# Mirror descent for MDPs, controlled diffusions and entropy annealing

David Šiška[1]

Joint work with D. Sethi[1] and Y. Zhang[2] on the controlled diffusions case and B. Kerimkulov[3], J.-M. Leahy[4], L. Szpruch[1,5] and Y. Zhang[2] on the MDPs in Polish spaces case

Advances in Stochastic Control and Reinforcement Learning, Banff International Research Station, Canada

11th April 2025

[1]University of Edinburgh, [2]Imperial College London, [3]Natwest, [4]PhysicsX, [5]Alan Turing Institute

# Outline

- Control problem, performance difference lemma, mirror descent
- Convexity (lack of) in optimisation objective for stochastic control
- Convergence of continuous time Mirror descent in continuous time control
- Annealing
- Discrete stepping mirror descent
- Overview of existing results and references

# Relaxed formulation of a stochastic control problem

Controlled state

$$dX_t^{x,\pi} = \left( \int_A b(X_t^{x,\pi}, a)\pi(da|X_t^{x,\pi}) \right) dt + \sigma(X_t^{x,\pi})dW_t, \quad t > 0; \quad X_0 = x \in \mathcal{O}\,. \tag{1}$$

Aim: *minimise*, over Markov policies $x \mapsto \pi(\cdot|x) \in \mathcal{P}(A)$, the objective:

$$v_0^\pi(x) := \mathbb{E}^{x,\pi}\left[ \int_0^{\tau_\mathcal{O}} \Gamma_t \int_A f(X_t, a)\,\pi(da|X_t)\,dt + \Gamma_{\tau_\mathcal{O}}\, g(X_{\tau_\mathcal{O}}) \right], \tag{2}$$

$$v_\tau^\pi(x) := v_0^\pi(x) + \tau\mathbb{E}^{x,\pi} \int_0^{\tau_\mathcal{O}} \Gamma_t \mathrm{KL}\left(\pi|\mu\right)(X_t)\,dt\,, \tag{3}$$

where

- $\mathcal{O} \subset \mathbb{R}^d$ open, bdd., $A$ a metric space, $\tau_\mathcal{O}$ the first exit time of $X_t^x$ from $\mathcal{O}$.
- $\Gamma_t^\pi = \exp\left(-\int_0^t \int_A c(X_s^{x,\pi}, a)\pi(da|X_s^{x,\pi})ds\right)$ controlled discount factor.
- For $\mu \in \mathcal{P}(A)$ define

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

# Performance difference

## Lemma 1

*For all* $\pi, \pi' \in \Pi_\mu$, $\tau \geq 0$ *and* $x \in \mathcal{O}$,

$$v_\tau^\pi(x) - v_\tau^{\pi'}(x) = \mathbb{E}^{x,\pi}\int_0^{\tau_\mathcal{O}} \Gamma_t \int_A \left(\mathcal{L}^a v_\tau^{\pi'}(X_t) + f(X_t, a) + \tau \ln \tfrac{d\pi'}{d\mu}(a|X_t)\right)(\pi - \pi')(da|X_t)\, dt$$
$$+ \tau \mathbb{E}^{x,\pi} \int_0^{\tau_\mathcal{O}} \Gamma_t KL\left(\pi \middle| \pi'\right)(X_t) dt\,.$$

*Moreover,* $\mathcal{L}^a$ *can be replaced by* $\overline{\mathcal{L}}^a$ *given by*

$$(\overline{\mathcal{L}}^a v)(x) = b(x,a)^\top Dv(x) - c(x,a)v(x) \quad \forall v\,.$$

*Consequence:*

$$\lim_{\varepsilon \searrow 0} \tfrac{1}{\varepsilon}\left(v_\tau^{\pi+\varepsilon(\pi'-\pi)} - v_\tau^\pi\right)(x) = \left\langle \mathcal{L}^\cdot v_\tau^\pi + f + \tau \ln \tfrac{\mathrm{d}\pi}{\mathrm{d}\mu}, \pi' - \pi \right\rangle_{x,\pi} = \left\langle \tfrac{\delta v_\tau^\pi(x)}{\delta \pi}, \pi' - \pi \right\rangle_{x,\pi},$$

where $\langle \cdot, \cdot \rangle_\pi : B_b(\mathcal{O} \times A) \times b\mathcal{M}(A \mid \mathbb{R}^d) \to \mathbb{R}$ is

$$\langle h, \hat{\pi} \rangle_{x,\pi} := \mathbb{E}^{x,\pi} \int_0^{\tau_\mathcal{O}} \Gamma_t h(X_t, a) \hat{\pi}(da|X_t)\, dt$$

and

$$\tfrac{\delta v_\tau^\pi(x)}{\delta \pi} = (\mathcal{L}^a v_\tau^\pi)(x') + f(a, x') + \tau \ln \tfrac{\mathrm{d}\pi}{\mathrm{d}\mu}(a|x')\,.$$

# Performance difference

Compact notation:

$$(v_\tau^{\pi} - v_\tau^{\pi'})(x) = \left\langle \tfrac{\delta v_\tau^{\pi'}}{\delta \pi}, \pi - \pi' \right\rangle_{x,\pi} + \tau \int_{\mathcal{O}} \mathrm{KL}(\pi|\pi')(x') d_x^{\pi}(dx')\,, \tag{4}$$

where

$$d_x^{\pi}(B) := \mathbb{E}^{x,\pi} \int_0^{\tau_{\mathcal{O}}} \Gamma_t \mathbf{1}_B(X_t)\, dt \text{ and } \langle h, \hat{\pi} \rangle_{x,\pi} := \mathbb{E}^{x,\pi} \int_0^{\tau_{\mathcal{O}}} \Gamma_t h(X_t, a) \hat{\pi}(da|X_t)\, dt\,.$$

# Performance difference MDP case

Infinite-horizon Markov decision model $(S, A, P, c, \gamma)$. Minimise over Markov policies $S \ni s \mapsto \pi(da|s) \in \mathcal{P}(A)$ the objective:

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

Q-function

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

The occupancy kernel

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

## Lemma 2 (Performance difference[1])

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

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

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

# Performance difference MDP case

*Consequence:* For $\pi, \pi' \in \Pi_\mu$ have

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

Compact notation:

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

where

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

and

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

*Aside:* Both "Q-learning" (MDPs) and "q-learning"[2] (continuous time controlled diffusions) are "*First variation*-learning".

