Policy Gradient Using Weak Derivatives for Reinforcement Learning

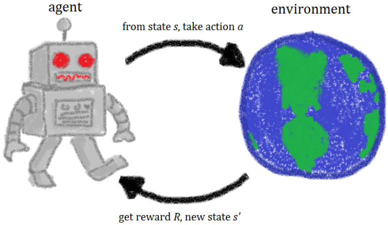
Sujay Bhatt[⋆] **Alec Koppel**[†] Vikram Krishnamurthy[⋆]
[⋆]Cornell University [†]Army Research Laboratory

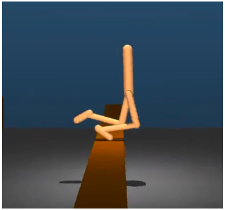
Learning Systems
IEEE Conference on Information Science and Systems

Mar. 20, 2019

Reinforcement Learning

- Reinforcement learning: data-driven control
  - $\Rightarrow$ unknown system model/cost function
  - $\Rightarrow$ parameterize policy/cost as stat. model for high dimensional spaces
- Recent successes:
  - $\Rightarrow$ AlphaGo Zero [Silver et al. '17]
  - $\Rightarrow$ Bipedal walker on terrain [Heess et al. '17]
  - $\Rightarrow$ Personalized web services [Theocharous et al. '15]





Problem Formulation

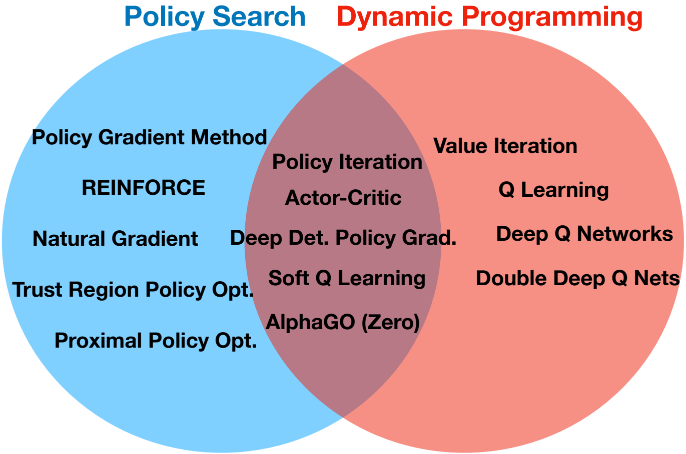
- Markov decision process (MDP) $(\mathcal{S}, \mathcal{A}, \mathbb{P}, R, \gamma)$
  - $\Rightarrow$ State space $\mathcal{S}$, action space $\mathcal{A}$ (high-dim. or even continuous)
  - $\Rightarrow$ Markov transition kernel $\mathbb{P}(s' \mid s, a) : \mathcal{S} \times \mathcal{A} \to \mathcal{P}(\mathcal{S})$
  - $\Rightarrow$ Reward $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, discount factor $\gamma \in (0, 1)$

- Stochastic policy $\pi : \mathcal{S} \to \mathcal{P}(\mathcal{A})$, i.e., $a_t \sim \pi(\cdot \mid s_t)$
- Infinite-horizon setting value function:

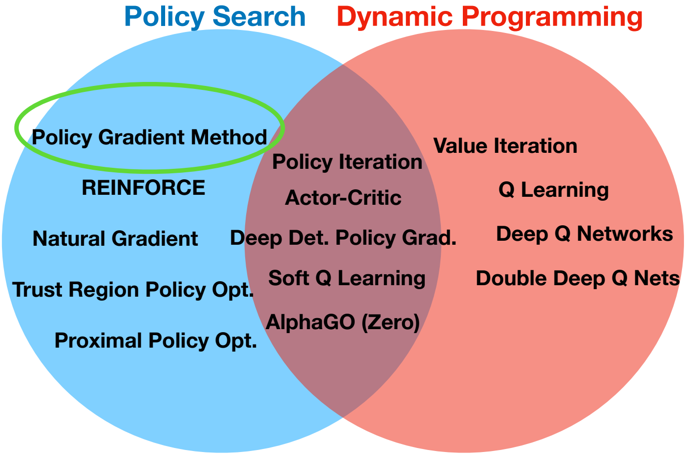
$$V(s) = \mathbb{E}\left(\sum_{t=0}^{\infty} \gamma^t \cdot R(s_t, a_t) \,\middle|\, s_0 = s \right),$$

