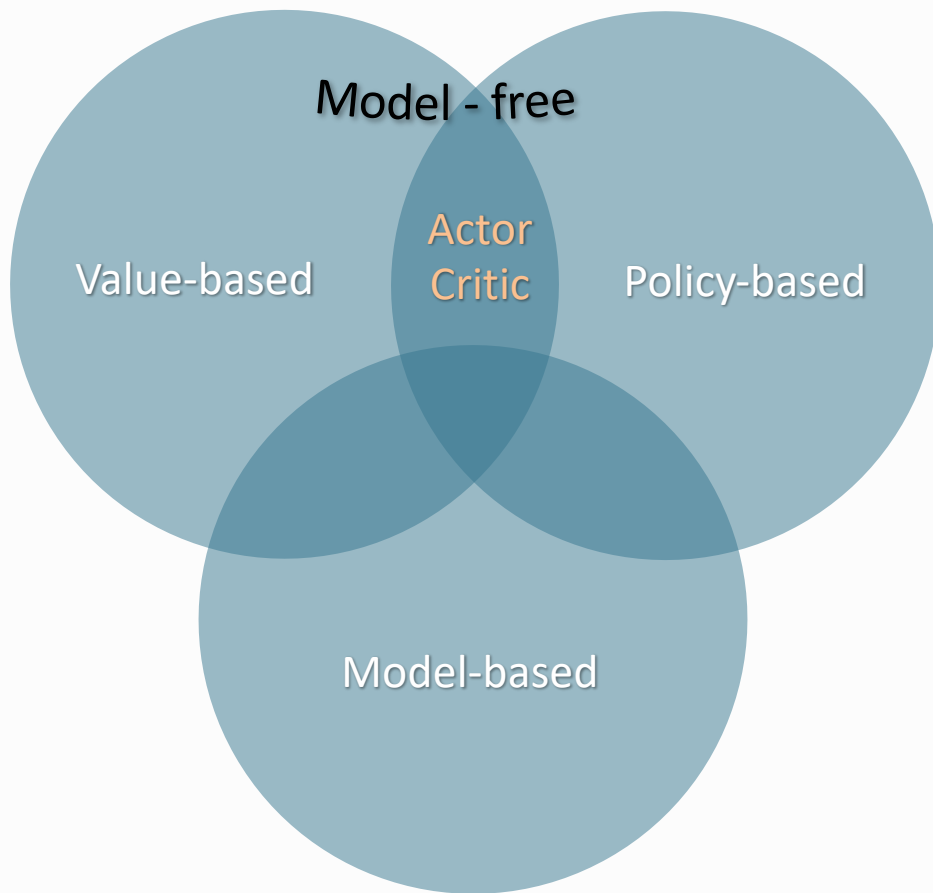


EPFL Summer School on Data Science, Optimization and Operations Research
August 15-20, 2021

Lecture 3: RL from Optimization Perspectives

Niao He, D-INFK, ETH Zurich

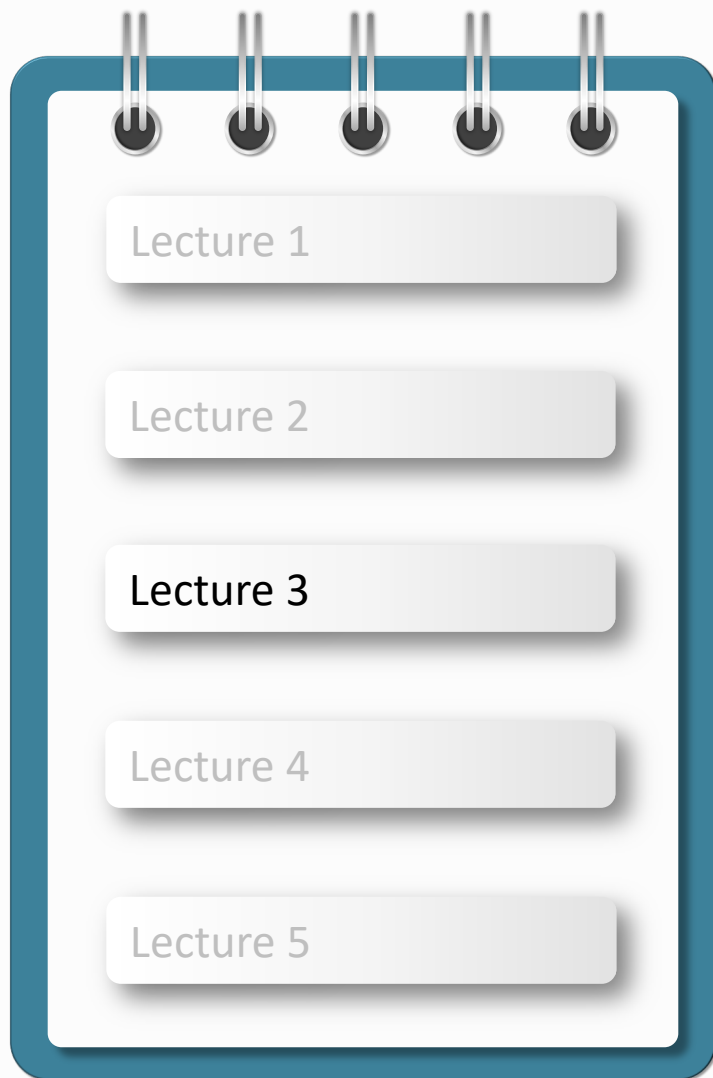
Recap: Reinforcement Learning Approaches



- **Value-based RL**
 - Estimate the optimal value function $Q^*(s, a)$
 - Example: Q-learning
- **Policy-based RL**
 - Search directly the optimal policy $\pi^*(\cdot | s)$
 - Example: Policy Gradient Method
- **Model-based RL**
 - First estimate the model P, R and then do planning

What are the convergence behaviors of these algorithms?

