



On the Hidden Biases of Policy Mirror Ascent in Continuous Action Spaces

Amrit Singh Bedi¹, Souradip Chakraborty¹, Anjaly Parayil², Brian Sadler³,
Pratap Tokekar¹, and Alec Koppel⁴

¹University of Maryland, College Park, MD, USA

²Microsoft Research, India

³DEVCOM Army Research Laboratory, Adelphi, MD, USA

⁴JP Morgan & Chase AI Research

Contact: amritbd@umd.edu

Motivation

Consider the policy optimization problem

$$\max_{\theta} J(\theta) := V^{\pi_\theta}(s_0)$$

Motivation

Consider the policy optimization problem

$$\max_{\theta} J(\theta) := V^{\pi_\theta}(s_0)$$

- θ is the policy parameter

Motivation

Consider the policy optimization problem

$$\max_{\theta} J(\theta) := V^{\pi_\theta}(s_0)$$

- θ is the policy parameter
- $V^{\pi_\theta}(s_0)$ is value function given by

$$V^\pi(s_0) = \mathbb{E} \left[\sum_{t=0}^{\infty} \gamma^t r_t \mid a_t = \pi(s_t) \right]$$

Motivation

Consider the policy optimization problem

$$\max_{\theta} J(\theta) := V^{\pi_\theta}(s_0)$$

- θ is the policy parameter
- $V^{\pi_\theta}(s_0)$ is value function given by

$$V^\pi(s_0) = \mathbb{E} \left[\sum_{t=0}^{\infty} \gamma^t r_t \mid a_t = \pi(s_t) \right]$$

Policy Gradient Algorithm (Sutton et al., 2000)

$$\nabla J(\theta) = \frac{1}{1-\gamma} \cdot \mathbb{E} [\nabla \log \pi_\theta(a \mid s) \cdot Q^{\pi_\theta}(s, a)]$$

Motivation

Consider the policy optimization problem

$$\max_{\theta} J(\theta) := V^{\pi_\theta}(s_0)$$

- θ is the policy parameter
- $V^{\pi_\theta}(s_0)$ is value function given by

$$V^\pi(s_0) = \mathbb{E} \left[\sum_{t=0}^{\infty} \gamma^t r_t \mid a_t = \pi(s_t) \right]$$

Policy Gradient Algorithm (Sutton et al., 2000)

$$\nabla J(\theta) = \frac{1}{1-\gamma} \cdot \mathbb{E} [\nabla \log \pi_\theta(a \mid s) \cdot Q^{\pi_\theta}(s, a)]$$

- Softmax/direct parametrization shows global convergence

(Bhandari & Russo, 2019; Zhang et al., 2020a; Agarwal et al., 2020)

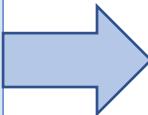
Motivation

Consider the policy optimization problem

$$\max_{\theta} J(\theta) := V^{\pi_\theta}(s_0)$$

- θ is the policy parameter
- $V^{\pi_\theta}(s_0)$ is value function given by

$$V^\pi(s_0) = \mathbb{E} \left[\sum_{t=0}^{\infty} \gamma^t r_t \mid a_t = \pi(s_t) \right]$$



Policy Gradient Algorithm (Sutton et al., 2000)

$$\nabla J(\theta) = \frac{1}{1-\gamma} \cdot \mathbb{E} [\nabla \log \pi_\theta(a \mid s) \cdot Q^{\pi_\theta}(s, a)]$$

- Softmax/direct parametrization shows global convergence
(Bhandari & Russo, 2019; Zhang et al., 2020a; Agarwal et al., 2020)
- General parametrizations in continuous state-action spaces
 - convergence to stationarity

(Bhatt et al., 2019; Zhang et al., 2020b)

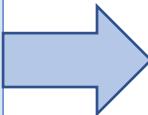
Motivation

Consider the policy optimization problem

$$\max_{\theta} J(\theta) := V^{\pi_\theta}(s_0)$$

- θ is the policy parameter
- $V^{\pi_\theta}(s_0)$ is value function given by

$$V^\pi(s_0) = \mathbb{E} \left[\sum_{t=0}^{\infty} \gamma^t r_t \mid a_t = \pi(s_t) \right]$$



