A Tutorial on Policy Gradient Methods

Lin Xiao and Lihong Li
(Facebook) (Amazon)

Mini-Tutorial on AI, Machine Learning and Optimization
SIAM Conference on Optimization

July 21, 2021

AI, machine learning, and optimization

- optimization: fundamental tool for AI and machine learning
- **importance of structure**
  - develop new algorithms and theory
    (simplex, IPM, mirror-descent, variance reduction, ...)
  - extend scope of fundamental algorithms and theory
    (also provide more insight of problem structure)
- this tutorial: **policy gradient methods**
  - among most effective methods for reinforcement learning
  - involve a variety of optimization algorithms and theory

focus: tabular setting, simple and unified convergence analysis

Outline

- **discounted finite Markov decision process (MDP)**
- (exact) policy gradient methods:
  - **policy gradient method with softmax parametrization** (non-uniform PŁ and smoothness, linear convergence)
  - **projected policy gradient method** (gradient mapping domination and $O(1/k)$ rate)
  - **natural policy gradient (NPG), mirror descent** ($O(1/k)$ rate and linear convergence)
  - **projected Q-descent method (new)**
- **summary** (insights on linear convergence)

Preview of results

basic conclusions

- convergence to global optimum despite nonconvexity
- constant stepsize: $O(1/k)$ sublinear convergence
- increasing stepsize: linear convergence

true for different classes of (exact) policy gradient methods

several new results

- $O(1/k)$ rate of projected policy gradient method
- simple analysis of linear convergence of NPG (mirror-descent)
- projected Q-descent method and its fast convergence

(please cite upcoming paper or this talk)

Markov decision process (MDP)

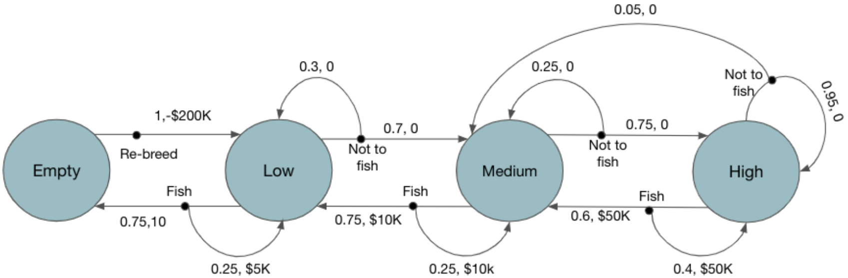
discounted finite MDP: $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, r, \gamma)$

- $\mathcal{S}$: finite state space
- $\mathcal{A}$: finite action space
- $P : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$: probability transition function
  - $P(s'|s, a)$: probability of transition to s' from s after action a
  - $P(\cdot|s, a) \in \Delta(\mathcal{S})$: distribution of state after action a in state s
- $r : \mathcal{S} \times \mathcal{A} \to \mathbf{R}$: reward function
- $\gamma \in (0, 1)$: discount factor for reward (usually close to 1)

powerful model with many applications in OR and RL

A simple example

whether to fish salmons this year:



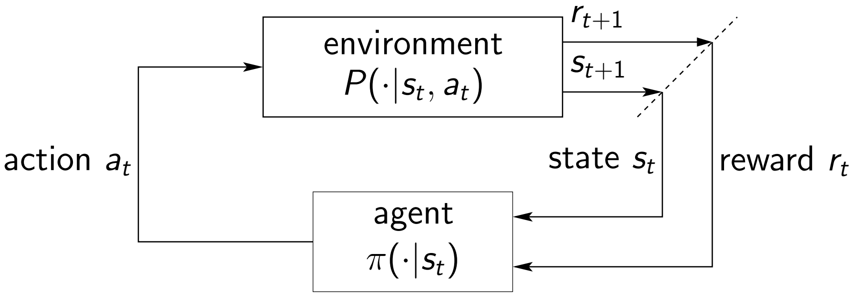
(image credit: Somnath Banerjee [Banerjee, 2021])

- 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

Policy

stationary policy: $\pi \in \Delta(\mathcal{A})^{|\mathcal{S}|}$ (independent of time t)

- $\pi(a|s)$: probability of taking action a in state s
- *deterministic policy:* $\pi(\cdot|s) = 1$ for some $a' \in \mathcal{A}$ and 0 for $a \neq a'$



notation: will also use $\pi_{s,a}$ for $\pi(a|s)$ and $\pi_s \in \Delta(\mathcal{A})$ for $\pi(\cdot|s)$

Value function

- **value function**: for each $s \in \mathcal{S}$

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

- **vectored value function**, $V(\pi) = [V_s(\pi)]_{s \in \mathcal{S}} \in \mathbf{R}^{|\mathcal{S}|}$, satisfies

$$V(\pi) = \sum_{t=0}^{\infty} \gamma^t P(\pi)^t R(\pi)$$

$P(\pi)$ and $R(\pi)$ linear functions of π

- $P(\pi) \in \mathbf{R}^{|\mathcal{S}| \times |\mathcal{S}|}$ where $P_{s,s'}(\pi) = \sum_{a \in \mathcal{A}} \pi_{s,a} P_{s,s'}(a)$
- $R(\pi) \in \mathbf{R}^{|\mathcal{S}|}$ where $R_s(\pi) = \sum_{a \in \mathcal{A}} \pi_{s,a} r_{s,a} = \langle \pi_s, r_s \rangle$

Policy evaluation

- vectored value function

$$\begin{aligned}
V(\pi) &= \sum_{t=0}^{\infty} \gamma^t P(\pi)^t R(\pi) \\
&= R(\pi) + \gamma P(\pi) R(\pi) + \gamma^2 P(\pi)^2 R(\pi) + \cdots \\
&= R(\pi) + \gamma P(\pi) \left(R(\pi) + \gamma P(\pi) R(\pi) + \gamma^2 P(\pi)^2 R(\pi) + \cdots \right) \\
&= R(\pi) + \lambda P(\pi) V(\pi)
\end{aligned}$$

- unique solution of linear equations (cost $|\mathcal{S}|^3$ solving directly)

$$V = R(\pi) + \gamma P(\pi) V$$

- closed-form expression:

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

well defined since $0 < \gamma < 1$ and $P(\pi)$ row stochastic

Optimality

- **optimal values** (maximizing discounted total reward)

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

- **optimal policy**

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

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

- **optimality equation (Bellman equation)**

$$V_s = \sup_{a \in \Delta\mathcal{A}} \left\{ r_{s,a} + \lambda \sum_{s' \in \mathcal{S}} P_{s,s'}(a) V_{s'} \right\}, \qquad \forall\, s \in \mathcal{S}$$

or in vector form: $V = \sup_{\pi \in \Delta(\mathcal{A})^{|\mathcal{S}|}} \{ R(\pi) + \lambda P(\pi) V \}$

Algorithms

- dynamic programming (DP)
  - value iteration
  - policy iteration
  - TD learning, Q-learning, ...
- linear programming
- (stochastic) policy gradient methods
  - REINFORCE, NPG, TRPO, actor-critic, ...
- direct policy optimization
  - evolution strategies, ...

Policy gradient methods

- **expected value function:** for any $\rho \in \Delta(\mathcal{S})$

$$V_\rho(\pi) := \mathop{\mathbf{E}}_{s\sim\rho}[V_s(\pi)] = \textstyle\sum_{s\in\mathcal{S}} \rho_s V_s(\pi) = \langle V(\pi), \rho\rangle$$

- **minimizing discounted total cost**

$$\min_{\pi\in\Delta(\mathcal{A})^{|\mathcal{S}|}} V_\rho(\pi) \qquad\qquad (\text{cost} = 1 - \text{reward} \in [0,1])$$

- focus of tutorial: methods with **exact gradient oracle**
  - policy gradient with soft-max parametrization
  - projected policy gradient method
  - natural policy gradient (mirror descent)

impractical, but **foundational** for stochastic methods

Structure of value function

- $V_\rho(\pi)$ **non-convex in general**

$$V_\rho(\pi) = \rho^T \left(I - \gamma P(\pi)\right)^{-1} R(\pi)$$

  for a concrete example, see, e.g., [Agarwal et al., 2021]

- **nonconvexity seems superficial**
  - admits linear programming formulation
  - scalar case: linear fractional function

$$f(x) = \frac{a^T x + b}{c^T x + d} \qquad \text{with} \quad c^T x + d > 0$$

    quasi-convex and quasi-concave
  - **policy gradient methods converge to global optima**

State-action value function

- **definition**

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

alternative definition:

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

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

$$V_s(\pi) = \mathop{\mathbf{E}}_{a \sim \pi_s} [Q_{s,a}(\pi)] = \sum_{a \in \mathcal{A}} \pi_{s,a} Q_{s,a}(\pi) = \langle Q_s(\pi), \pi_s \rangle$$

notations: moving s, a as subscripts, emphasizing function of π

Advantage function

- **definition**

$$A_{s,a}(\pi) := Q_{s,a}(\pi) - V_s(\pi)$$

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

$$\mathop{\mathbf{E}}_{a\sim\pi_s}[A_{s,a}(\pi)] = \sum_{a\in\mathcal{A}} \pi_{s,a} A_{s,a}(\pi) = \langle A_s(\pi), \pi_s \rangle = 0$$

proof:

$$\begin{aligned}
\big\langle A_s(\pi), \pi_s \big\rangle &= \big\langle Q_s(\pi) - V_s(\pi)\mathbf{1},\ \pi_s \big\rangle \\
&= \big\langle Q_s(\pi), \pi_s \big\rangle - V_s(\pi)\big\langle \mathbf{1}, \pi_s \big\rangle \\
&= V_s(\pi) - V_s(\pi) \\
&= 0
\end{aligned}$$

