Policy Mirror Descent with Dual Function Approximation

Lin Xiao
Meta FAIR (Fundamental AI Research)

ICCOPT 2025, USC

Outline

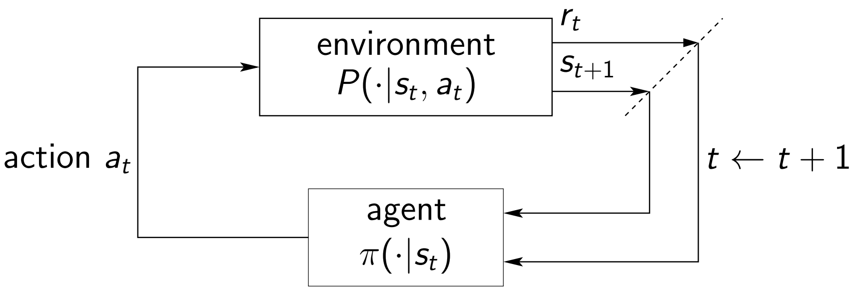
- Markov decision process (MDP)
- policy mirror descent (PMD) method (tabular case)
- PMD with dual function approximation

Outline

- Markov decision process (MDP)
- policy mirror descent (PMD) method (tabular case)
- PMD with dual function approximation

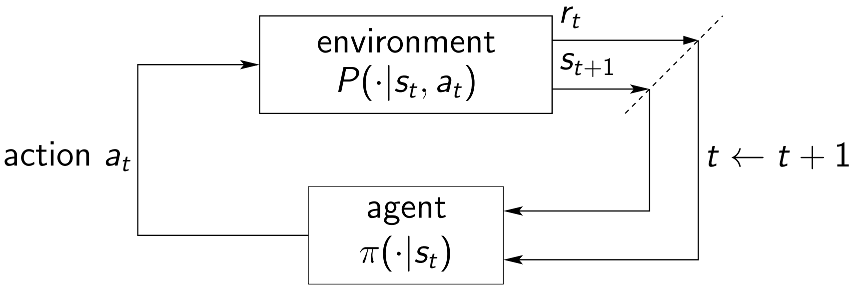
focus on optimization insights:
gradient dominance, preconditioning, mirror-descent, convex analysis

Markov decision process (MDP)



- $s_t \in \mathcal{S}$, finite state space (of the environment)
- $a_t \in \mathcal{A}$, finite action space (of the agent)
- $P(\cdot|s_t, a_t)$: transition probability function $P : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$
- $r_t = r(s_t, a_t)$: (random) reward encountered under (s_t, a_t)

Markov decision process (MDP)



- $s_t \in \mathcal{S}$, finite state space (of the environment)
- $a_t \in \mathcal{A}$, finite action space (of the agent)
- $P(\cdot|s_t, a_t)$: transition probability function $P : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$
- $r_t = r(s_t, a_t)$: (random) reward encountered under (s_t, a_t)

goal: choose policy π to maximize $\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)$ where $0 < \gamma < 1$

A simple example: salmon harvest

whether to fish salmons each year (sequential decision making)

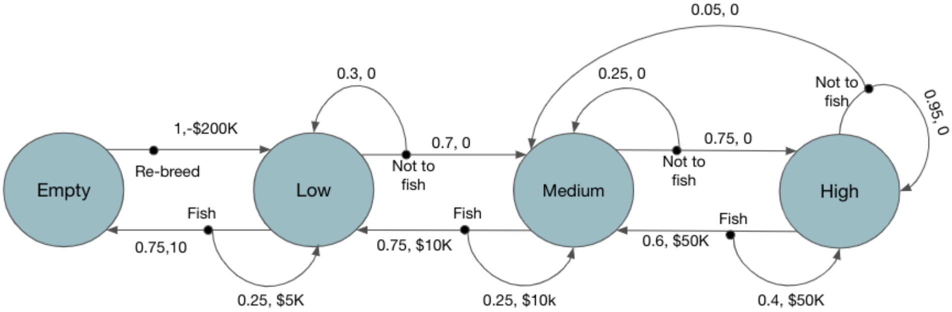


image credit: Somnath Banerjee (Towards Data Science, 2021)

- **state space** (salmon population): $\mathcal{S} = \{\text{empty, low, medium, high}\}$
- **action space**: $\mathcal{A} = \{\text{fish, not to fish, re-breed}\}$
- **transition probabilities** and **rewards** labeled in graph

Value function

- **value function** of discounted infinite-horizon MDP

$$V_s(\pi) := \mathbf{E}_{\substack{a_t \sim \pi(\cdot|s_t) \\ s_{t+1} \sim P(\cdot|s_t, a_t)}} \left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \,\middle|\, s_0 = s \right], \quad \forall\, s \in \mathcal{S}$$

Value function

- **value function** of discounted infinite-horizon MDP

$$V_s(\pi) := \mathbf{E}_{\substack{a_t \sim \pi(\cdot|s_t) \\ s_{t+1} \sim P(\cdot|s_t, a_t)}} \left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \,\middle|\, s_0 = s \right], \quad \forall\, s \in \mathcal{S}$$

- **vector form**: $V(\pi) \in \mathbf{R}^{|\mathcal{S}|}$

$$V(\pi) \;=\; \sum_{t=0}^{\infty} \gamma^t P(\pi)^t r(\pi) \;=\; \big(I - \gamma P(\pi)\big)^{-1} r(\pi)$$

both P and r linear in π (linear fractional)

- $P(\pi) \in \mathbf{R}^{|\mathcal{S}| \times |\mathcal{S}|}$ where $P_{s,s'}(\pi) = \sum_{a \in \mathcal{A}} \pi(a|s)\, P(s'|s,a)$
- $r(\pi) \in \mathbf{R}^{|\mathcal{S}|}$ where $r_s(\pi) = \sum_{a \in \mathcal{A}} \pi(a|s) r(s,a)$

Optimality

- **optimal values**

$$V_s^\star = \max_{\pi \in \Delta(\mathcal{A})^{|\mathcal{S}|}} V_s(\pi), \qquad \forall\, s \in \mathcal{S}$$

- **optimal policy**

$$\pi^\star(s) = \arg\max_{\pi \in \Delta(\mathcal{A})^{|\mathcal{S}|}} V_s(\pi), \qquad \forall\, s \in \mathcal{S}$$

exist stationary policy optimal for all s (e.g., [Puterman, 2005])

- **optimality conditions (Bellman equation)**

$$V_s = \max_{a \in \Delta\mathcal{A}} \left\{ r(s,a) + \gamma \sum_{s' \in \mathcal{S}} P(s'|s,a) V_{s'} \right\}, \qquad \forall\, s \in \mathcal{S}$$

Algorithms

- linear programming
- dynamic programming (DP)
  - value iteration
  - policy iteration
- reinforcement learning (& approximate DP)
  - Q-learning
  - TD learning
  - policy optimization (e.g., policy mirror descent)
  - actor-critic methods

many applications and growing importance in modern AI industry

Algorithms

- linear programming
- dynamic programming (DP)
  - value iteration
  - policy iteration
- reinforcement learning (& approximate DP)
  - Q-learning
  - TD learning
  - policy optimization (e.g., **policy mirror descent**)
  - actor-critic methods

many applications and growing importance in modern AI industry

Minimization formulation

- replace "reward" with "cost": $c(\pi) := -r(\pi) + \text{const}$

$$V(\pi) \;=\; \left(I - \gamma P(\pi)\right)^{-1} c(\pi)$$

- **expected cost** under *initial state distribution* $\rho \in \Delta(\mathcal{S})$

$$V_\rho(\pi) \;=\; \mathbf{E}_{s\sim\rho}\left[V_s(\pi)\right] \;=\; \rho^T \left(I - \gamma P(\pi)\right)^{-1} c(\pi)$$

Minimization formulation

- replace "reward" with "cost": $c(\pi) := -r(\pi) + \text{const}$

$$V(\pi) = \left(I - \gamma P(\pi)\right)^{-1} c(\pi)$$

- **expected cost** under *initial state distribution* $\rho \in \Delta(\mathcal{S})$

$$V_\rho(\pi) = \mathbf{E}_{s\sim\rho}\left[V_s(\pi)\right] = \rho^T\left(I - \gamma P(\pi)\right)^{-1} c(\pi)$$

- **minimizing expected cost**

$$\underset{\pi\in\Delta(\mathcal{A})^{|\mathcal{S}|}}{\text{minimize}} \quad V_\rho(\pi)$$

$V_\rho(\pi)$ **non-convex in general** (but has favorable structure)

Q-function

- **state-action cost-to-go**

$$Q_{s,a}(\pi) := \mathbf{E}_{\substack{a_t \sim \pi(\cdot|s_t) \\ s_{t+1} \sim P(\cdot|s_t, a_t)}} \left[\sum_{t=0}^{\infty} \gamma^t c(s_t, a_t) \,\middle|\, s_0 = s, \ a_0 = a \right]$$

Q-function

- **state-action cost-to-go**

$$Q_{s,a}(\pi) := \mathbf{E}_{\substack{a_t \sim \pi(\cdot|s_t) \\ s_{t+1} \sim P(\cdot|s_t, a_t)}} \left[\sum_{t=0}^{\infty} \gamma^t c(s_t, a_t) \,\middle|\, s_0 = s, \;\; a_0 = a \right]$$

alternatively

$$Q_{s,a}(\pi) = c(s, a) + \gamma \sum_{s' \in \mathcal{S}} P(s'|s, a) V_{s'}(\pi)$$

