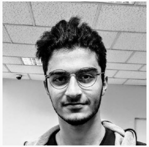
Towards Principled, Practical Policy Gradient for Bandits and Tabular MDPs

Sharan Vaswani (Simon Fraser University)

Joint work with: Michael Lu, Matin Aghaei, Anant Raj

RL Theory Workshop, 2024

Motivation

- Policy Gradient (PG) methods are widely used in practice.
- ✓ The policy gradient objective is non-concave. Under smoothness assumptions, PG methods can attain convergence to a stationary point.
- ✓ In certain settings (e.g. with a tabular parameterization), vanilla PG methods can achieve global convergence to the optimal policy [Agarwal et al., 2021, Mei et al., 2020, 2023, Yuan et al., 2022].

Motivation

- Policy Gradient (PG) methods are widely used in practice.

✓ The policy gradient objective is non-concave. Under smoothness assumptions, PG methods can attain convergence to a stationary point.

✓ In certain settings (e.g. with a tabular parameterization), vanilla PG methods can achieve global convergence to the optimal policy [Agarwal et al., 2021, Mei et al., 2020, 2023, Yuan et al., 2022].

- Prior theoretically principled PG methods:
  - × Require oracle-like knowledge about the environment (e.g. optimal action, the reward gap in multi-armed bandits) to set algorithm parameters, making them impractical.
  - × Use conservative choices of algorithm parameters and result in poor empirical performance.

Motivation

- Policy Gradient (PG) methods are widely used in practice.
- ✓ The policy gradient objective is non-concave. Under smoothness assumptions, PG methods can attain convergence to a stationary point.
- ✓ In certain settings (e.g. with a tabular parameterization), vanilla PG methods can achieve global convergence to the optimal policy [Agarwal et al., 2021, Mei et al., 2020, 2023, Yuan et al., 2022].
- Prior theoretically principled PG methods:
  - × Require oracle-like knowledge about the environment (e.g. optimal action, the reward gap in multi-armed bandits) to set algorithm parameters, making them impractical.
  - × Use conservative choices of algorithm parameters and result in poor empirical performance.
- **Aim**: Design practical PG algorithms while retaining theoretical guarantees.

Motivation

- Policy Gradient (PG) methods are widely used in practice.

✓ The policy gradient objective is non-concave. Under smoothness assumptions, PG methods can attain convergence to a stationary point.

✓ In certain settings (e.g. with a tabular parameterization), vanilla PG methods can achieve global convergence to the optimal policy [Agarwal et al., 2021, Mei et al., 2020, 2023, Yuan et al., 2022].

- Prior theoretically principled PG methods:
  - × Require oracle-like knowledge about the environment (e.g. optimal action, the reward gap in multi-armed bandits) to set algorithm parameters, making them impractical.
  - × Use conservative choices of algorithm parameters and result in poor empirical performance.
- **Aim**: Design practical PG algorithms while retaining theoretical guarantees.
- **This talk**: An optimization perspective on (stochastic) unregularized softmax policy gradient methods in the tabular setting (finite states/actions) with a focus on developing practical algorithms.

Outline

- **Problem Formulation**
- Softmax Policy Gradient
- Stochastic Softmax Policy Gradient
- Conclusion

Problem Formulation

- Infinite-horizon discounted MDP: $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, p, r, \rho, \gamma \rangle$ with finite states and actions ($S = |\mathcal{S}|$ and $A = |\mathcal{A}|$).

Problem Formulation

- Infinite-horizon discounted MDP: $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, p, r, \rho, \gamma \rangle$ with finite states and actions ($S = |\mathcal{S}|$ and $A = |\mathcal{A}|$).
- Distributions induced by policy π: For each state $s \in \mathcal{S}$, $\pi(\cdot|s)$ over actions. State occupancy measure: $d^\pi(s) = (1-\gamma) \sum_{\tau=0}^{\infty} \gamma^\tau \mathbb{P}(s_\tau = s \mid s_0 \sim \rho, a_\tau \sim \pi(\cdot|s_\tau))$.

Problem Formulation

- Infinite-horizon discounted MDP: $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, p, r, \rho, \gamma \rangle$ with finite states and actions ($S = |\mathcal{S}|$ and $A = |\mathcal{A}|$).
- Distributions induced by policy π: For each state $s \in \mathcal{S}$, $\pi(\cdot|s)$ over actions. State occupancy measure: $d^\pi(s) = (1-\gamma)\sum_{\tau=0}^{\infty} \gamma^\tau \mathbb{P}(s_\tau = s \mid s_0 \sim \rho, a_\tau \sim \pi(\cdot|s_\tau))$.
- Expected discounted return for π: $J(\pi) = \mathbb{E}_{s_0, a_0, \ldots}[\sum_{\tau=0}^{\infty} \gamma^\tau r(s_\tau, a_\tau)]$, where $s_0 \sim \rho, a_\tau \sim \pi(\cdot|s_\tau)$, and $s_{\tau+1} \sim p(\cdot|s_\tau, a_\tau)$.
- **Objective**: Given a set of feasible policies Π, $\max_{\pi \in \Pi} J(\pi)$. $\pi^* := \arg\max_{\pi \in \Pi} J(\pi)$.

Problem Formulation

- Infinite-horizon discounted MDP: $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, p, r, \rho, \gamma \rangle$ with finite states and actions ($S = |\mathcal{S}|$ and $A = |\mathcal{A}|$).
- Distributions induced by policy π: For each state $s \in \mathcal{S}$, $\pi(\cdot|s)$ over actions. State occupancy measure: $d^\pi(s) = (1-\gamma) \sum_{\tau=0}^{\infty} \gamma^\tau \mathbb{P}(s_\tau = s \mid s_0 \sim \rho, a_\tau \sim \pi(\cdot|s_\tau))$.
- Expected discounted return for π: $J(\pi) = \mathbb{E}_{s_0, a_0, \ldots}[\sum_{\tau=0}^{\infty} \gamma^\tau r(s_\tau, a_\tau)]$, where $s_0 \sim \rho, a_\tau \sim \pi(\cdot|s_\tau)$, and $s_{\tau+1} \sim p(\cdot|s_\tau, a_\tau)$.
- **Objective**: Given a set of feasible policies Π, $\max_{\pi \in \Pi} J(\pi)$. $\pi^* := \arg\max_{\pi \in \Pi} J(\pi)$.
- Softmax tabular parameterization: For parameters $\theta \in \mathbb{R}^{S \times A}$, the set Π consists of policies $\pi_\theta : \mathcal{S} \rightarrow \Delta_{\mathcal{A}}$ s.t. $\pi_\theta(a|s) = {}^{\exp(\theta(s,a))}/_{\sum_{a' \in \mathcal{A}} \exp(\theta(s,a'))}$.

Problem Formulation

- Infinite-horizon discounted MDP: $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, p, r, \rho, \gamma \rangle$ with finite states and actions ($S = |\mathcal{S}|$ and $A = |\mathcal{A}|$).
- Distributions induced by policy π: For each state $s \in \mathcal{S}$, $\pi(\cdot|s)$ over actions. State occupancy measure: $d^\pi(s) = (1-\gamma)\sum_{\tau=0}^{\infty} \gamma^\tau \mathbb{P}(s_\tau = s \mid s_0 \sim \rho, a_\tau \sim \pi(\cdot|s_\tau))$.
- Expected discounted return for π: $J(\pi) = \mathbb{E}_{s_0, a_0, \ldots}[\sum_{\tau=0}^{\infty} \gamma^\tau r(s_\tau, a_\tau)]$, where $s_0 \sim \rho, a_\tau \sim \pi(\cdot|s_\tau)$, and $s_{\tau+1} \sim p(\cdot|s_\tau, a_\tau)$.
- **Objective**: Given a set of feasible policies Π, $\max_{\pi \in \Pi} J(\pi)$. $\pi^* := \arg\max_{\pi \in \Pi} J(\pi)$.
- Softmax tabular parameterization: For parameters $\theta \in \mathbb{R}^{S \times A}$, the set Π consists of policies $\pi_\theta : \mathcal{S} \to \Delta_{\mathcal{A}}$ s.t. $\pi_\theta(a|s) = {}^{\exp(\theta(s,a))}/_{\sum_{a' \in \mathcal{A}} \exp(\theta(s,a'))}$.
- Abstract out the objective as $f(\theta) := J(\pi_\theta)$ with $f^* := \max_\theta f(\theta)$ to potentially extend the results to convex/constrained MDPs.

