A Note on Variance Reduction for Policy Gradient with Action-Dependent Factorized Baselines

Sheau Ni Chen
Department of Computer Science
National Yang Ming Chiao Tung University
nini1217.cs08@nycu.edu.tw

1 Introduction

Policy gradient methods are able to compute unbiased gradient estimate but suffer from high variance. To reduce variance without introducing bias, a baseline is often use. In Variance Reduction for Policy Gradient with Action-Dependent Factorized Baselines (Wu et al. [2018]), they derived a bias-free action-dependent baseline to further reduce variance.

At the time when the paper was published, a large body of work has investigated variance reduction technique but the factorizability of policy probability distribution has not been studied. Methods like Q-Prop (Gu et al. [2017]) make use of action-dependent control variate and off-policy data, it can therefore be more sample efficient compared with on-policy methods. However, it is computationally expensive. In contrast, policy gradient with action-dependent baselines proposed in the paper has little computational overhead and is more sample efficient compared to on-policy method with state-only baseline.

The key idea of the paper is decomposing actions into multiple factors in order for variance reduction. Particularly when the factors are conditionally independent, we can compute a separate baseline for each factor, which depend on all information except for the factor itself, and can also remove the influences of other factors.

2 Problem Formulation

2.1 Notations

The paper assumed a discrete-time Markov Decision Process (MDP), defined by $(\mathcal{S}, \mathcal{A}, \mathcal{P}, r, \rho_0, \gamma)$.

- $S \subseteq \mathbb{R}^n$ is an n-dimensional state space.
- $A \subseteq \mathbb{R}^m$ is a m-dimensional action space.
- $\mathcal{P} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}_+$ is a transition probability function.

- $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is a bounded reward function.
- $\rho_0 : \mathcal{S} \to \mathbb{R}_+$ is an initial state distribution.
- $\gamma \in (0, 1]$ is a discount factor

We will try to optimize a stochastic policy $\pi_\theta : \mathcal{S} \times \mathcal{A} \to \mathbb{R}_+$ parameterized by θ. Let $\eta(\pi_\theta) = \mathbb{E}_\tau[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)]$ denote the expected return of the whole trajectory $s_0 \sim \rho_0(s_0)$, $a_t \sim \pi_\theta(a_t|s_t)$, and $s_{t+1} \sim \mathcal{P}(s_{t+1}|s_t, a_t)$ for all t. Our goal is to find the optimal policy $arg\ max_\theta\ \eta(\pi_\theta)$.

Let $\hat{Q}(s_t, a_t)$ denote samples of cumulative discounted reward[1]. And $Q(s_t, a_t)$[2] means a function approximation of $\hat{Q}(s_t, a_t)$.

2.2 Assumptions

The authors didn't make any additional assumption to the MDP. But they assume that action dimensions are conditional independent except for section 3.7.

2.3 Preliminaries

2.3.1 The Score Function (SF) Estimator

Suppose we want to estimate $\nabla_\theta \mathbb{E}_x[f(x)]$ where $x \sim p_\theta(x)$, the family of distribution $\{p_\theta(x) : \theta \in \Theta\}$ has common support and $\log p_\theta(x)$ is continuous in θ. We have:

$$\nabla_\theta \mathbb{E}_x[f(x)] = \nabla_\theta \int p_\theta(x) f(x)\, dx \tag{1}$$

$$= \int p_\theta(x) \frac{\nabla_\theta p_\theta(x)}{p_\theta(x)} f(x)\, dx \tag{2}$$

$$= \int p_\theta(x) \nabla_\theta \log p_\theta(x)\ f(x)\, dx \tag{3}$$

$$= \mathbb{E}_x[\nabla_\theta \log p_\theta(x)\ f(x)] \tag{4}$$

2.3.2 Policy Gradient

The Policy Gradient Theorem states that

$$\nabla_\theta \eta(\pi_\theta) = \mathbb{E}_\tau \left[\sum_{t=0}^{\infty} \nabla_\theta \log \pi_\theta(a_t|s_t) \sum_{t'=0}^{\infty} \gamma^{t'-t} r_{t'} \right] \tag{5}$$

Define state visitation frequency $\rho_\pi(s) = \sum_{t=0}^{\infty} \gamma^t p(s_t = s)$ and $\hat{Q}(s_t, a_t) = \sum_{t=0}^{\infty} \gamma^{t'-t} r_{t'}$, we can write

$$\nabla_\theta \eta(\pi_\theta) = \mathbb{E}_{\rho_\pi, \pi} \left[\nabla_\theta \log \pi_\theta(a_t|s_t) \hat{Q}(s_t, a_t) \right] \tag{6}$$

To reduce variance without introducing bias, we can subtract off a quantity dependent on s_t from $\hat{Q}(s_t, a_t)$.

$$\nabla_\theta \eta(\pi_\theta) = \mathbb{E}_{\rho_\pi, \pi} \left[\nabla_\theta \log \pi_\theta(a_t|s_t) \left(\hat{Q}(s_t, a_t) - b(s_t) \right) \right] \tag{7}$$

[1] In the paper, "cumulative discounted return" is used, but I think they meant "return", which is "cumulative discounted reward".

