http://www.math.nyu.edu/faculty/goodman/teaching/MonteCarlo20/

Class notes: Monte Carlo methods
Week 11, Multi-distribution methods, rare event
tentative Jonathan Goodman
Tentative November 25, 2020

1 Rare events

Rare event simulation is the problem of finding the probability of very unlikely events. A direct approach would be to take n samples, take N to be the *hits*, the number of times the event happened, and take

$$\widehat{p} = \frac{N}{n} \ .$$

We saw that the accuracy is approximately

$$\text{relative accuracy} = \frac{\sigma_{\widehat{p}}}{p} \approx \frac{1}{\mathrm{E}[N]^{\frac{1}{2}}} \ .$$

This was assuming independent samples, but the MCMC error estimate is pessimistic too. To understand the formula, if you do $100,000 = 10^5$ samples and get 5 hits, then the accuracy is related to 5, not 10^5. Many applications call for estimating probabilities this small, or for doing other estimations that depend on this with such small probabilities.

A related problem is estimating the evidence integral

$$Z(Y) = \int L(Y|x)\pi(x)\,dx \ . \tag{1}$$

You could do this by sampling the prior (π), but if you have good data that strongly constrain x, then most samples of the prior are poor fits to the data. It is unlikely that a random $X \sim \pi$ resembles the x values that determine Z. We now have reasonably effective ways to sample the posterior

$$\rho(x|Y) = \frac{1}{Z(Y)} L(Y|x)\pi(x) \ .$$

It is ironic that Monte Carlo was invented as a way to estimate integrals, but estimating this integral is harder than sampling the distribution.

This class discusses two approaches to rare event simulation. One class of strategies is importance sampling motivated by *large deviation theory*. Large deviation theory is a group of theorems in probability theory that concern probabilities that depend on a parameter such as $n \to \infty$ or $\epsilon \to 0$. The theorems are motivated by the idea that $p_n \sim e^{-Cn}$ as $n \to \infty$. A typical theorem (see Section 2) says this in the weak form

$$\lim_{n\to\infty} -\frac{1}{n}\log(p_n) = C \ .$$

The proof of a large deviation theorem, like the one in Section 2, usually identifies a "mechanism", which is a description of a typical sample that meets the rare criterion. The hope is that this may be interpreted as an importance function (change of probability distribution) that makes the rare event more likely.

The other broad approach seeks very rare events as being the result of a sequence of reasonably likely events. You can view this as a sequence of hurdles that an event much go over to be in the target class. A simple example is in Section 3. These seem preferable to methods derived from large deviation theory because you don't have to start out with such a clear idea of the mechanism. Unfortunately, this is debatable. If you don't understand the mechanism, you are likely to create a sequence of hurdles that leads to unlikely rare events, events that are not typical of the rare event class you are looking for.

The method of Section 3 is an example of methods that use not just one or two probability distributions, but a sequence of them, ρ_k. For a well chosen sequence, neighboring distributions ρ_k and ρ_{k+1} have a reasonable "overlap", in one of several senses. This means that it is not hard to go from ρ_k to ρ_{k+1}. Two instances of this idea are *thermodynamic integration* for estimating integrals like the evidence integral, Section 4, and *simulated tempering* or *parallel tempering* of Section 5

2 Cramer's theorem

Cramer's theorem describes the probability that the sample mean of n i.i.d. random variables is completely wrong. More precisely, suppose $f(x)$ is the PDF of a one component random variable and

$$\mathrm{E}_f[X] = \mu \; .$$

The sample mean is

$$\overline{X}_n = \frac{1}{n} \sum_{k=1}^{n} X_k \; .$$

Take $a > \mu$ and ask for

$$p_n = \Pr\big[\, \overline{X}_n \geq a \big] \; .$$

Cramer's theorem is an approximate formula for P_n for large n, which applies if X has "light tails".

We give a derivation of the approximate formula for p_n that is similar to the usual derivation but has the advantage that is also identifies the prefactor. The prefactor contains a power of n, so the estimate of p_n cannot be accurate without it. The idea behind these proofs is to modify the distribution with an exponential *twise* (not a good term, but it stuck). The twisted distribution has expected value a. The prefactor comes from the central limit theorem.

The "twisted" distribution is the PDF

$$f_\lambda(x) = \frac{1}{Z(\lambda)} \, e^{\lambda x} \, f(x) \; . \tag{2}$$

The normalization factor is

$$Z(\lambda) = \int_{-\infty}^{\infty} e^{\lambda x} f(x)\, dx \ . \tag{3}$$

Cramer theory works as long as $Z(\lambda) < \infty$. This happens if $f(x)$ goes to zero at a fast enough exponential rate for the integral (3) to converge. Otherwise the probability p_n behaves in a different way and typical samples for $a > \mu$ have a different character. We define the un-normalized sum as

$$S_n = X_1 + \cdots + X_n \ . \tag{4}$$

The PDF will be $g_n(s)$, so $S_n \sim g_n(\cdot)$. The PDF is given by

$$g_n(s) = \int \cdots \int \delta(s - x_1 - \cdots - x_n)\, f(x_1) \cdots f(x_n)\, dx_1 \cdots dx_n \ . \tag{5}$$

We will use the central limit theorem and the twisted density (2) to estimate $g_n(na)$. Then one of our approximate integration methods (which we called "Watson's lemma") gives the approximate probability. The trick is to choose λ so that a is the expected value in the twisted distribution

$$a = \mathrm{E}_\lambda[X] = \frac{1}{Z(\lambda)} \int_{-\infty}^{\infty} x e^{\lambda x} f(x) dx \ . \tag{6}$$

Call this number $\lambda_*(a)$. It is "easy to see' that if λ_* exists, and if the problem is not trivial, the it is unique.