Discounted state visitation probability

- **definition**

$$d_{s,s'}(\pi) := (1-\gamma)\sum_{t=0}^{\infty}\gamma^t \Pr^\pi(s_t = s' \mid s_0 = s)$$
$$= (1-\gamma)\left[\left(I - \gamma P(\pi)\right)^{-1}\right]_{s,s'}$$

- **properties**
  - $\sum_{s'} d_{s,s'}(\pi) = 1$, therefore $d_s(\pi) \in \Delta(\mathcal{S})$ for all s and π
  - diagonals $d_{s,s}(\pi) \geq 1-\gamma$ (because $\Pr(s_0 = s \mid s_0 = s) = 1$)
  - $d_s(\pi) \in \Delta(\mathcal{S})$: visitation distribution when starting at s

- **expected state-visitation probability**

$$d_{\mu,s'}(\pi) := \mathop{\mathbf{E}}_{s\sim\mu}\left[d_{s,s'}(\pi)\right] = \sum_{s\in\mathcal{S}} \mu_s d_{s,s'}(\pi)$$

Distribution mismatch coefficients

- policy gradient methods often involve two distributions:
  - $\rho \in \Delta(\mathcal{S})$: used in $V_\rho(\pi)$ for performance evaluation
  - $\mu \in \Delta(\mathcal{S})$: initial state distribution to evaluate $\nabla V_\mu(\pi)$

- **distribution mismatch coefficient**

$$\left\| \frac{d_\rho(\pi)}{d_\mu(\pi')} \right\|_\infty := \max_{s \in \mathcal{S}} \frac{d_{\rho,s}(\pi)}{d_{\mu,s}(\pi')}$$

- since $d_{\mu,s}(\pi) \geq (1-\gamma)\mu_s$, for all $\pi' \in \Delta(\mathcal{A})^{|\mathcal{S}|}$

$$\left\| \frac{d_\rho(\pi)}{d_\mu(\pi')} \right\|_\infty \leq \frac{1}{1-\gamma} \left\| \frac{d_\rho(\pi)}{\mu} \right\|_\infty$$

- **assumption:** $\mu_s > 0$ for all $s \in \mathcal{S}$

Policy gradient

policy gradients are weighted Q-functions

$$\frac{\partial V_s(\pi)}{\partial \pi_{s',a'}} = \frac{1}{1-\gamma} d_{s,s'}(\pi) Q_{s',a'}(\pi)$$

policy gradient of expected value function

$$\frac{\partial V_\mu(\pi)}{\partial \pi_{s',a'}} = \frac{1}{1-\gamma} d_{\mu,s'}(\pi) Q_{s',a'}(\pi)$$

policy gradient with respect to π_s

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

Derivation of policy gradient

define $e_s \in \mathbf{R}^{|\mathcal{S}|}$ with $e_{s,s'} = 1$ if $s = s'$ and 0 otherwise

- value function: $V_s(\pi) = e_s^T V(\pi) = e_s^T \left(I - \gamma P(\pi)\right)^{-1} R(\pi)$
- visitation probability: $d_{s,s'}(\pi) = (1-\gamma) e_s^T \left(I - \gamma P(\pi)\right)^{-1} e_{s'}$

policy gradient: using matrix calculus $\frac{\partial X^{-1}}{\partial \alpha} = -X^{-1} \frac{\partial X}{\partial \alpha} X^{-1}$

$$\begin{aligned}
\frac{\partial V_s(\pi)}{\partial \pi_{s',a'}} &= e_s^T \left(I - \gamma P(\pi)\right)^{-1} \left(\frac{\partial R(\pi)}{\partial \pi_{s',a'}} + \gamma \frac{\partial P(\pi)}{\pi_{s',a'}} \left(I - \gamma P(\pi)\right)^{-1} R(\pi) \right) \\
&= e_s^T \left(I - \gamma P(\pi)\right)^{-1} e_{s'} \left(r_{s',a'} + \gamma P_{s',:}(a') V(\pi) \right) \\
&= \frac{1}{1-\gamma} d_{s,s'}(\pi) Q_{s',a'}(\pi)
\end{aligned}$$

Performance difference lemma

- original form [Kakade and Langford, 2002]

$$V_s(\pi) - V_s(\tilde{\pi}) = \frac{1}{1-\gamma} \mathop{\mathbf{E}}_{s' \sim d_s^{\pi}} \mathop{\mathbf{E}}_{a' \sim \pi(\cdot|s')} [A_{s',a'}(\tilde{\pi})]$$

(RHS: expectation of $A_{s',a'}(\tilde{\pi})$ w.r.t. distribution induced by π)

- useful variant:

$$V_s(\pi) - V_s(\tilde{\pi}) = \frac{1}{1-\gamma} \mathop{\mathbf{E}}_{s' \sim d_s^{\pi}} \langle Q_{s'}(\tilde{\pi}), \pi_{s'} - \tilde{\pi}_{s'} \rangle$$

equivalence:

$$\begin{aligned} \mathop{\mathbf{E}}_{a' \sim \pi(\cdot|s')} [A_{s',a'}(\tilde{\pi})] &= \left\langle A_{s'}(\tilde{\pi}),\ \pi_{s'} \right\rangle \\ &= \left\langle Q_{s'}(\tilde{\pi}) - V_{s'}(\tilde{\pi})\mathbf{1},\ \pi_{s'} \right\rangle \\ &= \left\langle Q_{s'}(\tilde{\pi}),\ \pi_{s'} \right\rangle - V_{s'}(\tilde{\pi}) \\ &= \left\langle Q_{s'}(\tilde{\pi}),\ \pi_{s'} - \tilde{\pi}_{s'} \right\rangle \end{aligned}$$

Proof of performance difference lemma

$$
\begin{aligned}
& V_s(\pi) - V_s(\tilde{\pi}) \\
&= \big\langle Q_s(\pi), \pi_s \big\rangle - \big\langle Q_s(\tilde{\pi}), \tilde{\pi}_s \big\rangle \\
&= \big\langle Q_s(\tilde{\pi}), \pi_s - \tilde{\pi}_s \big\rangle + \big\langle Q_s(\pi) - Q_s(\tilde{\pi}), \pi_s \big\rangle \\
&= \big\langle Q_s(\tilde{\pi}), \pi_s - \tilde{\pi}_s \big\rangle + \gamma \sum_{s \in \mathcal{S}} \pi_{s,a} \sum_{s' \in \mathcal{S}} P_{s,s'}(a) \big(V_{s'}(\pi) - V_{s'}(\tilde{\pi}) \big)
\end{aligned}
$$

let $u \in \mathbf{R}^{|\mathcal{S}|}$ such that $u_s = \big\langle Q_s(\tilde{\pi}), \pi_s - \tilde{\pi}_s \big\rangle$, then

$$
V(\pi) - V(\tilde{\pi}) = u + \gamma P(\pi) \big(V(\pi) - V(\tilde{\pi}) \big)
$$

therefore $V(\pi) - V(\tilde{\pi}) = \big(I - \gamma P(\pi) \big)^{-1} u$, written component-wise:

$$
V_s(\pi) - V_s(\tilde{\pi}) = \frac{1}{1-\gamma} \sum_{s' \in \mathcal{S}} d_{s,s'}(\pi) \big\langle Q_{s'}(\tilde{\pi}), \pi_{s'} - \tilde{\pi}_{s'} \big\rangle
$$

Outline

- discounted finite Markov decision process (MDP)
- (exact) policy gradient methods:
  - **policy gradient method with softmax parametrization**
    (non-uniform PŁ and smoothness, linear convergence)
  - projected policy gradient method
    (gradient mapping domination and $O(1/k)$ rate)
  - natural policy gradient (mirror descent)
    ($O(1/k)$ rate and linear convergence)
  - projected Q-descent method (new)
- summary (insights on linear convergence)

Parametrizations of policy

- **direct parametrization**
  - treat $\pi_{s,a}$ as variables in $\mathbf{R}^{|\mathcal{S}|\times|\mathcal{A}|}$
  - need explicit constraints: $\pi_s \in \Delta(\mathcal{A})$ for all $s \in \mathcal{S}$
  - complete policy class (contains optimal policy)
- **softmax parametrization:**

$$\pi_{s,a}(\theta) = \frac{\exp(f_{s,a}(\theta))}{\sum_{a'} \exp(f_{s,a'}(\theta))}$$

  - 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: $f_{s,a}(\theta) = \text{network}(\theta, \phi_{s,a})$ and $\theta \in \mathbf{R}^p$

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

Policy gradient theorem

define $J_s(\theta) = V_s(\pi(\theta))$, then [Sutton et al., 2000]:

$$\nabla J_s(\theta) = \frac{1}{1-\gamma} \sum_{s' \in \mathcal{S}} d_{s,s'}(\pi(\theta)) \sum_{a' \in \mathcal{A}} Q_{s',a'}(\pi(\theta)) \nabla \pi_{s',a'}(\theta)$$

- direct parametrization: $\frac{\partial \pi_{s,a}}{\partial \pi_{s',a'}} = \begin{cases} 1 & s = s' \quad a = a' \\ 0 & \text{otherwise} \end{cases}$, therefore

$$\frac{\partial V_s(\pi)}{\partial \pi_{s',a'}} = \frac{1}{1-\gamma} d_{s,s'}(\pi) Q_{s',a'}(\pi)$$

- softmax tabular parametrization [Agarwal et al., 2021]

$$\frac{\partial J_s(\theta)}{\partial \theta_{s',a'}} = \frac{1}{1-\gamma} d_{s,s'}(\pi(\theta)) \pi_{s',a'}(\theta) A_{s',a'}(\pi(\theta))$$

