

## Bayesian inference

The general problem we face in Bayesian inference is to compute the expectation of some function with respect to a probabilistic model with unknowns.

In the simplest case, we want the expectation of a single variable. In more complex cases, we want an expectation of a complex function of all variables.

Suppose  $p(\theta)$  is our distribution of interest (some  $\theta$  will be known, some not). We want:

$$Z = \mathbb{E}_{\theta \sim p} [f(\theta)] = \int d\theta p(\theta) f(\theta)$$

For some cases this integral will be available in closed form (eg., HMMs). For many (most?) cases, however, it will not.

Lets say that  $\theta$  is discrete, univariate. Then we can compute the expectation by just summing over all possible values. Obviously, though, this won't scale well to high-dimensional or non-discrete variables. But let's see what happens if we try...

### Integration by Summation

Suppose  $\theta$  is univariate, bounded continuous. Wlog,  $\theta \in [0, 1]$ . If we remember how we first learned integration, we can break  $[0, 1]$  into  $R$  equally-sized rectangles. Then, we have:

$$Z \approx \sum_{i=0}^{R-1} \frac{1}{R} p(i/R) f(i/R)$$

As  $R \rightarrow \infty$ ,  $Z$  becomes increasingly more accurate.

One way of thinking about this is that we have a set  $S$  containing  $R$ -many equally spaced points, and the integral is approximated by:

$$Z \approx \frac{1}{|S|} \sum_{\theta \in S} p(\theta) f(\theta)$$

Unfortunately, if  $\theta$  is  $D$ -dimensional, then we need to sum  $R^D$  values of  $\theta$ .

### Uniform Sampling

Instead of spacing  $\theta \in S$  evenly, let's space them randomly. This is the idea of "Monte Carlo" integration, which essentially means "randomized" integration. Uniform sampling is the simplest case. Let  $S$  be a random sampling of  $\theta$ s. Then, we still have:

$$Z \approx \frac{1}{|S|} \sum_{\theta \in S} p(\theta) f(\theta)$$

This scales better computationally, but still the number of samples required to guarantee that we get a close approximation is huge.

It's worth thinking about how hard this problem is. Think of a boat on a lake. We want to estimate the volume of the lake, but cannot see the bottom. We can drive the boat to any position in the lake and drop an anchor, thereby measuring the depth there. How can we approximate the volume? Uniform sampling says to drive randomly around the lake, dropping at the flip of a coin. But there are many chases in which we can do better.

### Importance Sampling

Here we use prior knowledge in the form of a helper distribution  $q$  that we expect to be “similar” to  $p$  and from which we *can* sample. It must have the same “support” as  $p$  (i.e., not zero too often). Then, we compute:

$$\begin{aligned} Z &= \mathbb{E}_{\theta \sim p}[f(\theta)] \\ &= \int d\theta p(\theta) f(\theta) \\ &= \int d\theta q(\theta) \frac{p(\theta)}{q(\theta)} f(\theta) \\ &= \mathbb{E}_{\theta \sim q} \left[ \frac{p(\theta)}{q(\theta)} f(\theta) \right] \end{aligned}$$

So instead of computing an expectation wrt  $p$ , we compute wrt  $q$ . And then we weight each example.

### Rejection Sampling

The idea in rejection sampling is similar to importance sampling. Let  $q$  be a proposal distribution that satisfies  $p(x) \leq Mq(x)$  for  $M < \infty$ . Now, draw points from  $q$  and accept them with probability  $p(x)/[Mq(x)]$ . Compute expectations only over the accepted points.