Stat 238, Spring 2025
Homework 3

Alexander Strang
Due: by **11:00 am** Tuesday, April 1, 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.

This Homework focuses on the regularizing effects of standard priors in multivariate linear regression problems. These problems are heavily algebraic, but this algebra is worth doing once for yourself and is highly scaffolded. Start early, and, if you get stuck, ask us questions in office hours or on Ed. These are such important and widely used problems that developing faculty with the basic machinery is a useful skill for real practice. For an example, at scale, using carefully crafted, empirically motivated normal priors, see Ron Cohen's work monitoring greenhouse gas emissions in the Bay Area from a distributed sensor network.

All problems on this assignment are eligible 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*: Instability of the MLE

Suppose that $Y|x \sim \mathcal{N}(Ax, C_y)$ for some $A \in \mathbb{R}^{m\times d}$, $x \in \mathbb{R}^d$, and C_y a valid $m \times m$ covariance matrix. Assume that C_y is invertible.

Let's study the MLE estimator in this case. This is an essential case since it allows explicit algebraic insight into the behavior of unstable estimators, and is the most common statistical estimator in applied problems, especially large applied problems.. Getting good at this algebra is a useful applied skill.

(a) Show that the MLE estimator for x given y is the solution to the least squares problem:

$$\hat{x}_{\text{MLE}}(y) = \text{argmin}\{\|Ax - y\|_{C_y^{-1}}{}^2\} \text{ where } \|u\|_M^2 = u^\intercal M u. \tag{1}$$

(b) Show that, there exists a linear change of coordinates, $\tilde{y} = Ty$ for some T invertible, $m \times m$ such that $\tilde{y}|x \sim \mathcal{N}(\tilde{A}x, I_{m\times m})$ where $I_{m\times m}$ is the $m \times m$ identity and where $\tilde{A} = TA$. Use this fact to argue that, we can assume, without loss of generality, that $C_y = \sigma_y^2 I$ for some σ_y^2.[1]

[1] It can be useful to keep some measure of the overall noise level, σ_y^2 to study how answers depend on the noise variance. Adopting this convention changes coordinates to "whiten" the noise, that is to consider i.i.d. Gaussian noise, while retaining a parameter that explicitly models the noise variance.

(c) Given ill-conditioned A[2], or singular A, maximum likelihood estimation is unstable. Show that, if we write $y = Ax + \zeta$ for $\zeta \sim \mathcal{N}(0, \sigma_y^2 I)$, then:

$$\hat{x}_{\text{MLE}}(y) = x + A^\dagger \zeta \text{ where } A^\dagger = (A^\intercal A)^{-1} A^\intercal \tag{2}$$

is the pseudo-inverse of A.[3] Expand the pseudo-inverse of A using its singular-value decomposition, $A = USV^\intercal$, to show that components of the noise vector ζ projected onto the singular vectors corresponding to the smallest singular value of A, $s_{\min}(A)$, may be greatly amplified when computing the MLE if $s_{\min}(A)$ is small.

(d) Since the MLE is unbiased, show that the standard deviation of the error in estimation, is:

$$\mathbb{E}_\zeta[\|\hat{x}_{\text{MLE}}(y)\|_2{}^2]^{1/2} = \langle A^\dagger A^\dagger, C_y \rangle \text{ where } \langle B, D \rangle = \sum_{i,j} b_{ij} d_{ij}. \tag{3}$$

Use Equation (3) to show that, if $C_y = \sigma_y^2 I$, then the standard error in the estimator is:

$$\mathbb{E}_\zeta[\|\hat{x}_{\text{MLE}}(y)\|_2{}^2]^{1/2} = \sigma_y \|A^\dagger\|_{\text{Fro}} \text{ where } \|M\|_{\text{Fro}}^2 = \sum_{i,j} m_{ij}^2. \tag{4}$$

Then, use the fact that the Frobenius norm of a matrix is the Euclidean norm of the vector of its singular values (see wiki on matrix norm facts), to show that, the standard error in the estimator is:

$$\mathbb{E}_\zeta[\|\hat{x}_{\text{MLE}}(y)\|_2{}^2]^{1/2} = \sigma_y \sqrt{\sum_{i=1}^{\text{rank(A)}} s_i(A)^{-2}} = \mathcal{O}(s_{\min}(A)^{-1}) \tag{5}$$

