A Note On Action-dependent Baseline For Variance Reduction

Qian-You Zhang
Department of Computer Science
National Yang Ming Chiao Tung University
qianyou.cs07@nycu.edu.tw

1 Introduction

Policy gradient gives us an unbiased way to update the parameters of our value network. While we are enjoying the nice properties it brings us, policy gradient does have a serious issue to be solved, which is high variance of gradient estimates. This high variance comes from sampling trajectories that are composed of a large number of steps. It would cause the training process harder and takes more iterations to converge. There are various ways to tackle the issue such as designing a baseline or including a additional critic network. The paper focuses on proposing and validating a optimal baseline function, which can reduce variance of gradient estimates most. Different from the classic state-dependent baseline in the literature, an action-dependent baseline is proposed by this paper. The key insights of introducing action into baseline is that it can exploit the additional information to further reduce variance. Specifically, their method computes a separate baseline for each action factor when these factors are conditionally independent. Also, they found that the action-dependent baseline can reach optimality with an additional assumption. Overall, I am impressed with the practical result and algorithm they derived. It gives a simple and computationally efficient way to find the action-dependent baseline that helps in increasing the speed of convergence.

2 Problem Formulation

2.1 Notation

This paper assumes a discrete-time Markov decision process (MDP), defined by $(\mathcal{S}, \mathcal{A}, \mathcal{P}, r, \rho_0, \gamma)$, in which $\mathcal{S} \subseteq \mathbb{R}^n$ is an n-dimensional state space, $\mathcal{A} \subseteq \mathcal{R}^m$ an m-dimensional action space, $\mathcal{P} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}_+$ a transition probability function, $r : \mathcal{S} \times \mathcal{A} \to \mathcal{R}$ a bounded reward function, $\rho_0 : \mathcal{S} \to \mathbb{R}_+$ an initial state distribution, and $\gamma \in (0, 1]$ a discount factor. The presented models are based on the optimization of a stochastic policy $\pi_\theta : \mathcal{S} \times \mathcal{A} \to \mathbb{R}_+$ parameterized by θ. Let $\eta(\pi_\theta)$ denote its expected return: $\eta(\pi_\theta) = \mathbb{E}_\tau[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)]$, where $\tau = (s_0, a_0, ...)$ denotes 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 $\arg\max_\theta \eta(\pi_\theta)$. We will use $\hat{Q}(s_t, a_t)$ to describe samples of cumulative discounted return, and $Q(a_t, s_t)$ to describe a function approximation of $\hat{Q}(s_t, a_t)$. We will use "Q-function" when describing an abstract action-value function.

2.2 Optimization Problem

Let b_i denotes the baseline for the ith factor, a_t^{-i} denotes all dimensions other than i in a_t. We set $b_i = b_i(s_t, a_t^{-i})$ to introduce all dimension of actions except a_t^i. I am going to show unbiasedness of gradient estimates with action-dependent baseline for each action factor and the following gradient estimator:

$$\mathbb{E}_{a_t}\Big[\nabla_\theta \log \pi_\theta(a_t^i|s_t) b_i(s_t, a_t^{-i})\Big] = 0 \tag{1}$$

$$\nabla_\theta \eta(\pi_\theta) = \mathbb{E}_{\rho_\pi,\pi}\Bigg[\sum_{i=1}^{m} \nabla_\theta \log \pi_\theta(a_t^i|s_t)\Big(\hat{Q}(s_t,a_t) - b_i(s_t, a^- i_t)\Big)\Bigg] \tag{2}$$

Also, I show the optimal state-dependent baseline $b^*(s_t)$ as well as optimal action-dependent baseline $b_i^*(s_t, a^- i_t)$ that minimize the variance of gradient estimates.

