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

Class notes: Monte Carlo methods
Week 2, Importance sampling, characteristic functions
Jonathan Goodman
September 16, 2020

1 Importance sampling

sec:is

Importance sampling is a Monte Carlo technique with many uses. One use is *variance reduction*. You find a different and probably more complicated way to estimate the same number. The complicated way is more work per sample, but needs fewer samples to achieve a given accuracy because its variance is lower. Another use is simplifying the problem of sampling. You find another probability density that is easier to sample than the one you started with, but close enough so that the change of distribution doesn't increase the variance too much. Designing importance sampling strategies for either purpose usually starts by understanding the original problem a little better.

This class introduces importance sampling and gives examples of these two ways it is applied. The material makes use of Gaussian approximation of probability densities and "quadratic exponential" approximations to functions being integrated. Expressions like $e^{-a(x-x_*)^2}$ are called *Gaussian* for probability densities of a random X. Otherwise, they are *quadratic exponentials*. If you want a Gaussian or quadratic exponential approximation, you first have to find x_*, which is the maximizer of the function or PDF. In probability, x_* is the *mode* of the PDF. It represents the "most likely way", or the *mechanism* by which something happens. The constant a in a PDF (or a generalization of it) determines the *fluctuations* around the "most likely way".

Importance sampling involves at least two probability distributions. Suppose $\rho(x)$ is a PDF for a d component random variable $X \in \mathbb{R}^d$. Using the notation from last week, suppose we seek to estimate

$$A = \mathrm{E}[V(X)] \; . \tag{1}$$ Ad

It will help to put the probability distribution as a subscript, as

$$A = \mathrm{E}_\rho[V(X)] = \int V(x)\rho(x)\,dx \; . \tag{2}$$ Epdf

Let $\sigma(x)$ be another PDF, and define the *likelihood ratio* to be

$$L(x) = \frac{\rho(x)}{\sigma(x)} \; . \tag{3}$$ L

Then

$$A = \int V(x)\rho(x)\,dx = \int V(x)\frac{\rho(x)}{\sigma(x)}\,\sigma(x)\,dx \; .$$

Therefore

$$A = \mathrm{E}_\sigma[\, V(X)L(X)\,] \,. \tag{4}$$ is

There are two ways to estimate A. One way is to generate N independent samples $X_k \sim \rho$ and use the *direct estimate*

$$\widehat{A}_{\rho,N} = \frac{1}{N} V(X_k) \,. \tag{5}$$ de

The other way is to generate N independent samples $X_k \sim \sigma$ and use the *importance sampling* estimate

$$\widehat{A}_{\sigma,N} = \frac{1}{N} V(X_k)L(X_k) \,. \tag{6}$$ ise

Importance sampling means using σ instead of ρ, and compensating by averaging V with the likelihood ratio.

Variance reduction is one application of importance sampling. In variance reduction, one seeks σ so that the variance of the importance sampled estimator (6) is less than the variance of the direct estimator (5).

$$\mathrm{var}_\sigma[\, X(X)L(X)\,] < \mathrm{var}_\rho[\, V(X)\,] \,.$$

Figure 1 is an fig:md example where the importance sampled error bar is smaller than the direct estimate almost by a factor of 8. This reduces the number of samples needed to reach a given accuracy by a factor of almost $8^2 = 64$.

Importance sampling may be used if σ is easier to sample than ρ. is to simplify the sampling process. It may be that $\rho(x)$ is complicated and hard to sample. Maybe we can find a simpler density $\sigma(x)$ that is easier to sample. The "fancy" sampler might be higher variance than the direct sampler, but the ease of sampling might make this a good tradeoff. This is particularly common when ρ is complicated but approximately Gaussian. We can take σ to be a Gaussian approximation to ρ.

The term *importance sampling* comes from the idea that the most common values of X under ρ might not be the most "important" ones. The alternative density σ may make these "important" values of X more likely. In fancy applications, importance sampling may be called *change of measure*. In finance, people talk about two "worlds". There is the "real world" where $X \sim \rho$ and $A = \mathrm{E}[V(x)]$. There is an alternative world (with alternative facts?) where $X \sim \sigma$ and $A = \mathrm{E}[VL]$.

