\,.\end{aligned} \tag{33}$$

Note convergence of $\pi_t(\cdot|s)$ in $\|\cdot\|_{\mathcal{M}(A)}$ follows from Pinsker's ineq. □.

Stability

Let $Q : \mathbb{R}_+ \to B_b(S \times A))$ be arbitrary. Consider:

$$\partial_t \pi_t(da|s) \tag{34}$$

$$= -\left(Q_t(s,a) + \tau \ln \frac{\mathrm{d}\pi_t}{\mathrm{d}\mu}(a|s) - \int_A \left(Q_t(s,a) + \tau \ln \frac{\mathrm{d}\pi_t}{\mathrm{d}\mu}(a|s) \right) \pi_t(da|s) \right) \pi_t(da|s)\,, \tag{35}$$

Theorem 8

Assume that $\pi \in C(\mathbb{R}_+; \Pi_\mu)$ satisfies (34) with some $Q : \mathbb{R}_+ \to B_b(S \times A)$. Then for all $\rho \in \mathcal{P}(S)$ and $t > 0$,

$$\min_{r \in [0,t]} V_\tau^{\pi_r}(\rho) - V_\tau^{\pi_\tau^*}(\rho) \le \frac{\tau}{(1-\gamma)(e^{\tau t} - 1)} \left(\int_S \mathrm{KL}(\pi_\tau^*(\cdot|s)|\pi_0(\cdot|s)) d_\rho^{\pi_\tau^*}(ds) + 2\kappa \int_0^t e^{\tau r} \left\| Q_\tau^{\pi_r} - Q_r \right\|_{L^1\left(S \times A, \rho_{\mathrm{ref}} \otimes \frac{\pi_r + \pi_{\mathrm{ref}}}{2}\right)} dr \right), \tag{36}$$

where π_τ^* is the optimal policy defined in (7), $\rho_{\mathrm{ref}} \in \mathcal{P}(S)$ is a reference measure such that $\rho \ll \rho_{\mathrm{ref}}$, $\pi_{\mathrm{ref}} \in \mathcal{P}(A|S)$ is a reference policy such that $\mu \ll \pi_{\mathrm{ref}}$, and $\kappa \ge 1$ is the concentrability coefficient defined by

$$\kappa := \left\| \frac{\mathrm{d}d_\rho^{\pi_\tau^*}}{\mathrm{d}\rho_{\mathrm{ref}}} \right\|_{L^\infty(S, \rho_{\mathrm{ref}})} + \left\| \frac{\mathrm{d}d_\rho^{\pi_\tau^*} \otimes \pi_\tau^*}{\mathrm{d}\rho_{\mathrm{ref}} \otimes \pi_{\mathrm{ref}}} \right\|_{L^\infty(S \times A, \rho_{\mathrm{ref}} \otimes \pi_{ref})}. \tag{37}$$

The estimate (36) holds with $Q_\tau^{\pi_r} - Q_r$ replaced by $Q_\tau^{\pi_r} - Q_r + F_r$ for any measurable $F : \mathbb{R}_+ \to B_b(S)$.

Discussion

In practice:

1. Need to approximate π by e.g. $\boldsymbol{\pi}(f(\cdot;\theta))$.
2. Need to approximate Q_τ^π.
3. Need to add e.g. regularity to evaluate quality of approximations.
4. Need to add "algorithm" (e.g. 1st order method - gradient flow) for learning parameters of Q_τ^π approximation.
5. Need to extend convergence (and stability) by considering the entire coupled system.

We do 1., 2. and partially 4. (least squares problem) in [Leahy et al., 2023].

References I

[Agazzi and Lu, 2020] Agazzi, A. and Lu, J. (2020). Global optimality of softmax policy gradient with single hidden layer neural networks in the mean-field regime. *arXiv preprint arXiv:2010.11858*.

[Bertsekas and Shreve, 2004] Bertsekas, D. P. and Shreve, S. (2004). *Stochastic optimal control: the discrete-time case*. Athena Scientific.

[Bhandari and Russo, 2019] Bhandari, J. and Russo, D. (2019). Global optimality guarantees for policy gradient methods. *arXiv preprint arXiv:1906.01786*.

[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.

References II

[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.

[Fatkhullin et al., 2023] Fatkhullin, I., Barakat, A., Kireeva, A., and He, N. (2023). Stochastic policy gradient methods: Improved sample complexity for Fisher-non-degenerate policies. *arXiv preprint arXiv:2302.01734*.

[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.

[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.

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

[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, 2022] Lan, G. (2022). Policy optimization over general state and action spaces. *arXiv preprint arXiv:2211.16715*.

[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.

[Leahy et al., 2022] Leahy, J.-M., Kerimkulov, B., Šiška, D., and Szpruch, L. (2022). Convergence of policy gradient for entropy regularized MDPs with neural network approximation in the mean-field regime. In *International Conference on Machine Learning*, pages 12222–12252. PMLR.

[Leahy et al., 2023] Leahy, J.-M., Kerimkulov, B., Šiška, D., Szpruch, Ł., and Zhang, Y. (2023). Fisher–rao gradient flow for entropy regularised continuous space mdps: a mirror descent approach. *In preparation*.

References IV

[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.

[Sutton and Barto, 2018] Sutton, R. S. and Barto, A. G. (2018). *Reinforcement learning*. 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*.

References V

[Yuan et al., 2022] Yuan, R., Du, S. S., Gower, R. M., Lazaric, A., and Xiao, L. (2022). Linear convergence of natural policy gradient methods with log-linear policies. *arXiv preprint arXiv:2210.01400*.

[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.

[Zhang et al., 2022] Zhang, M. S., Erdogdu, M. A., and Garg, A. (2022). Convergence and optimality of policy gradient methods in weakly smooth settings. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 9066–9073.

[Zhang et al., 2021] Zhang, Y., Chen, S., Yang, Z., Jordan, M., and Wang, Z. (2021). Wasserstein flow meets replicator dynamics: A mean-field analysis of representation learning in actor-critic. *Advances in Neural Information Processing Systems*, 34:15993–16006.