Accelerated Policy Gradient: On the Nesterov Momentum for Reinforcement Learning

Yen-Ju Chen[1]*, Nai-Chieh Huang[1]*, Ping-Chun Hsieh[1]

[1] Department of Computer Science, National Yang Ming Chiao Tung University, Hsinchu, Taiwan

* Equal Contribution

Contribution

- We propose Accerelated Policy Gradient (APG), which leverages the Nesterov's momentum scheme to accelerate the convergence performance of PG for RL.
- We show that APG enjoys a $\tilde{O}(1/t^2)$ convergence rate under softmax policy parameterization and bandit setting.
- Through numerical validation on both bandit and MDP problems, we confirm that APG exhibits $\tilde{O}(1/t^2)$ rate and hence substantially improves the convergence behavior over the standard PG.

Accelerated Policy Gradient (APG)

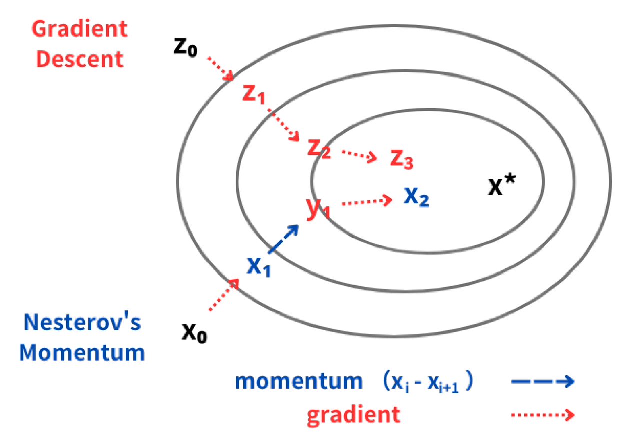
- Policy Gradient (PG)

$$\theta^{(t+1)} \leftarrow \theta^{(t)} + \eta \nabla_\theta V^{\pi_\theta}(\mu)\Big|_{\theta=\theta^{(t)}}$$

- Accelerated Policy Gradient (APG)

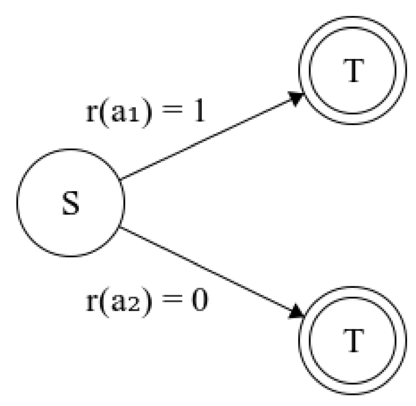
$$\theta^{(t)} \leftarrow \omega^{(t-1)} + \eta^{(t)} \nabla_\theta V^{\pi_\theta}(\mu)\Big|_{\theta=\omega^{(t-1)}}$$

$$\omega^{(t)} \leftarrow \theta^{(t)} + \frac{t-1}{t+2}(\theta^{(t)} - \theta^{(t-1)})$$



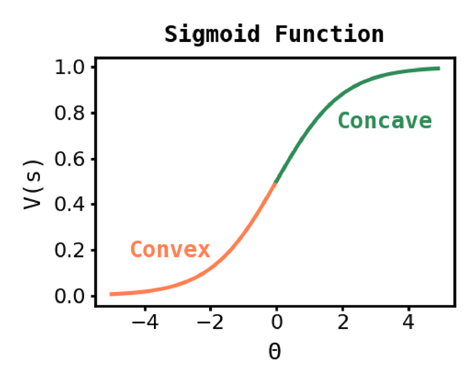
A Motivating Example: Insights of reaching $\tilde{O}(1/t^2)$

- **2-armed Bandit**
  - ▷ Policy parameters: $\theta = [\theta_1, \theta_2]$.
  - ▷ Reward: $r = [1, 0]$.