(e) What if we collect multiple i.i.d. samples $Y^{(n)} = \{Y_1, Y_2, \ldots, Y_n\}$ where $Y_j \sim \mathcal{N}(Ax, \sigma^2 I)$ are drawn i.i.d? Show that, this is equivalent to drawing a single sample $y^{(n)} = [y_1, y_2, ..., y_n] \sim \mathcal{N}(A^{(n)} x, \sigma^2 I_{nm \times nm})$ where $A^{(n)}$ is the block matrix:

$$A^{(n)} = \begin{bmatrix} A \\ A \\ \vdots \\ A \end{bmatrix}, \text{ and } I_{nm \times nm} = \begin{bmatrix} I_{m \times m} & & & \\ & I_{m \times m} & & \\ & & \ddots & \\ & & & I_{m \times m} \end{bmatrix} \tag{6}$$

and, thus, that:

$$\hat{x}_{\text{MLE}}(y^{(n)}) = x + A^\dagger \bar{\zeta}^{(n)} \text{ where } \bar{\zeta}^{(n)} = \frac{1}{n} \sum_{j=1}^{n} \zeta_j. \tag{7}$$

(f) Use your result from part (e) to show that, given n i.i.d. observations, the standard error in the estimator is:

$$\mathbb{E}_\zeta[\|\hat{x}_{\text{MLE}}(y)\|_2{}^2]^{1/2} = \frac{\sigma_y}{n^{1/2}} \|A^\dagger\|_{\text{Fro}} \tag{8}$$

Use this conclusion to argue that, to maintain a constant standard error as $s_{\min}(A)$ vanishes, you would need $n = \mathcal{O}(s_{\min}(A)^{-2})$ samples.

[2] A admits singular values of very different scales

[3] Hint: recall the normal equations used to solve any unweighted least-squares problem.

Problem 2*: Regularization via Prior Information

Problem 1 shows that, when the matrix A (or, TA if starting with $C_y \neq \sigma^2 I$), has small singular values, attempting to invert A via a MLE amplifies errors proportional to the reciprocal of the smallest singular value. Since many problems of interest use A with some very small singular values, direct MLE is hopelessly unstable for many real problems, even for reasonably large n.

Regularized estimators attempt to reduce the sample variance in the estimator by introducing a bias. The bias should move the estimator towards solutions we believe, a priori, are more plausible, or, in a direction such that the induced errors are acceptable. Commonly, naive MLE solutions can be rejected on the grounds that they return implausibly large estimated x. This leads to the general regularized estimator:

$$\hat{x}(y; \lambda, \mathcal{R}) = \text{argmin}\{\|Ax - y\|^2_{C_y^{-1}} + \lambda \mathcal{R}(x)\} \tag{9}$$

where $\mathcal{R} : \mathcal{X} \to \mathbb{R}$ is the regularizer. In most applications, $C_y \propto I$ (see the whitening argument in problem 1), and $\mathcal{R}(x)$ takes the form:

$$\mathcal{R}(x) = \|Rx\|_p{}^p \text{ where } \|u\|_p{}^p = \sum_{j=1}^{d} |u_j|^p \tag{10}$$

for some $p \in [0, \infty)$. If $\mathcal{R}$ takes the form of Equation (10) then we say that we are using L_p regularization.

(a) Show that, if R is invertible, then letting $\tilde{x} = Rx$, we can solve for the regularized estimator by solving the regularized problem:

$$\hat{\tilde{x}}(y; \lambda, \|\cdot\|_p{}^p) = \text{argmin}\{\|\tilde{A}\tilde{x} - y\|^2_{C_y^{-1}} + \lambda \|\tilde{x}\|_p{}^p\} \text{ with } \tilde{A} = AR^{-1} \tag{11}$$

then setting $\hat{x} = R^{-1}\hat{\tilde{x}}$. Argue, then, that, without loss of generality[4], we can change coordinates in x and y so that the regularized least-squares problem takes the canonical form:

$$\hat{x}(y; \lambda, \|\cdot\|_p{}^p) = \text{argmin}\{\|Ax - y\|_2{}^2 + \lambda \|x\|_p{}^p\}. \tag{12}$$