- Goal: find $\{a_t = \pi(s_t)\}$ to maximize $V_\pi(s) := \mathbb{E}[V(s) \mid a \sim \pi(s)]$
- $\max_{\pi \in \Pi} V_\pi(s)$ where Π is some family of distributions
  - $\Rightarrow$ E.g., Gaussian $\pi = \pi_\theta$ w/ $\theta \in \mathbb{R}^d$ $\Rightarrow \pi_\theta(\cdot \mid s) = \mathcal{N}(\phi(s)^\top \theta, \sigma^2)$
  - $\Rightarrow$ Define action-state value (Q) function $Q_\pi(s, a) = \mathbb{E}[V_\pi(s) \mid a_0 = a]$

Literature Landscape



Literature Landscape



Classic Policy Gradient Theorem

- Policy gradient formula [Sutton '00]

$$\nabla J(\theta) = \frac{1}{1-\gamma} \cdot \mathbb{E}_{(s,a)\sim\rho_\theta(\cdot,\cdot)}\left[\nabla \log \pi_\theta(a \,|\, s) \cdot Q_{\pi_\theta}(s,a)\right].$$

$\Rightarrow$ $\rho_\theta(s,a)$ $\Rightarrow$ ergodic dist. of Markov chain for fixed policy:

$$\rho_\theta(s,a) = (1-\gamma)\sum_{t=0}^{\infty}\gamma^t p(s_t = s \,|\, s_0, \pi_\theta) \cdot \pi_\theta(a \,|\, s).$$

- Estimating Score function: $\mathcal{O}(N)$ variance. for N samples

  $\Rightarrow$ See POMDPs, V. Krishnamurthy, Cambridge University Press, 2016

Literature Landscape

: Impact of Nondeterminism on Reproducibility in Deep Reinforcement Learning

Prabhat Nagarajan
partment of Computer Science
: University of Texas at Austin
›rabhatn@cs.utexas.edu

Garrett Warnell
Computational and Information Sciences Directorate
U.S. Army Research Laboratory
garrett.a.warnell.civ@mail.mil

Peter Stone
Department of Computer Science
The University of Texas at Austin
pstone@cs.utexas.edu

Abstract

While deep reinforcement lea
results reported in the literatur
in reproducibility can arise
to computational resources o
details. Another factor of part
ability to control for sources
is because DRL is faced with t

A Dissection of Overfitting and Generalization in Continuous Reinforcement Learning

Amy Zhang
McGill University
Facebook AI Research
amyzhang@fb.com

Nicolas Ballas
Facebook AI Research
ballasn@fb.com

Joelle Pineau
McGill University
Facebook AI Research
jpineau@fb.com

Abstract

ing are well known. How-
tic tools and remedies, was
work, we aim to offer new
of overfitting in deep Rein-

Deep Reinforcement Learning that Matters

Peter Henderson[1]*, Riashat Islam[1,2]*, Philip Bachman[2]
Joelle Pineau[1], Doina Precup[1], David Meger[1]
[1] McGill University, Montreal, Canada
[2] Microsoft Maluuba, Montreal, Canada
{peter.henderson, riashat.islam}@mail.mcgill.ca, phbachma@microsoft.com
{jpineau, dprecup}@cs.mcgill.ca, dmeger@cim.mcgill.ca

Abstract

In recent years, significant progress has been made in solving challenging problems across various domains using deep reinforcement learning (RL). Reproducing existing work and accurately judging the improvements offered by novel methods is vital to sustaining this progress. Unfortunately, reproducing results for state-of-the-art deep RL methods is seldom straightforward. In particular, non-determinism in standard benchmark environments, combined with variance intrinsic to the methods, can make reported results tough to interpret. Without significance metrics and tighter standardization of experimental reporting, it is difficult to determine whether improvements over the prior state-of-the-art are meaningful. In this paper, we investigate challenges posed by reproducibility, proper experimental techniques, and reporting procedures. We

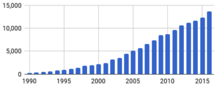
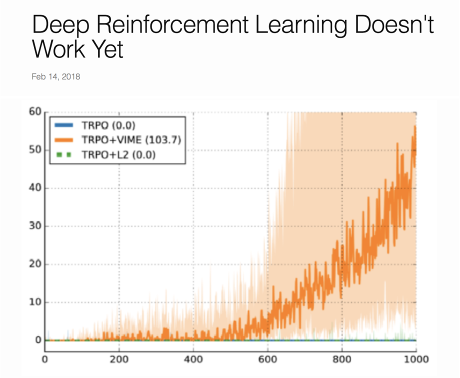


Figure 1: Growth of published reinforcement learning papers. Shown are the number of RL-related publications (y-axis) per year (x-axis) scraped from Google Scholar searches.