Problem Formulation

Properties of f:

- f is twice-differentiable but non-concave in θ.

Problem Formulation

Properties of f:

- f is twice-differentiable but non-concave in θ.
- f is *uniform* smooth i.e. there exists a constant $L \in (0, \infty)$ s.t. $\forall \theta$, $\nabla^2 f(\theta) \preceq L\, I_{SA}$. E.g. $L = \frac{5}{2}$ for bandit problems.

Problem Formulation

Properties of f:

- f is twice-differentiable but non-concave in θ.
- f is *uniform* smooth i.e. there exists a constant $L \in (0, \infty)$ s.t. $\forall \theta$, $\nabla^2 f(\theta) \preceq L\, I_{SA}$.
  E.g. $L = \frac{5}{2}$ for bandit problems.
- f is *non-uniform* smooth i.e. there exists a constant $L_1 \in (0, \infty)$ s.t. $\forall \theta$,
  $\nabla^2 f(\theta) \preceq L_1 \|\nabla f(\theta)\|\, I_{SA}$, i.e. optimization landscape is flatter closer to a stationary point.
  E.g. $L_1 = 3$ for bandit problems.

Problem Formulation

Properties of f:

- f is twice-differentiable but non-concave in θ.
- f is *uniform* smooth i.e. there exists a constant $L \in (0, \infty)$ s.t. $\forall \theta$, $\nabla^2 f(\theta) \preceq L\, I_{SA}$.
  E.g. $L = \frac{5}{2}$ for bandit problems.
- f is *non-uniform* smooth i.e. there exists a constant $L_1 \in (0, \infty)$ s.t. $\forall \theta$,
  $\nabla^2 f(\theta) \preceq L_1 \|\nabla f(\theta)\|\, I_{SA}$, i.e. optimization landscape is flatter closer to a stationary point.
  E.g. $L_1 = 3$ for bandit problems.
- f satisfies a *non-uniform Łojasiewciz condition*, i.e. for all θ, there exists a $C(\theta) \in (0, \infty)$ s.t. $\|\nabla f(\theta)\|_2 \geq C(\theta)\, [f^* - f(\theta)]$. E.g. $C(\theta) \propto \pi_\theta(a^*)$ for bandit problems.

Problem Formulation

Properties of f:

- f is twice-differentiable but non-concave in θ.
- f is *uniform* smooth i.e. there exists a constant $L \in (0, \infty)$ s.t. $\forall \theta$, $\nabla^2 f(\theta) \preceq L \, I_{SA}$.
  E.g. $L = \frac{5}{2}$ for bandit problems.
- f is *non-uniform* smooth i.e. there exists a constant $L_1 \in (0, \infty)$ s.t. $\forall \theta$, $\nabla^2 f(\theta) \preceq L_1 \, \|\nabla f(\theta)\| \, I_{SA}$, i.e. optimization landscape is flatter closer to a stationary point.
  E.g. $L_1 = 3$ for bandit problems.
- f satisfies a *non-uniform Łojasiewciz condition*, i.e. for all θ, there exists a $C(\theta) \in (0, \infty)$ s.t. $\|\nabla f(\theta)\|_2 \geq C(\theta) \, [f^* - f(\theta)]$. E.g. $C(\theta) \propto \pi_\theta(a^*)$ for bandit problems.

Sufficient exploration assumption for MDPs: Similar to Mei et al. [2020], we assume that the starting state distribution satisfies $\min_s \rho(s) > 0$ and hence $C_\infty := \max_\pi \left\| \frac{d^\pi_\rho}{\rho} \right\|_\infty < \infty$.
Allows us to exclusively focus on the optimization aspects of the problem.

Outline

- Problem Formulation
- **Softmax Policy Gradient**
- Stochastic Softmax Policy Gradient
- Conclusion

Softmax policy gradient

- Softmax policy gradient: At iteration $t \in [T]$, the SPG update is:

$$\theta_{t+1} = \theta_t + \eta_t \nabla f(\theta_t)\,,$$

where η_t is the step-size. For finite MDPs, $[\nabla f(\theta)]_{s,a} = \frac{d^{\pi_\theta}(s)\,\pi_\theta(a|s)\,A^{\pi_\theta}(s,a)}{1-\gamma}$.

Softmax policy gradient

- Softmax policy gradient: At iteration $t \in [T]$, the SPG update is:

$$\theta_{t+1} = \theta_t + \eta_t \nabla f(\theta_t)\,,$$

  where η_t is the step-size. For finite MDPs, $[\nabla f(\theta)]_{s,a} = \frac{d^{\pi_\theta}(s)\,\pi_\theta(a|s)\,A^{\pi_\theta}(s,a)}{1-\gamma}$.
- Assume $\nabla f(\theta)$ can be computed exactly. It is possible to account for the estimation error in the policy gradients [Agarwal et al., 2021].

Softmax policy gradient

- Softmax policy gradient: At iteration $t \in [T]$, the SPG update is:

$$\theta_{t+1} = \theta_t + \eta_t \nabla f(\theta_t)\,,$$

where η_t is the step-size. For finite MDPs, $[\nabla f(\theta)]_{s,a} = \frac{d^{\pi_\theta}(s)\, \pi_\theta(a|s)\, A^{\pi_\theta}(s,a)}{1-\gamma}$.
- Assume $\nabla f(\theta)$ can be computed exactly. It is possible to account for the estimation error in the policy gradients [Agarwal et al., 2021].

What is known for softmax PG*: For a target $\epsilon > 0$,

✓ SPG with $\eta_t = \frac{1}{L}$ and $T = O(1/\epsilon)$ ensures that $f^* - f(\theta_T) \leq \epsilon$ [Mei et al., 2020].

× In practice, using a step-size that depends on global smoothness constants is often too conservative and results in poor empirical performance.

*Natural policy gradient with an exact line-search/adaptive step-sizes can obtain a linear convergence rate [Bhandari and Russo, 2021, Khodadadian et al., 2021].

Softmax policy gradient

- Softmax policy gradient: At iteration $t \in [T]$, the SPG update is:

$$\theta_{t+1} = \theta_t + \eta_t \nabla f(\theta_t)\,,$$

where η_t is the step-size. For finite MDPs, $[\nabla f(\theta)]_{s,a} = \frac{d^{\pi_\theta}(s)\,\pi_\theta(a|s)\,A^{\pi_\theta}(s,a)}{1-\gamma}$.

- Assume $\nabla f(\theta)$ can be computed exactly. It is possible to account for the estimation error in the policy gradients [Agarwal et al., 2021].

What is known for softmax PG*: For a target $\epsilon > 0$,

✓ SPG with $\eta_t = \frac{1}{L}$ and $T = O(1/\epsilon)$ ensures that $f^* - f(\theta_T) \leq \epsilon$ [Mei et al., 2020].

× In practice, using a step-size that depends on global smoothness constants is often too conservative and results in poor empirical performance.

✓ Normalized SPG with an update: $\theta_{t+1} = \theta_t + \eta \frac{\nabla f(\theta)}{\|\nabla f(\theta)\|}$, $\eta = \frac{1}{2L_1}$ and $T = O(\log(1/\epsilon))$ ensures that $f^* - f(\theta_T) \leq \epsilon$ [Mei et al., 2021b].

