A Note on Natural Policy Gradient

Hsin-En, Su
Department of Computer Science
National Chiao Tung University
littlecrazymouse.cs06@nctu.edu.tw

1 Introduction

Policy-gradient methods has been well-received in approximating large Markov decision problems (MDPs). However, the standard gradient is non-covariant, which in other words, not invariant under transformation. This paper provide a natural gradient method which uses a covariant gradient, and draws a connection between natural gradient and policy iteration, proving that by following natural gradient, the policy is actually moving toward selecting greedy actions.

This paper also provide a function approximator that theoretically matches the natural gradient, proving that natural gradient not only works in the steepest direction, it is also a reasonable choice for such compatiable function approximator.

2 Problem Formulation

A finite MDP is a tuple (S, s_0, A, R, P) where S is the finite set of states, s_0 is the start state, A is the finite set of all possible actions, R is the reward function that maps state action pair to a real value $R : S \times A \rightarrow \mathbf{R}$ and P is the transition model that maps state action pair to a distribution of states. The problem concerned here is to train an agent to learn what action to perform in different states, this agent is characterized by a stochastic policy $\pi(a; s)$, which is the probability that the agent will take the action a in state s (The semicolon is used to distinguish the random variables from the parameters of the distribution). Here we assume that every policy π is erogdic, ie has a well-defined stationary distribution ρ^π, where $\rho^\pi(s)$ represent the probability that a certain state s is being visited by following the policy π.

To evaluate a policy's performance, the *average reward* $\eta(\pi)$ is defined as:

$$\eta(\pi) \equiv \sum_{s,a} \rho^\pi(s)\pi(a; s)R(s, a)$$

the state-action value $Q^\pi(s, a)$ is defined as :

$$Q^\pi(s, a) \equiv E_\pi \left\{ \sum_{t=0}^{\infty} R(s_t, a_t) - \eta(\pi) \mid s_0 = s, a_0 = a \right\}$$

where the s_t and a_t are the state and action at time t

and the value function $J^\pi(s)$ is:

$$J^\pi(s) \equiv E_{\pi(a';s)} \{Q^\pi(s, a')\}$$

Now since we've defined our evaluation metrics, we can set our goal as to find a policy π that maximizes the average reward $\eta(\pi)$. We consider the case where the policy is smoothly parameterized by some θ such that the policy $\pi(a; s, \theta)$ is represented as π_θ where $\theta \in \Re^m$.

Preprint. Under review.

A popular approach to solve the problem is *gradient ascent*, and thus we can find the exact gradient of the average reward as:

$$\nabla\eta(\pi_\theta) = \sum_{s,a} \rho^\pi(s)\nabla\pi(a;s,\theta)Q^\pi(s,a)$$

However we want difference of θ between each step to be small, thus we want to find a $d\theta$ that minimizes $\eta(\theta + d\theta)$ under the constraint that $|d\theta|^2$ is small, which $|d\theta|^2$ is defined as $\sum_{i,j} g_{ij}(\boldsymbol{\theta})d\theta_i d\theta_j$ in the nonorthonomal coordinate space. Using vector notation we can rewrite $\sum_{i,j} g_{ij}(\boldsymbol{\theta})d\theta_i d\theta_j$ as $d\theta^T G(\theta)d\theta$, where G is called Riemannian metric tensor.

Theorem 1. The steepest direction of $\eta(\pi_\theta)$ under the constraint that $d\boldsymbol{\theta} = \epsilon\boldsymbol{a}$, $|\boldsymbol{a}|^2 = \sum g_{ij}a_i a_j = a^T G(\theta)a = 1$ is

$$\tilde{\nabla}\eta(\pi_\theta) \equiv G(\pi_\theta)^{-1}\nabla\eta(\theta)$$

Proof:

$$d\boldsymbol{\theta} = \varepsilon\boldsymbol{a}$$

we want to find the $\boldsymbol{a}$ that minimizes $\eta(\pi_{\theta+d\theta})$, by using tylor series approximation we have

$$\eta(\pi_{\theta+d\theta}) \approx \eta(\pi_\theta) + \varepsilon\nabla\eta(\pi_\theta)^T\boldsymbol{a}$$

Solving the problem with constraint $|\boldsymbol{a}|^2 = \sum g_{ij}a_i a_j = 1$ by Lagrangian method, we have:

$$\frac{\partial}{\partial a_i}\left\{\nabla\eta(\pi_\theta)^T\boldsymbol{a} - \lambda\boldsymbol{a}^T G\boldsymbol{a}\right\} = 0$$

which is solved by

$$\nabla\eta(\pi_\theta) = 2\lambda G\boldsymbol{a}$$

thus we have

$$\tilde{\nabla}\eta(\pi_\theta) = G^{-1}\nabla\eta(\pi_\theta)$$

and $\tilde{\nabla}\eta(\pi_\theta))$ is called the natural gradient. As we can see, the normal form of gradient $\nabla\eta(\pi_\theta)$ is simply treating G as the Identity matrix, however this *ad hoc* solution is not ideal, as suggested in Amari [2000], we should define the metric based on the manifold that θ parameterized.

