Analysis and Improvement of Policy Gradient Estimation

Tingting Zhao Hirotaka Hachiya
Gang Niu Masashi Sugiyama

Tokyo Institute of Technology
2-12-1 O-okayama, Meguro-ku, Tokyo 152-8552, Japan.
{tingting@sg., hachiya@sg., gang@sg., sugi@} cs.titech.ac.jp
http://sugiyama-www.cs.titech.ac.jp/˜sugi

Abstract

Policy gradient is a useful model-free reinforcement learning approach, but it tends to suffer from instability of gradient estimates. In this paper, we analyze and improve the stability of policy gradient methods. We first prove that the variance of gradient estimates in the *PGPE* (policy gradients with parameter-based exploration) method is smaller than that of the classical REINFORCE method under a mild assumption. We then derive the optimal baseline for PGPE, which contributes to further reducing the variance. We also theoretically show that PGPE with the optimal baseline is more preferable than REINFORCE with the optimal baseline in terms of the variance of gradient estimates. Finally, we demonstrate the usefulness of the improved PGPE method through experiments.

1 Introduction

The goal of *reinforcement learning* (RL) is to find an optimal decision-making policy that maximizes the *return* (i.e., the sum of discounted rewards) through interaction with an unknown environment [17]. *Model-free* RL is a flexible framework in which decision-making policies are directly learned without going through explicit modeling of the environment. *Policy iteration* and *policy search* are two popular formulations of model-free RL[1].

In the policy iteration approach [7], the *value function* is first estimated and then policies are determined based on the learned value function. Policy iteration was demonstrated to work well in many real-world applications, especially in problems with discrete states and actions [18, 21, 1]. Although policy iteration can naturally deal with continuous states by function approximation [10], continuous actions are hard to handle due to the difficulty of finding maximizers of value functions with respect to actions. Moreover, since policies are indirectly determined via value function approximation, misspecification

[1]Policy iteration is originally a model-based RL approach, but it can be driven in a model-free mode by implicitly approximating an environment model with samples.

of value function models can lead to inappropriate policies even in very simple problems [19, 2]. Another limitation of policy iteration especially in physical control tasks is that control policies can vary drastically in each iteration. This causes severe instability in the physical system and thus is not favorable in practice.

Policy search is another approach to model-free RL that can overcome the limitations of policy iteration [22, 4, 8]. In the policy search approach, control policies are directly learned so that the return is maximized, for example, via a gradient method (called the *REINFORCE* method) [22], an EM algorithm [4], and a natural gradient method [8]. Among them, the gradient-based method is particularly useful in physical control tasks since policies are changed gradually. This ensures the stability of the physical system.

However, since the REINFORCE method tends to have a large variance in the estimation of the gradient directions, its naive implementation converges slowly [12, 14, 16]. Subtraction of the *optimal baseline* [20, 6] can ease this problem to some extent, but the variance of gradient estimates is still large. Furthermore, the performance heavily depends on the choice of an initial policy, and appropriate initialization is not straightforward in practice.

To cope with this problem, a novel policy gradient method called *policy gradients with parameter-based exploration* (PGPE) was proposed recently [16]. In PGPE, an initial policy is drawn from a prior probability distribution, and then actions are chosen deterministically. This construction contributes to mitigating the problem of initial policy choice and stabilizing gradient estimates [15]. Moreover, by subtracting a moving-average baseline, the variance of gradient estimates can be further reduced. Through robot-control experiments, PGPE was demonstrated to achieve more stable performance than existing policy-gradient methods.

The goal of this paper is to theoretically support the usefulness of PGPE, and to further improve its performance. More specifically, we first give bounds of the gradient estimates of the REINFORCE and PGPE methods. Our theoretical analysis shows that gradient estimates for PGPE have smaller variance than those for REINFORCE under a mild condition. We then show that the moving-average baseline for PGPE adopted in the original paper [16] has excess variance; we give the optimal baseline for PGPE that minimizes the variance, following the line of [20, 6]. We further theoretically show that PGPE with the optimal baseline is more preferable than REINFORCE with the optimal baseline in terms of the variance of gradient estimates. Finally, the usefulness of the improved PGPE method is demonstrated through experiments.

2 Policy Gradients for Reinforcement Learning

In this section, we review policy gradient methods.

2.1 Problem Formulation

Let us consider a Markov decision problem specified by $(\mathcal{S}, \mathcal{A}, P_T, P_I, r, \gamma)$, where $\mathcal{S}$ is a set of ℓ-dimensional continuous states, $\mathcal{A}$ is a set of continuous actions, $P_T(\boldsymbol{s'}|\boldsymbol{s}, a)$ is the transition probability density from current state $\boldsymbol{s}$ to next state $\boldsymbol{s'}$ when action a is taken, $P_I(\boldsymbol{s})$ is the probability of initial states, $r(\boldsymbol{s}, a, \boldsymbol{s'})$ is an immediate reward for transition from $\boldsymbol{s}$ to $\boldsymbol{s'}$ by taking action a, and $0 < \gamma < 1$ is the discounted factor for future rewards. Let $p(a|\boldsymbol{s}, \boldsymbol{\theta})$ be a stochastic policy with parameter $\boldsymbol{\theta}$, which represents the conditional probability density of taking action a in state $\boldsymbol{s}$.

Let $h = [\boldsymbol{s}_1, a_1, \ldots, \boldsymbol{s}_T, a_T]$ be a *trajectory* of length T. Then the *return* (i.e., the discounted sum of future rewards) along h is given by

$$R(h) := \sum_{t=1}^{T} \gamma^{t-1} r(\boldsymbol{s}_t, a_t, \boldsymbol{s}_{t+1}).$$

The expected return for parameter $\boldsymbol{\theta}$ is defined by

$$J(\boldsymbol{\theta}) := \int p(h|\boldsymbol{\theta}) R(h) \mathrm{d}h,$$

where

$$p(h|\boldsymbol{\theta}) = p(\boldsymbol{s}_1) \prod_{t=1}^{T} p(\boldsymbol{s}_{t+1}|\boldsymbol{s}_t, a_t) p(a_t|\boldsymbol{s}_t, \boldsymbol{\theta}).$$

The goal of reinforcement learning is to find the optimal policy parameter $\boldsymbol{\theta}^*$ that maximizes the expected return $J(\boldsymbol{\theta})$:

$$\boldsymbol{\theta}^* := \arg\max J(\boldsymbol{\theta}).$$

2.2 Review of the REINFORCE Algorithm

In the *REINFORCE* algorithm [22], the policy parameter $\boldsymbol{\theta}$ is updated via *gradient ascent*:

$$\boldsymbol{\theta} \longleftarrow \boldsymbol{\theta} + \varepsilon \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}),$$

where ε is a small positive constant. The gradient $\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta})$ is given by

$$\begin{aligned}
\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) &= \int \nabla_{\boldsymbol{\theta}} p(h|\boldsymbol{\theta}) R(h) \mathrm{d}h \\
&= \int p(h|\boldsymbol{\theta}) \nabla_{\boldsymbol{\theta}} \log p(h|\boldsymbol{\theta}) R(h) \mathrm{d}h \\
&= \int p(h|\boldsymbol{\theta}) \sum_{t=1}^{T} \nabla_{\boldsymbol{\theta}} \log p(a_t|\boldsymbol{s}_t, \boldsymbol{\theta}) R(h) \mathrm{d}h,
\end{aligned}$$

where we used the so-called 'log trick':

$$\nabla_{\boldsymbol{\theta}} p(h|\boldsymbol{\theta}) = p(h|\boldsymbol{\theta}) \nabla_{\boldsymbol{\theta}} \log p(h|\boldsymbol{\theta}).$$

Since $p(h|\boldsymbol{\theta})$ is unknown, the expectation is approximated by the empirical average:

$$\nabla_{\boldsymbol{\theta}} \widehat{J}(\boldsymbol{\theta}) = \frac{1}{N} \sum_{n=1}^{N} \sum_{t=1}^{T} \nabla_{\boldsymbol{\theta}} \log p(a_t^n | \boldsymbol{s}_t^n, \boldsymbol{\theta}) R(h^n),$$

where $h^n := [\boldsymbol{s}_1^n, a_1^n, \ldots, \boldsymbol{s}_T^n, a_T^n]$ is a roll-out sample.

Let us employ the Gaussian policy model with parameter $\boldsymbol{\theta} = (\boldsymbol{\mu}, \sigma)$, where $\boldsymbol{\mu}$ is the mean vector and σ is the standard deviation:

$$p(a|\boldsymbol{s}; \boldsymbol{\theta}) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(a - \boldsymbol{\mu}^\top \boldsymbol{s})^2}{2\sigma^2}\right).$$

Then the policy gradients are explicitly given as

$$\nabla_{\boldsymbol{\mu}} \log p(a|\boldsymbol{s}, \boldsymbol{\theta}) = \frac{a - \boldsymbol{\mu}^\top \boldsymbol{s}}{\sigma^2} \boldsymbol{s},$$
$$\nabla_{\sigma} \log p(a|\boldsymbol{s}, \boldsymbol{\theta}) = \frac{(a - \boldsymbol{\mu}^\top \boldsymbol{s})^2 - \sigma^2}{\sigma^3}.$$

A drawback of REINFORCE is that the variance of the above policy gradients is large [14, 16], which leads to slow convergence.

2.3 Review of the PGPE Algorithm

One of the reasons for large variance of policy gradients in the REINFORCE algorithm is that the empirical average is taken at each time step, which is caused by stochasticity of policies.

In order to mitigate this problem, another method called *policy gradients with parameter-based exploration* (PGPE) was proposed recently [16]. In PGPE, a linear *deterministic* policy (i.e., action a is chosen as $\boldsymbol{\theta}^\top \boldsymbol{s}$) is adopted, and stochasticity is introduced by considering $p(\boldsymbol{\theta}|\boldsymbol{\rho})$, a prior distribution over policy parameter $\boldsymbol{\theta}$ with hyper-parameter $\boldsymbol{\rho}$. Since entire history h is solely determined by a single sample of parameter $\boldsymbol{\theta}$ in this formulation, it is expected that the variance of gradient estimates can be reduced.

The expected return for hyper-parameter $\boldsymbol{\rho}$ is expressed as

$$J(\boldsymbol{\rho}) = \iint p(h|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\boldsymbol{\rho}) R(h) \mathrm{d}h \mathrm{d}\boldsymbol{\theta}.$$

Differentiating this with respect to $\boldsymbol{\rho}$, we have

$$\begin{aligned} \nabla_{\boldsymbol{\rho}} J(\boldsymbol{\rho}) &= \iint p(h|\boldsymbol{\theta}) \nabla_{\boldsymbol{\rho}} p(\boldsymbol{\theta}|\boldsymbol{\rho}) R(h) \mathrm{d}h \mathrm{d}\boldsymbol{\theta} \\ &= \iint p(h|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\boldsymbol{\rho}) \nabla_{\boldsymbol{\rho}} \log p(\boldsymbol{\theta}|\boldsymbol{\rho}) R(h) \mathrm{d}h \mathrm{d}\boldsymbol{\theta}, \end{aligned}$$

where the log trick for $\nabla_{\boldsymbol{\rho}} p(\boldsymbol{\theta}|\boldsymbol{\rho})$ is used. We then approximate the expectation over h and $\boldsymbol{\theta}$ by the empirical average:

$$\nabla_{\boldsymbol{\rho}} \widehat{J}(\boldsymbol{\rho}) = \frac{1}{N} \sum_{n=1}^{N} \nabla_{\boldsymbol{\rho}} \log p(\boldsymbol{\theta}^n|\boldsymbol{\rho}) R(h^n),$$

where each trajectory sample h^n is drawn from $p(h|\boldsymbol{\theta}^n)$ and parameter $\boldsymbol{\theta}^n$ is drawn from $p(\boldsymbol{\theta}^n|\boldsymbol{\rho})$.