Policy Gradient Algorithm (Sutton et al., 2000)

$$\nabla J(\theta) = \frac{1}{1-\gamma} \cdot \mathbb{E} [\nabla \log \pi_\theta(a \mid s) \cdot Q^{\pi_\theta}(s, a)]$$

- Softmax/direct parametrization shows global convergence
(Bhandari & Russo, 2019; Zhang et al., 2020a; Agarwal et al., 2020)
- General parametrizations in continuous state-action spaces
 - convergence to stationarity

(Bhatt et al., 2019; Zhang et al., 2020b)

There exists a Hidden Bias

Motivation

Consider the policy optimization problem

$$\max_{\theta} J(\theta) := V^{\pi_\theta}(s_0)$$

- θ is the policy parameter
- $V^{\pi_\theta}(s_0)$ is value function given by

$$V^\pi(s_0) = \mathbb{E} \left[\sum_{t=0}^{\infty} \gamma^t r_t \mid a_t = \pi(s_t) \right]$$

Policy Gradient Algorithm (Sutton et al., 2000)

$$\nabla J(\theta) = \frac{1}{1-\gamma} \cdot \mathbb{E} [\nabla \log \pi_\theta(a \mid s) \cdot Q^{\pi_\theta}(s, a)]$$

- Softmax/direct parametrization shows global convergence

(Bhandari & Russo, 2019; Zhang et al., 2020a; Agarwal et al., 2020)

- General parametrizations in continuous state-action spaces
 - convergence to stationarity

(Bhatt et al., 2019; Zhang et al., 2020b)

- Because of **Bounded score function** – standard assumption in the literature

Motivation

Consider the policy optimization problem

$$\max_{\theta} J(\theta) := V^{\pi_\theta}(s_0)$$

- θ is the policy parameter
- $V^{\pi_\theta}(s_0)$ is value function given by

$$V^\pi(s_0) = \mathbb{E} \left[\sum_{t=0}^{\infty} \gamma^t r_t \mid a_t = \pi(s_t) \right]$$

Policy Gradient Algorithm (Sutton et al., 2000)

$$\nabla J(\theta) = \frac{1}{1-\gamma} \cdot \mathbb{E} [\nabla \log \pi_\theta(a \mid s) \cdot Q^{\pi_\theta}(s, a)]$$

- Softmax/direct parametrization shows global convergence

(Bhandari & Russo, 2019; Zhang et al., 2020a; Agarwal et al., 2020)

- General parametrizations in continuous state-action spaces
 - convergence to stationarity

(Bhatt et al., 2019; Zhang et al., 2020b)

- Because of **Bounded score function** – standard assumption in the literature
- Does not hold for Gaussian policy parametrization

Motivation

Consider the policy optimization problem

$$\max_{\theta} J(\theta) := V^{\pi_\theta}(s_0)$$

- θ is the policy parameter
- $V^{\pi_\theta}(s_0)$ is value function given by

$$V^\pi(s_0) = \mathbb{E} \left[\sum_{t=0}^{\infty} \gamma^t r_t \mid a_t = \pi(s_t) \right]$$

Policy Gradient Algorithm (Sutton et al., 2000)

$$\nabla J(\theta) = \frac{1}{1-\gamma} \cdot \mathbb{E} [\nabla \log \pi_\theta(a \mid s) \cdot Q^{\pi_\theta}(s, a)]$$

- Softmax/direct parametrization shows global convergence

(Bhandari & Russo, 2019; Zhang et al., 2020a; Agarwal et al., 2020)

- General parametrizations in continuous state-action spaces
 - convergence to stationarity

(Bhatt et al., 2019; Zhang et al., 2020b)