(b) Adopting a regularized estimator expresses a preference relation over estimation procedures. In particular, it adopts a notion of a best estimator. As usual, any preference relation over decision procedures, can, with a fixed likelihood, be expressed equivalently as a Bayesian estimation procedure for a matched choice of prior. Show that:

1. Solving for the regularized estimator defined by equation (12) is the same, for any λ, as solving for:

$$\hat{x}(y; \delta, \|\cdot\|_p{}^p) = \text{argmin}\{\|Ax - y\|_2{}^2 \text{ given } \|x\|_p \leq \delta\} \tag{13}$$

for some δ that depends on λ. For what choice of prior is $\hat{x}(y; \delta, \|\cdot\|_p{}^p)$ the MAP estimator? In what sense is this a weakly informative prior? In what sense is it strongly informative?

[4]Beware what is hidden when we claim WLOG. Reducing problems without loss of generality to canonical forms is useful for analysis, but hides explicit degrees of freedom in the model implicitly in the terms that collect the necessary coordinate transformations. Here, the matrix A is storing all the information about C_y and R that was dropped by changing coordinates.

2. The previous equivalence only matches the optimal estimation procedure under two different approaches. If we adopt the objective function $f(x, y) = \|Ax - y\|_2{}^2 + \lambda\mathcal{R}(x)$ as a measure of the quality of an estimate, x, and prefer any estimate with a smaller objective, then the regularization implicitly expresses a complete prior. Find a prior such that, the preference rule, "*given y, prefer x to x' iff $p_{x|y} > p_{x'|y}$*"[5], is the same as adopting the preference rule, "*given y, prefer x to x' iff $f(x, y) > f(x'.y)$*". Show that the MAP estimator under this prior is the solution to the regularized least squares problem.

(c) Which of these two prior specifications do you prefer? Which more sensibly represents the regularized estimator as an optimal Bayesian estimator? Which expresses stronger beliefs about the unknown x?

Problem 3*: Bias in Regularized Estimators

How does regularization bias our estimates?

As in Problem 1, we will assume, WLOG (but with some loss of explicit clarity), that $C_y = \sigma_y^2 I$, and that $R = I$. We will, for this problem, assume that $p = 2$.

(a) Show that, the L_2 regularized estimator:

$$\hat{x}(y; \lambda) = \text{argmin}\{\|Ax - y\|_2{}^2 + \lambda\|x\|_2{}^2\} \tag{14}$$

returns the MAP estimator for a normal, normal model, where $x \sim \mathcal{N}(0, C_x)$ for some choice of C_x that depends on λ.

(b) The L_2 regularized estimator admits an explicit analytic solution. Show that the L_2 regularized estimator is the solution to the augmented least squares problem:

$$\hat{x}(y; \lambda) = \text{argmin}\{\|\hat{A}(\lambda)x - \hat{y}\|_2{}^2\} \tag{15}$$

where:

$$\hat{A} = \begin{bmatrix} A \\ \lambda I \end{bmatrix}, \quad \hat{y} = \begin{bmatrix} y \\ 0 \end{bmatrix} \tag{16}$$

(c) By converting into a standard least squares problem, we can borrow the standard solution, which follows from the augmented normal equations:

$$\hat{A}^\intercal \hat{A}\hat{x} = \hat{A}^\intercal \hat{y}. \tag{17}$$

Show that the augmented normal equations are solved by:

$$\hat{x}(y, \lambda) = (A^\intercal A + \lambda I)^{-1} A^\intercal y \tag{18}$$

then show that this equation also returns the posterior mean estimator for $x|y$ under an appropriate normal, normal model (see part (a)).

