Natural Policy Gradient

Yun-Rui Wu
Department of Computer Science
National Chiao Tung University
tugddr.cs06@nctu.edu.tw

1 Introduction

Please provide a clear overview of the selected paper. You may want to discuss the following aspects:

- The main research challenges tackled by the paper
  There have been a growing interest in direct policy-gradient methods for approximate planning in large Markov decision problems. In this paper, the author provide a method -natural gradient with connection to policy iteration, show that the natural gradient is moving toward choosing a greedy optimal action.
- The high-level technical insights into the problem of interest
  To use a matric based on manifold to minimize the distance which in terms of how much our parameters is adjusted
- The main contributions of the paper (compared to the prior works)
  Shows a drastic efficience improvement using the natural gradient rather than standard gradient
- Your personal perspective on the proposed method
  Using a matrix to replace the origin euclidean distance in standard gradient can decent better the true policy

2 Problem Formulation

- Your notations (e.g. MDPs, value functions, function approximators,...etc)
  Stationary distribution:
  $$\rho(s)^{\pi}$$
  Average reward:
  $$\nabla\eta(\theta) = \sum_{s,a} \rho(s)^{\pi} \nabla\pi(a;s,\theta)\mathcal{Q}^{\pi}(s,a)$$
  Fisher information matrix of this distribution:
  $$\mathcal{F}_s(\theta) \equiv E_{\pi(a;s,\theta)}[\frac{\partial log\pi(a;s,\theta)}{\partial\theta_i}\frac{\partial log\pi(a;s,\theta)}{\partial\theta_j}]$$
- The optimization problem of interest
  Using tradtional(standar) policy gradient may encounter a very serious business that when we update the policy, the updated parameter could be worse, for we know that the parameter updated in $\theta_{new} = \theta_{new} + \alpha\nabla\eta(\theta)$ we may find a proper α make a not worse parameters. This paper is focus on how to constraint the policy action probibility in a small change and make the steepest descent direction based on the underlying structure of the parameter space, that means natural policy gradient update a greedy policy rather than just better policy in standar policy gradient

Preprint. Under review.

- The technical assumptions The environment is a finite MDP, the agent's decision making procedure is characterized by a stochastic policy $\pi(a;s)$, which is the probability of taking action a in state s, and also make the assumption that every policy π is ergodic so that has a well-defined stationary distribution ρ^π

3 Theoretical Analysis

First, we define compatible function approximator $f^\pi(s,a;\omega)$:

$$\psi(s,a)^\pi = \nabla log\pi(a;s,\theta), f^\pi(s,a;\omega) = \omega^T\psi^\pi(s,a)$$

$\tilde{\omega}$ minimize the squared error

$$\epsilon(\omega,\pi) \equiv \sum_{s,a}\rho(s)^\pi\pi(a;s,\theta)(f^\pi(s,a;\omega) - \mathcal{Q}^\pi(s,a))^2 - (1)$$

Since $\tilde{\omega}$ the minimizes the squared error, it satisfies the condition, means $\partial\epsilon(\omega,\pi)/\partial\omega_i$,
that implies: (1) = 0

$$\sum_{s,a}\rho(s)^\pi\pi(a;s,\theta)f^\pi(s,a;\omega) = \sum_{s,a}\rho(s)^\pi\pi(a;s,\theta)\mathcal{Q}^\pi(s,a)$$

left hand side = $\mathcal{F}_s(\theta)\tilde{\omega}$, right hand side = $\nabla\eta$

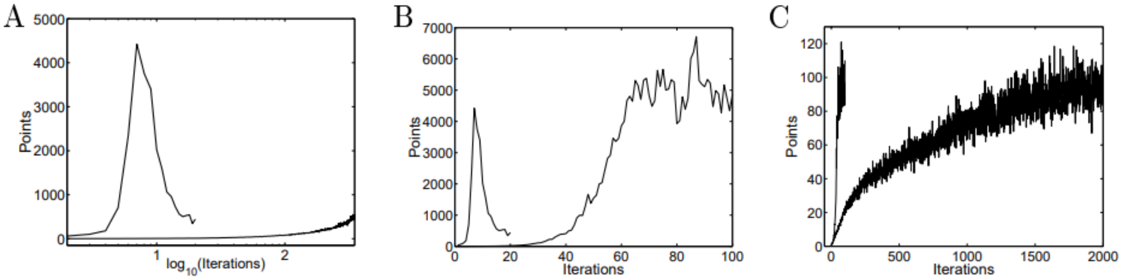
$$\tilde{\omega} = \mathcal{F}_s(\theta)^{-1}\nabla\eta(\theta)$$