[2] Jia, Y. and Zhou, X. Y. (2023). q-learning in continuous time. *JMLR*, 24.

# Deriving mirror descent

Let's say we have $\pi_{\text{old}}$. By Perf. Diff (4)

$$v_\tau^\pi(x) = v_\tau^{\pi_{\text{old}}}(x) + \left\langle \frac{\delta v_\tau^{\pi_{\text{old}}}}{\delta \pi}, \pi - \pi_{\text{old}} \right\rangle_{x,\pi} + \tau \int_{\mathcal{O}} \mathrm{KL}(\pi|\pi_{\text{old}})(x') d_x^\pi(dx') \,.$$

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

$$L^\pi(x) := v_\tau^{\pi_{\text{old}}}(x) + \left\langle \frac{\delta v_\tau^{\pi_{\text{old}}}}{\delta \pi}, \pi - \pi_{\text{old}} \right\rangle_{x,\pi_{\text{old}}} + \lambda \int_{\mathcal{O}} \mathrm{KL}(\pi|\pi_{\text{old}})(x') d_x^{\pi_{\text{old}}}(dx') \,.$$

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

$$\pi_{\text{new}}(da|x') = \arg\min_\pi \left( v_\tau^{\pi_{\text{old}}}(x) + \int_A \frac{\delta v_\tau^{\pi_{\text{old}}}}{\delta \pi}(x', a)(\pi - \pi_{\text{old}})(da|x') + \lambda \,\mathrm{KL}(\pi|\pi_{\text{old}})(x') \right) .$$

# Motivation for studying mirror descent

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

$$\theta_{\text{new}} = \theta_{\text{old}} + \eta \nabla_\theta v_\tau^{\pi_{\theta_{\text{old}}}}(x)\,.$$

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

Instead, re-write the mirror descent objective:

$$\begin{aligned}
L_{\text{MD}}(\theta) &= \left\langle \frac{\delta v_\tau^{\pi_{\theta_{\text{old}}}}}{\delta \pi}, \pi_\theta \right\rangle_{x, \pi_{\theta_{\text{old}}}} + \lambda \int_{\mathcal{O}} \text{KL}(\pi_\theta | \pi_{\theta_{\text{old}}})(x') d_x^{\pi_{\theta_{\text{old}}}}(dx') \\
&= \mathbb{E}^{x' \sim d_x^{\pi_{\theta_{\text{old}}}}}_{a \sim \pi_\theta(\cdot|x')} \left[ \frac{\delta v_\tau^{\pi_{\theta_{\text{old}}}}}{\delta \pi}(a, x') + \lambda \,\text{KL}(\pi_\theta | \pi_{\theta_{\text{old}}})(x') \right] \\
&= \mathbb{E}^{x' \sim d_x^{\pi_{\theta_{\text{old}}}}}_{a \sim \pi_{\theta_{\text{old}}}(\cdot|x')} \left[ \frac{\delta v_\tau^{\pi_{\theta_{\text{old}}}}}{\delta \pi}(a, x') \frac{\mathrm{d}\pi_\theta}{\mathrm{d}\pi_{\theta_{\text{old}}}}(a|x') + \lambda \,\text{KL}(\pi_\theta | \pi_{\theta_{\text{old}}})(x') \right].
\end{aligned}$$

Step $\eta > 0$:

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

# Connection to PPO

Mirror descent objective

$$L_{\mathrm{MD}}(\theta) = \mathbb{E}^{x' \sim d_x^{\pi_{\theta_{\mathrm{old}}}}}_{a \sim \pi_{\theta_{\mathrm{old}}}(\cdot|x')} \left[ \frac{\delta v_\tau^{\pi_{\theta_{\mathrm{old}}}}}{\delta \pi}(a, x') \frac{\mathrm{d}\pi_\theta}{\mathrm{d}\pi_{\theta_{\mathrm{old}}}}(a|x') + \lambda \, \mathrm{KL}(\pi_\theta | \pi_{\theta_{\mathrm{old}}})(x') \right].$$

Proximal policy optimization (PPO) [Schulman et al., 2017] replaces this with:

$$L_{\mathrm{PPO}}(\theta) = \mathbb{E}^{x' \sim d_x^{\pi_{\theta_{\mathrm{old}}}}}_{a \sim \pi_{\theta_{\mathrm{old}}}(\cdot|x')} \Bigg( \min \Bigg[ \frac{\delta v_\tau^{\pi_{\theta_{\mathrm{old}}}}}{\delta \pi}(a, x') \frac{\mathrm{d}\pi_\theta}{\mathrm{d}\pi_{\theta_{\mathrm{old}}}}(a|x'), \\ \mathrm{clip}\left(1 - \varepsilon, 1 + \varepsilon, \frac{\mathrm{d}\pi_\theta}{\mathrm{d}\pi_{\theta_{\mathrm{old}}}}(a|x')\right) \frac{\delta v_\tau^{\pi_{\theta_{\mathrm{old}}}}}{\delta \pi}(a, x') \Bigg] \Bigg).$$

