Stat 238, Spring 2025
Homework 2

Alexander Strang
Due: by **5:00 pm** Friday, February 28, 2025

Submission Instructions

Homework assignments may have a written portion and a code portion. Please follow the directions here Homework Guidelines Ed post when submitting, and check the link for homework policies.

Problems eligible for spot grading are marked with an asterisk. These are problems 1, 3, and 5. Please mark at least two problems for spot grading. Note, these will be graded for effort, not accuracy, so you should select the problems that show your best problem-solving effort. The goal is to target our attention to give you meaningful feedback on your work. All problems will be checked for completion and basic accuracy.

Problem 1*: Asymptotic Abnormality

We argued in class that, under "mild" regularity conditions, the posterior distribution for a continuous-valued unknown should converge to a normal distribution when we make many, i.i.d. measurements. Anytime someone asserts that their conditions are "mild", remember that, as is true with dining and spicy food, mild is in the eye (tongue?) of the beholder. Here, you will explore one of the common failure modes of the Bayesian central limit theorem. It highlights a key distinction: mathematically special cases may be the general case for specific applications. In our case, for "small sample" problems where the likelihood struggles to uniquely pin down the unknown, even with a very large number of observations.

For this problem, we will consider the following model motivated by tomography problems. In a tomography problem, we shine an intense particle source (the emitter) at an object, and record the shadows it casts when the beam of particles passes through it. For an example of how this might work, see this video on electron microscopes.[1] The three-dimensional distribution of density and absorbtivity within the object is then reconstructed by moving the beam around the object and recording how its shadow changes as we change the direction of the beam.

It is common, especially in microscopy settings, that we want to minimize the amount of radiation we expose our targets to. For this reason, we use as weak a beam as we can, and as sensitive a detector (in essence, the film recording the shadow) as we can. Engineers and physicists are remarkably good at building detectors. Using appropriate amplification methods, it is common to build detectors that can count single particle impacts. For example,

[1] This problem is motivated by a Bayesian statistics problem I am currently working on in Cryo-electron microscopy, where the goal is to reconstruct an image of a cell at atomic resolution.

if electromagnetic waves (light) is used, then a detector may count single photon strikes. If an electron beam is used, then it might count electron strikes.

Consider the following model. Let $[Y_1, Y_2]$ be the number of observed electron strikes in two adjacent pixels of the detector over the imaging period. Since the beam emitter releases particles continuously, the particles largely don't interact, the number of incident strikes in any pixel is independent and identically distributed across disjoint time intervals. The number of expected strikes per time period is proportional to the length of the time period, τ. This implies that the counting process must have exponential waiting times, so the number of observed strikes in any time period, in any pixel, is a Poisson random variable with some characteristic rate. Our goal is to estimate these rates. So, consider the likelihood model:

$$\textbf{Likelihood: } Y_1 \sim \text{Poiss}(\lambda_1\tau), \quad Y_2 \sim \text{Poiss}(\lambda_2\tau) \tag{1}$$

Here, τ plays the role that n usually does in our asymptotic theory (remember, the sum of two indpendent Poisson random variables is Poisson).

Usually, the goal is to infer the rates λ in order to reconstruct the shadow the image should have cast if we shone the beam on it forever. This is, up to first order physical models, the intensity of the original beam reduced by the integral of the density of the pbject along the line passing from the emitter to the detector. Repeating this measurement for many lines through the object (many rotations of the emitter) produces a large linear system that can be solved to resolve the distribution of density in the target object. Like many imaging techniques, this is a technological, mathematical, physical, and engineering marvel. It allows us to peer, with surprising precision, through objects, and to create a 3D reconstruction of them!