[2] $Q(a_t, s_t)$ is used in the paper (only in section 3.1), I assume it would not affect anything.

It can be shown by applying score function estimator:

$$\mathbb{E}_{a_t}\left[\nabla_\theta \log \pi_\theta(a_t|s_t) b(s_t)\right] = \nabla_\theta \mathbb{E}_{a_t}[b(s_t)] \tag{8}$$

$$= 0 \quad (b(s_t) \text{ is independent from } a_t) \tag{9}$$

The derivation of optimal state-dependent baseline would be shown in section 3.1

2.4 Problem to Solve

To design a method that decompose action into factors and calculate separate baselines of them in order to reduce variance when estimating gradient.

3 Theoretical Analysis

3.1 Derivation of Optimal State-Dependent Baseline

In appendix A of the paper, the authors gave a derivation of optimal state-dependent baseline by minimizing policy gradient estimate.

Define a random variable

$$g := \nabla_\theta \log \pi_\theta(a_t|s_t) \left(\hat{Q}(s_t, a_t) - b(s_t)\right), \quad a_t \sim \pi_\theta(a_t|s_t),\ s_t \sim \rho_\pi(s_t) \tag{10}$$

Then we can rewrite equation 7

$$\nabla_\theta \eta(\pi_\theta) = \mathbb{E}_{\rho_\pi, \pi}[g] \tag{11}$$

For convenience, we define

$$g_Q := \nabla_\theta \log \pi_\theta(a_t|s_t) \hat{Q}(s_t, a_t) \tag{12}$$

$$g_b := \nabla_\theta \log \pi_\theta(a_t|s_t) b(s_t) \tag{13}$$

$$\implies g = g_Q - g_b \tag{14}$$

The variance of policy gradient is

$$\mathrm{Var}(g) = \mathrm{Var}(g_Q - g_b) \tag{15}$$

$$= \mathrm{Var}(g_Q) + \mathrm{Var}(g_b) - 2\mathrm{Cov}(g_Q, g_b) \tag{16}$$

$$= \mathbb{E}_{\rho_\pi, \pi}\left[(g_Q)^T g_Q\right] - \mathbb{E}_{\rho_\pi, \pi}\left[g_Q\right]^T \mathbb{E}_{\rho_\pi, \pi}\left[g_Q\right] \tag{17}$$

$$+ \mathbb{E}_{\rho_\pi, \pi}\left[(g_b)^T g_b\right] - \mathbb{E}_{\rho_\pi, \pi}\left[g_b\right]^T \mathbb{E}_{\rho_\pi, \pi}\left[g_b\right] \tag{18}$$

$$- 2\left(\mathbb{E}_{\rho_\pi, \pi}\left[(g_Q)^T g_b\right] - \mathbb{E}_{\rho_\pi, \pi}\left[g_Q\right] \mathbb{E}_{\rho_\pi, \pi}\left[g_b\right]\right) \tag{19}$$

$$= \mathbb{E}_{\rho_\pi, \pi}\left[\nabla_\theta \log \pi_\theta(a_t|s_t)^T \nabla_\theta \log \pi_\theta(a_t|s_t) \hat{Q}(s_t, a_t)^2\right] \tag{20}$$

$$- \mathbb{E}_{\rho_\pi, \pi}\left[\nabla_\theta \log \pi_\theta(a_t|s_t) \hat{Q}(s_t, a_t)\right]^T \mathbb{E}_{\rho_\pi, \pi}\left[\nabla_\theta \log \pi_\theta(a_t|s_t) \hat{Q}(s_t, a_t)\right] \tag{21}$$

$$+ \mathbb{E}_{\rho_\pi, \pi}\left[\nabla_\theta \log \pi_\theta(a_t|s_t)^T \nabla_\theta \log \pi_\theta(a_t|s_t) b(s_t)^2\right] \tag{22}$$

$$- 2\mathbb{E}_{\rho_\pi, \pi}\left[\nabla_\theta \log \pi_\theta(a_t|s_t)^T \nabla_\theta \log \pi_\theta(a_t|s_t) \hat{Q}(s_t, a_t) b(s_t)\right] \tag{23}$$

$$(\mathbb{E}_{\rho_\pi, \pi}[g_b] = 0 \text{ (equation 8)} \implies \mathbb{E}_{\rho_\pi, \pi}[g_Q]\mathbb{E}_{\rho_\pi, \pi}[g_b] = 0)$$