**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

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

# Mirror descent for static minimization

# Mirror descent stepping to Mirror descent flow to Fisher–Rao

Aim: minimize $\mathcal{P}(A) \ni m \mapsto F(m) \in \mathbb{R} \cup \{+\infty\}$. Mirror stepping:

$$m^{n+1} = \operatorname*{argmin}_{m \in \mathcal{P}(A)} \langle \tfrac{\delta F(m^n)}{\delta m}, m - m^n \rangle + \lambda KL(m|m^n)\,, n \in \mathbb{N}\,, m^0 \in \mathcal{P}(A) \text{ given.}$$

This is "linear in measure plus KL" and given by

$$\tfrac{\mathrm{d}m^{n+1}}{\mathrm{d}m^n}(a) = \exp\left(-\tfrac{1}{\lambda}\tfrac{\delta F(m^n)}{\delta m}(a)\right) / \left\langle \exp\left(-\tfrac{1}{\lambda}\tfrac{\delta F(m^n)}{\delta m}\right), m^n \right\rangle .$$

Then

$$\ln \tfrac{\mathrm{d}m^{n+1}}{\mathrm{d}\mu} - \ln \tfrac{\mathrm{d}m^n}{\mathrm{d}\mu} = -\tfrac{1}{\lambda}\tfrac{\delta F(m^n)}{\delta m} - \ln \left\langle \exp\left(-\tfrac{1}{\lambda}\tfrac{\delta F(m^n)}{\delta m}\right), m^n \right\rangle .$$

Writing $z^n = \ln \tfrac{\mathrm{d}m^n}{\mathrm{d}\mu}$, interpolating, letting $\lambda \to \infty$ yields

$$\partial_s z_s = -\left(\tfrac{\delta F(\boldsymbol{\pi}(z_s))}{\delta m} - \langle \tfrac{\delta F(\boldsymbol{\pi}(z_s))}{\delta m}, \boldsymbol{\pi}(z_s) \rangle\right) = -\tfrac{\delta F(\boldsymbol{\pi}(z_s))}{\delta m}\,, s \geq 0\,,\ \boldsymbol{\pi}(z) = \tfrac{e^z}{\langle e^z, \mu \rangle}\mu\,.$$

Setting $m_s = \boldsymbol{\pi}(z_s)$ and chain rule

$$\begin{aligned} \partial_s m_s = \partial_s \boldsymbol{\pi}(z_s) &= \Big(\partial_s z_s - \langle \partial_s z_s, \boldsymbol{\pi}(z_s) \rangle\Big)\boldsymbol{\pi}(z_s) \\ &= -\tfrac{\delta F(m_s)}{\delta m} m_s\,. \end{aligned}$$

$$\partial_s m_s(da) = -\frac{\delta F(m_s)}{\delta m}(a)\, m_s(da)\,,\ s>0\,,\ \ m_0 \in \mathcal{P}(A) \text{ given.}$$

Energy dissipation:

$$\partial_s F(m_s) = \left\langle \frac{\delta F(m_s)}{\delta m}, \partial_s m_s \right\rangle = -\left\langle \left| \frac{\delta F(m_s)}{\delta m} \right|^2, m_s \right\rangle \le 0\,, s \ge 0\,.$$

## Definition (Convexity)

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

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

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

*Convergence if $\tau > 0$:* for $m^* = \mathrm{argmin}_{m\in\mathcal{P}(A)} F(A)$

$$\partial_s\, \mathrm{KL}(m^*|m_s) = \partial_s \left\langle \ln \tfrac{\mathrm{d}m^*}{\mathrm{d}\mu} - \ln \tfrac{\mathrm{d}m_s}{\mathrm{d}\mu}, m^* \right\rangle = -\left\langle \partial_s \ln \tfrac{\mathrm{d}m_s}{\mathrm{d}\mu}, m^* \right\rangle$$

$$= -\left\langle \frac{1}{\frac{\mathrm{d}m_s}{\mathrm{d}\mu}} \partial_s \tfrac{\mathrm{d}m_s}{\mathrm{d}\mu}, m^* \right\rangle = \left\langle \frac{\delta F(m_s)}{\delta m}, m^* \right\rangle$$

$$= \left\langle \frac{\delta F(m_s)}{\delta m}, m^* - m_s \right\rangle \le F(m^*) - F(m_s) - \tau\, \mathrm{KL}(m^*|m_s)\,.$$

Hence

$$e^{\tau s}\, \mathrm{KL}(m^*|m_s) \le \mathrm{KL}(m^*|m_0) + \int_0^s e^{\tau r}\, dr (F(m^*) - F(m_s))$$

and so

$$0 \le F(m_s) - F(m^*) \le \tfrac{\tau}{e^{\tau s}-1}\, \mathrm{KL}(m^*|m^0)\,.$$

# Lack of convexity in Markov controls and softmax

# Lack of convexity in Markov policies

Consider a simple LQR problem with the state given by just

$$dX_t = u(t, X_t)\, dt$$

and objective to *minimize*

$$v^u(s, x) = \int_s^T u(t, X_t)^2\, dt\,.$$

For

$$\mathcal{A} := \{u : [0, T] \times \mathbb{R} \mapsto \mathbb{R} : \text{ Lipschitz in space}\}$$

the map $\mathcal{A} \ni u \mapsto v^u(0, x)$ is *not* convex, e.g. [Giegrich et al., 2024, Proposition 2.4].

For

$$\mathcal{A} := \left\{\alpha : [0, T] \to \mathbb{R} : \int_0^T \alpha_t^2\, dt < \infty\right\}$$

we have $\mathcal{A} \ni \alpha \mapsto v^\alpha(0, x)$ *is* convex from Pontryagin optimality principle.