However, life is never quite that easy. It is extremely common that we don't see $Y_1 \sim \text{Poiss}(\lambda_1)$ and $Y_2 \sim \text{Poiss}(\lambda_2)$ where λ is the vector of rates we would see under our first order physical model. Instead, for nearby pixels, we see:

$$\begin{aligned} \textbf{Blurred Likelihood: } & Y_1 \sim \text{Poiss}(\nu_1\tau), \quad Y_2 \sim \text{Poiss}(\nu_2\tau) \\ & \nu_1 = 1/2(1+\epsilon)\lambda_1 + 1/2(1-\epsilon)\lambda_2 \\ & \nu_2 = 1/2(1-\epsilon)\lambda_1 + 1/2(1+\epsilon)\lambda_2 \end{aligned} \tag{2}$$

for some small ϵ. In other words, the ideal rates get blurred. This happens for a variety of reasons. First, scattering particles may make more than one bounce, as passing through the image, or may bounce off at an angle and strike a detector that is out of line with the emitter. Second, since we collect counts over time, the objects within the CT will usually move about over the course of imaging. Under thermal fluctuations, these motions are mostly Gaussian jitter, which blurs the image in nearby pixels. Finally, our beam may not be perfectly focused, introducing blur and aberration.

Let's see what small ϵ implies to illustrate the divide between "generic" mathematical and application.

(a) Let $B(\epsilon)$ denote the matrix such that $\nu = B(\epsilon)\lambda$. Solve for the Fisher information as a function of ϵ, ν, λ, and τ, setting $\tau = 1$.

(b) Use the Fisher information to write down the asymptotic normal approximation to the posterior. To do this, first solve for the Fisher information with $\tau = 1$, then replace

n with τ in the asymptotic normal theorem.[2] *You do not need to solve for the inverse Fisher information explicitly.*

(c) Find the eigenvalue decomposition of the covariance of the asymptotic normal approximation to the posterior, assuming $\lambda_1 = \lambda_2$. Identify the eigenvector corresponding to the largest eigenvalue, and the standard deviation along that direction. Identify the eigenvector corresponding to the smallest eigenvalue, and the standard deviation along that direction. Express your answers as functions of ϵ. *It will help to group the entries of the Fisher information into pairs of matching terms, and to find the eigenvalue decomposition for the generic form before plugging in. Your answer here should be simple. If it isn't, back up.*

(d) What happens as ϵ approaches zero, assuming $\lambda_1 = \lambda_2$? Identify the direction of maximal uncertainty in λ and the proposed standard deviation along that direction. Why must the effect you observe be true?[3] Would this effect change if $\lambda_1 \neq \lambda_2$?

(e) Derive the form (up to normalization) for the posterior distribution of λ given an observation $Y = y$ using independent, identical Gamma priors on the rates λ_1 and λ_2. You do not need to simplify your answer. You will use it numerically in parts (f - i).

(f) Write a code that can plot the normal approximation to the posterior and the true posterior over $\lambda_1 \geq 0, \lambda_2 \geq 0$. You should display the densities as colored surfaces whose colors indicate the height of the surfaces. You should be able to borrow much of this code from Lab 3.

(g) Write a code that plots the marginal distributions of these densities along the directions of maximal and minimal variance (w.r.t. the normal approximation).

(h) Plot the results for $\lambda = [0.2, 0.25]$ and $y = \text{round}(\tau\lambda)$ Use a Gamma prior with scale parameter 0.15 and shape parameter 2. Make plots for $\epsilon = 0.5$, 0.1, and 0.01. For each ϵ start with $\tau = 10$. Then, increase τ until the posterior normal approximation is decent.[4]

(i) Comment on how close the posterior is to its normal approximation in each case. Comment on how the match breaks down as ϵ decreases, and how large you needed to make τ in each case. Comment on which marginal distribution best illustrates the failure of the asymptotic theory, and whether the posterior does appear normal along any marginal.

Outro: Always beware an asymptotic gaurantee. Whenever you are given an asymptotic gauarantee you should ask, "How far in to the limit do I need to be before this approximation is any good?" It is often true that the asymptotic guarantees hold for all but some special cases, but the associated approximations fail spectacularly outside of the limit when we are close to the special cases. This moral will come back to haunt us when we build samplers.

