Examples paper 4 notes

Nicholas Wong

1 Monte Carlo integration

The focus of Monte Carlo methods is usually to estimate expectations. The expectation is defined as

$$F := \mathbb{E}_{X\sim p(x)}[f(X)] := \int f(x)\ p(x)\ dx. \tag{1}$$

That is, it is the average value of some function of a random variable X over its domain, weighted by the probability distribution $p(x)$. Put another way, it is the value of $f(X)$ when X is distributed according to the probability density function $p(x)$. Sometimes we just write $\mathbb{E}_X[f(X)]$, $\mathbb{E}_p[f(X)]$ or $\mathbb{E}[f(X)]$ when it is clear what we mean.

I hope it is then clear why if we do the following:

- Sample N values from the probability distribution $p(x)$ to get $x^{(1)}$, $x^{(2)}$, ..., $x^{(N)}$.
- Calculate the *empirical mean* or *sample mean*

$$\widehat{F} = \frac{1}{N}\sum_{i=1}^{N} f(x^{(i)}). \tag{2}$$

We should get an *estimate* of the expectation $F = \mathbb{E}[f(X)]$. This is called the Monte Carlo (MC) estimator.

This estimate asymptotically approaches the true value of F when we tend $N \to \infty$ because of the *Law of Large Numbers*[1]. On the other hand, if you take the expectation of $\widehat{F}$, noting that each of the $x^{(i)}$ can be thought as a random variable in itself distributed according to $p(x)$,

$$\mathbb{E}[\widehat{F}] = \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N} f(x^{(i)})\right] = \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left[f(x^{(i)})\right] = \frac{1}{N}\sum_{i=1}^{N} F = F. \tag{3}$$

This is what we call an *unbiased* estimator—its expectation is equal to the true value. This is one advantage of the MC estimator.

Note however that it is possible to have a *biased* estimator that nonetheless approaches the true value when $N \to \infty$, and sometimes we might even prefer biased estimators!

Another advantage of the MC estimator is its scaling with dimensionality. Think about numerical integration methods such as (finite) Riemann integration, trapezoid rule, Simpson's rule etc.[2] When x is one-dimensional, all is fine. However, the number of "bins" you need scales exponentially (cf. slide 5 in handout 3). MC estimators do not suffer from this problem: it does not care about how many dimensions x has. (In fact, the convergence rate of $\widehat{F}$ to F is $\sim 1/\sqrt{M}$)

[1] This law comes in the *strong* and *weak* flavours corresponding to different definitions of convergence.

[2] Aside in the form of a shameless plug: there are much better integration rules, such as Gauss quadrature, sparse grid integration etc. which potentially scales sub-exponentially with dimensions. These have elegant connections with *orthogonal polynomials*, which is the focus of equadratures, a software package I have been involved with heavily in my studies.

2 Importance sampling

Sometimes, the MC estimator is not ideal. For example, it may be difficult to sample from the distribution $p(x)$; nonetheless, it is not too difficult to evaluate $p(x)$ if you give me x. This rather bizarre sounding scenario is actually a really common issue in Bayesian inference. A concrete example is a massively simplified version of the stochastic volatility model

$$y = \varepsilon \exp(x), \tag{4}$$

where we imagine x as being a "hidden" variable and we can only observe y. Having a Gaussian prior on x (and knowing that ε is Gaussian distributed), the posterior distribution (likelihood × prior) of $p(x|y)$ is nonetheless non-Gaussian, but it is quite straightforward to evaluate $p(x|y)$ given a value of x and y.

To calculate F, The idea behind importance sampling is then

- Come up with a *trial* distribution $q(x)$ which is easy to sample from (in addition to being able to evaluate it).
- Sample N values from $q(x)$ to get $x^{(1)}$, $x^{(2)}$, ..., $x^{(N)}$.
- Evaluate
$$\bar{F} = \frac{1}{N}\sum_{i=1}^{N} f(x^{(i)})\frac{p(x^{(i)})}{q(x^{(i)})} \tag{5}$$

Again, we have convergence ($\bar{F} \to F$ as $N \to \infty$) and unbiasedness (Try to show this. What is the distribution of $x^{(i)}$?)

How do you choose $q(x)$? Often it boils down to choosing one which you can programmatically sample from (that is, one of the distributions from scipy.stats). Think about what other good qualities $q(x)$ should have (similarity to $p(x)$?)

2.1 Example

Now, an example where importance sampling can help. Imagine an artificial situation involving an incredibly biased coin, where we want to estimate $F = \mathbb{E}_P[X]$ where

$$P(X=1) = \frac{1}{1000}, \quad P(X=0) = \frac{999}{1000}, \tag{6}$$

where 1 represents heads and 0 represents tails, and P would be the probability mass function. Of course, we know the true value of $\mathbb{E}[X]$, which is 1/1000. However, let's pretend we don't and estimate it by MC. This involves flipping this coin for N times, counting the number of heads and then calculating

$$\widehat{F} = \frac{\text{number of heads of baised coin}}{N}. \tag{7}$$

However, actually doing this is quite inefficient, because one can imagine the coin mostly getting tails whenever you flip it. You might flip the coin hundreds of times and your MC estimate would still be 0, not a very good estimate!

Imagine instead trying to evaluate an importance sampling estimator using a *fair coin* with probability mass function Q:

$$Q(Y=1) = \frac{1}{2}, \quad Q(Y=0) = \frac{1}{2}. \tag{8}$$

Now,

$$\mathbb{E}_P[X] = \mathbb{E}_Q\left[Y\frac{P(Y)}{Q(Y)}\right] \quad \text{(No approximations here)} \tag{9}$$

the importance sampling estimator is thus

$$\bar{F} = \frac{1}{N}\sum_{i=1}^{N} y^{(i)}\frac{p(y^{(i)})}{q(y^{(i)})} = \frac{1}{N}\left(\sum_{y^{(i)}=1} 1 \cdot \frac{1/1000}{1/2} + \sum_{y^{(i)}=0} 0 \cdot \frac{999/1000}{1/2}\right) \tag{10}$$

So,

$$\bar{F} = \frac{\text{number of heads of \textit{unbiased} coin}}{N} \times \frac{1}{500}. \tag{11}$$

To see how much of a difference this makes, see the notebook in the Github repo!

Finally, revisit Q7 of the examples paper and convince yourself that this is a *biased* estimator. However, it still converges so it's fine, especially if you can only evaluate the probability densities up to a proportionality constant. The stochastic volatility model suffers from this problem actually.