Outline of Lecture Series



Introduction to RL

RL from Control Perspectives
- Value-based RL

RL from Optimization Perspectives
- *Policy-based RL*

RL from Learning Perspectives
- Deep RL

RL from Game Perspectives

Focus:

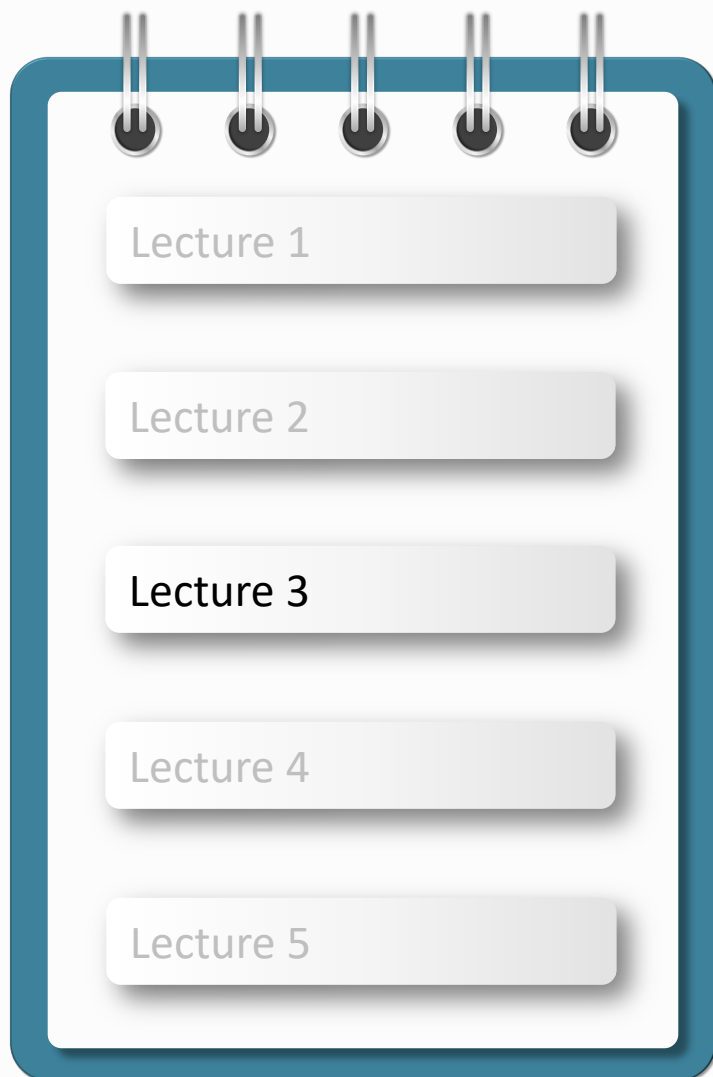
Convergence analysis of policy-based methods

- Policy gradient method
- Natural policy gradient method
- Variants w/o function approximation
- Variants w/o entropy regularization

Remark:

- **Assume** access to exact or stochastic gradient
- **Exploit** geometry and structure of RL objective

Outline of Lecture Series



Introduction to RL

RL from Control Perspectives
- Value-based RL

RL from Optimization Perspectives
- *Policy-based RL*

RL from Learning Perspectives
- Deep RL

RL from Game Perspectives

Focus:

Convergence analysis of policy-based methods

- Policy gradient method
- Natural policy gradient method
- Variants w/o function approximation
- Variants w/o entropy regularization

Gradient-based algorithms:

$$x_{k+1} = x_k + \alpha_k (\nabla f(x_k) + \epsilon_{k+1}), k = 0, 1, \dots$$

Notation Recap

MDP $(S, \mathcal{A}, P, R, \mu, \gamma)$

State value function:	$V^\pi(s) = \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t) s_0 = s \right]$
State-action value function:	$Q^\pi(s, a) = \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t) s_0 = s, a_0 = a \right]$
Optimal value function:	$V^*(s) := \max_{\pi} V^\pi(s), \quad Q^*(s, a) := \max_{\pi} Q^\pi(s, a)$
Optimal policy:	$\pi^*(s) = \operatorname{argmax}_{a \in \mathcal{A}} Q^*(s, a)$
Bellman equation:	$V^\pi(s) = \mathbb{E}_{a \sim \pi(s)} \left[R(s, a) + \gamma \mathbb{E}_{s' \sim P(\cdot s, a)} V^\pi(s') \right]$
Bellman optimality:	$Q^*(s, a) = R(s, a) + \mathbb{E}_{s' s, a} \left[\gamma \max_{a' \in \mathcal{A}} Q^*(s', a') \right]$
Policy gradient:	$\frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta} = \frac{1}{1 - \gamma} \mathbb{E}_{s \sim d_\mu^{\pi_\theta}, a \sim \pi_\theta(\cdot s)} \left[Q^{\pi_\theta}(s, a) \nabla \log \pi_\theta(a s) \right]$
State visitation distribution:	$d_\mu^\pi(s) = \mathbb{E}_{s_0 \sim \mu} \left[(1 - \gamma) \sum_{k=0}^{\infty} \gamma^k P(s_k = s s_0, \pi) \right]$

Policy Optimization

$$\max_{\theta \in \Theta} V^{\pi_\theta}(\mu) := \mathbb{E} \left[\sum_{k=0}^{\infty} \gamma^k r(s_k, a_k) \mid s_0 \sim \mu, s_{k+1} \sim P(\cdot | s_k, a_k), a_k \sim \pi_\theta(\cdot | s_k) \right]$$

Tabular MDPs

Direct
Parametrization

$$\pi_\theta(a|s) = \theta_{s,a}$$
$$\theta \in \Delta(\mathcal{A})^{|\mathcal{S}|}$$

Softmax
Parametrization

$$\pi_\theta(a|s) = \frac{\exp(\theta_{s,a})}{\sum_a \exp(\theta_{s,a})}$$

Function Approximation

Log-linear Policy

$$\pi_\theta(a|s) = \frac{\exp(\phi(s, a)^T \theta)}{\sum_a \exp(\phi(s, a)^T \theta)}$$

Neural Policy

$$\pi_\theta(a|s) = \frac{\exp(f_\theta(s, a))}{\sum_a \exp(f_\theta(s, a))}$$

$f_\theta(s, a)$ is a neural network

Policy Gradient Methods

$$\max_{\theta \in \Theta} V^{\pi_\theta}(\mu) := \mathbb{E} \left[\sum_{k=0}^{\infty} \gamma^k r(s_k, a_k) \middle| s_0 \sim \mu, s_{k+1} \sim P(\cdot | s_k, a_k), a_k \sim \pi_\theta(\cdot | s_k) \right]$$

Policy Gradient

$$\theta_{t+1} = \theta_t + \eta_t \nabla_\theta V^{\pi_{\theta_t}}(\mu)$$

Fisher information matrix

$$G^{\pi_\theta}(\mu) := \mathbb{E}_{s \sim d_\mu^{\pi_\theta}, a \sim \pi_\theta(\cdot | s)} [(\nabla_\theta \log \pi_\theta(a | s))(\nabla_\theta \log \pi_\theta(a | s))^T]$$

Natural Policy Gradient

(Kakade '01)

$$\theta_{t+1} = \theta_t + \eta_t [G^{\pi_{\theta_t}}(\mu)]^{-1} \nabla_\theta V^{\pi_{\theta_t}}(\mu)$$

Softmax NPG = soft policy iteration = Mirror Descent

$$\pi_{t+1}(a | s) \propto \pi_t(a | s) \exp(\eta_t Q^{\pi_t}(s, a) / (1 - \gamma))$$