Softmax policy gradient descent

unconstrained optimization

$$\min_{\theta \in \mathbf{R}^{|\mathcal{S}| \times |\mathcal{A}|}} \{ J_\rho(\theta) := V_\rho(\pi(\theta)) \}$$

policy gradient descent

$$\theta^{k+1} = \theta^k - \eta_k \nabla J_\mu(\theta^k)$$

- softmax tabular parametrization: $\pi_{s,a}(\theta) = \frac{\exp(\theta_{s,a})}{\sum_{a'} \exp(\theta_{s,a'})}$
- objective defined with $\rho \in \Delta(\mathcal{S})$
- gradient computed with $\mu \in \Delta(\mathcal{S})$:

$$\frac{\partial J_\mu(\theta)}{\partial \theta_{s,a}} = \frac{1}{1-\gamma} d_{\mu,s}(\pi(\theta)) \pi_{s,a}(\theta) A_{s,a}(\pi(\theta))$$

Convergence rate

Recall assumption of sufficient exploration: $\mu_s > 0$ for all $s \in S$

- [Mei et al., 2020]: with constant stepsize $\eta_k = \eta = (1-\gamma)^3/8$

$$J_\rho(\theta^k) - J_\rho^\star = O\left(\frac{|\mathcal{S}|}{(1-\gamma)^6 k}\right)$$

- [Mei et al., 2021]: normalized policy gradient descent

$$\theta^{k+1} = \theta^k - \eta \frac{\nabla J_\mu(\theta^k)}{\|\nabla J_\mu(\theta^k)\|_2}$$

geometric convergence:

$$J_\rho(\theta^{k+1}) - J_\rho^\star = O\left(\frac{1}{1-\gamma}\exp(-Ck)\right)$$

constants in $O(\cdot)$ depend on $\frac{1}{1-\gamma}$, $\left\|\frac{d_\mu(\pi^\star)}{\mu}\right\|_\infty$ and $\left\|\frac{d_\rho(\pi^\star)}{\mu}\right\|_\infty$

Smooth nonconvex optimization

$\min_{x \in \mathbf{R}^n} f(x)$ with gradient descent $x^{k+1} = x^k - \eta_k \nabla f(x^k)$

- **smoothness:** $\|\nabla f(x) - \nabla f(y)\|_2 \leq L\|x - y\|_2$, which implies

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2}\|y - x\|_2^2$$

- **descent property:** with stepsize $\eta_k = 1/L$,

$$f(x^k) - f(x^{k+1}) \geq \frac{1}{2L}\|\nabla f(x^k)\|_2^2$$

- **convergence rate:** $\min_{0 \leq k \leq K} \|\nabla f(x^k)\|_2^2 \leq \dfrac{2L\big(f(x^0) - f^\star\big)}{K + 1}$

Convergence to global optimum

- **gradient dominance** (Polyak-Łojasiewicz condition)

$$\frac{1}{2}\|\nabla f(x)\|_2^2 \geq \mu\big(f(x) - f^\star\big)$$

  – satisfied by strong convexity with convexity parameter μ

  – nonconvex f: guarantees convergence to global optimum

- **linear convergence to global optimum:**

$$f(x^k) - f(x^{k+1}) \;\geq\; \frac{1}{2L}\|\nabla f(x^k)\|_2^2 \;\geq\; \frac{\mu}{L}\big(f(x^k) - f^\star\big)$$

$$\implies\; f(x^{k+1}) - f^\star \leq \left(1 - \frac{\mu}{L}\right)\big(f(x^k) - f^\star\big)$$

$$\implies\; f(x^{k+1}) - f^\star \leq \left(1 - \frac{\mu}{L}\right)^{k+1}\big(f(x^0) - f^\star\big)$$

Convergence to global optimum

- **weak gradient dominance** (weak Łojasiewicz condition)

$$\|\nabla f(x)\|_2 \geq \sqrt{2\mu}\left(f(x) - f^\star\right)$$

- combined with smooth descent property:

$$f(x^k) - f(x^{k+1}) \;\geq\; \frac{1}{2L}\|\nabla f(x^k)\|_2^2 \;\geq\; \frac{\mu}{L}\left(f(x^k) - f^\star\right)^2$$

- $O(1/k)$ **convergence to global optimum:**

$$f(x^k) - f^\star \leq \frac{f(x^0) - f^\star}{1 + k \cdot \frac{\mu}{L}(f(x^0) - f^\star)}$$

Proof of $O(1/k)$ convergence

let $\delta_k = f(x^k) - f^\star$, then

$$\delta_k - \delta_{k+1} \geq \frac{\mu}{L}\delta_k^2$$

dividing both sides by $\delta_k \delta_{k+1}$ and using $\delta_k \geq \delta_{k+1}$:

$$\frac{1}{\delta_{k+1}} - \frac{1}{\delta_k} \geq \frac{\mu}{L}\frac{\delta_k}{\delta_{k+1}} \geq \frac{\mu}{L}$$

telescoping sum over iterations $0, 1, \ldots, k-1$:

$$\frac{1}{\delta_k} - \frac{1}{\delta_0} \geq k \cdot \frac{\mu}{L} \qquad \Longrightarrow \qquad \delta_k \leq \frac{1}{\frac{1}{\delta_0} + k \cdot \frac{\mu}{L}}$$

(cf. [Nesterov, 2004, Theorem 2.1.14])

Linear convergence under weak PŁ

- **weak gradient dominance** + **smoothness**:

$$f(x^k) - f(x^{k+1}) \;\geq\; \frac{1}{2L_k}\|\nabla f(x^k)\|_2^2 \;\geq\; \frac{\mu}{L_k}\big(f(x^k) - f^\star\big)^2$$

- **non-uniform smoothness** [Mei et al., 2021]

$$f(y) \leq f(x) + \langle \nabla f(x), y - x\rangle + \frac{L_k}{2}\|y - x\|_2^2$$

  - if $L_k \leq \beta\big(f(x^k) - f^\star\big)$ or $L_k \leq \beta'\|\nabla f(x^k)\|_2$, then with $\eta_k = \frac{1}{L_k}$,

$$f(x^k) - f(x^{k+1}) \;\geq\; \frac{\mu}{\beta}\big(f(x^k) - f^\star\big)$$

  - linear convergence: $f(x^k) - f^\star \leq \big(1 - \frac{\mu}{\beta}\big)^k\big(f(x^0) - f^\star\big)$
  - effectively, stepsize $\eta_k = \frac{1}{L_k}$ increases geometrically

Application: softmax policy gradient descent

minimize $J_\rho(\theta) = V_\rho(\pi(\theta))$:

$$\theta^{k+1} = \theta^k - \eta_k \nabla J_\mu(\theta^k)$$

- (weak) **gradient dominance** [Mei et al., 2020]

$$\|\nabla J_\mu(\theta)\|_2 \geq c_1 \left(J_\rho(\theta) - J_\rho^\star\right)$$

$\Rightarrow$ $O(1/k)$ convergence rate with constant stepsize

- **non-uniform smoothness** [Mei et al., 2021]

$$L_\rho(\theta) \leq c_2 \|\nabla J_\mu(\theta)\|_2$$

$\Rightarrow$ linear convergence with increasing stepsize, or normalized PG

$$\theta^{k+1} = \theta^k - \eta \frac{\nabla J_\mu(\theta^k)}{\|\nabla J_\mu(\theta^k)\|_2}$$

Key for linear convergence

- **(weak) PŁ condition (gradient dominance)**
  - linear quadratic regulators (LQR): [Fazel et al., 2018]
  - finite MDP: [Bhandari and Russo, 2019] [Agarwal et al., 2021]
  - can be derived from performance difference lemma (PDL)
- **non-uniform (gradient dominated) smoothness**
  - structure of softmax tabular parametrization

$$\pi_{s,a}(\theta) = \frac{\exp(\theta_{s,a})}{\sum_{a'} \exp(\theta_{s,a'})}$$

  - $\|\theta^k\| \to \infty$, converging to extremely flat landscape
  - can use increasing stepsize or **normalized gradient method**

(examples with different exponents given in [Mei et al., 2021])

Interpretation through Polyak stepsize I

unconstrained optimization

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

Polyak step size:

$$\eta_k = \frac{f(x^k) - f^\star}{\|\nabla f(x^k)\|^2}$$

two ways to derive

- from subgradient convergence analysis (assuming convexity of f)
- Newton-Raphson perspective [Gower et al., 2021]

$$f(x^k) + \langle \nabla f(x^k), x - x^k \rangle = f^\star$$

  - heavily under-determined linear system
  - minimizing $\|x - x^k\|^2$ with linear constraint gives Polyak stepsize

Interpretation through Polyak stepsize II

Polyak step size:

$$x^{k+1} = x^k - \frac{f(x^k) - f^\star}{\|\nabla f(x^k)\|^2} \nabla f(x^k)$$

not practical if $f^\star$ unknown, but has important implications:

- strong PŁ condition: $\|\nabla f(x)\|^2 \geq \mu(f(x) - f^\star)$

$$x^{k+1} = x^k - \eta \nabla f(x^k) \qquad \text{with} \quad \eta \sim \frac{1}{\mu}$$

- weak PŁ condition: $\|\nabla f(x)\| \geq \mu'(f(x) - f^\star)$

$$x^{k+1} = x^k - \eta \frac{\nabla f(x^k)}{\|\nabla f(x^k)\|} \qquad \text{with} \quad \eta \sim \frac{1}{\mu'}$$

Outline

- discounted finite Markov decision process (MDP)
- (exact) policy gradient methods:
  - policy gradient method with softmax parametrization (non-uniform P]L and smoothness, linear convergence)
  - **projected policy gradient method** (gradient mapping domination and $O(1/k)$ rate)
  - natural policy gradient (mirror descent) ($O(1/k)$ rate and linear convergence)
  - projected Q-descent method (new)
- summary (insights on linear convergence)

