Assignment 5, due October 14, before class

Please follow instructions on the assignments page

About this assignment: Part 1: The Laplace approximation and its use as a way of thinking about probability distributions. Part 2: More experience with likelihood moddeling.

The Laplace approximation

When someone talks about the *Laplace approximation*, it's likely they are talking about a function defined as the exponential of a "potential" function

$$f(x) = e^{-\phi(x)} \; . \tag{1}$$

The term "potential" refers to applications in statistical physics in which $\phi(x)$ is the potential energy of configuration x. The maximum of f is achieved at the minimum of ϕ, which we call the "potential minimum"

$$x_* = \arg\max_x f(x) = \arg\min_x \phi(x) \; .$$

A common model[1] in statistical physics is that the configuration, X, is random and the probability depends on the exponential of the potential

$$p(x) = \frac{1}{Z} e^{-\beta\phi(x)} \; . \tag{2}$$

This PDF is called the *Gibbs* (or *Gibbs Boltzmann*) distribution for potential ϕ with "inverse temperature" β. The definition (1) doesn't have the normalization factor and sets takes $\beta = 1$. It often happens that X is "confined" to be close to x_*, meaning that in the distribution (2), it is unlikely that $X - x_*$ is large (vague statements, admittedly). This is particularly true when β is large, which makes the difference between $\phi(x)$ and $\phi(x_*)$ matter more.

If X is confined to be near x_*, then it is possibly natural to approximate p or f using a Taylor approximation of ϕ. Let $H = D^2\phi(x_*)$ be the Hessian evaluated at x_*. The gradient $\nabla\phi(x_*)$ is zero because x_* is a local minimizer. Therefore, the lowest order non-trivial Taylor approximation is[2]

$$\phi(x) = \phi(x_*) + \frac{1}{2}\left(x - x_*\right)^T H \left(x - x_*\right) + O\Big(\left\|x - x_*\right\|^3\Big) \; . \tag{3}$$

The *Laplace approximation* is to approximate ϕ by the first two terms on the right.

$$p(x) \approx \frac{1}{Z} e^{-\frac{\beta}{2}(x-x_*)^T H (x-x_*)} \; . \tag{4}$$

The corresponding approximation to the simple exponential function (1) just factors out the (constant) factor $e^{-\phi(x_*)}$:

$$f(x) \approx e^{-\phi(x_*)}\, e^{-\frac{1}{2}(x-x_*)^T H (x-x_*)} \; . \tag{5}$$

[1] In equilibrium statistical physics, this is called the *canonical ensemble*

[2] Look up the "Big Oh" notation $O(\cdots)$ if you haven't seen it. It's described in the book by Bindel and me, and in many many other places. Look for something that starts with examples and intuition. This is a useful concept but not deep.

An application of the Laplace approximation to the Gibbs distribution (2) is (using the fact that the mean of a Gaussian is where the PDF is maximized and the covariance matrix is the inverse of the information matrix)

$$\begin{aligned} \mathrm{E}[X] &\approx x_* \\ \mathrm{cov}(X) &\approx \frac{1}{\beta} H^{-1} \end{aligned}$$

An application to the exponential function is (using the formula for the integral of the "quadratic exponential")

$$\begin{aligned} \int f(x)\,dx &\approx e^{-\phi(x_*)} \int_{\mathbb{R}^d} e^{-\frac{1}{2}(x-x_*)^T H (x-x_*)}\,dx \\ &= e^{-\phi(x_*)} \, \frac{1}{(2\pi)^{\frac{d}{2}}} \, \frac{1}{\sqrt{\det H}} \, . \end{aligned}$$

The region near x_* is the *potential well*. A "well" is a deep hole where you might find water or oil. It is "deep" if it's a long way to the bottom and hard to get out of. The "hard to get out of" aspect will is important also in computational statistics, we we will learn and then regret.

It is common in fitting problems that the likelihood surface has more than local maximum and the corresponding potential has several distinct potential wells. These might correspond to rather different parameter combinations that are (locally) best fits to the data. The *multiple well* problem in statistics, as in statistical physics, is quite annoying because partly because it happens a lot. There are strategies for dealing with the multiple well problem, but none of the strategies are very general or guaranteed to work. It is an active and important research area in computational statistics, physics and physical chemistry.