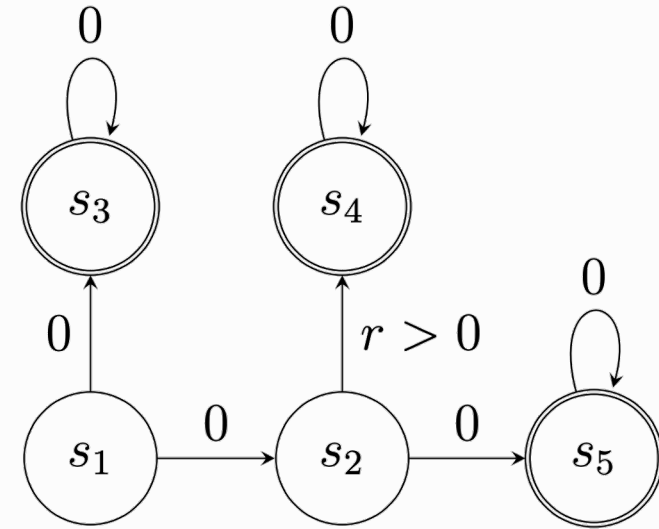
Non-concavity

- The objective is **non-concave** for direct or softmax parametrization. [Agarwal et al. 2020]

$$V^\pi(s_1) = \pi(\text{right}|s_1) \pi(\text{up}|s_2) \cdot r$$

$$V^{\pi_\theta}(s_1) = \frac{e^{\theta_{\text{right}, s_1}}}{e^{\theta_{\text{right}, s_1}} + e^{\theta_{\text{up}, s_1}}} \cdot \frac{e^{\theta_{\text{up}, s_2}}}{e^{\theta_{\text{right}, s_2}} + e^{\theta_{\text{up}, s_2}}} \cdot r$$

$$\theta = (\theta_{\text{right}, s_1}, \theta_{\text{up}, s_1}, \theta_{\text{right}, s_2}, \theta_{\text{up}, s_2})$$



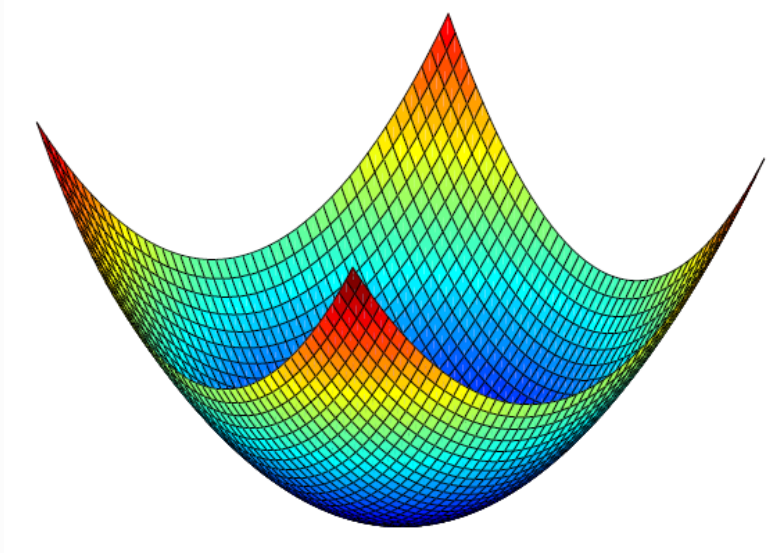
$$\theta_1 = (\log 1, \log 3, \log 3, \log 1), V^{\pi_{\theta_1}}(s_1) = \frac{1}{4} \cdot \frac{1}{4} \cdot r$$

$$\theta_2 = (-\log 1, -\log 3, -\log 3, -\log 1), V^{\pi_{\theta_2}}(s_1) = \frac{3}{4} \cdot \frac{3}{4} \cdot r$$

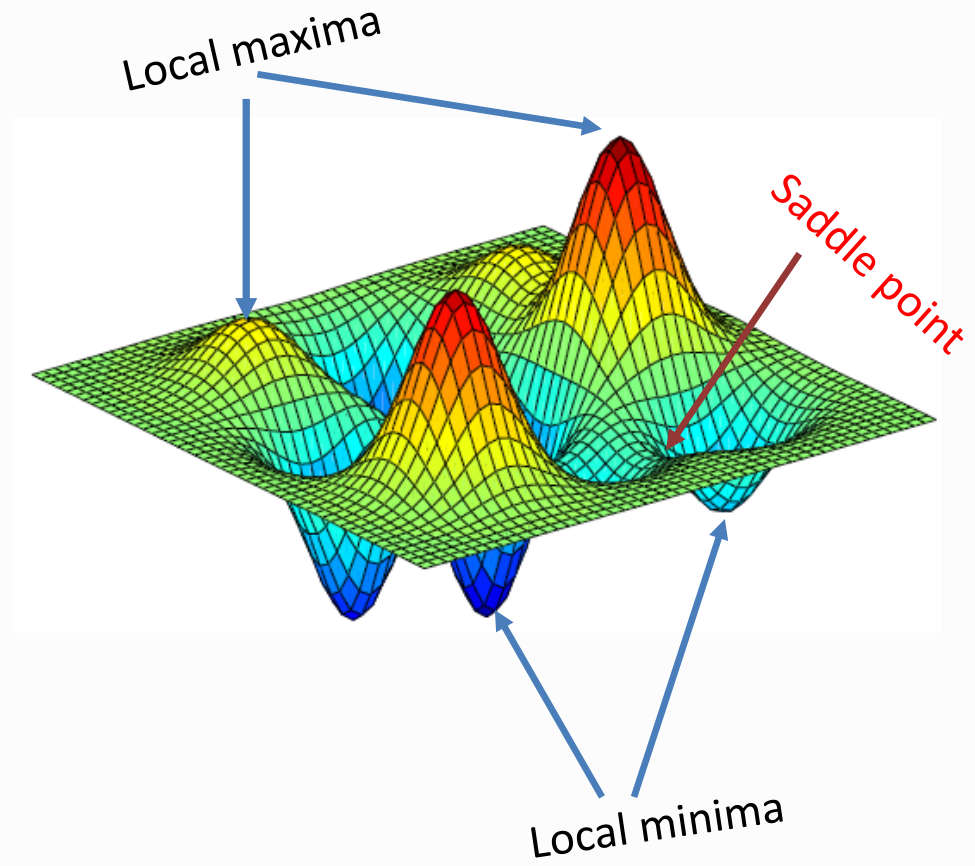
$$\bar{\theta} = (0, 0, 0, 0), V^{\pi_{\bar{\theta}}}(s_1) = \frac{1}{2} \cdot \frac{1}{2} \cdot r$$

$$\frac{V^{\pi_{\theta_1}}(s_1) + V^{\pi_{\theta_2}}(s_1)}{2} = \frac{5}{8}r > V^{\pi_{\bar{\theta}}}(s_1)$$

A Refresher on Nonconvex Optimization



Local minima = Global minima



Gradient Descent

$$\min_{x \in \mathbb{R}^n} f(x)$$

- Assume $f(x)$ is *L-smooth*, namely $\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|$.