Let us employ the Gaussian prior distribution with hyper-parameter $\boldsymbol{\rho} = (\boldsymbol{\eta}, \boldsymbol{\tau})$ to draw parameter vector $\boldsymbol{\theta}$, where $\boldsymbol{\eta}$ is the mean vector and $\boldsymbol{\tau}$ is the vector consisting of the standard deviation in each element:

$$p(\theta_i|\boldsymbol{\rho}_i) = \frac{1}{\tau_i \sqrt{2\pi}} \exp\left(-\frac{(\theta_i - \eta_i)^2}{2\tau_i^2}\right).$$

Then the derivative of $\log p(\boldsymbol{\theta}|\boldsymbol{\rho})$ with respect to η_i and τ_i are given as follows:

$$\nabla_{\eta_i} \log p(\boldsymbol{\theta}|\boldsymbol{\rho}) = \frac{\theta_i - \eta_i}{\tau_i^2},$$
$$\nabla_{\tau_i} \log p(\boldsymbol{\theta}|\boldsymbol{\rho}) = \frac{(\theta_i - \eta_i)^2 - \tau_i^2}{\tau_i^3}.$$

3 Variance of Gradient Estimates

In this section, we theoretically investigate the variance of gradient estimates in REINFORCE and PGPE.

For multi-dimensional state space, we consider the *trace* of the covariance matrix of gradient vectors. That is, for a random vector $\boldsymbol{A} = (A_1, \ldots, A_\ell)^\top$, we define

$$\begin{aligned}\mathbf{Var}(\boldsymbol{A}) &= \mathrm{tr}\left(\mathbb{E}\big[(\boldsymbol{A} - \mathbb{E}[\boldsymbol{A}])(\boldsymbol{A} - \mathbb{E}[\boldsymbol{A}])^\top\big]\right) \\ &= \sum_{m=1}^{\ell} \mathbb{E}\Big[(A_m - \mathbb{E}[A_m])^2\Big], \end{aligned} \tag{1}$$

where $\mathbb{E}$ denotes the expectation. Let

$$B = \sum_{i=1}^{\ell} \tau_i^{-2},$$

where ℓ is the dimensionality of state $\boldsymbol{s}$.

Below, we consider a subset of the following assumptions:

Assumption (A): $r(\boldsymbol{s}, a, \boldsymbol{s}') \in [-\beta, \beta]$ for $\beta > 0$.

Assumption (B): $r(\boldsymbol{s}, a, \boldsymbol{s}') \in [\alpha, \beta]$ for $0 < \alpha < \beta$.

Assumption (C): For $\delta > 0$, there exist two series $\{c_t\}_{t=1}^T$ and $\{d_t\}_{t=1}^T$ such that

$$\|\boldsymbol{s}_t\|_2 \geq c_t \quad \text{and} \quad \|\boldsymbol{s}_t\|_2 \leq d_t$$

hold with probability at least $(1-\delta)^{1/2N}$ respectively over the choice of sample paths, where $\|\cdot\|_2$ denotes the ℓ_2-norm.

Note that Assumption (B) is stronger than Assumption (A). Let

$$\mathcal{L}(T) = C_T \alpha^2 - D_T \beta^2/(2\pi),$$

where

$$C_T = \sum_{t=1}^{T} c_t^2 \quad \text{and} \quad D_T = \sum_{t=1}^{T} d_t^2.$$

First, we analyze the variance of gradient estimates in PGPE (the proofs of all the theorems are provided in Appendix):

Theorem 1. *Under Assumption (A), we have the following upper bounds:*

$$\mathbf{Var}\left[\nabla_{\boldsymbol{\eta}} \widehat{J}(\boldsymbol{\rho})\right] \leq \frac{\beta^2(1-\gamma^T)^2 B}{N(1-\gamma)^2},$$
$$\mathbf{Var}\left[\nabla_{\boldsymbol{\tau}} \widehat{J}(\boldsymbol{\rho})\right] \leq \frac{2\beta^2(1-\gamma^T)^2 B}{N(1-\gamma)^2}.$$

This theorem means that the upper bound of the variance of $\nabla_{\boldsymbol{\eta}} \widehat{J}(\boldsymbol{\rho})$ is proportional to β^2 (the upper bound of squared rewards), B (the trace of the inverse Gaussian covariance), and $(1-\gamma^T)^2/(1-\gamma)^2$, and is inverse-proportional to sample size N. The upper bound of the variance of $\nabla_{\boldsymbol{\tau}} \widehat{J}(\boldsymbol{\rho})$ is twice larger than that of $\nabla_{\boldsymbol{\eta}} \widehat{J}(\boldsymbol{\rho})$. When T goes to infinity, $(1-\gamma^T)^2$ will converge to 1.

Next, we analyze the variance of gradient estimates in REINFORCE:

Theorem 2. *Under Assumptions (B) and (C), we have the following lower bound with probability at least* $1-\delta$*:*

$$\mathbf{Var}\left[\nabla_{\boldsymbol{\mu}} \widehat{J}(\boldsymbol{\theta})\right] \geq \frac{(1-\gamma^T)^2}{N\sigma^2(1-\gamma)^2}\mathcal{L}(T).$$

Under Assumptions (A) and (C), we have the following upper bound with probability at least $(1-\delta)^{1/2}$*:*

$$\mathbf{Var}\left[\nabla_{\boldsymbol{\mu}} \widehat{J}(\boldsymbol{\theta})\right] \leq \frac{D_T \beta^2(1-\gamma^T)^2}{N\sigma^2(1-\gamma)^2}.$$

Under Assumption (A), we have

$$\mathbf{Var}\left[\nabla_{\sigma} \widehat{J}(\boldsymbol{\theta})\right] \leq \frac{2T\beta^2(1-\gamma^T)^2}{N\sigma^2(1-\gamma)^2}.$$

The upper bounds for REINFORCE are similar to those for PGPE, but they are monotone increasing with respect to trajectory length T. The lower bound for the variance of $\nabla_{\boldsymbol{\mu}}\widehat{J}(\boldsymbol{\theta})$ will be non-trivial if it is positive, i.e., $\mathcal{L}(T) > 0$. This can be fulfilled, e.g., if α and β satisfy

$$2\pi C_T \alpha^2 > D_T \beta^2.$$

Deriving a lower bound of the variance of $\nabla_{\sigma}\widehat{J}(\boldsymbol{\theta})$ is left open as future work.

Finally, we compare the variance of gradient estimates in REINFORCE and PGPE:

Theorem 3. *In addition to Assumptions (B) and (C), we assume $\mathcal{L}(T)$ is positive and monotone increasing with respect to T. If there exists T_0 such that $\mathcal{L}(T_0) \geq \beta^2 B\sigma^2$, then we have*

$$\mathbf{Var}[\nabla_{\boldsymbol{\mu}}\widehat{J}(\boldsymbol{\theta})] > \mathbf{Var}[\nabla_{\boldsymbol{\eta}}\widehat{J}(\boldsymbol{\rho})]$$

for all $T > T_0$, with probability at least $1 - \delta$.

The above theorem means that PGPE is more favorable than REINFORCE in terms of the variance of gradient estimates of the mean, if trajectory length T is large. This theoretical result would partially support the experimental success of the PGPE method [16].

4 Variance Reduction by Subtracting Baseline

In this section, we give a method to reduce the variance of gradient estimates in PGPE and analyze its theoretical properties.

4.1 Basic Idea of Introducing Baseline

It is known that the variance of gradient estimates can be reduced by subtracting a *baseline* b: for REINFORCE and PGPE, modified gradient estimates are given by

$$\nabla_{\boldsymbol{\theta}}\widehat{J}^b(\boldsymbol{\theta}) = \frac{1}{N}\sum_{n=1}^{N}(R(h^n) - b)\sum_{t=1}^{T}\nabla_{\boldsymbol{\theta}}\log p(a_t^n|\boldsymbol{s}_t^n, \boldsymbol{\theta}),$$

$$\nabla_{\boldsymbol{\rho}}\widehat{J}^b(\boldsymbol{\rho}) = \frac{1}{N}\sum_{n=1}^{N}(R(h^n) - b)\nabla_{\boldsymbol{\rho}}\log p(\boldsymbol{\theta}^n|\boldsymbol{\rho}).$$

The *adaptive reinforcement baseline* [22] was derived as the exponential moving average of the past experience:

$$b(n) = \gamma R(h^{n-1}) + (1-\gamma)b(n-1),$$

where $0 < \gamma \leq 1$. Based on this, an empirical gradient estimate with the moving-average baseline was proposed for REINFORCE [22] and PGPE [16].

The above moving-average baseline contributes to reducing the variance of gradient estimates. However, it was shown [6, 20] that the moving-average baseline is not optimal; the optimal baseline is, by definition, given as the minimizer of the variance of gradient estimates with respect to a baseline. Following this formulation, the optimal baseline for REINFORCE is given as follows [14]:

$$\begin{aligned} b^*_{\mathrm{REINFORCE}} &:= \arg\min_b \mathbf{Var}[\nabla_{\boldsymbol{\theta}} \widehat{J}^b(\boldsymbol{\theta})] \\ &= \frac{E[R(h)\| \sum_{t=1}^T \nabla_{\boldsymbol{\theta}} \log p(a_t|\boldsymbol{s}_t, \boldsymbol{\theta})\|^2]}{E[\| \sum_{t=1}^T \nabla_{\boldsymbol{\theta}} \log p(a_t|\boldsymbol{s}_t, \boldsymbol{\theta})\|^2]}. \end{aligned} \tag{2}$$

However, only the moving-average baseline was introduced to PGPE so far [16], which is suboptimal. Below, we derive the optimal baseline for PGPE, and study its theoretical properties.

4.2 Optimal Baseline for PGPE

Let b^*_{PGPE} be the optimal baseline for PGPE that minimizes the variance:

$$b^*_{\mathrm{PGPE}} := \arg\min_b \mathbf{Var}[\nabla_{\boldsymbol{\rho}} \widehat{J}^b(\boldsymbol{\rho})].$$

Then the following theorem gives the optimal baseline for PGPE:

Theorem 4. *The optimal baseline for PGPE is given by*

$$b^*_{\mathrm{PGPE}} = \frac{\mathbb{E}[R(h)\|\nabla_{\boldsymbol{\rho}} \log p(\boldsymbol{\theta}|\boldsymbol{\rho})\|^2]}{\mathbb{E}[\|\nabla_{\boldsymbol{\rho}} \log p(\boldsymbol{\theta}|\boldsymbol{\rho})\|^2]},$$

and the excess variance for a baseline b is given by

$$\mathbf{Var}[\nabla_{\boldsymbol{\rho}} \widehat{J}^b(\boldsymbol{\rho})] - \mathbf{Var}[\nabla_{\boldsymbol{\rho}} \widehat{J}^{b^*_{\mathrm{PGPE}}}(\boldsymbol{\rho})] = \frac{(b - b^*_{\mathrm{PGPE}})^2}{N} \mathbb{E}[\|\nabla_{\boldsymbol{\rho}} \log p(\boldsymbol{\theta}|\boldsymbol{\rho})\|^2].$$

The above theorem gives an analytic-form expression of the optimal baseline for PGPE. When expected return $R(h)$ and the squared norm of characteristic eligibility $\|\nabla_{\boldsymbol{\rho}} \log p(\boldsymbol{\theta}|\boldsymbol{\rho})\|^2$ are independent of each other, the optimal baseline is reduced to average expected return $\mathbb{E}[R(h)]$. However, the optimal baseline is generally different from the average expected return. The above theorem also shows that the excess variance is proportional to the squared difference of baselines $(b - b^*_{\mathrm{PGPE}})^2$ and the expected squared norm of characteristic eligibility $\mathbb{E}[\|\nabla_{\boldsymbol{\rho}} \log p(\boldsymbol{\theta}|\boldsymbol{\rho})\|^2]$, and is inverse-proportional to sample size N.

