STATS 305B: Midterm Exam

Write your name here:
MIDTERM SOLUTIONS

Instructions:

- The exam has 3 questions. There are a total of 98 points on the exam.
- Write on the exam. We will scan it. If you use the extra pages at the back, label clearly.
- You can bring handwritten notes on **one side** of an 8.5x11" piece of paper.
- Unless otherwise specified, you can write your answers using the "named distribution PDF shorthand," e.g. write the pdf of a Gaussian distribution with mean μ and variance σ^2 as $\mathcal{N}(x; \mu, \sigma^2)$.

Some tips:

1. It's usually a good idea to look through the whole exam before taking it to make sure there aren't missing pages; and so that you roughly know what you are up against
2. It's usually a good idea to skip questions if you are stuck and circle back.

Stanford Honor Code

1. The Honor Code is an undertaking of the students, individually and collectively:
    - that they will not give or receive aid in examinations; that they will not give or receive unpermitted aid in class work, in the preparation of reports, or in any other work that is to be used by the instructor as the basis of grading;
    - that they will do their share and take an active part in seeing to it that others as well as themselves uphold the spirit and letter of the Honor Code.
2. The faculty on its part manifests its confidence in the honor of its students by refraining from proctoring examinations and from taking unusual and unreasonable precautions to prevent the forms of dishonesty mentioned above. The faculty will also avoid, as far as practicable, academic procedures that create temptations to violate the Honor Code.
3. While the faculty alone has the right and obligation to set academic requirements, the students and faculty will work together to establish optimal conditions for honorable academic work.

Problem 1: *Exponential Family Distributions (33 pts)*

Consider the gamma distribution for a random variable $X \in \mathbb{R}_+$ with shape parameter $a > 0$ and rate parameter $b > 0$. Its pdf is,

$$\mathrm{Ga}(x; a, b) = \frac{b^a}{\Gamma(a)} x^{a-1} e^{-bx}.$$

(a) (6 pts) Write the gamma density in exponential family form. What are its natural parameters $\boldsymbol{\eta}$, sufficient statistics $\boldsymbol{t}(x)$, and log normalizer $A(\boldsymbol{\eta})$?

(b) (4 pts) Using the log normalizer, compute the expected value $\mathbb{E}[X]$ where $X \sim \mathrm{Ga}(a, b)$.

(c) (4 pts) Using the log normalizer, compute the variance $\mathrm{Var}[X]$ where $X \sim \mathrm{Ga}(a, b)$.

(d) (6 pts) Suppose $\lambda \sim \mathrm{Ga}(a, b)$ and $Y_i \mid \lambda \overset{\text{iid}}{\sim} \mathrm{Po}(\lambda)$ for $i = 1, \ldots, n$, where Po denotes the Poisson distribution with pmf $\mathrm{Po}(y; \lambda) = \frac{1}{y!} e^{-\lambda} \lambda^y$. Compute the posterior distribution, $p(\lambda \mid Y_1 = y_1, \ldots, Y_n = y_n)$.

(e) (7 pts) Compute the Fisher information $\mathcal{I}(\lambda)$ for the parameter λ of the Poisson distribution, $\mathrm{Po}(\lambda)$.

(f) (6 pts) Assume $Y_i \overset{\text{iid}}{\sim} \mathrm{Po}(\lambda^\star)$ for $i = 1, \ldots, n$ for some true mean $\lambda^\star$. Compare the mean and variance of the posterior distribution $p(\lambda \mid Y_1 = y_1, \ldots, Y_n = y_n)$ under the prior $\lambda \sim \mathrm{Ga}(a, b)$ to the mean and variance of the normal approximation of the sampling distribution of the maximum likelihood estimate, $\hat{\lambda}_{\mathsf{MLE}}$, in the large n regime ($n \gg a, b$).

