NEURAL INFORMATION PROCESSING SYSTEMS

Understanding the Effect of Stochasticity in Policy Optimization

Jincheng Mei[1,2,*], Bo Dai[2], Chenjun Xiao[1,2],

Csaba Szepesvari[3,1], Dale Schuurmans[2,1]

[1]University of Alberta, [2]Google Brain, [3]DeepMind

UNIVERSITY OF ALBERTA

Main contributions

Preferability of policy optimization algorithms:

Faster methods in true gradient settings are dominated by slower counterparts

Softmax policy gradient (PG), natural PG (NPG), geometry-aware normalized PG (GNPG)

Committal rate:

Necessary condition for almost sure convergence to globally optimal policy

Geometry-convergence trade-off:

Cannot achieve almost sure global convergence with faster than O(1/t) rates

Explaining initialization sensitivity and ensemble methods

Preferability of policy optimization algorithms

One-state Markov Decision Processes (MDPs), deterministic reward

$$\max_{\theta:[K]\to\mathbb{R}} \mathbb{E}_{a\sim\pi_\theta(\cdot)}[r(a)]$$

Softmax parameterization

$$\pi_\theta(a) = \frac{\exp\{\theta(a)\}}{\sum_{a'\in[K]} \exp\{\theta(a')\}}$$

Algorithms: true gradient settings

Softmax policy gradient (PG):
$$\theta_{t+1} \leftarrow \theta_t + \eta \cdot \frac{d\pi_{\theta_t}^\top r}{d\theta_t}$$

Natural PG (NPG):
$$\theta_{t+1} \leftarrow \theta_t + \eta \cdot r$$

Geometry-aware normalized PG (GNPG):
$$\theta_{t+1} \leftarrow \theta_t + \eta \cdot \frac{d\pi_{\theta_t}^\top r}{d\theta_t} \Big/ \left\| \frac{d\pi_{\theta_t}^\top r}{d\theta_t} \right\|_2$$

Results: true gradient settings

	Softmax PG	NPG	GNPG
True gradient	converges $\Theta(1/t)$ ✓✓	converges $O(e^{-c \cdot t})$ ✓✓✓	converges $O(e^{-c \cdot t})$ ✓✓✓

Algorithms: on-policy stochastic gradient settings

On-policy importance sampling estimator:

Definition 1 (On-policy IS). *At iteration* t*, sample one action* $a_t \sim \pi_{\theta_t}$*. The IS reward estimator* $\hat{r}_t$ *is constructed as* $\hat{r}_t(a) = \frac{\mathbb{I}\{a_t = a\}}{\pi_{\theta_t}(a)} \cdot r(a)$ *for all* $a \in [K]$.

Algorithms: on-policy stochastic gradient settings

On-policy importance sampling estimator:

Definition 1 (On-policy IS). *At iteration t, sample one action $a_t \sim \pi_{\theta_t}$. The IS reward estimator $\hat{r}_t$ is constructed as $\hat{r}_t(a) = \frac{\mathbb{I}\{a_t = a\}}{\pi_{\theta_t}(a)} \cdot r(a)$ for all $a \in [K]$.*

Softmax policy gradient (PG): $$\theta_{t+1} \leftarrow \theta_t + \eta \cdot \frac{d\pi_{\theta_t}^\top \hat{r}_t}{d\theta_t}$$

Natural PG (NPG): $$\theta_{t+1} \leftarrow \theta_t + \eta \cdot \hat{r}_t$$

Geometry-aware normalized PG (GNPG): $$\theta_{t+1} = \theta_t + \eta \cdot \frac{d\pi_{\theta_t}^\top \hat{r}_t}{d\theta_t} \Big/ \left\| \frac{d\pi_{\theta_t}^\top \hat{r}_t}{d\theta_t} \right\|_2$$

Results: reversed in two settings

	Softmax PG	NPG	GNPG
True gradient	converges $\Theta(1/t)$ ✓✓	converges $O(e^{-c \cdot t})$ ✓✓✓	converges $O(e^{-c \cdot t})$ ✓✓✓
Stochastic on-policy	converges in prob. ✓	fails w.p. > 0 ✗	fails w.p. > 0 ✗