Q-function

- **state-action cost-to-go**

$$Q_{s,a}(\pi) := \mathbf{E}_{\substack{a_t \sim \pi(\cdot|s_t) \\ s_{t+1} \sim P(\cdot|s_t, a_t)}} \left[\sum_{t=0}^{\infty} \gamma^t c(s_t, a_t) \,\middle|\, s_0 = s,\ a_0 = a \right]$$

alternatively

$$Q_{s,a}(\pi) = c(s,a) + \gamma \sum_{s' \in \mathcal{S}} P(s'|s,a) V_{s'}(\pi)$$

- useful relations (for all $s \in \mathcal{S}$):

$$V_s(\pi) = \mathbf{E}_{a \sim \pi_s}[Q_{s,a}(\pi)] = \langle Q_s(\pi), \pi_s \rangle$$

(**notation:** moving s, a as subscripts, emphasizing function of π)

Policy gradient

- **policy gradient**: weighted Q-function [Sutton et al., 1999]

$$\nabla_s V_\rho(\pi) = \frac{\partial V_\rho(\pi)}{\partial \pi_s} = \frac{1}{1-\gamma} d_{\rho,s}(\pi) Q_s(\pi)$$

where $\pi_s = [\pi_{s,a}]_{a \in \mathcal{A}}$ and $Q_s(\pi) = [Q_{s,a}]_{a \in \mathcal{A}}$

Policy gradient

- **policy gradient**: weighted Q-function [Sutton et al., 1999]

$$\nabla_s V_\rho(\pi) = \frac{\partial V_\rho(\pi)}{\partial \pi_s} = \frac{1}{1-\gamma} d_{\rho,s}(\pi) Q_s(\pi)$$

where $\pi_s = [\pi_{s,a}]_{a\in\mathcal{A}}$ and $Q_s(\pi) = [Q_{s,a}]_{a\in\mathcal{A}}$

- $d_\rho(\pi) \in \Delta(\mathcal{S})$: **discounted state-visitation distribution**

$$d_{\rho,s}(\pi) = (1-\gamma)\sum_{t=0}^{\infty} \gamma^t \, \mathbf{Prob}^\pi\big(s_t = s \,|\, s_0 \sim \rho\big)$$

  – simple lower bound: $d_{\rho,s}(\pi) \geq (1-\gamma)\rho_s$ for all π

Hint of structure

- **performance difference lemma** [Kakade and Langford, 2002]

$$V_\rho(\pi) - V_\rho(\tilde{\pi}) = \frac{1}{1-\gamma} \sum_s d_{\rho,s}(\pi) \left\langle Q_s(\tilde{\pi}),\ \pi_s - \tilde{\pi}_s \right\rangle$$

Hint of structure

- **performance difference lemma** [Kakade and Langford, 2002]

$$V_\rho(\pi) - V_\rho(\tilde{\pi}) = \frac{1}{1-\gamma} \sum_s d_{\rho,s}(\pi) \left\langle Q_s(\tilde{\pi}),\ \pi_s - \tilde{\pi}_s \right\rangle$$

- compare with Taylor expansion

$$V_\rho(\pi) - V_\rho(\tilde{\pi}) = \langle \nabla V_\rho(\tilde{\pi}), \pi - \tilde{\pi} \rangle + o(\|\pi - \tilde{\pi}\|)$$

$$\text{where} \quad \nabla_s V_\rho(\tilde{\pi}) = \frac{1}{1-\gamma} d_{\rho,s}(\tilde{\pi}) Q_s(\tilde{\pi})$$

Hint of structure

- **performance difference lemma** [Kakade and Langford, 2002]

$$V_\rho(\pi) - V_\rho(\tilde{\pi}) = \frac{1}{1-\gamma} \sum_s d_{\rho,s}(\pi) \left\langle Q_s(\tilde{\pi}),\ \pi_s - \tilde{\pi}_s \right\rangle$$

- compare with Taylor expansion

$$V_\rho(\pi) - V_\rho(\tilde{\pi}) = \langle \nabla V_\rho(\tilde{\pi}), \pi - \tilde{\pi} \rangle + o(\|\pi - \tilde{\pi}\|)$$

$$\text{where} \quad \nabla_s V_\rho(\tilde{\pi}) = \frac{1}{1-\gamma} d_{\rho,s}(\tilde{\pi}) Q_s(\tilde{\pi})$$

- **implications**
  - gradient dominance, convergence to global optima
  - larger step sizes (increasing geometrically) for fast convergence

Projected policy gradient method

optimization over Cartesian product of probability simplexes

$$\underset{\pi \in \Delta(\mathcal{A})^{|\mathcal{S}|}}{\text{minimize}} \quad V_\rho(\pi)$$

projection done for each s separately

$$\pi_s^{k+1} = \mathbf{proj}_{\Delta(\mathcal{A})} \left(\pi_s^k - \eta_k \nabla_s V_\rho(\pi^k)\right), \qquad s \in \mathcal{S}$$

- (weak) gradient dominance (by PDL) [Agarwal et al., 2021]
- convergence to global optima at $\mathcal{O}(1/k)$ rate [X., 2022]
- large constant depending on $|\mathcal{A}|$, $|\mathcal{S}|$, and $(1-\gamma)^5$

Projected policy gradient method

optimization over Cartesian product of probability simplexes

$$\underset{\pi \in \Delta(\mathcal{A})^{|\mathcal{S}|}}{\text{minimize}} \quad V_\rho(\pi)$$

projection done for each s separately

$$\pi_s^{k+1} = \mathbf{proj}_{\Delta(\mathcal{A})}\left(\pi_s^k - \eta_k \nabla_s V_\rho(\pi^k)\right), \qquad s \in \mathcal{S}$$

- (weak) gradient dominance (by PDL) [Agarwal et al., 2021]
- convergence to global optima at $\mathcal{O}(1/k)$ rate [X., 2022]
- large constant depending on $|\mathcal{A}|$, $|\mathcal{S}|$, and $(1-\gamma)^5$

key to improve: preconditioning & exploiting underlying geometry

Outline

- Markov decision process (MDP)
- **policy mirror descent (PMD) method** (tabular case)
  - mirror descent and convergence rate
  - preconditioning for MDP
  - sublinear and linear convergence
- PMD with dual function approximation

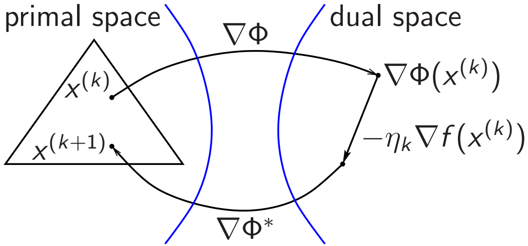
contents based on [Agarwal et al., 2021, Lan, 2021, Xiao, 2022]

Mirror descent

- convex optimization

$$\underset{x \in \mathcal{C}}{\text{minimize}} \quad f(x)$$

  - $\mathcal{C}$: closed convex set (nonempty interior)
  - f: convex over $\mathcal{C}$



- **mirror descent** [Nemirovski and Yudin, 1983]

$$x^{(k+1)} = \nabla\Phi^*\left(\nabla\Phi(x^{(k)}) - \eta_k \nabla f(x^{(k)})\right)$$

  - Φ: strictly convex and continuously differentiable over $\mathcal{C}$
  - Φ^*: conjugate function $\Phi^*(x^*) = \sup_{x \in \mathcal{C}} \{\langle x^*, x\rangle - \Phi(x)\}$

Mirror descent: primal form

- Bregman divergence

$$D_\Phi(x, y) = \Phi(x) - \Phi(y) - \langle \nabla \Phi(y), x - y \rangle$$

- nonlinear projected subgradient method [Beck and Teboulle, 2003]

$$x^{(k+1)} = \arg\min_{x \in \mathcal{C}} \left\{ \langle \nabla f(x^{(k)}),\ x \rangle + \frac{1}{\eta_k} D_\Phi(x, x^{(k)}) \right\}$$

  - equivalent to mirror descent (with some subtle conditions)
  - no explicit dependence on Φ^*, $\nabla \Phi^*$ (thus "primal" form)
  - $O(1/\sqrt{k})$ convergence rate in general convex setting

Mirror descent: examples

- **Euclidean geometry:** $\Phi(x) = \frac{1}{2}\|x\|_2^2$ and $D_\Phi(x, y) = \frac{1}{2}\|x - y\|_2^2$

$$x^{(k+1)} = \mathbf{proj}_{\mathcal{C}}\left(x^{(k)} - \eta_k \nabla f(x^{(k)})\right)$$

Mirror descent: examples

- **Euclidean geometry:** $\Phi(x) = \frac{1}{2}\|x\|_2^2$ and $D_\Phi(x,y) = \frac{1}{2}\|x-y\|_2^2$

$$x^{(k+1)} = \mathbf{proj}_{\mathcal{C}}\big(x^{(k)} - \eta_k \nabla f(x^{(k)})\big)$$