$$V(s) = \pi(a_1)\cdot r(a_1) + \pi(a_2)\cdot r(a_2) = \frac{\exp(\theta_1)}{\exp(\theta_1)+\exp(\theta_2)} = \frac{1}{1+\exp(\theta_2-\theta_1)} := \underbrace{\frac{1}{1+\exp(-\Theta)}}_{\text{sigmoid function}}, \text{ where } \Theta := \theta_1 - \theta_2$$

$\Rightarrow$ **Concave if $\Theta \geq 0$.**



Asymptotic Global Convergence of APG under Bandit Setting

- **Assumption 1 (Strict positivity of surrogate initial state distribution).** The surrogate initial state distribution satisfies $\min_s \mu(s) > 0$.
- **Assumption 2 (Bounded reward).** $r(s,a) \in [0,1], \forall s \in \mathcal{S}, a \in \mathcal{A}$.
- **Assumption 3 (Unique optimal action).** There is a unique optimal action a^* for each state $s \in \mathcal{S}$.
- **Theorem 1** Consider a tabular softmax parameterized policy π_θ. Under APG with $\eta^{(t)} = \frac{t}{t+1}\cdot\frac{1}{5}$, we have $V^{\pi_\theta^{(t)}}(s) \rightarrow V^*(s)$ as $t \rightarrow \infty$, for all $s \in \mathcal{S}$.

Convergence Rate of APG under Bandit Setting

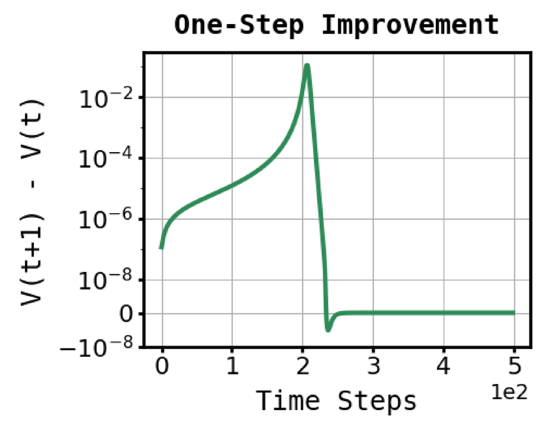
- **Lemma 1 (Local Concavity; Informal).** The function $\theta \rightarrow \pi_\theta^\top r$ is concave if $\theta_{a^*} - \theta_a > \delta$ for some $\delta > 0$, for all $a \neq a^*$.
- **Lemma 2 (Finite Time to Enter Local Concavity).** Consider a tabular softmax parameterized policy π_θ. Under APG with $\eta^{(t)} = \frac{t}{t+1}\cdot\frac{1}{5}$, given any $\delta > 0$, there exists a finite time T such that for all $t > T$, we have $\theta_{a^*} - \theta_a > \delta$, for all $a \neq a^*$.
- **Theorem 2** Consider a tabular softmax parameterized policy π_θ. Under APG with $\eta^{(t)} = \frac{t}{t+1}\cdot\frac{1}{5}$, there exists a finite time T such that for all $t > T$, we have:

$$\left(\pi^* - \pi_\theta^{(t)}\right)^\top r \leq \frac{|\mathcal{A}|-1}{(t-T)^2 + |\mathcal{A}| - 1} + \frac{10(2+T)\left(\left\|\theta^{(T)}\right\| + 2\ln(t-T)\right)^2}{t(t+1)}.$$

Difficulties in Analyzing Convergence: Non-Monotonic Improvement Under APG

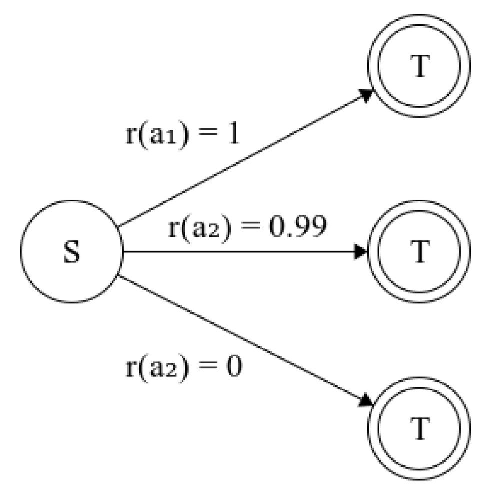
- **3-armed Bandit**
  - ▷ Highly sub-optimal policy initialization $\pi_0(a^*) \ll \pi_0(a_2) \ll \pi_0(a_3)$.
  - ▷ Attributed to the momentum term.

