Diverse Exploration via Conjugate Policies for Policy Gradient Methods

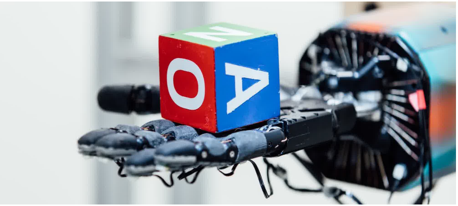
Andrew Cohen[1], Xingye Qiao[1], Lei Yu[1,2], Elliot Way[1], Xiangrong Tong[2]

SUNY Binghamton[1], Yantai University[2]

November 13, 2019

Introduction

- Reinforcement learning has great potential in enabling autonomous agents to robustly solve difficult problems.



- Common to these problems...
    - *Quickly* improve
        - → Explore!
    - Perform ’reasonably well’
        - → ~~Explore!~~
- **Diverse Exploration**: Explore the environment with a *diverse* set of ’good’ policies
    - First, explore in policy space

Key Contributions

- A variance analysis of the reparametrized policy gradient
- A diversity objective to reduce the gradient estimation variance
- An optimal solution via conjugate policies to the diversity objective
- DE algorithmic framework

Policy Gradients

- PG methods are state-of-the-art in training models with many parameters

$$J(\pi) = \mathbb{E}_{\tau}[R(\tau)] = \mathbb{E}_{s_0, a_0..}[\sum_{t=0}^{\infty} r(a_t, s_t)]$$

$$\nabla_{\theta}\mathbb{E}_{\tau\sim\pi}[R(\tau)] = \mathbb{E}_{\tau\sim\pi}[\sum_{t=0}^{T} \nabla_{\theta} \log(\pi(a_t|s_t;\theta))R_t(\tau)]$$

- Suffer from slow convergence, data inefficiency, high variance gradient estimates due to *a lack of exploration*

DE? But isn't PG on-policy?

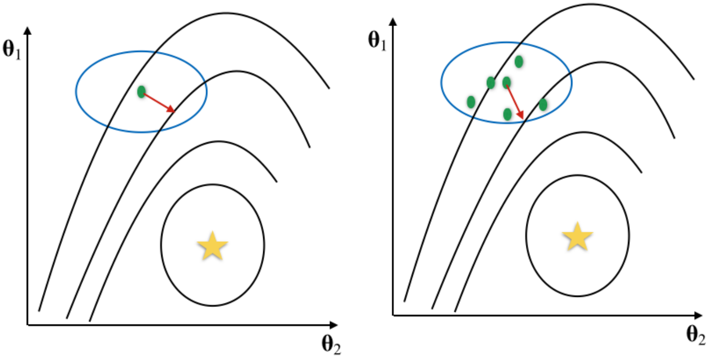
- Reparameterization allows incorporating diversity into PG methods without introducing off policy-bias

$$\mathbb{E}_{\substack{\theta\sim\mathcal{N}(\phi,\Sigma)\\ \tau\sim\pi_\theta}}[R(\tau)]$$

$$\nabla_{\phi,\Sigma}\mathbb{E}_{\substack{\epsilon\sim\mathcal{N}(0,I)\\ \tau\sim\pi}}[R(\tau)] = \mathbb{E}_{\substack{\epsilon\sim\mathcal{N}(0,I)\\ \tau\sim\pi}}[\sum_{t=0}^{T}\nabla_{\phi,\Sigma}\log(\pi(a_t|s_t;\phi+\epsilon\Sigma^{\frac{1}{2}}))R_t(\tau)]$$

What are "good" perturbations?

- *Local*
  - Small changes to the policy distribution yields similar performance.
- *Pairwise differences*
  - Diverse data from pairwise differences addresses shortcomings of PG methods.



Variance of gradient estimate

$$G_\epsilon := \mathbb{E}_{\tau\sim\pi_\epsilon}[\sum_{t=0}^{T}\gamma^t\nabla_\phi \log(\pi_\epsilon(a_t|s_t))R_t(\tau)]$$

$$\mathbb{V}_\epsilon(\frac{1}{k}\sum_{i=1}^{k}G_{\epsilon_i}) = \frac{1}{k^2}\sum_{i=1}^{k}\mathbb{V}_\epsilon(G_{\epsilon_i}) + \frac{2}{k^2}\sum_{i=1}^{k-1}\sum_{j=i+1}^{k} Cov(G_{\epsilon_i}, G_{\epsilon_j})$$

- $\frac{1}{k^2}\sum_{i=1}^{k}\mathbb{V}_\epsilon(G_{\epsilon_i}) = O(k^{-1})$
- Reduce *Cov* to reduce $\mathbb{V}_\epsilon(\frac{1}{k}\sum_{i=1}^{k}G_{\epsilon_i})$