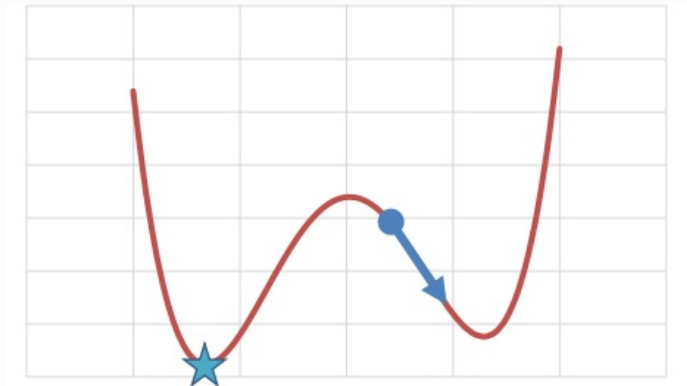
- Gradient Descent:**

$$x_{t+1} = x_t - \frac{1}{L} \nabla f(x_t)$$

- Convergence to a stationary point [Nesterov, 2004]:

$$\min_{t=1, \dots, T} \|\nabla f(x_t)\|^2 \leq \frac{2L(f(x_1) - f^*)}{T}$$

- Descent property: $f(x_t) - f(x_{t+1}) \geq \frac{1}{2L} \|\nabla f(x_t)\|_2^2$



Stochastic Gradient Descent

$$\min_{x \in \mathbb{R}^n} f(x) = \mathbb{E}[F(x, \xi)]$$

- Assume $f(x)$ is L -smooth and $\mathbb{E}\|\nabla F(x, \xi) - \nabla f(x)\|^2 \leq \sigma^2$.

- Stochastic Gradient Descent:**

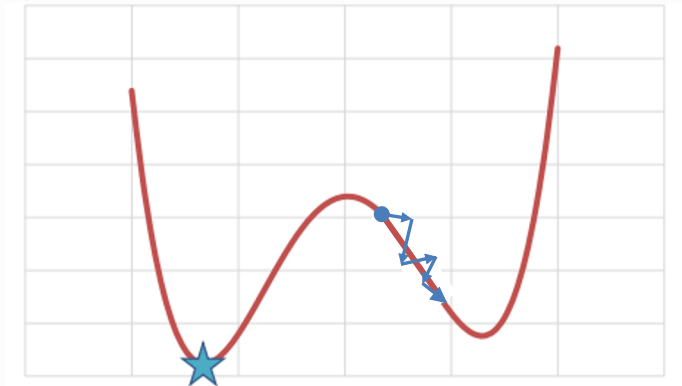
$$x_{t+1} = x_t - \eta_t \nabla F(x_t, \xi_t)$$

- Convergence rate:**

$$\eta_t = \min \left\{ \frac{1}{L}, \frac{\eta}{\sigma\sqrt{T}} \right\}$$

$$\mathbb{E}[\|\nabla f(\hat{x}_T)\|^2] \leq \frac{\sigma}{\sqrt{T}} \left(\frac{2(f(x_1) - f(x^*))}{\eta} + L\eta \right)$$

- Here $\hat{x}_T$ is selected uniformly at random from $\{x_t\}_{t=1, \dots, T}$
- Sample complexity: $O(\epsilon^{-4})$ samples to find a point x with $\mathbb{E}\|\nabla f(x)\| \leq \epsilon$.
- Lower bound: $O(\Delta L \sigma^2 \epsilon^{-4})$ stochastic oracle queries [Arjevani et al., 2019]



Proof of SGD Convergence

Proof

$$\begin{aligned} & \mathbb{E}[f(x_{t+1}) - f(x_t)] \\ & \leq \mathbb{E} \left[\nabla f(x_t)^T (x_{t+1} - x_t) + \frac{L}{2} \|x_{t+1} - x_t\|^2 \right] \\ & = \mathbb{E} \left[-\eta_t \nabla f(x_t)^T \nabla F(x_t, x_t) + \frac{L\eta_t^2}{2} \|\nabla F(x_t, \xi_t)\|^2 \right] \\ & \leq -(\eta_t - L\eta_t^2/2) \mathbb{E} \|\nabla f(x_t)\|^2 + \frac{L\sigma^2\eta_t^2}{2} \\ & \leq -\frac{\eta_t}{2} \mathbb{E} \|\nabla f(x_t)\|^2 + \frac{L\sigma^2\eta_t^2}{2} \end{aligned}$$
$$\mathbb{E} \|\nabla F(\hat{x}_T)\|^2 \leq \frac{2(f(x_1) - f(x^*)) + L\sigma^2 \sum_{t=1}^T \eta_t^2}{\sum_{t=1}^T \eta_t}$$

Smoothness of $f(x)$

Definition of x_{t+1}

Use $E(X^2) = E(X)^2 + V(X)$

Use $\eta_t \leq \frac{1}{L}$



Variance-reduced SGD

$$\min_{x \in \mathbb{R}^n} f(x) = \mathbb{E}[F(x, \xi)]$$

- Assume $f(x)$ is $\bar{L}$ -average smooth and $\mathbb{E} \|\nabla F(x, \xi) - \nabla f(x)\|^2 \leq \sigma^2$.
- **SGD with variance-reduced recursive gradient estimator:**

$$x_{t+1} = x_t - \eta_t g_t,$$
$$g_t = \begin{cases} \frac{1}{S} \sum_{k=1}^S \nabla F(x_t, \xi_t^k), & \text{if } t \equiv 0 \pmod{Q}; \\ (1 - \alpha) \left(g_{t-1} - \frac{1}{D} \sum_{k=1}^D \nabla F(x_{t-1}, \xi_t^k) \right) + \frac{1}{D} \sum_{k=1}^D \nabla F(x_t, \xi_t^k), & \text{otherwise,} \end{cases}$$

- Special cases: SPIDER [Fang et al., 2018], SARAH [Nguyen et al., 2017], STORM [Cutkosky & Orabona, 2019]
- Sample complexity: $O(\epsilon^{-3})$ samples to find a point x with $\mathbb{E} \|\nabla f(x)\| \leq \epsilon$.
- Lower bound: $O(\Delta \bar{L} \sigma \epsilon^{-3} + \sigma^2 \epsilon^{-2})$ stochastic oracle queries [Arjevani et al., 2019]