Projected policy gradient method

constrained optimization with direct parametrization

$$\min_{\pi \in \Delta(\mathcal{A})^{|\mathcal{S}|}} V_{\rho}(\pi) \qquad V_{\rho}(\pi) = \left\langle \rho,\ \left(I - \gamma P(\pi)\right)^{-1} R(\pi) \right\rangle$$

projected policy gradient method

$$\pi^{k+1} = \mathbf{Proj}_{\Delta(\mathcal{A})^{|\mathcal{S}|}} \left(\pi^{k} - \eta_k \nabla V_{\mu}(\pi^{k})\right)$$

projection can be done for each s separately

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

Convergence rate of PPG

- $O(1/\sqrt{K})$ convergence to global optimum [Agarwal et al., 2021]

$$\min_{0<k\leq K}\left\{V_\rho(\pi^k)-V_\rho^\star\right\}\leq\frac{64\gamma|\mathcal{S}||\mathcal{A}|}{\sqrt{K}(1-\gamma)^6}\left\|\frac{d_\rho(\pi^\star)}{\mu}\right\|_\infty^2$$

  - smoothness: $\|\nabla V_s(\pi)-\nabla V_s(\pi')\|_2\leq\frac{2\gamma|\mathcal{A}|}{(1-\gamma)^3}\|\pi-\pi'\|_2$
  - **variational gradient dominance** condition

$$V_\rho(\pi)-V_\rho^\star\leq\frac{1}{1-\gamma}\left\|\frac{d_\rho(\pi^\star)}{\mu}\right\|_\infty\max_{\pi'\in\Delta\mathcal{A}^{|\mathcal{S}|}}\left\langle\nabla V_\mu(\pi),\,\pi'-\pi\right\rangle$$

- new analysis as proximal gradient method
  - **gradient-mapping dominance** condition
  - $O(1/k)$ convergence with constant stepsize

Composite nonconvex optimization

$$\underset{x \in \mathbf{R}^n}{\text{minimize}} \quad \{F(x) := f(x) + \Psi(x)\}$$

- f smooth: $\|\nabla f(x) - \nabla f(y)\|_* \le L_f \|x - y\|$, which implies

$$|f(y) - f(x) - \langle \nabla f(x), y - x \rangle| \le \frac{L_f}{2}\|y - x\|^2$$

- Ψ: closed convex function, can be nonsmooth
  - nonsmooth regularization, such as $\Psi(x) = \lambda \|x\|_1$
  - indicator function: $\Psi(x) = 0$ if $x \in \mathcal{C}$ and $+\infty$ otherwise

- in context of projected policy gradient method
  - f: expected value function $V_\rho(\cdot)$
  - Ψ: indicator function of convex set $\Delta(\mathcal{A})^{|\mathcal{S}|}$

Proximal gradient method

- **proximal mapping** of Ψ

$$\mathbf{prox}_{\Psi}(x) = \arg\min_{y} \left\{ \Psi(y) + \tfrac{1}{2}\|y - x\|_2^2 \right\}$$

- **proximal gradient method** with constant step size $\eta = \frac{1}{L}$

$$\begin{aligned} x^{k+1} &= \arg\min_{x} \left\{ \left\langle \nabla f(x^k), x - x^k \right\rangle + \tfrac{L}{2}\|x - x^k\|^2 + \Psi(x) \right\} \\ &= \mathbf{prox}_{\frac{1}{L}\Psi}\left(x^k - \tfrac{1}{L}\nabla f(x^k) \right) \end{aligned}$$

- **Gradient mapping**

$$G_L(x^k) = L\left(x^k - \mathbf{prox}_{\frac{1}{L}\Psi}\left(x^k - \tfrac{1}{L}\nabla f(x^k) \right) \right)$$

therefore $x^{k+1} = x^k - \frac{1}{L}G_L(x^k)$ (if $\Psi \equiv 0$, then $G_L(x) = \nabla f(x)$)

Analysis of proximal gradient method

convergence rate:

$$\min_{0\leq k\leq K} \left\| G_L(x^k) \right\|_2 = O\left(\frac{1}{\sqrt{K+1}}\right)$$

- progress of each iteration (e.g., [Beck, 2017, Section 10.3])

$$F(x^k) - F(x^{k+1}) \geq \frac{1}{2L} \left\| G_L(x^k) \right\|_2^2$$

- sum over $k = 0, 1, \ldots, K-1$

$$F(x^0) - F(x^{K+1}) \geq \frac{1}{2L} \sum_{k=0}^{K} \left\| G_L(x^k) \right\|_2^2$$

- therefore

$$\min\nolimits_{0\leq k\leq K} \left\| G_L(x^k) \right\|_2^2 \leq \frac{2L\big(F(x^0) - F^\star\big)}{K+1}$$

Convergence to global optimum

- **gradient mapping domination**

$$\frac{1}{2}\|G_L(x)\|_2^2 \ge \mu\big(F(x^+) - F^\star\big)$$

where $x^+ = x - \frac{1}{L}G_L(x) = \mathbf{prox}_{\frac{1}{L}\psi}\left(x - \frac{1}{L}\nabla f(x)\right)$

- **linear convergence to global optimum**

$$F(x^k) - F(x^{k+1}) \;\ge\; \frac{1}{2L}\left\|G_L(x^k)\right\|_2^2 \;\ge\; \frac{\mu}{L}\big(F(x^{k+1}) - F^\star\big)$$

$$\Longrightarrow\;\; \left(1 + \frac{\mu}{L}\right)\big(F(x^{k+1}) - F^\star\big) \le F(x^k) - F^\star$$

$$\Longrightarrow\;\; F(x^{k+1}) - F^\star \le \left(1 + \frac{\mu}{L}\right)^{-(k+1)}\big(F(x^0) - F^\star\big)$$

Relation to KŁ or proximal-PŁ

- **Kurdyka-Łojasiewicz (KŁ)** with exponent 1/2

$$\min_{s\in\partial F(x)} \tfrac{1}{2}\|s\|^2 \geq \tilde{\mu}\big(F(x) - F^\star\big)$$

- **proximal PŁ** (equivalent to KŁ) [Karimi et al., 2016]

$$\tfrac{1}{2}P_L(x) \geq \mu\big(F(x) - F^\star\big)$$

$$P_L(x) := -2L\min_y \left\{\langle \nabla f(x), y - x\rangle + \tfrac{L}{2}\|y - x\|_2^2 + \Psi(y) - \Psi(x)\right\}$$

- **gradient mapping domination**

$$\tfrac{1}{2}\,\|G_L(x)\|_2^2 \geq \mu\big(F(x^+) - F^\star\big)$$

interlacing relations:

$$\tfrac{1}{2}\|P_L(x)\|_2^2 \geq \tfrac{1}{2}\,\|G_L(x)\|_2^2 \geq \mu\big(F(x) - F^\star\big) \geq \mu\big(F(x^+) - F^\star\big)$$

Convergence to global optimum

- **weak gradient mapping domination**

$$\|G_L(x)\|_2 \geq \sqrt{2\mu}\left(F(x^+) - F^\star\right)$$

- combined with proximal gradient descent property:

$$F(x^k) - F(x^{k+1}) \;\geq\; \frac{1}{2L}\|G_L(x^k)\|_2 \;\geq\; \frac{\mu}{L}\left(F(x^{k+1}) - F^\star\right)^2$$

- $O(1/k)$ **convergence to global optimum:**

$$F(x^k) - F^\star \leq \max\left\{\frac{1}{1 + k \cdot \frac{\mu}{4L}(F(x^0) - F^\star)}, \left(\frac{\sqrt{2}}{2}\right)^k\right\}(F(x^0) - F^\star)$$

Proof of $O(1/k)$ convergence I

let $\delta_k = F(x^k) - F^\star$, then

$$\delta_k - \delta_{k+1} \geq \frac{\mu}{L}\delta_{k+1}^2$$

dividing both sides by $\delta_k \delta_{k+1}$

$$\frac{1}{\delta_{k+1}} - \frac{1}{\delta_k} \geq \frac{\mu}{L}\frac{\delta_{k+1}}{\delta_k}$$

telescoping sum over iterations $0, 1, \ldots, k-1$:

$$\frac{1}{\delta_k} - \frac{1}{\delta_0} \geq \frac{\mu}{L}\sum_{i=0}^{k-1}\frac{\delta_{i+1}}{\delta_i}$$

cannot continue as before ($\geq \frac{\mu}{L}k$) because $\delta_{i+1} \leq \delta_i$

Proof of $O(1/k)$ convergence II

for any two constant $r, c \in (0,1)$, define

$$n(k,r) := \text{number of times } \tfrac{\delta_{i+1}}{\delta_i} \geq r \text{ for } 0 \leq i \leq k-1$$

- if $n(k,r) \geq ck$, then $\frac{\delta_{i+1}}{\delta_i} \geq r$ at least $\lceil ck \rceil$ times, thus

$$\frac{1}{\delta_k} - \frac{1}{\delta_0} \geq \frac{\mu}{L} rck \qquad \Longrightarrow \qquad \delta_k \leq \frac{1}{\frac{1}{\delta_0} + \frac{\mu}{L} rck} = O\left(\frac{1}{k}\right)$$

- if $n(k,r) < ck$, then $\frac{\delta_{i+1}}{\delta_i} < r$ at least $\lceil (1-c)k \rceil$ times

$$\delta_k \leq \delta_0 r^{(1-c)k} = \delta_0 \left(r^{1-c}\right)^k$$

combining two cases:

$$\delta_k \leq \min_{0<r,c<1} \max \left\{ \frac{1}{1+\frac{\mu}{L}\delta_0 rck}, \; \left(r^{1-c}\right)^k \right\} \delta_0$$

Linear convergence under weak KŁ?