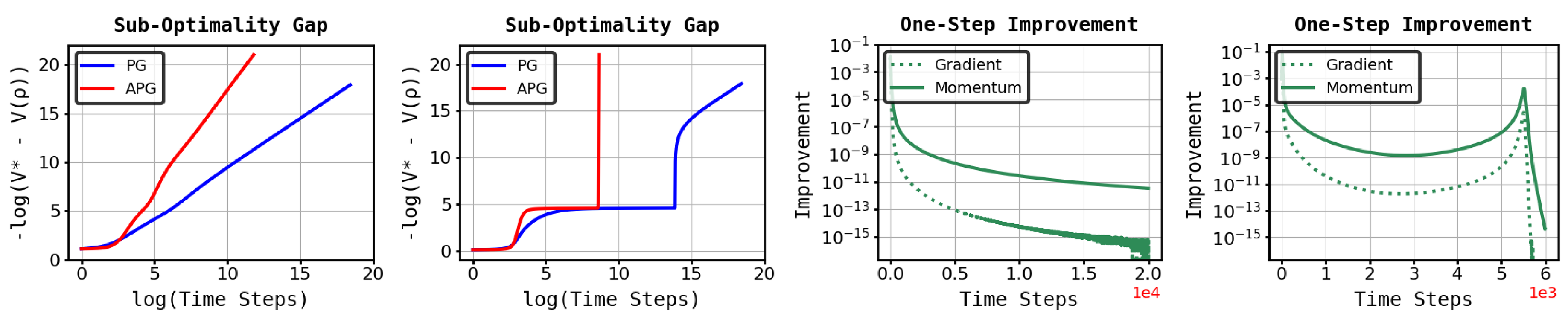
$\Rightarrow$ **Hard to reach first-order stationary points.**



Numerical Validation - Bandit Setting

- **3-armed Bandit**
  - ▷ Reward: $r = [1, 0.99, 0]$.
  - ▷ Uniform policy initialization: $\pi_0 = [1/3, 1/3, 1/3]$.
  - ▷ Hard policy initialization: $\pi_0 \approx [0.016, 0.117, 0.867]$.

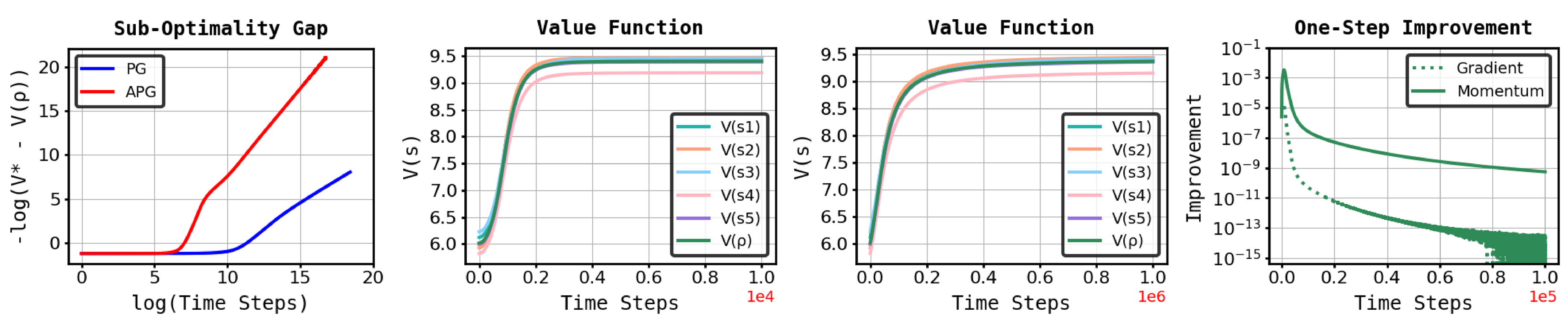




(a) Uniform Initialization (b) Hard Initialization (c) Uniform Initialization (d) Hard Initialization

