Stein Variational Policy Gradient

Chan, Kai-Chieh
Department of Computer Science
National Yang Ming Chiao Tung University
kai9988ckc.cs07@nycu.edu.tw

1 Introduction

The report is talking about Stein Variational Policy Gradient and the selected paper is "Liu et al., Stein Variational Policy Gradient, UAI 2017".And there are two reference papers ,"Liu and Wang,Stein Variational Gradient Descent: A General Purpose Bayesian Inference Algorithm,NIPS 2016 " and "Liu et al.,A Kernelized Stein Discrepancy for Goodness-of-fit Tests and Model Evaluation,ICML 2016".
The main contributions of the paper is that it combine Stein Variational Gradient Descent and Policy Gradient together.In our lecture, we are very familiar to Policy Gradient which is update the policy by the gradients approximated from sample trajectories. However, it may face high variance, slow convergence and insufficient exploration. In Stein Variational Policy Gradient method, it allows simultaneous exploitation and exploration of multiple policies.Instead of learning a single policy, SVPG use a set of "particles" to approximate the policy. And The report will show that how to update those particles to find the optimal policy.

2 Problem Formulation

Using the notations similar to our lecture in this report:

- **value function:** γ is a discount factor, $a_t \sim \pi(a_t|s_t)$ is drawn fome the ploicy, $s_{t+1}\sim P(s_{t+1}|s_t, a_t)$ is generated by the enviroment.

$$V^{\pi}(s_t) = \mathbb{E}_{a_t, s_t+1, \ldots}[\sum_{k=0}^{\infty} \gamma^k r(s_{t+k}, a_{t+k})]$$

$$V^{\pi_\theta}(\mu) = \mathbb{E}_{s_0, a_0, s_1, \ldots}[\sum_{k=0}^{\infty} \gamma^k r(s_k, a_k)]$$

- **KL Divergence:** p, q are two distribution, and KL Divergence can be define as:

$$D(q||p) = \mathbb{E}_{x\sim p(x)}[\log p(x) - \log q(x)]$$

- **kernal function:** $k(\vartheta, \theta)$

Consider the policy parameter θ as the random variable and $p(\theta)$ is the distribution of θ , suppose we have some prior domain knowledge about θ which is $q_0(\theta)$. And the optimization problem is:

$$\max_q \{\mathbb{E}_{\theta\sim q(\theta)}[V^{\pi_\theta}(\mu)] - \alpha D(q||q_0)\}$$

the goal is to maximum the expectation but also make $q(\theta)$ look like prior domain knowledge. Coefficient α is a temperature parameter. When $\alpha \to 0$, $q(\theta)$ will update according to expectation return more, when $\alpha \to \infty$, $q(\theta)$ will update according to prior distribution more.

If we don't have prior domain knowledge, we can use an uninformative prior $q_0(\theta) = const$, and the optimixation problem will become:

$$\max_q \{\mathbb{E}_{\theta\sim q(\theta)}[V^{\pi_\theta}(\mu)] + \alpha H(q)\}, \quad where \quad H(q) = \mathbb{E}_{\theta\sim p(\theta)}[-\log p(\theta)]$$

In the Stein Variational Policy Gradient, we will use Stein Variation Gradient Descent (SVGD) to update our policy. SVGD uses a set of "particles" $\{\theta_i\}_{i=1}^n$ to approximate a distribution, the benefits is that it can explorate many policies simultaneously and deterministic updates. Therefore, at the beginning, will have many policy particles θ_i, and gradually update those θ_i by the result of SVGD and make those particles approximate to optimal solution.

3 Theoretical Analysis

Our optimization problem is maximum $\mathbb{E}_{q(\theta)}[V^{\pi_\theta}(\mu)] - \alpha D(q||q_0)$. Taking the derivative of objective fuunction and setting it to zero, we can show what the optimal policy look like.

$$\begin{aligned}
\nabla_{q(\theta)}(\mathbb{E}_{\theta\sim q(\theta)}[V^{\pi_\theta}(\mu)] - \alpha D(q||q_0)) &= \nabla_{q(\theta)}(\mathbb{E}_{\theta\sim q(\theta)}[V^{\pi_\theta}(\mu)] - \alpha(\mathbb{E}_{\theta\sim q(\theta)}[\log q(\theta) + \log q_0(\theta)])) \\
&= \nabla_{q(\theta)}(\mathbb{E}_{\theta\sim q(\theta)}[V^{\pi_\theta}(\mu) - \alpha\log q(\theta) + \alpha\log q_0(\theta)]) \\
&= \nabla_{q(\theta)} \int_\theta q(\theta)V^{\pi_\theta}(\mu) - \alpha q(\theta)\log q(\theta) + \alpha q(\theta)\log q_0(\theta)\, d\theta \\
&= \int_\theta V^{\pi_\theta}(\mu) - \alpha\log q(\theta) - \alpha + \alpha\log q_0(\theta)\, d\theta \\
&= 0
\end{aligned}$$