× For finite MDPs, L_1 depends on C_∞ for which we can only obtain loose upper-bounds.

*Natural policy gradient with an exact line-search/adaptive step-sizes can obtain a linear convergence rate [Bhandari and Russo, 2021, Khodadadian et al., 2021].

Softmax policy gradient

Idea: Use a line-search to exploit the uniform smoothness and automatically set the step-size.

Softmax policy gradient

Idea: Use a line-search to exploit the uniform smoothness and automatically set the step-size.

Backtracking Armijo line-search: At every iteration t, start from an initial guess for the step-size ($\eta_{\max}$) and backtrack until the *Armijo condition* is satisfied.

$$f(\theta_t + \eta_t \nabla f(\theta_t)) \geq f(\theta_t) + h\, \eta_t \|\nabla f(\theta_t)\|_2^2 \,, \quad \text{(Armijo condition)}$$

where $h \in (0, 1)$ is a hyper-parameter.

- Above procedure guarantees that $\eta_t \geq \min\{{}^{2(1-h)}/_{L}, \eta_{\max}\}$.

Softmax policy gradient

Idea: Use a line-search to exploit the uniform smoothness and automatically set the step-size.

Backtracking Armijo line-search: At every iteration t, start from an initial guess for the step-size ($\eta_{\max}$) and backtrack until the *Armijo condition* is satisfied.

$$f(\theta_t + \eta_t \nabla f(\theta_t)) \geq f(\theta_t) + h\,\eta_t \|\nabla f(\theta_t)\|_2^2\,, \quad \text{(Armijo condition)}$$

where $h \in (0, 1)$ is a hyper-parameter.

- Above procedure guarantees that $\eta_t \geq \min\{^{2(1-h)}/_L, \eta_{\max}\}$.
- Theorem [LARV'24]: SPG with the backtracking Armijo line-search (with $h = \frac{1}{2}$) and $T = O(1/\epsilon)$ iterations ensures that $f^* - f(\theta_T) \leq \epsilon$
- *Proof*: Exploit the Łojasiewciz property with the standard proof for Armijo line-search on smooth functions. Guarantee that the non-uniform Łojasiewciz constant $C(\theta_t) > 0$ for all t.

Softmax policy gradient

Idea: Use a line-search to exploit the uniform smoothness and automatically set the step-size.

Backtracking Armijo line-search: At every iteration t, start from an initial guess for the step-size ($\eta_{\max}$) and backtrack until the *Armijo condition* is satisfied.

$$f(\theta_t + \eta_t \nabla f(\theta_t)) \geq f(\theta_t) + h\, \eta_t \|\nabla f(\theta_t)\|_2^2 \,, \quad \text{(Armijo condition)}$$

where $h \in (0, 1)$ is a hyper-parameter.

- Above procedure guarantees that $\eta_t \geq \min\{2(1-h)/L, \eta_{\max}\}$.
- Theorem [LARV'24]: SPG with the backtracking Armijo line-search (with $h = \frac{1}{2}$) and $T = O(1/\epsilon)$ iterations ensures that $f^* - f(\theta_T) \leq \epsilon$
- *Proof*: Exploit the Łojasiewcziz property with the standard proof for Armijo line-search on smooth functions. Guarantee that the non-uniform Łojasiewcziz constant $C(\theta_t) > 0$ for all t.

Q: Can we design a line-search to exploit the non-uniform smoothness and attain linear convergence for SPG?

Softmax policy gradient

Idea: If f is L_1 non-uniform smooth, then, $g(\theta) = \ln(f^* - f(\theta))$ is $O(L_1)$-uniform smooth (similar property holds for the logistic loss [Ji and Telgarsky, 2018]). Use backtracking Armijo line-search on $g(\theta)$.

Softmax policy gradient

Idea: If f is L_1 non-uniform smooth, then, $g(\theta) = \ln(f^* - f(\theta))$ is $O(L_1)$-uniform smooth (similar property holds for the logistic loss [Ji and Telgarsky, 2018]). Use backtracking Armijo line-search on $g(\theta)$.

Backtracking Armijo line-search: At every iteration t, start from an initial guess for the step-size ($\eta_{\max}$) and backtrack until the following condition is satisfied.

$$\ln(f^* - f(\theta_t + \eta_t \nabla f(\theta_t))) \leq \ln(f^* - f(\theta_t)) - h\, \eta_t \, \frac{\|\nabla f(\theta_t)\|_2^2}{f^* - f(\theta_t)} \quad \text{(Armijo condition for log-loss).}$$

- Above procedure guarantees that $\eta_t \geq \min\left\{\eta_{\max}, \frac{2(1-h)}{O(L_1)\,[f^* - f(\theta_t)]}\right\}$.

Softmax policy gradient

Idea: If f is L_1 non-uniform smooth, then, $g(\theta) = \ln(f^* - f(\theta))$ is $O(L_1)$-uniform smooth (similar property holds for the logistic loss [Ji and Telgarsky, 2018]). Use backtracking Armijo line-search on $g(\theta)$.

Backtracking Armijo line-search: At every iteration t, start from an initial guess for the step-size ($\eta_{\max}$) and backtrack until the following condition is satisfied.

$$\ln(f^* - f(\theta_t + \eta_t \nabla f(\theta_t))) \leq \ln(f^* - f(\theta_t)) - h\,\eta_t \frac{\|\nabla f(\theta_t)\|_2^2}{f^* - f(\theta_t)} \quad \text{(Armijo condition for log-loss).}$$

- Above procedure guarantees that $\eta_t \geq \min\left\{\eta_{\max}, \frac{2(1-h)}{O(L_1)\,[f^* - f(\theta_t)]}\right\}$.
- Theorem [LARV'24]: SPG with the backtracking line-search using the Armijo condition for the log-loss (with $h = \frac{1}{2}$) and and $T = O(\log(1/\epsilon))$ ensures that $f^* - f(\theta_T) \leq \epsilon$.

Softmax policy gradient

Idea: If f is L_1 non-uniform smooth, then, $g(\theta) = \ln(f^* - f(\theta))$ is $O(L_1)$-uniform smooth (similar property holds for the logistic loss [Ji and Telgarsky, 2018]). Use backtracking Armijo line-search on $g(\theta)$.

Backtracking Armijo line-search: At every iteration t, start from an initial guess for the step-size ($\eta_{\max}$) and backtrack until the following condition is satisfied.

$$\ln(f^* - f(\theta_t + \eta_t \nabla f(\theta_t))) \leq \ln(f^* - f(\theta_t)) - h\, \eta_t \, \frac{\|\nabla f(\theta_t)\|_2^2}{f^* - f(\theta_t)} \quad \text{(Armijo condition for log-loss).}$$

- Above procedure guarantees that $\eta_t \geq \min\left\{\eta_{\max}, \frac{2(1-h)}{O(L_1)\,[f^* - f(\theta_t)]}\right\}$.
- Theorem [LARV'24]: SPG with the backtracking line-search using the Armijo condition for the log-loss (with $h = \frac{1}{2}$) and and $T = O(\log(1/\epsilon))$ ensures that $f^* - f(\theta_T) \leq \epsilon$.
- × Similar to the Polyak step-size [Polyak, 1987], the above condition requires knowledge of f^*. In practice, if the rewards are in $[0, 1]$, estimate f^* by $\frac{1}{1-\gamma}$.

Softmax policy gradient

Idea: If f is L_1 non-uniform smooth, then, $g(\theta) = \ln(f^* - f(\theta))$ is $O(L_1)$-uniform smooth (similar property holds for the logistic loss [Ji and Telgarsky, 2018]). Use backtracking Armijo line-search on $g(\theta)$.