[2] This is valid since, you could imagine subdividing a time period of length τ into τ time periods of length 1 if τ is an integer. Then, the total number of observed counts over all time periods is sufficient, Poisson distributed, and the total number of intervals is τ. Notice that this arguments works whether or not τ is integer-valued since it should work for any choice of time unit, and, since it gives back the same answer no matter the length of the time intervals we use to subdivide τ.

[3] Think about the linear system defining ν as a function of λ. Is it invertible when $\epsilon - 0$?

[4] It may help to analyse the equation you derived for the standard deviation in the posterior normal approximation along the direction of maximal uncertainty. How must τ scale in ϵ to prevent the standard deviation from diverging?

Problem 2: Does Preference Imply Belief?

Introducing a joint model over both the unknown and the data can make certain decision problems easier to solve, and offers stronger, more consistent solutions. In class we posed a joint statistical model over the unknown and the data, a loss function, then showed that we could select a decision rule by minimizing the posterior loss conditional on the observation, for all possible observations. There is a deeper representation theorem here. It is largely credited to Leonard Savage (*The Foundations of Statistics*, 1954), who built on the works of Von Neumann and Morgenstern.

Savage argued that statistics, in particular, prior distributions over unknowns, can sensibly reflect unstated beliefs about the state of an unknown inherent in a loss function and set of preferences for different decision rules. In particular, he argued that, if a decision maker adopts a sufficiently self-consistent set of preferences over decision rules, a loss function over individual instances, and a likelihood, *then the preference over decision rules is equivalent to the preference order induced by computing the expected posterior loss for some uniquely specified a prior distribution.* This essentially runs our argument in reverse. Instead of deriving a particular decision rule (our most preferred rule) from a likelihood, prior, and loss, Savage showed that every likelihood, loss, and preference relation over decisions, implies a prior. The axioms needed to ensure the preference set could correspond to prior are frustratingly hard to find laid out in clear language. The Stanford Encyclopedia of Philosophy makes a reasonable attempt.

Savage's argument is the foundational pillar for most of *subjective* Bayesian statistics. It ensures that, if a user has reasonably self-consistent preferences over decision rules given a loss, then they are behaving as if they are Bayesian with respect to some prior. The prior may or may not match their actual beliefs about the world, but they behave as if it did. In this sense, the prior is really an expression of the user's preferences, and the mathematics of Bayesian statistics are a convenient formal system for performing computation. This framework is especially popular in economics, since it expresses revealed beliefs as the necessary consequence of some notion of utility or preference. From the "objective" statistician's perspective it is profoundly disturbing. If your prior is just an expression of your preferences, then any "inference" with respect to it is entirely subjective.

Let's work a bit with this idea. The two problems below ask you to walk between the three components of expected utility theory to show that fixing two (e.g. a statistical model and loss) implies the third (e.g. a preference relation over decision rules).

(a) Suppose that I am performing a simple binary hypothesis test. I adopt a likelihood ratio test with the threshold $k = 15$ (that is, I decide for hypothesis 1 if it would make the data 15 times more likely than under hypothesis 0). I state that the loss incurred by a false positive (decide 1, reality 0) is three times worse than a loss under a false negative (decide 0, reality 1). You ask why I chose $k = 15$, and I say that I selected $k = 15$ since it is my favorite among all other possible choices of thresholds. Assuming the rest of my preference relations satisfy Savage's axioms (unstated), what is my implied prior distribution over the two hypotheses?

(b) It is common, in a variety of on-line learning problems (problems where you collect data as you make decisions, and your decisions influence what data you see), to select options based on a posterior upper credible bound. A version of this rule would say, "If, given my current knowledge about the world, option j out of my n options could credibly have

the largest value, then I will pick j." To implement this rule, we look at the posterior probability distribution over the value of each option, select a probability $p \approx 1$ with which we hope to certify our bounds (the larger p the more credible the bound), and, for each option, compute the p^{th} percentile of its possible value. If V_j is the unknown value of option j, then this is an upper bound b_j such that $\Pr(V_j \leq b_j) = p$. You used a rule of this kind to optimize your research allocation in Lab 2.