Example- Gaussian Policy

- Because of **Bounded score function**

$$\pi_\theta(a|s) = \mathcal{N}(a|\varphi(s)^\top \theta, \sigma^2)$$

Motivation

Consider the policy optimization problem

$$\max_{\theta} J(\theta) := V^{\pi_\theta}(s_0)$$

- θ is the policy parameter
- $V^{\pi_\theta}(s_0)$ is value function given by

$$V^\pi(s_0) = \mathbb{E} \left[\sum_{t=0}^{\infty} \gamma^t r_t \mid a_t = \pi(s_t) \right]$$

Policy Gradient Algorithm (Sutton et al., 2000)

$$\nabla J(\theta) = \frac{1}{1-\gamma} \cdot \mathbb{E} [\nabla \log \pi_\theta(a \mid s) \cdot Q^{\pi_\theta}(s, a)]$$

- Softmax/direct parametrization shows global convergence

(Bhandari & Russo, 2019; Zhang et al., 2020a; Agarwal et al., 2020)

- General parametrizations in continuous state-action spaces
 - convergence to stationarity

(Bhatt et al., 2019; Zhang et al., 2020b)

Example- Gaussian Policy

- Because of **Bounded score function**

$$\nabla \log \pi_\theta(s, a) = \frac{(a - \varphi(s)^\top \theta) \varphi(s)}{\sigma^2}$$

Motivation

Consider the policy optimization problem

$$\max_{\theta} J(\theta) := V^{\pi_\theta}(s_0)$$

- θ is the policy parameter
- $V^{\pi_\theta}(s_0)$ is value function given by

$$V^\pi(s_0) = \mathbb{E} \left[\sum_{t=0}^{\infty} \gamma^t r_t \mid a_t = \pi(s_t) \right]$$

Policy Gradient Algorithm (Sutton et al., 2000)

$$\nabla J(\theta) = \frac{1}{1-\gamma} \cdot \mathbb{E} [\nabla \log \pi_\theta(a \mid s) \cdot Q^{\pi_\theta}(s, a)]$$

- Softmax/direct parametrization shows global convergence

(Bhandari & Russo, 2019; Zhang et al., 2020a; Agarwal et al., 2020)

- General parametrizations in continuous state-action spaces
 - convergence to stationarity

(Bhatt et al., 2019; Zhang et al., 2020b)

Example- Gaussian Policy

- Because of **Bounded score function**

$$\|\nabla \log \pi_\theta(s, a)\| \leq \mathcal{O}(D|a| + D^2\|\theta\|)$$

Motivation

Consider the policy optimization problem

$$\max_{\theta} J(\theta) := V^{\pi_\theta}(s_0)$$

- θ is the policy parameter
- $V^{\pi_\theta}(s_0)$ is value function given by

$$V^\pi(s_0) = \mathbb{E} \left[\sum_{t=0}^{\infty} \gamma^t r_t \mid a_t = \pi(s_t) \right]$$

Policy Gradient Algorithm (Sutton et al., 2000)

$$\nabla J(\theta) = \frac{1}{1-\gamma} \cdot \mathbb{E} [\nabla \log \pi_\theta(a \mid s) \cdot Q^{\pi_\theta}(s, a)]$$

- Softmax/direct parametrization shows global convergence
(Bhandari & Russo, 2019; Zhang et al., 2020a; Agarwal et al., 2020)
- General parametrizations in continuous state-action spaces
 - convergence to stationarity

(Bhatt et al., 2019; Zhang et al., 2020b)

Example- Gaussian Policy

- Because of **Bounded score function**

$$\|\nabla \log \pi_\theta(s, a)\| \leq \mathcal{O}(D|a| + D^2\|\theta\|)$$

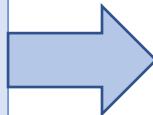
Motivation

Consider the policy optimization problem

$$\max_{\theta} J(\theta) := V^{\pi_\theta}(s_0)$$

- θ is the policy parameter
- $V^{\pi_\theta}(s_0)$ is value function given by

$$V^\pi(s_0) = \mathbb{E} \left[\sum_{t=0}^{\infty} \gamma^t r_t \mid a_t = \pi(s_t) \right]$$



Policy Gradient Algorithm (Sutton et al., 2000)

$$\nabla J(\theta) = \frac{1}{1-\gamma} \cdot \mathbb{E} [\nabla \log \pi_\theta(a \mid s) \cdot Q^{\pi_\theta}(s, a)]$$