Backtracking Armijo line-search: At every iteration t, start from an initial guess for the step-size ($\eta_{\max}$) and backtrack until the following condition is satisfied.

$$\ln(f^* - f(\theta_t + \eta_t \nabla f(\theta_t))) \leq \ln(f^* - f(\theta_t)) - h\,\eta_t \frac{\|\nabla f(\theta_t)\|_2^2}{f^* - f(\theta_t)} \quad \text{(Armijo condition for log-loss).}$$

- Above procedure guarantees that $\eta_t \geq \min\left\{\eta_{\max}, \frac{2(1-h)}{O(L_1)\,[f^* - f(\theta_t)]}\right\}$.
- Theorem [LARV'24]: SPG with the backtracking line-search using the Armijo condition for the log-loss (with $h = \frac{1}{2}$) and and $T = O(\log(1/\epsilon))$ ensures that $f^* - f(\theta_T) \leq \epsilon$.

× Similar to the Polyak step-size [Polyak, 1987], the above condition requires knowledge of f^*. In practice, if the rewards are in $[0, 1]$, estimate f^* by $\frac{1}{1-\gamma}$.

✓ Experimentally, on tabular MDPs, given a starting state distribution with fulll support, SPG + line-search can attain linear convergence and match the performance of policy iteration.

Outline

- Problem Formulation
- Softmax Policy Gradient
- **Stochastic Softmax Policy Gradient**
- Conclusion

Stochastic Softmax Policy Gradient

- Cannot compute the policy gradient exactly, and need to estimate it via interactions with the environment.

Stochastic Softmax Policy Gradient

- Cannot compute the policy gradient exactly, and need to estimate it via interactions with the environment.
- Require stochastic policy gradients $\nabla \widetilde{f}(\theta)$ that are unbiased and have bounded variance: $\forall \theta$,

$$\mathbb{E}[\nabla \widetilde{f}(\theta)] = \nabla f(\theta) \quad ; \quad \mathbb{E}\left\| \nabla \widetilde{f}(\theta) - \nabla f(\theta) \right\|_2^2 \leq \sigma^2 < \infty$$

Stochastic Softmax Policy Gradient

- Cannot compute the policy gradient exactly, and need to estimate it via interactions with the environment.
- Require stochastic policy gradients $\nabla \widetilde{f}(\theta)$ that are unbiased and have bounded variance: $\forall \theta$,

$$\mathbb{E}[\nabla \widetilde{f}(\theta)] = \nabla f(\theta) \quad ; \quad \mathbb{E}\left\| \nabla \widetilde{f}(\theta) - \nabla f(\theta) \right\|_2^2 \leq \sigma^2 < \infty$$

- Running example: Stochastic multi-armed bandits for which $f(\theta) = \langle \pi_\theta, r \rangle$.
  - At iteration t, sample action $a_t \sim \pi_{\theta_t}$ and construct the importance sampling (IS) reward estimate $\hat{r}_t(a) = \frac{\mathbb{1}\{a_t = a\}}{\pi_{\theta_t}(a)} R_t$ for each $a \in \mathcal{A}$, and calculate $\nabla \widetilde{f}(\theta) = \nabla_\theta \langle \pi_\theta, \hat{r}_t \rangle$.
  - $\nabla \widetilde{f}(\theta)$ is unbiased and has bounded variance.

Stochastic Softmax Policy Gradient

- Cannot compute the policy gradient exactly, and need to estimate it via interactions with the environment.
- Require stochastic policy gradients $\nabla \widetilde{f}(\theta)$ that are unbiased and have bounded variance: $\forall \theta$,

$$\mathbb{E}[\nabla \widetilde{f}(\theta)] = \nabla f(\theta) \quad ; \quad \mathbb{E}\left\|\nabla \widetilde{f}(\theta) - \nabla f(\theta)\right\|_2^2 \leq \sigma^2 < \infty$$

- Running example: Stochastic multi-armed bandits for which $f(\theta) = \langle \pi_\theta, r \rangle$.
    - At iteration t, sample action $a_t \sim \pi_{\theta_t}$ and construct the importance sampling (IS) reward estimate $\hat{r}_t(a) = \frac{\mathbb{1}\{a_t = a\}}{\pi_{\theta_t}(a)} R_t$ for each $a \in \mathcal{A}$, and calculate $\nabla \widetilde{f}(\theta) = \nabla_\theta \langle \pi_\theta, \hat{r}_t \rangle$.
    - $\nabla \widetilde{f}(\theta)$ is unbiased and has bounded variance.
- Can also construct such a gradient estimator for MDPs (rolling out trajectories and truncating them at a random stopping time (dependent on γ)).

Stochastic Softmax Policy Gradient

- Cannot compute the policy gradient exactly, and need to estimate it via interactions with the environment.
- Require stochastic policy gradients $\nabla \widetilde{f}(\theta)$ that are unbiased and have bounded variance: $\forall \theta$,

$$\mathbb{E}[\nabla \widetilde{f}(\theta)] = \nabla f(\theta) \quad ; \quad \mathbb{E}\left\|\nabla \widetilde{f}(\theta) - \nabla f(\theta)\right\|_2^2 \leq \sigma^2 < \infty$$

- Running example: Stochastic multi-armed bandits for which $f(\theta) = \langle \pi_\theta, r \rangle$.
  - At iteration t, sample action $a_t \sim \pi_{\theta_t}$ and construct the importance sampling (IS) reward estimate $\hat{r}_t(a) = \frac{\mathbb{1}\{a_t = a\}}{\pi_{\theta_t}(a)} R_t$ for each $a \in \mathcal{A}$, and calculate $\nabla \widetilde{f}(\theta) = \nabla_\theta \langle \pi_\theta, \hat{r}_t \rangle$.
  - $\nabla \widetilde{f}(\theta)$ is unbiased and has bounded variance.
- Can also construct such a gradient estimator for MDPs (rolling out trajectories and truncating them at a random stopping time (dependent on γ)).
- Stochastic softmax PG: At iteration t, construct $\nabla \widetilde{f}(\theta_t)$, and update the parameters as:

$$\theta_{t+1} = \theta_t + \eta_t \nabla \widetilde{f}(\theta_t)$$

.

Stochastic Softmax Policy Gradient

What is known for stochastic SPG*: For a target $\epsilon > 0$, Stochastic SPG:

- with $\eta_t \propto \|\nabla f(\theta_t)\|$ and $T = O(1/\epsilon^2)$ ensures that $\mathbb{E}[f^* - f(\theta_T)] \leq \epsilon$ [Mei et al., 2021a].
  - × The full gradient cannot be computed in the stochastic setting.
- with η_t that depends on $\mu \propto \mathbb{E}[\inf_{t \geq 1}[C(\theta_t)]^2]$ and $T = O(1/\epsilon^3)$ ensures that $\min_{t \in [T]} \mathbb{E}[f^* - f(\theta_t)] \leq \epsilon$ [Yuan et al., 2022].
  - × For bandit problems, $C(\theta) \propto \pi_\theta(a^*)$ and hence μ is unknown.

* Both natural policy gradient (NPG) and normalized SPG are too aggressive, do not explore enough and can commit to the sub-optimal action in the stochastic on-policy setting [Mei et al., 2021a, Chung et al., 2021].

Stochastic Softmax Policy Gradient

What is known for stochastic SPG*: For a target $\epsilon > 0$, Stochastic SPG:

- with $\eta_t \propto \|\nabla f(\theta_t)\|$ and $T = O(1/\epsilon^2)$ ensures that $\mathbb{E}[f^* - f(\theta_T)] \leq \epsilon$ [Mei et al., 2021a].
  - × The full gradient cannot be computed in the stochastic setting.
- with η_t that depends on $\mu \propto \mathbb{E}[\inf_{t \geq 1}[C(\theta_t)]^2]$ and $T = O(1/\epsilon^3)$ ensures that $\min_{t \in [T]} \mathbb{E}[f^* - f(\theta_t)] \leq \epsilon$ [Yuan et al., 2022].
  - × For bandit problems, $C(\theta) \propto \pi_\theta(a^*)$ and hence μ is unknown.

