STATS 305C: Practice Exam

Name:

Problem 1: *Gaussian models*

Consider the following model,

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

(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 $\boldsymbol{X} = \{\{x_{n,d}\} d = 1^D\}_{n=1}^N$. Derive a Gibbs sampler for the posterior distribution $p(\boldsymbol{\lambda} \mid \boldsymbol{X}; \alpha)$.

Solution:

(a) $\mathbb{E}[\lambda_d] = \alpha$ so when $\alpha > 1$, we the precisions are generally greater than one as well. Thus, the precision $(\sigma_d^2)^{-1} = \prod_{k=1}^d \lambda_k$ grows with d, and equivalently, the variance shrinks as d increases. Thus, we expect $x_{n,D}$ to be more concentrated around the prior mean of 0 than $x_{n,1}$.

(b) The joint probability is,

$$p(\boldsymbol{X}, \boldsymbol{\lambda}; \alpha) = \prod_{d=1}^{D} \text{Ga}(\lambda_d; \alpha, 1) \prod_{n=1}^{N} \prod_{d=1}^{D} \mathcal{N}\left(x_{n,d} \mid 0, \prod_{k=1}^{d} \lambda_k^{-1}\right)$$

Note that λ_k appears in the likelihood for all $x_{n,d}$ where $d \geq k$. Thus, the conditional distribution of λ_k is proportional to,

$$\begin{aligned} p(\lambda_k \mid -) &\propto \text{Ga}(\lambda_k; \alpha, 1) \prod_{n=1}^{N} \prod_{d=k}^{D} \mathcal{N}\left(x_{n,d} \mid 0, \prod_{j=1}^{d} \lambda_j^{-1}\right) \\ &\propto \lambda_k^{\alpha-1} e^{-\lambda_k} \prod_{n=1}^{N} \prod_{d=k}^{D} \lambda_k^{\frac{1}{2}} \exp\left\{-\frac{\lambda_k \left(\prod_{i=1}^{k-1} \lambda_i\right)\left(\prod_{j=k+1}^{d} \lambda_j\right) x_{n,d}^2}{2}\right\} \\ &\propto \text{Ga}(\lambda_k; \alpha', \beta') \end{aligned}$$

where

$$\begin{aligned} \alpha' &= \alpha + \frac{N(D-k+1)}{2} \\ \beta' &= 1 + \sum_{n=1}^{N} \sum_{d=k}^{D} \frac{\left(\prod_{i=1}^{k-1} \lambda_i\right)\left(\prod_{j=k+1}^{d} \lambda_j\right) x_{n,d}^2}{2} \end{aligned}$$

To implement a Gibbs sampler, iteratively sample each λ_k from its conditional, holding the others fixed.

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,

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

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

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

Solution: The first conditional is,

$$\begin{aligned} p(\lambda_g \mid \{x_{g,n}\}_{n=1}^N, \beta_g; \alpha) &\propto p(\lambda_g \mid \beta_g; \alpha) \prod_{n=1}^N p(x_{g,n} \mid \lambda_g) \\ &\propto \text{Ga}(\lambda_g; \alpha, \beta_g) \prod_{n=1}^N \text{Po}(x_{g,n} \mid \lambda_g) \\ &\propto \lambda_g^{\alpha-1} e^{-\beta_g \lambda_g} \prod_{n=1}^N \lambda_g^{x_{g,n}} e^{-\lambda_g} \\ &\propto \text{Ga}(\lambda_g; \alpha', \beta') \end{aligned}$$

where

$$\begin{aligned} \alpha' &= \alpha + \sum_{n=1}^N x_{g,n} \\ \beta' &= \beta_g + N \end{aligned}$$

The second is,

$$\begin{aligned} p(\beta \mid \{\lambda_g\}_{g=1}^G; \alpha_0, \beta_0, \alpha) &\propto \text{Ga}(\beta; \alpha_0, \beta_0) \prod_{g=1}^G \text{Ga}(\lambda_g; \alpha, \beta) \\ &\propto \beta^{\alpha_0 - 1} e^{-\beta_0 \beta} \prod_{g=1}^G \beta^\alpha e^{-\beta \lambda_g} \\ &\propto \text{Ga}(\beta; \alpha', \beta') \end{aligned}$$

where

$$\begin{aligned} \alpha' &= \alpha_0 + \alpha G \\ \beta' &= \beta_0 + \sum_{g=1}^G \lambda_g \end{aligned}$$

Problem 3: *Graphical models.*

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

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

Solution:

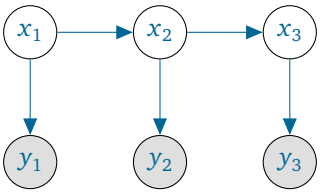


Figure 1: Note: This omits the hyperparameters, which could be drawn as black dots.