Find a loss function over the estimated value of an unknown that would justify this procedure. That is, find a loss function $l(V, \hat{v}(y))$ for which the Bayes optimal decision rule, for all choices of p, is to set $\hat{v}(Y)$, where $\hat{v}(Y)$ equals the p-upper credible bound on $V|Y = y$. *Hint: check the suggested reading from weeks 3 and 4, or the rule in HW 1 that suggested the median as an optimal posterior estimator.*

(c) *Optional: Making Sense of Savage.* Navigate to the Stanford Encyclopedia of Philosophy and try to restate each of Savage's axioms in your own words. Use as plain language as you can. Do these axioms make sense for an idealized preference relation?[5] Those pursuing Project Mini 1 may find this reading useful.

Problem 3*: Agnostics Hold Opinions

It's surprisingly hard to pose an uninformative prior over continuous valued unknowns (see BDA chapter 2.8). This fact is inevitable for continuous random variables.

Any continuous random variable must admit a density function. Density functions are not invariant under transformations of the underlying space, since the density is defined with respect to that space. Remember that a probability density is the expected number of samples in a small region of space, in the limit as the volume of the region vanishes. In other words, its units are probability mass *per volume*. Volumes depend on the units used to measure the unknown (i.e. the length of an interval changes if you change units), and, may change differently in different parts of a parameter space if we change coordinates nonlinearly. Thus, a uniform distribution under once choice of coordinates for the unknowns may not be uniform under another.

As a result, even our best attempts at picking an agnostic prior still imply a specific set of beliefs, namely, uniformity in a specific set of coordinates. This is an important lesson to remember. If you choose to use a Bayesian approach in a continuous space, you must always express some belief.

In this problem, you will practice using our best tool for finding "uninformative" priors.

(a) BDA 2.12: Suppose $Y|\theta \sim \text{Poisson}(\Theta)$. Apply the invariance principle to find Jeffreys' prior density for Θ.

(b) Use the invariance principle to find Jeffrey's prior for the standard deviation of a 1-dimensional Gaussian with known mean. How does your prior compare to:

- The choice recommended for scale parameters in BDA 2.8 (see *Pivotal Quantities*)?
- The conjugate prior family used for the variances of a standard normal? Is Jeffrey's prior an instance of the conjugate family? If so, for what parameters?

[5] They certainly don't match real behavior.

(c) BDA 2.17: The fact that uninformative densities are hard to find follows from the fact that density functions change shape under nonlinear transformations. Since the endpoint of most Bayesian inference is a posterior interval estimate, this fact also has consequences for the way we define intervals. In particular, since the highest posterior density interval (HPDI) depends on selecting regions of high density, it is not preserved under transformations.

Suppose the quantity $V = \frac{n}{\sigma^2}Y$ is observable, where Y is χ^2_n distributed. Suppose that σ has the (improper) non-informative prior density $p(\sigma) \propto \sigma^{-1}$ for all $\sigma > 0$.

- Find the corresponding prior density for σ^2 and $\log(\sigma)$. In what coordinate system is this prior uniform?
- Show that the 95% highest posterior density interval (HPDI) for σ^2 is not the same as the region obtained by squaring the endpoints of the 95% HPDI for σ.

Problem 4: Conjugacy Practice

(a) Show that, if $Y|\theta$ is exponentially distributed with rate θ, then the gamma distribution forms a conjugate prior for θ.

(b) Show that, a gamma prior is also conjugate if, given θ, the sequence $Y^{(n)} = \{Y_j\}_{j=1}^n$ are drawn i.i.d. from an exponential distribution with rate θ.

(c) Find all sufficient statistics in this scenario (a set of functions of $Y^{(n)}$ that uniquely specify the posterior distribution). How many do you need? Do they correspond to any standard summary statistics?

Problem 5*: Maximizing Mutual Information

Given a prior distribution for some unknowns, and a set of possible measurements, each with its own likelihood, what measurement should we pick to gain the most information about the unknown?