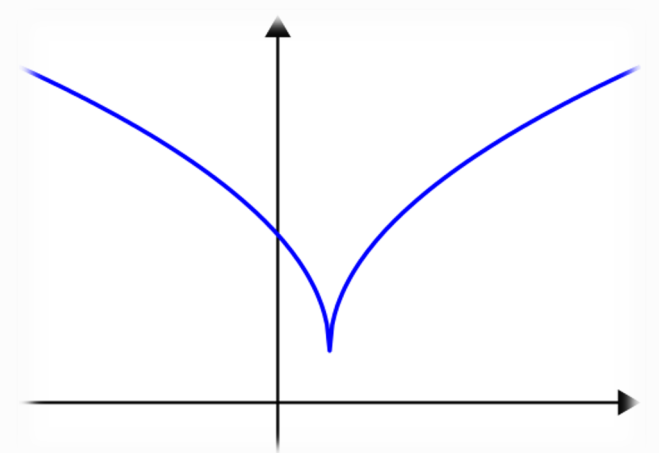
Convergence to global optimum under KL inequality

Under some “convex-like” assumptions, GD/SGD could converge to the global optimal solution:

- **Kurdyka-Lojasiewicz inequality** ($p < 1$)

$$\|\nabla f(x)\| \geq C \cdot (f(x) - f^*)^{1-p}, \quad \forall x$$

- **Polyak-Lojasiewicz condition** (gradient dominance): $p = \frac{1}{2}$
 - GD converges linearly to the global optimum. [Polyak, 1963]
- **Weak Lojasiewicz condition** (weak gradient dominance): $p = 0$
 - GD converges to the global optimum in the order of $O\left(\frac{1}{t}\right)$. [Nesterov, 2004]



Back to Policy Gradient Methods

- Policy Gradient Theorem:

$$\frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta} = \frac{1}{1-\gamma} \mathbb{E}_{s \sim d_\mu^{\pi_\theta}, a \sim \pi_\theta(\cdot|s)} [Q^{\pi_\theta}(s, a) \nabla \log \pi_\theta(a|s)]$$

- Construction of unbiased Monte-Carlo policy gradient: sample $\{s_t, a_t\}_{t=1, \dots, H} \sim P_\mu^{\pi_\theta}$

$$\begin{aligned} \widehat{Q}^{\pi_\theta}(s_t, a_t) &:= \sum_{t'=t}^{\infty} \gamma^{t'-t} r(s_{t'}, a_{t'}) \\ \widehat{\nabla V^{\pi_\theta}}(\mu) &:= \sum_{t=0}^{\infty} \gamma^t \widehat{Q}^{\pi_\theta}(s_t, a_t) \nabla \log \pi_\theta(a_t|s_t) \end{aligned}$$

- In general, only convergence to a stationary point when using GD/SGD methods.

Recent Breakthrough on Global Convergence of PG and NPG

- Tabular setting + softmax parametrization + exact gradient
[Bhandari and Russo, 2019; Mei et al., 2020&2021; Agarwal et al., 2020; Cen et al., 2020; Lan et al., 2021; ...]
 - $O(1/T)$ rate of PG and NPG [Mei et al., 2020]
 - $O(\exp(-T))$ rate of PG with entropy regularization [Mei et al., 2020]
 - $O(\exp(-T))$ rate of NPG with entropy regularization [Cen et al., 2020]
 - $O(\exp(-T))$ rate of NPG with general convex regularization [Lan et al., 2021]
 - $O(\exp(-T))$ rate of normalized PG [Mei et al., 2021]
- Linear function approximation setting + exact gradient
[Agarwal et al., 2020; Wang et al., 2019]
 - $O(1/\sqrt{T})$ rate of NPG with log-linear policy
 - $O(\exp(-T))$ rate of NPG with entropy regularization [Cayci, H., Srikant, 2020]

Key insight behind the convergence of PG

- Performance Difference Lemma [Kakade & Langford, 2002]

$$V^\pi(\mu) - V^{\pi'}(\mu) = \frac{1}{1 - \gamma} \mathbb{E}_{s \sim d_\mu^\pi, a \sim \pi(\cdot|s)} [A^{\pi'}(s, a)]$$

$$A^\pi(s, a) = Q^\pi(s, a) - V^\pi(s)$$

- Policy gradient with softmax tabular parametrization [Agarwal et al. 2020]

$$\frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta_{s,a}} = \frac{1}{1 - \gamma} d_\mu^{\pi_\theta}(s) \pi_\theta(a|s) A^{\pi_\theta}(s, a)$$

$$\pi_\theta(a|s) = \frac{\exp(\theta_{s,a})}{\sum_a \exp(\theta_{s,a})}$$

- Lojasiewicz inequality condition** [Mei et al., 2020]:

$$\left\| \frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta} \right\|_2 \geq \frac{\min_s \pi_\theta(a^*(s)|s)}{\sqrt{S} \cdot \|d_\mu^{\pi^*} / d_\mu^{\pi_\theta}\|_\infty} \cdot (V^*(\mu) - V^{\pi_\theta}(\mu))$$

Proof of Performance Difference Lemma

$$V^\pi(\mu) - V^{\pi'}(\mu) = \frac{1}{1-\gamma} \mathbb{E}_{s \sim d_\mu^\pi, a \sim \pi(\cdot|s)} [A^{\pi'}(s, a)]$$

Proof:

$$\begin{aligned} V^\pi(s) - V^{\pi'}(s) &= \mathbb{E}_{\tau \sim \mathbf{Pr}^\pi(\tau|s_0=s)} \left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \right] - V^{\pi'}(s) \\ &= \mathbb{E}_{\tau \sim \mathbf{Pr}^\pi(\tau|s_0=s)} \left[\sum_{t=0}^{\infty} \gamma^t \left(r(s_t, a_t) + V^{\pi'}(s_t) - V^{\pi'}(s_t) \right) \right] - V^{\pi'}(s) \\ &\stackrel{(a)}{=} \mathbb{E}_{\tau \sim \mathbf{Pr}^\pi(\tau|s_0=s)} \left[\sum_{t=0}^{\infty} \gamma^t \left(r(s_t, a_t) + \gamma V^{\pi'}(s_{t+1}) - V^{\pi'}(s_t) \right) \right] \\ &\stackrel{(b)}{=} \mathbb{E}_{\tau \sim \mathbf{Pr}^\pi(\tau|s_0=s)} \left[\sum_{t=0}^{\infty} \gamma^t \left(r(s_t, a_t) + \gamma \mathbb{E}[V^{\pi'}(s_{t+1})|s_t, a_t] - V^{\pi'}(s_t) \right) \right] \\ &\stackrel{(c)}{=} \mathbb{E}_{\tau \sim \mathbf{Pr}^\pi(\tau|s_0=s)} \left[\sum_{t=0}^{\infty} \gamma^t \left(Q^{\pi'}(s_t, a_t) - V^{\pi'}(s_t) \right) \right] \\ &= \mathbb{E}_{\tau \sim \mathbf{Pr}^\pi(\tau|s_0=s)} \left[\sum_{t=0}^{\infty} \gamma^t A^{\pi'}(s_t, a_t) \right] = \frac{1}{1-\gamma} \mathbb{E}_{s' \sim d_s^\pi} \mathbb{E}_{a \sim \pi(\cdot|s)} \gamma^t A^{\pi'}(s', a), \end{aligned}$$



