Stat210A: Theoretical Statistics **Lecture Date: September 22, 2014**

Recitation 4: Sequential Experiments (cont.) and Sampling

Lecturer: Tamara Broderick *Scribe: Roy Dong*

1 Sequential Experiments

We quickly finish discussing sequential experiments.

Theorem 1 (Theorem 5.4)**.** *Let $X_1, X_2, \cdots \sim f_\theta$ be i.i.d., where f_θ is a density with respect to some measure μ. Let $\{N = n\} = \{(X_1, \ldots, X_n) \in A_n\}$. Then $(N, X_1, \ldots, X_N)$ have joint density:*

$$(n, x_1, \ldots, x_n) \mapsto \left[\prod_{i=1}^{n} f_\theta(x_i)\right] 1_{A_n}(x_1, \ldots, x_n)$$

Remark 2. Note that the event $\{N = n\}$ is independent of X_m for $m > n$ and θ, conditioned on $X_1, \ldots, X_n$.

This statement has the following interpretation. For any $n \in \mathbf{N}$ and measurable $A \subset \mathbf{R}^n$:

$$P(N = n, (X_1, \ldots, X_n) \in A) = \int_{R^n} 1_A(x_1, \ldots, x_n) \left[\prod_{i=1}^{n} f_\theta(x_i)\right] 1_{A_n}(x_1, \ldots, x_n) \mu^n(dx)$$

Here, μ^n is the product measure of μ n times. If $\mu(\mathbf{R}) = 1$ (which can be assumed without loss of generality since $\mathbf{R}$ is σ-finite), this can also be considered as an integration with respect to the measure μ^∞, where this is the product measure of μ countably many times.

Some of this can also be thought of as with respect to the product of the counting measure and μ^∞. Take care when doing this, though, as the dimension of the $(X_1, \ldots, X_N)$ changes as N changes.

Theorem 3 (Theorem 5.4b)**.** *Suppose $f_\theta(x) = e^{\eta(\theta) \cdot T(x) - B(\theta)} h(x)$, where $\eta(\theta)$ is s-dimensional. Then, the joint density of a sequential experiment with stopping time N is:*

$$\exp\left(\eta(\theta) \cdot \sum_{i=1}^{n} T(x) - nB(\theta)\right) \left(\prod_{i=1}^{n} h(x_i)\right) 1_{A_n}(x_1, \ldots, x_n)$$

Remark 4. Note that this is an exponential family as well, with canonical parameters $(\eta_1(\theta), \ldots, \eta_s(\theta), -B(\theta))$. That is, now N is a random variable and the sufficient statistics are given by:

$$\left(\sum_{i=1}^{N} T_1(X_i), \ldots, \sum_{i=1}^{N} T_s(X_i), N\right)$$

Thus, $\tilde{h}(x, n) = \left(\prod_{i=1}^{n} h(x_i)\right) 1_{A_n}(x_1, \ldots, x_n)$. Note that, for this new exponential family, the partition function is trivial, i.e. $\tilde{B}(\theta) = 0$.

Example 5 (Bernoulli)**.** Let:

$$f_\theta(x_i) = \theta^{x_i}(1-\theta)^{1-x_i} = (1-\theta)\exp\left(x_i \log\left(\frac{\theta}{1-\theta}\right)\right)$$

(Here $0 < \theta < 1$ and $x_i \in \{0, 1\}$.) Also, let our stopping time be given by:

$$\{N = n\} = \left\{\sum_{i=1}^{n}(1 - X_i) = r, X_n = 0\right\} = A_n$$

Note that this is the first time we have r failures. (We have r failures, and the most recent observation, X_n, was also a failure.) Then:

$$\begin{aligned} P(N = n, X_1 = x_1, \ldots, X_n = x_n) &= \exp\left(\sum_{i=1}^{n} x_i \log\left(\frac{\theta}{1-\theta}\right) + n\log(1-\theta)\right) 1_{A_n}(x_1, \ldots, x_n) \quad (1) \\ &= \exp\left((n-r)\log\left(\frac{\theta}{1-\theta}\right) + n\log(1-\theta)\right) 1_{A_n}(x_1, \ldots, x_n) \quad (2) \\ &= \theta^{n-r}(1-\theta)^r 1_{A_n}(x_1, \ldots, x_n) \quad (3) \end{aligned}$$