Next, we analyze the contribution of the optimal baseline to the variance with respect to mean parameter $\boldsymbol{\eta}$ in PGPE:

Theorem 5. *If $r(\boldsymbol{s}, a, \boldsymbol{s}') \geq \alpha > 0$, we have the following lower bound:*

$$\mathbf{Var}[\nabla_{\boldsymbol{\eta}} \widehat{J}(\boldsymbol{\rho})] - \mathbf{Var}[\nabla_{\boldsymbol{\eta}} \widehat{J}^{b^*_{\mathrm{PGPE}}}(\boldsymbol{\rho})] \geq \frac{\alpha^2 (1 - \gamma^T)^2 B}{N(1 - \gamma)^2}.$$

Under Assumption (A), we have the following upper bound:

$$\mathbf{Var}[\nabla_{\boldsymbol{\eta}}\widehat{J}(\boldsymbol{\rho})] - \mathbf{Var}[\nabla_{\boldsymbol{\eta}}\widehat{J}^{b^*_{\mathrm{PGPE}}}(\boldsymbol{\rho})] \leq \frac{\beta^2(1-\gamma^T)^2 B}{N(1-\gamma)^2}.$$

This theorem shows that the lower and upper bounds of the excess variance are proportional to α^2 and β^2 (the bounds of squared immediate rewards), B (the trace of the inverse Gaussian covariance), and $(1-\gamma^T)^2/(1-\gamma)^2$, and are inverse-proportional to sample size N. When T goes to infinity, $(1-\gamma^T)^2$ will converge to 1.

4.3 Comparison with REINFORCE

Next, we analyze the contribution of the optimal baseline for REINFORCE, and compare it with that for PGPE. It was shown [6, 20] that the excess variance for a baseline b in REINFORCE is given by

$$\begin{aligned}&\mathbf{Var}[\nabla_{\boldsymbol{\theta}}\widehat{J}^{b}(\boldsymbol{\theta})] - \mathbf{Var}[\nabla_{\boldsymbol{\theta}}\widehat{J}^{b^*_{\mathrm{REINFORCE}}}(\boldsymbol{\theta})]\\ &\quad= \frac{(b - b^*_{\mathrm{REINFORCE}})^2}{N}\mathbb{E}\left[\left\|\sum_{t=1}^{T}\nabla_{\boldsymbol{\theta}}\log p(a_t|\boldsymbol{s}_t,\boldsymbol{\theta})\right\|^2\right].\end{aligned}$$

Based on this, we have the following theorem.

Theorem 6. *Under Assumptions (B) and (C), we have the following bounds with probability at least* $1-\delta$*:*

$$\frac{C_T\alpha^2(1-\gamma^T)^2}{N\sigma^2(1-\gamma)^2} \leq \mathbf{Var}[\nabla_{\boldsymbol{\mu}}\widehat{J}(\boldsymbol{\theta})] - \mathbf{Var}[\nabla_{\boldsymbol{\mu}}\widehat{J}^{b^*_{\mathrm{REINFORCE}}}(\boldsymbol{\theta})] \leq \frac{\beta^2(1-\gamma^T)^2 D_T}{N\sigma^2(1-\gamma)^2}.$$

The above theorem shows that the lower and upper bounds of the excess variance are monotone increasing with respect to trajectory length T.

In the aspect of the amount of reduction in the variance of gradient estimates, Theorem 5 and Theorem 6 show that the optimal baseline for REINFORCE contributes more than that for PGPE.

Finally, based on Theorem 1 and Theorem 5 and based on Theorem 2 and Theorem 6, we have the following theorem:

Theorem 7. *Under Assumptions (B) and (C), we have*

$$\begin{aligned}\mathbf{Var}[\nabla_{\boldsymbol{\eta}}\widehat{J}^{b^*_{\mathrm{PGPE}}}(\boldsymbol{\rho})] &\leq \frac{(1-\gamma^T)^2}{N(1-\gamma)^2}(\beta^2-\alpha^2)B,\\ \mathbf{Var}[\nabla_{\boldsymbol{\mu}}\widehat{J}^{b^*_{\mathrm{REINFORCE}}}(\boldsymbol{\theta})] &\leq \frac{(1-\gamma^T)^2}{N\sigma^2(1-\gamma)^2}(\beta^2 D_T - \alpha^2 C_T),\end{aligned}$$

where the latter inequality holds with probability at least $1-\delta$*.*

This theorem shows that the upper bound of the variance of gradient estimates for REINFORCE with the optimal baseline is still monotone increasing with respect to trajectory length T. On the other hand, since $(1-\gamma^T)^2 \leq 1$, the above upper bound of the variance of gradient estimates in PGPE with the optimal baseline can be further upper-bounded as

$$\mathbf{Var}[\nabla_{\boldsymbol{\eta}} \widehat{J}^{b^*_{\mathrm{PGPE}}}(\boldsymbol{\rho})] \leq \frac{(\beta^2-\alpha^2)B}{N(1-\gamma)^2},$$

which is independent of T. Thus, when trajectory length T is large, the variance of gradient estimates in REINFORCE with the optimal baseline may be significantly larger than the variance of gradient estimates in PGPE with the optimal baseline.

5 Experiments

In this section, we experimentally investigate the usefulness of the proposed method, PGPE with the optimal baseline.

5.1 Illustration

Let the state space $\mathcal{S}$ be one-dimensional and continuous, and the initial state is randomly chosen from the standard normal distribution. The action space $\mathcal{A}$ is also set to be one-dimensional and continuous. The transition dynamics of the environment is set at

$$s_{t+1} = s_t + a_t + \varepsilon,$$

where $\varepsilon \sim \mathcal{N}(0, 0.5^2)$ is stochastic noise and $\mathcal{N}(\mu, \sigma^2)$ denotes the normal distribution with mean μ and variance σ^2. The immediate reward is defined as

$$r = \exp\left(-s^2/2 - a^2/2\right) + 1,$$

which is bounded as $1 < r \leq 2$.

5.1.1 Variance and Bias

First, we illustrate the variance of gradient estimates of the following methods:

- **REINFORCE:** REINFORCE without any baselines.
- **REINFORCE-OB:** REINFORCE with the optimal baseline.
- **PGPE:** PGPE without any baselines.
- **PGPE-MB:** PGPE with the moving-average baseline.
- **PGPE-OB:** PGPE with the optimal baseline.

Table 1: Variance and bias of estimated parameters for toy data.

Method	$T = 10$				$T = 50$			
	Variance		Bias		Variance		Bias	
	μ, η	σ, τ	μ, η	σ, τ	μ, η	σ, τ	μ, η	σ, τ
REINFORCE	13.2570	26.9173	-0.3102	-1.5098	188.3860	278.3095	-1.8126	-5.1747
REINFORCE-OB	0.0914	0.1203	0.0672	0.1286	0.5454	0.8996	-0.2988	-0.2008
PGPE	0.9707	1.6855	-0.0691	0.1319	1.6572	3.3720	-0.1048	-0.3293
PGPE-MB	0.2127	0.3238	0.0828	-0.1295	0.4123	0.8332	0.0925	-0.2556
PGPE-OB	0.0372	0.0685	-0.0164	0.0512	0.0850	0.1815	0.0480	-0.0779
PGPE-MB-SyS	0.1070	0.8087	0.0850	0.2625	0.2717	1.7883	0.1022	0.1124
PGPE-OB-SyS	0.0908	0.1084	-0.0854	0.0640	0.2865	0.3009	0.0460	0.1602

For fair comparison, all of these methods use the same parameter setup: the mean and standard deviation of the Gaussian distribution is set at $\mu = -1.5$ and $\sigma = 1$, and the length of the trajectory is set at $T = 10$ or 50. The discount factor is set at $\gamma = 0.9$, and the number of episodic samples is set at $N = 100$.

Table 1 summarizes the variance of gradient estimates over 100 runs, showing that the variance of REINFORCE is overall larger than PGPE. A notable difference between REINFORCE and PGPE is that the variance of REINFORCE significantly grows as T increases, whereas that of PGPE is not influenced that much by T. This well agrees with our theoretical analysis in Section 3. The results also show that the variance of PGPE-OB is much smaller than that of PGPE-MB. REINFORCE-OB contributes highly to reducing the variance especially when T is large, which also well agrees with our theory. However, PGPE-OB still provides much smaller variance than REINFORCE-OB.

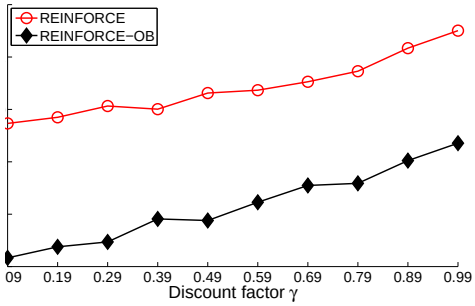
We also investigate the bias of gradient estimates of each method. Here, we regard gradients estimated with $N = 1000$ as true gradients, and compute the bias of gradient estimates. The results are also included in Table 1, showing that introduction of baselines does not increase the bias; rather, it tends to reduce the bias.

Figure 1 shows the variance of gradient estimates with respect to the mean parameter as functions of discounted factor γ, in $\log_{10}$-scale. The graphs show that, as discount factor γ gets close to 1, the variance increases. This well agrees with our theoretical analysis in Section 3. Among the compared methods, PGPE-OB has the smallest variance overall.

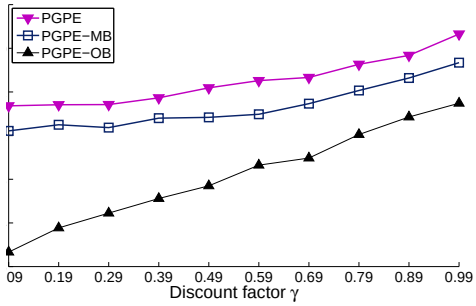
5.1.2 Symmetric Sampling for PGPE

In order to improve the convergence property of the PGPE method, a heuristic of using a pair of symmetric samples called the *symmetric sampling method* was introduced [16]. Here, we numerically investigate its effect on the variance of gradient estimates.

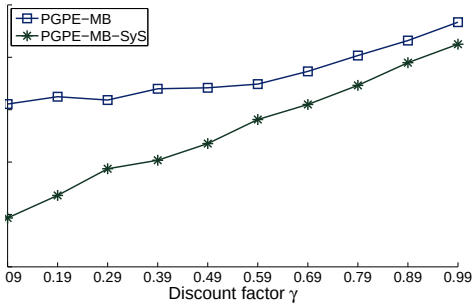
In the symmetric sampling method, perturbation sample $\boldsymbol{\epsilon}_n$ is drawn from distribution $\mathcal{N}(0, \boldsymbol{\tau}^2)$, and then symmetric parameter samples are created as $\boldsymbol{\theta}_n^+ = \boldsymbol{\eta} + \boldsymbol{\epsilon}_n$ and $\boldsymbol{\theta}_n^- = \boldsymbol{\eta} - \boldsymbol{\epsilon}_n$. Let R_n^+ and R_n^- be returns obtained by $\boldsymbol{\theta}_n^+$ and $\boldsymbol{\theta}_n^-$, respectively. Based on these two returns, gradients with respect to $\boldsymbol{\eta}$ are calculated using the difference between the



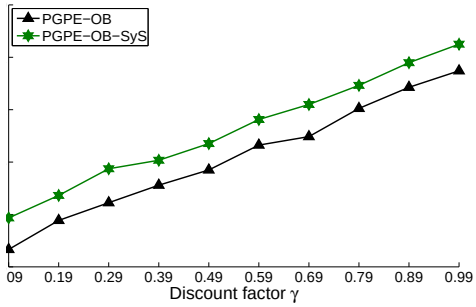
(a) REINFORCE and REINFORCE-OB



(b) PGPE, PGPE-MB, and PGPE-OB



(c) PGPE-MB and PGPE-MB-SyS



(d) PGPE-OB and PGPE-OB-SyS

Figure 1: Variance of gradient estimates with respect to the mean parameter as functions of discount factor γ for toy data.

two returns as

$$\nabla_{\boldsymbol{\eta}} \widehat{J}(\boldsymbol{\rho}) \approx \frac{1}{N} \sum_{n=1}^{N} \frac{\boldsymbol{\epsilon}_n (R_n^+ - R_n^-)}{2\boldsymbol{\eta}^2}.$$

On the other hand, gradients with respect to $\boldsymbol{\tau}$ can not be directly computed from symmetric parameter samples since $\boldsymbol{\theta}^+$ and $\boldsymbol{\theta}^-$ are equally probable under given $\boldsymbol{\tau}$. To cope with this problem, the difference of the mean of the two returns and the baseline is used as

$$\nabla_{\boldsymbol{\tau}} \widehat{J}(\boldsymbol{\rho}) \approx \frac{1}{N} \sum_{n=1}^{N} \nabla_{\boldsymbol{\tau}} \log p(\boldsymbol{\theta}^n | \boldsymbol{\rho}) \left(\frac{R_n^+ + R_n^-}{2} - b \right).$$

Note that, since the symmetric sampling method produces two parameters $\boldsymbol{\theta}^+$ and $\boldsymbol{\theta}^-$, it requires two trajectory samples in every update.