Q: Can we design a practical stochastic SPG method that ensures global convergence and does not require unknown problem-dependent constants?

* Both natural policy gradient (NPG) and normalized SPG are too aggressive, do not explore enough and can commit to the sub-optimal action in the stochastic on-policy setting [Mei et al., 2021a, Chung et al., 2021].

Stochastic Softmax Policy Gradient

What is known for stochastic SPG*: For a target $\epsilon > 0$, Stochastic SPG:

- with $\eta_t \propto \|\nabla f(\theta_t)\|$ and $T = O(1/\epsilon^2)$ ensures that $\mathbb{E}[f^* - f(\theta_T)] \leq \epsilon$ [Mei et al., 2021a].
  - × The full gradient cannot be computed in the stochastic setting.
- with η_t that depends on $\mu \propto \mathbb{E}[\inf_{t \geq 1}[C(\theta_t)]^2]$ and $T = O(1/\epsilon^3)$ ensures that $\min_{t \in [T]} \mathbb{E}[f^* - f(\theta_t)] \leq \epsilon$ [Yuan et al., 2022].
  - × For bandit problems, $C(\theta) \propto \pi_\theta(a^*)$ and hence μ is unknown.

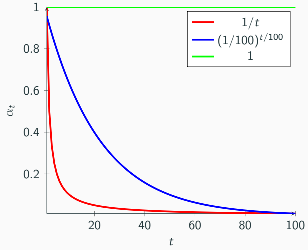
Q: Can we design a practical stochastic SPG method that ensures global convergence and does not require unknown problem-dependent constants?

Observation: Problem is equivalent to constructing a step-size schedule for SGD when minimizing a smooth, non-convex function satisfying a gradient domination condition (with parameter μ) without the knowledge of μ.

* Both natural policy gradient (NPG) and normalized SPG are too aggressive, do not explore enough and can commit to the sub-optimal action in the stochastic on-policy setting [Mei et al., 2021a, Chung et al., 2021].

Digression – SGD with exponentially decreasing step-sizes

- Idea: Use exponentially decreasing step-sizes [Li et al., 2021, Vaswani et al., 2022]. Specifically, for a fixed T, $\eta_t := \eta_0 \, \alpha_t$ where $\eta_0 = \frac{1}{L}$ and $\alpha_t = \alpha^t$ where $\alpha := \left(\frac{1}{T}\right)^{1/T}$.
- Exponential step-sizes lie between the constant and $1/t$ decreasing step-sizes, implying that for $t \in [T]$, $\alpha_t \in \left[\frac{1}{t}, 1\right]$.

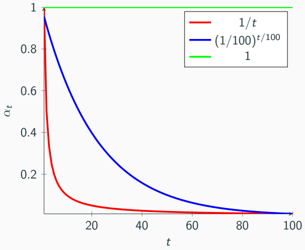


Digression – SGD with exponentially decreasing step-sizes

- Idea: Use exponentially decreasing step-sizes [Li et al., 2021, Vaswani et al., 2022]. Specifically, for a fixed T, $\eta_t := \eta_0 \, \alpha_t$ where $\eta_0 = \frac{1}{L}$ and $\alpha_t = \alpha^t$ where $\alpha := \left(\frac{1}{T}\right)^{1/T}$.
- Exponential step-sizes lie between the constant and $1/t$ decreasing step-sizes, implying that for $t \in [T]$, $\alpha_t \in \left[\frac{1}{t}, 1\right]$.

✓ When minimizing smooth, non-convex functions satisfying the Polyak Łojasiewcziz (PL) condition (with constant μ), SGD with exponentially decreasing step-sizes requires $O(\log(1/\epsilon) + \sigma^2/\epsilon^2)$ iterations to ensure an ϵ sub-optimality [Li et al., 2021].

✓ The step-sizes do not require knowledge of μ.



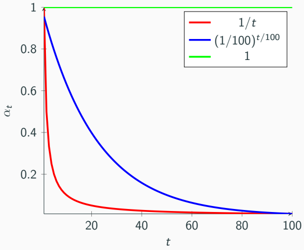
Digression – SGD with exponentially decreasing step-sizes

- Idea: Use exponentially decreasing step-sizes [Li et al., 2021, Vaswani et al., 2022]. Specifically, for a fixed T, $\eta_t := \eta_0 \, \alpha_t$ where $\eta_0 = \frac{1}{L}$ and $\alpha_t = \alpha^t$ where $\alpha := \left(\frac{1}{T}\right)^{1/T}$.
- Exponential step-sizes lie between the constant and $1/t$ decreasing step-sizes, implying that for $t \in [T]$, $\alpha_t \in \left[\frac{1}{t}, 1\right]$.

✓ When minimizing smooth, non-convex functions satisfying the Polyak Łojasiewciz (PL) condition (with constant μ), SGD with exponentially decreasing step-sizes requires $O(\log(1/\epsilon) + \sigma^2/\epsilon^2)$ iterations to ensure an ϵ sub-optimality [Li et al., 2021].

✓ The step-sizes do not require knowledge of μ.

× Compared to the PL condition, the softmax policy optimization objective only satisfies a weaker (non-uniform) gradient domination condition.



Stochastic Softmax Policy Gradient

Theorem [LARV'24]: For a given $\epsilon \in (0,1)$, running stochastic SPG with exponentially decreasing step-sizes $\eta_t = \eta_0\, \alpha^t$ where $\eta_0 = \frac{1}{L}$ and $\alpha = \left(\frac{1}{T}\right)^{\frac{1}{T}}$, results in the following convergence:
If $\mathbb{E}[f^* - f(\theta_t)] > \epsilon$ for all $t \in [1, T]$, $\mu \propto \mathbb{E}[\inf_{t \geq 1}[C(\theta_t)]^2] > 0$ and $\kappa := \frac{L}{\mu}$, then,

$$\mathbb{E}[f^* - f(\theta_{T+1})] \leq [f^* - f(\theta_1)]\, C_1 \exp\left(-\frac{\alpha\, \epsilon\, T}{\kappa \ln(T)}\right) + \frac{C_1\, C_2}{2\, L} \frac{\ln^2(T)\, \sigma^2}{\epsilon^2\, T}$$

Setting $T = \tilde{O}\left(1/\epsilon + \sigma^2/\epsilon^3\right)$ iterations ensures $\min_{t \in [1, T+1]} \mathbb{E}[f^* - f(\theta_t)] \leq \epsilon$.

Stochastic Softmax Policy Gradient

Theorem [LARV'24]: For a given $\epsilon \in (0,1)$, running stochastic SPG with exponentially decreasing step-sizes $\eta_t = \eta_0 \, \alpha^t$ where $\eta_0 = \frac{1}{L}$ and $\alpha = \left(\frac{1}{T}\right)^{\frac{1}{T}}$, results in the following convergence:
If $\mathbb{E}[f^* - f(\theta_t)] > \epsilon$ for all $t \in [1, T]$, $\mu \propto \mathbb{E}[\inf_{t \geq 1}[C(\theta_t)]^2] > 0$ and $\kappa := \frac{L}{\mu}$, then,

$$\mathbb{E}[f^* - f(\theta_{T+1})] \leq [f^* - f(\theta_1)] \, C_1 \exp\left(-\frac{\alpha \, \epsilon \, T}{\kappa \, \ln(T)}\right) + \frac{C_1 \, C_2}{2\,L} \frac{\ln^2(T) \, \sigma^2}{\epsilon^2 \, T}$$

Setting $T = \tilde{O}\left({}^{1}/_{\epsilon} + {}^{\sigma^2}/_{\epsilon^3}\right)$ iterations ensures $\min_{t \in [1, T+1]} \mathbb{E}[f^* - f(\theta_t)] \leq \epsilon$.

