11 Sampling Methods

Definition. ***Sampling***

1. ***Motivation*** *Approximate inference methods on models where exact inference is intractable. We consider approximate inference based on numerical sampling, i.e. Monte Carlo stuff*

2. ***Goal*** *Computing expectation of some function $f(\mathbf{z})$ with respect to some probability distribution $p(\mathbf{z})$ which is usually too complex to be evaluated analytically*

$$\mathbb{E}\{f\} = \int f(\mathbf{z})p(\mathbf{z})d\mathbf{z}$$

3. ***Approach*** *Obtain a set of samples $\mathbf{z}^{(l)}$ where $l = 1, \cdots, L$ drawn independently from distribution $p(\mathbf{z})$. We then approximate the expectation by a finite sume*

$$\hat{f} = \frac{1}{L} = \sum_{l=1}^{L} f(\mathbf{z}^{(l)}) \qquad \textbf{\textit{where}} \qquad \mathbb{E}\left\{\hat{f}\right\} = \mathbb{E}\{f\} \quad var\left\{\hat{f}\right\} = \frac{1}{L} var\{f\}$$

Definition. ***Baysian Monte Carlo***

1. ***Goal*** *Want to approximate $\mathbb{E}_{\theta\sim p(\theta)}\{f(\theta)\}$. But we cannot compute $p(\theta)$ directly, instead we can compute $g = \frac{p}{c}$ for some normalizing constant c. We want to find*

$$\mathbb{E}\{f\} = \frac{\int f(\theta)g(\theta)d\theta}{\int g(\theta)d\theta}$$

2. ***Idea*** *Generate a set of samples $\{\ \theta_i\ \}_{i=1}^{N}$ from $p(\theta)$ and*

$$\mathbb{E}\{f\} \simeq \frac{1}{N}\sum_{i=1}^{N} f(\theta_i)$$

3. ***Context of Bayesian*** *f is the prediction. p is some posterior distribution we cannot evaluate directly. $g = p(\mathcal{D}|\theta)p(\theta)$ is product of prior and likelihood which is computable, we can then sample*

$$p = p(\theta|\mathcal{D}) \overset{bayes}{=} \frac{p(\mathcal{D}|\theta)p(\theta)}{\int p(\mathcal{D}|\theta)p(\theta)d\theta} = \frac{g(\theta)}{\int g(\theta)d\theta} = \frac{g}{c} \qquad c \in \mathbb{R}$$

In this case the estimator is unbiased and variance shrinks by a factor of sample size

$$\mathbb{E}_{\theta\sim p}\left\{\hat{f}(\theta)\right\} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_{\theta\sim p}\{f(\theta)\} = \mathbb{E}_{\theta\sim p}\{f(\theta)\}$$

$$var_{\theta\sim p}\left\{\hat{f}(\theta)\right\} = \frac{1}{N^2}\sum_{i=1}^{N} var_{\theta\sim p}\{f(\theta)\} = \frac{1}{N} var_{\theta\sim p}\{f(\theta)\}$$

Definition. ***Ancestral Sampling*** *(8.1.2)*

1. ***Idea*** *Sample $\hat{x}_1, \cdots, \hat{x}_K$ from a joint distribution $p(x_1, \cdots, x_K)$ which factorizes into a directed acyclic graph.*

2. ***Algorithm*** *Order varaibles such that they follow topological ordering. Start with drawfing a sample $\hat{x}_1$ from distribution $p(x_1)$, then work through each node in oder such that for node x_i, we draw sample from $p(x_i|parent(x_i))$, where the parent variables are set to sampled values.*