- **simplex:** $\mathcal{C} = \Delta$ **(ubiquitous in ML, especially RL)**
  - $\Phi(x) = \sum_i x_i \log x_i$ for $x \in \Delta$ and ∞ otherwise
  - $D_\Phi(x,y) = D_{KL}(x\|y) = \sum_i x_i \log(x_i/y_i)$
  - algorithm

$$x_i^{(k+1)} = \frac{x_i^{(k)} \exp\big(-\eta_k \nabla_i f(x^{(k)})\big)}{\sum_j x_j^{(k)} \exp\big(-\eta_k \nabla_j f(x^{(k)})\big)}, \qquad i = 1, \dots, n$$

Mirror descent: examples

- **Euclidean geometry:** $\Phi(x) = \frac{1}{2}\|x\|_2^2$ and $D_\Phi(x,y) = \frac{1}{2}\|x-y\|_2^2$

$$x^{(k+1)} = \mathbf{proj}_{\mathcal{C}}\big(x^{(k)} - \eta_k \nabla f(x^{(k)})\big)$$

- **simplex:** $\mathcal{C} = \Delta$ **(ubiquitous in ML, especially RL)**
  - $\Phi(x) = \sum_i x_i \log x_i$ for $x \in \Delta$ and ∞ otherwise
  - $D_\Phi(x,y) = D_{KL}(x\|y) = \sum_i x_i \log(x_i/y_i)$
  - algorithm

$$x_i^{(k+1)} = \frac{x_i^{(k)} \exp\big(-\eta_k \nabla_i f(x^{(k)})\big)}{\sum_j x_j^{(k)} \exp\big(-\eta_k \nabla_j f(x^{(k)})\big)}, \qquad i = 1, \ldots, n$$

(**subtleties:** Δ has empty interior, $\nabla\Phi^* \neq (\nabla\Phi)^{-1}$, etc.)

Mirror descent: convergence rate

- convex optimization: $\text{minimize}_{x \in \mathcal{C}} \;\; f(x)$
- mirror descent (primal form)

$$x^{(k+1)} = \arg\min_{x \in \mathcal{C}} \left\{ \langle \nabla f(x^{(k)}), x \rangle + \frac{1}{\eta_k} D_\Phi(x, x^{(k)}) \right\}$$

  – $\mathcal{O}(1/\sqrt{k})$ convergence rate [Beck and Teboulle, 2003]

- fast convergence rate under **relative smoothness**
  – $O(1/k)$ convergence rate [Birnbaum et al., 2011]; independent recent work [Bauschke et al., 2017, Lu et al., 2018]
  – linear rate under **relative strong convexity** [Lu et al., 2018]

Relative smoothness and strong convexity

- **relative smoothness** with parameter β

$$f(x) \leq f(y) + \langle \nabla f(y), x - y \rangle + \beta\, D_\Phi(x, y)$$

equivalently

$$D_f(x, y) \leq \beta\, D_\Phi(x, y)$$

- **relative strong convexity** with parameter α

$$D_f(x, y) \geq \alpha\, D_\Phi(x, y)$$

- relatively smooth and strongly convex ($\alpha \leq \beta$)

$$\alpha\, D_\Phi(x, y) \leq D_f(x, y) \leq \beta\, D_\Phi(x, y)$$

Analysis of mirror descent I

- **three-point descent lemma** [Chen and Teboulle, 1993]
  if φ convex and
  $$x^+ = \arg\min_{u \in \mathcal{C}} \left\{ \varphi(u) + D_\Phi(u, x) \right\}$$
  then for any $u \in \mathcal{C}$,
  $$\varphi(x^+) + D_\Phi(x^+, x) \le \varphi(u) + D_\Phi(u, x) - D_\Phi(u, x^+)$$

Analysis of mirror descent I

- **three-point descent lemma** [Chen and Teboulle, 1993]
  if φ convex and
  $$x^+ = \arg\min_{u \in \mathcal{C}} \left\{ \varphi(u) + D_\Phi(u, x) \right\}$$
  then for any $u \in \mathcal{C}$,
  $$\varphi(x^+) + D_\Phi(x^+, x) \le \varphi(u) + D_\Phi(u, x) - D_\Phi(u, x^+)$$

- applying to MD update with $\varphi(\cdot) = \langle \nabla f(x^{(k)}), \cdot \rangle$
  $$\left\langle \nabla f(x^{(k)}), x^{(k+1)} - u \right\rangle + \frac{1}{\eta_k} D_\Phi(x^{(k+1)}, x^{(k)}) \le \frac{1}{\eta_k} D_\Phi(u, x^{(k)}) - \frac{1}{\eta_k} D_\Phi(u, x^{(k+1)})$$

Analysis of mirror descent II

by subtracting and adding $\langle \nabla f(x^{(k)}), x^{(k)} \rangle$,

$$\underbrace{\langle \nabla f(x^{(k)}), x^{(k+1)} - x^{(k)} \rangle + \frac{1}{\eta_k} D_\Phi(x^{(k+1)}, x^{(k)})}_{\mathbf{A}} + \underbrace{\langle \nabla f(x^{(k)}), x^{(k)} - u \rangle}_{\mathbf{B}}$$

$$\leq \frac{1}{\eta_k} D_\Phi(u, x^{(k)}) - \frac{1}{\eta_k} D_\Phi(u, x^{(k+1)})$$

Analysis of mirror descent II

by subtracting and adding $\langle \nabla f(x^{(k)}), x^{(k)} \rangle$,

$$\underbrace{\langle \nabla f(x^{(k)}), x^{(k+1)} - x^{(k)} \rangle + \frac{1}{\eta_k} D_\Phi(x^{(k+1)}, x^{(k)})}_{\mathbf{A}} + \underbrace{\langle \nabla f(x^{(k)}), x^{(k)} - u \rangle}_{\mathbf{B}}$$
$$\leq \frac{1}{\eta_k} D_\Phi(u, x^{(k)}) - \frac{1}{\eta_k} D_\Phi(u, x^{(k+1)})$$

- **relative smoothness** ($\frac{1}{\eta_k} \geq \beta$): $\mathbf{A} \geq f(x^{(k+1)}) - f(x^{(k)})$
- **relative strong convexity**: $\mathbf{B} \geq f(x^{(k)}) - f(u) + \alpha D_\Phi(u, x^{(k)})$

combining together,

$$f(x^{(k+1)}) - f(u) \leq \left(\frac{1}{\eta_k} - \alpha \right) D_\Phi(u, x^{(k)}) - \frac{1}{\eta_k} D_\Phi(u, x^{(k+1)})$$

Analysis of mirror descent III

using constant step size $\eta_k = 1/\beta$:

$$f(x^{(k+1)}) - f(u) \leq (\beta - \alpha) D_\Phi(u, x^{(k)}) - \beta D_\Phi(u, x^{(k+1)})$$

Analysis of mirror descent III

using constant step size $\eta_k = 1/\beta$:

$$f(x^{(k+1)}) - f(u) \le (\beta - \alpha) D_\Phi(u, x^{(k)}) - \beta D_\Phi(u, x^{(k+1)})$$

rate of convergence: [Lu et al., 2018]

- if $\alpha = 0$, then sublinear convergence

$$f(x^{(k)}) - f(u) \le \frac{\beta}{k} D_\Phi(u, x^0)$$

- if $\alpha > 0$, then linear convergence

$$f(x^{(k)}) - f(u) \le \left(1 - \frac{\alpha}{\beta}\right)^k D_\Phi(u, x^0)$$

Policy mirror descent (PMD)

- weighted divergence: for arbitrary $\mu \in \Delta(\mathcal{S})$

$$D_\mu(\pi, \pi') = \mathbf{E}_{s\sim\mu}\left[D(\pi_s, \pi'_s)\right] = \sum_{s\in\mathcal{S}} \mu_s D(\pi_s, \pi'_s)$$

- (preconditioned) **policy mirror-descent**

$$\pi^{k+1} = \arg\min_{\pi\in\Delta(\mathcal{A})^{|\mathcal{S}|}} \left\{ \eta_k \left\langle \nabla V_\rho(\pi^k),\ \pi \right\rangle + \frac{1}{1-\gamma} D_{d_\rho(\pi^k)}(\pi, \pi^k) \right\}$$

Policy mirror descent (PMD)

- weighted divergence: for arbitrary $\mu \in \Delta(\mathcal{S})$

$$D_\mu(\pi, \pi') = \mathbf{E}_{s\sim\mu}\big[D(\pi_s, \pi'_s)\big] = \textstyle\sum_{s\in\mathcal{S}} \mu_s D(\pi_s, \pi'_s)$$

- (preconditioned) **policy mirror-descent**

$$\pi^{k+1} = \arg\min_{\pi\in\Delta(\mathcal{A})^{|\mathcal{S}|}} \Big\{ \eta_k \big\langle \nabla V_\rho(\pi^k),\ \pi \big\rangle + \tfrac{1}{1-\gamma} D_{d_\rho(\pi^k)}(\pi, \pi^k) \Big\}$$

- plug in policy gradient $\nabla_s V_\rho(\pi^k) = \frac{1}{1-\gamma} d_{\rho,s}(\pi^k) Q_s(\pi^k)$

$$\boxed{\pi_s^{k+1} = \arg\min_{\pi_s\in\Delta(\mathcal{A})} \Big\{ \eta_k \big\langle Q_s(\pi^k),\ \pi_s \big\rangle + D(\pi_s, \pi_s^k) \Big\}, \quad \forall s \in \mathcal{S}}$$