As an example, consider the problem calculating moments of a standard normal

$$\mu_{2n} = \mathrm{E}_{\mathcal{N}(0,1)}\big[\, Z^{2n}\big] \,.$$

This is a big number for large n, so large values of Z must be important. For this reason, we try an alternative density $\sigma = \mathcal{N}(0, \sigma^2)$ (be careful of this conflict of notation). The probability densities involved are

$$\rho(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} \,, \quad \sigma(z) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{z^2}{2\sigma^2}} \,.$$

The likelihood ratio is

$$L(z) = \frac{\frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}}{\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{z^2}{2\sigma^2}}} = \sigma e^{-\frac{z^2}{2}\left(1-\frac{1}{\sigma^2}\right)} \ . \tag{7}$$ Glr

Therefore

$$\mu_{2n} = \sigma \mathrm{E}_{0,\sigma^2}\left[Z^{2n} e^{-\frac{Z^2}{2}\left(1-\frac{1}{\sigma^2}\right)} \right] \ . \tag{8}$$ isGm

The direct way to evaluate μ_{2n} is to generate N independent samples $Z_k \sim \mathcal{N}(0,1)$ and average the numbers Z_k^{2n}. An important sampling strategy would be to generate N independent samples $Z_k \sim \mathcal{N}(0,\sigma^2)$, average the numbers $Z_k^{2n} e^{-\frac{Z_k^2}{2}\left(1-\frac{1}{\sigma^2}\right)}$, then multiply by σ.

Here is what is going on in this importance sampling strategy. We want Z values that are larger than typical standard normal values, so we sample from a density with a larger variance. This gives samples Z_k that are larger than they should be. If we just average Z_k^{2n}, the average would be too large. Instead we *re-weight* the samples with the likelihood ratio (7). For $\sigma > 1$ (which we plan to use), the coefficient of Z^2 is negative, so the weights

$$W_k = \sigma e^{-\frac{Z_k^2}{2}\left(1-\frac{1}{\sigma^2}\right)}$$

are likely to be small. This makes the importance sampled estimator (8) have the same expected value as the direct sampler. We find the precise formula for the weights W_k by doing some algebra with probability densities.

The posted code `ImportanceSamplingDemo.py` implements this importance sampling algorithm. You will see that I made it starting from a code from Week 1. This is for estimating $\mathrm{E}\left[Z^8 \right]$, but you should experiment with higher moments and different σ stretch factors. Experiment with higher σ values for higher moments. Note the last column for relative error. If you look at the row for $\sigma = 2$, you will see the error is -1.1, which may not seem small. But the actual answer is $A = 105 + 3 \cdot 5 \cdot 7$, so the relative error is $-.013$. Most of the errors are within the error bars, but not all. The error bar for $\sigma = 1$ (which is the direct estimate) is about ten. All the larger σ have smaller error bars. The optimum seems to be $\sigma = 3$, but nearby σ give almost as good performance.

```
[JonathansMBP20:MonteCarlo20/classMaterials/week2] jg% python3 ImportanceSamplingDemo.py

   importance sampling for the 8-th moment with 10000 samples

   sigma    estimate      error       error bar   relative error

   1.00    1.048e+02    -1.856e-01    1.011e+01    -1.768e-03
   1.50    1.065e+02     1.501e+00    2.482e+00     1.430e-02
   2.00    1.039e+02    -1.084e+00    1.540e+00    -1.033e-02
   2.50    1.054e+02     4.305e-01    1.323e+00     4.100e-03
   3.00    1.049e+02    -5.022e-02    1.280e+00    -4.782e-04
   3.50    1.040e+02    -1.007e+00    1.302e+00    -9.586e-03
   4.00    1.050e+02    -2.056e-02    1.348e+00    -1.958e-04
   5.00    1.021e+02    -2.917e+00    1.473e+00    -2.778e-02
   7.00    1.047e+02    -2.677e-01    1.775e+00    -2.550e-03
  10.00    1.050e+02     4.203e-02    2.169e+00     4.003e-04
```

Figure 1: Output from `ImportanceSamplingDemo.py` using importance sampling to estimate Gaussian moments. fig:md

2 Bayesian Reasoning

sec:B

Here is a brief explanation of the Bayesian point of view in statistics. The course will come back to this in future weeks. Today, it is motivation for the Bayesian Information Criterion (*BIC*) integrals of subsection 3.2. sec:BIC

Bayesian statistics is an approach to parameter estimation from data. Suppose the parameters to be estimates are $(X_1, \ldots, X_d)$. Suppose the data are $(Y_1, \ldots, Y_N)$. Bayesian approach is based on a model of how the data are made. First, the parameters are chosen from a probability distribution called the *prior*. Before the data are made, we have $X \sim \pi(x)$. Then the data are made from a probability distribution that depends on the parameters,

$$Y \sim \rho(y|X) \; .$$

The joint distribution of parameters and data is

$$(X, Y) \sim \pi(x)\rho(y|x) \; .$$

The *posterior* density is the conditional density of the parameters, conditional on the data, which is

$$X \sim \rho(x|Y) = \frac{\pi(x)\rho(Y|x)}{Z(Y)} \; . \tag{9}$$ p

This formula is *Bayes' rule* for conditional probability (look it up if necessary). The posterior is the conditional density of the parameters, conditioned on the data you have. The Bayesian philosophy is that this posterior represents your knowledge of the parameters given the data. The word *prior* means "before" and *posterior* means "after". This is before and after seeing the data.