The trick is to see that the integrand in (5) is equal to zero unless $s = x_1 + \cdots + x_n$. Therefore, in the integral,

$$1 = e^{-s + x_1 + \cdots + x_n} = e^{-s} e^{x_1} \cdots e^{x_n} \ .$$

We also multiply by n factors of $Z(\lambda)\frac{1}{Z(\lambda)}$. Finally, we use/abuse the common convention

$$F(\lambda) = \log(Z(\lambda)) \ , \quad Z^n = e^{-nF(\lambda} \ .$$

The result is

$$g_n(s) = e^{-s + nF(\lambda)} \int \cdots \int \delta(s - x_1 - \cdots - x_n)\, f_\lambda(x_1) \cdots f_\lambda(x_n)\, dx_1 \cdots dx_n \ .$$

This is true for any λ. We rewrite it and put in λ_*:

$$g_n(an) = e^{-(a - F(\lambda_*))n} \int \cdots \int \delta(an - x_1 - \cdots - x_n)\, f_{\lambda_*}(x_1) \cdots f_{\lambda_*}(x_n)\, dx_1 \cdots dx_n \ . \tag{7}$$

Here is the point of the choice of λ_*. The PDF f_{λ_*} has mean value a and the integral on the right is the probability density that na is exactly the sum of of the X_k. For large n, the PDF of $\sum X_k$ is approximately normal (in the twisted f_{λ_*} distribution) with mean an and variance $n\sigma^2_{\lambda_*}$, with

$$\sigma^2_{\lambda_*} = \int (x - a)^2 f_{\lambda_*}(x)\, ds \ . \tag{8}$$

The Gaussian PDF consists only of the pre-factor when evaluated at its mean. The pre-factor, for variance $n\sigma^2_{\lambda_*}$, is

$$\frac{1}{\sqrt{2\pi n\sigma_{\lambda_*}}} \; .$$

Therefore,

$$g_n(an) \approx \frac{1}{\sqrt{2\pi\sigma^2_{\lambda_*(a)}}} \, \frac{1}{\sqrt{n}} \, e^{-n(a-F(\lambda_*))} \; . \tag{9}$$

Exercise 1 asks you to estimate p_n from this.

I don't know how Cramer came to the idea of exponential twist. You can see the structure in examples. A physicist might find the exponential twist (2) natural. If you want $\mathrm{E}[X]$ to move from μ to a, you have to "pull", which means apply a force. The potential corresponding to a force λ is λx. The probability p_n is an integral over the set $\sum x_k \geq an$. An approximate formula for p_n would be an approximate integration method. You can view the central limit theorem as an approximate integration method, and one of the only approximate integration methods that gets better when the dimension increases. Our other methods, Watson's lemma and the Laplace method, require care in high dimensions.

The proof suggests an importance sampling method with

$$(X_1, \cdots, X_n) \sim \prod_{k=1}^{n} f_{\lambda_*}(x_k) \; . \tag{10}$$

The twisted estimator is: first make M samples from the twisted distribution (10), then use the twisted estimator

$$\widehat{p} = \frac{1}{M} \sum_{k=1}^{M} L(\vec{X_k}) \, \mathbf{1}_{\sum X_{k,j} > na} \; . \tag{11}$$

The factor $\mathbf{1}_{**}$ is the indicator function that is equal to 1 if $\sum_j X_{k,j} > a$.

There are some technical things we need to do to implement this sophisticated estimation strategy. For one thing, we have to compute the function $Z(\lambda)$ and solve the equation (6). Actually, we can differentiate to get

$$F'(\lambda) = \frac{1}{Z(\lambda)} Z'(\lambda) = \frac{1}{Z(\lambda)} \int x e^{\lambda x} f(x) \, dx \; .$$

Therefore, (6) is equivalent to $f'(\lambda) = a$. These integrals may be calculated and the equation solved by numerical integration and One variable Newton's method. Those calculations would be fast compared to the work it takes to sample. This is a general feature of large deviation inspired methods – some deterministic computational problem to solve to find the "mechanism". There, that is just identifying λ_*. In other instances (e.g., work by Professor Vanden Einden and others) it requires solving a partial differential equation. Even then, the PDE solve is cheap compared to sampling.

Once we have λ_*, we have to sample the twisted distribution (2). After a hard class on Monte Carlo methods, we expect to know enough to do this efficiently. Rejection sampling is often used here. A piece of philosophy is that we can do hard MC methods because we are able to sample from distributions efficiently.