Policy Search is High Variance

- From Alexander Irpan's blog, software engineer at Google Brain:



- High sample path variance precludes practicality of Deep RL
  $\Rightarrow$ 30% failure rate is counted as working, publishable

New Policy Gradient Theorem via Weak Derivatives

- Policy gradient formula [Bhatt et al '19]

$$\nabla J(\theta) = \frac{1}{1-\gamma}\Big[\mathbb{E}_{(s,a)\sim\pi_\theta^{\oplus}(\cdot,\cdot)}\{g(\theta,s)\cdot Q_{\pi_\theta}(s,a)\} - \mathbb{E}_{(s,a)\sim\pi_\theta^{\ominus}(\cdot,\cdot)}\{g(\theta,s)\cdot Q_{\pi_\theta}(s,a)\}\Big].$$

  $\Rightarrow$ $g(\theta, s)$ $\Rightarrow$ normalizing constant, ensures $\pi^{\oplus}$, $\pi^{\ominus}$ valid distributions
- Note: no score function by differentiating w.r.t. policy directly!

  $\Rightarrow$ uses Hahn-Jordan signed decomposition of measures
- **Contribution**: Policy search via new expression for policy gradient

  $\Rightarrow$ establish almost sure convergence of these algs.

  $\Rightarrow$ yields **lower variance gradient estimates** for Gaussian policy

  $\Rightarrow$ Observe faster convergence on Pendulum

Weak Derivative Parameterization

- Consider Gaussian policy $\pi_\theta(\cdot \mid s) = \mathcal{N}(\theta^T \phi(s), \sigma^2)$
  $\Rightarrow$ mean is modulated by the optimization parameter θ.
- Jordan decomposition is as follows:

$$\nabla \pi_\theta(\cdot \mid s) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{(a - \theta^T \phi(s))^2}{2\sigma^2}\right) \times \frac{1}{\sigma^2}(a - \theta^T \phi(s)) \cdot \phi(s).$$
$$:= g(\theta, s)\left(\pi_\theta^{\oplus}(\cdot \mid s) - \pi_\theta^{\ominus}(\cdot \mid s)\right),$$

$\Rightarrow$ constant $g(\theta, s) = \frac{\phi(s)}{2\sqrt{2\pi\sigma^2}}$. Positive & negative measures:

$$\pi_\theta^{\oplus}(\cdot \mid s) = \frac{1}{\sigma^2}(a - \theta^T \phi(s)) \cdot \exp\left(\frac{(a - \theta^T \phi(s))^2}{2\sigma^2}\right),$$
$$\pi_\theta^{\ominus}(\cdot \mid s) = \frac{1}{\sigma^2}(\theta^T \phi(s) - a) \cdot \exp\left(\frac{(a - \theta^T \phi(s))^2}{2\sigma^2}\right).$$

- Note $\pi_\theta^{\oplus}(\cdot \mid s)$ and $\pi_\theta^{\ominus}(\cdot \mid s)$ are orthogonal Rayleigh distributions
  $\Rightarrow \pi_\theta^{\oplus}(\cdot \mid s)$ on $\mathbb{1}(a > \theta^T \phi(s))$; $\pi_\theta^{\ominus}(\cdot \mid s)$ domain $\mathbb{1}(a < \theta^T \phi(s))$.