The notation above follows the practice in Bayesian statistics. The prior is often called π. All other probability densities are denoted with the same letter,

ρ in this case. The conditional distribution of the data given the parameters is $\rho(y|x)$. The posterior of the parameters given the data is $\rho(x|y)$. For computer scientists and people who program in C++, this would be called *polymorphism*, in which the same function name can refer to different functions depending on the types of the function arguments. Python does not have this kind of polymorphism, but Bayesian statistics does.

The denominator $Z(Y)$ in the posterior density (9) may be found using the fact that the posterior is a probability density as a function of x. That is

$$\int \rho(x|Y)\,dx = 1\ , \quad \text{for all } Y\ .$$

In view of (9), this gives

$$Z(Y) = \int \pi(x)\rho(Y|x)\,dx\ . \tag{10}$$ ei

In the context of Bayesian statistics, this integral may be called the *evidence integral*. In Week 1, a similar integral formula for the normalization constant was called the *partition function*. The integral is complicated in part because probability densities like $\rho(y|x)$ depend on parameters x is complicated ways. As an example of this, look how the normal PDF depends on the standard deviation. Numerical estimation of the evidence integral turns out (see later weeks for details) to be one of the hardest technical challenges of practical Bayesian statistics.

Here's an example of a situation where the Bayes approach seems appropriate. Suppose there is a "population" whose blood pressure is Gaussian with mean μ and variance σ_P^2. [Obviously oversimplified model: "blood pressure" is more complicated than just one number, and whichever number you choose, it would not be Gaussian in any earthly population.] Suppose you select one person from that population "at random" and measure her blood pressure, giving Y. Blood pressure measure measurements are unbiased and have error that is Gaussian with standard deviation σ_M.

The statistics question is: given the measurement Y, what is your "posterior" state of knowledge about the actual blood pressure X. Suppose, for example, that Y is much less than μ. That means the measurement suggests that the person has a very low blood pressure, much lower than the mean. Since it's unlikely for a person to have such a low blood pressure, you may suspect instead that there was a large measurement error. Your prior belief (the prior) influences your posterior belief, but the measurement influences it also. The Bayes point of view is a systematic way to take both priors and new data into account.

Here's how it works out in this simple Gaussian example. The probability

densities are

$$\begin{aligned}
\pi &= \mathcal{N}(\mu, \sigma_P^2) \\
\pi(x) &= \frac{1}{\sqrt{2\pi\sigma_P^2}}\, e^{-\frac{(x-\mu)^2}{2\sigma_P^2}} \\
\rho(\cdot|x) &= \mathcal{N}(x, \sigma_M^2) \\
\rho(y|x) &= \frac{1}{\sqrt{2\pi\sigma_M^2}}\, e^{-\frac{(y-x)^2}{2\sigma_M^2}} \\
\rho(x,y) &= \pi(x)\rho(y|x) \\
\rho(x,y) &= \frac{1}{2\pi\sigma_P\sigma_M} e^{-\left(\frac{(x-\mu_P)^2}{2\sigma_P^2} + \frac{(y-x)^2}{2\sigma_M^2}\right)} \\
\rho(x|y) &= \frac{1}{Z(y)}\rho(x,y) \\
\rho(x|y) &= \frac{1}{Z(y)}\,\frac{1}{2\pi\sigma_P\sigma_M} e^{-\left(\frac{(x-\mu_P)^2}{2\sigma_P^2} + \frac{(y-x)^2}{2\sigma_M^2}\right)} \,. \qquad (11)
\end{aligned}$$

gp

We learn from this that the posterior density is a "quadratic exponential" (an exponential of a quadratic function of x). A quadratic exponential that is a probability density (the posterior is a density) may be expressed in terms of its mean and variance. I write $\mu_a(y)$ for the mean of the posterior density, which is called the *posterior mean*. The subscript a is for "after" – you can't use μ_P for both the prior and the posterior means. The posterior distribution depends on the data, so the posterior mean can depend on y. The posterior variance will be denoted σ_a^2. This turns out not to depend on the data – which is an unusual feature of the all-Gaussian model.