In equation 21 to 23 of the paper, the authors write

$$\begin{aligned}
\mathrm{Var}(g) &= \mathbb{E}_{\rho_\pi,\pi}\left[(g-\mathbb{E}_{\rho_\pi,\pi}[g])^T(g-\mathbb{E}_{\rho_\pi,\pi}[g])\right] && (24)\\
&= \mathbb{E}_{\rho_\pi,\pi}\left[\nabla_\theta \log \pi_\theta(a_t|s_t)^T \nabla_\theta \log \pi_\theta(a_t|s_t)\right] b(s_t)^2 && (25)\\
&\quad - 2\mathbb{E}_{\rho_\pi,\pi}\left[\nabla_\theta \log \pi_\theta(a_t|s_t)^T \nabla_\theta \log \pi_\theta(a_t|s_t)\hat{Q}(s_t,a_t)\right] b(s_t) && (26)
\end{aligned}$$

There are two mistakes. One is that $b(s_t)$ should be inside of the expectation because s_t is drawn from ρ_π. The other one is the $\mathrm{Var}(g_Q)$ term (equation 20 and 21) is missing (it will disappear when taking derivative).

To minimize variance $\mathrm{Var}(g)$, we need to have $\frac{\partial}{\partial b}[\mathrm{Var}(g)] = 0$ when baseline $b(s_t)$ is the optimal one $b^*(s_t)$.

$$\begin{aligned}
0 &= \frac{\partial}{\partial b}[\mathrm{Var}(g)]\Big|_{b(s_t)=b^*(s_t)} && (27)\\
&= 2\mathbb{E}_\pi\left[\nabla_\theta \log \pi_\theta(a_t|s_t)^T \nabla_\theta \log \pi_\theta(a_t|s_t)\right] b^*(s_t) && (28)\\
&\quad - 2\mathbb{E}_\pi\left[\nabla_\theta \log \pi_\theta(a_t|s_t)^T \nabla_\theta \log \pi_\theta(a_t|s_t)\hat{Q}(s_t,a_t)\right] && (29)
\end{aligned}$$

Then we can solve the optimal state-dependent baseline

$$b^*(s_t) = \frac{\mathbb{E}_\pi\left[\nabla_\theta \log \pi_\theta(a_t|s_t)^T \nabla_\theta \log \pi_\theta(a_t|s_t)\hat{Q}(s_t,a_t)\right]}{\mathbb{E}_\pi\left[\nabla_\theta \log \pi_\theta(a_t|s_t)^T \nabla_\theta \log \pi_\theta(a_t|s_t)\right]} \quad (30)$$

In the paper (equation 25 to 27), the authors write $\mathbb{E}_{\rho_\pi,\pi}$. However, since we need to have optimal value for every s_t, we should use $\mathbb{E}_\pi$.

3.2 Baselines for Policies with Conditionally Independent Factors

Assume that the factors of policy are conditionally independent. We also assume a m-dimensional action space and a_t^i denote the i-th factor of a_t, we have:

$$\pi_\theta(a_t|s_t) = \prod_{i=1}^{m} \pi_\theta(a_t^i|s_t) \quad (31)$$

Hence

$$\begin{aligned}
\nabla_\theta \eta(\pi_\theta) &= \mathbb{E}_{\rho_\pi,\pi}\left[\nabla_\theta \log \pi_\theta(a_t^i|s_t)\hat{Q}(s_t,a_t)\right] && (32)\\
&= \mathbb{E}_{\rho_\pi,\pi}\left[\nabla_\theta \log\left(\prod_{i=1}^{m} \pi_\theta(a_t^i|s_t)\right)\hat{Q}(s_t,a_t)\right] && (33)\\
&= \mathbb{E}_{\rho_\pi,\pi}\left[\nabla_\theta \sum_{i=1}^{m} \log \pi_\theta(a_t^i|s_t)\hat{Q}(s_t,a_t)\right] && (34)\\
&= \mathbb{E}_{\rho_\pi,\pi}\left[\sum_{i=1}^{m} \nabla_\theta \log \pi_\theta(a_t^i|s_t)\hat{Q}(s_t,a_t)\right] && (35)
\end{aligned}$$