- ✓ The rate is *noise-adaptive* and depends on σ. Recovers $O(1/\epsilon)$ convergence in the exact setting (when $\sigma = 0$). The $O(1/\epsilon^3)$ rate matches that of SGD when minimizing smooth non-convex functions satisfying the Łojasiewcziz condition [Fontaine et al., 2021].
- ✓ The algorithm does not require unknown problem-dependent constants.

Stochastic Softmax Policy Gradient

Theorem [LARV'24]: For a given $\epsilon \in (0,1)$, running stochastic SPG with exponentially decreasing step-sizes $\eta_t = \eta_0 \, \alpha^t$ where $\eta_0 = \frac{1}{L}$ and $\alpha = \left(\frac{1}{T}\right)^{\frac{1}{T}}$, results in the following convergence:
If $\mathbb{E}[f^* - f(\theta_t)] > \epsilon$ for all $t \in [1, T]$, $\mu \propto \mathbb{E}[\inf_{t \geq 1} [C(\theta_t)]^2] > 0$ and $\kappa := \frac{L}{\mu}$, then,

$$\mathbb{E}[f^* - f(\theta_{T+1})] \leq [f^* - f(\theta_1)] \, C_1 \, \exp\left(-\frac{\alpha \, \epsilon \, T}{\kappa \, \ln(T)}\right) + \frac{C_1 \, C_2}{2 \, L} \frac{\ln^2(T) \, \sigma^2}{\epsilon^2 \, T}$$

Setting $T = \tilde{O}\left({}^{1}/_{\epsilon} + {}^{\sigma^2}/_{\epsilon^3}\right)$ iterations ensures $\min_{t \in [1, T+1]} \mathbb{E}[f^* - f(\theta_t)] \leq \epsilon$.

- ✓ The rate is *noise-adaptive* and depends on σ. Recovers $O(1/\epsilon)$ convergence in the exact setting (when $\sigma = 0$). The $O(1/\epsilon^3)$ rate matches that of SGD when minimizing smooth non-convex functions satisfying the Łojasiewcziz condition [Fontaine et al., 2021].
- ✓ The algorithm does not require unknown problem-dependent constants.
- Ensuring $\mu > 0$ requires that $\pi_{\theta_t}(a^*) > 0$. This is true for any finite T.
- × The rate depends on μ which depends on the initialization/trajectory and can be small.

Stochastic Softmax Policy Gradient

Theorem [LARV'24]: For a given $\epsilon \in (0,1)$, running stochastic SPG with exponentially decreasing step-sizes $\eta_t = \eta_0 \, \alpha^t$ where $\eta_0 = \frac{1}{L}$ and $\alpha = \left(\frac{1}{T}\right)^{\frac{1}{T}}$, results in the following convergence:
If $\mathbb{E}[f^* - f(\theta_t)] > \epsilon$ for all $t \in [1, T]$, $\mu \propto \mathbb{E}[\inf_{t \geq 1}[C(\theta_t)]^2] > 0$ and $\kappa := \frac{L}{\mu}$, then,

$$\mathbb{E}[f^* - f(\theta_{T+1})] \leq [f^* - f(\theta_1)] \, C_1 \, \exp\left(-\frac{\alpha \, \epsilon \, T}{\kappa \, \ln(T)}\right) + \frac{C_1 \, C_2}{2\,L} \frac{\ln^2(T) \, \sigma^2}{\epsilon^2 \, T}$$

Setting $T = \tilde{O}\left(1/\epsilon + \sigma^2/\epsilon^3\right)$ iterations ensures $\min_{t \in [1, T+1]} \mathbb{E}[f^* - f(\theta_t)] \leq \epsilon$.

- ✓ The rate is *noise-adaptive* and depends on σ. Recovers $O(1/\epsilon)$ convergence in the exact setting (when $\sigma = 0$). The $O(1/\epsilon^3)$ rate matches that of SGD when minimizing smooth non-convex functions satisfying the Łojasiewcziz condition [Fontaine et al., 2021].
- ✓ The algorithm does not require unknown problem-dependent constants.
- Ensuring $\mu > 0$ requires that $\pi_{\theta_t}(a^*) > 0$. This is true for any finite T.
- × The rate depends on μ which depends on the initialization/trajectory and can be small.
- × Slower rate (in terms of T) compared to [Mei et al., 2021a, 2023].

Stochastic Softmax Policy Gradient

Observation [Mei et al., 2023]: In the bandit setting, stochastic gradients satisfy the strong growth condition (SGC) [Schmidt and Roux, 2013, Vaswani et al., 2019] meaning that there exists a problem-dependent constant $\varrho \geq 1$ s.t $\forall \theta$,

$$\mathbb{E}\left\|\nabla \widetilde{f}(\theta)\right\|_2^2 \leq \varrho \left\|\nabla f(\theta)\right\|$$

- As $\|\nabla f(\theta)\| \to 0$, $\|\nabla \widetilde{f}(\theta)\| \to 0 \implies$ the variance decreases closer to a stationary point.

Stochastic Softmax Policy Gradient

Observation [Mei et al., 2023]: In the bandit setting, stochastic gradients satisfy the strong growth condition (SGC) [Schmidt and Roux, 2013, Vaswani et al., 2019] meaning that there exists a problem-dependent constant $\varrho \geq 1$ s.t $\forall \theta$,

$$\mathbb{E}\left\|\nabla \widetilde{f}(\theta)\right\|_2^2 \leq \varrho \, \|\nabla f(\theta)\|$$

- As $\|\nabla f(\theta)\| \to 0$, $\|\nabla \widetilde{f}(\theta)\| \to 0 \implies$ the variance decreases closer to a stationary point.
- ✓ Do not need to decrease the step-size. Running stochastic SPG with a constant step-size $\eta \propto 1/\varrho$ and $T = O(1/\epsilon)$ ensures that $\mathbb{E}[f^* - f(\theta_T)] \leq \epsilon$ [Mei et al., 2023]. Moreover, the algorithm ensures that $\pi_{\theta_t}(a^*) > 0$ for all t and $\lim_{t \to \infty} \pi_{\theta_t}(a^*) \to 1$.

Stochastic Softmax Policy Gradient

Observation [Mei et al., 2023]: In the bandit setting, stochastic gradients satisfy the strong growth condition (SGC) [Schmidt and Roux, 2013, Vaswani et al., 2019] meaning that there exists a problem-dependent constant $\varrho \geq 1$ s.t $\forall \theta$,

$$\mathbb{E}\left\|\nabla \tilde{f}(\theta)\right\|_2^2 \leq \varrho \left\|\nabla f(\theta)\right\|$$

- As $\|\nabla f(\theta)\| \to 0$, $\|\nabla \tilde{f}(\theta)\| \to 0 \implies$ the variance decreases closer to a stationary point.
- ✓ Do not need to decrease the step-size. Running stochastic SPG with a constant step-size $\eta \propto 1/\varrho$ and $T = O(1/\epsilon)$ ensures that $\mathbb{E}[f^* - f(\theta_T)] \leq \epsilon$ [Mei et al., 2023]. Moreover, the algorithm ensures that $\pi_{\theta_t}(a^*) > 0$ for all t and $\lim_{t \to \infty} \pi_{\theta_t}(a^*) \to 1$.
- × For bandit problems, $\varrho \propto \Delta := \min_{i \neq a^*} |r(a^*) - r(i)|$. The mean reward vector r is unknown in the stochastic setting, and the resulting algorithm cannot be implemented.