A feature of Gaussian distributions is that the mean is the point with the highest probability density. Therefore, we can find $\mu_a(y)$ by maximizing $\rho(x|y)$ over x. We see that x appears only in the exponent, so we maximize $\rho(x|y)$ by minimizing the exponent, which is a quadratic function of x. In the last line, the minimizer is called $\mu_a(y)$:

$$\begin{aligned}
\min_x \left[\frac{(x-\mu_P)^2}{2\sigma_P^2} + \frac{(y-x)^2}{2\sigma_M^2}\right] \Longrightarrow \partial_x \left[\frac{(x-\mu_P)^2}{2\sigma_P^2} + \frac{(y-x)^2}{2\sigma_M^2}\right] &= 0 \\
\frac{x-\mu_P}{\sigma_P^2} + \frac{x-y}{\sigma_M^2} &= 0 \\
\frac{x}{\sigma_P^2} + \frac{x}{\sigma_M^2} &= \frac{\mu_P}{\sigma_P^2} + \frac{y}{\sigma_M^2} \\
\mu_a(y) &= \left[\frac{1}{\sigma_P^2} + \frac{1}{\sigma_M^2}\right]^{-1} \left(\frac{\mu_P}{\sigma_P^2} + \frac{y}{\sigma_M^2}\right) .
\end{aligned}$$

This may be understood as saying the posterior mean is a weighted average of

the prior mean and the measurement. The weights are

$$w_P = \left[\frac{1}{\sigma_P^2} + \frac{1}{\sigma_M^2}\right]^{-1} \frac{1}{\sigma_P^2}\ , \quad w_M = \left[\frac{1}{\sigma_P^2} + \frac{1}{\sigma_M^2}\right]^{-1} \frac{1}{\sigma_M^2}\ . \tag{12}$$ pw

These are positive numbers that have $w_P + w_M = 1$. The posterior mean is

$$\mu_a(y) = w_p \mu_p + w_M y\ . \tag{13}$$ pm

How much weight is given to the prior and the data depends on their *precisions*. The *precision* is the reciprocal of the variance. According to the weight formulas, more precision implies more weight. If the measurement has high precision (low variance) relative to the precision of the prior, then the measurement gets more weight.

We find the variance by finding the coefficient of x^2 in the exponential. From (11), the exponent is

$$\frac{1}{2}\left(\frac{1}{\sigma_P^2} + \frac{1}{\sigma_M^2}\right) x^2 + \text{ less than quadratic in } x\ .$$

This says that the posterior precision is the sum of the prior and measurement precisions:

$$\frac{1}{\sigma_a^2} = \frac{1}{\sigma_P^2} + \frac{1}{\sigma_M^2}\ .$$

We may understand this as saying that the prior and the data are both information we use to form our posterior belief. The posterior variance is the inverse of the posterior precision:

$$\sigma_a^2 = \left(\frac{1}{\sigma_P^2} + \frac{1}{\sigma_M^2}\right)^{-1}\ . \tag{14}$$ pv

Altogether, the posterior is Gaussian with mean μ_a and variance σ_a^2. Thus, the posterior density (11) may be written as

$$\rho(x|y) = \frac{1}{\sqrt{2\pi\sigma_a^2}} e^{-\frac{(x-\mu_a)^2}{2\sigma_z^2}}\ .$$

Gaussian examples illustrate the definitions, but they are not typical of general problems in Bayesian statistics. Even when it is possible to write a closed form expression for the posterior density (9), it is a computational challenge to learn what this formula says about the posterior distribution of X. The term *Markov chain Monte Carlo* (*MCMC*) was invented by statisticians who realized that these sampling methods made it possible to understand the posterior. This course will spend many weeks on MCMC.

The Bayesian view of statistics may be contrasted with the *frequentist* view. The Bayes approach needs a prior distribution on the unknown parameters. Sometimes there is a prior that is informed by lots of prior information. The

distribution of blood pressure in a population would be an example of a well justified prior. More commonly, the prior reflects some general qualitative beliefs about the parameters. For example, we may assume the components of X are uniformly distributed in some intervals that make sense on general grounds. In this case, we hope that the prior is *uninformative*, which means that the posterior does not depend strongly on arbitrary specifics of the prior. The extreme uninformative prior is the *flat* prior. In the flat prior, $\pi(x)$ is assumed to be constant. The constant value is irrelevant. In this case, the posterior would be given by (9), but with $\pi(x)$ removed. Needless to say, a flat prior is not a true probability density. The posterior may or may not make sense with a flat prior. There is a phrase associated with other physical modeling: "replacing ignorance with fiction". I think this applies also to deciding what prior to use. The Wikipedia page on Bayesian statistics, and many books on the subject, talk about *conjugate priors*. These have no justification other than mathematical convenience. If a conjugate prior gives an answer different from a flat prior, then this difference is a step away from the right answer. My advice: never use a conjugate prior. One of the strongest arguments for the frequentist point of view is that it does not require a prior.

The first step in a frequentist statistical analysis is to present a *point estimate*, $\widehat{X}$, which is a "best guess" at the true parameter combination x. The point estimate is a function, so we can write $\widehat{X}(Y)$. An advantage of point estimates is that you can estimate properties of a distribution without estimating other properties. We used this in Week 1, when we estimated the mean and variance of a random variable that was called $V(X)$. It is possible to estimate the mean and variance without estimating the distribution of V. A disadvantage of a point estimate is that it's just a number or collection of d numbers that are the best guess at the d components of X.