Solution:

(a)

$$\text{Ga}(x; a, b) = \exp\{a \ln x - bx + a \ln b - \ln \Gamma(a)\} x^{-1}$$

Sufficient statistics: $\boxed{\boldsymbol{t}(x) = (\ln x, -x)}$

Natural parameters: $\boxed{\boldsymbol{\eta} = (a, b)}$

Log normalizer: $\boxed{A(\boldsymbol{\eta}) = A(a, b) = \ln \Gamma(a) - a \ln b}$

* The exponential family form is not unique and there are many correct solutions.

(b)

$$\mathbb{E}[X] = -\mathbb{E}[-X] = -\frac{\partial}{\partial b} A(a, b) = \boxed{\frac{a}{b}}$$

(c)

$$\text{Var}[X] = \text{Var}[-X] = \frac{\partial^2}{\partial b^2} A(a, b) = \boxed{\frac{a}{b^2}}$$

(d)

$$\begin{aligned} p(\lambda \mid Y_1 = y_1, \ldots, Y_n = y_n) &\propto p(\lambda, y_1, \ldots, y_n) \\ &\propto p(\lambda) \prod_{i=1}^{n} p(y_i \mid \lambda) \\ &\propto \lambda^{a-1} e^{-b\lambda} e^{-n\lambda} \lambda^{\sum_{i=1}^{n} y_i} \\ &\propto \boxed{\text{Ga}\left(\lambda; a + \sum_{i=1}^{n} y_i, b + n\right)} \end{aligned}$$

(e)

$$\mathcal{L}(\lambda) = -\lambda + y \; ln\lambda - \ln y!$$

$$\begin{aligned} \frac{\partial}{\partial \lambda} \mathcal{L}(\lambda) &= -1 + \frac{y}{\lambda} \\ \frac{\partial^2}{\partial \lambda^2} \mathcal{L}(\lambda) &= -\frac{y}{\lambda^2} \end{aligned}$$

$$\mathcal{I}(\lambda) = -\mathbb{E}\left[\frac{\partial^2}{\partial \lambda^2} \mathcal{L}(\lambda)\right] = -\mathbb{E}\left[\frac{y}{\lambda^2}\right] = \boxed{\frac{1}{\lambda}}$$

(f) First we find the MLE:

$$\mathcal{L}(\lambda) = -n\lambda + \sum_{i=1}^{n} y_i \ln \lambda - \ln y_i$$

$$\frac{\partial}{\partial \lambda}\mathcal{L}(\lambda) = -n + \frac{1}{\lambda}\sum_{i=1}^{n} y_i$$

$$\hat{\lambda}_{MLE} = \frac{1}{n}\sum_{i=1}^{n} y_i = \bar{y}$$

The Fisher information for n samples is

$$\mathcal{I}\left(\hat{\lambda}_{MLE}\right) = -\mathbb{E}\left[-\frac{\sum_{i=1}^{n} y_i}{\lambda^2}\right] = \frac{n}{\lambda}$$

$$\boxed{\mathbb{E}[\hat{\lambda}_{MLE}] = \lambda, \quad \mathrm{Var}[\hat{\lambda}_{MLE}] = \mathcal{I}(\hat{\lambda}_{MLE})^{-1} = \frac{\lambda}{n}}$$

$$\boxed{\text{Posterior mean} : \frac{a + \sum_{i=1}^{n} y_i}{b+n}, \quad \text{Posterior variance} : \frac{a + \sum_{i=1}^{n} y_i}{(b+n)^2}}$$

When $y_i \overset{\text{iid}}{\sim} \text{Po}(\lambda)$, the expected value of the posterior mean is $(a + n\lambda)/(b + n)$ which goes to λ in the large n limit. Likewise, the expected value of the posterior variance is $(a + \lambda n)/(b + n)^2$, which goes to λ/n. In this limit, the posterior mean and variance match the mean and variance of the MLE in expectation. The Bernstein-von Mises theorem formalizes this relationship.