- **weak gradient dominance** + **smoothness**:

$$F(x^k) - F(x^{k+1}) \;\geq\; \frac{1}{2L_k}\|G_{L_k}(x^k)\|_2^2 \;\geq\; \frac{\mu}{L_k}\big(F(x^{k+1}) - F^\star\big)^2$$

- **non-uniform smoothness** (extending [Mei et al., 2021])?

$$f(y) \leq f(x) + \langle \nabla f(x), y - x\rangle + \frac{L_k}{2}\|y - x\|_2^2$$

if $L_k \leq \beta\big(F(x^{k+1}) - F^\star\big)$ or $L_k \leq \beta' \|G_{L_k}(x^k)\|_2$, then

$$F(x^k) - F(x^{k+1}) \;\geq\; \frac{\mu}{\beta}\big(F(x^{k+1}) - F^\star\big)$$

  - does not work with compact feasible set
  - will revisit as special case of mirror descent

Application: projected policy gradient method

- composite nonconvecx optimization formulation

$$\min_{\pi \in \mathbf{R}^{|\mathcal{S}||\mathcal{A}|}} \left\{ F(\pi) := V_\rho(\pi) + \Psi(\pi) \right\}$$

  - V_ρ: smooth with Lipschitz constant $\frac{2\gamma|\mathcal{A}|}{(1-\gamma)^3}$ [Agarwal et al., 2021]
  - Ψ: indicator of convex set $\Delta(\mathcal{A})^{|\mathcal{S}|}$

- projected policy gradient method

$$\pi^{k+1} = \mathbf{prox}_{\eta\Psi}\left(\pi^k - \eta \nabla V_\mu(\pi^k)\right) = \pi^k - \eta G_{1/\eta}(\pi^k)$$

  [Agarwal et al., 2021]: $O(1/\sqrt{k})$ convergence rate

- new $O(1/k)$ convergence rate
  - smoothness + (weak) gradient mapping domination

Gradient mapping domination

- variational gradient domination [Agarwal et al., 2021, Lemma 4]

$$V_\rho(\pi) - V_\rho(\pi^\star) \leq \frac{1}{1-\gamma} \left\| \frac{d_\rho(\pi^\star)}{\mu} \right\|_\infty \max_{\pi' \in \Delta(\mathcal{A})^{|\mathcal{S}|}} \left\langle \nabla V_\mu(\pi),\, \pi - \pi' \right\rangle$$

- [Nesterov, 2013, Theorem 1]: if $\pi^+ = \pi - \eta G_{1/\eta}(\pi)$, then

$$\max_{\pi' \in \Delta(\mathcal{A})^{|\mathcal{S}|}} \left\langle \nabla V_\mu(\pi^+),\, \pi^+ - \pi' \right\rangle \leq (1 + L\eta) \cdot \left\| G_{1/\eta}(\pi) \right\|_2 \cdot \left\| \pi^+ - \pi' \right\|_2$$

(holds for $\min_x f(x) + \Psi(x)$ where f smooth and Ψ convex)

- using $\eta = 1/L$ and noticing $\|\pi_s^+ - \pi_s'\|_2 \leq \sqrt{2}$ for all $s \in \mathcal{S}$,

$$\max_{\pi' \in \Delta(\mathcal{A})^{|\mathcal{S}|}} \left\langle \nabla V_\mu(\pi^+),\, \pi^+ - \pi' \right\rangle \leq 2\sqrt{2|\mathcal{S}|} \cdot \left\| G_{1/\eta}(\pi) \right\|_2$$

Improved convergence rate

- (weak) gradient mapping domination:

$$V_\rho(\pi^+) - V_\rho(\pi^\star) \le \frac{2\sqrt{2|\mathcal{S}|}}{1-\gamma} \left\| \frac{d_\rho(\pi^\star)}{\mu} \right\|_\infty \|G_L(\pi)\|_2$$

  (recall general form $\|G_L(x)\|_2 \ge \sqrt{2\mu}\big(F(x^+) - F^\star\big)$)

- $O(1/k)$ convergence rate according to previous analysis
- iteration complexity for $V_\rho(\pi^k) - V_\rho^\star \le \epsilon$

$$\max\left\{ \frac{128|\mathcal{S}||\mathcal{A}|}{\epsilon\,(1-\gamma)^6} \left\| \frac{d_\rho(\pi^\star)}{\mu} \right\|_\infty^2,\ 3\log\frac{1}{(1-\gamma)\epsilon} \right\}$$

Outline

- discounted finite Markov decision process (MDP)
- (exact) policy gradient methods:
  - policy gradient method with softmax parametrization
    (non-uniform PŁ and smoothness, linear convergence)
  - projected policy gradient method
    (gradient mapping domination and $O(1/k)$ rate)
  - **natural policy gradient (mirror descent)**
    ($O(1/k)$ rate and linear convergence)
  - projected Q-descent method (new)
- summary (insights on linear convergence)

Natural policy gradient (NPG) method

- recall problem definition

$$J_\rho(\theta) := V_\rho(\pi(\theta))$$

- Fisher information matrix (induced by π)

$$F_\rho(\theta) = \mathop{\mathbf{E}}_{s\sim d_\rho(\pi(\theta))} \mathop{\mathbf{E}}_{a\sim\pi_s(\theta)} \left[\nabla \log \pi_{s,a}(\theta) \left(\nabla \log \pi_{s,a}(\theta)\right)^T\right]$$

- preconditioned policy gradient method

$$\theta^{k+1} = \theta^t + \eta F_\rho(\theta^k)^{-1} \nabla J_\rho(\theta)$$

[Kakade, 2001]

NPG with softmax parametrization

- softmax parametrization: $\pi : \mathbf{R}^{|\mathcal{S}|\times|\mathcal{A}|} \to \Delta(\mathcal{A})^{|\mathcal{S}|}$

$$\pi_{s,a}(\theta) = \frac{\exp(\theta_{s,a})}{\sum_{a'\in\mathcal{A}} \exp(\theta_{s,a'})}$$

- NPG update (e.g., [Agarwal et al., 2021])

$$\theta^{k+1} = \theta^k - \frac{\eta}{1-\gamma} A(\pi(\theta^k))$$

- corresponding policy update

$$\pi_{s,a}(\theta^{k+1}) = \pi_{s,a}(\theta^k) \frac{\exp\left(-\eta Q_{s,a}(\pi(\theta^k))/(1-\gamma)\right)}{Z_s(\theta^k)}$$

$Z_s(\theta^k)$: normalization factor such that $\sum_{a\in\mathcal{A}} \pi_{s,a}(\theta^{k+1}) = 1$

Equivalent mirror descent update

using direct parametrization:

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

where $_{KL}$ denotes Kullback-Leibler divergence:

$$D_{KL}(p||q) = \sum_{a \in \mathcal{A}} p_a \log \frac{p_a}{q_a}$$

convergence to global optima (without regularization)

- [Shani et al., 2020]: $O(1/\sqrt{k})$ convergence rate
- [Agarwal et al., 2021]: $O(1/k)$ rate
- [Lan, 2021]: $O(1/k)$ rate and linear convergence
- [Khodadadian et al., 2021]: linear convergence (adaptive stepsize)

Mirror descent method

- convex optimization problem

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

- **Bregman divergence** of reference function h (strictly convex)

$$D_h(x, y) := h(x) - h(y) - \langle \nabla f(y) x - y \rangle$$

- **mirror descent** (proximal gradient form)

$$x^{k+1} = \underset{x \in \mathcal{C}}{\arg\min} \left\{ \eta_k \langle \nabla f(x^k), x \rangle + D(x, x^k) \right\}$$

  - originally due to [Nemirovski and Yudin, 1983]
  - proximal (sub-)gradient form [Beck and Teboulle, 2003]
  - $O(1/\sqrt{k})$ convergence rate in general convex setting

Examples of mirror descent method

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

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

$$\begin{aligned} x^{k+1} &= \arg\min_{x \in \mathcal{C}} \left\{ \eta_k \langle \nabla f(x^k), x \rangle + \frac{1}{2} \left\| x - x^k \right\|_2^2 \right\} \\ &= \mathbf{Proj}_{\mathcal{C}} \left(x^k - \eta_k \nabla f(x^k) \right) \end{aligned}$$

- simplex: $\mathcal{C} = \Delta$, $h = \sum_i x_i \log x_i$ and $D(x, y) = D_{KL}(x \| y)$

$$x_i^{k+1} = x_i \frac{\exp(-\eta_k \nabla_i f(x^k))}{Z(x^k)}$$

where $Z(x^k)$ is normalization factor such that $x^{k+1} \in \Delta$

Relative smoothness and strong convexity

question: can we have faster rate than $O(1/\sqrt{k})$ if f smooth?