PMD with exact Q-function

- **closed-form update**

$$\pi_{s,a}^{(k+1)} = \frac{\pi_{s,a}^{(k)} \exp\left(-\eta_k Q_{s,a}(\pi^{(k)})\right)}{\sum_{a' \in \mathcal{A}} \pi_{s,a'}^{(k)} \exp\left(-\eta_k Q_{s,a'}(\pi^{(k)})\right)}, \qquad \forall (s,a) \in \mathcal{S} \times \mathcal{A}$$

PMD with exact Q-function

- **closed-form update**

$$\pi_{s,a}^{(k+1)} = \frac{\pi_{s,a}^{(k)} \exp\left(-\eta_k Q_{s,a}(\pi^{(k)})\right)}{\sum_{a' \in \mathcal{A}} \pi_{s,a'}^{(k)} \exp\left(-\eta_k Q_{s,a'}(\pi^{(k)})\right)}, \qquad \forall (s,a) \in \mathcal{S} \times \mathcal{A}$$

- **convergence to global optima**
  - Shani et al. [2020]: $O(1/\sqrt{k})$ convergence rate
  - Agarwal et al. [2021]: $O(1/k)$ rate
  - Lan [2021]: $O(1/k)$ and linear convergence (diminishing regu.)
  - Khodadadian et al. [2021]: linear rate (adaptive stepsize)
  - X. [2022]: linear rate with geometrically increasing stepsize

PMD with exact Q-function

- **closed-form update**

$$\pi_{s,a}^{(k+1)} = \frac{\pi_{s,a}^{(k)} \exp\left(-\eta_k Q_{s,a}(\pi^{(k)})\right)}{\sum_{a' \in \mathcal{A}} \pi_{s,a'}^{(k)} \exp\left(-\eta_k Q_{s,a'}(\pi^{(k)})\right)}, \qquad \forall (s,a) \in \mathcal{S} \times \mathcal{A}$$

- **convergence to global optima**
  - Shani et al. [2020]: $O(1/\sqrt{k})$ convergence rate
  - Agarwal et al. [2021]: $O(1/k)$ rate
  - Lan [2021]: $O(1/k)$ and linear convergence (diminishing regu.)
  - Khodadadian et al. [2021]: linear rate (adaptive stepsize)
  - X. [2022]: linear rate with geometrically increasing stepsize

non-convex, no relative smoothness or strong convexity

Analysis of PMD I

$$\pi_s^{k+1} = \arg\min_{p \in \Delta(\mathcal{A})} \left\{ \eta_k \langle Q_s(\pi^k), p \rangle + D_{KL}(p, \pi_s^k) \right\}, \quad \forall s \in \mathcal{S}$$

- **three-point descent lemma** [Chen and Teboulle, 1993]:
  if φ convex and

  $$x^+ = \arg\min_{u \in \mathcal{C}} \{ \varphi(u) + D(u, x) \}$$

  then for any $u \in \mathcal{C}$,

  $$\varphi(x^+) + D(x^+, x) \le \varphi(u) + D(u, x) - D(u, x^+)$$

Analysis of PMD I

$$\pi_s^{k+1} = \arg\min_{p \in \Delta(\mathcal{A})} \left\{ \eta_k \left\langle Q_s(\pi^k), p \right\rangle + D_{KL}(p, \pi_s^k) \right\}, \quad \forall s \in \mathcal{S}$$

- **three-point descent lemma** [Chen and Teboulle, 1993]:
  if φ convex and
  $$x^+ = \arg\min_{u \in \mathcal{C}} \left\{ \varphi(u) + D(u, x) \right\}$$
  then for any $u \in \mathcal{C}$,
  $$\varphi(x^+) + D(x^+, x) \leq \varphi(u) + D(u, x) - D(u, x^+)$$
- applying to PMD update with $\varphi(\cdot) = \eta_k \langle Q_s(\pi^k), \cdot \rangle$,
  $$\eta_k \langle Q_s(\pi^k), \pi_s^{k+1} - p \rangle + D(\pi_s^{k+1}, \pi_s^k) \leq D(p, \pi_s^k) - D(p, \pi_s^{k+1})$$

Analysis of PMD II

$$\langle Q_s(\pi^k), \pi_s^{k+1} - p \rangle + \frac{1}{\eta_k} D(\pi_s^{k+1}, \pi_s^k) \le \frac{1}{\eta_k} D(p, \pi_s^k) - \frac{1}{\eta_k} D(p, \pi_s^{k+1})$$

Analysis of PMD II

$$\langle Q_s(\pi^k), \pi_s^{k+1} - p \rangle + \frac{1}{\eta_k} D(\pi_s^{k+1}, \pi_s^k) \le \frac{1}{\eta_k} D(p, \pi_s^k) - \frac{1}{\eta_k} D(p, \pi_s^{k+1})$$

descent property

- letting $p = \pi_s^k$ yields

$$\langle Q_s(\pi^k), \pi_s^{k+1} - \pi_s^k \rangle \le 0, \qquad \forall s \in \mathcal{S}$$

Analysis of PMD II

$$\langle Q_s(\pi^k), \pi_s^{k+1} - p\rangle + \tfrac{1}{\eta_k} D(\pi_s^{k+1}, \pi_s^k) \le \tfrac{1}{\eta_k} D(p, \pi_s^k) - \tfrac{1}{\eta_k} D(p, \pi_s^{k+1})$$

descent property

- letting $p = \pi_s^k$ yields

$$\langle Q_s(\pi^k), \pi_s^{k+1} - \pi_s^k\rangle \le 0, \qquad \forall\, s \in \mathcal{S}$$

- $V_s(\pi^{k+1}) \le V_s(\pi^k)$ for all $s \in \mathcal{S}$ because of PDL:

$$V_\rho(\pi^{k+1}) - V_\rho(\pi^k) = \frac{1}{1-\gamma} \mathbf{E}_{s\sim d_\rho(\pi^{k+1})} \big\langle Q_s(\pi^k), \pi_s^{k+1} - \pi_s^k \big\rangle \le 0$$

independent of step size (like linear or concave function)

Analysis of PMD III

- apply **three-point descent lemma** with $u = \pi_s^\star$

$$\underbrace{\langle Q_s(\pi^k), \pi_s^{k+1} - \pi_s^k \rangle}_{\mathbf{A}} + \underbrace{\langle Q_s(\pi^k), \pi_s^k - \pi_s^\star \rangle}_{\mathbf{B}} \leq \frac{1}{\eta_k} D(\pi_s^\star, \pi_s^k) - \frac{1}{\eta_k} D(\pi_s^\star, \pi_s^{k+1})$$

$\frac{1}{\eta_k} D(\pi_s^{k+1}, \pi_s^k)$ ignored (but was necessary with relative smoothness)

Analysis of PMD III

- apply **three-point descent lemma** with $u = \pi_s^\star$

$$\underbrace{\langle Q_s(\pi^k), \pi_s^{k+1} - \pi_s^k \rangle}_{\mathbf{A}} + \underbrace{\langle Q_s(\pi^k), \pi_s^k - \pi_s^\star \rangle}_{\mathbf{B}} \le \frac{1}{\eta_k} D(\pi_s^\star, \pi_s^k) - \frac{1}{\eta_k} D(\pi_s^\star, \pi_s^{k+1})$$

$\frac{1}{\eta_k} D(\pi_s^{k+1}, \pi_s^k)$ ignored (but was necessary with relative smoothness)

- taking expectation w.r.t. $d_\rho(\pi^\star)$ and **apply PDL twice**:

$$\mathbf{E}_{s \sim d_\rho(\pi^\star)}[\mathbf{A}] \ge \left\| \frac{d_\rho(\pi^\star)}{d_\rho(\pi^{k+1})} \right\|_\infty (1-\gamma)\big(V_\rho(\pi^{k+1}) - V_\rho(\pi^k)\big)$$

$$\mathbf{E}_{s \sim d_\rho(\pi^\star)}[\mathbf{B}] = (1-\gamma)\big(V_\rho(\pi^k) - V_\rho(\pi^\star)\big)$$

Convergence rate of PMD

- arbitrary constant step size $\eta_k = \eta$

$$V_\rho(\pi^k) - V_\rho(\pi^\star) \le \frac{1}{k(1-\gamma)}\left(\delta_0 + \frac{1}{\eta}D_0\right)$$

  - holds for any $\rho \in \Delta(\mathcal{S})$

Convergence rate of PMD

- arbitrary constant step size $\eta_k = \eta$

$$V_\rho(\pi^k) - V_\rho(\pi^\star) \le \frac{1}{k(1-\gamma)} \left(\delta_0 + \frac{1}{\eta} D_0 \right)$$

  – holds for any $\rho \in \Delta(\mathcal{S})$

- increasing step size $\eta_{k+1} = \eta_k/\gamma$ [X. 2022]

$$V_{\rho^\star}(\pi^k) - V_{\rho^\star}(\pi^\star) \le \gamma^k \left(\delta_0 + \frac{1}{\gamma \eta_0} D_0 \right)$$

  – for $\rho \neq \rho^\star$: linear convergence with slower rate
  – as $\eta_k \to \infty$, PMD becomes **policy iteration**