Problem 2: *Logistic Regression (30 pts)*

Consider a logistic regression model,

$$Y \mid X = x \sim \mathrm{Bern}(\sigma(\beta x))$$

where $\beta \in \mathbb{R}$ and $X \in \mathbb{R}$ are scalar weights and covariates, respectively, and $\sigma(a) = 1/(1+e^{-a})$ is the logistic function. Suppose you observe the following set of independent observations,

X_i	Y_i
-1	0
+1	1

Table 1: Two data points for Problem 2

(a) (8 pts) Write the log likelihood function. Simplify as much as possible.

(b) (3 pt) Is the MLE finite? Explain your answer.

(c) (8 pts) Now introduce a Gaussian prior, $\beta \sim \mathrm{N}(0,1)$. Compute the derivative of the log joint probability with respect to β.

(d) (8 pts) Sketch a plot (by hand) of the log joint probability and its derivative, both as functions of β. Be sure to label key points such as the y-intercept of the derivative, and make sure that your graphs are consistent with each other, i.e. satisfy the fundamental theorem of calculus.

(e) (3 pts) Is the *maximum a posteriori* (MAP) estimate finite? Is it unique? Justify your answer.

Solution:

(a)

$$\begin{aligned}\log \Pr(Y \mid X;\beta) &= \log(1-\sigma(-\beta)) + \log(\sigma(\beta)\\ &= 2\log\Big(\frac{1}{1+\exp(-\beta)}\Big)\\ &= \boxed{2\log(\sigma(\beta)).}\end{aligned}$$

Another derivation based on the notes sees that

$$\log \Pr(Y \mid X;\beta) = \beta - \log(1+\exp(\beta)) - \log(1+\exp(-\beta)).$$

(b) No, the data points are perfectly separable, and so the MLE needs to be infinite to make the sigmoid a step function and give both data points probability 1 of being correctly classified.

(c)

$$\log p(Y,\beta|X) \propto \frac{-\beta^2}{2} + 2\log\Big(\sigma(\beta)\Big),$$

and so

$$\boxed{\frac{\partial \log p(Y,\beta|X)}{\partial \beta} = -\beta + 2\Big(1-\sigma(\beta)\Big).}$$

An equivalent way of writing is

$$\frac{\partial \log p(Y,\beta|X)}{\partial \beta} = -\beta + 2 - \frac{2}{1+\exp(-\beta)}.$$

Since $\sigma(\beta) = \frac{1}{1+\exp(-\beta)}$, it follows that an equivalent way of writing this expression is

$$\begin{aligned}\frac{\partial \log p(Y,\beta|X)}{\partial \beta} &= -\beta + 2\Big(\frac{\exp(-\beta)}{1+\exp(-\beta)}\Big)\\ &= -\beta + \frac{2}{1+\exp(\beta)}.\end{aligned}$$

(d) See fig. 1

(e) The MAP is finite and unique because it is the maximum of a concave function.

