

# STATS 305B: Midterm Exam

Write your name here:

## Instructions:

- The exam has 3 questions. There are a total of 100 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  $\mu$  and variance  $\sigma^2$  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 (35 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,

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

(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  $X \sim \text{Ga}(a, b)$ .

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

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

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

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

**Problem 2: Logistic Regression (30 pts)**

Consider a logistic regression model,

$$Y | X = x \sim \text{Bern}(\sigma(\beta x))$$

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,  $\beta \sim N(0, 1)$ . Compute the derivative of the log joint probability with respect to  $\beta$ .

(d) (8 pts) Sketch a plot (by hand) of the log joint probability and its derivative, both as functions of  $\beta$ . 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.

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

The Lasso problem is an  $L_1$  penalized least squares problem,

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

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

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

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

- (a) (5 pts) Find a setting of  $\lambda$  such that the MAP estimate of model (2) is the same as the minimizer of eq. 1. Your solution should be in terms of  $\gamma$  and  $\sigma^2$ .

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

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

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

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

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

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

- (c) (9 pts) Derive a closed-form expression for the conditional distribution  $p(\boldsymbol{\beta} \mid \mathbf{y}, \mathbf{X}, \boldsymbol{\nu}; \lambda, \sigma^2)$ . *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,  $p(v_j^{-1} | \boldsymbol{\beta}, \mathbf{y}, \mathbf{X}; \lambda, \sigma^2) = \text{IG}(v_j^{-1}; \mu_j, \xi_j)$ , where  $\text{IG}(\mu_j, \xi_j)$  denotes the *inverse Gaussian* distribution with mean  $\mu_j$  and shape  $\xi_j$ , which are functions of  $\beta_j, \lambda$ , and  $\sigma^2$ . 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  $\hat{\boldsymbol{\beta}}_{\text{MAP}}$  has exact zeros. Are samples from the posterior distribution also sparse? Explain your answer.

Blank page for extra work (please label clearly for GradeScope scanning).

Blank page for extra work (please label clearly for GradeScope scanning).

Blank page for extra work (please label clearly for GradeScope scanning).