Results: reversed in two settings

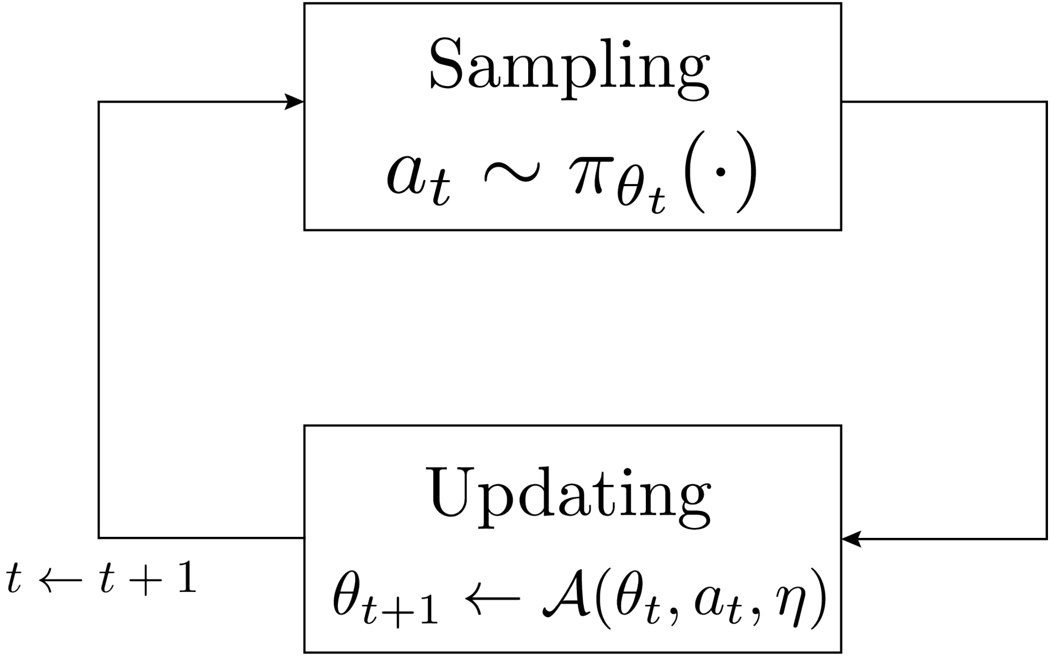
	Softmax PG	NPG	GNPG
True gradient	converges $\Theta(1/t)$ ✓✓	converges $O(e^{-c \cdot t})$ ✓✓✓	converges $O(e^{-c \cdot t})$ ✓✓✓
Stochastic on-policy	converges in prob. ✓	fails w.p. > 0 ✗	fails w.p. > 0 ✗

Same reason:

- **faster rate with true gradients**
- **failure with on-policy stochastic gradients**