Policy Search with Weak Derivatives

- Unbiasedly estimate $Q_{\pi_\theta^\oplus}(s,a)$ and $Q_{\pi_\theta^\ominus}(s,a)$ [Paternain 2018]:
  $\Rightarrow$ Draw $T' \sim \text{Geom}(1-\gamma)$, i.e., $P(T'=t) = (1-\gamma)\gamma$
  $\Rightarrow$ Monte Carlo rollout $\mathcal{R}^\oplus = (s_0^\oplus, a_0^\oplus, \cdots, s_{T'}^\oplus, a_{T'}^\oplus)$ and $\mathcal{R}^\ominus = (s_0^\ominus, a_0^\ominus, \cdots, s_{T'}^\ominus, a_{T'}^\ominus)$ associated w/ positive/negative measures

$$\hat{Q}_{\pi_\theta^\oplus}(s,a) = \sum_{t=0}^{T'} \gamma^t R(s_t^\oplus, a_t^\oplus) \Big| s_0 = s, a_0 = a$$

$$\hat{Q}_{\pi_\theta^\ominus}(s,a) = \sum_{t=0}^{T'} \gamma^t R(s_t^\ominus, a_t \ominus) \Big| s_0 = s, a_0 = a$$

Policy Search with Weak Derivatives

- Unbiasedly estimate $Q_{\pi_\theta^\oplus}(s,a)$ and $Q_{\pi_\theta^\ominus}(s,a)$ [Paternain 2018]:
  $\Rightarrow$ Draw $T' \sim \text{Geom}(1-\gamma)$, i.e., $P(T'=t) = (1-\gamma)\gamma$
  $\Rightarrow$ Monte Carlo rollout $\mathcal{R}^\oplus = (s_0^\oplus, a_0^\oplus, \cdots, s_{T'}^\oplus, a_{T'}^\oplus)$ and
  $\mathcal{R}^\ominus = (s_0^\ominus, a_0^\ominus, \cdots, s_{T'}^\ominus, a_{T'}^\ominus)$ associated w/ positive/negative measures

$$\hat{Q}_{\pi_\theta^\oplus}(s,a) = \sum_{t=0}^{T'} \gamma^t R(s_t^\oplus, a_t^\oplus) \Big| s_0 = s, a_0 = a$$

$$\hat{Q}_{\pi_\theta^\ominus}(s,a) = \sum_{t=0}^{T'} \gamma^t R(s_t^\ominus, a_t\ominus) \Big| s_0 = s, a_0 = a$$

- Draw (s,a) from $\rho_\theta(\cdot,\cdot)$:
  $\Rightarrow$ Draw $T \sim \text{Geom}(1-\gamma)$
  $\Rightarrow$ Rollout a trajectory $(s_0, a_0, s_1, \cdots, s_T, a_T)$
  $\Rightarrow$ Evaluate the gradient at (s_T, a_T)

$$\hat{\nabla} J(\theta) = \frac{g(\theta_T, s_T)}{1-\gamma} \left[\hat{Q}_{\pi_\theta^\oplus}(s_T, a_T) - \hat{Q}_{\pi_\theta^\ominus}(s_T, a_T) \right]$$

Policy Search with Weak Derivatives

- Unbiasedly estimate $Q_{\pi_\theta^\oplus}(s,a)$ and $Q_{\pi_\theta^\ominus}(s,a)$ [Paternain 2018]:
  $\Rightarrow$ Draw $T' \sim \text{Geom}(1-\gamma)$, i.e., $P(T'=t) = (1-\gamma)\gamma$
  $\Rightarrow$ Monte Carlo rollout $\mathcal{R}^\oplus = (s_0^\oplus, a_0^\oplus, \cdots, s_{T'}^\oplus, a_{T'}^\oplus)$ and $\mathcal{R}^\ominus = (s_0^\ominus, a_0^\ominus, \cdots, s_{T'}^\ominus, a_{T'}^\ominus)$ associated w/ positive/negative measures

$$\hat{Q}_{\pi_\theta^\oplus}(s,a) = \sum_{t=0}^{T'} \gamma^t R(s_t^\oplus, a_t^\oplus) \Big| s_0 = s, a_0 = a$$

$$\hat{Q}_{\pi_\theta^\ominus}(s,a) = \sum_{t=0}^{T'} \gamma^t R(s_t^\ominus, a_t \ominus) \Big| s_0 = s, a_0 = a$$

- Draw (s,a) from $\rho_\theta(\cdot,\cdot)$:
  $\Rightarrow$ Draw $T \sim \text{Geom}(1-\gamma)$
  $\Rightarrow$ Rollout a trajectory $(s_0, a_0, s_1, \cdots, s_T, a_T)$
  $\Rightarrow$ Evaluate the gradient at (s_T, a_T)

$$\hat{\nabla} J(\theta) = \frac{g(\theta_T, s_T)}{1-\gamma} \left[\hat{Q}_{\pi_\theta^\oplus}(s_T, a_T) - \hat{Q}_{\pi_\theta^\ominus}(s_T, a_T) \right]$$