We numerically compare the variance of gradient estimates of the following methods:

- **PGPE-MB-SyS:** PGPE-MB with symmetric sampling.
- **PGPE-OB-SyS:** PGPE-OB with symmetric sampling.

In the previous experiments, the number of episodic samples for non-symmetric sampling methods was set at $N = 100$. If the number of sampled parameters is the same, the symmetric sampling methods will require twice as many trajectory samples (i.e., $N = 200$)

as non-symmetric sampling counterparts since the symmetric sampling methods produce two parameters θ^+ and θ^-. For fair comparison, we only use the half number of sampled parameters for the symmetric sampling methods, which requires $N = 100$ trajectory samples.

The bottom half of Table 1 shows the numerical results. In terms of the variance of gradient estimates with respect to mean parameter η, PGPE-MB-SyS has smaller variance than PGPE-MB. Thus, symmetric sampling contributes to reducing the variance for the PGPE-MB method, which agrees with the experimental results reported in [16]. However, PGPE-OB (without symmetric sampling) has smaller variance than PGPE-OB-SyS, indicating that symmetric sampling increases the variance for the PGPE-OB method. As for the variance of gradient estimates with respect to deviation parameter τ, symmetric sampling tends to increase the variance both for the PGPE-MB and PGPE-OB methods.

5.1.3 Variance and Policy Parameter Change through Entire Policy-Update Process

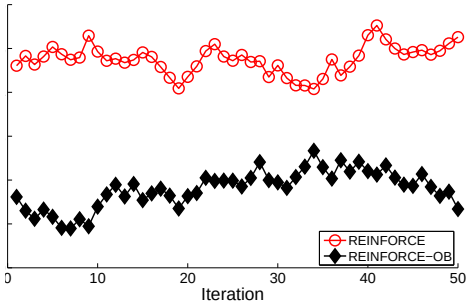
Next, we investigate the variance of gradient estimates when policy parameters are updated over iterations[2].

In this experiment, we set $T = 20$, and the variance is computed from 50 runs. We set $N = 10$ for all the methods, and policies are updated over 50 iterations. In order to evaluate the variance in a stable manner, we repeat the above experiments 20 times with random choice of initial mean parameter μ from $[-3.0, -0.1]$, and investigate the average variance of gradient estimates with respect to mean parameter μ over 20 trials.

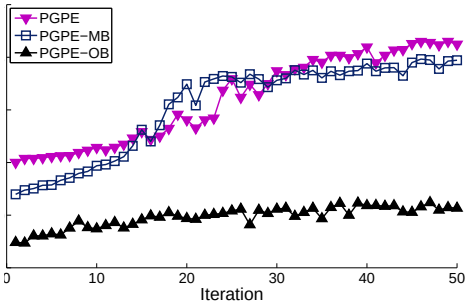
The results are summarized in Figure 2. Figure 2(a) compares the variance of REINFORCE with/without baselines, whereas Figure 2(b) compares the variance of PGPE with/without baselines. These plots show that introduction of baselines contributes highly to the reduction of the variance over iterations. Figure 2(c) compares the variance of PGPE-MB and PGPE-MB-SyS, showing that symmetric sampling contributes highly to stabilization. Figure 2(d) compares the variance of PGPE-OB and PGPE-OB-SyS, showing that the variance of PGPE-OB (without symmetric sampling) is smaller than that of PGPE-OB-SyS. Overall, in terms of the variance of gradient estimates, PGPE-OB compares favorably with other methods.

Next, we investigate how policy parameters change over 50 iterations. We set $N = 10$ and $T = 10$, and set the initial mean parameter at $\eta = -1.6$, -0.8, or -0.1, and initial deviation parameter at $\tau = 1$. Figure 3 depicts the contour of the expected return and illustrates changes of policy parameters over iterations for PGPE-MB and PGPE-OB. In the graphs, the maximum of the return surface is located at the middle bottom. Figure 3(a) shows that update directions of PGPE-MB are unstable and the three paths do not converge even after 50 iterations. On the other hand, Figure 3(b) shows that

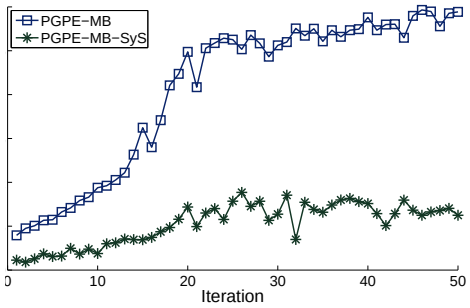
[2]If the deviation parameter σ takes a negative value during the policy-update process, we set it at 0.05.



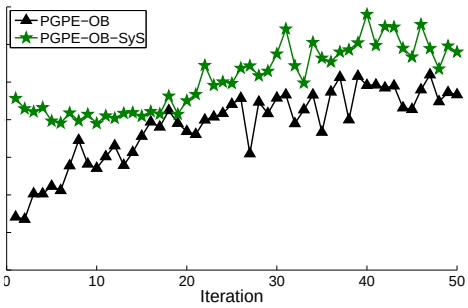
(a) REINFORCE and REINFORCE-OB



(b) PGPE, PGPE-MB, and PGPE-OB

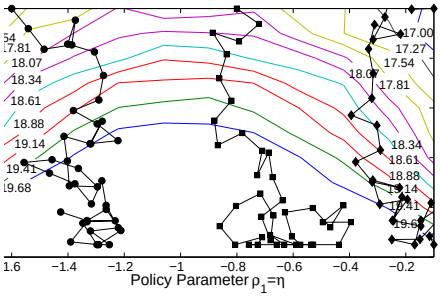


(c) PGPE-MB and PGPE-MB-SyS

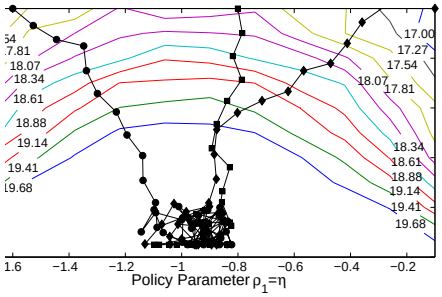


(d) PGPE-OB and PGPE-OB-SyS

Figure 2: Variance of gradient estimates with respect to the mean parameter through policy-update iterations for toy data.



(a) PGPE-MB



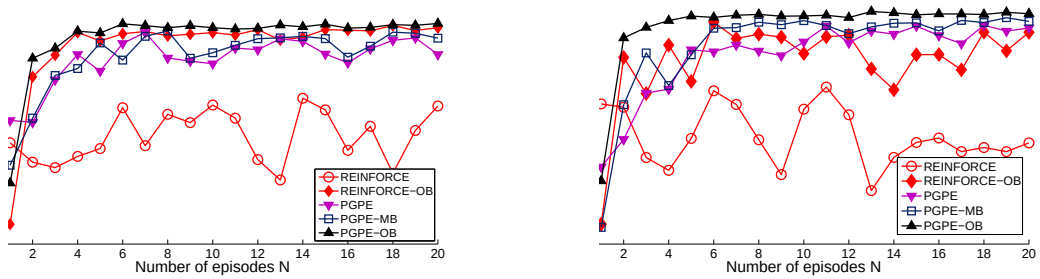
(b) PGPE-OB

Figure 3: Policy parameter change through policy-update iterations for toy data.

PGPE-OB gives much more reliable update directions and the three paths converge to a maximum point rapidly.

5.1.4 Performance of Learned Policies

Finally, we evaluate returns obtained by each method. The trajectory length is fixed at $T = 20$, and the maximum number of policy-update iterations is set at 50. We investigate average returns over 20 runs as functions of the number of episodic samples N. We have two experimental results for different initial policies. Figure 4(a) shows the



(a) Good initial policy (b) Poor initial policy

Figure 4: Average returns over 20 runs as functions of the number of episodic samples N for toy data.

results when initial mean parameter μ is chosen randomly from $[-1.6, -0.1]$, which tends to perform well. The graph shows that PGPE-OB performs the best, especially when $N < 5$; then REINFORCE-OB follows with a small margin. PGPE-MB and plain PGPE also work reasonably well, although they are slightly unstable due to larger variance. Plain REINFORCE is highly unstable, which is caused by the huge variance of gradient estimates (see Figure 2 again).

Figure 4(b) describes the results when initial mean parameter μ is chosen randomly from $[-3.0, -0.1]$, which tends to result in poorer performance. In this setup, difference among the compared methods is more significant than the case with good initial policies. Overall, plain REINFORCE performs very poorly, and even REINFORCE-OB tends to be outperformed by the PGPE methods. This means that REINFORCE is very sensitive to the choice of initial policies. Among the PGPE methods, PGPE-OB works very well and converges quickly.

5.2 Cart-Pole Balancing

Here, we evaluate the performance of our proposed method in a more complex task of *cart-pole balancing* [3]. A pole is hanged to the roof of a cart (see Figure 5), and the goal is to swing up the pole by moving the cart properly and try to keep the pole at the top.

The state space $\mathcal{S}$ is two-dimensional and continuous, which consists of the angle $\varphi \in [0, 2\pi]$ and angular velocity $\dot{\varphi} \in [-3\pi, 3\pi]$ of the pole. The action space $\mathcal{A}$ is one-dimensional and continuous, which corresponds to the force applied to the cart (note that we can *not* directly control the pole, but only indirectly through moving the cart). We use the Gaussian policy model for REINFORCE and linear policy model for PGPE, where state $\boldsymbol{s}$ is non-linearly transformed to a feature space via a basis function vector.

We use 20 Gaussian kernels with standard deviation $\sigma = 0.5$ as the basis functions, where the kernel centers are distributed over the following grid points:

$$\{0, \pi/2, \pi, 3\pi/2\} \times \{-3\pi, -3\pi/2, 0, 3\pi/2, 3\pi\}.$$

For the position of pole, we use the polar system where $\varphi = 0$ and $\varphi = 2\pi$ are treated as

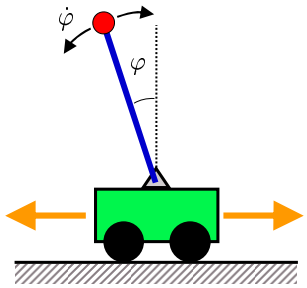


Figure 5: Cart-pole balancing.

the same. That is, for the i-th Gaussian center $(c_i, \dot{c}_i)$, the basis function $\phi_i(\boldsymbol{s})$ is given by

$$\phi_i(\boldsymbol{s}) = \exp\left(-\frac{((\cos(\varphi)-\cos(c_i))^2+(\sin(\varphi)-\sin(c_i))^2)/4+(\dot{\varphi}-\dot{c}_i)^2/(6\pi)^2}{2\sigma^2}\right).$$

The dynamics of the pole (i.e., the update rule of the angle and the angular velocity) is given by

$$\begin{aligned}\varphi_{t+1} &= \varphi_t + \dot{\varphi}_{t+1}\Delta t,\\ \dot{\varphi}_{t+1} &= \dot{\varphi}_t + \frac{9.8\sin(\varphi_t) - \alpha w l \dot{\varphi}_t^2 \sin(2\varphi_t)/2 + \alpha\cos(\varphi_t)a_t}{4l/3 - \alpha w l \cos^2(\varphi_t)}\Delta t,\end{aligned}$$

where $\alpha = 1/W + w$ and a_t is the action taken at time t. We set the problem parameters as: the mass of the cart $W = 8$[kg], the mass of the pole $w = 2$[kg], and the length of the pole $l = 0.5$[m]. We set the time step Δt for the position and velocity updates at 0.01[s] and action selection at 0.1[s]. The reward function is defined as

$$r(\boldsymbol{s}_t, a_t, \boldsymbol{s}_{t+1}) = \cos(\varphi_{t+1}).$$

That is, the higher the pole is, the more rewards we can obtain. The initial policy is chosen randomly, and the initial-state probability density is set to be uniform. The agent collects $N = 100$ episodic samples with trajectory length $T = 40$, and the discount factor is set at $\gamma = 0.9$.