Stochastic Softmax Policy Gradient

Observation [Mei et al., 2023]: In the bandit setting, stochastic gradients satisfy the strong growth condition (SGC) [Schmidt and Roux, 2013, Vaswani et al., 2019] meaning that there exists a problem-dependent constant $\varrho \geq 1$ s.t $\forall \theta$,

$$\mathbb{E}\left\|\nabla \widetilde{f}(\theta)\right\|_2^2 \leq \varrho \, \|\nabla f(\theta)\|$$

- As $\|\nabla f(\theta)\| \to 0$, $\|\nabla \widetilde{f}(\theta)\| \to 0 \implies$ the variance decreases closer to a stationary point.
- ✓ Do not need to decrease the step-size. Running stochastic SPG with a constant step-size $\eta \propto 1/\varrho$ and $T = O(1/\epsilon)$ ensures that $\mathbb{E}[f^* - f(\theta_T)] \leq \epsilon$ [Mei et al., 2023]. Moreover, the algorithm ensures that $\pi_{\theta_t}(a^*) > 0$ for all t and $\lim_{t \to \infty} \pi_{\theta_t}(a^*) \to 1$.
- × For bandit problems, $\varrho \propto \Delta := \min_{i \neq a^*} |r(a^*) - r(i)|$. The mean reward vector r is unknown in the stochastic setting, and the resulting algorithm cannot be implemented.

Q: Can we design a practical stochastic SPG method that achieves the faster $O(1/\epsilon)$ rate and does not require unknown problem-dependent constants?

Stochastic Softmax Policy Gradient

Observation: Stochastic SPG with exponential step-sizes can adapt to the decreasing σ_t.

Theorem [LARV'24]: For a given $\epsilon \in (0,1)$, running stochastic SPG with unbiased stochastic gradients that are bounded, i.e. $\|\nabla \widetilde{f}(\theta)\| \leq B$, satisfy the SGC with $\varrho \geq 1$ and using exponentially decreasing step-sizes $\eta_t = \eta_0 \, \alpha^t$ where $\eta_0 < \frac{1}{L_1^2 B}$ and $\alpha = \left(\frac{1}{T}\right)^{\frac{1}{T}}$ results in the following convergence:

If $\mathbb{E}[f^* - f(\theta_t)] > \epsilon$ for all $t \in [1, T]$ and $T_0 := T \max\left\{\frac{\ln(\varrho\,\eta_0)}{\ln(T)}, 0\right\}$, then,

$$\mathbb{E}[f^* - f(\theta_{T+1})] \leq [f^* - f(\theta_1)]\, C_1 \exp\left(-\frac{\alpha\,\epsilon\, T}{\kappa \ln(T)}\right) + \frac{C_2 \sum_{t=1}^{T_0 - 1} \mathbb{E}[f^* - f(\theta_t)]}{\epsilon^2\, T^2}$$

Stochastic Softmax Policy Gradient

Observation: Stochastic SPG with exponential step-sizes can adapt to the decreasing σ_t.

Theorem [LARV'24]: For a given $\epsilon \in (0,1)$, running stochastic SPG with unbiased stochastic gradients that are bounded, i.e. $\|\nabla \widetilde{f}(\theta)\| \leq B$, satisfy the SGC with $\varrho \geq 1$ and using exponentially decreasing step-sizes $\eta_t = \eta_0 \, \alpha^t$ where $\eta_0 < \frac{1}{L_1^2 B}$ and $\alpha = \left(\frac{1}{T}\right)^{\frac{1}{T}}$ results in the following convergence:
If $\mathbb{E}[f^* - f(\theta_t)] > \epsilon$ for all $t \in [1, T]$ and $T_0 := T \max\left\{\frac{\ln(\varrho\, \eta_0)}{\ln(T)}, 0\right\}$, then,

$$\mathbb{E}[f^* - f(\theta_{T+1})] \leq [f^* - f(\theta_1)]\, C_1 \exp\left(-\frac{\alpha\, \epsilon\, T}{\kappa \ln(T)}\right) + \frac{C_2 \sum_{t=1}^{T_0 - 1} \mathbb{E}[f^* - f(\theta_t)]}{\epsilon^2\, T^2}$$

- Best case: Have knowledge of ϱ and can set $\eta_0 \leq 1/\varrho$. $T_0 = 0$ and setting $T = \tilde{O}(1/\epsilon)$ ensures that $\min_{t \in [1, T+1]} \mathbb{E}[f^* - f(\theta_t)] \leq \epsilon$. Matches the result in [Mei et al., 2023].
- Worst case: Since ρ is unknown, setting η_0 to be large can result in $T_0 = O(T)$. Ensuring $\min_{t \in [1, T+1]} \mathbb{E}[f^* - f(\theta_t)] \leq \epsilon$ requires $T = \tilde{O}(1/\epsilon^3)$ iterations.

Stochastic Softmax Policy Gradient

Observation: Stochastic SPG with exponential step-sizes can adapt to the decreasing σ_t.

Theorem [LARV'24]: For a given $\epsilon \in (0,1)$, running stochastic SPG with unbiased stochastic gradients that are bounded, i.e. $\|\nabla \widetilde{f}(\theta)\| \leq B$, satisfy the SGC with $\varrho \geq 1$ and using exponentially decreasing step-sizes $\eta_t = \eta_0 \, \alpha^t$ where $\eta_0 < \frac{1}{L_1^2 B}$ and $\alpha = \left(\frac{1}{T}\right)^{\frac{1}{T}}$ results in the following convergence:

If $\mathbb{E}[f^* - f(\theta_t)] > \epsilon$ for all $t \in [1, T]$ and $T_0 := T \max\left\{\frac{\ln(\varrho\,\eta_0)}{\ln(T)}, 0\right\}$, then,

$$\mathbb{E}[f^* - f(\theta_{T+1})] \leq [f^* - f(\theta_1)]\, C_1 \exp\left(-\frac{\alpha\,\epsilon\, T}{\kappa\, \ln(T)}\right) + \frac{C_2 \sum_{t=1}^{T_0 - 1} \mathbb{E}[f^* - f(\theta_t)]}{\epsilon^2\, T^2}$$

- Best case: Have knowledge of ϱ and can set $\eta_0 \leq 1/\varrho$. $T_0 = 0$ and setting $T = \tilde{O}(1/\epsilon)$ ensures that $\min_{t\in[1,T+1]} \mathbb{E}[f^* - f(\theta_t)] \leq \epsilon$. Matches the result in [Mei et al., 2023].
- Worst case: Since ρ is unknown, setting η_0 to be large can result in $T_0 = O(T)$. Ensuring $\min_{t\in[1,T+1]} \mathbb{E}[f^* - f(\theta_t)] \leq \epsilon$ requires $T = \tilde{O}(1/\epsilon^3)$ iterations.

✓ Using exponential step-sizes makes stochastic SPG robust to ϱ.

Stochastic Softmax Policy Gradient for Bandits

✓ For stochastic multi-armed bandit problems with rewards in $[0, 1]$, setting $\eta_0 \leq \frac{1}{18}$ and using importance-weighted reward estimates ensures the convergence rate on the previous slide.

Stochastic Softmax Policy Gradient for Bandits

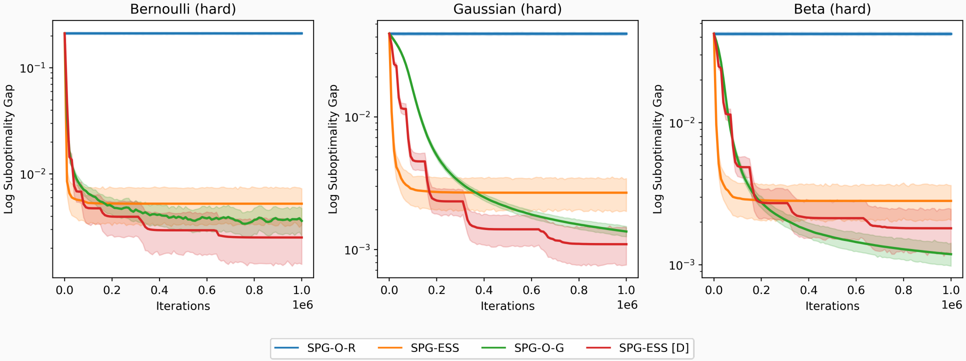
- ✓ For stochastic multi-armed bandit problems with rewards in $[0, 1]$, setting $\eta_0 \leq \frac{1}{18}$ and using importance-weighted reward estimates ensures the convergence rate on the previous slide.
- ✓ The result does not require the knowledge of problem-dependent constants (e.g. reward gap, variance or distribution of the rewards) nor does it require any explicit exploration.