# Lack of convexity in softmax parametrization

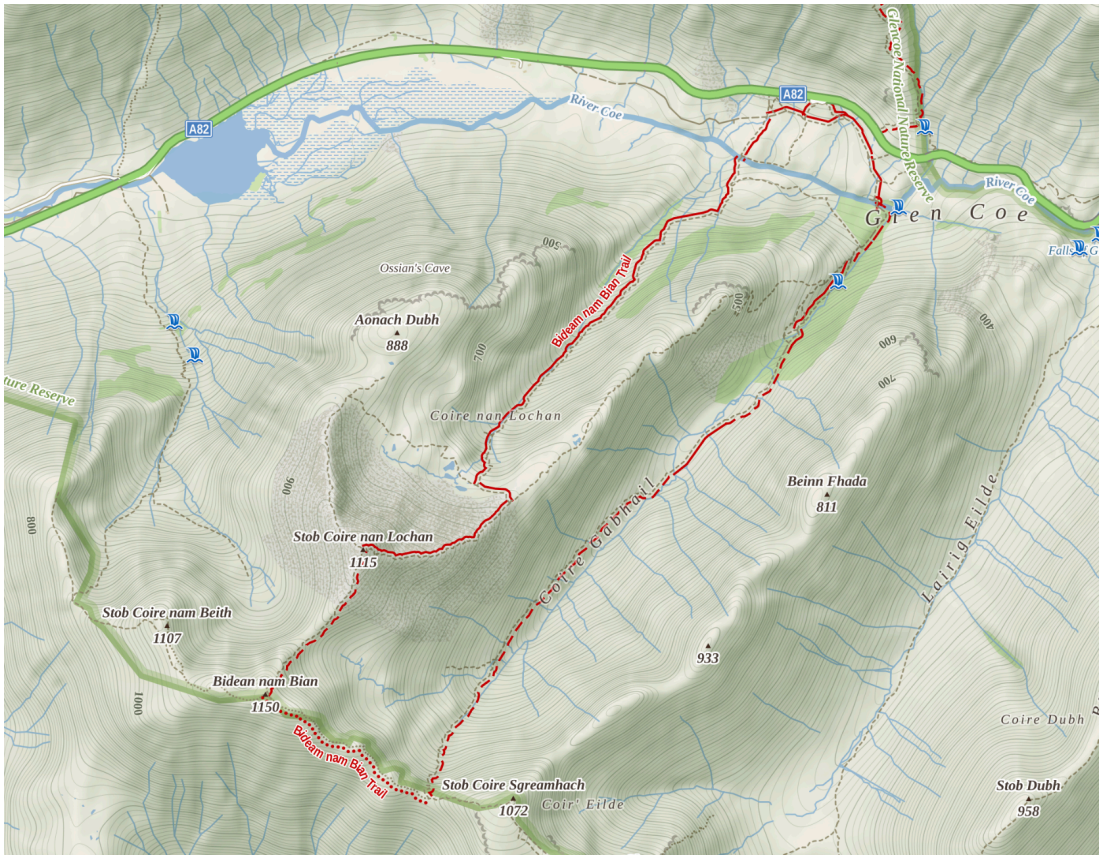
Consider *minimizing*, over $\theta \in \mathbb{R}^{|A|}$ the objective

$$v(\theta) := \int_A c(a)\pi(da;\theta),$$

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

Then the map $\theta \mapsto v(\theta)$ is in general *not* convex [Mei et al., 2020, Propn. 1].

# Gradient dominance




Non-uniform Polyak-Łojasiewicz: there is $\mu : A \to (0, \infty)$ s.t. for all $a \in A$

$$0 \leq f(a) - f(a^*) \leq \mu(a) |\nabla f(a)|^2 .$$

Easily see that with $\frac{d}{ds} a_s = -\nabla f(a_s)$ you get

$$0 \leq f(a_s) - f(a^*) \leq (f(a_0) - f(a^*)) \exp\Big( - \int_0^s \mu^{-1}(a_r)\, dr \Big) .$$

# Convergence of continuous time Mirror descent for continuous time control

# Gradient flow

*Mirror flow*:

$$\partial_s Z_s(x,a) = -\frac{\delta v_{\tau_s}^{\boldsymbol{\pi}(Z_s)}}{\delta \pi}(x,a)\,, \quad \boldsymbol{\pi}(Z)(da|x) \propto e^{Z(x,a)}\,, \quad s>0\,,\ Z_0 \text{ given}\,. \tag{5}$$

## Theorem 3

*Under appropriate assumptions for each $Z_0 \in B_b(\mathcal{O}\times A)$ and $\boldsymbol{\tau} \in C([0,\infty);(0,\infty))$, there exists a unique $Z \in \cap_{S>0} C^1([0,S];B_b(\mathcal{O}\times A))$ satisfying (5).*

## Theorem 4

*Let $\boldsymbol{\tau} \in C^1([0,\infty);(0,\infty))$, and $Z \in \cap_{S>0} C^1([0,S];B_b(\mathcal{O}\times A))$ be the solution to (5). Then for all $x \in \mathcal{O}$ and for all $s>0$,*

$$\begin{aligned}\partial_s v_{\boldsymbol{\tau}_s}^{\boldsymbol{\pi}(Z_s)}(x) = -\mathbb{E}^{x,\boldsymbol{\pi}(Z_s)} \int_0^{\tau_{\mathcal{O}}} \Gamma_t \int_A \left|\frac{\delta v_{\boldsymbol{\tau}_s}^{\boldsymbol{\pi}(Z_s)}}{\delta \pi}(a|X_t)\right|^2 \boldsymbol{\pi}(Z_s)(da|X_t)dt \\ + (\partial_s \boldsymbol{\tau}_s)\mathbb{E}^{\mathbb{P}^x,\boldsymbol{\pi}(Z_s)} \int_0^{\tau_{\mathcal{O}}} \Gamma_t^{\boldsymbol{\pi}(Z_s)} KL(\boldsymbol{\pi}(Z_s)|\mu)(X_t)dt\,.\end{aligned} \tag{6}$$