We investigate average returns over 10 trials as the functions of policy-update iterations. The return at each trial is computed over 100 test episodic samples (which are not used for policy learning). The experimental results are plotted in Figure 6, showing that the improvement of both plain REINFORCE and REINFORCE-OB tend to be slow, and all PGPE methods outperformed REINFORCE methods overall. Among the PGPE methods, the proposed PGPE-OB converges faster than PGPE-MB and plain PGPE. Moreover, the use of symmetric sampling further improves the performance. Overall, PGPE equipped with both the optimal baseline and symmetric sampling (PGPE-OB-SyS) gives the best performance.

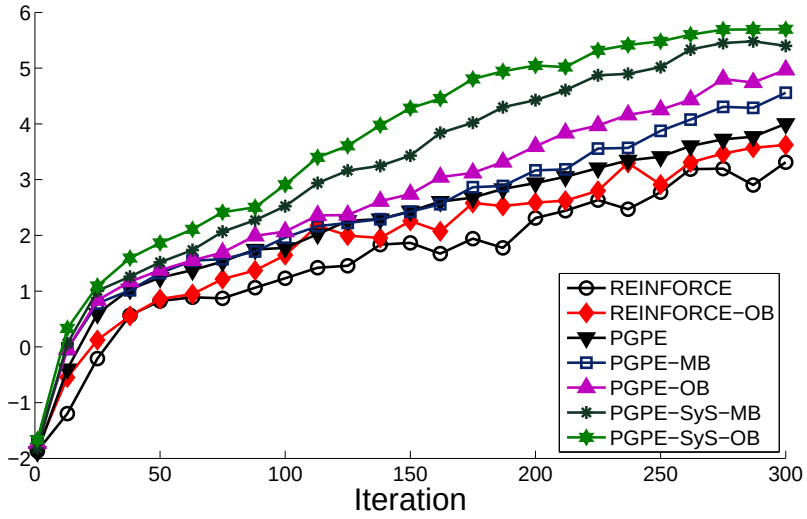


Figure 6: Agerage returns over 10 runs as function of the number of iterations.

6 Conclusions

In this paper, we analyzed and improved the stability of the policy gradient method called PGPE (policy gradients with parameter-based exploration). We theoretically showed that, under a mild condition, PGPE provides more stable gradient estimates than the classical REINFORCE method. We also derived the optimal baseline for PGPE, and theoretically showed that PGPE with the optimal baseline is more preferable than REINFORCE with the optimal baseline in terms of the variance of gradient estimates. Finally, we demonstrated the usefulness of PGPE with optimal baseline through experiments. We also experimentally showed that the use of symmetric sampling further improves the performance.

Although we focused on the gradient based approach to optimize the distribution of policy parameters, there are many alternative heuristic approaches such as the *genetic algorithm* (GA), *estimation of distribution algorithm* (EDA), and *hill climbing* (HC). GA is a heuristic approach inspired by mutation, selection, and crossover [5]. In GA, the population of randomly generated individuals are initially constructed. Then, in each iteration, multiple individuals are selected from the current population based on a fitness function, and new populations are formed by crossover between selected individuals with mutations. GAs could be applied to optimizing policy parameters by regarding individuals and the fitness function as policy parameters and the reward function respectively.

EDA is an outgrowth of GAs [11]. In EDAs, the probability distribution of populations is estimated from selected individuals and new populations are sampled from the distribution. EDAs would be more stable since the difficulty of designing crossover and mutation is diminished. Similarly to GAs, EDA could be applied to optimizing policy parameters. However, estimating a high-dimensional distribution of policy parameters is highly challenging.

HC is an optimization technique which belongs to the class of local search methods [9]. HC iteratively finds a better parameter by comparing the value of all neighbors of the current parameter. HC could be in principle applied to optimizing policy parameters, but it would not be computationally efficient to compare the return of all neighbored parameters.

These heuristic optimization techniques would be useful approaches to RL problem. Thus, an important future work along this line is to combine the meta-heuristics with the gradient-based method.

Another challenging issue to be discussed in the reinforcement learning field is the trade-off between exploration and exploitation. PGPE is not an exception since choosing similar policy parameters many times and collecting data is not efficient especially when data collection is expensive and time consuming. Therefore, in our future work, we will investigate the trade-off between exploration and exploitation in the framework of PGPE.

In real-world problems, often not all the state variables can be observed. In such cases, it is natural to consider *partially observable* settings. Thus, an important future direction is to formulate the PGPE problem in the framework of partially observable MDPs.

Recently, the combination of parameter-based exploration and natural policy gradient has been proposed to speed up the policy gradient methods [13]. We will extend the current theoretical analysis so that the above *natural PGPE* method can also be analyzed.

Acknowledgments

T. Z. and G. N. were supported by the MEXT scholarship and the GCOE program, H. H. was supported by the FIRST program, and M. S. was supported by MEXT KAKENHI 23120004.

Appendix

In the appendix, we give proofs of the theorems. First, we give some preliminaries.

If $X \sim \chi^2(k)$, then the non-central moments are given by

$$\mathbb{E}\left[X^n\right] = 2^n \frac{\Gamma(n+k/2)}{\Gamma(k/2)} = k(k+2)\cdots(k+2n-2),$$

where $\Gamma(z)$ is the Gamma function defined as

$$\Gamma(z) := \int_0^{+\infty} t^{z-1} e^{-t} dt.$$

The Gamma function satisfies $\Gamma(z+1) = z\Gamma(z)$, $\Gamma(1/2) = \sqrt{\pi}$, and $\Gamma(1) = 1$.

If $X \sim \mathcal{N}(\mu, \sigma^2)$, central absolute moments (the moments of $|X-\mu|$) are given by

$$\mathbb{E}\left[|x-\mu|^p\right] = \begin{cases} \sigma^p (p-1)!! \sqrt{2/\pi}, & p \text{ is odd}, \\ \sigma^p (p-1)!! & p \text{ is even}, \end{cases}$$

where $n!!$ denotes the double factorial defined by

$$n!! := \begin{cases} n \cdot (n-2) \cdots 5 \cdot 3 \cdot 1 & n \text{ is positive odd,} \\ n \cdot (n-2) \cdots 6 \cdot 4 \cdot 2 & n \text{ is positive even,} \\ 1 & n = 1 \text{ or } 0. \end{cases}$$

A Proof of Theorem 1

For notational brevity, we denote the i-th component of $\boldsymbol{f}(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\eta}} \log p(\boldsymbol{\theta} \mid \boldsymbol{\rho})$ and the i-th component of $\boldsymbol{g}(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\tau}} \log p(\boldsymbol{\theta} \mid \boldsymbol{\rho})$ as

$$f_i(\boldsymbol{\theta}) = \nabla_{\eta_i} \log p(\boldsymbol{\theta} \mid \boldsymbol{\rho}) = \frac{\theta_i - \eta_i}{\tau_i^2},$$
$$g_i(\boldsymbol{\theta}) = \nabla_{\tau_i} \log p(\boldsymbol{\theta} \mid \boldsymbol{\rho}) = \frac{(\theta_i - \eta_i)^2 - \tau_i^2}{\tau_i^3}.$$

Proof. According to Eq.(1), we have

$$\begin{aligned} \mathbf{Var}[R(h)\boldsymbol{f}(\boldsymbol{\theta})] &\leq \sum_{i=1}^{\ell} \mathbb{E}\left[(Rf_i)^2\right] \\ &= \sum_{i=1}^{\ell} \int p(\theta_i) \left(\sum_{t=1}^{T} \gamma^{t-1} r(\boldsymbol{s}_t, a_t, \boldsymbol{s}_{t+1})\right)^2 \left(\frac{\theta_i - \eta_i}{\tau_i^2}\right)^2 d\theta_i \\ &\leq \sum_{i=1}^{\ell} \int p(\theta_i) \left(\sum_{t=1}^{T} \gamma^{t-1} \beta\right)^2 \left(\frac{\theta_i - \eta_i}{\tau_i^2}\right)^2 d\theta_i \\ &= \sum_{i=1}^{\ell} \int p(\theta_i) \left(\frac{\beta(1-\gamma^T)}{1-\gamma}\right)^2 \left(\frac{\theta_i - \eta_i}{\tau_i^2}\right)^2 d\theta_i \\ &= \sum_{i=1}^{\ell} \frac{\beta^2(1-\gamma^T)^2}{\tau_i^2(1-\gamma)^2} \mathbb{E}\left[\left(\frac{\theta_i - \eta_i}{\tau_i}\right)^2\right]. \end{aligned}$$

Let $\psi_i = ((\theta_i - \eta_i)/\tau_i)^2$ for $i = 1, \ldots, \ell$. We could know that $\psi_i \sim \chi^2(1)$ and $\mathbb{E}[\psi_i] = 1$ since $\theta_i \sim \mathcal{N}(\eta_i, \tau_i^2)$, and thus

$$\mathbf{Var}[R(h)\boldsymbol{f}(\boldsymbol{\theta})] \leq \frac{\beta^2(1-\gamma^T)^2 B}{(1-\gamma)^2}.$$

Hence the first part of Theorem 1 follows due to

$$\mathbf{Var}\left[\nabla_{\boldsymbol{\eta}} \hat{\mathcal{J}}(\boldsymbol{\rho})\right] = \frac{1}{N} \mathbf{Var}[R(h)\boldsymbol{f}(\boldsymbol{\theta})].$$

Similarly,

$$\begin{aligned}\mathbf{Var}[R(h)\boldsymbol{g}(\boldsymbol{\theta})] &\leq \sum_{i=1}^{\ell} \mathbb{E}\left[(Rg_i)^2\right] \\ &\leq \sum_{i=1}^{\ell} \frac{\beta^2(1-\gamma^T)^2}{\tau_i^2(1-\gamma)^2} \mathbb{E}\left[\left(\left(\frac{\theta_i-\eta_i}{\tau_i}\right)^2 - 1\right)^2\right].\end{aligned}$$

Let $\psi_i = ((\theta_i - \eta_i)/\tau_i)^2$ for $i = 1, \ldots, \ell$. Since $\theta_i \sim \mathcal{N}(\eta_i, \tau_i^2)$, we could know that

$$\mathbb{E}\left[(\psi_i - 1)^2\right] = \mathbb{E}\left[\psi_i^2\right] - 2\mathbb{E}[\psi_i] + 1 = 2.$$

Hence

$$\mathbf{Var}[R(h)\boldsymbol{g}(\boldsymbol{\theta})] \leq \frac{2\beta^2(1-\gamma^T)^2 B}{(1-\gamma)^2}.$$

Notice that

$$\mathbf{Var}\left[\nabla_{\boldsymbol{\tau}} \hat{\mathcal{J}}(\boldsymbol{\rho})\right] = \frac{1}{N}\mathbf{Var}[R(h)\boldsymbol{g}(\boldsymbol{\theta})],$$

which completes the proof. □

B Proof of Theorem 2

To begin with, we note that $\boldsymbol{\mu}$ is a vector and σ is a scalar in REINFORCE. We denote the i-th component of $\boldsymbol{f}(h) = \sum_{t=1}^{T} \nabla_{\boldsymbol{\mu}} \log p(a_t \mid \boldsymbol{s}_t, \boldsymbol{\theta})$ and the scalar function $g(h)$ as

$$\begin{aligned} f_i(h) &= \sum_{t=1}^{T} \nabla_{\mu_i} \log p(a_t \mid \boldsymbol{s}_t, \boldsymbol{\theta}) = \sum_{t=1}^{T} \frac{a_t - \boldsymbol{\mu}^\top \boldsymbol{s}_t}{\sigma^2} s_{t,i}, \\ g(h) &= \sum_{t=1}^{T} \nabla_{\sigma} \log p(a_t \mid \boldsymbol{s}_t, \boldsymbol{\theta}) = \sum_{t=1}^{T} \frac{(a_t - \boldsymbol{\mu}^\top \boldsymbol{s}_t)^2 - \sigma^2}{\sigma^3}, \end{aligned}$$

where all functions above are parameterized by $\boldsymbol{\theta}$.