- Softmax/direct parametrization shows global convergence
(Bhandari & Russo, 2019; Zhang et al., 2020a; Agarwal et al., 2020)
- General parametrizations in continuous state-action spaces
 - convergence to stationarity

(Bhatt et al., 2019; Zhang et al., 2020b)

- Because of **Bounded score function**
- Results in **Hidden bias** in the convergence rate

$$\mathcal{O}(1/\sqrt{K}) + \mathcal{O}(\lambda)$$

Exploration Tolerance parameter (Bedi et al., 2021)

Motivation

Question:

“Which policy parameterizations achieve exact convergence to stationarity in continuous state and action space?”

- Because of **Bounded score function**
- Results in **Hidden bias** in the convergence rate

$$\mathcal{O}(1/\sqrt{K}) + \mathcal{O}(\lambda)$$

Exploration Tolerance parameter (Bedi et al., 2021)

Proposed Algorithm: Stochastic Recursive Mirror Ascent

Idea: heavy-tailed policy parametrization

$$\pi_\theta(a|s) = \frac{1}{\sigma\pi(1 + ((a - \varphi(s)^\top\theta)/\sigma)^2)}$$

Bounded score function $\|\nabla \log \pi_\theta(s, a)\| \leq \frac{D}{\sigma}$

Proposed Algorithm: Stochastic Recursive Mirror Ascent Challenges

Proposed Algorithm: Stochastic Recursive Mirror Ascent Challenges

- Unstability in search directions
 - We *propose* mirror-ascent style updates

Proposed Algorithm: Stochastic Recursive Mirror Ascent Challenges

- Unstability in search directions
 - We *propose* mirror-ascent style updates

$$\theta_{k+1} = \arg \max_{\theta} \left\{ \langle \hat{\mathbf{g}}_k, \theta \rangle - \frac{1}{\eta} D_{\psi}(\theta, \theta_k) \right\}$$

Proposed Algorithm: Stochastic Recursive Mirror Ascent Challenges

- Unstability in search directions
 - We *propose* mirror-ascent style updates

$$\theta_{k+1} = \arg \max_{\theta} \left\{ \langle \hat{\mathbf{g}}_k, \theta \rangle - \frac{1}{\eta} D_{\psi}(\theta, \theta_k) \right\}$$

- But due to non-convexity, it requires *increasing batch size* of samples
 - We utilize momentum style updates

Proposed Algorithm: Stochastic Recursive Mirror Ascent Challenges

- Unstability in search directions
 - We *propose* mirror-ascent style updates

$$\theta_{k+1} = \arg \max_{\theta} \left\{ \langle \hat{\mathbf{g}}_k, \theta \rangle - \frac{1}{\eta} D_{\psi}(\theta, \theta_k) \right\}$$

- But due to non-convexity, it requires *increasing batch size* of samples
 - We utilize momentum style updates

$$\hat{\mathbf{g}}_k = (1 - \beta)(\hat{\mathbf{g}}_{k-1} - \tilde{\nabla} J(\theta_{k-1}, \xi_k(\theta_k))) + \nabla J(\theta_k, \xi_k(\theta_k))$$

Theoretical Result

Main Theorem

- To achieve $\min_{1 \leq k \leq K} \mathbb{E} [\|\text{Bregman Gradient}\|] \leq \epsilon$
- SRMA requires $K \geq \mathcal{O}\left(\frac{1}{\epsilon^4}\right)$ with $\mathcal{O}(1)$ samples at each k