Global Convergence of Softmax Policy Gradient

- Tabular setting + softmax parametrization + exact gradient

Theorem [Mei et al., 2020] Set stepsize $\eta = (1 - \gamma)^3/8$ and assume $\min \mu(s) > 0$, then for all $t \geq 1$,

$$V^*(\mu) - V^{\pi_{\theta_t}}(\mu) \leq \frac{16S}{c^2(1 - \gamma)^6 t} \cdot \left\| \frac{d\pi^*}{\mu} \right\|_{\infty}^2 \left\| \frac{1}{\mu} \right\|_{\infty}$$

where $c = \min_{s,t} \pi_{\theta_t}(a^*(s)|s) > 0$.

Proof: Smoothness + Lojasiewicz condition $\Rightarrow O\left(\frac{1}{t}\right)$ rate of GD with constant stepsize

Policy Gradient Methods with Entropy Regularization

- **Regularized objective:**

$$\min_{\pi} V_{\lambda}^{\pi}(\mu) := V^{\pi}(\mu) + \lambda H^{\pi}(\mu)$$

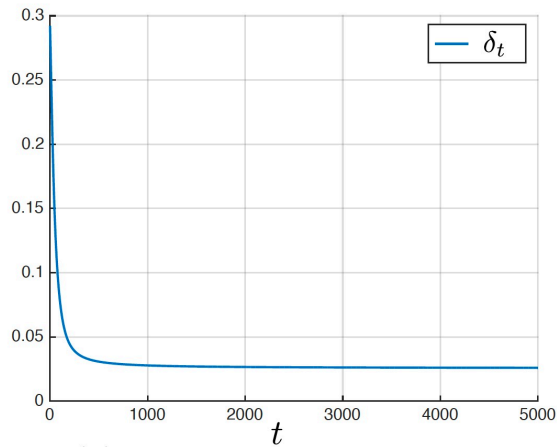
- Value function: $V^{\pi}(\mu) := \mathbb{E}_{\pi} \left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) | s_0 \sim \mu \right]$
- Entropy regularizer: $H^{\pi}(\mu) := \mathbb{E}_{\pi} \left[\sum_{t=0}^{\infty} -\gamma^t \log \pi(a_t | s_t) | s_0 \sim \mu \right]$
- The regularized objective is still nonconcave.
- $\lambda > 0 \Rightarrow$ encourage exploration, $\min_{a,s} \pi_{\theta}(a|s) > 0$

- **Entropy-regularized policy gradient:**

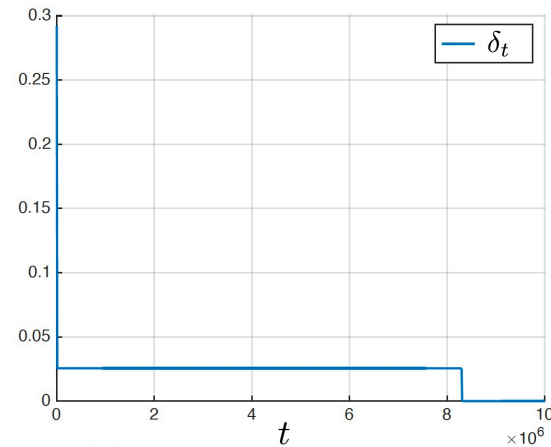
$$\theta_{t+1} = \theta_t + \eta_t \nabla_{\theta} V_{\lambda}^{\pi_{\theta_t}}(\mu)$$

The effect of entropy regularization

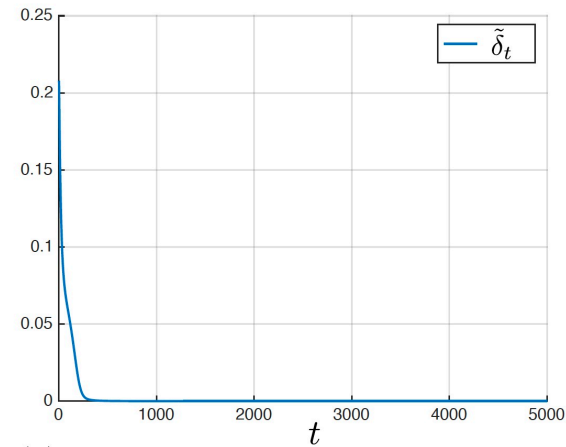
- Using entropy regularization leads to faster convergence.



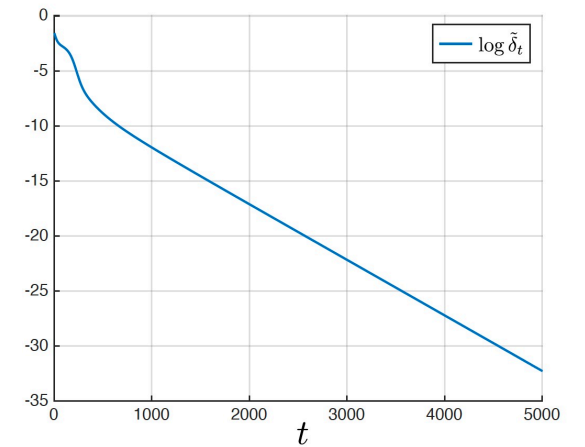
(a) Softmax policy gradient.



(b) Softmax policy gradient.



(c) Entropy regularized softmax PG.



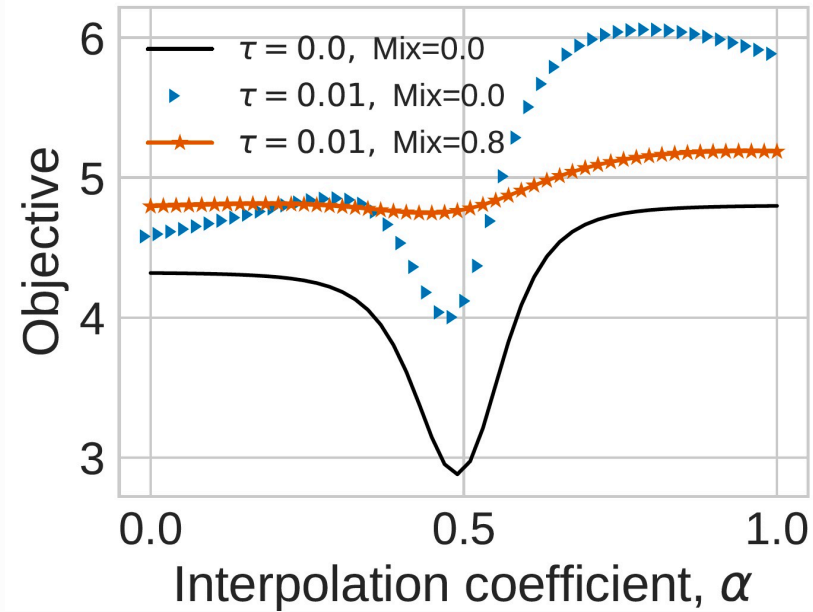
(d) Entropy regularized softmax PG.