Let a_t^{-i} denote all dimensions other than i in a_t and let $b_i(s_t, a_t^{-i})$ denote the baseline of the i-th factor. Due to score function estimator and conditional independent, we have

$$\mathbb{E}_{a_t}\left[\nabla_\theta \log \pi_\theta(a_t^i|s_t) b_i(s_t, a_t^{-i})\right] = \nabla_\theta \mathbb{E}_{a_t}\left[b_i(s_t, a_t^{-i})\right] \tag{36}$$

(Apply score function estimator)

$$= \mathbb{E}_{a_t^{-i}}\left[\nabla_\theta \mathbb{E}_{a_t^i}[b_i(s_t, a_t^{-i})]\right] \tag{37}$$

(By conditional independent assumption)

$$= 0 \qquad (a_t^{-i} \text{ does not depend on } a_t^i) \tag{38}$$

Hence we can use the following gradient estimator:

$$\nabla_\theta \eta(\pi_\theta) = \mathbb{E}_{\rho_\pi, \pi}\left[\sum_{i=1}^{m} \nabla_\theta \log \pi_\theta(a_t^i|s_t)\left(\hat{Q}(s_t, a_t) - b_i(s_t, a_t^{-i})\right)\right] \tag{39}$$

In section 3.7, we can also show that it also applies to general policy structures where the conditional independent assumption does not hold.

3.3 Optimal Action-Dependent Baseline

In this section, we are going to derive the optimal action-dependent baseline.

Define $z_i := \nabla_\theta \log \pi_\theta(a_t^i|s_t)$.

And policy gradient of the component is

$$\nabla_\theta \eta_i(\pi_\theta) := \mathbb{E}_{\rho_\pi, \pi}\left[\nabla_\theta \log \pi_\theta(a_t^i|s_t)\left(\hat{Q}(s_t, a_t) - b_i(s_t, a_t^{-i})\right)\right] \tag{40}$$

In the paper, the authors use ∇ in the above equation instead of ∇_θ. (θ is missing)

For simplicity, we make the following assumption:

$$\nabla_\theta \log \pi_\theta(a_t^i|s_t)^T \nabla_\theta \log \pi_\theta(a_t^j|s_t) \equiv z_i^T z_j = 0, \ \ \forall i \neq j \tag{41}$$

which mean different subsets of parameters strongly influence different action dimensions or factors. This assumption is made only for the analysis to be clean. Even without this assumption, the proposed baseline is bias-free.

To derive the optimal action-dependent baseline, we denote

$$g_i := \nabla_\theta \log \pi_\theta(a_t^i|s_t)\left(\hat{Q}(s_t, a_t) - b_i(s_t, a_t^{-i})\right), \quad a_t \sim \pi_\theta(a_t|s_t), s_t \sim \rho_\pi(s_t) \tag{42}$$

such that

$$\nabla_\theta \eta(\pi_\theta) = \mathbb{E}_{\rho_\pi, \pi}[g] = \mathbb{E}_{\rho_\pi, \pi}\left[\sum_{i=1}^{m} g_i\right] \tag{43}$$

Denote mean correlation term

$$M_{ij} := \mathbb{E}_{\rho_\pi, \pi}\left[z_i \hat{Q}(s_t, a_t)\right]^T \mathbb{E}_{\rho_\pi, \pi}\left[z_j \hat{Q}(s_t, a_t)\right] \tag{44}$$

and let $M = \sum_i \sum_j M_{ij}$. Note that M does not depend on $b_i(\cdot)$.

Under the above assumptions, we have:

$$\mathrm{Var}(\sum_{i=1}^{m} g_i) = \sum_i \mathrm{Var}(g_i) + \sum_i \sum_{j \neq i} \mathrm{Cov}(g_i, g_j) \tag{45}$$

$$= \sum_i \mathrm{Var}(g_i) + \sum_i \sum_{j \neq i} \left(\mathbb{E}_{\rho_\pi, \pi}[g_i^T g_j] - \mathbb{E}_{\rho_\pi, \pi}[g_i]^T \mathbb{E}_{\rho_\pi, \pi}[g_j] \right) \tag{46}$$

(There's no bracket in the second term in the paper, I assume it is a typo)

$$= \sum_i \mathrm{Var}(g_i) - \sum_i \sum_{j \neq i} \mathbb{E}_{\rho_\pi, \pi}[g_i]^T \mathbb{E}_{\rho_\pi, \pi}[g_j] \tag{47}$$

(by assumption $z_i z_j = 0,\ \forall i \neq j$)

$$= \sum_i \mathrm{Var}(g_i) - \sum_i \sum_{j \neq i} M_{ij} \tag{48}$$

(By equation 38, $\mathbb{E}_{a_t}[\nabla_\theta \log \pi_\theta(a_t^i | s_t) b_i(s_t, a_t^{-i})] = 0$)