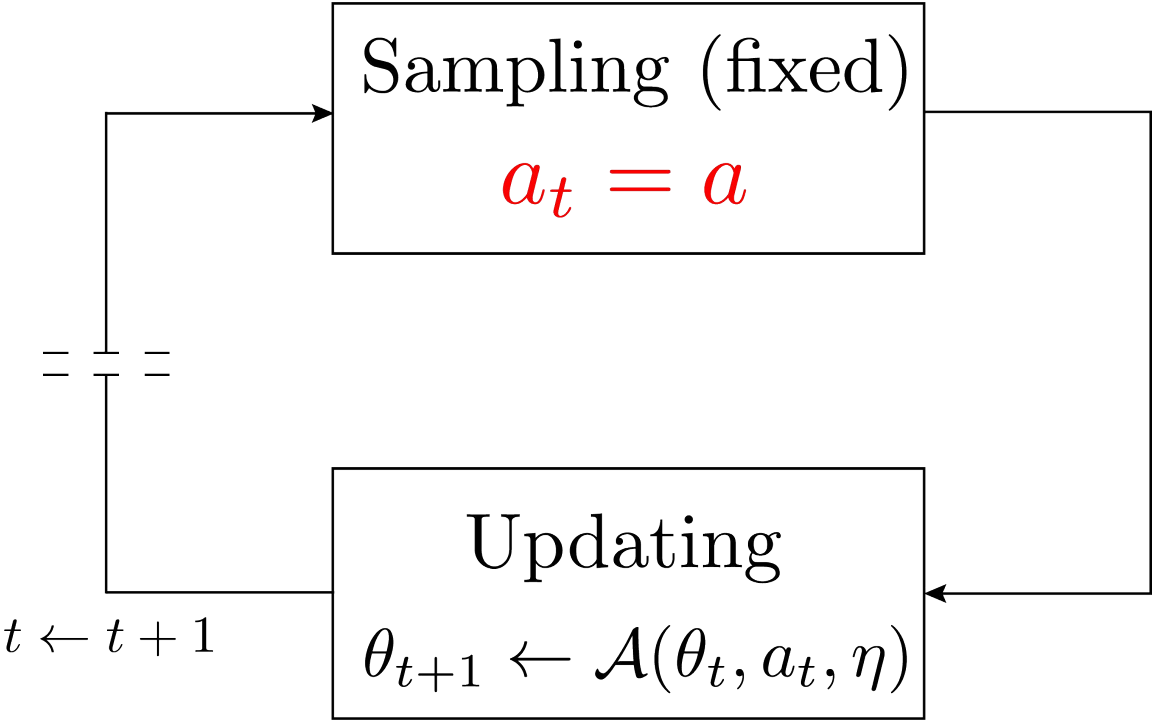
Difficulty: coupling in on-policy setting

Coupled circle between “sampling” and “updating”



Change stochastic behaviours to deterministic behaviours

Decouple the circle between “sampling” and “updating”



Committal rate: larger -- more aggressive

Fix sampling one action forever, measure **the aggressiveness of an update**

$$\kappa(\mathcal{A}, a) = \sup \left\{ \alpha \geq 0 : \limsup_{t \to \infty} t^{\alpha} \cdot [1 - \pi_{\theta_t}(a)] < \infty \right\}$$

Committal rate: larger -- more aggressive

Fix sampling one action forever, measure **the aggressiveness of an update**

$$\kappa(\mathcal{A}, a) = \sup \left\{ \alpha \geq 0 : \limsup_{t \to \infty} t^{\alpha} \cdot [1 - \pi_{\theta_t}(a)] < \infty \right\}$$

Examples:

$\pi_{\theta_t}(a) = 1 - 1/(t \cdot \log(t))$	$\kappa(\mathcal{A}, a) = 1$
$\pi_{\theta_t}(a) = 1 - 1/e^t$	$\kappa(\mathcal{A}, a) = \infty$
$1 - \pi_{\theta_t}(a) \in \Omega(1)$	$\kappa(\mathcal{A}, a) = 0$

Necessary condition for almost sure convergence

The following is a necessary condition for almost sure convergence to optimal policy, for any on-policy policy optimization algorithm:

$$\max_{a:r(a)<r(a^*),\pi_{\theta_1}(a)>0} \kappa(\mathcal{A}, a) \le 1$$

Necessary condition for almost sure convergence

The following is a necessary condition for almost sure convergence to optimal policy, for any on-policy policy optimization algorithm:

$$\max_{a: r(a) < r(a^*), \pi_{\theta_1}(a) > 0} \kappa(\mathcal{A}, a) \leq 1$$

Reason: If $\kappa(\mathcal{A}, a) > 1$, then $\Pr(a_t = a \text{ for all } t \geq 1 | a_t \sim \pi_{\theta_t}(\cdot)) > 0$

Aggressive updates could fail by sampling one action forever.

lack of exploration

“vicious circle” between sampling and updating

Necessary condition for almost sure convergence

The following is a necessary condition for almost sure convergence to optimal policy, for any on-policy policy optimization algorithm:

$$\max_{a: r(a) < r(a^*), \pi_{\theta_1}(a) > 0} \kappa(\mathcal{A}, a) \leq 1$$

Reason: If $\kappa(\mathcal{A}, a) > 1$, then $\Pr(a_t = a \text{ for all } t \geq 1 | a_t \sim \pi_{\theta_t}(\cdot)) > 0$

$$\prod_{t=1}^{\infty} \pi_{\theta_t}(a) > 0 \text{ if and only if } \sum_{t=1}^{\infty} (1 - \pi_{\theta_t}(a)) < \infty$$

