

# STATS 305C: Practice Exam

**Name:**

**Problem 1: Gaussian models**

Consider the following model,

$$\begin{aligned}x_{n,d} &\stackrel{\text{iid}}{\sim} \mathcal{N}(0, \sigma_d^2) && \text{for } n = 1, \dots, N; d = 1, \dots, D \\ \sigma_d^2 &= \prod_{k=1}^d \lambda_k^{-1} && \text{for } d = 1, \dots, D \\ \lambda_d &\stackrel{\text{iid}}{\sim} \text{Ga}(\alpha, 1) && \text{for } d = 1, \dots, D\end{aligned}$$

- (a) Suppose  $\alpha > 1$ . Describe how this *multiplicative inverse gamma* prior affects the distribution of the data,  $x_{n,d}$ . For example, how does the distribution of  $x_{n,1}$  generally compare to that of  $x_{n,D}$ ?
- (b) Let  $\boldsymbol{\lambda} = \{\lambda_k\}_{k=1}^K$  and  $\mathbf{X} = \{\{x_{n,d}\}_{d=1}^D\}_{n=1}^N$ . Derive a Gibbs sampler for the posterior distribution  $p(\boldsymbol{\lambda} \mid \mathbf{X}; \alpha)$ .

**Problem 2: Hierarchical models.**

Recall the probability density function of the gamma distribution,  $p(\lambda; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta\lambda}$ , where  $\Gamma(\cdot)$  is the gamma function. Now consider the following hierarchical model,

$$\begin{aligned} \beta_g &\sim \text{Ga}(\alpha_0, \beta_0) \\ \lambda_g &\sim \text{Ga}(\alpha, \beta_g) && \text{for } g = 1, \dots, G \\ x_{g,n} &\sim \text{Po}(\lambda_g) && \text{for } g = 1, \dots, G; n = 1, \dots, N. \end{aligned}$$

Using the Poisson probability mass function  $p(x | \lambda) = \frac{1}{x!} \lambda^x e^{-\lambda}$ , derive a Gibbs sampling algorithm for this hierarchical model. Specifically, derive the conditional distributions,

- $p(\lambda_g | \{x_{g,n}\}_{n=1}^N, \beta_g; \alpha)$ ,
- $p(\beta_g | \lambda_g; \alpha_0, \beta_0)$ .

**Problem 3: Graphical models.**

(a) Draw the graphical model corresponding to this joint probability distribution,

$$p(\{x_n, y_n\}_{n=1}^N; \alpha, \beta, \gamma) = p(x_1 | \alpha) \left[ \prod_{n=2}^N p(x_n | x_{n-1}; \beta) \right] \left[ \prod_{n=1}^N p(y_n | x_n; \gamma) \right].$$

(b) Write the joint distribution corresponding to the graphical model in Figure 1.

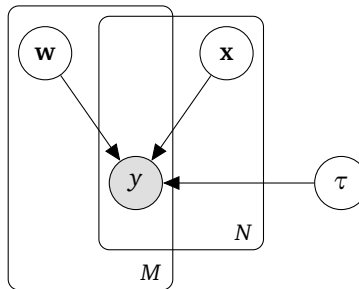


Figure 1

**Problem 4:** *Continuous latent variable models*

Canonical correlation analysis is a technique for paired datasets  $\{(\mathbf{x}_n, \mathbf{y}_n)\}_{n=1}^N$  where  $\mathbf{x}_n \in \mathbb{R}^{D_x}$  and  $\mathbf{y}_n \in \mathbb{R}^{D_y}$ . Like PCA, it can be viewed as a limiting case of a linear Gaussian model latent variable model,

$$\begin{aligned} \mathbf{z}_n &\sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \\ \mathbf{x}_n &\sim \mathcal{N}(\mathbf{W}_x \mathbf{z}_n + \mathbf{b}_x, \Sigma_x) \\ \mathbf{y}_n &\sim \mathcal{N}(\mathbf{W}_y \mathbf{z}_n + \mathbf{b}_y, \Sigma_y). \end{aligned}$$

Derive the conditional distribution  $p(\mathbf{y}_n | \mathbf{x}_n; \boldsymbol{\theta})$  where  $\boldsymbol{\theta} = (\mathbf{W}_x, \mathbf{W}_y, \mathbf{b}_x, \mathbf{b}_y, \Sigma_x, \Sigma_y)$ .