Note: this proof is omitted in the original paper, however I think this is an important proof for readers who first learned about natural gradient.

3 Theoretical Analysis

3.1 Natural Gradient with Fisher Information Matrix

Since we mention that simply picking G as the Identity matrix is not necessariliy ideal for the natrual gradient, in this section we want to present a different way to pick the matrix G and justify the choice we made.

Since the average reward $\boldsymbol{\eta(\pi_\theta)}$ is actually a function on the set of distributions $\boldsymbol{\{\pi_\theta : \theta \in \Re^m\}}$. For each state s, there corresponds a probability manifold, where the distribution $\boldsymbol{\pi(a; s, \theta)}$ is a point on this manifold with coordinates $\boldsymbol{\theta}$. The Fisher information matrix of this distribution $\boldsymbol{\pi(a; s, \theta)}$ is

$$\boldsymbol{F_s(\theta) \equiv E_{\pi(a;s,\theta)}\left[\frac{\partial \log \pi(a;s,\theta)}{\partial\theta_i}\frac{\partial \log \pi(a;s,\theta)}{\partial\theta_j}\right]}$$

The Fisher information matrix is clearly positive definite, and is shown in Amari [2000] that the Fisher information matrix up to a scale, is an invariant metric on the space of the parameters of probability distributions. The distance between two points in the same regardless the choice of the coordinates, unlike $\boldsymbol{G = I}$.

Since the average reward is defined on a set of these distributions, the straightforward choice we make for the metric is:

$$\boldsymbol{F(\theta) \equiv E_{\rho^\pi(s)}\left[F_s(\theta)\right]}$$

and thus the natural gradient we use is defined as:

$$\boldsymbol{\tilde{\nabla}\eta(\theta) \equiv F(\theta)^{-1}\nabla\eta(\theta)}$$

3.2 A compatible Function Approximator

To show that this choice is sensible, we consider the case that $Q^{\pi}(s,a)$ is approximated by some compatible function approximator $f^{\pi}(s,a;\omega)$ parameterized by ω, and we define the function approximator f as:

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

and the squared error $\epsilon(\omega,\pi)$ of the approximator as:

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

Theroem 2. Let $\tilde{\omega}$ minimize the squared error $\epsilon\left(\omega,\pi_{\theta}\right)$. Then

$$\tilde{\omega} = \tilde{\nabla}\eta(\theta)$$

Proof:

Since $\tilde{\omega}$ minimizes the squared error, it satisfies the condition $\partial\epsilon/\partial\omega_i = 0$ which implies

$$\sum_{s,a} \rho^{\pi}(s)\pi(a;s,\theta)\psi^{\pi}(s,a)\left(\psi^{\pi}(s,a)^T\tilde{\omega} - Q^{\pi}(s,a)\right) = 0$$

or equivalently:

$$\left(\sum_{s,a} \rho^{\pi}(s)\pi(a;s,\theta)\psi^{\pi}(s,a)\psi^{\pi}(s,a)^T\right)\tilde{\omega} = \sum_{s,a} \rho^{\pi}(s)\pi(a;s,\theta)\psi^{\pi}(s,a)Q^{\pi}(s,a)$$

By definition of ψ^{π}, $\nabla\pi(a;s,\theta) = \pi(a;s,\theta)\psi^{\pi}(s,a)$ and so the right hand side is equal to $\nabla\eta$. Also by definition of ψ^{π}, $F(\theta) = \sum_{s,a}\rho^{\pi}(s)\pi(a;s,\theta)\psi^{\pi}(s,a)\psi^{\pi}(s,a)^T$ Substitution leads to:

$$F(\theta)\tilde{\omega} = \nabla\eta(\theta)$$

Solving for $\tilde{\omega}$ gives $\tilde{\omega} = F(\theta)^{-1}\nabla\eta(\theta)$, and the result follows from the definition of the natural gradient.