One-state MDP with 20 actions [Mei et al., 2020]

The effect of entropy regularization

- Using entropy regularization yields better optimization landscape and geometry.

Start				4.5
5.0				



Gridworld, log-linear policy, exact policy gradient [Ahmed et al., 2019]

Linear Convergence of Entropy-regularized Softmax Policy Gradient

- Tabular setting + softmax parametrization + exact gradient + entropy regularization

Theorem [Mei et al., 2020] Set stepsize $\eta = (1 - \gamma)^3 / (8 + \lambda(4 + \log A))$ and assume $\min \mu(s) > 0$, then for all $t \geq 1$,

$$V_{\lambda}^*(\mu) - V_{\lambda}^{\pi_{\theta_t}}(\mu) \leq \left\| \frac{1}{\mu} \right\|_{\infty} \cdot \frac{1 + \lambda \log A}{(1 - \gamma)^2} e^{-C(t-1)}$$

where C depends on $\min_{s,t} \pi_{\theta_t}(a^*(s)|s) > 0$.

Proof: Smoothness + Polyak-Lojasiewicz condition $\Rightarrow$ linear rate of GD with constant stepsize

Entropy Regularization as an Optimization Aid

- Tabular setting + softmax parametrization + exact gradient

$$\theta_{t+1} = \theta_t + \eta_t \nabla_{\theta} V_{\lambda}^{\pi_{\theta_t}}(\mu)$$

$$\pi_{\theta}(a|s) = \frac{\exp(\theta_{s,a})}{\sum_a \exp(\theta_{s,a})}$$

- **Global convergence**

$\lambda = 0$: without entropy regularization

PG: $O(1/T)$ rate [Mei et al., '20]

NPG: $O(1/T)$ rate [Agarwal et al., '20]

$$\left\| \frac{\partial V^{\pi_{\theta}}(\mu)}{\partial \theta} \right\|_2 \geq C(\theta) \min_s \pi(a^*(s)|s) [V^*(\mu) - V^{\pi_{\theta}}(\mu)]$$

$\lambda > 0$: with entropy regularization

PG: $O(\exp(-T))$ rate [Mei et al., '20]

NPG: $O(\exp(-T))$ rate [Cen et al., '20, Lan '21]

$$\left\| \frac{\partial V_{\lambda}^{\pi_{\theta}}(\mu)}{\partial \theta} \right\|_2 \geq C(\theta) \min_{a,s} \pi_{\theta}(a|s) [V_{\lambda}^*(\mu) - V_{\lambda}^{\pi_{\theta}}(\mu)]^{1/2}$$

What about function approximation setting?

Policy Optimization

$$\min_{\pi} V_{\lambda}^{\pi}(\mu) := V^{\pi}(\mu) + \lambda H^{\pi}(\mu)$$

Tabular MDPs

Direct Parametrization

$$\pi_{\theta}(a|s) = \theta_{s,a}$$
$$\theta \in \Delta(\mathcal{A})^{|\mathcal{S}|}$$

Softmax Parametrization

$$\pi_{\theta}(a|s) = \frac{\exp(\theta_{s,a})}{\sum_a \exp(\theta_{s,a})}$$

Function Approximation

Log-linear Policy

$$\pi_{\theta}(a|s) = \frac{\exp(\phi(s,a)^T \theta)}{\sum_a \exp(\phi(s,a)^T \theta)}$$

Neural Policy

$$\pi_{\theta}(a|s) = \frac{\exp(f_{\theta}(s,a))}{\sum_a \exp(f_{\theta}(s,a))}$$

$f_{\theta}(s,a)$ is a neural network

Natural Policy Gradient

- Fisher info matrix:

$$G^{\pi_{\theta}}(\mu) := \mathbb{E}_{s \sim d_{\mu}^{\pi_{\theta}}, a \sim \pi_{\theta}(\cdot|s)} [(\nabla_{\theta} \log \pi_{\theta}(a|s))(\nabla_{\theta} \log \pi_{\theta}(a|s))^T]$$

- NPG update:

$$\theta_{t+1} = \theta_t + \eta_t [G^{\pi_{\theta_t}}(\mu)]^{-1} \cdot \nabla_{\theta} V_{\lambda}^{\pi_{\theta_t}}(\mu)$$

- Mirror Descent:

$$\theta_{t+1} = \operatorname{argmax}_{\theta} \left\{ \nabla_{\theta} V_{\lambda}^{\pi_{\theta_t}}(\mu)^T (\theta - \theta_t) - \frac{1}{2\eta_t} D(\theta, \theta_t) \right\}$$

- Bregman divergence: $D(\theta, \theta_t) = (\theta - \theta_t)^T G^{\pi_{\theta_t}}(\mu) (\theta - \theta_t)$
- Quadratic approximation of KL divergence: $KL(P^{\pi_{\theta_t}}(\tau) || P^{\pi_{\theta}}(\tau))$

Compatible Function Approximation

- Define the function

$$L(w, \theta) = \mathbb{E}_{s \sim d_\mu^{\pi_\theta}, a \sim \pi_\theta(\cdot|s)} \left[\left(\nabla_\theta^\top \log \pi_\theta(a|s) w - (Q_\lambda^{\pi_\theta}(s, a) - \lambda \log \pi_\theta(a|s)) \right)^2 \right]$$

$$w_\theta^* \in \arg \min_{w \in \mathbb{R}^d} L(w, \theta) \quad \longleftrightarrow \quad w_\theta^* = \frac{1}{1 - \gamma} [G^{\pi_\theta}(\mu)]^{-1} \cdot \nabla_\theta V_\lambda^{\pi_\theta}(\mu)$$

- Compute natural policy gradient by minimizing $L(w, \theta_t)$.
- Approximation error: $\min_w L(w, \theta_t) \leq \epsilon_{app}, \forall t$
- Measures the expressiveness of the function approximation