- **relative smoothness** with parameter β

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

  equivalently

$$D_f(x, y) \leq \beta\, D_h(x, y)$$

- **relative strong convexity** with parameter α

$$D_f(x, y) \geq \alpha\, D_h(x, y)$$

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

$$\alpha\, D_h(x, y) \leq D_f(x, y) \leq \beta\, D_h(x, y)$$

Mirror descent: faster convergence rate

- convex optimization problem

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

assumption: f relatively smooth with respect to h

- mirror descent

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

  - $O(1/k)$ convergence rate [Birnbaum et al., 2011]
  - independent recent works: [Bauschke et al., 2017], [Lu et al., 2018], [Zhou et al., 2019]
  - linear rate under relative strong convexity [Lu et al., 2018]

Analysis of mirror descent I

- **three-point descent lemma** [Chen and Teboulle, 1993]:
  if ϕ convex and
  $$x^+ = \arg\min_{u \in \mathcal{C}} \{\phi(u) + D_h(u, x)\}$$
  then for any $u \in \mathcal{C}$,
  $$\phi(x^+) + D_h(x^+, x) \le \phi(u) + D_h(u, x) - D_h(u, x^+)$$
- applying to MD update with $\phi(\cdot) = \eta_k \langle \nabla f(x^k), \cdot \rangle$
  $$\underbrace{\langle \nabla f(x^k), x^{k+1} - u \rangle + \tfrac{1}{\eta_k} D_h(x^{k+1}, x^k)}_{a} \le \tfrac{1}{\eta_k} D_h(u, x^k) - \tfrac{1}{\eta_k} D_h(u, x^{k+1})$$
- let $u = x^k$ and $\eta_k \le \frac{1}{\beta}$, then $a \ge f(x^{k+1}) - f(x^k)$ by relative smooth
  $$\textbf{descent property:} \quad f(x^{k+1}) - f(x^k) \le -\tfrac{1}{\eta_k} D_h(x^k, x^{k+1}) \le 0$$

Analysis of mirror descent II

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

$$\underbrace{\langle \nabla f(x^k), x^{k+1} - x^k \rangle + \frac{1}{\eta_k} D_h(x^{k+1}, x^k)}_{a} + \underbrace{\langle \nabla f(x^k), x^k - u \rangle}_{b}$$

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

- **relative smoothness** $(\eta_k \leq \frac{1}{\beta})$: $\; a \geq f(x^{k+1}) - f(x^k)$
- **(relative strong) convexity**: $\; b \geq f(x^k) - f(u) + \alpha D_h(u, x^k)$

combining together,

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

Analysis of mirror descent III

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

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

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

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

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

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

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

Policy mirror descent (PMD)

- expected divergence: for arbitrary $\rho \in \Delta(\mathcal{S})$

$$D_\rho(\pi || \pi') = \mathop{\mathbf{E}}_{s\sim\rho}\left[D(\pi_s || \pi'_s)\right] == \sum_{s\in\mathcal{S}} \rho_s D(\pi_s || \pi'_s)$$

- mirror descent with weighted divergence

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

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

$$\pi^{k+1} = \mathop{\arg\min}_{\pi\in\Delta(\mathcal{A})^{|\mathcal{S}|}} \left\{ \frac{1}{1-\gamma} \sum_{s\in\mathcal{S}} d_{\mu,s}(\pi^k) \Big(\eta_k \left\langle Q_s(\pi^k),\ \pi_s \right\rangle + D(\pi_s, \pi_s^k) \Big) \right\}$$

$$\Longrightarrow \boxed{\pi_s^{k+1} = \mathop{\arg\min}_{\pi_S\in\Delta(\mathcal{A})} \left\{ \eta_k \left\langle Q_s(\pi^k),\ \pi_s \right\rangle + D(\pi_s, \pi_s^k) \right\}, \quad \forall\, s \in \mathcal{S}}$$

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
  $$x^+ = \arg\min_{u \in \mathcal{C}} \{ \phi(u) + D(u, x) \}$$
  then for any $u \in \mathcal{C}$,
  $$\phi(x^+) + D(x^+, x) \le \phi(u) + D(u, x) - D(u, x^+)$$

- applying to PMD update with $\phi(\cdot) = \eta_k \langle Q_s(\pi^k), \cdot \rangle$,
  $$\eta_k \langle Q_s(\pi^k), \pi_s^{k+1} - p \rangle + D(\pi_s^{k+1}, \pi_s^k) \le D(p, \pi_s^k) - D(p, \pi_s^{k+1})$$

Analysis of PMD II

last inequality on previous slide

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

- **Q-descent property**: letting $p = \pi_s^k$ yields

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

- let $\pi = \pi_s^\star$, and subtract and add π_s^k in inner product:

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

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

Analysis of PMD III

taking expectation w.r.t. visitation distribution $d_\rho(\pi^\star)$

$$\underbrace{\mathbf{E}_{s\sim d_\rho(\pi^\star)}\left[\langle Q_s(\pi^k), \pi_s^{k+1} - \pi_s^k\rangle\right]}_{a} + \underbrace{\mathbf{E}_{s\sim d_\rho(\pi^\star)}\left[\langle Q_s(\pi^k), \pi_s^{k} - \pi_s^\star\rangle\right]}_{b}$$
$$\le \frac{1}{\eta_k} D_{d_\rho(\pi^\star)}(\pi^\star, \pi^k) - \frac{1}{\eta_k} D_{d_\rho(\pi^\star)}(\pi^\star, \pi^{k+1})$$

- performance difference lemma: $b = (1-\gamma)\big(V_\rho(\pi^k) - V_\rho(\pi^\star)\big)$
- part a:

$$\begin{aligned}
a &= \textstyle\sum_{s\in\mathcal{S}} d_{\rho,s}(\pi^\star)\big\langle Q_s(\pi^k), \pi_s^{k+1} - \pi_s^k\big\rangle \\
&= \textstyle\sum_{s\in\mathcal{S}} \frac{d_{\rho,s}(\pi^\star)}{d_{\rho,s}(\pi^{k+1})} d_{\rho,s}(\pi^{k+1})\big\langle Q_s(\pi^k), \pi_s^{k+1} - \pi_s^k\big\rangle \\
(Q\text{-descent}) \quad &\ge \left\|\frac{d_\rho(\pi^\star)}{d_\rho(\pi^{k+1})}\right\|_\infty \textstyle\sum_{s\in\mathcal{S}} d_{\rho,s}(\pi^{k+1})\big\langle Q_s(\pi^k), \pi_s^{k+1} - \pi_s^k\big\rangle \\
(\text{PDL}) \quad &= \left\|\frac{d_\rho(\pi^\star)}{d_\rho(\pi^{k+1})}\right\|_\infty (1-\gamma)\big(V_\rho(\pi^{k+1}) - V_\rho(\pi^k)\big)
\end{aligned}$$

Analysis of PMD IV

let $\theta_{k+1} = \left\| \frac{d_\rho(\pi^\star)}{d_\rho(\pi^{k+1})} \right\|_\infty$, $\delta_k = V_\rho(\pi^k) - V_\rho(\pi^\star)$, $D_k = D_{d_\rho(\pi^\star)}(\pi^\star, \pi^k)$, then

$$\boxed{\theta_{k+1}\big(\delta_{k+1} - \delta_k\big) + \delta_k \leq \frac{1}{(1-\gamma)\eta_k} D_k - \frac{1}{(1-\gamma)\eta_k} D_{k+1}}$$

need to upper bound θ_{k+1}

- ρ uniform distribution: $\rho_s = 1/|\mathcal{S}|$ for all $s \in \mathcal{S}$

$$\theta_{k+1} \leq \frac{1 + \gamma|\mathcal{S}|}{1-\gamma}$$

- $\rho = \rho^\star$: stationary distribution of optimal policy $\pi^\star$

$$\theta_{k+1} \leq \max_{s \in \mathcal{S}} \frac{\rho_s}{(1-\gamma)\rho_s} = \frac{1}{1-\gamma}$$

Convergence rate with $\rho^\star$

using $\theta_{k+1} = \left\| \frac{d_\rho(\pi^\star)}{d_\rho(\pi^{k+1})} \right\|_\infty \leq \frac{1}{1-\gamma}$,

$$\boxed{\delta_{k+1} + \frac{1}{\eta_k} D_{k+1} \leq \gamma \delta_k + \frac{1}{\eta_k} D_k}$$

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

$$\delta_k \leq \frac{1}{k(1-\gamma)} \left(\delta_0 + \frac{1}{\eta} D_0 \right)$$

- increasing step size $\eta_{k+1} = \eta_k / \gamma$

$$\delta_{k+1} + \frac{1}{\gamma \eta_{k+1}} D_{k+1} \leq \gamma^k \left(\delta_0 + \frac{1}{\gamma \eta_0} D_0 \right)$$

Iteration complexity with $\rho^\star$

theorem: in order to find π such that $V_{\rho^\star}(\pi) \le V^\star_{\rho^\star} + \epsilon$, number of iterations of PMD bounded by (assuming $\pi^0_{s,a} = 1/|\mathcal{A}|$ for all s, a)

- with constant step size $\eta_k = \eta \ge (1-\gamma)\log|\mathcal{A}|$

$$\frac{2}{(1-\gamma)^2\epsilon}$$

  same as Agarwal et al. [2021] (for natural policy gradient)

- with $\eta_0 \ge \frac{1-\gamma}{\gamma}\log|\mathcal{A}|$ and $\eta_{k+1} = \eta_k/\gamma$

$$\frac{1}{1-\gamma}\log\frac{2}{(1-\gamma)\,\epsilon}$$

  slightly tighter than Lan [2021] (diminishing regularization)

Convergence rate with uniform ρ

using $\theta_{k+1} \le \frac{1+\gamma|\mathcal{S}|}{1-\gamma}$,

$$\boxed{(1+\gamma|\mathcal{S}|)\delta_{k+1} \le \gamma(1+|\mathcal{S}|)\delta_k + \frac{1}{\eta_k}D_k - \frac{1}{\eta_k}D_{k+1}}$$

- constant stepsize $\eta_k = \eta$

$$\delta_k \le \frac{1}{k(1-\gamma)}\left((1+\gamma|\mathcal{S}|)\delta_0 + \frac{1}{\eta}D_0\right)$$