It is natural to wonder whether the central limit theorem may be thought of as a Laplace approximation. Yes, and in several different ways. If you have seen the proof of the central limit theorem using Fourier analysis, go through it carefully and you will find a Laplace approximation.

1. (*Stirling Laplace approximation*) *Stirling's approximation* (or Stirling's *formula*) is an approximation to $n!$ when n is large. One derivation starts with an integral formula for $n!$ and makes a Laplace approximation to the integral. The integral formula is $n! = \int f(t)\,dt$. You put f in the form $e^{-\phi(t)}$ then make a quadratic approximation to ϕ. In this $d = 1$ example it is possible to make plots to visualize the Laplace approximation. In these applications, it is useless to approximate f by a quadratic. for one thing, f is positive and the quadratic isn't. Going deeper, the quadratic approximation to f has a tiny range of validity – just a small region about the potential minimum, but the region of validity of the quadratic approximation to ϕ is large enough (when it is) to include the whole region where the integrand is significantly different from zero.

   (a) Verify that

   $$n! = \int_0^\infty t^n e^{-1}\,dt \, .$$

   Hint. if you know it for $(n-1)!$, you can get it for $n!$ by integration by parts, using $e^{-t} = -\frac{d}{dt} e^{-t}$. You can verify directly for $n = 0$ or $n = 1$.

   (b) Write the integrand in exponential form: $t^n e^{-1} = e^{-\phi(t)}$. Find

   $$t_* = \arg\min_x \phi(t) \, .$$

   Make the local Taylor approximation

   $$\phi(t) \approx \widetilde{\phi}(t) = \phi(t_*) + \frac{1}{2}\phi''(t_*)(t - t_*)^2 \, . \tag{6}$$

Find an expression for t_* and $\phi''(t_*)$.

(c) Explore the accuracy of the approximation of ϕ by plotting ϕ and $\widetilde{\phi}$ (defined by (6)) on the same graph. Do one for a moderate value of n and one for a rather larger value. Do you observe the approximation improving as n increases?

(d) Explore the accuracy of $e^{-\widetilde{\phi}(t)}$ as an approximation to $f(t) = t^n e^{-t}$ in a region of t values around t_*. Use a few n values. Observe that the interval of t values where the approximation is valid starts to contain most of the interval where f is large. Thus the regions of t where the approximation is poor do not effect the integral very much.

(e) Find

$$S(n) = \int_{-\infty}^{\infty} e^{-\widetilde{\phi}(t)}\, dt = \sqrt{2\pi n}\, n^n e^{-n} \ .$$

Stirling's approximation is $n! \approx S(n)$. Sometimes it is simplified to[3] $n! \approx n^n e^{-n}$, possibly in the form $\log(n!) \approx n\log(n) - n$.

2. (*Poisson distribution*) The *Poisson* random variable, N, with parameters λ and T represents the number of "hits" (or "arrivals") of a *Poisson process* with rate parameter λ in time interval $[0, T]$. The Poisson probabilities are $p_n(\lambda, T) = \Pr_{\lambda,T}(N = n)$. (The right side of this is, in speaking, "the probability that the random variable N is equal to the integer n, with parameters λ and T".) It turns out that p_n depends only on $\mu = \lambda T$. This should be "obvious" in retrospect. Divide the time interval $[0, T]$ into M small intervals of size Δt, where $T = M\Delta t$. We will take $\Delta t \to 0$ and $M \to \infty$ with T fixed. Define independent *Bernoulli* random variables

$$H_k = \begin{cases} 1 & \text{if there is a "hit" in } [k\Delta t, (k+1)\Delta t] \\ 0 & \text{otherwise} \end{cases}$$