The overall variance is minimized when each component variance is minimized since M does not depend on baseline b.

$$\mathrm{Var}(g_i) = \mathbb{E}_{\rho_\pi, \pi}\left[z_i^T z_i \left(\hat{Q}(s_t, a_t) - b_i(s_t, a_t^{-i}) \right)^2 \right] \tag{49}$$

$$- \mathbb{E}_{\rho_\pi, \pi}\left[z_i \left(\hat{Q}(s_t, a_t) - b_i(s_t, a_t^{-i}) \right) \right]^T \mathbb{E}_{\rho_\pi, \pi}\left[z_i \left(\hat{Q}(s_t, a_t) - b_i(s_t, a_t^{-i}) \right) \right] \tag{50}$$

$$= \mathbb{E}_{\rho_\pi, \pi}\left[z_i^T z_i \left(\hat{Q}(s_t, a_t)^2 - 2\hat{Q}(s_t, a_t) b_i(s_t, a_t^{-i}) + b_i(s_t, a_t^{-i})^2 \right) \right] \tag{51}$$

(In the paper, the "square" mark of $b_i(s_t, a_t^{-i})^2$ is put in a wrong place.)

$$- \mathbb{E}_{\rho_\pi, \pi}\left[z_i \hat{Q}(s_t, a_t) \right]^T \mathbb{E}_{\rho_\pi, \pi}\left[z_i \hat{Q}(s_t, a_t) \right] \tag{52}$$

(By equation 38, $\mathbb{E}_{a_t}[\nabla_\theta \log \pi_\theta(a_t^i | s_t) b_i(s_t, a_t^{-i})] = 0$)

$$= \mathbb{E}_{\rho_\pi, \pi}\left[z_i^T z_i \hat{Q}(s_t, a_t)^2 \right] - M_{ii} \tag{53}$$

(They do not depend on b_i)

$$+ \mathbb{E}_{\rho_\pi, a_t^{-i}}\left[-2\mathbb{E}_{a_t^i}\left[z_i^T z_i \hat{Q}(s_t, a_t) \right] b_i(s_t, a_t^{-i}) + \mathbb{E}_{a_t^i}\left[z_i^T z_i \right] b_i(s_t, a_t^{-i})^2 \right] \tag{54}$$

(conditional independent)

Minimize this variance

$$\frac{\partial}{\partial b_i}\left[\mathrm{Var}\left(\sum_i g_i \right) \right] = \frac{\partial}{\partial b_j}\left[\mathrm{Var}(g_j) \right] = 0, \quad \forall j \tag{55}$$

$$\Longrightarrow b_i^*(s_t, a_t^{-i}) = \frac{\mathbb{E}_{a_t^i}\left[z_i^T z_i \hat{Q}(s_t, a_t) \right]}{\mathbb{E}_{a_t^i}\left[z_i^T z_i \right]} \tag{56}$$

Therefore the optimal action-dependent baseline is

$$b_i^*(s_t, a_t^{-i}) = \frac{\mathbb{E}_{a_t^i}\left[\nabla_\theta \log \pi_\theta(a_t^i | s_t)^T \nabla_\theta \log \pi_\theta(a_t^i | s_t) \hat{Q}(s_t, a_t) \right]}{\mathbb{E}_{a_t^i}\left[\nabla_\theta \log \pi_\theta(a_t^i | s_t)^T \nabla_\theta \log \pi_\theta(a_t^i | s_t) \right]} \tag{57}$$

3.4 Suboptimality of the optimal state-dependent baseline

In this chapter, we are going to answer how much variance we can reduce by using action-dependent baseline over a traditional state-dependent baseline.

Let $\mathrm{Var}^*(\sum_i g_i)$ be the variance resulting from the optimal state-dependent baseline and $\mathrm{Var}(\sum_i g_i)$ be variance resulting from another baseline $b = (b_i(s_t, a_t^{-i}))_{i\in[m]}$, which may be sub-optimal or action independent.