Note that Line 2 holds because we are restricting ourselves to the event A_n. We can marginalize this over X to get:

$$P(N = n) = \binom{n-1}{n-r}\theta^{n-r}(1-\theta)^r$$

The $\binom{n-1}{n-r}$ term is because the last X_n is fixed as a failure, and we must find $n - r$ successes in the previous $n - 1$ trials. Doing so recovers the negative binomial distribution, which was the purpose of this example.

2 Sampling

2.1 Motivation

Recall the definition of Bayes risk:

$$\begin{aligned} R(\delta) &= \int R(\theta, \delta)p(\theta)d\theta \quad (4) \\ &= \int \left(\int L(\theta, \delta(x))p(x|\theta)dx\right)p(\theta)d\theta \quad (5) \\ &= \int \left(\int L(\theta, \delta(x))p(\theta|x)d\theta\right)p(x)dx \quad (6) \end{aligned}$$

This uses Bayes theorem:

$$p(\theta|x) = \frac{p(x|\theta)p(\theta)}{p(x)}$$

Here, $p(\theta|x)$ is called the posterior, $p(x|\theta)$ the likelihood, $p(\theta)$ the prior, $p(x)$ the evidence.

Chaining these together might yield crazy distributions that are hard to sample. One can't just easily use a uniform random variable generator. This motivates sampling theory. Here, we want to compute

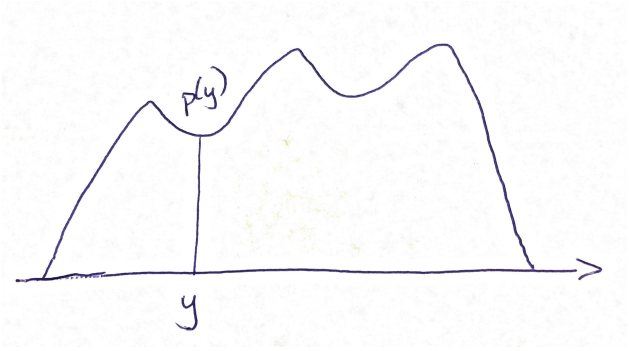
$\int f(\theta)p(\theta|x)d\theta$, which would, for example, give us the moments of θ as a random variable given our observed data. For sampling theory, we consider the more general problem of estimating $\int f(y)p(y)dy$. Here, p is easy to evaluate but hard to sample.

Suppose $Y^{(1)}, \ldots, Y^{(n)} \sim p$ are i.i.d. The Strong Law of Large Numbers (SLLN) gives us that:

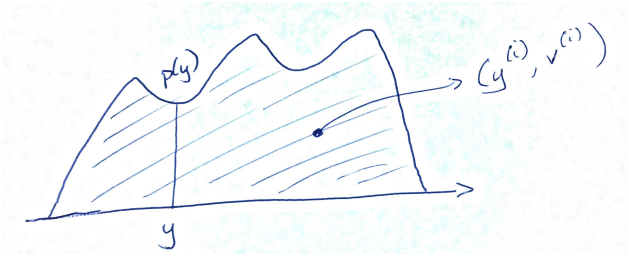
$$\frac{1}{n}\sum_{i=1}^{n} f(Y^{(i)}) \xrightarrow{\text{a.s.}} \int f(y)p(y)dy$$

(Recall that $X_n \xrightarrow{\text{a.s.}} X$ means that $P(\lim_{n\to\infty} X_n = X) = 1$.)

Again, the situation is that p is easy to evaluate but hard to sample.