The probability of a hit in sub-interval k is

$$\Pr(H_k = 1) = \lambda \Delta t = p \ .$$

If the overall arrival rate is λ hits per unit time, and if Δt is small, then we make the approximation that there are either no hits or one hit in the sub-interval because two or more would be much less likely than one hit, which already is unlikely. The total number of hits is

$$N = \sum_{k=0}^{M-1} H_k \ . \tag{7}$$

A sum of independent Bernoulli random variables with a common p is called *binomial*. The exact probabilities for this binomial random variable are

$$\Pr_b(N = n) = \binom{M}{n} p^n (1-p)^{M-n} \ . \tag{8}$$

[3] You get some intuition from the fact that $n^n = n \cdot n \cdots n$, which is more than $n! = n \cdot (n-1) \cdots 1$. The ratio is $\frac{n}{n}\frac{n-1}{n} \cdots \frac{1}{n}$. That product is something like e^{-n}.

The *binomial coefficients* are

$$\binom{M}{n} = \frac{M \cdot (M-1) \cdot \cdots \cdot (M-n+1)}{n \cdot (n-1) \cdots}$$

$$\binom{M}{0} = 1$$

$$\binom{M}{1} = M$$

$$\binom{M}{2} = \frac{M(M-1)}{2}$$

etc.

When M is large, we make approximations

$$\binom{M}{2} = \frac{1}{2}M^2 \cdot \frac{M-1}{M} \approx \frac{1}{2}M^2$$

$$\binom{M}{3} = \frac{1}{6}M^3 \cdot \frac{M-1}{M}\frac{M-2}{M} \approx \frac{1}{6}M^3$$

etc.

Also, if a is a fixed number,

$$\left(1 + \frac{a}{M}\right)^M \approx e^a \, .$$

All these approximations have (relative) error terms of order $\frac{1}{M}$. The Poisson probabilities are the limit of the binomial probabilities as M goes to infinity

$$P_P(n) = \lim_{M \to \infty} P_n(n) \, .$$

(a) Show that (do these in order to see the pattern forming)

$$\Pr_P(N = 0) = e^{-\lambda T}$$

$$\Pr_P(N = 1) = \lambda T e^{-\lambda T}$$

$$\Pr_P(N = 2) = \frac{1}{2}(\lambda T)^2 e^{-\lambda T}$$

$$P_P(n) = \frac{1}{n!}(\lambda T)^n e^{-\lambda T}$$

(b) This "rain" picture of the Poisson process (every interval of length has probability $\lambda \Delta t$ to get a raindrop, all intervals independent) makes it easy to understand facts about the process. Use it to show that $\mathrm{E}_P[N] = \lambda T$ and that $\mathrm{var}(N) = \lambda T$. For the second, the variance of N is the sum of the variances of the H_k, which are independent. It is common to write μ for λT. The Poisson probabilities are

$$p_n = \frac{\mu^n}{n!} e^{-\mu} \, . \tag{9}$$

The mean and variance formulas you are asked to verify imply the following identities for

certain infinite sums:

$$\mu = e^{-\mu} \sum_{n=0}^{\infty} n \frac{\mu^n}{n!}$$
$$\mu = \left(e^{-\mu} \sum_{n=0}^{\infty} n^2 \frac{\mu^n}{n!} \right) - \mu^2 \; .$$

If you do this (verify these formulas directly using tricks for calculating sums), please don't hand that in.

3. (*Laplace-like approximation for rare events*) A *rare event* is an event that happens with low probability. It is often important to determine $\Pr(A) = p_A$ with good relative accuracy even when p_A is very small. You could estimate p_A by direct simulation: generate samples and count *hits*. A hit is an sample event where A happens. This can be inefficient if p_A small because most samples are not hits and therefore do not contribute to the estimate of p_A. *Rare event simulation* refers to methods that estimate p_A more effectively than direct simulation.

