Proof of Policy Gradient Theorem in Simple Cases

April 15, 2018

1 The Theorem

Let $f(x)$ be any function, and $p(\theta|x)$ be any parameterized distribution over x which is differentiable with respect to θ. The gradient of the expectation of $f(x)$ can be derived as:

$$\begin{aligned}
\nabla_\theta \mathbb{E}_{x\sim p(\theta|x)}[f(x)] &= \nabla_\theta \int_\Omega f(x)p(\theta|x)dx \\
&= \int_\Omega f(x)\nabla_\theta p(\theta|x)dx \\
&= \int_\Omega f(x)\frac{\nabla_\theta p(\theta|x)}{p(\theta|x)}p(\theta|x)dx \\
&= \int_\Omega f(x)\nabla_\theta(\log p(\theta|x))p(\theta|x)dx \\
&= \mathbb{E}_{x\sim p(\theta|x)}[f(x)\nabla_\theta(\log p(\theta|x))] \\
&\simeq \frac{1}{N}\sum_{i=1}^{N} f(x_i)\nabla_\theta(\log p(\theta|x_i))
\end{aligned}$$

where Ω is the domain of x. The last line means that we approximate the expectation by sampling in practice. $x_i(i = 1, 2, \cdots N)$ are independent samples drawn from $p(\theta|x)$.

2 With Baseline

Adding an arbitrary baseline b which is not a function of x or θ to $f(x)$ does not change the gradient, because:

$$\begin{aligned}
\nabla_\theta \int_\Omega (f(x)+b)p(\theta|x)dx &= \nabla_\theta \int_\Omega f(x)p(\theta|x)dx + b\nabla_\theta \int_\Omega p(\theta|x)dx \\
&= \nabla_\theta \int_\Omega f(x)p(\theta|x)dx + b\nabla_\theta 1 \\
&= \nabla_\theta \int_\Omega f(x)p(\theta|x)dx
\end{aligned}$$

Using an appropriate baseline can often reduce the variance of the estimation of the gradient. However, theoretically deriving an optimal baseline is difficult, so in our work we simply come up with a baseline by intuition, which turned out to work relatively well.