Outline

- Markov decision process (MDP)
- policy mirror descent (PMD) method (tabular case)
- **PMD with dual function approximation**
  - challenge with function approximation
  - affine-restricted Legendre functions and Bregman divergence
  - dual approximation policy optimization (DAPO)

joint work with **Zhihan Xiong** and **Maryam Fazel** [Xiong et al., 2024]

Parametrizations of policy

- **softmax parametrization** (ensuring $\pi_s^\theta \in \Delta(\mathcal{A})$)

$$\pi_{s,a}^\theta = \frac{\exp(f_{s,a}(\theta))}{\sum_{a'} \exp(f_{s,a'}(\theta))}, \qquad (s,a) \in \mathcal{S} \times \mathcal{A}$$

  - softmax tabular policy class: $f_{s,a}(\theta) = \theta_{s,a}$ and $\theta \in \mathbf{R}^{|\mathcal{S}|\times|\mathcal{A}|}$
  - log-linear policy class: $f_{s,a}(\theta) = \langle \theta, \phi_{s,a} \rangle$ and $\theta \in \mathbf{R}^p$
  - neural policy class: $f_{s,a}(\theta) = \mathrm{network}(\theta, \phi_{s,a})$ and $\theta \in \mathbf{R}^p$

  last two classes may be incomplete (usually $p \ll |\mathcal{S}|\,|\mathcal{A}|$)

Parametrizations of policy

- **softmax parametrization** (ensuring $\pi_s^\theta \in \Delta(\mathcal{A})$)

$$\pi_{s,a}^\theta = \frac{\exp(f_{s,a}(\theta))}{\sum_{a'} \exp(f_{s,a'}(\theta))}, \qquad (s,a) \in \mathcal{S} \times \mathcal{A}$$

  - softmax tabular policy class: $f_{s,a}(\theta) = \theta_{s,a}$ and $\theta \in \mathbf{R}^{|\mathcal{S}|\times|\mathcal{A}|}$
  - log-linear policy class: $f_{s,a}(\theta) = \langle \theta, \phi_{s,a} \rangle$ and $\theta \in \mathbf{R}^p$
  - neural policy class: $f_{s,a}(\theta) = \mathrm{network}(\theta, \phi_{s,a})$ and $\theta \in \mathbf{R}^p$

  last two classes may be incomplete (usually $p \ll |\mathcal{S}|\,|\mathcal{A}|$)

- **notations**
  - $\pi^{(k)}$ means $\pi^{\theta^{(k)}}$, $Q_s^{(k)}$ means $Q_s(\pi^{(k)})$ means $Q_s(\pi^{\theta^{(k)}})$

Policy gradient under parametrization

- can directly use SGD to minimize $V_\rho(\pi^\theta)$ over θ
  - unbiased estimate of stochastic gradient (e.g., REINFORCE)
  - not taking full advantage of MDP structure
- structured estimate of policy gradient [Sutton et al., 1999]

$$\nabla_\theta V_\rho(\pi^\theta) = \frac{1}{1-\gamma} \sum_{s' \in \mathcal{S}} d_{\rho,s'}(\pi^\theta) \sum_{a' \in \mathcal{A}} Q^\theta_{s',a'} \nabla_\theta \pi^\theta_{s',a'}$$

- lack preconditioning (as in PMD), thus slower convergence

Policy gradient under parametrization

- can directly use SGD to minimize $V_\rho(\pi^\theta)$ over θ
  - unbiased estimate of stochastic gradient (e.g., REINFORCE)
  - not taking full advantage of MDP structure
- structured estimate of policy gradient [Sutton et al., 1999]

$$\nabla_\theta V_\rho(\pi^\theta) = \frac{1}{1-\gamma} \sum_{s' \in \mathcal{S}} d_{\rho,s'}(\pi^\theta) \sum_{a' \in \mathcal{A}} Q^\theta_{s',a'} \nabla_\theta \pi^\theta_{s',a'}$$

- lack preconditioning (as in PMD), thus slower convergence

this work: adapting PMD with function approximation

Challenge with function approximation

- recall tabular case

$$\pi_s^{(k+1)} = \arg\min_{p \in \Delta} \left\{ \eta_k \left\langle Q^{(k)}, p \right\rangle + D_\Phi\left(p, \pi_s^{(k)}\right) \right\}$$

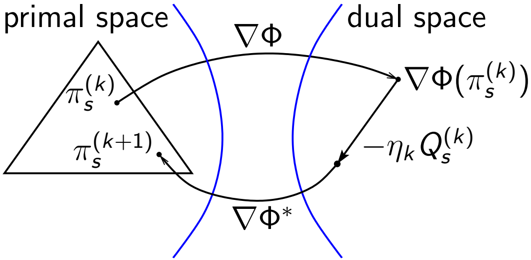
key to use three-point descent lemma: **convex in** p

- PMD with function approximation

$$\theta^{(k+1)} = \arg\min_{\theta \in \Theta} \; \mathbf{E}_{s \sim d_\rho^{(k)}} \left[\left\langle \widehat{Q}_s^{(k)}, \; \pi_s^\theta \right\rangle + D_\Phi\left(\pi_s^\theta, \pi_s^{(k)}\right) \right]$$

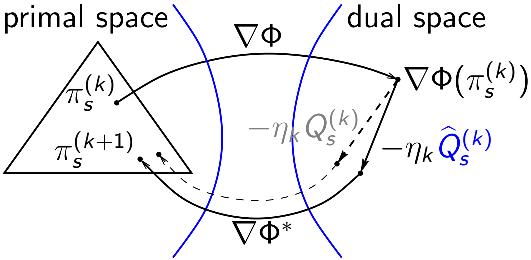
cannot use three-point descent lemma due to **nonconvexity**

PMD in primal-dual form



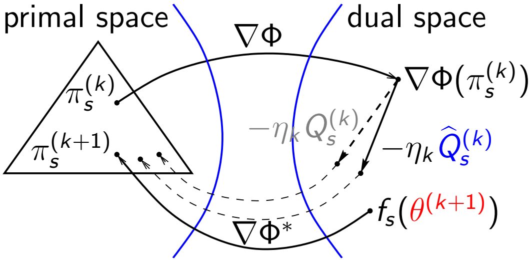
$$\pi_s^{(k+1)} = \nabla\Phi^*\Big(\nabla\Phi\big(\pi^{(k)}\big) - \eta_k Q_s^{(k)}\Big)$$

PMD in primal-dual form



$$\pi_s^{(k+1)} = \nabla\Phi^*\Big(\nabla\Phi\big(\pi^{(k)}\big) - \eta_k \widehat{Q}_s^{(k)}\Big)$$

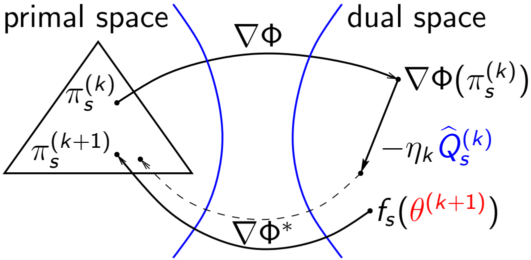
PMD in primal-dual form



$$f_s(\theta^{(k+1)}) \approx \nabla\Phi(\pi^{(k)}) - \eta_k \widehat{Q}_s^{(k)}$$

$$\pi_s^{(k+1)} = \nabla\Phi^*\big(f_s(\theta^{(k+1)})\big)$$

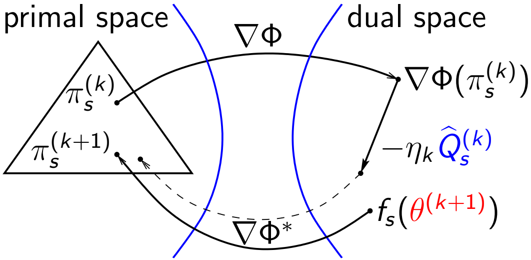
PMD in primal-dual form



$$f_s(\theta^{(k+1)}) \approx \nabla\Phi(\pi^{(k)}) - \eta_k \widehat{Q}_s^{(k)}$$

$$\pi_s^{(k+1)} = \nabla\Phi^*\big(f_s(\theta^{(k+1)})\big)$$

PMD in primal-dual form



$$f_s(\theta^{(k+1)}) \approx \nabla\Phi(\pi^{(k)}) - \eta_k \widehat{Q}_s^{(k)}$$

$$\pi_s^{(k+1)} = \nabla\Phi^*\big(f_s(\theta^{(k+1)})\big)$$

- technicalities of mirror map on simplex (empty interior)
- how to measure approximation error in dual space?