And we use the following notation:

$$Z_i := Z_i(s_t, a_t^{-i}) = \mathbb{E}_{a_t^i}\left[\nabla_\theta \log \pi_\theta(a_t^i|s_t)^T \nabla_\theta \log \pi_\theta(a_t^i|s_t)\right] \tag{58}$$

$$Y_i := Y_i(s_t, a_t^{-i}) = \mathbb{E}_{a_t^i}\left[\nabla_\theta \log \pi_\theta(a_t^i|s_t)^T \nabla_\theta \log \pi_\theta(a_t^i|s_t)\hat{Q}(s_t, a_t)\right] \tag{59}$$

$$X_i := X_i(s_t, a_t^{-i}) = \mathbb{E}_{a_t^i}\left[\nabla_\theta \log \pi_\theta(a_t^i|s_t)^T \nabla_\theta \log \pi_\theta(a_t^i|s_t)\hat{Q}(s_t, a_t)^2\right] \tag{60}$$

The variance can be written as:

$$\mathrm{Var}(\sum_i g_i) = \sum_i \mathrm{Var}(g_i) - \sum_i \sum_{j\neq i} M_{ij} \tag{61}$$

$$= \sum_i \left(\mathbb{E}_{\rho_\pi, a_t}\left[z_i^T z_i \left(\hat{Q}(s_t, a_t)^2 - 2\hat{Q}(s_t, a_t) b_i(s_t, a_t^{-i}) + b_i(s_t, a_t^{-i})^2\right)\right] - M_{ii}\right) \tag{62}$$

$$+ \left(\sum_i M_{ii} - M\right) \tag{63}$$

(By equation 51.)

$$= \sum_i \left(\mathbb{E}_{\rho_\pi, a_t^{-i}}\left[X_i - 2b_i(s_t, a_t^{-i}) Y_i + b_i(s_t, a_t^{-i})^2 Z_i\right]\right) - M \tag{64}$$

(Separate a_t^i from a_t^{-i})

By equation 57, the optimal action-dependent baseline is

$$b_i^*(s_t, a_t) = \frac{\mathbb{E}_{a_t^i}\left[\nabla_\theta \log \pi_\theta(a_t^i|s_t)^T \nabla_\theta \log \pi_\theta(a_t^i|s_t)\hat{Q}(s_t, a_t)\right]}{\mathbb{E}_{a_t^i}\left[\nabla_\theta \log \pi_\theta(a_t^i|s_t)^T \nabla_\theta \log \pi_\theta(a_t^i|s_t)\right]} = \frac{Y_i}{Z_i} \tag{65}$$

Therefore by substituting $\frac{Y_i}{Z_i}$ into b_i in equation 64, the variance of gradient with optimal action-dependent baseline can be written as:

$$\mathrm{Var}^*\left(\sum_i g_i\right) = \sum_i \mathbb{E}_{\rho_\pi, a_t^{-i}}\left[X_i - \frac{Y_i^2}{Z_i}\right] - M \tag{66}$$

Finally, define the variance improvement $I_b := \mathrm{Var}(\sum_i g_i) - \mathrm{Var}^*(\sum_i g_i)$, it can be calculated as:

$$I_b := \sum_i \left(\mathbb{E}_{\rho_\pi, a_t^{-i}} \left[X_i - 2b_i(s_t, a_t^{-i}) Y_i + b_i(s_t, a_t^{-i})^2 Z_i \right] - \mathbb{E}_{\rho_\pi, a_t^{-i}} \left[X_i - \frac{Y_i^2}{Z_i} \right] \right) \tag{67}$$

$$= \sum_i \mathbb{E}_{\rho_\pi, a_t^{-i}} \left[b_i(s_t, a_t^{-i})^2 Z_i - 2b_i(s_t, a_t) Y_i + \frac{Y_i^2}{Z_i} \right] \tag{68}$$

$$= \sum_i \mathbb{E}_{\rho_\pi, a_t^{-i}} \left[\left(b_i(s_t, a_t^{-i}) \sqrt{Z_i} - \frac{Y_i}{\sqrt{Z_i}} \right)^2 \right] \tag{69}$$

(Z_i is non-negative because it is a inner product of a vector with itself)

$$= \sum_i \mathbb{E}_{\rho_\pi, a_t^{-i}} \left[Z_i \left(b_i(s_t, a_t^{-i}) - \frac{Y_i}{Z_i} \right)^2 \right] \tag{70}$$

$$= \sum_i \mathbb{E}_{\rho_\pi, a_t^{-i}} \left[Z_i \left(b_i(s_t, a_t^{-i}) - b_i^*(s_t, a_t^{-i}) \right)^2 \right] \tag{71}$$

$$= \sum_i \mathbb{E}_{\rho_\pi, a_t^{-i}} \left[\mathbb{E}_{a_t^i} \left[\nabla_\theta \log \pi_\theta(a_t^i | s_t)^T \nabla_\theta \log \pi_\theta(a_t^i | s_t) \right] \left((b_i(s_t, a_t^{-i}) - b_i^*(s_t, a_t^{-i})) \right)^2 \right] \tag{72}$$

