The Law of Large Numbers and Policy Gradients

David Meyer

dmm@{1-4-5.net,uoregon.edu,...}

Last update: June 14, 2018

1 Introduction

The *Strong Law of Large Numbers* (LLN) is usually stated as follows:

Let $x_1, x_2, \ldots, x_M$ be a sequence of independent and identically distributed (i.i.d) random variables, each having a finite mean $\mu_i = E[x_i]$.

Then with probability one

$$\frac{1}{M} \sum_{i=1}^{M} x_i \to E[x] \tag{1}$$

as $M \to \infty$.

A complementary theorem, Ergodic Theorem, is stated as follows: Let $\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(M)}$

be M samples from a Markov chain that is *aperiodic*, *irreducible*, and *positive recurrent*[1], and $E[g(\theta)] < \infty$.

Then with probability one

$$\frac{1}{M} \sum_{i=1}^{M} g(\theta_i) \to E[g(\theta)] = \int_{\Theta} g(\theta)\, \pi(\theta)\, d\theta \tag{2}$$

as $M \to \infty$ and where π is the stationary distribution of the Markov chain.

[1] In this case, the chain is said to be *ergodic*.

2 The LLN and Likelihood Ratio Policy Gradients

Suppose that $r(x)$ is a performance measure that depends on some random variable X, and $q(x;\theta)$ is the is the probability that $X = x$, parameterized by $\theta \in \mathbb{R}^K$. Under mild regularity conditions, the gradient with respect to θ of the expected performance $\eta(\theta)$ can be seen to be the following:

$$\eta(\theta) = \mathbb{E}_{x\sim q(xl\theta)}[r(x)] \qquad \text{\# definition of } \eta(\theta) \tag{3}$$

$$= \sum_x r(x) \cdot q(x;\theta) \qquad \text{\# definition of expectation} \tag{4}$$

$$\nabla\eta(\theta) = \sum_x r(x) \nabla_\theta q(x;\theta) \qquad \text{\# take the derivative of both sides} \tag{5}$$

$$= \sum_x r(x) \frac{\nabla_\theta q(x;\theta)}{q(x;\theta)} q(x;\theta) \qquad \text{\# multiply by } 1 = \frac{q(x;\theta)}{q(x;\theta)} \tag{6}$$

$$= \mathbb{E}_{x\sim q(x;\theta)}\, r(x) \frac{\nabla_\theta q(x;\theta)}{q(x;\theta)} \qquad \text{\# definition of expectation} \tag{7}$$

So our gradient $\nabla_\theta \eta(\theta) = \mathbb{E}_{x\sim q(x;\theta)}\, r(x) \frac{\nabla_\theta q(x;\theta)}{q(x;\theta)}$, which means we can estimate the expectation (gradient) with

$$\hat{\eta}(\theta) = \frac{1}{N} \sum_{i=1}^{N} r(x) \frac{\nabla_\theta q(x;\theta)}{q(x;\theta)}$$

Now, given the law of large numbers we know

$$\hat{\eta}(\theta) \to \eta(\theta) \text{ with probability one}$$

This means our gradient estimator ($\hat{\eta}(\theta)$) is *unbiased* since its expected value equals the true gradient. Specifically:

$$\mathbb{E}[\hat{\eta}(\theta)] = \nabla\eta(\theta) \tag{8}$$