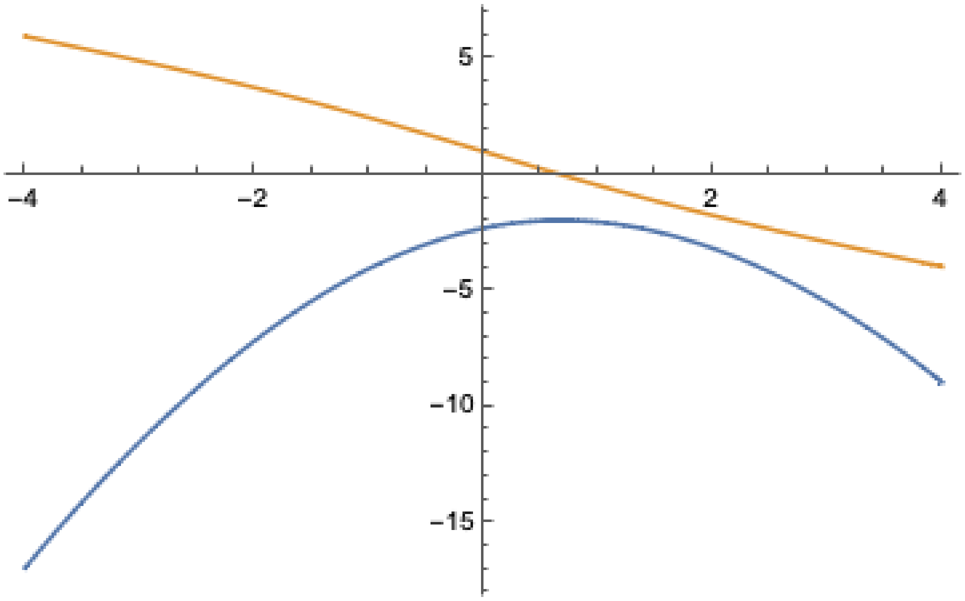

Figure 1: Solution for Problem 2d. Note that the derivative is equal to 1 at $\beta = 0$.

Problem 3: *The Bayesian Lasso (35 pts)*

The Lasso problem is an L_1 penalized least squares problem,

$$\mathscr{L}(\boldsymbol{\beta}) = \frac{1}{2}\sum_{i=1}^{n}(y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta})^2 + \gamma \sum_{j=1}^{p} |\beta_j|. \tag{1}$$

where $y_i \in \mathbb{R}$, $\boldsymbol{x}_i \in \mathbb{R}^p$, and $\boldsymbol{\beta} \in \mathbb{R}^p$. From a Bayesian perspective, minimizing $\mathscr{L}(\boldsymbol{\beta})$ is equivalent to *maximum a posteriori* (MAP) estimation in the following probabilistic model,

$$\begin{aligned} \beta_j &\overset{\text{iid}}{\sim} \text{Lap}(\lambda) \\ y_i &\overset{\text{ind}}{\sim} \text{N}(\boldsymbol{x}_i^\top \boldsymbol{\beta}, \sigma^2), \end{aligned} \tag{2}$$

where Lap(λ) denotes a Laplace distribution with density $\text{Lap}(\beta; \lambda) = \frac{\lambda}{2} e^{-\lambda|\beta|}$.

(a) (5 pts) Find a setting of λ such that the MAP estimate of model (2) is the same as the minimizer of eq. 1. Your solution should be in terms of γ and σ^2.

(b) (8 pts) The Laplace density can also be written as a *scale mixtures of Gaussians*,

$$\text{Lap}(\beta; \lambda) = \frac{\lambda}{2} e^{-\lambda|\beta|} = \int_0^\infty \text{N}(\beta; 0, \nu) \cdot \text{Exp}\left(\nu; \tfrac{\lambda^2}{2}\right) \mathrm{d}\nu = \int_0^\infty \frac{1}{\sqrt{2\pi\nu}} e^{-\frac{\beta^2}{2\nu}} \cdot \frac{\lambda^2}{2} e^{-\frac{\lambda^2 \nu}{2}} \, \mathrm{d}\nu,$$

where we have substituted the density of the exponential distribution to obtain the last equality.

Let $\boldsymbol{y} = \{y_i\}_{i=1}^n$ and $\boldsymbol{X} = \{\boldsymbol{x}_i\}_{i=1}^n$. Use the integral representation above to write a joint distribution,

$$p(\boldsymbol{\beta}, \boldsymbol{\nu}, \boldsymbol{y} \mid \boldsymbol{X}; \lambda, \sigma^2)$$

on an extended space that includes the *augmentation variables* $\boldsymbol{v} = (v_1, \ldots, v_p)$, such that the marginal distribution $p(\boldsymbol{\beta}, \boldsymbol{y} \mid \boldsymbol{X}; \lambda, \sigma^2)$ matches that of the generative model described in eq. (2).