Proof. Since

$$\begin{aligned}\mathbf{Var}[\nabla_{\boldsymbol{\mu}} \hat{\mathcal{J}}(\boldsymbol{\theta})] &= \frac{1}{N}\mathbf{Var}[R(h)\boldsymbol{f}(h)], \\ \mathbf{Var}[\nabla_{\sigma} \hat{\mathcal{J}}(\boldsymbol{\theta})] &= \frac{1}{N}\mathbf{Var}[R(h)g(h)],\end{aligned}$$

we can just focus on the bounds of $\mathbf{Var}[R(h)\boldsymbol{f}(h)]$ and $\mathbf{Var}[R(h)g(h)]$.

The upper bound of $\mathbf{Var}[R(h)\boldsymbol{f}(h)]$:

$$\begin{aligned}
\mathbf{Var}[R(h)\boldsymbol{f}(h)] &\leq \sum_{i=1}^{\ell} \mathbb{E}\left[(Rf_i)^2\right] \\
&= \mathbb{E}\left[R^2 \boldsymbol{f}^\top \boldsymbol{f}\right] \\
&= \int_h p(h) \left(\sum_{t=1}^{T} \gamma^{t-1} r(\boldsymbol{s}_t, a_t, \boldsymbol{s}_{t+1})\right)^2 \\
&\quad \times \left(\sum_{t=1}^{T} \frac{a_t - \boldsymbol{\mu}^\top \boldsymbol{s}_t}{\sigma^2} \boldsymbol{s}_t\right)^\top \left(\sum_{t=1}^{T} \frac{a_t - \boldsymbol{\mu}^\top \boldsymbol{s}_t}{\sigma^2} \boldsymbol{s}_t\right) dh \\
&\leq \frac{\beta^2 (1-\gamma^T)^2}{\sigma^2 (1-\gamma)^2} \mathbb{E}\left[\left(\sum_{t,t'=1}^{T} \frac{(a_t - \boldsymbol{\mu}^\top \boldsymbol{s}_t)(a_{t'} - \boldsymbol{\mu}^\top \boldsymbol{s}_{t'})}{\sigma^2} \boldsymbol{s}_t^\top \boldsymbol{s}_{t'}\right)\right].
\end{aligned}$$

Let $\xi_t = (a_t - \boldsymbol{\mu}^\top \boldsymbol{s}_t)/\sigma$ for $t = 1, \ldots, T$. Then, $\xi_1, \ldots, \xi_T$ are independent standard normal variables because of $a_t \sim \mathcal{N}(\boldsymbol{\mu}^\top \boldsymbol{s}_t, \sigma^2)$. Since all $\nabla_{\boldsymbol{\mu}} \log p(a_t \mid \boldsymbol{s}_t, \boldsymbol{\theta})$ in $\boldsymbol{f}(h)$ are parameterized by the states $\boldsymbol{s}_t$, and the stochasticity of ξ_t comes only from a_t, it is sufficient to consider fixed states. Given $\{\boldsymbol{s}_t\}_{t=1}^T$, $\xi_1 \boldsymbol{s}_1, \ldots, \xi_T \boldsymbol{s}_T$ are ℓ-dimensional independent normal variables with zero means, that is, $\mathbb{E}[\xi_t \boldsymbol{s}_t] = \mathbf{0}$. Hence,

$$\begin{aligned}
\mathbb{E}\left[\left(\sum_{t,t'=1}^{T} \frac{(a_t - \boldsymbol{\mu}^\top \boldsymbol{s}_t)(a_{t'} - \boldsymbol{\mu}^\top \boldsymbol{s}_{t'})}{\sigma^2} \boldsymbol{s}_t^\top \boldsymbol{s}_{t'}\right)\right] &= \mathbb{E}\left[\left(\sum_{t,t'=1}^{T} \xi_t \xi_{t'} \boldsymbol{s}_t^\top \boldsymbol{s}_{t'}\right)\right] \\
&= \sum_{t=1}^{T} \mathbb{E}\left[\xi_t^2 \boldsymbol{s}_t^\top \boldsymbol{s}_t\right] + \sum_{t,t'=1, t\neq t'}^{T} \mathbb{E}[\xi_t \boldsymbol{s}_t]^\top \mathbb{E}[\xi_{t'} \boldsymbol{s}_{t'}] \\
&= \sum_{t=1}^{T} \|\boldsymbol{s}_t\|^2 \mathbb{E}\left[\xi_t^2\right].
\end{aligned}$$

Since $\xi_t \sim \mathcal{N}(0,1)$, we have $\xi_t^2 \sim \chi^2(1)$ and $\mathbb{E}[\xi_t^2] = 1$. Consequently,

$$\begin{aligned}
\mathbf{Var}[R(h)\boldsymbol{f}(h)] &\leq \frac{\beta^2 (1-\gamma^T)^2}{\sigma^2 (1-\gamma)^2} \sum_{t=1}^{T} \|\boldsymbol{s}_t\|^2 \mathbb{E}\left[\xi_t^2\right] \\
&= \frac{\beta^2 (1-\gamma^T)^2}{\sigma^2 (1-\gamma)^2} \sum_{t=1}^{T} \|\boldsymbol{s}_t\|^2 \\
&\leq \frac{D_T \beta^2 (1-\gamma^T)^2}{\sigma^2 (1-\gamma)^2},
\end{aligned}$$

with probability at least $(1-\delta)^{1/2N}$.

The upper bound of $\mathbf{Var}[R(h)g(h)]$:

$$\begin{aligned}\mathbf{Var}[R(h)g(h)] &\le \mathbb{E}\left[(Rg)^2\right] \\ &= \int_h p(h)\left(\sum_{t=1}^{T}\gamma^{t-1}r(\boldsymbol{s}_t,a_t,\boldsymbol{s}_{t+1})\right)^2\left(\sum_{t=1}^{T}\frac{(a_t-\boldsymbol{\mu}^\top\boldsymbol{s}_t)^2-\sigma^2}{\sigma^3}\right)^2 dh \\ &\le \frac{\beta^2(1-\gamma^T)^2}{\sigma^2(1-\gamma)^2}\mathbb{E}\left[\left(\sum_{t=1}^{T}\left(\frac{a_t-\boldsymbol{\mu}^\top\boldsymbol{s}_t}{\sigma}\right)^2-T\right)^2\right].\end{aligned}$$

Let $\xi_t=(a_t-\boldsymbol{\mu}^\top\boldsymbol{s}_t)/\sigma$ for $t=1,\ldots,T$. Then $\xi_1,\ldots,\xi_T$ are independent standard normal variables. Let $\kappa=\sum_{t=1}^{T}\xi_t^2$. Then we have $\kappa\sim\chi^2(T)$ and

$$\mathbb{E}\left[(\kappa-T)^2\right]=\mathbb{E}\left[\kappa^2\right]-2T\mathbb{E}[\kappa]+T^2=2T.$$

Hence

$$\mathbf{Var}[R(h)g(h)]\le\frac{2T\beta^2(1-\gamma^T)^2}{\sigma^2(1-\gamma)^2}.$$

The lower bound of $\mathbf{Var}[R(h)f(h)]$: By the same technique used in the corresponding upper bound, we can prove that with probability at least $(1-\delta)^{1/2N}$,

$$\sum_{i=1}^{\ell}\mathbb{E}\left[(Rf_i)^2\right]\ge\frac{C_T\alpha^2(1-\gamma^T)^2}{\sigma^2(1-\gamma)^2}.$$

On the other hand, based on the existence of $\{d_t\}_{t=1}^T$, there must be $\{d_{t,i}\}_{t=1}^T$ for $i=1,\ldots,\ell$, such that $d_t^2=\sum_{i=1}^{\ell}d_{t,i}^2$ and the inequality $|s_{t,i}|\le d_{t,i}$ holds with probability at least $(1-\delta)^{1/2N\ell}$. Let $\xi_{t,i}=\mathrm{sgn}(s_{t,i})(a_t-\boldsymbol{\mu}^\top\boldsymbol{s}_t)/\sigma$ for $t=1,\ldots,T$ and $i=1,\ldots,\ell$. Then all $\xi_{t,i}$ are independent standard normal variables. Let $\kappa_i=\sum_{t=1}^{T}\xi_{t,i}|s_{t,i}|$ and $\zeta_i=\sum_{t=1}^{T}\xi_{t,i}d_{t,i}$. Then $\kappa_i\sim\mathcal{N}(0,\sum_{t=1}^{T}s_{t,i}^2)$ for fixed $s_{1,i},\ldots,s_{T,i}$, $\zeta_i\sim\mathcal{N}(0,\sum_{t=1}^{T}d_{t,i}^2)$, and $\mathbb{E}[|\kappa_i|\mid s_{1,i},\ldots,s_{T,i}]\le\mathbb{E}[|\zeta_i|]$ holds with probability at least $(1-\delta)^{1/2N\ell}$ over the choice of $s_{1,i},\ldots,s_{T,i}$ according to the underlying $p(h)$. When $\int_h p(h)Rf_i dh>0$, with

probability at least $(1-\delta)^{1/2N\ell}$,

$$\begin{aligned}
\int_h p(h)Rf_i dh &\leq \int_{\{h|f_i(h)>0\}} p(h)Rf_i dh \\
&\leq \frac{\beta(1-\gamma^T)}{1-\gamma} \int_{\{h|f_i(h)>0\}} p(h) f_i dh \\
&= \frac{\beta(1-\gamma^T)}{1-\gamma} \int_{\{h|\sum_{t=1}^T \xi_{t,i}|s_{t,i}|>0\}} p(h) \sum_{t=1}^T \xi_{t,i}|s_{t,i}|dh \\
&= \frac{\beta(1-\gamma^T)}{1-\gamma} \int_0^{+\infty} p(\kappa_i)\kappa_i d\kappa_i \\
&= \frac{\beta(1-\gamma^T)}{1-\gamma} \left(\frac{1}{2}\mathbb{E}[|\kappa_i|]\right) \\
&= \frac{\beta(1-\gamma^T)}{1-\gamma} \left(\frac{1}{2}\mathbb{E}_{s_{1,i},\ldots,s_{T,i}}\Big[\mathbb{E}_{\kappa_i}[|\kappa_i| \mid s_{1,i},\ldots,s_{T,i}]\Big]\right) \\
&\leq \frac{\beta(1-\gamma^T)}{1-\gamma} \left(\frac{1}{2}\mathbb{E}[|\zeta_i|]\right) \\
&= \frac{\beta(1-\gamma^T)}{1-\gamma} \frac{\sqrt{\sum_{t=1}^T d_{t,i}^2}}{\sqrt{2\pi}}.
\end{aligned}$$

When $\int_h p(h)Rf_i dh < 0$, with probability at least $(1-\delta)^{1/2N\ell}$,

$$\int_h p(h)Rf_i dh \geq -\frac{\beta(1-\gamma^T)}{1-\gamma} \frac{\sqrt{\sum_{t=1}^T d_{t,i}^2}}{\sqrt{2\pi}}.$$

Therefore,

$$\begin{aligned}
\sum_{i=1}^{\ell} (\mathbb{E}[Rf_i])^2 &= \sum_{i=1}^{\ell} \left(\int_h p(h)Rf_i dh\right)^2 \\
&\leq \sum_{i=1}^{\ell} \frac{\beta^2(1-\gamma^T)^2}{\sigma^2(1-\gamma)^2} \frac{\sum_{t=1}^T d_{t,i}^2}{2\pi} \\
&= \frac{\beta^2(1-\gamma^T)^2}{2\pi\sigma^2(1-\gamma)^2} \sum_{t=1}^T \sum_{i=1}^{\ell} d_{t,i}^2 \\
&= \frac{\beta^2(1-\gamma^T)^2}{2\pi\sigma^2(1-\gamma)^2} \sum_{t=1}^T d_t^2 \\
&= \frac{D_T\beta^2(1-\gamma^T)^2}{2\pi\sigma^2(1-\gamma)^2},
\end{aligned}$$

with probability at least $(1-\delta)^{1/2N}$.