3. ***Summary*** *We make one pass through set of variables in order $\mathbf{z}_i, \cdots, \mathbf{z}_M$, by sampling from conditional distribution $p(\mathbf{z}_i|parent(\mathbf{z}_i))$ where*

$$p(\mathbf{z}) = \prod_i p(\mathbf{z}_i|parent(\mathbf{z}_i))$$

Definition. ***Sampling from standard distribution*** *Given $z \sim p(z)$ we want to determine a function $f(\cdot)$ such that $y = f(z)$ is a desired distribution. The following relationship holds*

$$p(y) = p(z)\left|\frac{dz}{dy}\right|$$

We usually want to convert a pseudo-random number generator with uniform distribution $[0, 1]$ to a desired distribution. So $p(z) = 1$, goal is to pick f by integration

$$f = h^{-1} \qquad z = h(y) = \int_{-\infty}^{y} p(y)dy$$

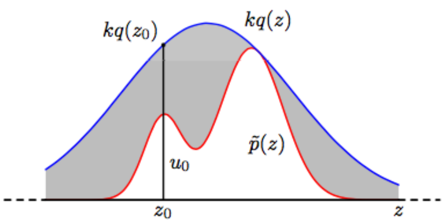
This method has the limitation that it requires inverting a indefinite integral and that we are able to sample from $p(z)$

Definition. ***Rejection Sampling***
Given we cannot sample from $p(\mathbf{z})$ directly. But we do have

1. ***A readily computable function*** *$\tilde{p}(z)$, such that $p(z) = \frac{1}{Z}\tilde{p}(z)$.*

2. ***A proposal distribution*** *$q(z)$, from which we can sample easily*

that satisfies the inequality $kq(z) \geq \tilde{p}(z)$



We generate $z_0 \sim q(z)$ *and* $u_0 \sim Unif[0, kq(z_0)]$. *We reject sample* z_0 *if* $u_0 > \tilde{p}(z_0)$ *otherwise accept the sample. In this way, the accepted samples falls uniformly under the curve of* $\tilde{p}(z)$, *and hence the corresponding accepted samples are distributed according to* $p(z)$. *Note rejection sampling rejects exponentially more samples (hence inefficient) for high dimensional distributions. It also has problem with picking the right* k *and a good upper bound distribution* q.

Definition. ***Metropolis-Hasting Algorithm*** *We want to maintain a record of current state* $\mathbf{z}^\tau$ *and the proposal distribution* $q(\mathbf{z}|\mathbf{z}^{(\tau)})$ *depends on the current state. A sequence* $\{\mathbf{z}^{(1)}, \mathbf{z}^{(2)}, \cdots\}$ *forms a markov chain. Again we want to sample from* $p(\mathbf{z}) = \frac{\tilde{p}\mathbf{z}}{c}$ *for some* $c \in \mathbb{R}$. *At each iteration of the algorithm, we generate candidate* $\mathbf{z}^* \sim q(\mathbf{z}|\mathbf{z}^{(\tau)})$ *and accept the sample with probability*

$$A(\mathbf{z}^*, \mathbf{z}^{(\tau)}) = \min\left(1, \frac{\tilde{p}(\mathbf{z}^*)q(\mathbf{z}^{(\tau)}|\mathbf{z}^*)}{\tilde{p}(\mathbf{z}^{(\tau)})q(\mathbf{z}^*|\mathbf{z}^{(\tau)})}\right)$$

this can be achieved by choosing a random number $u \sim Unif(0,1)$ *and accpet the sample if* $A(\mathbf{z}^*, \mathbf{z}^{(\tau)}) > u$. *In general, the algorithm favors accepting samples with*

1. $\tilde{p}(\mathbf{z}^*) \geq \tilde{p}(\mathbf{z}^{(\tau)})$, *i.e. if transition yields higher value of* $p(\mathbf{z})$
2. $q(\mathbf{z}^{(\tau)}|\mathbf{z}^*) \geq q(\mathbf{z}^*|\mathbf{z}^{(\tau)})$, *i.e. if current state is easier to get back to*

We can prove that distribution of $\mathbf{z}^{(\tau)}$ *tends to* $p(\mathbf{z})$ *as* $\tau \to \infty$. *Notice, the proposal distribution is not that important, as the long term convergence is guaranteed*