# Exponential convergence

*Mirror flow*:

$$\partial_s Z_s(x,a) = -\frac{\delta v_{\tau_s}^{\boldsymbol{\pi}(Z_s)}}{\delta \pi}(x,a)\,, \quad \boldsymbol{\pi}(Z)(da|x) \propto e^{Z(x,a)}\,, \quad s>0\,,\ Z_0 \in B_b(\mathcal{O}\times A) \text{ given}\,.$$

## Theorem 5

*There exists $C>0$ such that for all decreasing $\boldsymbol{\tau} \in C^1([0,\infty);(0,\infty))$we have for all $s>0$, $x \in \overline{\mathcal{O}}$ and $\tau>0$,*

$$v_{\boldsymbol{\tau}_s}^{\boldsymbol{\pi}(Z_s)}(x) - v_\tau^*(x) \le \frac{C}{\tau}\left(\frac{1+\tau}{\int_0^s e^{\int_0^{s'}\boldsymbol{\tau}_r dr}ds'} + \frac{\int_0^s (\boldsymbol{\tau}_{s'}-\tau)e^{\int_0^{s'}\boldsymbol{\tau}_r dr}ds'}{\int_0^s e^{\int_0^{s'}\boldsymbol{\tau}_r dr}ds'}\right)\,. \tag{7}$$

*Assume further that $A$ is of finite cardinality. Then for all $x \in \mathcal{O}$ and $s>0$,*

$$v_{\boldsymbol{\tau}_s}^{\boldsymbol{\pi}(Z_s)}(x) - v_\tau^*(x) \le C\left(\frac{1}{\int_0^s e^{\int_0^{s'}\boldsymbol{\tau}_r dr}ds'} + \frac{\int_0^s (\boldsymbol{\tau}_{s'}-\tau)e^{\int_0^{s'}\boldsymbol{\tau}_r dr}ds'}{\int_0^s e^{\int_0^{s'}\boldsymbol{\tau}_r dr}ds'}\right)\,. \tag{8}$$

# Exponential convergence

*Mirror flow*:

$$\partial_s Z_s(x,a) = -\frac{\delta v_{\tau_s}^{\boldsymbol{\pi}(Z_s)}}{\delta \pi}(x,a)\,, \quad \boldsymbol{\pi}(Z)(da|x) \propto e^{Z(x,a)}\,, \ s>0\,, \ Z_0 \in B_b(\mathcal{O}\times A) \text{ given}\,.$$

## Corollary 6

*There exists $C>0$ s.t. for all $\tau>0$, and for all $s>0$, $x\in\overline{\mathcal{O}}$*

$$0 \le v_\tau^{\boldsymbol{\pi}(Z_s)}(x) - v_\tau^*(x) \le C\frac{1}{e^{\tau s}-1}\,. \tag{9}$$

*Assume further that $A$ is of finite cardinality. There exists $C>0$ s.t. for all $\tau>0$, and for all $s>0$, $x\in\overline{\mathcal{O}}$*

$$0 \le v_\tau^{\boldsymbol{\pi}(Z_s)}(x) - v_\tau^*(x) \le C\frac{\tau}{e^{\tau s}-1}\,. \tag{10}$$

In the finite cardinality case as $\lim_{\tau\to 0}\frac{\tau}{e^{\tau s}-1} = \frac{1}{s}$.

# Annealing

# Annealing: the unregularized problem

Unregularized Hamiltonian $H : \mathcal{O} \times \mathbb{R} \times \mathbb{R}^d \to \mathbb{R}$ such that for all $(x, u, p) \in \mathcal{O} \times \mathbb{R} \times \mathbb{R}^d$,

$$H(x, u, p) := \inf_{a \in A} \Big( b(x, a)^\top p - c(x, a)u + f(x, a) \Big) \,. \tag{11}$$

## Assumption 1

1. *$A$ is a nonempty, compact and separable metric space. For all $x \in \mathcal{O}$, $b(x, \cdot)$, $c(x, \cdot)$ and $f(x, \cdot)$ are continuous on $A$.*
2. *If a set $\mathcal{C} \subset A$ satisfies $\mu(\mathcal{C}) = 1$, then $\mathcal{C}$ is dense in $A$.*

The HJB for the unregularized control problem

$$\inf_{a \in A} \Big( (\mathcal{L}^a v)(x) + f(x, a) \Big) = 0, \quad \text{a.e. } x \in \mathcal{O}; \quad v(x) = g(x), \quad x \in \partial\mathcal{O} \,, \tag{12}$$

# Annealing: general case

## Theorem 7

*The HJB* (12) *has a unique solution in* $W^{2,p^*}(\mathcal{O})$ *which is* $v_0^* = \inf_\pi v_0^\pi$. *Moreover, there exists* $C \geq 0$ *such that for all* $\tau > 0$,

$$0 \leq v_0^{\pi_\tau^*} - v_0^* \leq v_\tau^* - v_0^* \leq C \| \left(H_\tau(\cdot, v_0^*(\cdot), Dv_0^*(\cdot)) - H(\cdot, v_0^*(\cdot), Dv_0^*(\cdot))\right)^+ \|_{L^{p^*}(\mathcal{O})}.$$

*Finally,* $\lim_{\tau \to 0} \|(H_\tau(\cdot, v_0^*(\cdot), Dv_0^*(\cdot)) - H(\cdot, v_0^*(\cdot), Dv_0^*(\cdot)))^+\|_{L^{p^*}(\mathcal{O})} = 0$ *and consequently,* $\lim_{\tau \to 0} v_0^{\tau_\tau^*} = v_0^*$ *uniformly on* $\overline{\mathcal{O}}$.