(c) (9 pts) Derive a closed-form expression for the conditional distribution $p(\boldsymbol{\beta} \mid \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{v}; \lambda, \sigma^2)$. *Note that this is a conditional of the augmented model.*

(d) (8 pts) It can be shown that under the augmented model, the auxiliary variables also have tractable conditionals. Specifically, $p(v_j^{-1} \mid \boldsymbol{\beta}, \boldsymbol{y}, \boldsymbol{X}; \lambda, \sigma^2) = \text{IG}(v_j^{-1}; \mu_j, \xi_j)$, where $\text{IG}(\mu_j, \xi_j)$ denotes the *inverse Gaussian* distribution with mean μ_j and shape ξ_j, which are functions of β_j, λ, and σ^2. Use this fact to outline a Gibbs sampling algorithm targeting the posterior distribution of the augmented model. Feel free to reference your answers from part c.

(e) (5 pts) One reason the Lasso is so popular is that it yields sparse estimates – i.e., the posterior mode $\hat{\boldsymbol{\beta}}_{\mathsf{MAP}}$ has exact zeros. Are samples from the posterior distribution also sparse? Explain your answer.

Solution:

(a) Note that the MAP estimate is equivalent to,

$$\begin{aligned}\arg\max -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - x_i^T\beta)^2 - \lambda\sum_{j=1}^{p}|\beta_j| &= \arg\min \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - x_i^T\beta)^2 + \lambda\sum_{j=1}^{p}|\beta_j| \\ &= \arg\min \frac{1}{2}\sum_{i=1}^{n}(y_i - x_i^T\beta)^2 + \lambda\sigma^2\sum_{j=1}^{p}|\beta_j|\end{aligned}$$

since $\sigma^2 > 0$. Thus,

$$\boxed{\lambda = \frac{\gamma}{\sigma^2}.}$$

(b)

$$p(v,\beta,y|X;\lambda,\sigma^2) = \Big(\prod_{j=1}^{p}\text{Exp}(v_j;\lambda^2/2)\mathcal{N}(\beta_j;0,v_j)\Big)\Big(\prod_{i=1}^{n}\mathcal{N}(y_i;x_i^T\beta,\sigma^2)\Big)$$

(c) To get the complete conditional for β, we just have to look at the terms in the joint that use β. Doing so, we see that all the terms are quadratic, and so β will come from a multivariate Gaussian. The quadratic term takes the form

$$J = \text{diag}(1/v_j) + \frac{\sum_{i=1}^{n} x_i x_i^T}{\sigma^2}$$

while the linear term takes the form

$$h = \frac{\sum_{i=1}^{n} y_i x_i}{\sigma^2}.$$

In matrix form, if we let X be an $n \times p$ matrix, and $\mathbf{y}$ be a vector in $\mathbb{R}^n$, then we can write these natural parameters as

$$\begin{aligned} J &= \text{diag}(1/v_j) + \frac{X^TX}{\sigma^2} \\ h &= \frac{X^Ty}{\sigma^2}\end{aligned}$$

By completing the square, it follows that the complete conditional for β is

$$\boxed{p(\beta \mid X, y, v) = \text{MVN}\big(\beta \mid J^{-1}h, J^{-1}\big).}$$

(d)

```
Algorithm 1: Gibbs Sampler for Bayesian Lasso
Input: Data $\mathbf{X}, \mathbf{y}, \lambda, \sigma^2$, number of iterations $T$
Output: Samples from the posterior distribution of $\boldsymbol{\beta}$
Function GibbsSamplerBayesianLasso($\mathbf{X}, \mathbf{y}, N$):
    Initialize $\boldsymbol{\beta}^{(0)}, \nu$ randomly
    for $t \leftarrow 1$ to $T$ do
        for $j \leftarrow 1$ to $p$ do
            Sample $1/\nu_j^{(t)}$ from its IG conditional given $\boldsymbol{\beta}^{(t-1)}$(see problem statement)
        end
        Sample $\boldsymbol{\beta}^{(t)}$ from its MVN conditional given $\boldsymbol{\nu}^{(t)}$ (see soln. 3c)
    end
return $\{\boldsymbol{\beta}^{(t)}\}_{t=1}^{T}$
```

(e) No, we can see from part d that the samples are normal with non-degenerate precision.