2.1 Model selection

sec:ms

Model selection is a part of statistics where you decide which of several candidate models is most appropriate for the data you have. Typically, the candidate models range from simple to more complex, with the more complex models being more complex extensions of simple models.

For example, suppose your data are (t_j, Y_j), for $j = 1, \ldots, n$. These may represent observations Y_j taken at fixed times t_j. The Y_j are the true values modified by observation noise, which we take to be independent Gaussian as above. A model might take the form

$$Y_j = f(t_j, X_0, \ldots, X_{d-1}) + Z_j \ .$$

The "functional form" $f(t, X)$ contains d fitting parameters $X_0, \ldots, X_{d-1}$. The "observation errors" Z_j are assumed to be independent Gaussians with mean zero and variance σ^2. The *residuals* for a given parameter combination X are the differences between the model "prediction" and the observational data, which is

$$Z_j = Y_j - f(t_j, X) \ .$$

Calibrating a model means choosing fitting parameters. We write the best fit as $\widehat{X}$. The ides, then, is that

$$\widehat{Y}_j - f(t_j, \widehat{X})$$

is our best estimate of the true y_j that we are unable to observe exactly. A frequentist might give the point estimate $\widehat{X}$ as the best guess of the true parameter combination. A "true" Bayesian would not give a point estimate at all.

The *least squares* fitting criterion is to find the parameter combination that minimizes the sum of the squares of the residuals

$$R_d = \min_X \sum_j \left(Y_j = f(t_j, X)\right)^2 \ .$$

Let us denote the optimal parameter combination by $\widehat{X}$. The model prediction would be

$$\widehat{Y}(t) = f(t, \widehat{X}) \ .$$

One goal of statistical modeling is to make accurate predictions. We don't only want the best parameters, we want the best model. That's model selection.

We could choose the model to minimize the fitting residual R_d. But that leads to *overfitting*. More parameters lead to smaller R_d (better data fitting), but are worse at prediction. For polynomial models, it is a simple mathematical fact that $R_d < R_{d-1}$, except if $\widehat{X}_{d-1} = 0$ in the d parameter model, which is very unlikely even if the true parameter is equal to zero. The extreme case is $d = n$, where the least squares polynomial interpolates the data exactly. This means that the parameters "fit the noise" (measurement errors) as well as the "signal". The result is predictions based on noise. We need a criterion that decides how much fitting error should we allow in the interest of simple models with fewer parameters.

Elementary classical statistics has an answer that applies to polynomials. You do a hypothesis test to see whether the hypothesis that the new parameter is zero is rejected by the data. This is not backed by theory that I'm aware of, and it does not even apply in most problems where the parameter appears in a non-linear way. Bayesian model selection is an alternative.

The Bayesian model is that model d is chosen with probability π_d. Then the parameters X are chosen with a prior $\pi_d(x)$. The the data are chosen with a data model $Y \sim \rho_d(y|X)$. The posterior of parameters and data is

$$\rho(x, d|Y) = \frac{\rho_d(Y|x)\pi_d(x)\pi_d}{Z(Y)}$$

The posterior probability of model d is

$$rho_d(Y) = \int \rho(x, d|Y)\, dx = \frac{1}{Z(Y)} \left[\int_x \rho_d(Y|x)\pi_d(x)\, dx\right] \pi_d \ .$$

The quantity in square brackets is the *evidence integral*

$$E_d = \int_x \rho_d(Y|x)\pi_d(x)\, dx \ .$$

3 Rare events and large deviations

A rare event is an event that is unlikely to happen. Suppose X is a random object and S is some set of outcomes. A direct estimate of $A = \Pr(S) = \Pr(X \in S)$ is to make N independent samples of X and count the "hits", which are samples X_k with $X_k \in S$:

$$H = \#\{X_k \in S\} \ .$$

The estimate is the fraction of samples that are hits:

$$\widehat{A} = \frac{H}{N} \ . \tag{15}$$ dpe

It is expensive to estimate the probability of rare events by direct simulation because most of the samples are "wasted" because they are not hits. The accuracy of a Monte Carlo estimate typically scales like the square root of the number of samples. For the direct rare event estimator, it depends on the expected number of hits, which is far smaller than the number of samples, see exercise 3. More simply, if you generate $N = 1000$ samples and get no hits, then you can't tell whether $A = 10^{-3}$ or $A = 10^{-5}$ or smaller. You do not know, even approximately, the order of magnitude of the rare event probability.

Large deviation theory may be used to find importance sampling strategies for rare events that reduce the variance (20) of direct sampling. A *large deviation* happens when a random variable is much larger or much smaller than usual. Large deviations are rare events. *Large deviation theory* is a collection of methods for making rough theoretical estimates of the probabilities of large deviations.