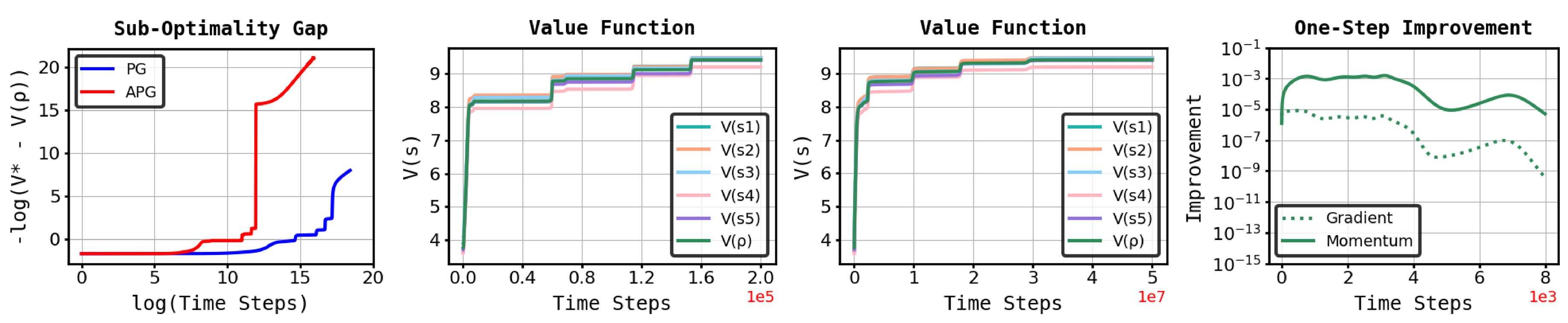
Numerical Validation - MDP Setting under Uniform Policy Initialization

- **MDP with 5 states and 5 actions**
  - ▷ Uniform policy initialization: $\pi_0(a|s) = 1/|\mathcal{A}|, \;\forall (s,a) \in \mathcal{S}\times\mathcal{A}$.



Numerical Validation - MDP Setting under Hard Policy Initialization

- **MDP with 5 states and 5 actions**
  - ▷ Hard policy initialization: $\pi_0(a|s) < \pi_0(a'|s)$, if $Q^*(s,a) > Q^*(s,a')$, $\forall (s,a,a') \in \mathcal{S}\times\mathcal{A}\times\mathcal{A}$.



Challenges of Convergence Analysis of APG for the General MDPs

- Due to the lack of monotonic improvement in the value function, it is not possible to apply the monotone convergence theorem on the limiting value function.
- True gradients with respect to a specific action under bandit setting:

$$\frac{\partial \pi_\theta^\top r}{\partial \theta_a} = \pi_\theta(a)\cdot(r(a) - \pi_\theta^\top r).$$

- True gradients with respect to a specific state-action under the general MDP setting:

$$\frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta_{s,a}}\Bigg|_{\theta=\theta} = \frac{1}{1-\gamma}\cdot d_\mu^{\pi_\theta}(s)\cdot\pi_\theta(a|s)\cdot(Q^{\pi_\theta}(s,a) - V^{\pi_\theta}(s)).$$

- The interdependence of Q^{π_θ} and V^{π_θ} in the MDP setting poses challenges in tracking the gradient value and analyzing the limiting value function.

https://github.com/NYCU-RL-Bandits-Lab/APG

{mru.11,naich.cs09,pinghsieh}@nycu.edu.tw