Now we showed that $F(\theta)$ is a reasonable option for the natural gradient when we pick $\omega^T\psi^{\pi}(s,a)$ as our funciton approximator.

3.3 Greedy Policy Improvemnt

To further demonstrate the advantage of using natural gradient with Fisher information matrix instead of an Identity matrix, we show that by following $\tilde{\nabla}\eta(\theta)$, the natural gradient is moving toward the best action rather than just a good action. To show this we first consider the case where policies are in the exponential family such that $\pi(a;s,\theta) \propto \exp\left(\theta^T\phi_{sa}\right)$, where ϕ_{sa} is some feature vector in $\Re^m$. We first show that a sufficiently large step in the direction of the natural gradient $F(\theta)^{-1}\nabla\eta(\theta)$ is equivalent to taking a greedy improvement step.

Theorem 3. For $\pi(a;s,\theta) \propto \exp\left(\theta^T\phi_{sa}\right)$, assume that $\tilde{\nabla}\eta(\theta)$ is non-zero and that $\tilde{\omega}$ minimizes the approximation error. Let $\pi_{\infty}(a;s) = \lim_{\alpha\to\infty}\pi(a;s,\theta+\alpha\tilde{\nabla}\eta(\theta))$ Then $\pi_{\infty}(a;s) \neq 0$ if and only if $a \in \mathrm{argmax}_{a'} f^{\pi}\left(s,a';\tilde{\omega}\right)$

Recall that

$$f^{\pi}(s,a;\tilde{\omega}) = \tilde{\nabla}\eta(\theta)^T\psi^{\pi}(s,a)$$

and

$$\psi(s,a)^{\pi} = \nabla\log\pi(a;s,\theta) \propto \nabla\theta^T\phi_{sa} = \phi_{sa}$$

so we know

$$\mathrm{argmax}_{a'} f^{\pi}\left(s,a';\tilde{\omega}\right) = \mathrm{argmax}_{a'}\tilde{\nabla}\eta(\theta)^T\phi_{sa'}$$

After a gradient step, we have $\pi(a;s,\theta+\alpha\tilde{\nabla}\eta(\theta)) \propto \exp(\theta^T\phi_{sa} + \alpha\tilde{\nabla}\eta(\theta)^T\phi_{sa})$, and if the step size $\alpha\to\infty$, it is clear that the term $\alpha\tilde{\nabla}\eta(\theta)^T\phi_{sa}$ dominates, so that $\pi_{\infty}(a,s) = 0$ if and only if $a \notin \mathbf{argmax}_{a'} \nabla\eta(\theta)^T \phi_{sa'}$. This simply means that the policy will now only choose the best action, as if it was updated by some greedy approach.

Note: this theorem is referred as *theorem 2* in the original paper, however I don't think everything is correctly written, so this proof is a revised version.

Next, we want to have the case in the general parameterized policy. The following theorem shows that the natural gradient is locally moving toward the best action defined by the function approximator.

Theorem 3. Assume that $\tilde{\omega}$ minimizes the approximation error and let the update to the parameter be $\theta' = \theta + \alpha\tilde{\nabla}\eta(\theta)$. Then

$$\pi\left(a; s, \theta'\right) = \pi(a; s, \theta)\left(1 + f^{\pi}(s, a; \tilde{\omega})\right) + O\left(\alpha^2\right)$$

Proof. The change in θ, $\Delta\theta$, is $\alpha\tilde{\nabla}\eta(\theta)$, so by theorem 2, $\Delta\theta = \alpha\tilde{\omega}$. To first order

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

where we have used the definition of ψ and f

4 Conclusion

This paper provide an interesting insight into why we should use natural gradient with Fisher information matrix, As shown in **setion 3.3**, by following such gradient the policy is actually moving toward the greedy policy. However there's two thing that I think that is suitable for future studies. First, which is that choosing a greedy action does not necessarily improve the performance, and many detailed studies have gone into understanding this failure Bertsekas and Tsitsiklis [1996]. Second, this paper provided a form of function approximator that is aligned to the natural gradient, however I wonder if the gradient is compatiable to other different form of function approximators.

References

Shun-ichi Amari. Natural gradient works efficiently in learning. *Neural Computation*, 10, 11 2000. doi: 10.1162/089976698300017746.

Dimitri P. Bertsekas and John N. Tsitsiklis. *Neuro-Dynamic Programming*. Athena Scientific, 1st edition, 1996. ISBN 1886529108.