**Problem 5: The Bayesian Lasso**

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

$$\mathcal{L}(\mathbf{w}) = \sum_{n=1}^N \|y_n - \mathbf{x}_n^\top \mathbf{w}\|_2^2 + \lambda_0 \sum_{d=1}^D |w_d|. \quad (1)$$

From a Bayesian perspective, minimizing  $\mathcal{L}(\mathbf{w})$  is equivalent to *maximum a posteriori* (MAP) estimation in the following Bayesian model,

$$\begin{aligned} w_d &\stackrel{\text{iid}}{\sim} \text{Lap}(\lambda) \\ y_n &\stackrel{\text{iid}}{\sim} \mathcal{N}(\mathbf{x}_n^\top \mathbf{w}, \sigma^2), \end{aligned} \quad (2)$$

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

- (a) 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  $\lambda_0$  and  $\sigma^2$ .
- (b) The Laplace density can also be written as a *scale mixtures of Gaussians*,

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

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

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

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

- (c) What algorithm would you use to perform Bayesian inference to approximate the posterior distribution  $p(\mathbf{w}, \mathbf{v} \mid \mathbf{X}, \mathbf{y}; \lambda, \sigma^2)$ ? Sketch out the steps involved.

**Problem 6: Mixture Models**

Consider the following *zero-inflated Poisson regression* model where  $w, x_n \in \mathbb{R}_+$ ,  $y_n \in \mathbb{N}$ , and  $z_n \in \{0, 1\}$ ,

$$\begin{aligned}w &| \alpha, \beta \sim \text{Gamma}(\alpha, \beta) \\z_n &| \gamma \stackrel{\text{iid}}{\sim} \text{Bern}(\gamma) \\y_n &| x_n, z_n, w \stackrel{\text{iid}}{\sim} \text{Poisson}(wx_n z_n).\end{aligned}$$

- (a) Sketch the probability mass function of the marginal distribution  $p(y_n | x_n, w, \gamma)$  for  $\gamma \in \{0, 0.5, 1\}$ , assuming  $wx_n = 5$ . What is  $p(y_n = 0 | x_n, w, \gamma)$ ? (Note:  $0! = 1$  and  $0^0 = 1$ .)

- (b) Compute the conditional distribution  $p(z_n = 1 | y_n, x_n, w, \gamma)$ .

- (c) Compute the expected log probability,

$$\mathcal{L}(w) = \mathbb{E}_{p(z|y,x,w',\gamma)} \left[ \log p(\{y_n, x_n, z_n\}_{n=1}^N, w | \alpha, \beta, \gamma) \right],$$

where  $w'$  denotes a fixed weight. For notational simplicity, let  $q_n \triangleq p(z_n = 1 | y_n, x_n, w', \gamma)$  denote the solution to part (c), and drop terms in  $\mathcal{L}(w)$  that are constant with respect to  $w$ .

- (d) Assume  $\alpha > 1$ . Solve for  $w^* = \arg \max \mathcal{L}(w)$  using the fact that the mode of the  $\text{Gamma}(a, b)$  distribution is at  $(a - 1)/b$  when  $a > 1$ .

**Problem 7: Mixed Membership Models**

Latent Dirichlet allocation (LDA) corresponds to the following generative model,

$$\begin{aligned}
 \boldsymbol{\eta}_k &\sim \text{Dir}(\boldsymbol{\phi}) && \text{for } k = 1, \dots, K \\
 \boldsymbol{\pi}_n &\sim \text{Dir}(\boldsymbol{\alpha}) && \text{for } n = 1, \dots, N \\
 z_{n,\ell} &\sim \boldsymbol{\pi}_n && \text{for } n = 1, \dots, N; \ell = 1, \dots, L \\
 x_{n,\ell} &\sim \boldsymbol{\eta}_{z_{n,\ell}} && \text{for } n = 1, \dots, N; \ell = 1, \dots, L
 \end{aligned}$$

where  $\boldsymbol{\eta}_k \in \Delta_V$  are the *topics* (i.e. distributions over words) and  $\boldsymbol{\pi}_n \in \Delta_K$  are the *topic proportions* (i.e. distributions over topics).

However, this model fails to capture correlations in the topic proportions; for example, that a “finance” topic and a “government” topic may often co-occur in the same document. *Correlated topic models* address this limitation by replacing the Dirichlet prior on  $\boldsymbol{\pi}_n$  with a logistic normal prior,

$$\begin{aligned}
 \boldsymbol{\pi}_n &= \text{softmax}(\mathbf{u}_n) = \left[ \frac{e^{u_{n1}}}{1 + \sum_{k=1}^{K-1} e^{u_{nk}}}, \dots, \frac{e^{u_{n,K-1}}}{1 + \sum_{k=1}^{K-1} e^{u_{nk}}}, \frac{1}{1 + \sum_{k=1}^{K-1} e^{u_{nk}}} \right]^\top \\
 \mathbf{u}_n &\sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})
 \end{aligned}$$

where  $\mathbf{u}_n \in \mathbb{R}^{K-1}$ . The correlations in  $\mathbf{u}_n$  due to the multivariate normal prior induce correlations in  $\boldsymbol{\pi}_n$  as well.