- increasing stepsize $\eta_{k+1} \ge \frac{1+\gamma|\mathcal{S}|}{\gamma+\gamma|\mathcal{S}|}\eta_k$

$$\delta_{k+1} + \frac{1}{\gamma(1+|\mathcal{S}|)\eta_{k+1}}D_k \le \left(1 - \frac{1-\gamma}{1+\gamma|\mathcal{S}|}\right)^k \left(\delta_0 + \frac{1}{\gamma(1+|\mathcal{S}|)\eta_0}D_0\right)$$

Iteration complexity with uniform ρ

theorem: in order to find π such that $V_\rho(\pi) \leq V_\rho^\star + \epsilon$, number of iterations of PMD bounded by (assuming $\pi^0_{s,a} = 1/|\mathcal{A}|$ for all s, a)

- with constant step size $\eta_k = \eta \geq \frac{1-\gamma}{1+\gamma|\mathcal{S}|} \log|\mathcal{A}|$

$$\frac{2(1+\gamma|\mathcal{S}|)}{(1-\gamma)^2\epsilon}$$

  worse by factor $(1+\gamma|\mathcal{S}|)$ than [Agarwal et al., 2021]

- with $\eta_0 \geq \frac{1-\gamma}{\gamma(1+\gamma|\mathcal{S}|)} \log|\mathcal{A}|$ and $\eta_{k+1} \geq \frac{1+\gamma|\mathcal{S}|}{\gamma+\gamma|\mathcal{S}|}\eta_k$

$$\frac{1+\gamma|\mathcal{S}|}{1-\gamma} \log \frac{2}{(1-\gamma)\,\epsilon}$$

  worse by factor $(1+\gamma|\mathcal{S}|)$ than for complexity with $\rho^\star$

Superlinear convergence I

rewrite the "master inequality" as

$$\boxed{\delta_{k+1} + \frac{D_{k+1}}{\theta_{k+1}(1-\gamma)\eta_k} \leq \left(1 - \frac{1}{\theta_{k+1}}\right)\left(\delta_k + \frac{D_k}{(\theta_{k+1}-1)(1-\gamma)\eta_k}\right)}$$

increasing stepisze: if $\eta_k \geq \frac{\theta_k}{\theta_{k+1}-1}\eta_{k-1}$, then

$$\begin{aligned}\delta_{k+1} + \frac{D_{k+1}}{\theta_{k+1}(1-\gamma)\eta_k} &\leq \left(1 - \frac{1}{\theta_{k+1}}\right)\left(\delta_k + \frac{D_k}{\theta_k(1-\gamma)\eta_{k-1}}\right) \\ &\leq \prod_{i=0}^{k}\left(1 - \frac{1}{\theta_{i+1}}\right)\left(\delta_0 + \frac{D_0}{\theta_0(1-\gamma)\eta_{-1}}\right)\end{aligned}$$

therefore, $\theta_k \to 1$ implies **superlinear convergence**

- if $D_k = D_{d_\rho(\pi^\star)}(\pi^\star, \pi^k) \to 0$, then $\pi^k \to \pi^\star$, thus $\left\|\frac{d_\rho(\pi^\star)}{d_\rho(\pi^k)}\right\|_\infty \to 1$
- but it may not hold even though $\frac{D_k}{\theta_k(1-\gamma)\eta_{k-1}} \to 0$, since $\eta_k \to \infty$

Superlinear convergence II

sufficient conditions for $\theta_k = \left\| \frac{d_\rho(\pi^\star)}{d_\rho(\pi^k)} \right\|_\infty \to 1$

- convergence of $P(\pi^k)$ (cf. [Puterman, 2005, Corollary 6.4.10]

$$\lim_{k \to \infty} \left\| P(\pi^k) - P(\pi^\star) \right\| = 0$$

  because $d_\rho(\pi) = \left(I - \gamma P(\pi)\right)^{-T} \rho$

- there exists $0 < C < \infty$ such that for all $k = 1, 2, \ldots$

$$\left\| P(\pi^k) - P(\pi^\star) \right\| \leq C \left(V_\rho(\pi^k) - V_\rho^\star \right)$$

  then quadratic convergence (cf. [Puterman, 2005, Theorem 6.4.8])

these conditions may not hold even if $V_\rho(\pi^k) \to V_\rho^\star$ at fast rate

Outline

- discounted finite Markov decision process (MDP)
- (exact) policy gradient methods:
  - policy gradient method with softmax parametrization (non-uniform PŁ and smoothness, linear convergence)
  - projected policy gradient method (gradient mapping domination and $O(1/k)$ rate)
  - natural policy gradient (mirror descent) ($O(1/k)$ rate and linear convergence)
  - **projected Q-descent method (new)**
- summary (insights on linear convergence)

Projected Q-descent (PQD)

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

$$D_\rho(\pi, \pi') = \mathop{\mathbf{E}}_{s\sim\rho}\left[\tfrac{1}{2}\|\pi_s - \pi'_s\|_2^2\right] = \sum_{s\in\mathcal{S}} \rho_s \cdot \tfrac{1}{2}\|\pi_s - \pi'_s\|_2^2 \ \leq 1$$

- mirror descent method

$$\begin{aligned}
\pi^{k+1} &= \mathop{\arg\min}_{\pi\in\Delta(\mathcal{A})^{|\mathcal{S}|}} \left\{\eta_k \left\langle \nabla V_\mu(\pi^k),\ \pi\right\rangle + \tfrac{1}{1-\gamma} D_{d_\mu(\pi^k)}(\pi, \pi^k)\right\} \\
&= \mathop{\arg\min}_{\pi\in\Delta(\mathcal{A})^{|\mathcal{S}|}} \left\{\tfrac{1}{1-\gamma} \sum_{s\in\mathcal{S}} d_{\mu,s}(\pi^k)\left(\eta_k \left\langle Q_s(\pi^k),\ \pi_s\right\rangle + \tfrac{1}{2}\|\pi_s - \pi_s^k\|_2^2\right)\right\} \\
&= \mathbf{Proj}_{\Delta(\mathcal{A})^{|\mathcal{S}|}}\left(\pi^k - \eta_k Q(\pi^k)\right)
\end{aligned}$$

- equivalent to separate projections for each state

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

Convergence of PQD

projected Q-descent:

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

from analysis of general mirror descent:

- constant step size $\eta_k = \eta$

$$\delta_k \le \frac{1}{k(1-\gamma)}\left(\delta_0 + \frac{1}{\eta} D_{\rho^\star}(\pi^\star, \pi^0)\right) \le \frac{1}{k(1-\gamma)}\left(\frac{1}{1-\gamma} + \frac{1}{\eta}\right)$$

- increasing step size $\eta_{k+1} = \eta_k/\gamma$

$$\delta_k + \frac{1}{\gamma\eta_k} D_{\rho^\star}(\pi^\star, \pi^k) \le \gamma^k\left(\delta_0 + \frac{1}{\gamma\eta_0} D_{\rho^\star}(\pi^\star, \pi^0)\right) \le \gamma^k\left(\frac{1}{1-\gamma} + \frac{1}{\gamma\eta_0}\right)$$

Iteration complexity of PQD

theorem in order to find π^k such that $V_{\rho^\star}(\pi^k) \leq V^\star_{\rho^\star} + \epsilon$, number of iterations of PQD bounded by

- with constant step size $\eta_k = \eta \geq (1-\gamma)$

$$\frac{2}{(1-\gamma)^2\epsilon}$$

- with $\eta_0 \geq \frac{1-\gamma}{\gamma}$ and $\eta_{k+1} = \eta_k/\gamma$

$$\frac{1}{1-\gamma}\log\frac{2}{(1-\gamma)\,\epsilon}$$

(new) dimension-independent iteration complexity, same as NPG!

Outline

- discounted finite Markov decision process (MDP)
- (exact) policy gradient methods:
  - policy gradient method with softmax parametrization (non-uniform PŁ and smoothness, linear convergence)
  - projected policy gradient method (gradient mapping domination and $O(1/k)$ rate)
  - natural policy gradient (mirror descent) ($O(1/k)$ rate and linear convergence)
  - projected Q-descent method (new)
- **summary (insights on linear convergence)**

Hint of structure

policy gradient

$$\frac{\partial V_s(\pi)}{\partial \pi_{s'}} = \frac{1}{1-\gamma} d_{s,s'}(\pi) Q_{s'}(\pi)$$

- recall performance difference lemma:

$$V_s(\pi) - V_s(\tilde{\pi}) = \frac{1}{1-\gamma} \sum_{s'} d_{s,s'}(\pi) \left\langle Q_{s'}(\tilde{\pi}), \pi_{s'} - \tilde{\pi}_{s'} \right\rangle$$

- looks like a linear function!