This is a special form of a Bayesian decision theory problem. It is widely used to optimize experimental design, and if iterated, can construct optimal codes, compression algorithms, and can maximize channel capacity (communication rate). It can also be used to construct optimal search procedures that iteratively select the measurement which is expected to provide the most information about the unknown given our current knowledge about the unknown. You've seen versions of this question in Labs 2 and 3. Now we have enough tools to solve it elegantly for a moderately general class of models. Our main tool here will be conjugate normal, normal models. If you need a refresher, check your work from Lab 3, or Chapters 2.5 and 3.3 in BDA.

Let $X \sim p_X$ denote an unknown, real-valued state vector taking values in $\mathbb{R}^d$. Let $\mathcal{M}$ equal a set of possible measurements we could make (questions we could ask about the unknown), and $m \in \mathcal{M}$ denote a particular measurement. After measurement m, we observe $Y \sim p_{Y|X,m}$. Note that, since different measurements have different likelihoods, they will induce distinct posteriors.

To choose the measurement m we consider the following three strategies:

1. Choose a measure of uncertainty $U[p]$ that can evaluate the degree of uncertainty in a distribution, p. For example, $U[p]$ might be some measure of the total variance in the distribution p (trace of the covariance matrix). Then, select m to minimize the expected uncertainty left over after a measurement: $m_* = \text{argmin}_{m\in\mathcal{M}}\{\mathbb{E}_Y[U[p_{X|Y,m}]]\}$.

2. Choose the measure that maximizes the mutual information between the observed outcome and the unknown. Mutual information is, equivalently, the expected reduction in uncertainty after a measurement, for an appropriate choice of uncertainty measure, and, a measure of how easy it is to show, in a hypothesis test, that the unknown and the measurement are coupled (not independent). That is, select $m_* = \text{argmax}_{m\in\mathrm{M}}\{I(X;Y|m)\}$ where $I(X;Y)$ is the mutual information between two random quantities.

3. We could, instead, take an adversarial approach, where our goal is to select the measurement that is expected to lead to the largest change in our beliefs. That is, we want the measurement that is expected to teach us the most, by inducing the largest change when moving from prior to posterior. Let $D(p||q)$ be a divergence (that is, a measure of how distinct two distributions are). Then, select $m_* = \text{argmax}_{m\in\mathcal{M}}\{\mathbb{E}[D(p_{X|Y,m}||p_X)]\}$.

Amazingly, all three approaches are the same if we adopt the right definition of uncertainty, information, and divergence. In particular, if we adopt the Shannon entropy, H, as our measure of uncertainty, and the KL divergence as our measure of divergence, then all three methods are equivalent to maximizing the mutual information between the measurement and the unknown, where information is an expected reduction in entropy. This remarkable result gives a simple aim:

$$\begin{aligned}&\textbf{Find: } m_* = \text{argmax}_{m\in\mathcal{M}}\{I(X;Y|m)\} = \text{argmin}_{m\in\mathcal{M}}\{\mathbb{E}_Y[H[p_{X|Y,m}]]\}\\&\textbf{where: } H[p] = -\mathbb{E}_{X\sim p}[\log(p(X))].\end{aligned} \tag{3}$$

You ran a version of this algorithm in your second lab, but used standard deviation for U instead of entropy. In this problem, you will show that the information-theoretic framing gives reasonable, simple, and easily implemented answers for the normal-normal model. This answer justifies standard approaches in signal processing and experimental design, namely, matched filtering.

Consider a Bayesian model on a d-dimensional unknown $X \in \mathbb{R}^d$ of the form:

$$\begin{aligned}&\textbf{Prior: } X \sim \mathcal{N}(\hat{x}, C_x)\\&\textbf{Likelihood: } Y = m^\intercal X + \zeta, \text{ where } m \in \mathcal{M} \subset \mathbb{R}^d \text{ and } \zeta \sim \mathcal{N}(0, \sigma_y^2).\end{aligned} \tag{4}$$

This is a standard measurement model in many applied settings, namely, we make one-dimensional measurements by evaluating some inner-product (projection) of the state vector, and assume that our measurements are corrupted by Gaussian noise. Here $m \in \mathcal{M}$ is the measurement vector. Our freedom to choose different measurements is represented by the freedom to select different projections of the state. It should be intuitive that the best direction will depend on C_x and will aim to constrain the covariance along directions with large variance.