3.1 The Laplace method

sec:Lm

The *Laplace method* is an approximation technique for certain integrals. It is the basis for many Gaussian approximations, as we will see. Some rare event/large deviation problems can be understood using it. It illustrates the philosophy that you start by looking for the most likely way an extreme event happens. I explain it using a simple example, which is a derivation of Stirling's approximation. The general principles should be clear from the example.

The derivation starts with the integral formula

$$n! = \int_0^\infty x^n e^{-x} \, dx \ .$$

Stirling's approximation is an approximate formula that is valid when n is large. When n is large, and x is not large, the integrand increases rapidly with x. On the other hand, the "exponentials beat powers" principle implies that eventually e^{-x} "wins" and the integrand starts decreasing. Most of "the mass" of the integral is in the intermediate range neither term dominates. First, we identify the x values where the integrand is large, starting by finding x_*, where the

integrand is maximized. Then we approximate the integrand in a neighborhood around x_*. Inside this neighborhood, the integrand is well approximated by a "quadratic exponential" function (see below). This integral can be calculated explicitly and gives the Stirling approximation. The rest of the integral is much smaller by comparison.

You find the max of the integrand by setting the derivative to zero and solving

$$\begin{aligned} \frac{d}{dx}\, x^n e^{-x} &= n x^{n-1} e^{-x} - x^n e^{-x} \\ 0 &= (x-n) x^{n-1} e^{-x} \\ x_* &= n \, . \end{aligned}$$

Now we use a Taylor series to approximate the integrand for x near x_*. The key to the Laplace method is to approximate the exponent, not the integrand. For this, we write

$$x^n e^{-x} = e^{-\phi(x)} \, , \quad \phi(x) = x - n \log(x) \, . \tag{16}$$ phi

The $-$ sign in $e^{-\phi}$ follows tradition and makes it possible to talk about this using the intuitive language of statistical physics. For that reason, we call $\phi(x)$ the *potential*. The *Gibbs* distribution (of the *Gibbs Boltzmann* distribution) corresponding to potential ϕ (and $k_B T = 1$) is $\rho(x) = \frac{1}{Z} e^{-\phi(x)}$. The probability density goes down as the potential goes up. The normalization constant Z is called the *partition function* and is determined by the normalization of the probability density

$$\int \rho(x)\, dx = 1 \implies Z = \int e^{-\phi(x)}\, dx \, .$$

The *ground state* or the *minimum energy state* is x_* with $\phi(x_*) = \min$. You find the state with highest probability density by minimizing the potential.

We have already calculated in our example that the minimizer of our potential is $x_* = n$:

$$\min_x \phi(x) = \phi_{\min} = \phi(x_*) = n - n \log(n) \, .$$

For a Taylor series of ϕ for $x \approx x_* = n$, we calculate the derivatives of (16) phi (ϕ'''

is used later):

$$\begin{aligned}
\phi(x_*) &= n - n\log(n) \\
\phi'(x) &= 1 - \frac{n}{x} \\
\phi'(x_*) &= 0 \quad \text{(because } x_* \text{ s a minimizer)} \\
\phi''(x) &= \frac{n}{x^2} \\
\phi''(x_*) &= \frac{1}{n} \\
\phi'''(x) &= -\frac{2n}{x^3} \\
\phi'''(x_*) &= -\frac{2}{n^2} \; .
\end{aligned}$$

The quadratic Taylor approximation for x near x_* is

$$\phi(x) \approx \phi(x_*) + \frac{1}{2}\phi''(x_*)(x - x_*)^2 = n - n\log(n) + \frac{(x-n)^2}{2n} \; . \qquad (17) \quad \boxed{\texttt{pa}}$$

The corresponding integral is

$$\begin{aligned}
n! &= \int e^{-\phi(x)}\, dx \\
&\approx \int e^{-\left[n - n\log(n) + \frac{(x-n)^2}{2n}\right]}\, dx \\
&= e^{n\log(n)} e^{-n} \int e^{-\frac{(x-n)^2}{2n}}\, dx \\
&\approx n^n e^{-n} \int_{-\infty}^{\infty} e^{-\frac{(x-n)^2}{2n}}\, dx \\
n! &\approx \sqrt{2\pi n}\, n^n e^{-n} \; . \qquad (18) \quad \boxed{\texttt{sa}}
\end{aligned}$$

In the next to last step we changed the range of integration from $[0, \infty)$ to $(-\infty, \infty)$. This is a good approximation because if n is large, then the integral over $(-\infty, 0]$ is much smaller than the integral over $[0, \infty)$. If you think of the integrand as a Gaussian probability density, then the mean is n and the variance is n, so the standard deviation is $\sqrt{n}$. This makes any point with $x \le 0$ at least $\sqrt{n}$ standard deviations away from the mean. For example, with $n = 10$, that's at least $\sqrt{10} = 3.15$ standard deviations, which is a lot for a Gaussian.