- Policy Gradient update: $\theta_{k+1} = \theta_k + \alpha_k \hat{\nabla} J(\theta_k)$

Convergence Guarantee

- Asymptotic convergence to stationary points:

Theorem (Convergence with Diminishing Stepsize)

Let $\{\theta_k\}_{k\geq 0}$ *be the sequence of parameters of the policy* π_{θ_k} *given by RPG.*
If the stepsize $\{\alpha_k\}$ *satisfies*

$$\sum_{k=0}^{\infty} \alpha_k = \infty, \quad \sum_{k=0}^{\infty} \alpha_k^2 < \infty,$$

then we have $\theta_k \to \theta^*$ *where* θ^* *satisfies* $J(\theta^*) = 0$

- Avoid assumption on boundedness of iterates

  $\Rightarrow$ violated for most parameterizations, including Gaussian

Convergence Guarantee

- Convergence rate with diminishing stepsize

Theorem (Rate with Diminishing Stepsize)

Let $\{\theta_k\}_{k\geq 0}$ *be the sequence of parameters of the policy* π_{θ_k}. *Let the stepsize be* $\alpha_k = k^{-a}$ *where* $a \in (0,1)$. *Let*

$$K_\epsilon = \min\left\{k : \inf_{0\leq m\leq k} \mathbb{E}[\|\nabla J(\theta_m)\|^2] \leq \epsilon\right\} \leq \mathcal{O}(\epsilon^{-\frac{1}{2}})$$

$\Rightarrow$ Recover the $O(1/\sqrt{k})$ optimal rate of SGA for nonconvex opt.

Convergence Guarantee

Corollary

Let γ denote the discount factor and K_ϵ denote the iteration complexity. The average sample complexity

$$\left(\frac{1+\gamma}{1-\gamma}\right) K_\epsilon.$$

- Number of samples needed depends on discount factor

Convergence Guarantee

Theorem

The expected variance of the gradient estimates $\hat{\nabla} J(\theta)$ obtained using weak derivatives is given as:

$$\mathbb{E}\left\{\mathrm{var}^{WD}(\hat{\nabla} J(\theta)\right\} \leq \frac{2 \cdot M^2 \cdot G_{WD}}{(1-\gamma)^5},$$

where $G_{WD} = \mathbb{E}_{s\sim\pi_\theta}\left(\|g(\theta,s)\|^2\right)$. On the other hand, if score function is used instead of weak derivatives, the variance is

$$\mathbb{E}\left\{\mathrm{var}^{SF}(\hat{\nabla} J(\theta))\right\} \leq \frac{M^2 \cdot G_{SF}}{(1-\gamma)^5},$$

where $G_{SF} = \mathbb{E}_{(s,a)\sim\pi_\theta(a \mid s)}\left\{\|\nabla \pi_\theta(a \mid s)\|^2\right\}$.

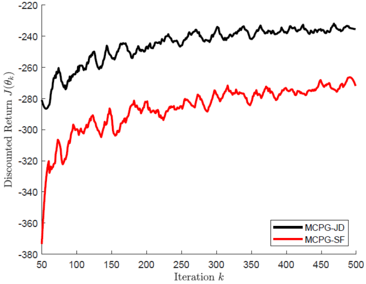
Corollary

For Gaussian policy $\pi_\theta(\cdot \mid s) = \mathcal{N}(\theta^T \phi(s), \sigma^2)$, we have $G_{WD} = \frac{1}{2\cdot\pi} G_{SF}$.

Pendulum Experiments

- Compare with Score function $\Rightarrow$ akin to REINFORCE [Williams '92]

  $\Rightarrow$ fixed Q function horizon estimate



  $\Rightarrow$ lower variance translates to faster learning in practice

  $\Rightarrow$ further experiments needed, hopefully during Sujay's postdoc

Conclusion

- Policy gradient method $\Rightarrow$ foundation of many RL methods
  $\Rightarrow$ scales gracefully to large problems, but afflicted with high variance
- We derive new policy gradient theorem based on Hahn-Jordan decomp.
  $\Rightarrow$ new policy search algorithms from this foundation
  $\Rightarrow$ provably convergent and lower variance than score function
- Experimentally observe these properties of policy search on pendulum
  $\Rightarrow$ solidified foundation for additional variance reduction techniques