   (a) Suppose $\widehat{p}_A = \frac{N}{n}$, where N is the number of hits in n independent trials. Let $\widehat{\sigma}_{p_A}$ be the standard deviation of this direct estimator. Show that the relative accuracy is approximately (for large n)

   $$\frac{\sigma_{\widehat{p}_A}}{p_A} \approx \frac{1}{\sqrt{\mathrm{E}[\#\mathrm{hits}]}} \; . \tag{10}$$

   (b) Write a simulator for the Poisson process using exponential random variables. That is, let $R_k \sim \exp(\lambda)$ be independent exponentials with rate parameter λ (generated using the log trick). These are the *inter-arrival* times for the Poisson process. The arrival times (or raindrops) for the process are $S_1 = R_1$, $S_2 = S_1 + R_2$, $\cdots$ $S_k = S_{k-1} + R_k$. The Poisson random variable with rate λ and time T is the number of arrivals before time T:

   $$N = \max \{ \, k \text{ with } S_k < T \, \} \; .$$

   Test your simulator by using a histogram to estimate the Poisson probabilities p_n and making a plot of the estimated p_n and the formula (9). Choose parameters with μ significantly larger than 1 and significantly smaller than 50. Plot the histogram out to n where p_n is "vanishingly small". It may take many simulations to get good agreement. Try to make a histogram where you can clearly see small deviations and one where the deviations are too small to be seen.

   (c) The rare event will be the event where the number of arrivals is much more than the expected number: $N \gg \mu$. Choose an $M > 1$ (maybe $M = 2$, but think about other values) and $\mu \sim 25$ (again, think about other values). Make sure the code has M and μ as variables so you can change either value by changing one line. If Claude (or a different AI) does your coding, make sure they do this. Specifically, take the rare event A to be $A = \{ \, N > M\mu \, \}$. Estimate p_A and σ_{p_A} from simulation. Look for parameter combinations (M and μ) so that $p_A \sim 10^{-2} = 1\%$ and $p_A \sim 10^{-4}$. A typical laptop running scalar Python may not be able to estimate probabilities that small accurately in a reasonable amount of time. See whether that's true.

   (d) Write the Poisson probabilities (9) is $p_n = e^{-\phi(n)}$. Suppose we want to probability that $N > n_* = M\mu$ with $M > 1$ and μ sort of large. Make a Taylor approximation of ϕ

   $$\phi(n) \approx \widetilde{\phi}(n) = \phi(n_*) + \phi'(n_*)(n - n_*) \; . \tag{11}$$

   As with the Laplace approximation (when it applies), this is supposed to be accurate "where it matters", meaning that the n values where it is not accurate ($n > n_*$) have such small

probability that the error is irrelevant. Make some plots plotting ϕ and $\widetilde{\phi}$, and p_n and $\widetilde{p}_n = e^{-\widetilde{\phi}_n}$ to see whether this approximation is accurate for estimating rare events of the form A.

(e) The approximation (11) makes $\widetilde{p}_n$ a geometric series whose sum you can find in closed form. Use this to get an approximate formula for p_A. The formula predicts values of p_A even when they are far smaller than can be estimated using direct simulation. Use the estimator from part (b) to explore the accuracy of this approximate formula.

4. (*Error correlations*) Suppose a variable y is a linear function of d "regressor" variables $a_1, \cdots, a_d$ with weights $\theta_1, \cdots, \theta_d$.

$$y = \theta_1 a_1 + \cdots + \theta_d a_d \ .$$

To determine the weights, you do n experiments with various combinations of the explanatory variables. Without measurement errors, the results would be

$$y_k = \theta_1 a_{1,k} + \cdots + \theta_d a_{d,k} \ .$$

Suppose[4] that the resulting observations are $Y_k = y_k + \epsilon_k$, where the measurement errors ϵ_k are i.i.d. gaussian with a common variance σ_2. Formulate this as a linear least squares problem

$$\widehat{\theta} = \arg\min_{\theta} \|A\theta - Y\|_2 \ .$$

Identify the elements of the matrix A in terms of the experimental design variables a_{jk}. Show that, in this data model, this is equivalent to maximum likelihood estimation. Show that the estimation errors $\widehat{\theta}_j - \theta_j$ are correlated and find a formula for the estimate covariance matrix $\Sigma_{\widehat{\theta}}$. Show that the relative estimation errors may be much larger than the relative observation errors if A has a wide range of singular values (is *ill conditioned*).