The formula (18) is *Stirling's approximation.* [If this were not a math class, I might have said *Stirling's formula* and written $n! = \sqrt{2\pi n}\, n^n e^{-n}$.] We have already seen that even the simpler "formula" $n! = n^n e^{-n}$ explains much of the behavior of the factorial function. We found this just by minimizing the potential

$$n! = \int e^{-\phi(x)}\, dx \sim n^n e^{-n} = e^{-\phi(x_*)} \; . \qquad (19) \quad \boxed{\texttt{ca}}$$

The "only thing" missing from this is the *prefactor*, which is $\sqrt{2\pi n}$. The prefactor is less important because it is algebraic, which is small compared to the exponential factors n^n or e^n. We also saw that it takes more work to figure the prefactor than the exponential part. For the exponential part, you just minimize the potential function. For the prefactor you need the second derivatives, the Gaussian approximation, and the change in the region of integration.

When the Laplace method works, it works because the range of integration consists of two overlapping regions. One region is points near x_* where the potential is small. The potential is large enough in the other region, and $e^{-\phi(x)}$ is so small, that the integral over the far region is tiny compared to the integral over the near region. The key is that the quadratic Taylor approximation is valid in the near region. This is the dichotomy behind the Laplace method: either $\phi(x) \approx \phi(x_*) + \frac{1}{2}\phi''(x_*)(x - x_*)^2$, or $e^{-\phi(x)}$ is too small to matter.

You can see, informally, that the Laplace dichotomy is valid in our derivation of Stirling's approximation. The error in the two term Taylor approximation of $\phi(x)$ about x_* depends on the third derivative. The theme of Taylor series remainder inequalities from calculus is that the error is on the order of the first neglected term. In this case, the first neglected term is the third derivative term. One form of the remainder estimate is

$$\phi(x) = \left[\phi(x_*) + \frac{1}{2}\phi''(x_*)(x - x_*)^2\right] + \frac{1}{6}\phi'''(\xi)(x - x_*)^3 \ , \ \ |\xi - x_*| \leq |x - x_*| \ .$$

Informally, replace the unknown ξ with the nearby x_*, and you get

$$\phi(x) - \left[\phi(x_*) + \frac{1}{2n}(x - x_*)^2\right] \sim \frac{1}{3n^2}\,|x - x_*|^3 \ .$$

We can calculate $|x - x_*|$ that makes the error term equal to ϵ:

$$\begin{aligned} \frac{1}{3n^2}\,|x - x_*|^3 &= \epsilon \\ |x - x_*| &= 3^{\frac{1}{3}} n^{\frac{2}{3}} \epsilon^{\frac{1}{3}} \ . \end{aligned}$$

Then we can estimate how much ϕ has changed when $x - x_*$ is this big:

$$\phi(x) - \phi(x_*) \approx \frac{1}{2n}(x - x_*)^2 = \frac{3^{\frac{2}{3}}}{2}\epsilon^{\frac{2}{3}} n^{\frac{1}{3}} \ .$$

The Laplace dichotomy is true in this case because $n^{\frac{1}{3}}$ is a positive power of n. If $|x - x_*| < 3^{\frac{1}{3}} n^{\frac{2}{3}} \epsilon^{\frac{1}{3}}$, then the quadratic Taylor approximation of ϕ is accurate (within ϵ). If $|x - x_*| > 3^{\frac{1}{3}} n^{\frac{2}{3}} \epsilon^{\frac{1}{3}}$, then $\phi(x) > \phi(x_*) + \frac{3^{\frac{2}{3}}}{2}\epsilon^{\frac{2}{3}} n^{\frac{1}{3}}$, and (with $a = \frac{3^{\frac{2}{3}}}{2} = 1.04\cdots$)

$$e^{-\phi(x)} < e^{-\phi(x_*)}\, e^{-a\epsilon^{\frac{2}{3}} n^{\frac{1}{3}}} \ .$$

This shows that $e^{-\phi(x)}$ is less than $e^{-\phi(x_*)}$ by a factor of $e^{-a\epsilon^{\frac{2}{3}} n^{\frac{1}{3}}}$. For any $\epsilon > 0$, this goes to zero as $n \to \infty$. The integral from outside the near region is smaller by a factor goes to zero as n goes to infinity.

This simple check is not the whole story, but it is the essential part. If you're so inclined and have the training in $\epsilon - \delta$ analysis, you can make a proof based on these calculations and other (simpler) inequalities based on ϕ being convex. The theorem of *Stirling's approximation* is that it's relative error goes to zero as n goes to infinity.

$$\frac{\sqrt{2\pi n}\, n^n e^{-n} - n!}{n!} \to 0 \text{ as } n \to \infty .$$

3.2 Bayesian Information Criterion

sec:BIC

The *Bayesian information criterion*, or *BIC* is a way to decide which model is better for the data.