- (a) Without doing any math, sketch the density of  $\boldsymbol{\pi}_n \in \Delta_3$  when  $\boldsymbol{\mu} = [0, 0]^\top$  and  $\boldsymbol{\Sigma} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ . Do the same for  $\boldsymbol{\mu} = [0, 0]^\top$  and  $\boldsymbol{\Sigma} = \begin{bmatrix} 1 & \frac{1}{2} \\ \frac{1}{2} & 1 \end{bmatrix}$ . Explain your reasoning.

- (b) Try to derive CAVI updates for this model. Where do you run into trouble and why?



**Problem 8: Variational autoencoders**

Consider the following *deep mixture model*,

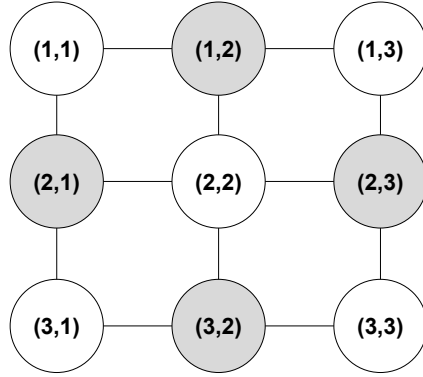
$$\begin{aligned}z_n &\sim \pi \\ \mathbf{x}_n &\sim \mathcal{N}(\boldsymbol{\mu}_{z_n}, \boldsymbol{\Sigma}_{z_n}) \\ \mathbf{y}_n &\sim \mathcal{N}(f(\mathbf{x}_n; \mathbf{w}), \sigma^2 \mathbf{I})\end{aligned}$$

where  $z_n \in \{1, \dots, K\}$  is a discrete latent variable,  $\mathbf{x}_n \in \mathbb{R}^M$  is a continuous latent variable,  $\mathbf{y}_n \in \mathbb{R}^D$  is an observed data point, and  $f : \mathbb{R}^M \mapsto \mathbb{R}^D$  is a neural network with weights  $\mathbf{w}$ . The generative model parameters are  $\boldsymbol{\theta} = (\pi, \{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}_{k=1}^K, \mathbf{w})$ .

- (a) Suppose you wanted to perform fixed form variational inference to approximate the posterior,  $p(z_n, \mathbf{x}_n \mid \mathbf{y}_n; \boldsymbol{\theta}) \approx q(z_n, \mathbf{x}_n; \boldsymbol{\phi})$ , with variational parameters  $\boldsymbol{\phi}$ . What challenges might you encounter when trying to maximize the local ELBO,  $\mathcal{L}_n(\boldsymbol{\theta}, \boldsymbol{\phi})$ , using stochastic gradient ascent and the reparameterization trick (i.e. the pathwise gradient estimator)?
- (b) Suggest an alternative to the reparameterization trick that could allow you to fit  $\boldsymbol{\theta}$  and  $\boldsymbol{\phi}$ . What challenges might this alternative present?
- (c) Rewrite the generative model by marginalizing over  $z_n$  to obtain a collapsed model  $p(\mathbf{x}_n, \mathbf{y}_n; \boldsymbol{\theta})$ , and assume a variational posterior  $q(\mathbf{x}_n; \boldsymbol{\phi})$ . Can you use the reparameterization trick now?

**Problem 9: State space models**

In class we studied state space models for sequential data, like hidden Markov models and linear dynamical systems. Here we will consider similar models for 2-dimensional data. Suppose we observe an image  $\mathbf{y} \in \mathbb{R}^{H \times W}$  which we believe to be a noisy version of an underlying binary image  $\mathbf{x} \in \{0, 1\}^{H \times W}$ . Given  $\mathbf{y}$ , we wish to recover the true image  $\mathbf{x}$  which it was derived from. We formulate this as a probabilistic inference problem. We will assume the image is square and start by constructing a graph which connects neighboring pixels. The graph for  $H = W = 3$  is shown below, with the node labels corresponding to the indices in the vectors  $\mathbf{y}$  and  $\mathbf{x}$ .



Our prior on  $\mathbf{x}$  will be given as an *Ising model*, which encodes our belief that nearby pixels are likely to be similar:

$$p(\mathbf{x}) = \frac{1}{Z(\theta)} \prod_{(ij,kl) \in \mathcal{E}} \psi_{\theta}(x_{ij}, x_{kl})$$