One side: high committal rate → instability

The following is a necessary condition for almost sure convergence to optimal policy, for any on-policy policy optimization algorithm:

$$\max_{a: r(a) < r(a^*), \pi_{\theta_1}(a) > 0} \kappa(\mathcal{A}, a) \leq 1$$

Verification:

$$\kappa(\mathrm{NPG}, a) = \infty$$

$$\kappa(\mathrm{GNPG}, a) = \infty$$

$$\kappa(\mathrm{PG}, a) = 1$$

Another side: fast rate → high committal rate

If $O(1/t^{\alpha})$ with positive probability, then $\kappa(\mathcal{A}, a^*) \geq \alpha$

Another side: fast rate → high committal rate

If $O(1/t^{\alpha})$ with positive probability, then $\kappa(\mathcal{A}, a^*) \geq \alpha$

Reason: $$\left(\pi^* - \pi_{\theta_t}\right)^{\top} r \geq \left(1 - \pi_{\theta_t}(a^*)\right) \cdot \Delta$$

A tension between aggressiveness and stability.

Geometry-Convergence Trade-off

If $\kappa(\mathcal{A}, a^*) = \kappa(\mathcal{A}, a)$ for at least one sub-optimal action a, then the algorithm can achieve **at most one** of the following two properties:

(1) converges to a globally optimal policy almost surely

(2) converges to a deterministic policy at a rate faster than O(1/t) w.p. > 0

Geometry-Convergence Trade-off

If $\kappa(\mathcal{A}, a^*) = \kappa(\mathcal{A}, a)$ for at least one sub-optimal action a, then the algorithm can achieve **at most one** of the following two properties:

(1) converges to a globally optimal policy almost surely

(2) converges to a deterministic policy at a rate faster than O(1/t) w.p. > 0

Can achieve none of them: $\theta_{t+1} \leftarrow \theta_t$ ("staying")

Difference with Omega(log T) bandit lower bounds: holds for deterministic reward settings.

Geometry-Convergence Trade-off

“If $\kappa(\mathcal{A}, a^*) = \kappa(\mathcal{A}, a)$ for at least one sub-optimal action a” is **necessary, otherwise the trade-off could be bypassed.**

Geometry-Convergence Trade-off

“If $\kappa(\mathcal{A}, a^*) = \kappa(\mathcal{A}, a)$ for at least one sub-optimal action a” is **necessary, otherwise the trade-off could be bypassed.**

If it is broken (“oracle baseline”): $r(a_1) = 1, \quad r(a_2) = -1$

$$\kappa(\mathrm{NPG}, a_1) = \infty, \quad \kappa(\mathrm{NPG}, a_2) = 0$$

Then NPG **achieves both** almost sure global convergence and a $O(e^{-c \cdot t})$ rate.

Explaining practical observations

Same method, different random seeds, very different performances

Practical methods like PPO are aggressive NPG based

On-policy NPG in its nature will converge to different places w.p. > 0

Explaining practical observations

Same method, different random seeds, very different performances

Practical methods like PPO are aggressive NPG based

On-policy NPG in its nature will converge to different places w.p. > 0

Ensemble methods: run log(1/delta) parallel instances, pick the best one

W.p. 1 - delta, the best one converges to the optimal policy

One-state MDPs to general MDPs

The results generalize to general finite Markov Decision Processes (MDPs).

Conclusions

Preferability of policy optimization algorithms:

Faster methods in true gradient settings are dominated by slower counterparts

Softmax policy gradient (PG), natural PG (NPG), geometry-aware normalized PG (GNPG)

Committal rate:

Necessary condition for almost sure convergence to globally optimal policy

Geometry-convergence trade-off:

Cannot achieve almost sure global convergence with faster than O(1/t) rates

Explaining initialization sensitivity and ensemble methods