4 Assignment 2, due February 16

sec:assignment

Always check the class message board on the NYU Classes site from home.nyu.edu before doing any work on the assignment.
Corrections: none yet.

Gm

1. Let Z be a standard normal, which has PDF

$$\frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} .$$

Denote the moments by

$$\mu_k = \mathrm{E}[Z^k] .$$

Use integration by parts (Hint: $\frac{d}{dz} e^{-\frac{z^2}{2}} = -z e^{-\frac{z^2}{2}}$) to show that

$$\int_{-\infty}^{\infty} z^{2n} e^{-\frac{z^2}{2}}\, dz = (2n-1) \int_{-\infty}^{\infty} z^{2n-2} e^{-\frac{z^2}{2}}\, dz .$$

Conclude that

$$\mu_{2n} = (2n-1)(2n-3)\cdots 3 .$$

This is written $(2n-1)!!$. Do not confuse this with $((2n-1)!)!$. For example, $5!! = 5 \cdot 3 = 15$, while $(5!)! = 120! =$ big .

ex:gsh

2. Here's a heuristic that estimates a good stretch factor σ in the Gaussian moment importance sampling example. We set σ to be the size of a typical Z that contributes to μ_{2n}. A typical value of Z^{2n} is about $\mu_{2n} = (2n-1)!!$. Therefore, a typical value of $|Z|$ may be about $[(2n-1)!!]^{\frac{1}{2n}}$. Use Stirling's formula and

$$(2n-1)!! = \frac{(2n)!}{2^n n!}$$

to show that this typical $|Z|$ might be around $\sqrt{2n}$. Use this to explain the observation that $\sigma \approx 3$ is a good value in the experiment of Figure 1. fig:md

ex:dsh

3. Show that when $\Pr(X \in S)$ is small, then the direct estimator (15) has relative accuracy given by the expected number of hits:

$$\frac{\sigma_{\widehat{A}}}{A} \approx \frac{1}{\sqrt{\mathrm{E}[H]}} \; . \tag{20}$$

rad

ex:hb

4. As we say a lot, low dimensional intuition can lead to exponentially bad algorithms in high dimension. Suppose you want $X \in \mathbb{R}^d$ uniformly distributed in the unit ball $|x| \leq 1$. One way is to "propose" X uniformly distributed in the "bounding box" and then accept the first proposal that is inside the ball. The bounding box is defined by $-x \leq x_j \leq 1$ for $j = 1, \ldots, d$. You can make such X by taking components $X_j = -1 + 2U_j$, where the U_j are independent and uniformly distributed in $[0, 1]$ as given by a random number generator. The rejection algorithm is to generate such X until the first time $|X| \leq 1$. The number of trials needed for this grows exponentially with d. For this exercise, we assume instead that the proposals are independent uniformly distributed Gaussians. More precisely, $X_j \sim \mathcal{N}(0, 1)$, i.i.d. Here are some steps to estimate $\Pr|X| \leq 1)$. These include a new approximate integration method, which sometimes is called *Watson's lemma*. We rely on the following "polar coordinates" integration of radially symmetric functions in d dimensions. If $f(r)$ is a suitable function of $r \geq 0$, then

$$\int_{\mathbb{R}^d} f(|x|) \, dx_1 \cdots dx_d = \omega_{d-1} \int_0^\infty r^{d-1} f(r) \, dr \; .$$

Here, ω_{d-1} is the "surface area" of the unit sphere $|x| = 1$ in d dimensions.

(a) Find a formula for ω_{d-1} using the fact that

$$\int \cdots \int \left[\frac{1}{\sqrt{2\pi}} e^{-\frac{x_1^2}{2}} \right] \cdots \left[\frac{1}{\sqrt{2\pi}} e^{-\frac{x_d^2}{2}} \right] dx_1 \cdots dx_d = 1 \; .$$

The formula has the form

$$\omega_d = C_d \int_0^\infty r^d e^{-\frac{r^2}{2}} \, dr$$

and you have a formula for C_d. ($d - 1$ changed to d to make the formulas simpler.)

(b) Use the Laplace method to find an approximate formula for the integral that holds when d is large.

(c) Show that the probability $|X| \leq 1$ is

$$P_d = \frac{\omega_{d-1}}{(2\pi)^{\frac{d}{2}}} \int_0^1 e^{-\phi(r)} \, dr \; ,$$

where

$$\phi(r) = \frac{r^2}{2} - (d-1)\log(r) \ .$$

Find where in the interval $[0,1]$ the potential ϕ has a minimum and make a linear approximation (using ϕ and ϕ') of ϕ about that point. Integrate with the linear approximation t estimate P_d.

(d) Justify the approximation by showing that ϕ is large where the approximation is poor. Part (c) is the *Watson's lemma* method that is appropriate many problems where the minimum of the potential is at the boundary of the region of integration.