The optimal distribution of policy parameter θ is

$$q(\theta) \propto exp(\frac{1}{\alpha}V^{\pi_\theta}(\mu)) \times q_0(\theta)$$

we can interpret $exp(\frac{1}{\alpha}V^{\pi_\theta}(\mu))$ as likelihood function, $q_0(\theta)$ as prior distribution and $q(\theta)$ as posterior, this is equivalent to a Bayesian formulation.

Traditional Bayesian formulations use MCMC to drawn the sample from prior distribution and it may suffer from slow convergence as high variance, therefore, the paper use Stein Variational Gradient Descent for Bayesian inference. SVGD use a set of particles $\{\theta_i\}_{i=1}^n$ approximate the target posterior diestibutions.

Stein's Idendity: Let $p(x)$ be a continuously differentiable density supported on $X \subset \mathbb{R}^d$, and $\phi(x) = [\phi_1(x), \phi_2(x), ..., \phi_d(x)]^T$ a smooth vector function, A_p is Stein's idendity, assuming zero boundary conditions on ϕ we have:

$$\mathbb{E}_{x\sim p}[A_p\phi(x)] = 0, \quad where \quad A_p\phi(x) = \nabla_x \log p(x)\phi(x)^T + \nabla_x\phi(x)$$

proof:

$$\begin{aligned}
\mathbb{E}_{x\sim p}[\nabla_x \log p(x)\phi(x)^T + \nabla_x\phi(x)] &= \mathbb{E}_{x\sim p}[\frac{\nabla_x p(x)}{p(x)}\phi(x)^T + \nabla_x\phi(x)] \\
&= \int_{-\infty}^{\infty} p(x)[\frac{\nabla_x p(x)}{p(x)}\phi(x)^T + \nabla_x\phi(x)]\,dx \\
&= \int_{-\infty}^{\infty} \nabla_x p(x)\phi(x)^T + p(x)\nabla_x\phi(x)\,dx \\
&= \int_{-\infty}^{\infty} \nabla_x p(x)\phi(x)\,dx \\
&= p(x)\phi(x)|_{-\infty}^{\infty} \\
&= 0
\end{aligned}$$

(In "Liu and Wang,Stein Variational Gradient Descent: A General Purpose Bayesian Inference Algorithm,NIPS 2016" ,page 2. $A_p\phi(x) = \phi(x)\nabla_x log\ p(x)^T + \nabla_x\phi(x)$ should be corrected to $A_p\phi(x) = \nabla_x log\ p(x)\phi(x)^T + \nabla_x\phi(x)$)

Stein's Discrepancy: Let $q(x)$ be another density, the magnitude of $\mathbb{E}_{x\sim p}[A_p\phi(x)]$ relates to how different p and q are, and can be define as Stein discrepancy $\mathbb{S}(q,p)$, by considering the maximum violation of Stein's identity for ϕ in some proper function set *F*:

$$\mathbb{S}(q,p) = \max_{\phi\in F}\{[\mathbb{E}_{x\sim q} tr(A_p\phi(x))]^2\}$$

Kernelized Stein Discrepancy(KSD) maximizing ϕ in the unit ball of reproducing kernel Hilbert space (RKHS). KSD is defined as:

$$\mathbb{S}(q,p) = \max_{\phi\in H^d}\{[\mathbb{E}_{x\sim q} tr(A_p\phi(x))]^2\}, \quad ||\phi||_{H^d} \le 1$$

Theorem 3.1 Let $T(x) = x + \epsilon\phi(x)$ and $q_{[T]}(z)$ is the density of $z = T(x)$ where $x\sim q(x)$,

$$\nabla_\epsilon D(q,p)|_{\epsilon=0} = -\mathbb{E}_{x\sim q}[tr(A_p\phi(x))]$$

To proof Theorem 3.1, we first show Lemma 3.2

Lemma 3.2 Let q and p be two smooth densities, $T_\epsilon(x)$ is an one-to-one transform on X, $T: X\to X$, $q_{[T]}$ is the density of $z = T_\epsilon(x)$,when $x\sim q$, then we have:

$$\nabla_\epsilon D(q,p)|_{\epsilon=0} = -\mathbb{E}_{x\sim q}[\nabla_{T(x)}\log p(T(x))^T\nabla_\epsilon T(x) + tr(((\nabla_x T(x))^{-1})^T\cdot\nabla_\epsilon\nabla_x T(x))]$$

proof:

Let $z = T(x), x \sim q$, then density of z is:

$$q_{[T]}(z) = q(T^{-1}(z))\cdot|det(\nabla_z T^{-1}(z))|$$

$$q_{[T^{-1}]}(x) = q(T(x))\cdot|det(\nabla_x T(x))|$$

By the change of variable, we have:

$$D(q_{[T]}||p) = D(q||p_{[T^{-1}]}) \quad , and \quad \nabla_\epsilon D(q_{[T]}||p) = -\mathbb{E}_{x\sim q}[\nabla_\epsilon \log p_{[T^{-1}]}(x)]$$

$$
\begin{aligned}
D(q_{[T]}||p) &= D(q||p_{[T^{-1}]}) \\
&= -\mathbb{E}_{x\sim q}[(\log p_{[T^{-1}]}(x) - \log q(x))] \\
\nabla_\epsilon D(q_{[T]}||p) &= -\nabla_\epsilon \mathbb{E}_{x\sim q}[(\log p_{[T^{-1}]}(x) - \log q(x))] \\
&= -\mathbb{E}_{x\sim q}[\nabla_\epsilon \log p_{[T^{-1}]}(x)]
\end{aligned}
$$

then compute $\nabla_\epsilon \log p_{[T^{-1}]}(x)$:

$$
\begin{aligned}
\nabla_\epsilon \log p_{[T^{-1}]}(x) &= \nabla_\epsilon \log(\ p(T(x)) \cdot |det(\nabla_x T(x))|\ ) \\
&= \nabla_\epsilon [\log(\ p(T(x))) + \log(|det(\nabla_x T(x))|)\ ] \\
&= \nabla_{T(x)} \log p(T(x))^T \nabla_\epsilon T(x) + tr(\frac{\nabla_{(\nabla_x T(x))}|det(\nabla_x T(x))|}{|det(\nabla_x T(x))|} \cdot \nabla_\epsilon \nabla_x T(x)) \\
&= \nabla_{T(x)} \log p(T(x))^T \nabla_\epsilon T(x) + tr(\frac{det(\nabla_x T(x)) \cdot ((\nabla_x T(x))^{-1})^T}{det(\nabla_x T(x))} \cdot \nabla_\epsilon \nabla_x T(x)) \\
&= \nabla_{T(x)} \log p(T(x))^T \nabla_\epsilon T(x) + tr(((\nabla_x T(x))^{-1})^T \cdot \nabla_\epsilon \nabla_x T(x))
\end{aligned}
$$

Therefore:

$$
\nabla_\epsilon D(q,p)|_{\epsilon=0} = -\mathbb{E}_{x\sim q}[\nabla_{T(x)} \log p(T(x))^T \nabla_\epsilon T(x) + tr(((\nabla_x T(x))^{-1})^T \cdot \nabla_\epsilon \nabla_x T(x))]
$$

(In "Liu and Wang,Stein Variational Gradient Descent: A General Purpose Bayesian Inference Algorithm,NIPS 2016" ,page 11. I think that it is leak of a Transpose in $(\nabla_x T(x))^{-1}$, so the lemma is $\nabla_\epsilon D(q,p)|_{\epsilon=0} = -\mathbb{E}_{x\sim q}[\nabla_{T(x)} \log p(T(x))^T \nabla_\epsilon T(x) + tr(((\nabla_x T(x))^{-1})^T \cdot \nabla_\epsilon \nabla_x T(x))]$)

By the result of Lemma 3.2. When $T(x) = x + \epsilon\phi(x)$ and $\epsilon = 0$, we have:

- $T(x) = x$
- $\nabla_\epsilon T(x) = \phi(x)$
- $\nabla_x T(x) = I$
- $\nabla_\epsilon \nabla_x T(x) = \nabla_x \phi(x)$

Then rewrite the result of Lemma 3.2. We can get Theorem 3.1:

$$
\begin{aligned}
\nabla_\epsilon D(q,p)|_{\epsilon=0} &= -\mathbb{E}_{x\sim q}[\nabla_{T(x)} \log p(T(x))^T \nabla_\epsilon T(x) + tr(((\nabla_x T(x))^{-1})^T \cdot \nabla_\epsilon \nabla_x T(x))] \\
&= -\mathbb{E}_{x\sim q}[\nabla_x \log p(x)^T \phi(x) + tr((I)^{-1} \cdot \nabla_x \phi(x))] \\
&= -\mathbb{E}_{x\sim q}[tr(\nabla_x \log p(x) \phi(x)^T + \nabla_x \phi(x))] \qquad (1) \\
&= -\mathbb{E}_{x\sim q}[tr(A_p \phi(x))]
\end{aligned}
$$

Lemma 3.3 Consider all the perturbation directions ϕ in the ball $B = \{\phi \in H^d : ||\phi||^2_{H^d} \leq \mathbb{S}(q,p)\}$ of RKHS H^d, the directions that maximizing the negative gradient is Theorem 3.1 is :

$$
\phi^*_{q,p}(\cdot) = \mathbb{E}_{x\sim q}[k(x,\cdot)\nabla_x \log p(x) + \nabla_x k(x,\cdot)]
$$

the negative gradient in Theorem 3.1 is similar to KSD.