Second, we want choose best action: $\pi(s,a;\theta) \propto exp(\theta^T\phi_{sa}), \phi_{sa}$ is some feature vecor in R^m, means the probability manifold of $\pi(s,a;\omega)$ could be curved, so a translation of a point by a tangent vector would not necessarily keep the point on the manifold.
Assume that $\tilde{\nabla}\eta(\theta)$ is non-zero and that $\tilde{\omega}$ minimizes the approximation error.
let $\pi_\infty(a;s) = lim_{\alpha\to\infty}(a;s,\theta+\alpha\tilde{\nabla}\eta(\theta))$, and by the previous result:
$f^\pi(s,a;\omega) = \tilde{\nabla}\eta(\theta)^T\psi^\pi(s,a)$
After a gradient step, $\pi(s,a;\theta+\alpha\tilde{\nabla}\eta(\theta)) \propto exp(\theta^T\phi_{sa} + \alpha\tilde{\nabla}\eta(\theta)^T\psi^\pi(s,a))$
if $\alpha \to \infty$, $\tilde{\nabla}\eta(\theta)^T\psi^\pi(s,a)$ will dominates: means $\pi_\infty(a;s) \neq 0$ if and only if $a \notin argmax_{a'}\tilde{\nabla}\eta(\theta)^T\psi^\pi(s,a)$, so if we update policy by natural policy gradient, we get greedy policy like $f^\pi(s,a;\omega)$
Third, assume that $\tilde{\omega}$ minimizes the approximation error and let the update to the parameter be $\theta' = \theta + \alpha\tilde{\nabla}\eta(\theta)$. Then,
so that, in general, we update policy by natural policy gradient like in Second, but don't need

$$\begin{aligned}
\pi(a;s,\theta') &= \pi(a;s,\theta) + \frac{\partial\pi(a;s,\theta)^T}{\partial\theta}\Delta\theta + O(\Delta\theta^2) \\
&= \pi(a;s,\theta)(1+\psi(s,a)^T\Delta\theta) + O(\Delta\theta^2) \\
&= \pi(a;s,\theta)(1+\alpha\psi(s,a)^T\tilde{\omega}) + O(\alpha^2) \\
&= \pi(a;s,\theta)(1+\alpha f^\pi(s,a;\tilde{\omega})) + O(\alpha^2)\,,
\end{aligned}$$

consider special policy

Experiments:
this paper take a look at the challenging MDP of Tetris in [1]:
The top curve in A is the result in [1], and low curve is standar policy gradient, we can know it's a better point using the specific method in [1] in few iterations. In B, the right curve is using natural policy gradient in this paper, it has a better performance than left curve, method in [1], just do a little



more iterations than [1]. Last, the curve in C, we reduce the dimention of parameter which define in [1], by reduce height to 10, and we observe that, the left curve is natural policy gradient, and right curve is standar policy gradient, them finally reach similar points, however, the left use large amount of iteration to achieve it. This means, maybe standar policy gradient can reach a best performance, but natural policy gradient also does, and faster. This is what natural policy gradient want to prove, since it always take best parameter to update the policy, standar policy gradient, just take better.

4 Conclusion

Please provide succinct concluding remarks for your report. You may discuss the following aspects:

- The potential future research directions
  This 2002's paper, prove the natural policy gradient, can be use in general. and we take a look of $\mathcal{F}_s(\theta)$, in traditional policy gradient, we take parameter varience is $\triangle\theta$, and hope it not be too large. We know KL divergence, also can measure the distribution of $\pi(\theta)$ and $\pi(\theta + \triangle\theta)$, and it's may have a relation between them.

References

D. P. Bertsekas and J. N. Tsitsiklis. Neuro-Dynamic Programming. Athena Scientific, 1996.