2.3 Assumptions

Assumption 2.3.1. *Conditionally independent action factors:*

$$\pi_\theta(a_t|s_t) = \prod_{i=1}^{m} \pi_\theta(a_t^i|s_t) \tag{3}$$

Assumption 2.3.2.

$$\nabla_\theta \log \pi_\theta(a_t^i|s_t)^T \nabla_\theta \log \pi_\theta(a_t^j|s_t) \equiv z_i^T z_j = 0, \forall i \neq j \tag{4}$$

2.4 Preliminaries

2.4.1 The Score Function Estimator

Score function has a nice property that we will use, that is, expected value of score function is exact zero. It's useful when proving the unbiasedness of gradient estimates. Suppose that we want to estimate $\nabla_\theta \mathbb{E}_x[f(x)]$ where $x \sim p_\theta(x)$, and the family of distributions $\{p_\theta(x) : \theta \in \Theta\}$ has common support. Further suppose that $\log p_\theta(x)$ is continuous in θ. In this case we have

$$\begin{aligned}\nabla_\theta \mathbb{E}_x[f(x)] &= \nabla_\theta \int p_\theta(x) f(x) dx = \int p_\theta(x) \frac{\nabla_\theta p_\theta(x)}{p_\theta(x)} f(x) dx \\ &= \int p_\theta(x) \nabla_\theta \log p_\theta(x) f(x) dx = \mathbb{E}_x[\nabla_\theta \log p_\theta(x) f(x)].\end{aligned} \tag{5}$$

2.4.2 Policy Gradient

The Policy Gradient Theorem (Sutton et al., 2000) states that

$$\nabla_\theta \eta(\pi_\theta) = \mathbb{E}_\tau\left[\sum_{t=0}^{\infty} \nabla_\theta \log \pi_\theta(a_t|s_t) \sum_{t'=t}^{\infty} \gamma^{t'=t} r_{t'}\right]. \tag{6}$$

Define $\rho_\pi(s) = \sum_{t=0}^{\infty} \gamma^t p(s_t = s)$ as the state visitation frequency, and $\hat{Q}(s_t, a_t) = \sum_{t'=t}^{\infty} \gamma^{t'-t} r_{t'}$. We can rewrite the above equation as

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

After applying state-dependent baseline into the above equation, we got

$$\nabla_\theta \eta(\pi_\theta) = \mathbb{E}_{\rho_\pi,\pi}\Big[\nabla_\theta \log \pi_\theta(a_t|s_t)\Big(\hat{Q}(s_t,a_t) - b(s_t)\Big)\Big]. \tag{8}$$

3 Theoretical Analysis

3.1 Unbiasedness of imported action-dependent baseline

Lemma 3.1.1. *For continuous function $f(x)$, if $f(x)$ is independent from x s.t. $\forall x : y = f(x) = C \in$ const, then $\nabla_\theta \mathbb{E}_x[f(x)] = 0$.*

Proof.

$$\begin{aligned}\nabla_\theta \mathbb{E}_x[f(x)] &= \nabla_\theta \int p_\theta(x) f(x) dx \\ &= \nabla_\theta \int p_\theta(x) C dx \\ &= \nabla_\theta C \cdot \int p_\theta(x) dx \\ &= \nabla_\theta C = 0\end{aligned}$$

□

Using Equation (7) and Assumption 2.3.1, we show the form of policy gradient for factorized action

$$\nabla_\theta \eta(\pi_\theta) = \mathbb{E}_{\rho_\pi, \pi}\big[\nabla_\theta \log \pi_\theta(a_t|s_t)\hat{Q}(s_t, a_t)\big] = \mathbb{E}_{\rho_\pi, \pi}\left[\sum_{i=1}^{m} \nabla_\theta \log \pi_\theta(a_t^i|s_t)\hat{Q}(s_t, a_t)\right] \tag{9}$$

According to Lemma 3.1.1 and score function estimator, we have

$$\begin{aligned}\mathbb{E}_{a_t}\Big[\nabla_\theta \log \pi_\theta(a_t^i|s_t) b_i(s_t, a_t^{-i})\Big] &= \nabla_\theta \mathbb{E}_{a_t}\Big[b_i(s_t, a_t^{-i})\Big] \\ &= \nabla_\theta \mathbb{E}_{a_t^{-i}}\left[\mathbb{E}_{a_t^i}\Big[b_i(s_t, a_t^{-i})\Big]\right] \\ &= \mathbb{E}_{a_t^{-i}}\left[\nabla_\theta \mathbb{E}_{a_t^i}\Big[b_i(s_t, a_t^{-i})\Big]\right] \\ &= 0\end{aligned} \tag{10}$$