Some convex analysis

Legendre function: proper, closed convex function ϕ satisfying

- $\mathcal{D} := \operatorname{int}(\operatorname{dom}\phi)$ nonempty
- ϕ differentiable and strictly convex on $\mathcal{D}$
- $\lim_{n\to\infty} \|\nabla\phi(x_n)\| = \infty$ if $\{x_n\} \in \mathcal{D}$ converges to boundary of $\mathcal{D}$

Some convex analysis

Legendre function: proper, closed convex function ϕ satisfying

- $\mathcal{D} := \operatorname{int}(\operatorname{dom}\phi)$ nonempty
- ϕ differentiable and strictly convex on $\mathcal{D}$
- $\lim_{n\to\infty} \|\nabla\phi(x_n)\| = \infty$ if $\{x_n\} \in \mathcal{D}$ converges to boundary of $\mathcal{D}$

conjugate: $\phi^*(x^*) = \sup_{x \in \operatorname{dom}\phi} \{\langle x^*, x\rangle - \phi(x)\}$

property: $\nabla\phi^* = (\nabla\phi)^{-1}$, i.e.,

$$\nabla\phi^*(\nabla\phi(x)) = x, \qquad \nabla\phi(\nabla\phi^*(x^*)) = x^*$$

Some convex analysis

Legendre function: proper, closed convex function ϕ satisfying

- $\mathcal{D} := \text{int}\,(\text{dom}\phi)$ nonempty
- ϕ differentiable and strictly convex on $\mathcal{D}$
- $\lim_{n\to\infty} \|\nabla\phi(x_n)\| = \infty$ if $\{x_n\} \in \mathcal{D}$ converges to boundary of $\mathcal{D}$

conjugate: $\phi^*(x^*) = \sup_{x \in \text{dom}\phi} \{\langle x^*, x\rangle - \phi(x)\}$

property: $\nabla\phi^* = (\nabla\phi)^{-1}$, i.e.,

$$\nabla\phi^*(\nabla\phi(x)) = x, \qquad \nabla\phi(\nabla\phi^*(x^*)) = x^*$$

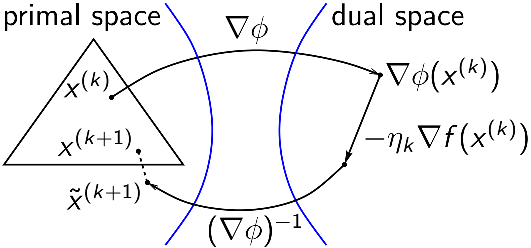
example: $\phi(x) = \begin{cases} \sum_i x_i \log(x_i) & \text{if } x \in \mathbf{R}^n_+ \\ \infty & \text{otherwise} \end{cases}$

Mirror descent with Legendre function

negative entropy on $\mathbf{R}^n_+$: range $(\nabla\phi)^{-1} = \mathbf{R}^n_+$, thus **need projection**

Mirror descent with Legendre function

negative entropy on $\mathbf{R}^n_+$: range $(\nabla\phi)^{-1} = \mathbf{R}^n_+$, thus **need projection**

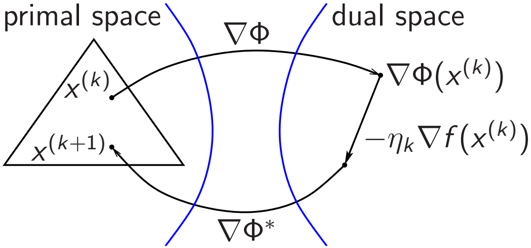


$$\tilde{x}^{(k+1)} = (\nabla\phi)^{-1}\big(\nabla\phi(x^{(k)}) - \eta_k \nabla f(x^{(k)})\big)$$

$$x^{(k+1)} = \arg\min_{x\in\Delta} D_\phi\big(x, \tilde{x}^{(k+1)}\big)$$

avoid technicality of dealing with empty interior [e.g., Bubeck, 2015]

Mirror descent without projection

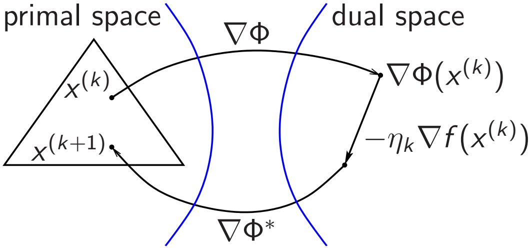


$$x^{(k+1)} = \nabla\Phi^*\left(\nabla\Phi(x^{(k)}) - \eta_k \nabla f(x^{(k)})\right)$$

example: $\Phi(x) = \sum_i x_i \log(x_i)$ if $x \in \Delta$ and ∞ otherwise

- $\mathbf{dom}\Phi$ has *empty interior*, $\partial\Phi(x) = \{\log(x) + c\mathbf{1} \mid c \in \mathbf{R}\}$
- $\Phi^*(x^*) = \log\left(\sum_i \exp(x_i^*)\right)$, $\nabla\Phi^*(x^*) = \frac{\exp(x^*)}{\|\exp(x^*)\|_1} \in \Delta$

Mirror descent without projection



$$x^{(k+1)} = \nabla\Phi^*\left(\nabla\Phi(x^{(k)}) - \eta_k \nabla f(x^{(k)})\right)$$

example: $\Phi(x) = \sum_i x_i \log(x_i)$ if $x \in \Delta$ and ∞ otherwise

- $\mathbf{dom}\Phi$ has *empty interior*, $\partial\Phi(x) = \{\log(x) + c\mathbf{1} \mid c \in \mathbf{R}\}$
- $\Phi^*(x^*) = \log\left(\sum_i \exp(x_i^*)\right)$, $\nabla\Phi^*(x^*) = \frac{\exp(x^*)}{\|\exp(x^*)\|_1} \in \Delta$

technicality matters when decomposing the operations

More convex analysis

- **affine-restricted Legendre function:** $\Phi(x) = \phi(x) + \delta(x|\mathcal{L})$
  - ϕ of Legendre type
  - $\mathcal{L}$ an affine subspace; $\delta(x|\mathcal{L}) = 0$ if $x \in \mathcal{L}$ and ∞ otherwise

More convex analysis

- **affine-restricted Legendre function:** $\Phi(x) = \phi(x) + \delta(x|\mathcal{L})$
  - ϕ of Legendre type
  - $\mathcal{L}$ an affine subspace; $\delta(x|\mathcal{L}) = 0$ if $x \in \mathcal{L}$ and ∞ otherwise

- **lemma:** suppose $\mathrm{int}\,(\mathrm{dom}\phi) \cap \mathcal{L} \neq \emptyset$, then for any $\nabla\Phi \in \partial\Phi$

$$\nabla\Phi^*\big(\nabla\Phi(x)\big) = x, \quad \big\langle\nabla\Phi(\nabla\Phi^*(x^*)), x - y\big\rangle = \langle x^*, x - y\rangle$$

More convex analysis

- **affine-restricted Legendre function:** $\Phi(x) = \phi(x) + \delta(x|\mathcal{L})$
  - ϕ of Legendre type
  - $\mathcal{L}$ an affine subspace; $\delta(x|\mathcal{L}) = 0$ if $x \in \mathcal{L}$ and ∞ otherwise

- **lemma:** suppose $\operatorname{int}(\operatorname{dom}\phi) \cap \mathcal{L} \neq \emptyset$, then for any $\nabla\Phi \in \partial\Phi$

$$\nabla\Phi^*\big(\nabla\Phi(x)\big) = x, \quad \big\langle \nabla\Phi(\nabla\Phi^*(x^*)), x - y\big\rangle = \langle x^*, x - y\rangle$$

- **lemma:** Bregman divergence D_Φ well defined, as for any $\nabla\Phi \in \partial\Phi$

$$\big\langle \nabla\Phi(y), x - y\big\rangle = \langle \nabla\phi(y), x - y\rangle$$

More convex analysis

- **affine-restricted Legendre function:** $\Phi(x) = \phi(x) + \delta(x|\mathcal{L})$
  - ϕ of Legendre type
  - $\mathcal{L}$ an affine subspace; $\delta(x|\mathcal{L}) = 0$ if $x \in \mathcal{L}$ and ∞ otherwise

- **lemma:** suppose $\operatorname{int}(\operatorname{dom}\phi) \cap \mathcal{L} \neq \emptyset$, then for any $\nabla\Phi \in \partial\Phi$

$$\nabla\Phi^*\big(\nabla\Phi(x)\big) = x, \quad \big\langle \nabla\Phi(\nabla\Phi^*(x^*)), x - y \big\rangle = \langle x^*, x - y \rangle$$

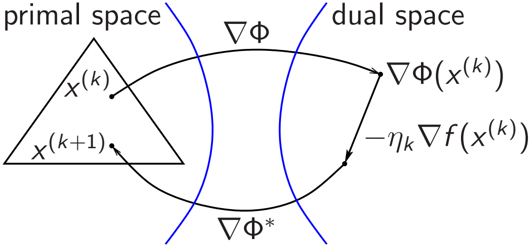
- **lemma:** Bregman divergence D_Φ well defined, as for any $\nabla\Phi \in \partial\Phi$

$$\big\langle \nabla\Phi(y), x - y \big\rangle = \langle \nabla\phi(y), x - y \rangle$$

- **lemma:** Bregman divergence of conjugate pairs

$$D_{\Phi^*}(x^*, y^*) = D_\Phi(\nabla\Phi^*(y^*), \nabla\Phi^*(x^*))$$

Mirror descent without projection

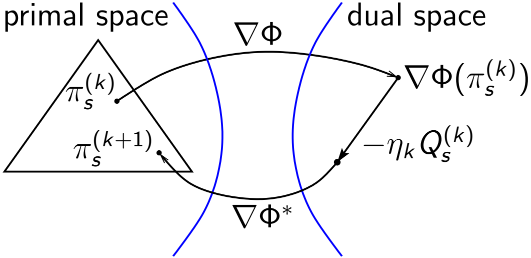


$$x^{(k+1)} = \nabla\Phi^*\big(\nabla\Phi(x^{(k)}) - \eta_k \nabla f(x^{(k)})\big)$$

theory on affine-restricted Legendre function:

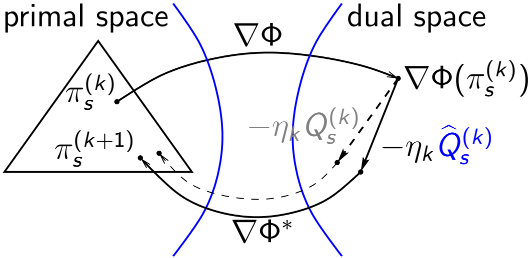
- support negative entropy restricted on Δ (empty interior)
- enable convergence analysis of PMD with function approximation

PMD with function approximation



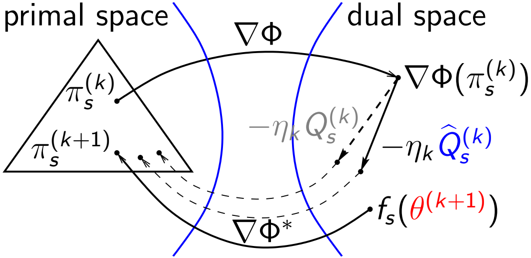
$$\pi_s^{(k+1)} = \nabla\Phi^*\left(\nabla\Phi(\pi^{(k)}) - \eta_k Q_s^{(k)}\right)$$

PMD with function approximation



$$\pi_s^{(k+1)} = \nabla\Phi^*\Big(\nabla\Phi\big(\pi^{(k)}\big) - \eta_k \widehat{Q}_s^{(k)}\Big)$$

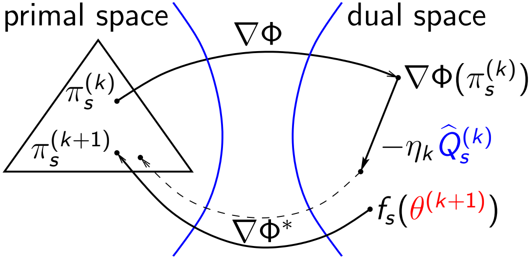
PMD with function approximation



$$f_s(\theta^{(k+1)}) \approx \nabla\Phi(\pi^{(k)}) - \eta_k \widehat{Q}_s^{(k)}$$

$$\pi_s^{(k+1)} = \nabla\Phi^*\big(f_s(\theta^{(k+1)})\big)$$

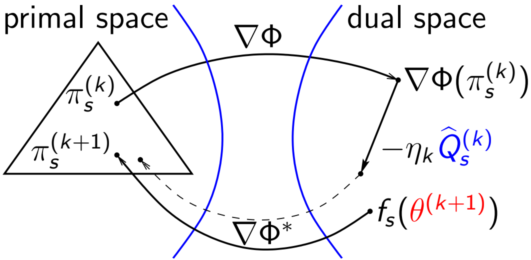
PMD with function approximation



$$f_s(\theta^{(k+1)}) \approx \nabla\Phi(\pi^{(k)}) - \eta_k \widehat{Q}_s^{(k)}$$

$$\pi_s^{(k+1)} = \nabla\Phi^*\left(f_s(\theta^{(k+1)})\right)$$

PMD with function approximation

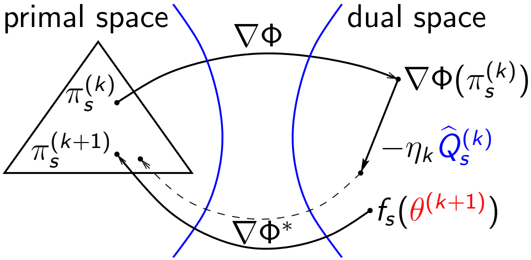


$$f_s(\theta^{(k+1)}) \approx \nabla\Phi(\pi^{(k)}) - \eta_k \widehat{Q}_s^{(k)}$$
$$\pi_s^{(k+1)} = \nabla\Phi^*\big(f_s(\theta^{(k+1)})\big)$$

how to measure approximation error in dual space?

i.e., choose loss function to minimize over θ (neural networks)

Function approximation in dual space

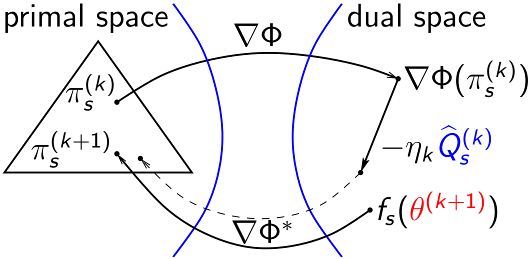


- standard approach: L_2-approximation

$$\theta^{(k+1)} = \arg\min_{\theta \in \Theta} \mathbf{E}_{s \sim d_\rho^{(k)}} \Big[\Big\| f_s(\theta) - \Big(\nabla\Phi(\pi_s^{(k)}) - \eta_k \widehat{Q}_s^{(k)} \Big) \Big\|_2^2 \Big]$$

  – compatible approximation [Sutton et al., 1999, Agarwal et al., 2021]
  – adopted in recent work by [Alfano et al., 2024]

Dual approximation policy optimization (DAPO)



- measure approximation error using D_{Φ^*} [Xiong et al., 2024]

$$\theta^{(k+1)} = \underset{\theta\in\Theta}{\arg\min}\ \mathbf{E}_{s\sim d_\rho^{(k)}}\left[D_{\Phi^*}\left(\nabla\Phi(\pi_s^{(k)}) - \eta_k \widehat{Q}_s^{(k)},\ f_s(\theta)\right)\right]$$

in practice: inexact minimization using a few steps of SGD

DAPO algorithm

input: initial policy $\pi^{(0)}$ parametrized by $\theta^{(0)}$

for $k = 0, \ldots, K$ **do**

critic update: find $\widehat{Q}^{(k)}$ that approximates $Q(\pi^{(k)})$

actor update:

$$\theta^{(k+1)} \approx \operatorname*{arg\,min}_{\theta \in \Theta} \; \mathbf{E}_{s \sim d_\rho^{(k)}} \Big[D_{\Phi^*}\big(\nabla \Phi(\pi_s^{(k)}) - \eta_k \widehat{Q}_s^{(k)},\; f_s(\theta) \big) \Big]$$

policy update: $\pi_s^{(k+1)} = \nabla \Phi^* \big(f_s(\theta^{(k+1)}) \big), \;\; s \in \mathcal{S}$

end for

Convergence analysis of DAPO

assumptions: distribution mismatch coefficients $\leq \vartheta_\rho$, . . . , and

$$\mathbf{E}_{s\sim d_\rho^{(k)}}\Big[\big\|\widehat{Q}_s^{(k)} - Q_s^{(k)}\big\|_\infty\Big] \leq \epsilon_{\mathrm{critic}}$$

$$\mathbf{E}_{s\sim d_\rho^{(k)}}\Big[D_{\Phi^*}\big(\nabla\Phi(\pi_s^{(k)}) - \eta_k \widehat{Q}_s^{(k)},\ f_s(\theta^{(k+1)})\big)\Big] \leq \eta_k\, \epsilon_{\mathrm{actor}}$$

Convergence analysis of DAPO

assumptions: distribution mismatch coefficients $\leq \vartheta_\rho$, ..., and

$$\mathbf{E}_{s\sim d_\rho^{(k)}}\left[\left\|\widehat{Q}_s^{(k)} - Q_s^{(k)}\right\|_\infty\right] \leq \epsilon_{\text{critic}}$$

$$\mathbf{E}_{s\sim d_\rho^{(k)}}\left[D_{\Phi^*}\left(\nabla\Phi(\pi_s^{(k)}) - \eta_k \widehat{Q}_s^{(k)},\ f_s(\theta^{(k+1)})\right)\right] \leq \eta_k\, \epsilon_{\text{actor}}$$

sublinear convergence

$$\frac{1}{K}\sum_{k=0}^{K-1}\left(V_\rho^{(k)} - V_\rho^\star\right) \leq \frac{1}{K}\left(\frac{D_0^\star}{(1-\gamma)\eta} + \frac{V_{d_\rho^\star}^{(0)}}{1-\gamma}\right) + \frac{\vartheta_\rho\sqrt{\epsilon_{\text{actor}}} + (2-\gamma)\vartheta_\rho\epsilon_{\text{critic}}}{(1-\gamma)^2}$$

linear convergence

$$V_\rho^{(K)} - V_\rho^\star \leq \left(1 - \frac{1}{\vartheta_\rho}\right)^K\left(V_\rho^{(0)} - V_\rho^\star + \frac{D_0^\star/(\vartheta_\rho - 1)}{(1-\gamma)\eta_0}\right) + \frac{\vartheta_\rho^2\sqrt{\epsilon_{\text{actor}}} + 2\vartheta_\rho^2\epsilon_{\text{critic}}}{1-\gamma}$$