5. (*Variance modeling*[5]) A certain physical process yields independent samples X_j from a PDF with parameters θ:

$$X_k \sim f(\cdot, \theta) \ , \quad \text{i.i.d.}$$

You want to estimate θ using a large sample set $X_1, \cdots, X_n$. The traditional method is to put the X_k into bins of width Δx, and bin centers x_j with $x_{j+1} = x_j + \Delta x$, resulting in bin counts

$$N_j = \# \left\{ k \ \text{ with } \ x_j - \frac{\Delta x}{2} \le X_k \le x_j + \frac{\Delta x}{2} \right\} \ .$$

The resulting density estimates are

$$f(x_j) \approx \widehat{F}_j = \frac{N_j}{n \Delta x} \ .$$

With this, you can look for parameter combinations θ where the theoretical PDF $p(\cdot, \theta)$ matches the histogram. This can be done using ordinary least squares fitting:

$$\widehat{\theta}_{\text{hom}} = \arg\min_{\theta} \sum_j \left(\widehat{F}_j - f(x_j, \theta) \right)^2 \ . \tag{12}$$

The subscript is for *homoskedasticity*, which means equal variance. As we know, this is equivalent to maximum likelihood estimation (the best possible) if the "measurement errors" satisfy $\widehat{F}_j = f(x_j, \theta_*) + \epsilon_j$, where $\epsilon_j \sim \mathcal{N}(0, \sigma^2)$, i.i.d.

[4] See Exercise 5 for a critique of these assumptions.

[5] Exercises 5 and 6 arose from a technical paper I reviewed.

(a) Show that the bin counts N_j are Bernoulli random variables. Show that if Δx is small, the N_j are approximately Poisson distributed with

$$\mu_j = \mathrm{E}[N_j] \approx n\Delta x\, f(x_j, \theta_*) \ .$$

Use this to find an approximate formula for $\sigma_j^2 = \mathrm{var}(\,\widehat{F}_j\,)$.

(b) Assume a central limit theorem approximation $\epsilon_j = \widehat{F}_j - f(x_j) \sim \mathcal{N}(0, \sigma_j^2)$. Describe the weighted least squares problem that arises from maximum likelihood estimation of θ_*. Denote the resulting maximum likelihood estimate by $\widehat{\theta}_{\rm het}$, where "het" is for "heteroskedasticity" (unequal variances).

(c) The fitting residuals[6] are

$$r_j = \widehat{F}_j - f(x_j, \widehat{\theta}_{\rm het})$$

The model of part (b) suggests that these residuals are random but on the order of σ_j. What does this suggest about the relative size of the fitting residuals? In particular, are the "outlier" bin probabilities estimated with more or less relative accuracy?

(d) In the homoskedasticity model related to (12), what do you expect about the rough sizes of fitting residuals? How should fitting residuals differ from those of the heteroskedastic model? Where would the residuals be larger or smaller?

6. (*Likelihood modeling*) Write the likelihood function that arises from the true data model $X_k \sim f(\cdot, \theta_*)$. Describe the optimization problem that results from maximizing the log likelihood. Assume the derivatives $\nabla_\theta f(x, \theta)$ are known (very reasonable assumption in many/most cases).

(a) Describe a gradient descent method for solving this optimization problem. The algorithm is deterministic, given the data.

(b) Describe a stochastic gradient descent method for the same problem, which uses just one data item for each iteration.

7. (*Computational experiments with PDF fitting*) A *Gaussian mixture model* is a model that makes f a weighted sum of several Gaussians with different means and variances. If there are m such Gaussians, the model is

$$f(x) = \sum_{i=1}^{m} w_i \gamma(\cdot, \mu_i, \sigma_i^2) \ .$$

Pure mathematicians often use γ to denote the Gaussian probability density:

$$\gamma(x, \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}} \ .$$

Download the dataset `obs.py`, which is a `pandas` file with fake data observations. Fit a two gaussian ($m = 2$) mixture model using each of the three methods: two from Exercise 5 and one from Exercise 6.

[6] *Residual* in this context means the extent to which an equation is not satisfied. It is usually possible to evaluate residuals, as it is here, but is usually not possible to evaluate the estimation error $\widehat{\theta} - \theta_*$.