Accordingly, applying baseline b_i into each action factor, we get

$$\nabla_\theta \eta(\pi_\theta) = \mathbb{E}_{\rho_\pi, \pi}\left[\sum_{i=1}^{m} \nabla_\theta \log \pi_\theta(a_t^i|s_t)\Big(\hat{Q}(s_t, a_t) - b_i(s_t, a_t^{-i})\Big)\right] \tag{11}$$

3.2 Optimal state-dependent baseline and action-dependent baseline

3.2.1 Optimal state-dependent baseline

Denote g to be the random variable that estimates the policy gradient with state-dependent baseline:

$$g(s_t) \coloneqq \nabla_\theta \log \pi_\theta(a_t|s_t)\Big(\hat{Q}(s_t, a_t) - b(s_t)\Big), a_t \sim \pi_\theta(a_t|s_t) \tag{12}$$

The variance of the policy gradient with state-dependent baseline is:

$$\begin{aligned}\text{Var}(g) &= \mathbb{E}_{a_t}[g^T g] - \mathbb{E}_{a_t}[g]^T \mathbb{E}_{a_t}[g] \\ &= \mathbb{E}_{a_t}[g^T g] - \mathbb{E}_{a_t}[\nabla_\theta \log \pi_\theta(a_t|s_t)\hat{Q}(s_t, a_t)]^T \mathbb{E}_{a_t}[\nabla_\theta \log \pi_\theta(a_t|s_t)\hat{Q}(s_t, a_t)] \\ &= \mathbb{E}_{a_t}\Big[\nabla_\theta \log \pi_\theta(a_t|s_t)^T \nabla_\theta \log \pi_\theta(a_t|s_t)\Big] b(s_t)^2 \\ &\quad - 2\mathbb{E}_{a_t}\Big[\nabla_\theta \log \pi_\theta(a_t|s_t)^T \nabla_\theta \log \pi_\theta(a_t|s_t)\hat{Q}(s_t, a_t)\Big] b(s_t) \\ &\quad + \mathbb{E}_{a_t}\Big[\nabla_\theta \log \pi_\theta(a_t|s_t)^T \nabla_\theta \log \pi_\theta(a_t|s_t)\hat{Q}(s_t, a_t)^2\Big] \\ &\quad - \mathbb{E}_{a_t}[\nabla_\theta \log \pi_\theta(a_t|s_t)\hat{Q}(s_t, a_t)]^T \mathbb{E}_{a_t}[\nabla_\theta \log \pi_\theta(a_t|s_t)\hat{Q}(s_t, a_t)]\end{aligned} \tag{13}$$

Then, we take the derivative of Var(g) with respect to b in order to find $b^*(s_t)$ that minimizes variance:

$$\begin{aligned}\frac{\partial}{\partial b}[\text{Var}(g)] &= 2\mathbb{E}_{a_t}\Big[\nabla_\theta \log \pi_\theta(a_t|s_t)^T \nabla_\theta \log \pi_\theta(a_t|s_t)\Big] b(s_t) \\ &\quad - 2\mathbb{E}_{a_t}\Big[\nabla_\theta \log \pi_\theta(a_t|s_t)^T \nabla_\theta \log \pi_\theta(a_t|s_t)\hat{Q}(s_t, a_t)\Big] \\ &= 0\end{aligned} \tag{14}$$

$$\implies b^*(s_t) = \frac{\mathbb{E}_{a_t}\Big[\nabla_\theta \log \pi_\theta(a_t|s_t)^T \nabla_\theta \log \pi_\theta(a_t|s_t)\hat{Q}(s_t,a_t)\Big]}{\mathbb{E}_{a_t}\Big[\nabla_\theta \log \pi_\theta(a_t|s_t)^T \nabla_\theta \log \pi_\theta(a_t|s_t)\Big]} \tag{15}$$