Convergence analysis of DAPO

assumptions: distribution mismatch coefficients $\leq \vartheta_\rho, \ldots,$ and

$$\mathbf{E}_{s \sim d_\rho^{(k)}} \Big[\big\| \widehat{Q}_s^{(k)} - Q_s^{(k)} \big\|_\infty \Big] \leq \epsilon_{\text{critic}}$$

$$\mathbf{E}_{s \sim d_\rho^{(k)}} \Big[D_{\Phi^*} \big(\nabla \Phi(\pi_s^{(k)}) - \eta_k \widehat{Q}_s^{(k)},\ f_s(\theta^{(k+1)}) \big) \Big] \leq \eta_k\, \epsilon_{\text{actor}}$$

sublinear convergence

$$\frac{1}{K} \sum_{k=0}^{K-1} \left(V_\rho^{(k)} - V_\rho^\star \right) \leq \frac{1}{K} \left(\frac{D_0^\star}{(1-\gamma)\eta} + \frac{V_{d_\rho^\star}^{(0)}}{1-\gamma} \right) + \frac{\vartheta_\rho \sqrt{\epsilon_{\text{actor}}} + (2-\gamma)\vartheta_\rho \epsilon_{\text{critic}}}{(1-\gamma)^2}$$

linear convergence

$$V_\rho^{(K)} - V_\rho^\star \leq \left(1 - \frac{1}{\vartheta_\rho} \right)^K \left(V_\rho^{(0)} - V_\rho^\star + \frac{D_0^\star / (\vartheta_\rho - 1)}{(1-\gamma)\eta_0} \right) + \frac{\vartheta_\rho^2 \sqrt{\epsilon_{\text{actor}}} + 2\vartheta_\rho^2 \epsilon_{\text{critic}}}{1-\gamma}$$

(provide convergence analysis for practical methods: SAC, MDPO)

Final technicality

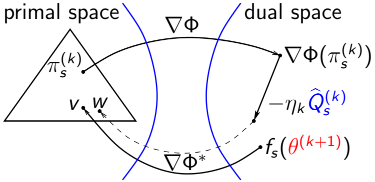
- cannot use **three-point descent lemma** with $\pi_s^\theta = \nabla\Phi^*(f_s(\theta))$

Final technicality

- cannot use **three-point descent lemma** with $\pi_s^\theta = \nabla\Phi^*(f_s(\theta))$
- working directly with **three-point identity**

$$D_\Phi(u,v) + D_\Phi(v,w) - D_\Phi(u,w) = \left\langle \nabla\Phi(v) - \nabla\Phi(w),\, v-u \right\rangle$$

$$u = \pi_s^\star$$
$$v = \nabla\Phi^*(f_s(\theta^{(k+1)}))$$
$$w = \nabla\Phi^*\left(\nabla\Phi(\pi_s^{(k)}) - \eta_k \widehat{Q}_s^{(k)}\right)$$

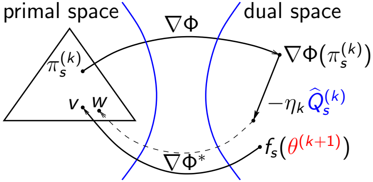


Final technicality

- cannot use **three-point descent lemma** with $\pi_s^\theta = \nabla\Phi^*(f_s(\theta))$
- working directly with **three-point identity**

$$D_\Phi(u,v) + D_\Phi(v,w) - D_\Phi(u,w) = \left\langle \nabla\Phi(v) - \nabla\Phi(w),\, v - u \right\rangle$$

$$u = \pi_s^\star$$
$$v = \nabla\Phi^*(f_s(\theta^{(k+1)}))$$
$$w = \nabla\Phi^*\left(\nabla\Phi(\pi_s^{(k)}) - \eta_k \widehat{Q}_s^{(k)}\right)$$



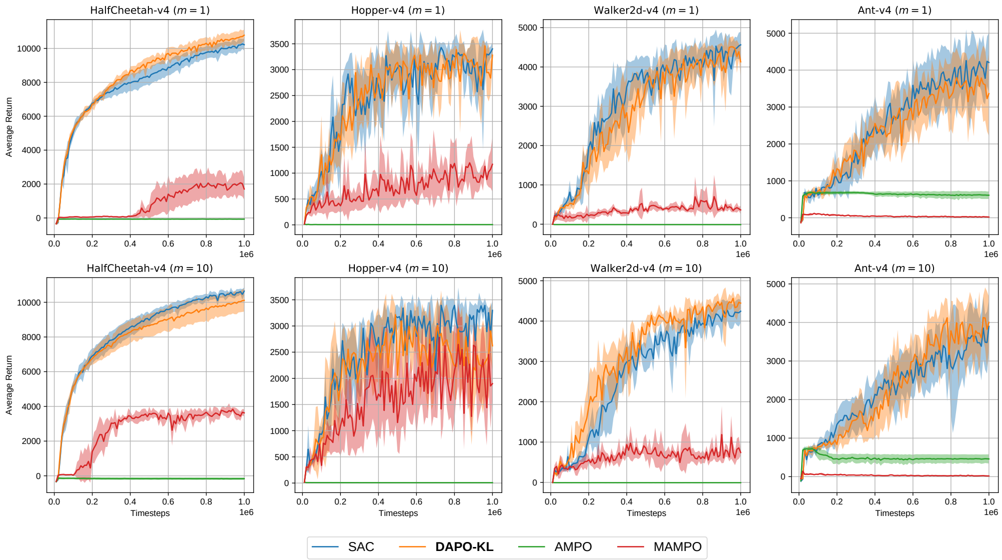
- bound inner product with $D(v,w)$

$$\left\langle \nabla\Phi(v) - \nabla\Phi(w),\, v - u \right\rangle \le \left(1 + \left\| \frac{u}{v} \right\|_\infty\right) \left(D_\Phi(v,w) + \sqrt{2D_\Phi(v,w)}\right)$$

  - need Pinsker's inequality on simplex (not with projection)
  - $D_\Phi(v,w) \le \eta_k \epsilon_{\mathrm{actor}}$ directly controlled by function approximation

Numerical experiments

average reward on MuJoCo benchmarks (m: number of SGD steps)



Summary

- **policy mirror descent (PMD)**
  - convergence to global optima despite non-convexity (PDL)
  - sublinear/linear convergence depending on step size rules
- **PMD with dual function approximation (DAPO)**
  - convex analysis of affine-restricted Legendre functions
  - measure approximation loss using dual Bregman divergence
  - analysis without three-point descent lemma

Summary

- **policy mirror descent (PMD)**
  - convergence to global optima despite non-convexity (PDL)
  - sublinear/linear convergence depending on step size rules
- **PMD with dual function approximation (DAPO)**
  - convex analysis of affine-restricted Legendre functions
  - measure approximation loss using dual Bregman divergence
  - analysis without three-point descent lemma
- **insights for optimization**
  - importance of exploiting MDP structure (PDL)
  - power of general techniques: mirror descent, convex analysis

References I

A. Agarwal, S. M. Kakade, J. D. Lee, and G. Mahajan. On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *Journal of Machine Learning Research*, 22(98):1–76, 2021.

C. Alfano, R. Yuan, and P. Rebeschini. A novel framework for policy mirror descent with general parameterization and linear convergence. *Advances in Neural Information Processing Systems*, 36, 2024.

H. H. Bauschke, J. Bolte, and M. Teboulle. A descent lemma beyond Lipschitz gradient continuity: first-order method revisited and applications. *Mathematics of Operations Research*, 42(2):330–348, 2017.

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

References II

B. Birnbaum, N. R. Devanur, and L. Xiao. Distributed algorithms via gradient descent for Fisher market. In *Proceedings of the 12th ACM Conference on Electronic Commerce*, pages 127–136, San Jose, California, USA, 2011.

S. Bubeck. *Convex Optimization: Algorithms and Complexity*. Number 8:3-4 in Foundations and Trends in Machine Learning. now Publishers Inc., 2015.

G. Chen and M. Teboulle. Convergence analysis of a proximal-like minimization algorithm using Bregman functions. *SIAM Journal on Optimization*, 3(3): 538–543, 1993.

S. Kakade and J. Langford. Approximately optimal approximate reinforcement learning. In *Proceedings of the 19th International Conference on Machine Learning (ICML)*, volume 2, pages 267–274, 2002.

S. Khodadadian, P. R. Jhunjhunwala, S. M. Varma, and S. T. Maguluri. On the linear convergence of natural policy gradient algorithm. arXiv preprint, arXiv:2105.01424, 2021.

References III

G. Lan. Policy mirror descent for reinforcement learning: Linear convergence, new sampling complexity, and generalized problem classes. Preprint, arXiv:2102.00135, 2021.

H. Lu, R. M. Freund, and Y. Nesterov. Relatively smooth convex optimization by first-order methods, and applications. *SIAM Journal on Optimization*, 28(1): 333–354, 2018.

A. Nemirovski and D. Yudin. *Problem Complexity and Method Efficiency in Optimization*. Wiley Interscience, 1983.

M. L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley and Sons, Inc., 2005.

L. Shani, Y. Efroni, and S. Mannor. Adaptive trust region policy optimization: Global convergence and faster rates for regularized mdps. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence*, pages 5668–5675. AAAI Press, 2020.

References IV

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

L. Xiao. On the convergence rates of policy gradient methods. *Journal of Machine Learning Research*, 23(282):1–36, 2022.

Z. Xiong, M. Fazel, and L. Xiao. Dual approximation policy optimization. *arXiv preprint arXiv:2410.01249*, 2024.