Finally, with probability at least $(1-\delta)^{1/N}$, we have

$$\begin{aligned}\mathbf{Var}[R(h)\boldsymbol{f}(h)] &= \sum_{i=1}^{\ell} \mathbb{E}\left[(Rf_i)^2\right] - (\mathbb{E}[Rf_i])^2 \\ &\geq \frac{(1-\gamma^T)^2}{\sigma^2(1-\gamma)^2}\mathcal{L}(T).\end{aligned}$$

□

C Proof of Theorem 3

Proof. According to Theorem 1 and Theorem 2, we could know that if there exists T_0 such that

$$\frac{(1-\gamma^T)^2}{N\sigma^2(1-\gamma)^2}\mathcal{L}(T_0) \geq \frac{\beta^2(1-\gamma^T)^2 B}{N(1-\gamma)^2},$$

we could get

$$\mathcal{L}(T_0) \geq \beta^2 B\sigma^2.$$

Under our assumption that $\mathcal{L}(T) > 0$ and $\mathcal{L}(T)$ is monotonically increasing with respect to T, we will have that whenever

$$\exists T_0, \mathcal{L}(T_0) \geq \beta^2 B\sigma^2,$$

there must be

$$\forall T > T_0, \mathbf{Var}[\nabla_{\boldsymbol{\mu}}\widehat{J}(\boldsymbol{\theta})] > \mathbf{Var}[\nabla_{\boldsymbol{\eta}}\widehat{J}(\boldsymbol{\rho})].$$

□

D Proof of Theorem 4

We denote $\boldsymbol{f}(\boldsymbol{\theta})$ and its i-th component $\boldsymbol{f}_i(\boldsymbol{\theta})$ as

$$\begin{aligned}\boldsymbol{f}(\boldsymbol{\theta}) &= \left(\nabla_{\boldsymbol{\eta}}\log p(\boldsymbol{\theta} \mid \boldsymbol{\rho})^\top, \nabla_{\boldsymbol{\tau}}\log p(\boldsymbol{\theta} \mid \boldsymbol{\rho})^\top\right)^\top = \nabla_{\boldsymbol{\rho}}\log p(\boldsymbol{\theta} \mid \boldsymbol{\rho}), \\ \boldsymbol{f}_i(\boldsymbol{\theta}) &= (\nabla_{\eta_i}\log p(\boldsymbol{\theta} \mid \boldsymbol{\rho}), \nabla_{\tau_i}\log p(\boldsymbol{\theta} \mid \boldsymbol{\rho}))^\top = \nabla_{\boldsymbol{\rho}_i}\log p(\boldsymbol{\theta} \mid \boldsymbol{\rho}).\end{aligned}$$

Note that we still have

$$\begin{aligned}\mathbf{Var}\left[\nabla_{\boldsymbol{\rho}}\widehat{J}^b(\boldsymbol{\rho})\right] &= \mathbf{Var}\left[\nabla_{\boldsymbol{\eta}}\widehat{J}^b(\boldsymbol{\rho})\right] + \mathbf{Var}\left[\nabla_{\boldsymbol{\tau}}\widehat{J}^b(\boldsymbol{\rho})\right] \\ &= \frac{1}{N}\mathbf{Var}[(R(h)-b)\nabla_{\boldsymbol{\eta}}\log p(\boldsymbol{\theta} \mid \boldsymbol{\rho})] \\ &\quad + \frac{1}{N}\mathbf{Var}[(R(h)-b)\nabla_{\boldsymbol{\tau}}\log p(\boldsymbol{\theta} \mid \boldsymbol{\rho})] \\ &= \frac{1}{N}\mathbf{Var}[(R(h)-b)\boldsymbol{f}(\boldsymbol{\theta})].\end{aligned}$$

Proof. According to Eq.(1), we have

$$\begin{aligned}\mathbf{Var}[(R(h)-b)\boldsymbol{f}(\boldsymbol{\theta})] &= \sum_{i=1}^{\ell}\mathbb{E}[(R-b)^2\boldsymbol{f}_i^\top\boldsymbol{f}_i] - (\mathbb{E}[(R-b)\boldsymbol{f}_i])^\top(\mathbb{E}[(R-b)\boldsymbol{f}_i]) \\ &= \sum_{i=1}^{\ell}\mathbb{E}[R^2\boldsymbol{f}_i^\top\boldsymbol{f}_i] - 2\mathbb{E}[Rb\boldsymbol{f}_i^\top\boldsymbol{f}_i] + \mathbb{E}[b^2\boldsymbol{f}_i^\top\boldsymbol{f}_i] \\ &\quad - (\mathbb{E}[R\boldsymbol{f}_i] - \mathbb{E}[b\boldsymbol{f}_i])^\top(\mathbb{E}[R\boldsymbol{f}_i] - \mathbb{E}[b\boldsymbol{f}_i]).\end{aligned}$$

Noticing that

$$\begin{aligned}\mathbb{E}[b\boldsymbol{f}_i] &= \int p(\theta_i \mid \boldsymbol{\rho}_i) b \nabla_{\boldsymbol{\rho}_i} \log p(\theta_i \mid \boldsymbol{\rho}_i) d\theta_i \\ &= \int b \nabla_{\boldsymbol{\rho}_i} p(\theta_i \mid \boldsymbol{\rho}_i) d\theta_i \\ &= b \nabla_{\boldsymbol{\rho}_i} \int p(\theta_i \mid \boldsymbol{\rho}_i) d\theta_i \\ &= b \nabla_{\boldsymbol{\rho}_i} 1 \\ &= b(\nabla_{\eta_i} 1, \nabla_{\tau_i} 1)^\top \\ &= (0,0)^\top,\end{aligned}$$

we have

$$\mathbf{Var}[(R(h)-b)\boldsymbol{f}(\boldsymbol{\theta})] = \mathbb{E}[R^2\boldsymbol{f}^\top\boldsymbol{f}] - 2\mathbb{E}[Rb\boldsymbol{f}^\top\boldsymbol{f}] + \mathbb{E}[b^2\boldsymbol{f}^\top\boldsymbol{f}] - \mathbb{E}[R\boldsymbol{f}]^\top\mathbb{E}[R\boldsymbol{f}]. \quad (3)$$

The optimal baseline is obtained by minimizing the variance, so that differentiating it with respect to b and setting the result to zero will give us the optimal baseline for PGPE:

$$b^*_{\mathrm{PGPE}} = \frac{\mathbb{E}[R\boldsymbol{f}^\top\boldsymbol{f}]}{\mathbb{E}[\boldsymbol{f}^\top\boldsymbol{f}]}.$$

Subsequently,

$$\begin{aligned}&\mathbf{Var}[(R-b)\boldsymbol{f}] - \mathbf{Var}[(R-b^*_{\mathrm{PGPE}})\boldsymbol{f}] \\ &\quad= -2\mathbb{E}[Rb\boldsymbol{f}^\top\boldsymbol{f}] + \mathbb{E}[b^2\boldsymbol{f}^\top\boldsymbol{f}] + 2\mathbb{E}[Rb^*_{\mathrm{PGPE}}\boldsymbol{f}^\top\boldsymbol{f}] - \mathbb{E}[b^{*2}_{\mathrm{PGPE}}\boldsymbol{f}^\top\boldsymbol{f}] \\ &\quad= -2\mathbb{E}[Rb\boldsymbol{f}^\top\boldsymbol{f}] + \mathbb{E}[b^2\boldsymbol{f}^\top\boldsymbol{f}] + 2\frac{\mathbb{E}[R\boldsymbol{f}^\top\boldsymbol{f}]}{\mathbb{E}[\boldsymbol{f}^\top\boldsymbol{f}]}\mathbb{E}[R\boldsymbol{f}^\top\boldsymbol{f}] - \left(\frac{\mathbb{E}[R\boldsymbol{f}^\top\boldsymbol{f}]}{\mathbb{E}[\boldsymbol{f}^\top\boldsymbol{f}]}\right)^2\mathbb{E}[\boldsymbol{f}^\top\boldsymbol{f}] \\ &\quad= b^2\mathbb{E}[\boldsymbol{f}^\top\boldsymbol{f}] - 2b\mathbb{E}[R\boldsymbol{f}^\top\boldsymbol{f}] + \frac{(\mathbb{E}[R\boldsymbol{f}^\top\boldsymbol{f}])^2}{\mathbb{E}[\boldsymbol{f}^\top\boldsymbol{f}]} \\ &\quad= \left(b - \frac{\mathbb{E}[R\boldsymbol{f}^\top\boldsymbol{f}]}{\mathbb{E}[\boldsymbol{f}^\top\boldsymbol{f}]}\right)^2\mathbb{E}[\boldsymbol{f}^\top\boldsymbol{f}] \\ &\quad= \left(b - b^*_{\mathrm{PGPE}}\right)^2\mathbb{E}[\boldsymbol{f}^\top\boldsymbol{f}],\end{aligned}$$

which leads to

$$\begin{aligned}\mathbf{Var}[\nabla_{\boldsymbol{\rho}}\widehat{J}^{b}(\boldsymbol{\rho})] - \mathbf{Var}[\nabla_{\boldsymbol{\rho}}\widehat{J}^{b^*_{\mathrm{PGPE}}}(\boldsymbol{\rho})] &= \frac{1}{N}\mathbf{Var}[(R-b)\boldsymbol{f}] - \frac{1}{N}\mathbf{Var}[(R-b^*_{\mathrm{PGPE}})\boldsymbol{f}] \\ &= \frac{(b-b^*_{\mathrm{PGPE}})^2}{N}\mathbb{E}[\boldsymbol{f}^\top\boldsymbol{f}].\end{aligned}$$

□

E Proof of Theorem 5

We denote the i-th component of $\boldsymbol{f}(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\eta}} \log p(\boldsymbol{\theta} \mid \boldsymbol{\rho})$ as

$$f_i(\boldsymbol{\theta}) = \nabla_{\eta_i} \log p(\boldsymbol{\theta} \mid \boldsymbol{\rho}) = \frac{\theta_i - \eta_i}{\tau_i^2}.$$

Proof. By the same technique used in the proof of Theorem 4, we know, when the baseline $b = 0$,

$$\mathbf{Var}[\nabla_{\boldsymbol{\eta}}\widehat{J}(\boldsymbol{\rho})] - \mathbf{Var}[\nabla_{\boldsymbol{\eta}}\widehat{J}^{b^*_{\mathrm{PGPE}}}(\boldsymbol{\rho})] = \frac{\left(\mathbb{E}[R\boldsymbol{f}^\top\boldsymbol{f}]\right)^2}{N\mathbb{E}[\boldsymbol{f}^\top\boldsymbol{f}]}.$$

On one hand,

$$\begin{aligned}\mathbb{E}[\boldsymbol{f}^\top\boldsymbol{f}] &= \sum_{i=1}^{\ell}\mathbb{E}[f_i^2] \\ &= \sum_{i=1}^{\ell}\mathbb{E}\left[\left(\frac{\theta_i-\eta_i}{\tau_i^2}\right)^2\right] \\ &= \sum_{i=1}^{\ell}\frac{1}{\tau_i^2}\mathbb{E}\left[\left(\frac{\theta_i-\eta_i}{\tau_i}\right)^2\right].\end{aligned}$$

Let $\psi_i = ((\theta_i - \eta_i)/\tau_i)^2$ for $i = 1, \ldots, \ell$. We could know that $\psi_i \sim \chi^2(1)$ and $\mathbb{E}[\psi_i] = 1$ since $\theta_i \sim \mathcal{N}(\eta_i, \tau_i^2)$, and thus

$$\mathbb{E}[\boldsymbol{f}^\top\boldsymbol{f}] = \sum_{i=1}^{\ell}\frac{1}{\tau_i^2} = B.$$