Theorem 1

- $G_\epsilon := \mathbb{E}_{\tau\sim\pi_\epsilon}[\sum_{t=0}^{T}\gamma^t\nabla_\phi \log(\pi_\epsilon(a_t|s_t))R_t(\tau)]$
- *Local* perturbations: $Cov(\nabla_\phi \log(\pi_{\epsilon_i}), \nabla_\phi \log(\pi_{\epsilon_j}))$ drives $Cov(G_{\epsilon_i}, G_{\epsilon_j})$.
- **Two perturbations minimize covariance iff they maximize KL divergence.**

Theorem

Let ϵ_i and ϵ_j be two perturbations such that $\|\epsilon_i\|_2$ = $\left\|\epsilon_j\right\|_2 = \delta_\epsilon$. Then, (1) the trace of $\mathcal{C}ov(\nabla_\phi \log(\pi_{\epsilon_j}), \nabla_\phi \log(\pi_{\epsilon_i}))$ is minimized and (2) $\frac{1}{2}(\epsilon_j - \epsilon_i)^T\hat{F}(\epsilon_i)(\epsilon_j - \epsilon_i)$ the estimated KL divergence $D_{KL}(\pi_{\epsilon_i}||\pi_{\epsilon_j})$ is maximized, when $\epsilon_i = -\epsilon_j$ and they are along the direction of the eigenvector of $F(\epsilon_i)$ with the largest eigenvalue.

Objective

- From Theorem 1, maximize pairwise KL divergence between perturbations.
- $\tilde{D}_{KL}(\phi||\phi+\epsilon) = \frac{1}{2}\epsilon^T F_{\phi}\epsilon$
- $$\mathcal{P}^* = \arg\max_{\mathcal{P}} \sum_{i=1}^{k-1}\sum_{j=i+1}^{k} \tilde{D}_{KL}(\phi+\epsilon_j||\phi+\epsilon_i) \textit{ subject to } |\mathcal{P}| = k \leq n$$
- (Thm 2): *Conjugate perturbations* have maximal pairwise divergence
  - **Orthogonal wrt FIM**: $\epsilon_i A \epsilon_j = 0$

Approach and Baselines

- TRPO to estimate the natural gradient descent direction using perturbed gradient estimator over *k* perturbations.
- Conjugate gradient descent in NGD.
  - Free conjugate perturbations!
- Perturbation radius: $\tilde{D}_{KL}(\phi||\phi+\epsilon) \leq \delta_P$
- Experimental baselines:
  - RP: Gaussian parameter space noise $\epsilon \sim \mathcal{N}(0, I)$
  - TRPO: Sampling perturbations from the zero matrix

Performance and Covariance Results

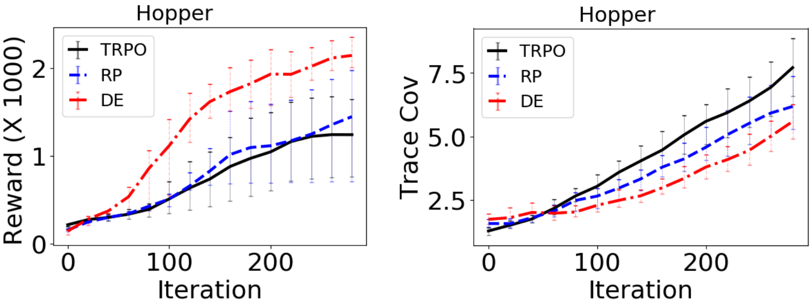


Figure: Comparison between TRPO, RP (TRPO with Random Perturbations), and DE (TRPO with Diverse Exploration) on average performance of all behavior policies and trace of the covariance matrix of perturbed gradient estimates, across iterations of learning on Hopper. Reported values are the average and interquartile range over 10 runs.

Decreasing the Number of Perturbations

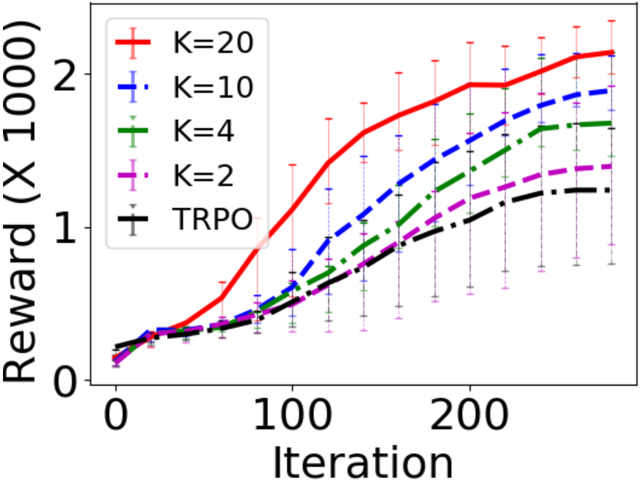


Figure: Average performance of all behavior policies for DE on Hopper with a decreasing number of perturbed policies and TRPO.

Summary

- A variance analysis of the reparametrized policy gradient
- A diversity objective to reduce the gradient estimation variance
- An optimal solution via conjugate policies to the diversity objective