[5] that is, prefer estimates with higher posterior probability

(d) Extend this solution to the case when we draw n i.i.d. vectors y using the same arguments from problem 1 part (e). Show that:

$$\hat{x}(y^{(n)}, \lambda) = \left(A^\intercal A + \frac{\lambda}{n} I\right)^{-1} A^\intercal \bar{y}^{(n)} \text{ where } \bar{y}^{(n)} = \frac{1}{n}\sum_{j=1}^{n} y_j. \tag{19}$$

(e) Use Equation (19) to solve for the bias:

$$b(x, \lambda) = \mathbb{E}_{Y^{(n)}|x}[\hat{x}(Y^{(n)}, \lambda) - x]. \tag{20}$$

Show that the bias takes the form $M(\lambda, n)x$ for some matrix $M(\lambda, n)$ that depends on λ, A, and n. Then, expand using the singular value decomposition of A, $A = USV^\intercal$ to show the following facts about the bias:[6]

1. The bias is proportional to $\|x\|_2$, and, the relative bias $(\mathbb{E}_{Y^{(n)}|x}[\hat{x}(Y^{(n)}, \lambda) - x]/\|x\|_2$ is constant in $\|x\|_2$ for fixed $x/\|x\|_2$. That is, the relative bias depends only on the direction of the vector x, but the absolute bias grows with larger x.
2. Find the direction $x/\|x\|_2$ of inputs that produce the largest bias (in magnitude). Show that it aligns to a singular vector of A, and that, the direction associated with the largest bias (strongest regularization effect/shrinkage) is the direction that was expected to contribute the most to the expected magnitude of the error in the MLE estimator (the direction that is most amplified by $A^\dagger$, and, that is least well-resolved by the likelihood alone).
3. The regularized estimator is always conservative in the sense that the bias always has a negative projection onto x. That is, the regularization "shrinks" the estimator. To study the "shrinkage" induced bu the regularization, decompose x onto the singular vectors of A. Show that the bias in every component "shrinks" each component, the degree of shrinkage (magnitude) of the bias is monotonically increasing in λ, and is monotonically decreasing in the associated singular value.
4. Show that the degree of shrinkage (magnitude of the bias) along each component is monotonically decreasing in n. In particular, show that the estimator is asymptotically unbiased for large n, converges to the MLE, and, thus, must be consistent.
5. Argue that, if, as is generally true, the singular values of A span many orders of magnitude, then, for most reasonable λ there exist a set of directions J along which $\lambda \gg ns_j(A)^2$ if $j \in J$ and a set of directions j along which $\lambda \ll ns_j(A)^2$ if $j \in J$. Show that, along the former directions, the expected regularized estimator is strongly shrunk to zero, while along the latter the expected regularized estimator is essentially unbiased. How are the strongly shrunk directions related to the directions that made the MLE unstable?[7]

[6] It may help to rotate your coordinate system for x by working in the coordinate system $\tilde{x} = V^\intercal x$. This will diagonalize the problem.

[7] This is standard for most regularized least squares problems. The estimator is essentially unbiased along a subset of directions where the likelihood is sharply resolved and is highly biased towards zero along a subset of directions where the likelihood is not specific. It also reveals a general phenomena in practical Bayesian inference with weakly informative priors: the likelihood dominates along some well resolved direction, while the posterior is constrained in the remaining directions by the prior.

Problem 4*: Variance-Bias Trade-offs

What λ should we use?

Increasing λ stabilizes inference, but expresses stronger beliefs/constraints about/on the unknown, so produces larger biases. This is a bias-variance trade-off.

A variety of heuristic rules are widely used to balance the bias-variance trade-off. For example, Morozov's discrepancy principle suggests that the user should pick λ as large as possible while keeping the likelihood of $\hat{x}(y; \lambda, \mathcal{R})$ sufficiently large. This usually amounts to starting with large λ, and gradually decreasing λ, until the discrepancy between $A\hat{x}(y; \lambda)$ and y is plausibly small (that is, on the scale expected under $y \sim \mathcal{N}(Ax, C_y)$. This approach adopts the heuristic "*regularize as much as possible while retaining fidelity to the data.*"

In this problem, you will show that, if the regularized estimator is derived under a Bayesian approach, then the prior induces a λ that optimally balances the estimation errors introduced by bias and variance. We will analyze the bias-variance trade-off in a L_2 regularized estimator, then show that, if we adopted a normal prior on x, the regularization parameter λ can be selected automatically based on our prior beliefs.

This problem proceeds under the same assumptions as Problem 3.

(a) Compute the sample covariance $\text{Var}_{Y^{(n)}|x}[\hat{x}(Y^{(n)}, \lambda)]$.