On the other hand, when $\mathbb{E}[R\boldsymbol{f}^\top\boldsymbol{f}] > 0$, we have

$$\begin{aligned}\mathbb{E}[R\boldsymbol{f}^\top\boldsymbol{f}] &= \sum_{i=1}^{\ell}\int p(\theta_i)R\left(\frac{\theta_i-\eta_i}{\tau_i^2}\right)^2 d\theta_i \\ &\leq \sum_{i=1}^{\ell}\frac{\beta(1-\gamma^T)}{\tau_i^2(1-\gamma)}\int p(\theta_i)\left(\frac{\theta_i-\eta_i}{\tau_i}\right)^2 d\theta_i \\ &= \sum_{i=1}^{\ell}\frac{\beta(1-\gamma^T)}{\tau_i^2(1-\gamma)}\mathbb{E}[\psi_i] \\ &= \frac{\beta(1-\gamma^T)B}{(1-\gamma)},\end{aligned}$$

while $\mathbb{E}[R\boldsymbol{f}^\top\boldsymbol{f}] < 0$, we have

$$\mathbb{E}[R\boldsymbol{f}^\top\boldsymbol{f}] \geq -\frac{\beta(1-\gamma^T)B}{(1-\gamma)}.$$

Hence,

$$\frac{\left(\mathbb{E}[R\boldsymbol{f}^\top\boldsymbol{f}]\right)^2}{\mathbb{E}[\boldsymbol{f}^\top\boldsymbol{f}]} \leq \frac{\beta^2(1-\gamma^T)^2B}{(1-\gamma)^2}.$$

Similarly,

$$\frac{\left(\mathbb{E}[R\boldsymbol{f}^\top\boldsymbol{f}]\right)^2}{\mathbb{E}[\boldsymbol{f}^\top\boldsymbol{f}]} \geq \frac{\alpha^2(1-\gamma^T)^2B}{(1-\gamma)^2},$$

which completes the proof. □

F Proof of Theorem 6

We denote $\boldsymbol{f}(h) = \sum_{t=1}^T \nabla_{\boldsymbol{\mu}} \log p(a_t \mid \boldsymbol{s}_t, \boldsymbol{\theta})$.

Proof. It is easy to prove that, when $b = 0$,

$$\mathbf{Var}[\nabla_{\boldsymbol{\mu}}\widehat{J}(\boldsymbol{\theta})] - \mathbf{Var}[\nabla_{\boldsymbol{\mu}}\widehat{J}^{b^*_{\text{REINFORCE}}}(\boldsymbol{\theta})] = \frac{(\mathbb{E}[R\boldsymbol{f}^\top\boldsymbol{f}])^2}{N\mathbb{E}[\boldsymbol{f}^\top\boldsymbol{f}]}.$$

From the proof of Theorem 2, we could have

$$\mathbb{E}[\boldsymbol{f}^\top\boldsymbol{f}] = \frac{1}{\sigma^2}\sum_{t=1}^T \|\boldsymbol{s}_t\|^2.$$

On the other hand,

$$\begin{aligned}
\mathbb{E}[R\boldsymbol{f}^\top\boldsymbol{f}] &= \int_h p(h)\left(\sum_{t=1}^T \gamma^{t-1} r(\boldsymbol{s}_t, a_t, \boldsymbol{s}_{t+1})\right)\left(\sum_{t=1}^T \frac{a_t - \boldsymbol{\mu}^\top\boldsymbol{s}_t}{\sigma^2}\boldsymbol{s}_t\right)^\top\left(\sum_{t=1}^T \frac{a_t - \boldsymbol{\mu}^\top\boldsymbol{s}_t}{\sigma^2}\boldsymbol{s}_t\right) dh \\
&\leq \frac{\beta(1-\gamma^T)}{\sigma^2(1-\gamma)}\mathbb{E}\left[\left(\sum_{t,t'=1}^T \frac{(a_t - \boldsymbol{\mu}^\top\boldsymbol{s}_t)(a_{t'} - \boldsymbol{\mu}^\top\boldsymbol{s}_{t'})}{\sigma^2}\boldsymbol{s}_t^\top\boldsymbol{s}_{t'}\right)\right] \\
&= \frac{\beta(1-\gamma^T)}{\sigma^2(1-\gamma)}\sum_{t=1}^T \|\boldsymbol{s}_t\|^2.
\end{aligned}$$

Similarly,

$$\mathbb{E}[R\boldsymbol{f}^\top\boldsymbol{f}] \geq \frac{\alpha(1-\gamma^T)}{\sigma^2(1-\gamma)}\sum_{t=1}^T \|\boldsymbol{s}_t\|^2.$$

Therefore,

$$\frac{\alpha^2(1-\gamma^T)^2\sum_{t=1}^T\|\boldsymbol{s}_t\|^2}{\sigma^2(1-\gamma)^2}\leq\frac{(\mathbb{E}[R\boldsymbol{f}^\top\boldsymbol{f}])^2}{\mathbb{E}[\boldsymbol{f}^\top\boldsymbol{f}]}\leq\frac{\beta^2(1-\gamma^T)^2\sum_{t=1}^T\|\boldsymbol{s}_t\|^2}{\sigma^2(1-\gamma)^2},$$

and subsequently, with probability at least $(1-\delta)^{1/N}$, we have

$$\frac{C_T\alpha^2(1-\gamma^T)^2}{\sigma^2(1-\gamma)^2}\leq\frac{(\mathbb{E}[R\boldsymbol{f}^\top\boldsymbol{f}])^2}{\mathbb{E}[\boldsymbol{f}^\top\boldsymbol{f}]}\leq\frac{\beta^2(1-\gamma^T)^2D_T}{\sigma^2(1-\gamma)^2}.$$

From this, the theorem follows. □

G Proof of Theorem 7

Proof. According to Theorem 5, we know

$$\mathbf{Var}[\nabla_{\boldsymbol{\eta}}\widehat{J}^{b^*_{\text{PGPE}}}(\boldsymbol{\rho})]\leq\mathbf{Var}\left[\nabla_{\boldsymbol{\eta}}\widehat{J}(\boldsymbol{\rho})\right]-\frac{\alpha^2(1-\gamma^T)^2B}{N(1-\gamma)^2}.$$

According to Theorem 1, we have

$$\mathbf{Var}\left[\nabla_{\boldsymbol{\eta}}\widehat{J}(\boldsymbol{\rho})\right]\leq\frac{\beta^2(1-\gamma^T)^2B}{N(1-\gamma)^2}.$$

Hence,

$$\mathbf{Var}[\nabla_{\boldsymbol{\eta}}\widehat{J}^{b^*_{\text{PGPE}}}(\boldsymbol{\rho})]\leq\frac{(1-\gamma^T)^2}{N(1-\gamma)^2}(\beta^2-\alpha^2)B.$$

According to Theorem 6, we know that

$$\mathbf{Var}[\nabla_{\boldsymbol{\mu}}\widehat{J}^{b^*_{\text{REINFORCE}}}(\boldsymbol{\theta})]\leq\mathbf{Var}\left[\nabla_{\boldsymbol{\mu}}\widehat{J}(\boldsymbol{\theta})\right]-\frac{C_T\alpha^2(1-\gamma^T)^2}{N\sigma^2(1-\gamma)^2}$$

will holds with probability at leat $(1-\delta)^{1/2}$. Furthermore, according to Theorem 2, we have the following upper bound with probability at least $(1-\delta)^{1/2}$:

$$\mathbf{Var}\left[\nabla_{\boldsymbol{\mu}}\widehat{J}(\boldsymbol{\theta})\right]\leq\frac{D_T\beta^2(1-\gamma^T)^2}{N\sigma^2(1-\gamma)^2}.$$

Eventually, we arrive at the upper bound for REINFORCE with the optimal baseline:

$$\mathbf{Var}[\nabla_{\boldsymbol{\mu}}\widehat{J}^{b^*_{\text{REINFORCE}}}(\boldsymbol{\theta})]\leq\frac{(1-\gamma^T)^2}{N\sigma^2(1-\gamma)^2}(D_T\beta^2-C_T\alpha^2),$$

with probability at least $1-\delta$. □

References

[1] N. Abe, P. Melville, C. Pendus, C. K. Reddy, D. L. Jensen, V. P. Thomas, J. J. Bennett, G. F. Anderson, B. R. Cooley, M. Kowalczyk, M. Domick, and T. Gardinier. Optimizing debt collections using constrained reinforcement learning. In *Proceedings of the 16th ACM SGKDD Conference on Knowledge Discovery and Data Mining*, pages 75–84, 2010.

[2] J. Baxter, P. Bartlett, and L. Weaver. Experiments with infinite-horizon, policy-gradient estimation. *Journal of Artificial Intelligence Research*, 15:351–381, 2001.

[3] M. Bugeja. Non-linear swing-up and stabilizing control of an inverted pendulum system. In *Proceedings of IEEE Region 8 EUROCON*, volume 2, pages 437–441, 2003.

[4] P. Dayan and G. E. Hinton. Using expectation-maximization for reinforcement learning. *Neural Computation*, 9(2):271–278, 1997.

[5] D. E. Goldberg. *Genetic Algorithms in Search, Optimization and Machine Learning*. Addison Wesley, 1989.

[6] E. Greensmith, P. L. Bartlett, and J. Baxter. Variance reduction techniques for gradient estimates in reinforcement learning. *Journal of Machine Learning Research*, 5:1471–1530, 2004.

[7] L. P. Kaelbling, M. L. Littman, and A. W. Moore. Reinforcement learning: A survey. *Journal of Artificial Intelligence Research*, 4:237–285, 1996.

[8] S. Kakade. A natural policy gradient. In T. G. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems 14*, pages 1531–1538, Cambridge, MA, 2002. MIT Press.

[9] J. Koza, K. Martin, S. Matthew, M. William, Y. Jessen, and L. Guido. *Genetic Programming IV: Routine Human-Competitive Machine Intelligence*. Kluwer Academic Publishers, 2003.

[10] M. G. Lagoudakis and R. Parr. Least-squares policy iteration. *Journal of Machine Learning Research*, 4:1107–1149, 2003.

[11] P. Larrañaga and J. A. Lozano. *Estimation of distribution algorithms: A new tool for evolutionary computation*. Kluwer Academic Publishers, Boston, 2002.

[12] P. Marbach and J. N. Tsitsiklis. Approximate gradient methods in policy-space optimization of Markov reward processes. *Discrete Event Dynamic Systems*, 13(1-2):111–148, 2004.

[13] A. Miyamae, Y. Nagata, I. Ono, and S. Kobayashi. Natural policy gradient methods with parameter-based exploration for control tasks. In *Advances in Neural Information Processing Systems*, volume 2, pages 437–441, 2010.

[14] J. Peters and S. Schaal. Policy gradient methods for robotics. In *Processing of the IEEE/RSJ International Conferece on Inatelligent Robots and Systems(IROS)*, 2006.

[15] T. Rückstiess, F. Sehnke, T. Schaul, D. Wierstra, Y. Sun, and J. Schmidhuber. Exploring parameter space in reinforcement learning. *Paladyn*, 1(1):14–24, 2010.

[16] F. Sehnke, C. Osendorfer, T. Rückstiess, A. Graves, J. Peters, and J. Schmidhuber. Parameter-exploring policy gradients. *Neural Networks*, 23(4):551–559, 2010.

[17] R. S. Sutton and G. A. Barto. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA, USA, 1998.

[18] G. Tesauro. TD-gammon, a self-teaching backgammon program, achieves master-level play. *Neural Computation*, 6(2):215–219, 1994.

[19] L. Weaver and J. Baxter. Reinforcement learning from state and temporal differences. Technical report, Department of Computer Science, Australian National University, 1999.

[20] L. Weaver and N. Tao. The optimal reward baseline for gradient-based reinforcement learning. In *Processings of the Seventeeth Conference on Uncertainty in Artificial Intelligence*, pages 538–545, 2001.

[21] J. D. Williams and S. Young. Partially observable Markov decision processes for spoken dialog systems. *Computer Speech and Language*, 21(2):231–422, 2007.

[22] R. J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8:229–256, 1992.