# Annealing: finite cardinality A

## Theorem 8

*Suppose* $A = \{a_1, \ldots, a_N\}$. *Let* $\tau_s = 1/(1+s)$. *Then there exists* $C > 0$ *such that*

$$0 \leq v_0^{\pi(Z_s)}(x) - v_0^*(x) \leq \frac{C}{s}, \quad \forall x \in \overline{\mathcal{O}}, s > 1\,.$$

# Annealing: example setting

## Assumption 2

*There exists* $C \geq 0$ *and* $\tau_{\max} \in (0,1)$ *such that for all* $\tau \in (0, \tau_{\max}]$ *and* $x \in \mathcal{O}$,

$$H_\tau(x, v_0^*(x), Dv_0^*(x)) - H(x, v_0^*(x), Dv_0^*(x)) \leq C\tau \ln \tfrac{1}{\tau} .$$

## Proposition 3

*Suppose that* $A \subset \mathbb{R}^k$ *is a nonempty convex and compact set, and* $\mu \in \mathcal{P}(A)$ *is the uniform distribution on* $A$. *Assume further that* $b \in C(\overline{\mathcal{O}} \times A; \mathbb{R}^d)$, $c \in C(\overline{\mathcal{O}} \times A; \mathbb{R})$ *and* $f \in C(\overline{\mathcal{O}} \times A; \mathbb{R})$ *are such that for all* $x \in \overline{\mathcal{O}}$,

$$A \ni a \mapsto h(x,a) := b(x,a)^\top Dv_0^*(x) - c(x,a)v_0^*(x) + f(x,a) \in \mathbb{R}$$

*admits a unique minimiser in the interior of* $A$ *and is twice differentiable with derivative* $D^2_{aa}h \in C(\overline{\mathcal{O}} \times A; \mathbb{R}^{k \times k})$. *Then Assumption 2 holds.*

## Theorem 9

*Let* $\tau_s = 1/\sqrt{1+s}$ *for all* $s > 0$. *Then* $\lim_{s\to\infty} v_0^{\pi(Z_s)}(x) = v_0^*(x)$ *for all* $x \in \overline{\mathcal{O}}$.
*If Assumption 2 holds then there exists* $C > 0$ *such that*

$$0 \le v_0^{\pi(Z_s)}(x) - v_0^*(x) \le \frac{C \ln s}{\sqrt{s}}, \quad \forall x \in \overline{\mathcal{O}}, s > 1\,.$$

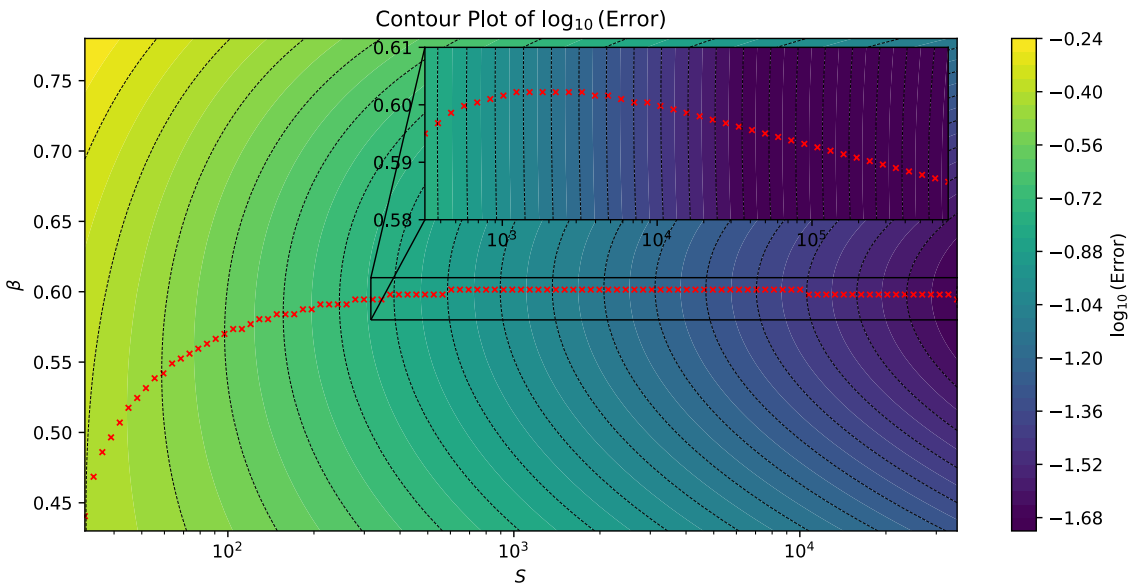



Figure: Theoretical error with scheduler $\tau_s = (1+s)^{-\beta}$

# Discrete stepping mirror descent

# Mirror descent stepping: Energy dissipation

$$m^{n+1} = \operatorname*{argmin}_{m\in\mathcal{P}(A)} \langle \tfrac{\delta F(m^n)}{\delta m}, m - m^n\rangle + \lambda KL(m|m^n)\,, n \in \mathbb{N}\,, m^0 \in \mathcal{P}(A) \text{ given.}$$

So for all $m$

$$\langle \tfrac{\delta F(m^n)}{\delta m}, m - m^n\rangle + \lambda KL(m|m^n) \geq \langle \tfrac{\delta F(m^n)}{\delta m}, m^{n+1} - m^n\rangle + \lambda KL(m^{n+1}|m^n)$$

hence with $m = m^n$

$$0 \geq \langle \tfrac{\delta F(m^n)}{\delta m}, m^{n+1} - m^n\rangle + \lambda KL(m^{n+1}|m^n)\,.$$