Theoretical Result

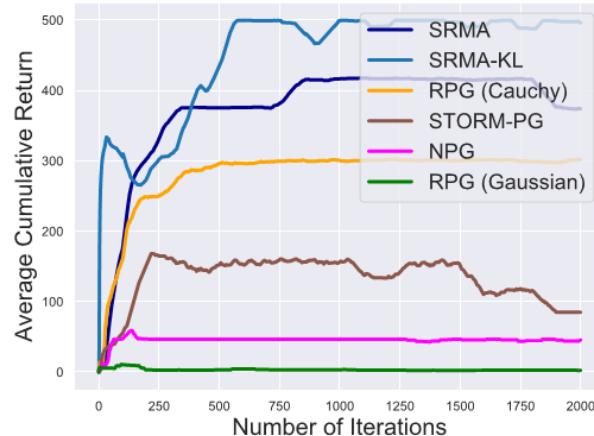
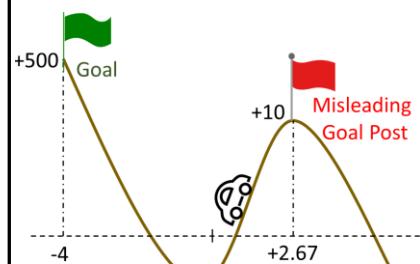
Main Theorem

- To achieve $\min_{1 \leq k \leq K} \mathbb{E} [\|\text{Bregman Gradient}\|] \leq \epsilon$
- SRMA requires $K \geq \mathcal{O}\left(\frac{1}{\epsilon^4}\right)$ with $\mathcal{O}(1)$ samples at each k

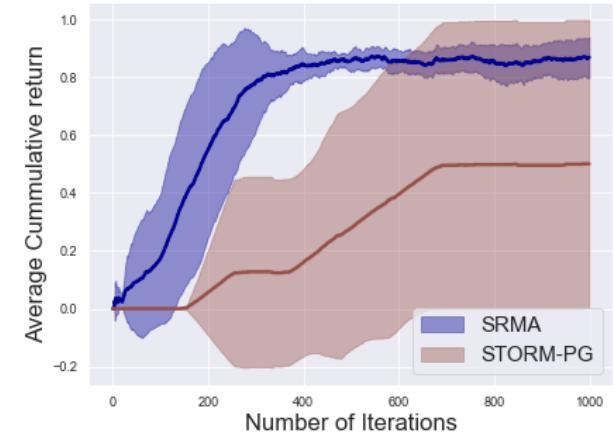
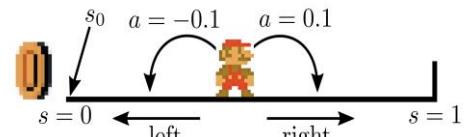
Algorithms	Hidden bias	Bregman term	SC
SVRPO (Xu et al., 2017)	Yes	No	N/A
SVRPG (Papini et al., 2018)	Yes	No	$\mathcal{O}(\epsilon^{-4})$
STORM-PG (Yuan et al., 2020)	Yes	No	$\mathcal{O}(\epsilon^{-4})$
RPG (Zhang et al., 2020b)	Yes	No	$\mathcal{O}(\epsilon^{-4})$
SRMA (This work)	No	Yes	$\mathcal{O}(\epsilon^{-4})$

Experiments

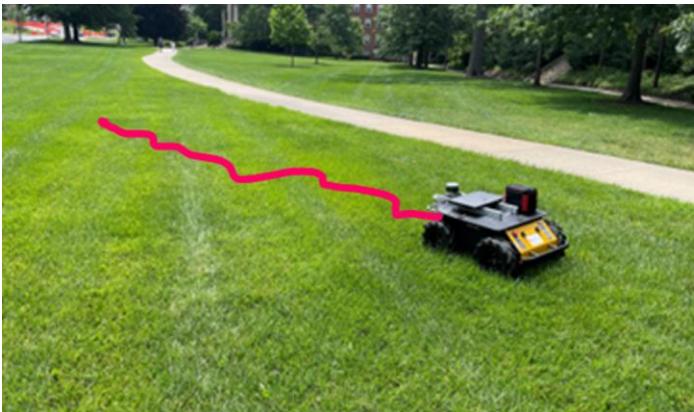
Pathological Mountain Car



Mario 1D (Sparse Reward Settings)



Real World Experiments* (additional)



(a) Reaching goal



(b) Obstacle avoidance



(c) Uneven terrain

Poster #907, Hall E
Thu 21 Jul 6 pm-8pm EST

Thank You
Contact: amritbd@umd.edu