# Study Week Report 5

## Paper Review

### Regularized Policy Gradients: Direct Variance Reduction in Policy Gradient Estimation

- ACML 2015

### Author


- Tingting Zhao & Jucheng Yang, Tianjin University of Science and Technology
- Gang Niu, The University of Tokyo
- Ning Xie,Tongji University


### Content

- the author stress that in policy gradient , the **estimate of gradient** is of high variance. So they propose a direct method to decrease the vairance, that put the variance of gradient into evalution function as a punishment term. The common policy gradient loss function is $E(R_\tau)$ , and they change it into $E_\theta[R_\tau] - \lambda \cdot Var(\nabla_\theta E_\theta[R_\tau])$.
- Author design three experiments to testify the algorithm : A manmade function , Mountain Car , Stoke Based Rendering System

### Thoughts

- its aim is to decrease the variance of the gradient estimate, and may have similar effect as directly decrease variance.
- the last experiment , **Stoke Based Rendering System**, is impressive

## Two methods for policy gradient

### Solving Linear equation

- $f(x) = f(x_s) + (x - x_s) \cdot \nabla f(x_s) + o(x - x_s)^2$ when $x- > x_s$ we can take it as $f(x) = f(x_s) + (x - x_s) \cdot \nabla f(x_s)$. Thus to calculate $\nabla f(x)$ we just need to get $f(x_s + \delta_i)$ for n times. and we can get $n$ equations with the form $f(x_s + \delta_i) - f(x_s) = \nabla \cdot \delta_i$. calculate its inverse is suffice to get its gradient.

### Sampling

- to calculate $\nabla(\int_\tau p_\theta(\tau) \cdot R_\tau) = \int_\tau p_\theta(\tau) \cdot \nabla log(p_\theta(\tau)) \cdot R_\tau = E(log(p_\theta(\tau)) \cdot R_\tau)$. And just to sample according this formula to get the gradient.

## Proposed Method

### Optimize with hard constraint

First We try to use the KKT to convert the variance limit but failed.

Formalization consider the start point $s_0$, $Max(E(R_{s_0}))$ s.t. $Var(R_{s_0}) \leq C$. By KKT of lagarian , we can have the equal form of set of equation

$\nabla[E(R_{s_0}) - \lambda(Var(R_{s_0}) - C)] = 0$, $\lambda \cdot [Var(R_{s_0}) - C] = 0$, $Var(R_{s_0}) - C \leq 0$

and if we limit the answer to be interior of the possible solution we can relax the condition to $\nabla E(R_{s_0}) = 0$, $Var(R_{s_0}) - C < 0$ However it's not trivial to solve $\nabla E(R_{s_0}) = 0$.

### Penalty Item soft constriant

Then we consider the soft bound

**Notice that different from ordinary Reinforcement Learning, we are now specifically optimizing $E(R_{s_0})$ and $Var(R_{s_0})$ $s_0$ is the start point, and the other states is of no concern to us.**

Now we try to optimize $max(E(R_{s_0}) - \lambda\sqrt{E(R_{s_0}^2) - E(R_{s_0})^2})$

Given a $\epsilon$-greedy $\theta$ parameterized strategy $\pi_\theta$, we directly sample the gradient and then take the gradient of it, Pay attention we are optimizing the $\epsilon$-greedy policy. and we decrease the $\epsilon$ as progressing further.

### Mathmatical Calculations

let the benifit function $J(\theta) = \int_\tau p_\theta(\tau) \cdot R_\tau - \lambda\sqrt{\int_\tau p_\theta(\tau) \cdot R_\tau^2 - (\int_\tau p_\theta(\tau) \cdot R_\tau)^2}$

$\nabla_\theta J(\theta) = \int_\tau (p_\theta(\tau)\nabla log(p_\theta(\tau)) \cdot R_\tau) - \lambda \cdot \frac{\int_\tau (p_\theta(\tau)\nabla log(p_\theta(\tau)) \cdot R_\tau^2) - 2\int_\tau (p_\theta(\tau)\nabla log(p_\theta(\tau)) \cdot R_\tau) \cdot \int_\tau p_\theta(\tau) \cdot R_\tau}{2\sqrt{\int_\tau p_\theta(\tau) \cdot R_\tau^2 - (\int_\tau p_\theta(\tau) \cdot R_\tau)^2}}$

writing into expectation form $\nabla J(\theta) = E(log(p_\theta(\tau)) \cdot R_\tau) - \lambda\frac{E[log(p_\theta(\tau)) \cdot R_\tau^2] - 2E[R_\tau] \cdot E[log(p_\theta(\tau)) \cdot R_\tau]}{2\sqrt{E[R_\tau^2] - E[R_\tau]^2}}$

Thus, we need to sample four quantities in one turn.

**Mathmatical Interpretation for $\lambda$**

- by Chipchoff's inequality $P(|X - E(X)| > \epsilon) < \frac{Var(X)}{\epsilon^2}$
- we take outcome by pessimistic result less than $E(x) - \lambda$ with probalbility at most $\frac{Var(X)}{\epsilon^2}$ this bound is tight without other information besides second momentum

**Designed Experiment**

**Aim of experiment**

To show that use our algorithm will get a better worst-case reward.

1. A man-made MDP with a high-risk act that averagely benefits more. And an act that gain with a lower risk but gain less reward
2. ...