When $b = b^*(s)$, where $b^*(s)$ is optimal state-dependent derived in section 3.1, we have:

$$I_{b=b^*(s)} := \sum_i \mathbb{E}_{\rho_\pi, a_t^{-i}} \left[Z_i \left(b_i^*(s_t) - b_i^*(s_t, a_t^{-i}) \right)^2 \right] \tag{73}$$

$$= \sum_i \mathbb{E}_{\rho_\pi, a_t^{-i}} \left[Z_i \left(\frac{\sum_j Y_j}{\sum_j Z_j} - \frac{Y_i}{Z_i} \right)^2 \right] \tag{74}$$

(Conditional independent assumption)

$$= \sum_i \mathbb{E}_{\rho_\pi, a_t^{-i}} \left[\frac{1}{Z_i} \left(\frac{Z_i}{\sum_j Z_j} \sum_j Y_j - Y_i \right)^2 \right] \tag{75}$$

(Value inside expectation is non-negative $\rightarrow$ improvement is non-negative)

The difference will be particularly large when the Q function is highly sensitive to action.

3.5 Marginalization of the Global action-value function

Optimal baselines derived above are too computationally expensive to use in practice. Therefore the authors proposed various baselines and analyzed their computational cost in the paper.

3.5.1 Marginalized Q Baseline

In practice, we often use $b(s_t) = \mathbb{E}_{a_t}[\hat{Q}(s_t, a_t)] = V(s_t)$. Similarly, the authors proposed to use $b_i(s_t, a_t^{-1}) = \mathbb{E}_{a_t^i}[\hat{Q}(s_t, a_t)]$, which is action-dependent.

In particular, when log probability of each policy factor is loosely correlated with action-value function, i.e.

$$\mathbb{E}_{a_t^i} \left[z_i^T z_i \hat{Q}(s_t, a_t) \right] \approx \mathbb{E}_{a_t^i} \left[z_i^T z_i \right] \mathbb{E}_{a_t^i} \left[\hat{Q}(s_t, a_t) \right] \tag{76}$$

Then the proposed baseline is close to the optimal baseline, i.e.

$$I_{b=\mathbb{E}_{a_t^i}[\hat{Q}(s_t,a_t)]} = \sum_i \mathbb{E}_{\rho_\pi, a_t^{-i}} \left[Z_i (\mathbb{E}_{a_t^i} \left[\hat{Q}(s_t, a_t) \right] - b^*(s_t, a_t^{-i})^2 \right] \tag{77}$$

(By equation 71)

$$= \sum_i \mathbb{E}_{\rho_\pi, a_t^{-i}} \left[Z_i \left(\mathbb{E}_{a_t^i} \left[\hat{Q}(s_t, a_t) \right] - \frac{\mathbb{E}_{a_t^i} \left[z_i^T z_i \hat{Q}(s_t, a_t) \right]}{\mathbb{E}_{a_t^i}[z_i^T z_i]} \right)^2 \right] \tag{78}$$

$$\approx 0 \tag{79}$$

This has another benefit that it only requires learning one function approximator, and can simply use it to obtain baseline value for each action coordinate.

3.5.2 Monte Carlo Marginalized Q Baseline

After fitting $Q_{\pi_\theta}(s_t, a_t)$, we can sample M i-th coordinate action $\alpha_j \sim \pi_\theta(a_t^i | s_t)\ \forall j \in [0, M)$ and obtain baseline through Monte Carlo estimates:

$$b_i(s_t, a_t^{-i}) = \frac{1}{M} \sum_{j=0}^{M} Q_{\pi_\theta} \left(s_t, (a_t^{-i}, \alpha_j) \right) \tag{80}$$

In fact, any function can be used to aggregate the samples $Q_{\pi_\theta}(s_t, (a_t^{-i}, \alpha_j))$, for discrete action dimensions, we can use max instead of mean.

3.5.3 Mean Marginalized Q Baseline

Monte Carlo method can be computationally expensive, therefore the proposed another baseline to reduce computational burden:

$$b_i(s_t, a_t^{-i}) = Q_{\pi_\theta} \left(s_t, (a_t^{-i}, (\bar{a})_t^{i3}) \right) \tag{81}$$

where $(\bar{a})_t^i = \mathbb{E}_{\pi_\theta}[a_t^i]$ is the average action of coordinate i.

3.6 Final Algorithm

In the following algorithm, we assume that action dimensions are conditional independent.