(b) Use the singular value decomposition of A to show that the sample covariance is aligned to the singular vectors of A (the principal components, e.g. eigenvectors of the covariance are parallel to the singular vectors of A). Provide a formula for the standard deviation along each principal component of the sample covariance.

(c) Using your result from part (b), show that:

1. The sample standard deviations are ordered in decreasing order from the smallest singular values of A to the largest.
2. Increasing λ monotonically reduces the sample standard deviation along each principal component. Thus, while larger λ leads to larger bias, it always reduces the sample variance in the estimator.
3. For large n, the sample standard deviations are all $\mathcal{O}(n^{-1/2})$, and, increasing n reduces the regularizing effect of λ on the sample covariance in the estimator.

(d) Show that, by explicitly expanding the error $\hat{x} - x = \hat{x} - \mathbb{E}[\hat{x}] + \mathbb{E}[\hat{x}] - x$, for any estimator, whose sampling covariance is independent of the true x, the expected standard error in the estimator squared, $\mathbb{E}_X[\mathbb{E}_{Y^{(n)}|x}[\|\hat{x}(Y^{(n)}) - x\|^2]]$, can be expanded:

$$\mathbb{E}_X[\mathbb{E}_{Y^{(n)}|x}[\|\hat{x}(Y^{(n)}) - x\|^2]] = \mathbb{E}_X[\|b(X)\|^2] + \text{trace}(\text{Var}_{Y^{(n)}}[\hat{x}(Y^{(n)})]). \tag{21}$$

Interpret the left hand side as a standard notion of risk (choose Bayesian, posterior, or frequentist and specify the loss function used), to show that a natural notion of risk can be decomposed into the expected magnitude of the bias plus the sampling variance in the estimator. This establishes the bias-variance trade-off explicitly since increasing λ increases the first term, while decreasing the second.

(e) Using the decision-theoretic rational for the posterior mean estimator, and the relation between the posterior mean and MAP estimators in a normal, normal model, argue that,

if $X \sim \mathcal{N}(0, \sigma_x^2 I)$, then the choice of λ that optimizes the trade-off expressed in equation (21) can be derived directly from the prior variance.

(f) Weak Information: If $X \sim \mathcal{N}(0, \sigma_x^2 I)$, but we use $\hat{x}(Y^{(n)}, \lambda)$ for λ derived from $\sigma' > \sigma_x$ (use a weaker prior than the truth), then will our estimator be more or less biased, and more or less variable, then the optimal estimator?

Problem 5*: Robust Inference via Fat-Tailed Priors

In Problem 3 you established that, using a normal prior introduced a bias that increased proportional to the truth. This is a natural consequence of adopting a normal prior, whose tails decay very rapidly. The larger the truth, the more evidence it takes to override the fast decaying tails of the normal prior. In this sense, even for large σ_x, a normal prior induces a strong regularizing effect, and, as noted above, is rarely "weakly informative." This is particularly true for the standard setting where A is ill-conditioned, so some directions are well resolved by the likelihood alone, and are largely unaffected by the prior, while others are essentially unresolved by the likelihood, and are entirely resolved by the prior.

In many applications, we want a regularizer whose tails decay more slowly than the tails of a normal distribution. These are more willing to allow the possibility of outliers, so demand less evidence to convince that the true unknown is large. This is useful if, the true prior distribution is heavy-tailed, or, if we want a regularizer that is more amenable to outliers. Adopting a heavy-tailed regularizer produces an estimator that is more sample efficient when attempting to estimate large unknowns, but is more variable.

Throughout this problem we will assume that $Y^{(n)}|X = x \sim p_{Y|x}$ i.i.d., that the regularity conditions assumed in our unit on consistency and asymptotics apply, and that our model is well-specified. Under those conditions we may assume that, for large n, the likelihood is sharply peaked about the true unknown x_*, the posterior is very small outside of a neighborhood of x near to x_*, and, that the posterior near x_* admits the second-order Taylor expansion:

$$\begin{aligned}\frac{1}{n}\log(p_{X|y}(x|Y^{(n)})) \simeq \\ \frac{1}{n}\left(\nabla_x \log(p_X(x))|_{x=x_*}(x-x_*) + \frac{1}{2}(x-x_*)^\intercal \nabla_x^2 \log(p_X(x))|_{x=x_*}(x-x_*)\right)\ldots \\ -\frac{1}{2}(x-x_*)^\intercal I[x_*](x-x_*)\end{aligned} \tag{22}$$

where $\simeq$ is meant up to an additive constant, $I[x_*]$ is the Fisher information, and convergence occurs in probability over $Y^{(n)}$ as n diverges.[8]