Here,  $\mathcal{E}$  is the edge set of the pixel graph,  $Z(\theta)$  is a normalizing constant, and  $\psi_{\theta} : \{0, 1\} \times \{0, 1\} \rightarrow \mathbb{R}_{++}$  is defined by:

$$\psi_{\theta}(x_{ij}, x_{kl}) = \begin{cases} e^{\theta} & , x_{ij} = x_{kl} \\ 1 & , x_{ij} \neq x_{kl} \end{cases}$$

where  $\theta > 0$  is a hyperparameter. We assume a Gaussian noise model, which gives us a likelihood over  $\mathbf{y}$  given  $\mathbf{x}$  as:

$$p(\mathbf{y} | \mathbf{x}) = \prod_{i=1}^H \prod_{j=1}^W \mathcal{N}(y_{ij} | x_{ij}, \sigma^2)$$

where  $\sigma^2$  is a hyperparameter. Given  $\mathbf{y}$ , we will obtain our denoised image by sampling from the posterior  $p(\mathbf{x} | \mathbf{y})$  using Gibbs sampling

- (a) Given a pixel  $(i, j)$ , let  $\mathcal{E}(i, j)$  denote its neighbors in the pixel graph. Similarly, given  $\mathbf{x} \in \{0, 1\}^{H \times W}$ , let  $N_1(i, j, \mathbf{x}_{-ij}) = \sum_{(k,l) \in \mathcal{E}(i,j)} x_{k,l}$  denote the number of neighbors of pixel  $(i, j)$  set to 1 and let  $N_0(i, j, \mathbf{x}_{-ij}) = \sum_{(k,l) \in \mathcal{E}(i,j)} 1 - x_{kl}$  denote the number of neighbors of pixel  $(i, j)$  set to 0.

Show that the complete conditional of  $x_{ij}$  is given by:

$$p(x_{ij} = 1 \mid \mathbf{x}_{-ij}, \mathbf{y}) = \frac{e^{\phi_1}}{e^{\phi_0} + e^{\phi_1}}$$

where

$$\begin{aligned}\phi_1 &= \theta N_1(i, j, \mathbf{x}_{-ij}) + \log \mathcal{N}(y_{ij} \mid x_{ij} = 1, \sigma^2) \\ \phi_0 &= \theta N_0(i, j, \mathbf{x}_{-ij}) + \log \mathcal{N}(y_{ij} \mid x_{ij} = 0, \sigma^2)\end{aligned}$$

- (b) Suppose we also incorporate a prior  $p(\theta)$  on  $\theta$ , e.g.  $p(\theta) = \text{Gamma}(\theta; \alpha, \beta)$ . It is not possible to derive a closed form for  $\theta$ 's complete conditional  $p(\theta \mid \mathbf{x}, \mathbf{y})$ . Explain what we may do instead to approximately sample from this conditional. Why might this be computationally challenging for large images (i.e. when  $D$  is large)?
- (c) [Bonus] Consider the pixel graph, and let  $\mathcal{S}$  be a maximal set of nodes such that  $\mathcal{E}(i, j) \cap \mathcal{S} = \emptyset$  for all  $(i, j) \in \mathcal{S}$ . For the example graph, we could use  $\mathcal{S}$  as the shaded set of nodes, so  $\mathcal{S} = \{(1, 2), (2, 1), (2, 3), (3, 2)\}$ . Explain why we have  $\mathbf{x}_{\mathcal{S}} \perp\!\!\!\perp \mathbf{x}_{-\mathcal{S}} \mid \mathbf{y}, \theta$  and how we can exploit this for an efficient parallel block Gibbs update.

**Problem 10: Bayesian nonparametrics**

In class we said that a Dirichlet random variable equal in distribution to a normalized vector of independent gamma random variables,

$$\begin{aligned}\gamma_k &\stackrel{\text{ind}}{\sim} \text{Ga}(\alpha_k, 1) \\ \boldsymbol{\pi} &= \left[ \frac{\gamma_1}{\sum_{k=1}^K \gamma_k}, \dots, \frac{\gamma_K}{\sum_{k=1}^K \gamma_k} \right]^\top \\ \Rightarrow \boldsymbol{\pi} &\sim \text{Dir}(\boldsymbol{\alpha}).\end{aligned}$$

It turns out there are many other useful properties of the gamma distribution, like

$$\gamma_k \stackrel{\text{ind}}{\sim} \text{Ga}(\alpha_k, 1) \Rightarrow \sum_{k=1}^K \gamma_k \sim \text{Ga}\left(\sum_{k=1}^K \alpha_k, 1\right).$$

Moreover, the normalized vector of gammas is independent of the sum,  $\boldsymbol{\pi} \perp \sum_{k=1}^K \gamma_k$ . Finally, the gamma is also related to the beta distribution,

$$\begin{aligned}\gamma_k &\sim \text{Ga}(\alpha_k, 1); \quad k \in \{0, 1\} \\ \beta &= \frac{\gamma_1}{\gamma_0 + \gamma_1} \\ \Rightarrow \beta &\sim \text{Beta}(\alpha_1, \alpha_0).\end{aligned}$$

Use these properties to derive a stick breaking procedure for sampling a (finite) Dirichlet distribution with concentration  $\boldsymbol{\alpha} \in \mathbb{R}_+^K$ .