Stein Variational Gradient Descent(SVGD): SVGD iteratively updates particles to decrease the KL divergence between the particles θ_i and the targets distribution q. Each update can be written as $\theta_i \leftarrow \theta_i + \epsilon\phi^*(\theta_i)$. And we choose maximumly decrease the KL divergence, so that ϕ^* is :

$$\phi^* \leftarrow \max_{\phi \in H}\{-\frac{d}{d\epsilon}D(\rho[\epsilon\phi]||q)\}, \quad ||\phi||_H \leq 1$$

where $\rho[\epsilon\phi]$ denotes the distribution of $\theta' = \theta + \epsilon\phi^*(\theta)$. Let $k(x, x')$ is the positive definite kernel associated with the RKHS. We have show that this optimization problem have a closed form solution,

$$\phi^*(\theta) = \mathbb{E}_{\vartheta\sim\rho}[\nabla_\vartheta \log q(\theta)k(\vartheta, \theta) + \nabla_\vartheta k(\vartheta, \theta)]$$

By sampling, we can get the empirical averaging $\hat{\phi}$,

$$\hat{\phi}(\theta) = \frac{1}{n}\sum_{j=1}^{n}[\nabla_{\theta_j} \log q(\theta_j)k(\theta_j, \theta_i) + \nabla_{\theta_j} k(\theta_j, \theta_i)]$$

Stein Variational Policy Gradient: Back to our optimization problem, we have show that the optimal distribution of policy parameter θ is $q(\theta) \propto exp(\frac{1}{\alpha}V^{\pi_\theta}(\mu)) \times q_0(\theta)$. So we replace $q(\theta)$ with $exp(\frac{1}{\alpha}V^{\pi_\theta}(\mu)) \times q_0(\theta)$ from the above result. And we can get each update is:

$$\hat{\phi}(\theta) = \frac{1}{n}\sum_{j=1}^{n}[\nabla_{\theta_j}(\frac{1}{\alpha}V^{\pi_{\theta_j}}(\mu) + q_0(\theta_j))k(\theta_j, \theta_i) + \nabla_{\theta_j} k(\theta_j, \theta_i)]$$

Then the pseudocode of SVPG is shown below:

Algorithm 1 Stein Variational Policy Gradient

Input: Learning rate ϵ, kernel $k(x, x\prime)$, temperature, initial policy particles $\{\theta_i\}$.
for iteration $t = 0, 1, ..., T$ **do**
  for particle $i = 0, 1, ..., n$ **do**
    Compute $\nabla_{\theta_i} V^{\pi_{\theta_i}}(\mu)$
  end for
  for particle $i = 0, 1, ..., n$ **do**
    $\Delta\theta_i \leftarrow \frac{1}{n}\sum_{j=1}^{n}[\nabla_{\theta_j}(\frac{1}{\alpha}V^{\pi_{\theta_j}}(\mu) + \log q_0(\theta_j))k(\theta_j, \theta_i) + \nabla_{\theta_j} k(\theta_j, \theta_i)]$
    $\theta_i \leftarrow \theta_i + \epsilon\Delta\theta_i$
  end for
end for

$\nabla_\theta V^{\pi_\theta}(\mu)$ can be computed by the method we learn from the lecture, and we can use baseline or advantage function. Temperature α provides a tradeoff between exploitation and exploration.

4 Conclusion

The paper show some experiment and domonstrate that SVPG have better explore compare to original policy gradient methods. And I think this is interest because when I did the programming

homeword in HW2, I found that my agent was leck of explorate and always stuck in some policy which is obviously can be improved, moreover need a lot of time to imporve the policy. Therefore, I think exploration is important and this method give us a different aspect of how to exploration with many particles.

The potential future research directions of this method can be like choose different kernel functions(This paper use Gaussian RBF kernel to do their experiment). And how to choose kernel might be important and need to be discuss. Other potential future research maybe like how to adjust the tradeoff between explorate and exploitation.

After finishing this report, I am grad to have a chance to learn other policy gradient method. Although I have to spend many time to comprehend those formula, I still learn a lot from this theory project.