3 Histogram bifurcation

Histogram bifurcation methods estimate p_n by estimating s_r defined by

$$\Pr(S_n > s_r) = 2^{-r} \ .$$

Suppose we choose a one-step sample size m and choose m independent samples with $X_{k,j} \sim f$. Then we can estimate s_1 by

$$\#\left\{X_{1,j} > s_1\right\} = \frac{1}{2}\, m \ .$$

This says that half the sample is above s_1 and half is below. Define the "Cramer set"

$$C_s \subset \mathbb{R}^n = \left\{ \vec{X} \mid \sum_{j=1}^{n} X_j > s \right\} \ .$$

We are looking for s_r so that

$$\Pr(C_{s_r}) = 2^{-r} \ .$$

If we estimate this by sampling from the un-twisted distribution, we will be doing direct rare-event sampling, which is a bad idea. Instead, we use the fact that

$$\Pr(\ \vec{X} \in C_{s_{r+1}} \mid \vec{X} \in C_{s_r}) = \frac{1}{2} \ . \tag{12}$$

This suggests a way to estimate s_{r+1} from s_r.

Let ρ_s be the probability distribution defined by the constraint C_s

$$\rho_s(\vec{x}) = \frac{1}{\Pr(C_s)} \begin{cases} \prod f(x_j) & \text{if } \sum_{J=1}^{m} x_j > s \\ 0 & \text{otherwise .} \end{cases}$$

Suppose we have $\widehat{s}_r$, which is an estimate of s_r. Then we calculate $\widehat{s}_{r+1}$ by using MCMC to get m samples from $\rho_{\widehat{s}_r}$ and then choosing $\widehat{s}_{r+1}$ so that half of the samples are greater than $\widehat{s}_{r+1}$. You can say that $\widehat{s}_{r+1}$ bifurcates (cuts in half) the histogram of S for $\vec{X} \sim \rho_{s_r}$.

The proof of Cramer's theorem suggests that there is an efficient Gibbs sampler for ρ_s that works by resampling X_i successively for $i = 1, \cdots, n$. Suppose you want to resample component X_i. You make a proposal $Y \sim f$ and accept if the new sample is in C_r:

$$\left(\sum_{j=1}^{n} X_j \right) + Y - X_i > s \ .$$

Otherwise you reject Y and keep the old value X_i. The probability that Y is accepted is related to the "overlap" between f and f_λ. This overlap is "healthy" even when p_n is exponentially small. This makes it practical to simulate ρ_s with reasonable auto-correlation time.

These repeated bifurcation methods (and related method below) suffer from a buildup of error The error $\widehat{s}_{r+1}$ depends on the error in $\widehat{s}_r$ and in the new error from one more bifurcation. An overall error bound for the estimate of p_n would have to take into account the number of bifurcations needed to get there. Since the bifurcation errors are approximately independent, this might scale like the square root of the number of bifurcation steps. The bifurcation method is not cheap, but is it much cheaper than the direct method. It could be optimized – maybe the factors should not be $\frac{1}{2}$. Maybe all the simulations should not have the same length?

Finally, a piece of philosophy. The Cramer theory and algorithm depend on the integrals (6) being finite. This fails in "fat tailed" distributions that do not have very fat tails. For example, it fails for the log-normal distribution, which has finite moments of all orders. When the exponential moments (6) become infinite, then the mechanism of $S > na$ changes. It is likely that the bifurcation algorithm also starts to perform poorly in that case, because the Gibbs sampler has a harder distribution. The bifurcation algorithm requires less knowledge from analysis than the importance sampler, but it does require some knowledge.

4 Thermodynamic integration

Thermodynamic integration is a method for estimating integrals like (1). It uses a sequence of stages, as the bifurcation method does. Each stage gains a factor, so you get many factors using many stages. The idea is to define a sequence of likelihood distributions $L_\beta(x)$ We then define

$$Z(\beta) = \int L_\beta(x)\pi(x)\,dx\ .$$

Note the simplified notation with that data Y not given explicitly. The likelihoods should be chosen, for $0 \le \beta \le 1$ so that $Z(0)$ is known and $Z(1)$ is the target. The

5 Simulated tempering

6 Exercises

1. Use Watson's lemma and the formula (9) to estimate $p_n(a)$. The exponential factor is the same but the constant and the power of n in the pre-factor change. Be careful to integrate with respect to s and not a.

2. Discuss the details of implementing the Cramer inspired importance sampling estimator (11).

   (a) Write a formula for the likelihood ratio.

   (b) Use Cramer theory (arguments like Section 2) to estimate $\text{var}(\widehat{p})$. Don't worry about the constant. Just show that the relative accuracy goes like a power of n rather than an exponential as the direct estimator does.

3. Code the single component Gibbs sampler from Section 3. Take $f(x)$ to make X uniformly distributed in $[-1, 1]$. Use this to estimate

$$\text{E}[, S \mid S > s] \ .$$

   You will have to several histogram splittings to get to large values of s. Compute the auto-correlation function and estimate the autocorrelation time of the MCMC sequence S_k. The surprising result is that $\tau < 1$.