As a thought experiment, one way to get i.i.d. draws according to p is to sample uniformly underneath in the interior of the density's graph. We'd sample the points $(Y^{(i)}, V^{(i)})$.



Then, the sequence $(Y^{(i)}) \sim p$ is i.i.d. To convince yourself of this, just note that:

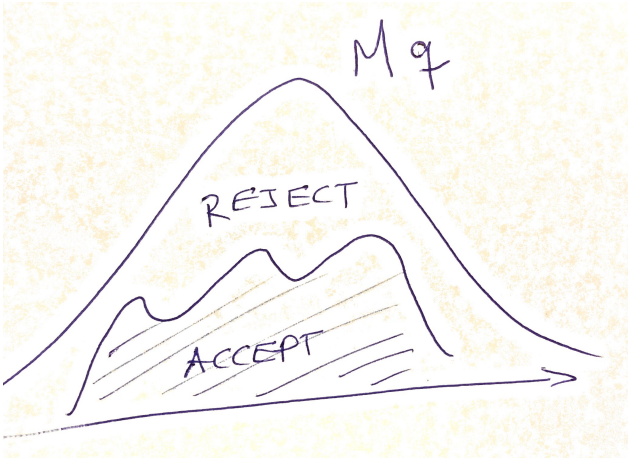
$$P(Y \in A) = \int 1_A(y)p(y)dy = \int 1_A(y)\left(\int_0^{p(y)} 1dv\right)dy$$

We have the uniform distribution underneath the graph.

However, sampling uniformly in the interior is often as difficult as sampling from the original distribution.

2.2 Rejection Sampling

Suppose we have a probability density q that is both easy to sample and can 'enclose' p, i.e. there exists an M such that for all y, $p(y) \leq Mq(y)$. (Note that this enforces that $M \geq 1$.) The intuition is we sample uniformly from the area enclosed by Mq, and accept if the sample is enclosed by p, and reject it otherwise.

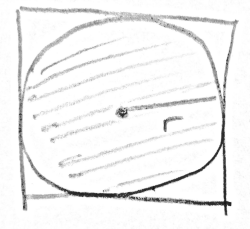


The pseudo-code for rejection sampling can be seen below:

$i \leftarrow 1$;
while $i \leq n$ **do**
  $y^{(i)} \sim q$;
  $u \sim \text{Unif}[0, 1]$;
  if $u < \frac{p(y^{(i)})}{Mq(y^{(i)})}$ **then**
    accept the sample;
    $i \leftarrow i + 1$;
  else
    reject the sample;
  end
end

Note that we could be throwing a way a lot of samples, and this could be wasteful. The acceptance rate is the area of the acceptance region divided by the total area underneath Mq. This is $1/M$. Thus, a smaller M is better.

Consider the example where we wish to sample uniformly in the volume of a sphere. An easy thing to sample from would be an enclosing box. (We can just draw i.i.d. uniform for each coordinate.)



However, the acceptance rate gets arbitrarily bad as the dimension goes to infinity. The volume of an n-dimensional sphere is: $\pi^{n/2}r^n/\Gamma(n/2+1)$, and the volume of an n-dimensional enclosing box is $(2r)^n$. The acceptance rate is thus proportional to $\left(\frac{\sqrt{\pi}}{2}\right)^n$, and

$$\left(\frac{\sqrt{\pi}}{2}\right)^n \to 0$$

Importance sampling helps us deal with the curse of dimensionality.

2.3 Importance Sampling

Denote the value we are estimating with $\phi := \int f(y)p(y)dy$. Consider another estimator:

$$\hat{\phi}_q = \frac{1}{n}\sum_{i=1}^{n} f(Y^{(i)})w(Y^{(i)})$$

Here, as before, $Y^{(i)} \sim q$ is i.i.d. Also, $w(y^{(i)}) = p(y^{(i)})/q(y^{(i)})$.