Algorithm 1 Policy gradient for factorized policies using action-dependent baselines

Require: Number of iterations N, batch size B, initial policy parameters θ

Initialize action-value function estimate $Q_{\pi_\theta}(s_t, a_t) \equiv 0$ and policy π_θ

for j in $\{1, \ldots, N\}$ **do**

Collect samples: $(s_t, a_t)_{t \in \{1, \ldots, B\}}$

Compute baseline: $b_i(s_t, a_t^{-i}) = \mathbb{E}_{a_t^i}[\hat{Q}(s_t, a_t)]$ for $i \in \{1, \ldots, m\}$

Compute advantage $\hat{A}_i(s_t, a_t) := \hat{Q}(s_t, a_t) - b_i(s_t, a_t^{-i}),\ \forall t$

Perform a policy update step on θ using $\hat{A}_i(s_t, a_t)$

Update action-value function approximation with current batch: $Q_{\pi_\theta}(s_t, a_t)$

end for

3.7 Baselines for General Actions

In this section, we consider a more general case. Assume there are m factors a_t^i to a_t^m, which together form action a_t.

Conditioned on s_t, the factors of action a_t drawn from $\pi_\theta(a_t|s_t)$ form a certain directed acyclic graphic model, including fully dependent case. Consider $f(i)$ denotes the indices of the parents of the i-th factor, which means $f(i)$ is the set og factors that the i-th factor depends on. And let $D(i)$ denote the indices of descendants of i in the graphic model (including i itself) and $D(i)$ is the set of factors that the i-th factor affects.

Without loss of generality, assume that the following factorization holds:

$$\pi_\theta(a_t, s_t) = \prod_{i=1}^{m} \pi_\theta(a_t^i|s_t, a_t^{f(i)}) \tag{82}$$

In this case, we can set the i-th baseline to be $b_i(s_t, a_t^{[m]\setminus D(i)})$, where $[m] = \{1, 2, \ldots, m\}$. In other words, the i-th baseline depends on the factor it does not affect.

The gradient estimator can therefore be written as

$$\nabla_\theta \eta(\pi_\theta) = \mathbb{E}_{\rho_\pi, \pi}\left[\sum_{i=1}^{m} \nabla_\theta \log \pi_\theta(a_t^i|s_t, a_t^{f(i)}) \left(\hat{Q}(s_t, a_t) - b_i(s_t, a_t^{[m]\setminus D(i)})\right)\right] \tag{83}$$

In the most general case without any conditional independence assumption, we have

$$f(i) = \{1, 2, \ldots, i-1\} \tag{84}$$

$$D(i) = \{i, i+1, \ldots, m\} \tag{85}$$

And we can write

$$\nabla_\theta \eta(\pi_\theta) = \mathbb{E}_{\rho_\pi, \pi}\left[\sum_{i=1}^{m} \nabla_\theta \log \pi_\theta(a_t^i|s_t, a_t^1, \ldots, a_t^{-1}) \left(\hat{Q}(s_t, a_t) - b_i(s_t, a_t^1, \ldots, a_t^{-1})\right)\right] \tag{86}$$

The analysis in section 3.3 and section 3.4 transfers also to this general case.

The final algorithm with general actions is:

Algorithm 2 Policy gradient for general factorization policies using action-dependent baselines

Require: Number of iterations N, batch size B, initial policy parameters θ

Initialize baseline $b_i(s_t, a_t^{[m]\setminus D(i)}) \equiv 0$, for $i \in \{1, \ldots, m\}$ and policy π_θ

for j in $\{1, \ldots, N\}$ **do**

Collect samples: $(s_t, a_t)_{t \in \{1, \ldots, B\}}$

Compute advantage $\hat{A}_i(s_t, a_t) := \hat{Q}(s_t, a_t) - b_i(s_t, a_t^{[m]\setminus D(i)}),\ \forall t$

Perform a policy update step on θ using $\hat{A}_i(s_t, a_t)$

Update action-value function approximation with current batch: $b_i(s_t, a_t^{[m]\setminus D(i)})$

end for

4 Conclusion

We can conclude that considering actions in baseline can indeed further reduce variance compared to state-only baseline without introducing bias, and it can be applied to both conditionally independent cases and general cases.

References

Cathy Wu, Aravind Rajeswara, Yan Duan, Vikash Kumar, Alexandre M Bayen, Sham Kakade, Igor Mordatch, and Pieter Abbeel. *Variance Reduction for Policy Gradient with Action-Dependent Factorized Baselines*. ICLR, 2018.

Shixiang Gu, Timothy Lillicrap, Zoubin Ghahramani, Richard E Turner, and Sergey Levine. *Q-prop: Sample-efficient policy gradient with an off-policy critic*. ICLR, 2017.