**Definition (L-smoothness)**

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

$$F(m) - F(m') \leq \left\langle \tfrac{\delta F(m')}{\delta m}, m - m' \right\rangle + L\,\mathrm{KL}(m|m')\,.$$

Then if $\lambda \geq L$ we get

$$F(m^{n+1}) - F(m^n) \leq \left\langle \tfrac{\delta F(m^n)}{\delta m}, m^{n+1} - m^n \right\rangle + \lambda\,\mathrm{KL}(m^{n+1}|m^n) \leq 0\,.$$

# Mirror descent stepping: Convergence

## Lemma 10 (Bregman proximal inequality)

*Let $\mathcal{C} \subseteq \mathcal{P}(A)$ be convex and let $G : \mathcal{C} \to \mathbb{R}$ be convex. Fix $m' \in \mathcal{C}$ and define*

$$\hat{m} \in \operatorname*{argmin}_{m \in \mathcal{C}} G(m) + \mathrm{KL}(m|m') \,.$$

*Then*

$$G(m) + \mathrm{KL}(m|m') \geq G(\hat{m}) + \mathrm{KL}(m|\hat{m}) + \mathrm{KL}(\hat{m}|m') \,.$$

Using this:

$$\tfrac{1}{\lambda} \langle \tfrac{\delta F(m^n)}{\delta m}, m - m^n \rangle + \mathrm{KL}(m|m^n) \geq \tfrac{1}{\lambda} \langle \tfrac{\delta F(m^n)}{\delta m}, m^{n+1} - m^n \rangle + \mathrm{KL}(m|m^{n+1}) + \mathrm{KL}(m^{n+1}|m^n)$$

this, convexity and L-smoothness, $\lambda \geq L$, $\mathrm{KL}(\cdot|\cdot) \geq 0$:

$$\begin{aligned} \lambda(\mathrm{KL}(m^*|m^{n+1}) - \mathrm{KL}(m^*|m^n)) &\leq \langle \tfrac{\delta F(m^n)}{\delta m}, m^* - m^n \rangle - \langle \tfrac{\delta F(m^n)}{\delta m}, m^{n+1} - m^n \rangle - \lambda \,\mathrm{KL}(m^{n+1}|m^n) \\ &\leq F(m^*) - F(m^n) - \tau \,\mathrm{KL}(m^*|m^n) - F(m^{n+1}) + F(m^n) \end{aligned}$$

Hence

$$\mathrm{KL}(m^*|m^{n+1}) + \tfrac{1}{\lambda}[F(m^{n+1}) - F(m^*)] \leq \tfrac{\lambda-\tau}{\lambda} \,\mathrm{KL}(m^*|m^n)) \leq \cdots \leq \left(\tfrac{\lambda-\tau}{\lambda}\right)^{n+1} \mathrm{KL}(m^*|m^0)) \,.$$

# Mirror decent stepping for MDP & continuous time control

Pointwise in $x \in \mathcal{O}$ optimization:

$$\pi^{n+1}(\cdot|x) = \arg\min_{m \in \mathcal{P}(A)} \left( \int_A \frac{\delta v_\tau^{\pi^n}}{\delta \pi}(x,a)(m(da) - \pi^n(da|x)) + \lambda \, \mathrm{KL}(m|\pi^n)(x) \right). \tag{13}$$

MDP case:

- Can prove linear convergence in general using Bregman proximal inequality and perf. diff. (for both L-smoothness and convexity). Proof relies on MDP structure: for $F \leq 0$:

$$\int_S F(s') d_\rho^s(ds') = \int_S F(s') P_\pi^0(ds'|s) + \sum_{k=1}^{\infty} \int_S \gamma^k F(s') P_\pi^k(ds'|s) \leq \int_S F(s') \delta_s(ds') = F(s) \, .$$

- Under assumptions that help handle occupation measure can prove exponential convergence: need $d_\rho^{\pi_\tau^*} \ll \rho$ and $\xi := (1-\gamma)^{-1} \left\| \frac{\mathrm{d} d_\rho^{\pi_\tau^*}}{\mathrm{d}\rho} \right\|_{B_b(S)} < \infty$.

Continuous time control:

- Can prove convergence, no rate (time disc. estimate not uniform in time vs. exponential convergence on continuous time flow) .
- The argument for exponential convergence "works" but the assumptions on occupations measures seem impossible.

Thank you!

Questions?

# Existing results

# Mirror descent

Static optimization problem:

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

# 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].

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 for MDPs with discrete state and action spaces

- [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.
- [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 $L^\infty$ and $L^2$ approx errors respectively.

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

# NPG / Mirror descent for MDPs with 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].

- Discrete step mirror descent and Fisher–Rao flow: Exponential convergence for entropy regularized MDPs in Polish state & action spaces [Kerimkulov et al., 2023].

# Continuous time relaxed control and RL

- Exploratory LQR: [Wang et al., 2020], [Giegrich et al., 2024], [Szpruch et al., 2024].
- Continuous-time RL algorithms via MDPs: [Munos, 2006], [Munos and Bourgine, 1998], [Munos, 2000], [Han and E, 2016], [Tallec et al., 2019], [Pham and Warin, 2022], [Hamdouche et al., 2022].
- Continuous-time RL algorithms robust to time discretization: [Jia and Zhou, 2022a, Jia and Zhou, 2022b], [Jia and Zhou, 2023], [Zhao et al., 2024].
- Convergence in convex open loop setting: [Šiška and Szpruch, 2020], [Sethi and Šiška, 2022], [Kerimkulov et al., 2022], [Kerimkulov et al., 2024].
- Mean field: [Guo et al., 2022, Wei and Yu, 2023], [Pham and Warin, 2023], [Frikha et al., 2023],
  - *Convergence of policy gradient in LQR: [Carmona et al., 2019], [Wang et al., 2021],*
  - *Full error analysis in LQR: [Frikha et al., 2024].*

# References I


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

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

[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*.