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

  not really; but has both convex- and concave-like properties
  - gradient (mapping) dominance ensures global optimality
  - descent property does not depends on stepsize

Quasi-convexity and quasi-concavity

- for any $\pi \in \Delta(\mathcal{A})^{|\mathcal{S}|}$ define $\lambda \in \mathbf{R}_+^{|\mathcal{S}|\times|\mathcal{A}|}$ as

$$\lambda_{s,a} = \sum_{s' \in \mathcal{S}} \rho_{s'} \sum_{t=0}^{\infty} \gamma^t \overset{\pi}{\Pr}\left(s_t = s, a_t = a | s_0 = s'\right)$$

  - λ feasible for dual linear program (e.g., [Puterman, 2005])
  - $\pi_{s,a} = \frac{\lambda_{s,a}}{\sum_{a' \in \mathcal{A}} \lambda_{s,a'}}$ and $V_\rho(\pi) = \langle r, \lambda \rangle = \sum_{s \in \mathcal{S}, a \in \mathcal{A}} r_{s,a} \lambda_{s,a}$
- **proof of quasi-convexity:** for any $c \geq 0$,

$$\left\{\pi \in \Delta(\mathcal{A})^{|\mathcal{S}|} \mid V_\rho(\pi) \leq c\right\}$$

  is image under linear fractional transform of

$$\left\{\lambda \in \mathbf{R}_+^{|\mathcal{S}|\times|\mathcal{A}|} \mid \lambda \text{ feasible to dual LP and } \langle r, \lambda \rangle \leq c\right\}$$

  (see, e.g., [Boyd and Vandenberghe, 2004, Section 2.3.3])

Connection with policy iteration

- **policy iteration**

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

- **natural policy gradient (exponentiated Q-descent)**

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

- **projected Q-descent**

$$\pi_s^{k+1} = \arg\min_{p \in \Delta(\mathcal{A})} \left\{ \eta_k \left\langle Q_s(\pi^k), p \right\rangle + \tfrac{1}{2}\|p - \pi_s^k\|_2^2 \right\}, \quad \forall s \in \mathcal{S}$$

NPG and PQD equivalent to policy iteration as $\eta_k \to \infty$

Comparison of convergence rates

- **policy iteration**

$$\|V(\pi^k) - V^\star\|_\infty \le \gamma^k \|V(\pi^0) - V^\star\|_\infty \le \frac{\gamma^k}{1-\gamma}$$

- **natural policy gradient** with $\eta_{k+1} = \eta_k/\gamma$

$$\begin{aligned} V_{\rho^\star}(\pi^k) - V^\star_{\rho^\star} &\le \gamma^k \left(V_{\rho^\star}(\pi^0) - V^\star_{\rho^\star} + \frac{1}{\gamma\eta_0} \mathbf{E}_{s\sim\rho^\star}\left[D_{KL}(\pi^\star_s, \pi^0_s) \right] \right) \\ &\le \gamma^k \left(\frac{1}{1-\gamma} + \frac{1}{\gamma\eta_0} \log|\mathcal{A}| \right) \end{aligned}$$

- **projected Q-descent** with $\eta_{k+1} = \eta_k/\gamma$

$$\begin{aligned} V_{\rho^\star}(\pi^k) - V^\star_{\rho^\star} &\le \gamma^k \left(V_{\rho^\star}(\pi^0) - V^\star_{\rho^\star} + \frac{1}{\gamma\eta_0} \mathbf{E}_{s\sim\rho^\star}\left[\frac{1}{2} \left\|\pi^\star_s - \pi^0_s\right\|_2^2 \right] \right) \\ &\le \gamma^k \left(\frac{1}{1-\gamma} + \frac{1}{\gamma\eta_0} \right) \end{aligned}$$

Insights from [Bhandari and Russo, 2020]

- define $\hat{\pi}^{k+1}$ as function of stepsize $\eta > 0$:

$$\hat{\pi}^{k+1}(\eta) := \arg\min_{\pi \in \Delta(\mathcal{A})^{|\mathcal{S}|}} \left\{ \eta \left\langle Q(\pi^k), \pi \right\rangle + D(\pi, \pi^k) \right\}$$

  notice that $\hat{\pi}^{k+1}(\eta) \to \pi_{\mathsf{PI}}^{k+1}$ as $\eta \to \infty$

- exact line search

$$\pi^{k+1} = \hat{\pi}^{k+1}(\eta_\star), \qquad \eta_\star = \arg\inf_{\eta \in (0,\infty]} V_\rho\left(\hat{\pi}^{k+1}(\eta)\right)$$

  then $V_\rho\left(\pi^{k+1}\right) \leq V_\rho\left(\pi_{\mathsf{PI}}^{k+1}\right)$, known to converge linearly

Insights from [Bhandari and Russo, 2020]

- in fact only need η_k to satisfy

$$V_\rho(\pi^k) - V_\rho\left(\hat{\pi}^{k+1}(\eta_k)\right) \le \tfrac{1}{2}\left(V_\rho(\pi^k) - \inf_\eta V_\rho\left(\hat{\pi}^{k+1}(\eta)\right)\right)$$

and can replace $\frac{1}{2}$ by any fixed factor less than 1

- Frank-Wolfe (Conservative PI of [Kakade and Langford, 2002])

$$\pi^{k+1} = (1-\alpha)\pi^k + \alpha\pi_{\mathsf{PI}}^{k+1} \qquad \pi_{\mathsf{PI}}^{k+1} = \arg\min_{\pi\in\Delta(\mathcal{A})^{|\mathcal{S}|}} \left\langle Q(\pi^k), \pi \right\rangle$$

linear convergence

$$\|V(\pi^k) - V^\star\|_\infty \le \left(1 - \alpha(1-\gamma)\right)^k \|V(\pi^0) - V^\star\|_\infty$$

this tutorial: linear convergence with stepsizes $\eta_{k+1} = \eta_k/\gamma$

Role of regularization

work with regularized cost function:

$$\min_{\pi \in \Delta(\mathcal{A})^{|\mathcal{S}|}} \quad V_\rho(\pi) + \tau \mathop{\mathbf{E}}_{s \sim \rho}[h_s(\pi_s)]$$

where h_s convex or strongly convex
recent work: [Cen et al., 2020], [Lan, 2021], [Zhan et al., 2021], ...

simple take-aways:

- not necessary for linear convergence; increasing stepsizes suffice (see also [Mei et al., 2021, Lan, 2021, Khodadadian et al., 2021])
- but can improve the rate of convergence (regularized objective)
- good for approximate and stochastic policy gradient methods?

Summary I

- **policy gradient methods**
  - policy gradient with softmax parametrization
  - projected policy gradient method
  - natural policy gradient (mirror descent)
  - projected Q-descent method
- **convergence to global optimality** despite nonconvexity
  - constant stepsize: $O(1/k)$ sublinear convergence
  - increasing stepsize: $O(\gamma^k)$ linear convergence
- **interesting optimization theory**
  - unconstrained: gradient domination, nonuniform smoothness
  - constrained: gradient mapping domination
  - mirror descent: preconditioning with weighted divergence

Summary II

structure of discounted-reward finite MDP

- generalized monotonicity ([Liu et al., 2019, Lan, 2021]) (same as gradient domination? both due to PDL)
- Q-descent property, independent of stepsize
- work with Q-functions rather than gradient (preconditioning)

extensions

- stochastic policy gradient methods and their sample complexities ([Shani et al., 2020], [Lan, 2021], . . .)
- function approximations, their optimality and efficiency
- . . .

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.

S. Banerjee. Real world applications of markov decision process. Blog post at towardsdatascience.com, January 2021.

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. *First-Order Methods in Optimization*. MOS-SIAM Series on Optimization. SIAM, 2017.

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

J. Bhandari and D. Russo. Global optimality guarantees for policy gradient methods. arXiv preprint, arXiv:1906.01786, 2019.

J. Bhandari and D. Russo. A note on the linear convergence of policy gradient methods. arXiv preprint, arXiv:2007.11120, 2020.

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. P. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

S. Cen, C. Cheng, Y. Chen, Y. Wei, and Y. Chi. Fast global convergence of natural policy gradient methods with entropy regularization. arXiv preprint, arXiv:2007.06558, 2020.

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.

M. Fazel, R. Ge, S. Kakade, and M. Mesbahi. Global convergence of policy gradient methods for the linear quadratic regulator. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1467–1476. PMLR, 10–15 Jul 2018.

R. M. Gower, A. Defazio, and M. Rabbat. Stochastic Polyak stepsize with a moving target. arXiv preprint, arXiv:2106.11851, 2021.

S. Kakade. A natural policy gradient. In *Proceedings of the 14th International Conference on Neural Information Processing Systems (NIPS'01)*, pages 1531–1538, 2001.

References III

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.

H. Karimi, J. Nutini, and M. Schmidt. Linear convergence of gradient and proximal-gradient methods under the Polyak-łojasiewicz condition. In P. Frasconi, N. Landwehr, G. Manco, and J. Vreeken, editors, *Machine Learning and Knowledge Discovery in Databases (ECML PKDD 2016)*, volume 9851 of *Lectur Notes in Computer Sciencce*. Springer, 2016.

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.

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

B. Liu, Q. Cai, Z. Yang, and Z. Wang. Neural trust region/proximal policy optimization attains globally optimal policy. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

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.

References IV

J. Mei, C. Xiao, C. Szepesvári, and D. Schuurmans. On the global convergence rates of softmax policy gradient methods. In *Proceedings of the 37 th International Conference on Machine Learning (ICML)*, 2020.

J. Mei, Y. Gao, B. Dai, C. Szepesvári, and D. Schuurmans. Leveraging non-uniformity in first-order non-convex optimization. In *Proceedings of the 38 th International Conference on Machine Learning (ICML)*, 2021.

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

Y. Nesterov. *Introductory Lecture on Convex Optimization: A Basic Course*. Kluwer Academic Publishers, 2004.

Y. Nesterov. Gradient methods for minimizing composite functions. *Mathematical Programming*, 140:125–161, 2013.

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

References V

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.

R. S. Sutton, D. McAllester, S. Singh, and Y. Mansour. Policy gradient methods for reinforcement learning with function approximation. In *Advances in Neural Information Processing Systems*, volume 12, pages 1057–1063. MIT Press, 2000.

W. Zhan, S. Cen, B. Huang, Y. Chen, J. D. Lee, and Y. Chi. Policy mirror descent for regularized reinforcement learning: A generalized framework with linear convergence. arXiv preprint, arXiv:2105.11066, 2021.

Y. Zhou, Y. Liang, and L. Shen. A simple convergence analysis of bregman proximal gradient algorithm. *Computational Optimization and Applications*, 93:903–912, 2019.