(Entropy-regularized) NPG with Linear Function Approximation

Natural Policy Gradient

$$w_t \in \arg \min_{w \in \mathbb{R}^d} L(w, \theta_t)$$

$$\theta_{t+1} = \theta_t + \eta w_t$$

$$L(w, \theta) = \mathbb{E}_{s \sim d_\mu^{\pi_\theta}, a \sim \pi_\theta(\cdot|s)} \left[\left(\nabla_\theta^\top \log \pi_\theta(a|s) w - (Q_\lambda^{\pi_\theta}(s, a) - \lambda \log \pi_\theta(a|s)) \right)^2 \right]$$

- Softmax parametrization with linear function approximation: $\pi_\theta(a|s) = \frac{\exp(\phi_{s,a}^T \theta)}{\sum_a \exp(\phi_{s,a}^T \theta)}$
- Equivalently NPG update: $\pi_{t+1}(a|s) \leftarrow \frac{\pi_t(a|s) \exp(w_t \cdot \phi_{s,a})}{Z}$

Convergence of NPG with Linear Function Approximation

- Sublinear convergence up to approximation error:

Theorem [Agarwal et al., 2020] Set stepsize $\eta = \sqrt{2 \log |A| / (\beta W^2 T)}$,

$$\min_{t < T} \left\{ V^{\pi^*}(\rho) - V^{\pi_t}(\rho) \right\} \leq \frac{W}{1 - \gamma} \frac{\sqrt{2\beta \log |A|}}{T} + \frac{\sqrt{4|A| \|d^*/\nu\|_\infty \cdot \epsilon_{app}}}{(1 - \gamma)^{3/2}}$$

Key assumptions:

- Distribution mismatch coefficient: $\|d^*/\nu\|_\infty \leq C^*$
- Relative condition number: $\mathbb{E} \left[\sup_{w \in \mathbb{R}^d} \frac{w^\top \Sigma_{d^*}^{(t)} w}{w^\top \Sigma_\nu^{(t)} w} \right] \leq \kappa. \quad \Sigma_\nu = \mathbb{E}_{s,a \sim \nu} [\phi_{s,a} \phi_{s,a}^\top]$
- Policy smoothness: $\log \pi_\theta(a|s)$ is β -smooth, $\|w\| \leq W$

Linear Convergence of Entropy-regularized NPG

- Linear convergence up to approximation error:

Theorem [Cayci, H., Srikant, 2021] Set stepsize $\eta = \frac{(1-\gamma)\sigma^2 r_{\min}}{(r_{\max} + \lambda \log A)^2}$

$$\Phi(\pi_T) := \sum_s d_{\mu}^{\pi^*} D_{KL}(\pi^* || \pi_T) \leq (1 - \eta\lambda)^T \log |\mathcal{A}| + \frac{\sqrt{C^* \cdot \epsilon_{app}}}{\lambda},$$

$$V_{\lambda}^{\pi^*}(\mu) - V_{\lambda}^{\pi_T}(\mu) \leq (1 - \eta\lambda)^T \frac{\log |\mathcal{A}|}{\eta(1 - \gamma)} + \frac{\sqrt{C^* \cdot \epsilon_{app}}}{\lambda\eta(1 - \gamma)},$$

where $\sigma := \min_t \sigma_1 \left(G_{\lambda}^{\pi_{\theta_t}}(\mu) \right) > 0$.

Key assumptions:

- Concentrability coefficient: $C_t = \mathbb{E}_{s \sim d_{\mu}^{\pi_t}, a \sim \pi_t(\cdot|s)} \left[\left(\frac{d_{\mu}^{\pi^*}(s) \pi^*(a|s)}{d_{\mu}^{\pi_t}(s) \pi_t(a|s)} \right)^2 \right] \leq C^* < \infty$.
- Regularity of the parametric model : $F(\mu) = \mathbb{E}_{s \sim \mu, a \sim \text{Unif}(\mathcal{A})} \left[(\phi_{s,a} - \bar{\phi}_s)(\phi_{s,a} - \bar{\phi}_s)^{\top} \right] \succ 0$.

Proof Sketch

Entropy
Regularization

$$\Phi(\pi_{t+1}) - \Phi(\pi_t) \leq -\eta\lambda\Phi(\pi_t) - \eta(1 - \gamma) \left(V_{\lambda}^{\pi^*}(\mu) - V_{\lambda}^{\pi_t}(\mu) \right)$$

$$- \eta E_{s \sim d_{\mu}^{\pi^*}, a \sim \pi^*(\cdot|s)} \left[w_t^{\top} \nabla_{\theta} \log \pi_t(a|s) - Q_{\lambda}^{\pi_t}(s, a) + \lambda \log \pi_t(a|s) \right]$$

$$- \eta E_{s \sim d_{\mu}^{\pi^*}} \left[V_{\lambda}^{\pi_t}(s) \right] + \frac{1}{2} \eta^2 \|w_t\|_2^2$$

Approximation error

Negative with small η

Entropy Regularization as an Optimization Aid

- **Linear function approximation** + softmax parametrization + exact gradient

$$\theta_{t+1} = \theta_t + \eta_t [G^{\pi_{\theta_t}}(\mu)]^{-1} \cdot \nabla_{\theta} V_{\lambda}^{\pi_{\theta_t}}(\mu)$$

$$\pi_{\theta}(a|s) = \frac{\exp(\phi_{s,a}^T \theta)}{\sum_a \exp(\phi_{s,a}^T \theta)}$$

- **Global convergence** (up to approximation error)

$\lambda = 0$: without entropy regularization

NPG: $O(1/\sqrt{T})$ rate [Agarwal et al., '20]

$\lambda > 0$: **with entropy regularization**

NPG: $O(\exp(-T))$ rate [Cayci, H., Srikant, '21]

- Extension to Q-NPG variants

Summary

- Convergence to stationary point for policy gradient methods in general.
- Convergence to global optimum for policy gradient methods in tabular setting (despite nonconvexity).
- Using entropy regularization leads to faster convergence and encourages exploration.
- Exploiting the geometry: Lojasiewicz inequality, mirror descent, etc.

Open Questions

- Nonlinear function approximations?
- Stochastic policy gradient methods?
- Global optimality and sample complexity?

Reference

- **[Mei et al., 2020]** On the Global Convergence Rates of Softmax Policy Gradient Methods. In ICML 2021.
- **[Agarwal et al., 2021]** On the Theory of Policy Gradient Methods: Optimality, Approximation, and Distribution Shift. JMLR 2021.
- **[Cayci, H., Srikant, 2021]** Linear Convergence of Entropy-Regularized Natural Policy Gradient with Linear Function Approximation. arXiv preprint arXiv:2106.04096, 2021.
- A Tutorial on Policy Gradient Methods. SIAM Conference on Optimization. Lin Xiao and Lihong Li.