[Carmona et al., 2019] Carmona, R., Laurière, M., and Tan, Z. (2019). Linear-quadratic mean-field reinforcement learning: convergence of policy gradient methods. *arXiv preprint arXiv:1910.04295*.

[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.

# References II


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

[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.

[Frikha et al., 2023] Frikha, N., Germain, M., Laurière, M., Pham, H., and Song, X. (2023). Actor-critic learning for mean-field control in continuous time. *arXiv preprint arXiv:2303.06993*.

[Frikha et al., 2024] Frikha, N., Pham, H., and Song, X. (2024). Full error analysis of policy gradient learning algorithms for exploratory linear quadratic mean-field control problem in continuous time with common noise. *arXiv preprint arXiv:2408.02489*.

[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.

# References III


[Guo et al., 2022] Guo, X., Xu, R., and Zariphopoulou, T. (2022). Entropy regularization for mean field games with learning. *Mathematics of Operations Research*.

[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.

[Hamdouche et al., 2022] Hamdouche, M., Henry-Labordere, P., and Pham, H. (2022). Policy gradient learning methods for stochastic control with exit time and applications to share repurchase pricing. *Applied Mathematical Finance*, 29(6):439–456.

[Han and E, 2016] Han, J. and E, W. (2016). Deep learning approximation for stochastic control problems. *arXiv preprint arXiv:1611.07422*.

[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.

[Jia and Zhou, 2022a] Jia, Y. and Zhou, X. Y. (2022a). Policy evaluation and temporal-difference learning in continuous time and space: A martingale approach. *Journal of Machine Learning Research*, 23(154):1–55.

# References IV


[Jia and Zhou, 2022b] Jia, Y. and Zhou, X. Y. (2022b). Policy gradient and actor-critic learning in continuous time and space: Theory and algorithms. *The Journal of Machine Learning Research*, 23(1):12603–12652.

[Jia and Zhou, 2023] Jia, Y. and Zhou, X. Y. (2023). q-learning in continuous time. *The Journal of Machine Learning Research*, 24:161–1.

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

[Kerimkulov et al., 2022] Kerimkulov, B., Leahy, J.-M., Šiška, D., and Szpruch, Ł. (2022). Convergence of policy gradient for entropy regularized MDPs with neural network approximation in the mean-field regime. *arXiv preprint arXiv:2201.07296*.

[Kerimkulov et al., 2023] Kerimkulov, B., Leahy, J.-M., Šiška, D., Szpruch, Ł., and Zhang, Y. (2023). A Fisher–Rao gradient flow for entropy-regularised Markov decision processes in Polish spaces. *arXiv preprint arXiv:2310.02951*.

[Kerimkulov et al., 2024] Kerimkulov, B., Šiška, D., Szpruch, Ł., and Zhang, Y. (2024). Mirror descent for stochastic control problems with measure-valued controls. *arXiv preprint arXiv:2401.01198*.

[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.

# References V


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

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

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

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

[Munos, 2000] Munos, R. (2000). A study of reinforcement learning in the continuous case by the means of viscosity solutions. *Machine Learning*, 40(3):265–299.

[Munos, 2006] Munos, R. (2006). Policy gradient in continuous time. *Journal of Machine Learning Research*, 7(May):771–791.

# References VI


[Munos and Bourgine, 1998] Munos, R. and Bourgine, P. (1998). Reinforcement learning for continuous stochastic control problems. In *Advances in Neural Information Processing Systems*, pages 1029–1035.

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

[Pham and Warin, 2022] Pham, H. and Warin, X. (2022). Mean-field neural networks-based algorithms for McKean–Vlasov control problems. *arXiv preprint arXiv:2212.11518*.

[Pham and Warin, 2023] Pham, H. and Warin, X. (2023). Actor critic learning algorithms for mean-field control with moment neural networks. *arXiv preprint arXiv:2309.04317*.

[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*.

[Sethi and Šiška, 2022] Sethi, D. and Šiška, D. (2022). The modified MSA, a gradient flow and convergence. *arXiv preprint arXiv:2212.05784*.

[Šiška and Szpruch, 2020] Šiška, D. and Szpruch, Ł. (2020). Gradient flows for regularized stochastic control problems. *arXiv preprint arXiv:2006.05956*.

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

# References VII


[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.

[Szpruch et al., 2024] Szpruch, L., Treetanthiploet, T., and Zhang, Y. (2024). Optimal scheduling of entropy regularizer for continuous-time linear-quadratic reinforcement learning. *SIAM Journal on Control and Optimization*, 62(1):135–166.

[Tallec et al., 2019] Tallec, C., Blier, L., and Ollivier, Y. (2019). Making Deep Q-learning methods robust to time discretization. *arXiv preprint arXiv:1901.09732*.

[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.

[Wang et al., 2020] Wang, H., Zariphopoulou, T., and Zhou, X. Y. (2020). Reinforcement learning in continuous time and space: A stochastic control approach. *Journal of Machine Learning Research*, 21(198):1–34.

[Wang et al., 2021] Wang, W., Han, J., Yang, Z., and Wang, Z. (2021). Global convergence of policy gradient for linear-quadratic mean-field control/game in continuous time. In *International Conference on Machine Learning*, pages 10772–10782. PMLR.

# References VIII


[Wei and Yu, 2023] Wei, X. and Yu, X. (2023). Continuous-time q-learning for mean-field control problems. *arXiv preprint arXiv:2306.16208*.

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

[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.

[Zhao et al., 2024] Zhao, H., Tang, W., and Yao, D. (2024). Policy optimization for continuous reinforcement learning. *Advances in Neural Information Processing Systems*, 36.