To proceed we will need to know the entropy of a Gaussian distribution. If $Z \sim \mathcal{N}(\hat{z}, C_z)$ then $H[p_Z] = \frac{1}{2}\log(\det(C_z)) + c_d$ where c_d is a constant that only depends on the dimension of Z, and where $\det(\cdot)$ denotes the determinant.[6]

[6] $c_d = \frac{d}{2}(1 + \log(2\pi))$.

Using this fact, let's select the information theoretic optimal measurement direction, m, for any normally distributed state vector evaluated with a linear measurment and normal measurement error.

(a) Derive the formula for $p_{X|Y=y,m}$ for any observation y and measurement direction m. You may reference your notes from Lab 3. Remember, the normal-normal model is conjugate, so the posterior will remain normal. This means, you only need to recall the formulas that update the mean and covariance in X after observing $Y = y$ for a measurement m.

(b) Show that the change in entropy between the prior p_X and the posterior $p_{X|Y=y,m}$ does not depend on the change in the means of either distribution, or the dimensional constant c_d.

(c) Show, in addition, that the change in entropy, $H[p_X] - H[p_{X|Y=y,m}]$, does not depend on the actual measured value, but is a deterministic function of m alone![7] Use this fact to give a closed formula for the information acquisition function, $I(X;Y|m) = \mathbb{E}_{Y|m}[H[p_X] - H[p_{X|Y,m}]]$.

(d) Using rules for logs and determinants[8], show that the information acquisition function (information gained if we use the measurement m), can be written:

$$\begin{aligned} I(X;Y|m) &= \frac{1}{2}\log(1 + \mathrm{SNR}(m; C_x, \sigma_\zeta^2)) \\ \text{where: } \mathrm{SNR}(m; C_x, \sigma_y^2)) &= \frac{m^\intercal C_x m}{\sigma_y^2}. \end{aligned} \tag{5}$$

(e) Show that maximizing the information acquisition function over possible measurements is equivalent to maximizing the signal-to-noise ratio, $\mathrm{SNR}(m; C_x, \sigma_y^2))$ in the measurement, and, that this is a valid signal-to-noise ratio by arguing that $m^\intercal C_x m$ is the variance in Y given m in the absence of noise.

(f) Maximizing the signal-to-noise ratio with respect to m is equivalent to maximizing the quadratic form $m^\intercal C_x m$ over $\mathcal{M}$. Since covariance matrices are positive semi-definite, if $\mathcal{M}$ is convex, then the solution will always lie at the boundary of the set of available measurements. Suppose, for simplicity, that $\mathcal{M} = \{m \text{ such that } \|m\|^2 \leq 1\}$. That is, the unit ball. Show then, that m_* must be parallel to the leading principal component of C_x (the eigenvector of C_x whose eigenvalue is the largest among all eigenvalues of C_x).[9]

[7] This is a striking feature of the normal-normal model. It means we can pick the best sequence of measurements before collecting data. It follows since the rules that update the precision of the posterior depend only on the measurement set-up, not the observed outcome.

[8] It will help to remember that you can always factor a covariance matrix into a product of two symmetric matrices (it's "square root"), that the determinant of a product is the same as the product of the determinants, in any order, and that the determinant is the product of the eigenvalues of a matrix. If you get stuck, look up the formula for the determinant of a rank-one perturbation of the identity matrix.

[9] Trying to solve this optimization problem for other $\mathcal{M}$ produces elegant geometric optimization problems. For example, when $\mathcal{M}$ is a polytope, the best choice of m is the solution to a quadratic programming problem.

Problem 0: Spot Grading

Select two problems from the problems marked with an asterisk for spot grading.