3.2.2 Optimal action-dependent baseline

Denote $g_i(s_t)$ to be the random variable that estimates the policy gradient with action-dependent baseline:

$$g_i(s_t) \coloneqq \nabla_\theta \log \pi_\theta(a_t^i|s_t)\Big(\hat{Q}(s_t,a_t) - b(s_t,a_t^{-i})\Big), a_t \sim \pi_\theta(a_t|s_t) \tag{16}$$

Recall the Assumption 2.3.2:

$$\nabla_\theta \log \pi_\theta(a_t^i|s_t)^T \nabla_\theta \log \pi_\theta(a_t^j|s_t) \equiv z_i^T z_j = 0, \forall i \neq j \tag{17}$$

Under the above assumption, the total variance of the policy gradient with action-dependent baseline is:

$$\begin{aligned}
\text{Var}(\sum_{i=1}^m g_i) &= \sum_i \text{Var}(g_i) + \sum_i \sum_{j\neq i} \text{Cov}(g_i, g_j) \\
&= \sum_i \text{Var}(g_i) + \sum_i \sum_{j\neq i} \mathbb{E}_{a_t}[g_i^T g_j] - \mathbb{E}_{a_t}[g_i]^T \mathbb{E}_{a_t}[g_j] \\
&= \sum_i \text{Var}(g_i) + 0 - \sum_i \sum_{j\neq i} \mathbb{E}_{a_t}[g_i]^T \mathbb{E}_{a_t}[g_j] \\
&= \sum_i \text{Var}(g_i) - \sum_i \sum_{j\neq i} M_{ij}
\end{aligned} \tag{18}$$

where we denote the mean correction term $M_{ij} \coloneqq \mathbb{E}_{a_t}\Big[z_i\hat{Q}(s_t,a_t)\Big]^T \mathbb{E}_{a_t}\Big[z_j\hat{Q}(s_t,a_t)\Big]$. Also let $M = \sum_i \sum_{j\neq i} M_{ij}$. Note that M does not depend on b_i, and thus does not affect the optimal value.

Observing the above equation, the overall variance is minimized when each component variance is minimized. Therefore, we find variance of g_i:

$$\begin{aligned}
\text{Var}(g_i) &= \mathbb{E}_{a_t}\Big[z_i^T z_i\Big(\hat{Q}(s_t,a_t) - b_i(s_t,a_t^{-i})\Big)^2\Big] \\
&\quad - \mathbb{E}_{a_t}\Big[z_i\Big(\hat{Q}(s_t,a_t) - b_i(s_t,a_t^{-i})\Big)\Big]^T \mathbb{E}_{a_t}\Big[z_i\Big(\hat{Q}(s_t,a_t) - b_i(s_t,a_t^{-i})\Big)\Big] \\
&= \mathbb{E}_{a_t}\Big[z_i^T z_i \hat{Q}(s_t,a_t)^2\Big] \\
&\quad + \mathbb{E}_{a_t^{-i}}\Big[-2b_i(s_t,a_t^{-i})\mathbb{E}_{a_t^i}\Big[z_i^T z_i \hat{Q}(s_t,a_t)\Big] + b_i(s_t,a_t^{-i})^2 \mathbb{E}_{a_t^i}[z_i^T z_i]\Big] - M_{ii}
\end{aligned} \tag{19}$$

We take the derivative of $\text{Var}(g_i)$ with respect to b in order to find $b^*(s_t)$ that minimizes variance:

$$\frac{\partial}{\partial b_i}[\text{Var}(g_i)] = 0 \tag{20}$$

$$\implies b_i^*(s_t,a_t^{-i}) = \frac{\mathbb{E}_{a_t^i}\Big[z_i^T z_i \hat{Q}(s_t,a_t)\Big]}{\mathbb{E}_{a_t^i}\big[z_i^T z_i\big]} \tag{21}$$

4 Conclusion

On the basis of the original paper, I did my best to extend and reorganize the equations used while keeping the original idea of the authors. It's a pity that I didn't have enough time to mention the practical results derived by the authors. There's definitely a lot of room for improvement in this report. For example, assumption 2.3.2 is like coming from nowhere because I did not give a decent justification for the assumption in this report such as in what situations that this assumption would be satisfied. Also, what would the results be like while removing some assumptions. I think these are future research directions I should work on.