Is this unbiased? Yes. Consider:

$$E(\hat{\phi}_q) = \int f(y^{(i)})\frac{p(y^{(i)})}{q(y^{(i)})}q(y)dy = \int f(y)p(y)dy = \phi$$

Implicitly, we assume that $p \ll q$ here. Thus, using the SLLN: $\hat{\phi}_q \xrightarrow{\text{a.s.}} \phi$ as $n \to \infty$.

This works for any q such that $p \ll q$. But intuitively, some q should behave better than others. If most of the mass of q is allocated where p has very little mass, then the convergence may not be very good. To compare different distributions q, consider the variance.

$$\text{Var}_q(\hat{\phi}_q) = \frac{1}{n}\text{Var}_q\left(f(Y)\frac{p(Y)}{q(Y)}\right)$$

Thus:

$$n\text{Var}_q(\hat{\phi}_q) = E_q\left(f(Y)\frac{p(Y)}{q(Y)}\right)^2 - \left[E_q\left(f(Y)\frac{p(Y)}{q(Y)}\right)\right]^2 = \int\left(f(y)^2\frac{p(y)^2}{q(y)}\right)dy - \phi^2$$

(Note that one $q(y)$ cancels with the density we integrate with respect to.)

Let's minimize this variance. Note that Jensen's inequality yields:

$$E_q\left(f(Y)\frac{p(Y)}{q(Y)}\right)^2 \geq \left[E_q\left(|f(Y)|\frac{p(Y)}{q(Y)}\right)\right]^2 = \left(\int |f(y)|p(y)dy\right)^2$$

($p \geq 0$ and $q \geq 0$ since they are densities.) This bound is achieved with $q(y) \propto |f(y)|p(y)$. (This is easy to check, noting that the normalizing constant is $\int |f(y)|p(y)dy$.)

Of course, we usually can't choose $q(y) \propto |f(y)|p(y)$ since the normalizing constant is almost exactly the integral we want to solve in the first place, but it's a good rule of thumb for what you should be aiming for.

2.3.1 Unnormalized densities

What if p and/or q are not normalized? Recall Bayes's theorem. The evidence, $p(x)$ can be hard to calculate. q might be easy to sample from but hard to normalize. (For example, suppose $X \sim U[0,1], Y \sim N(X,1), Z \sim \text{Beta}(X,|Y|)$.)

Suppose neither p nor q are normalized. Let $a = \int p(y)dy$ and $b = \int q(y)dy$. Also, suppose $Y^{(i)} \sim q/b$ are i.i.d. Consider a new estimator:

$$\tilde{\phi}_g = \frac{\sum_{i=1}^n f(y^{(i)})w(y^{(i)})}{\sum_{i=1}^n w(y^{(i)})}$$

We can re-write this:

$$\tilde{\phi}_g = \frac{\frac{1}{n}\sum_{i=1}^n f(y^{(i)})\frac{p(y^{(i)})/a}{q(y^{(i)})/b}}{\frac{1}{n}\sum_{i=1}^n \frac{p(y^{(i)})/a}{q(y^{(i)})/b}}$$

The numerator contains the distributions we wanted (normalized), and is just normalized importance sampling. With SLLN:

$$\frac{1}{n}\sum_{i=1}^n f(y^{(i)})\frac{p(y^{(i)})/a}{q(y^{(i)})/b} \xrightarrow{\text{a.s.}} \int f(y)p(y)dy$$

Choosing the special case where $f(y) \equiv 1$, then:

$$\frac{1}{n}\sum_{i=1}^n \frac{p(y^{(i)})/a}{q(y^{(i)})/b} \xrightarrow{\text{a.s.}} 1$$

The denominator converges almost surely to 1. Thus $\tilde{\phi}_g \xrightarrow{\text{a.s.}} \phi$. (We can do this for almost sure convergence because the reasoning is almost entirely real analysis. We must be more careful if we only had convergence in probability or convergence in distribution.)

3 Closing remarks

There are other methods. There is adaptive rejection sampling, which refines q to be closer to p across iterations. This still suffers from the curse of dimensionality. There is also adaptive importance sampling. (Not covered now.)