Stochastic Softmax Policy Gradient for Bandits

- ✓ For stochastic multi-armed bandit problems with rewards in $[0, 1]$, setting $\eta_0 \leq \frac{1}{18}$ and using importance-weighted reward estimates ensures the convergence rate on the previous slide.
- ✓ The result does not require the knowledge of problem-dependent constants (e.g. reward gap, variance or distribution of the rewards) nor does it require any explicit exploration.

Stochastic Softmax Policy Gradient for Bandits

- ✓ For stochastic multi-armed bandit problems with rewards in $[0, 1]$, setting $\eta_0 \leq \frac{1}{18}$ and using importance-weighted reward estimates ensures the convergence rate on the previous slide.
- ✓ The result does not require the knowledge of problem-dependent constants (e.g. reward gap, variance or distribution of the rewards) nor does it require any explicit exploration.



Outline

- Problem Formulation
- Softmax Policy Gradient
- Stochastic Softmax Policy Gradient
- **Conclusion**

Conclusion

- ✓ Developed practical, principled variants of (stochastic) softmax PG in the tabular setting.
- ✓ Similar results for softmax PG with entropy regularization.
- × Step-sizes guaranteeing convergence of stochastic SPG are still quite conservative.

Conclusion

- ✓ Developed practical, principled variants of (stochastic) softmax PG in the tabular setting.
- ✓ Similar results for softmax PG with entropy regularization.
- × Step-sizes guaranteeing convergence of stochastic SPG are still quite conservative.

  Q: Can we use larger (constant) step-sizes (beyond those dependent on smoothness, SGC) and still guarantee theoretical convergence?

Conclusion

- ✓ Developed practical, principled variants of (stochastic) softmax PG in the tabular setting.
- ✓ Similar results for softmax PG with entropy regularization.
- × Step-sizes guaranteeing convergence of stochastic SPG are still quite conservative.

Q: Can we use larger (constant) step-sizes (beyond those dependent on smoothness, SGC) and still guarantee theoretical convergence?

Yes! Recent paper (with Jincheng Mei, Bo Dai, Alekh Agarwal, Anant Raj, Dale Schuurmans, Csaba Szepesvári) shows that stochastic SPG with *any* (potentially large) constant step-size guarantees that $\lim_{t\to\infty} \pi_{\theta_t}(a^*) \to 1$.

Conclusion

✓ Developed practical, principled variants of (stochastic) softmax PG in the tabular setting.

✓ Similar results for softmax PG with entropy regularization.

× Step-sizes guaranteeing convergence of stochastic SPG are still quite conservative.

Q: Can we use larger (constant) step-sizes (beyond those dependent on smoothness, SGC) and still guarantee theoretical convergence?

Yes! Recent paper (with Jincheng Mei, Bo Dai, Alekh Agarwal, Anant Raj, Dale Schuurmans, Csaba Szepesvári) shows that stochastic SPG with *any* (potentially large) constant step-size guarantees that $\lim_{t \to \infty} \pi_{\theta_t}(a^*) \to 1$.

Open questions: Do not have a handle on the algorithm's non-asymptotic behaviour or the convergence rate.

Conclusion

- ✓ Developed practical, principled variants of (stochastic) softmax PG in the tabular setting.
- ✓ Similar results for softmax PG with entropy regularization.
- × Step-sizes guaranteeing convergence of stochastic SPG are still quite conservative.

  Q: Can we use larger (constant) step-sizes (beyond those dependent on smoothness, SGC) and still guarantee theoretical convergence?

  Yes! Recent paper (with Jincheng Mei, Bo Dai, Alekh Agarwal, Anant Raj, Dale Schuurmans, Csaba Szepesvári) shows that stochastic SPG with *any* (potentially large) constant step-size guarantees that $\lim_{t \to \infty} \pi_{\theta_t}(a^*) \to 1$.

  Open questions: Do not have a handle on the algorithm's non-asymptotic behaviour or the convergence rate.

Future work:

- Generalize to (non)-linear policy parameterization.
- Generalize beyond softmax policies.

Questions?

Papers: https://arxiv.org/abs/2405.13136

Contact: vaswani.sharan@gmail.com, michael_lu_3@sfu.ca

Alekh Agarwal, Sham M Kakade, Jason D Lee, and Gaurav Mahajan. On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *J. Mach. Learn. Res.*, 22(98):1–76, 2021.

Jalaj Bhandari and Daniel Russo. On the linear convergence of policy gradient methods for finite mdps. In *International Conference on Artificial Intelligence and Statistics*, pages 2386–2394. PMLR, 2021.

Wesley Chung, Valentin Thomas, Marlos C Machado, and Nicolas Le Roux. Beyond variance reduction: Understanding the true impact of baselines on policy optimization. In *International Conference on Machine Learning*, pages 1999–2009. PMLR, 2021.

Xavier Fontaine, Valentin De Bortoli, and Alain Durmus. Convergence rates and approximation results for sgd and its continuous-time counterpart. In *Conference on Learning Theory*, pages 1965–2058. PMLR, 2021.

Ziwei Ji and Matus Telgarsky. Risk and parameter convergence of logistic regression. *arXiv preprint arXiv:1803.07300*, 2018.

Sajad Khodadadian, Prakirt Raj Jhunjhunwala, Sushil Mahavir Varma, and Siva Theja Maguluri. On the linear convergence of natural policy gradient algorithm. In *2021 60th IEEE Conference on Decision and Control (CDC)*, pages 3794–3799. IEEE, 2021.

Xiaoyu Li, Zhenxun Zhuang, and Francesco Orabona. A second look at exponential and cosine step sizes: Simplicity, adaptivity, and performance. In *International Conference on Machine Learning*, pages 6553–6564. PMLR, 2021.

Jincheng Mei, Chenjun Xiao, Csaba Szepesvari, and Dale Schuurmans. On the global convergence rates of softmax policy gradient methods. In *International Conference on Machine Learning*, pages 6820–6829. PMLR, 2020.

Jincheng Mei, Bo Dai, Chenjun Xiao, Csaba Szepesvari, and Dale Schuurmans. Understanding the effect of stochasticity in policy optimization. *Advances in Neural Information Processing Systems*, 34:19339–19351, 2021a.

Jincheng Mei, Yue Gao, Bo Dai, Csaba Szepesvari, and Dale Schuurmans. Leveraging non-uniformity in first-order non-convex optimization. In *International Conference on Machine Learning*, pages 7555–7564. PMLR, 2021b.

Jincheng Mei, Zixin Zhong, Bo Dai, Alekh Agarwal, Csaba Szepesvari, and Dale Schuurmans. Stochastic gradient succeeds for bandits. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 24325–24360. PMLR, 23–29 Jul 2023. URL `https://proceedings.mlr.press/v202/mei23a.html`.

Boris T Polyak. Introduction to optimization. 1987.

Mark Schmidt and Nicolas Le Roux. Fast convergence of stochastic gradient descent under a strong growth condition. *arXiv preprint arXiv:1308.6370*, 2013.

Sharan Vaswani, Francis Bach, and Mark Schmidt. Fast and faster convergence of sgd for over-parameterized models and an accelerated perceptron. In *The 22nd international conference on artificial intelligence and statistics*, pages 1195–1204. PMLR, 2019.

Sharan Vaswani, Benjamin Dubois-Taine, and Reza Babanezhad. Towards noise-adaptive, problem-adaptive (accelerated) stochastic gradient descent. In *International Conference on Machine Learning*, pages 22015–22059. PMLR, 2022.

Rui Yuan, Robert M. Gower, and Alessandro Lazaric. A general sample complexity analysis of vanilla policy gradient, 2022.

Backup Slides

Softmax PG: Experimental Results