So, if n is large, then the MAP estimator will converge, in probability to the maximizer of Equation (22), which, provided $I[x_*]$ is invertible, is, to lowest order in n:

$$(\hat{x}_{\text{MAP}}(Y^{(n)}) - x_*) \xrightarrow[n\to\infty]{} -(nI[x_*])^{-1}\nabla_x \log(p_X(x))|_{x=x_*} \tag{23}$$

where convergence occurs in probability.[9] This equation provides a more general method for analyzing the consistency of posterior point estimators.

[8] Notice that equation 22 recovers the asymptotic normal result, but now includes a correction associated with the vanishing influence of the prior.

[9] Technically, this statement should be written $n(\hat{x}_{\text{MAP}}(Y^{(n)}) - x_*) \xrightarrow[n\to\infty]{} -I[x_*]^{-1}\nabla_x \log(p_X(x))|_{x=x_*}$

In particular, we can use equation (23) to approximate the convergence rate in n of posterior estimators to the true unknown, x_*, as a function of x_*, for different likelihoods and priors.

(a) For consistency with the rest of this HW, we will assume a normal likelihood. Show that the Fisher information for $Y|x \sim \mathcal{N}(Ax_*, C_y)$ is a fixed matrix that is independent of x_*. What is the Fisher information when $C_y = \sigma_y^2 I$? Have you seen this matrix earlier in this HW? If so, what role did it play?

(b) Adopt a normal prior. That is, $X \sim \mathcal{N}(0, \sigma_x^2 I)$ where, without loss of generality, we assume a coordinate system where the entries of X are independent (see Problem 2). Simplify the right-hand side of equation (23) in this case. Show that the right-hand side has the form $-\frac{1}{n}Mx$ for some matrix M that depends on A, σ_x, and σ_y. Try to express the entries of M in terms of a signal-to-noise or noise-to-signal ratio (covariance in Ax versus the noise variance).

(c) Adopt a prior of the form $p_X(x) \propto \exp(-\frac{\lambda}{p}\|x\|_p{}^p)$. Show that, for $p = 1$, $\nabla_x \log(p_X(x))|_{x=x_*} = -\lambda \times \text{sign}(x)$ so the convergence rate of the MAP estimator is independent of the magnitude of the true unknown x_*. In this sense the L_1 prior (Laplace prior), is more robust than the normal prior. It demands less information to believe in large x_*.

(d) Repeat your analysis for a Student's-t prior, $p_X(x) \propto (\nu + \|x\|_2{}^2)^{-(\nu+1)/2}$. This is a prior with very heavy tails. How does the convergence rate of the MAP estimator to x_* in n depend on $\|x_*\|$ now? How does it depend on ν? Try to make sense of your result by reasoning about the tail decay rate of this prior.

(e) Consider the three examples (Normal, Laplace, Student's-t). Use $\nu = 1$ (Cauchy), and $\nu = 3$ for the Student's-t distribution. For your experiments, use a 1-dimensional example with $A = 1$, $\sigma_y = 1$, $\sigma_x = 1$, and $\lambda = 1$. Fix $x_* = 1, 2, 4, 8$, and 16. Sample $Y^{(n)}$ from the corresponding likelihood for $n = 10^m$ where $m \in [0, 4]$ are evenly spaced. To sample, first sample a single sequence $Z^{(n)} \sim \mathcal{N}(0, 1)$ with $n = 10^4$, then set $Y^{(n)} = x_* + Z^{(n)}$ for each desired n and x_*. This fixes the random seed, so that all ensuing simulations only differ in the choice of prior. Make a sequence of comparison plots showing the error in the MAP estimator for the four prior specifications listed above, as a function of m, for the five x_* values. Do your results match your previous analysis/intuition about the regularizing effects of the priors?

Problem 0: Spot Grading

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