(b) Write the joint distribution corresponding to the graphical model in Figure 2.

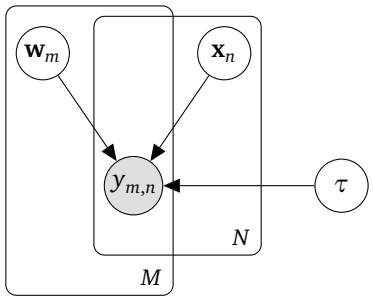


Figure 2

Solution:

$$p(\tau)\prod_{m=1}^{M}p(\boldsymbol{w}_m)\prod_{n=1}^{N}p(\boldsymbol{x}_n)\prod_{m=1}^{M}\prod_{n=1}^{N}p(y_{m,n}\mid \boldsymbol{w}_m,\boldsymbol{x}_n)$$

Problem 4: *Continuous latent variable models*

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

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

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

Solution: Note that

$$p(\boldsymbol{y}_n \mid \boldsymbol{x}_n; \boldsymbol{\theta}) = \int p(\boldsymbol{y}_n \mid \boldsymbol{z}_n; \boldsymbol{\theta})\, p(\boldsymbol{z}_n \mid \boldsymbol{x}_n; \boldsymbol{\theta})\, \mathrm{d}\boldsymbol{z}_n$$

since $\boldsymbol{y}_n \perp\!\!\!\perp \boldsymbol{x}_n \mid \boldsymbol{z}_n$. The conditional in the integrand is,

$$\begin{aligned}p(\boldsymbol{z}_n \mid \boldsymbol{x}_n; \boldsymbol{\theta}) &\propto p(\boldsymbol{z}_n)\, p(\boldsymbol{x}_n \mid \boldsymbol{z}_n; \boldsymbol{\theta}) \\ &= \mathcal{N}(\boldsymbol{z}_n; \boldsymbol{0}, \boldsymbol{I})\, \mathcal{N}(\boldsymbol{x}_n \mid \boldsymbol{W}_x \boldsymbol{z}_n + \boldsymbol{b}_x; \boldsymbol{\Sigma}_x) \\ &\propto \exp\left\{-\frac{1}{2}\boldsymbol{z}_n^\top \boldsymbol{J}_z \boldsymbol{z}_n + \boldsymbol{h}_z^\top \boldsymbol{z}_n\right\}\end{aligned}$$

where

$$\begin{aligned}\boldsymbol{J}_z &= \boldsymbol{I} + \boldsymbol{W}_x^\top \boldsymbol{\Sigma}_x^{-1} \boldsymbol{W}_x \\ \boldsymbol{h}_z &= \boldsymbol{W}_x^\top \boldsymbol{\Sigma}_x^{-1} (\boldsymbol{x}_n - \boldsymbol{b}_x).\end{aligned}$$

Completing the square, $p(\boldsymbol{z}_n \mid \boldsymbol{x}_n; \boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\mu}_z, \boldsymbol{\Sigma}_z)$ where

$$\boldsymbol{\mu}_z = \boldsymbol{J}_z^{-1} \boldsymbol{h}_z = (\boldsymbol{I} + \boldsymbol{W}_x^\top \boldsymbol{\Sigma}_x^{-1} \boldsymbol{W}_x)^{-1} \boldsymbol{W}_x^\top \boldsymbol{\Sigma}_x^{-1} (\boldsymbol{x}_n - \boldsymbol{b}_x) \tag{1}$$

$$\boldsymbol{\Sigma}_z = \boldsymbol{J}_z^{-1} = (\boldsymbol{I} + \boldsymbol{W}_x^\top \boldsymbol{\Sigma}_x^{-1} \boldsymbol{W}_x)^{-1} \tag{2}$$

Finally, we obtain the predictive distribution by marginalizing over $\boldsymbol{z}_n$ under this Gaussian conditional distribution,

$$p(\boldsymbol{y}_n \mid \boldsymbol{x}_n; \boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{y}_n \mid \boldsymbol{W}_y \boldsymbol{\mu}_z + \boldsymbol{b}_y, \boldsymbol{\Sigma}_y + \boldsymbol{W}_y \boldsymbol{\Sigma}_z \boldsymbol{W}_y^\top).$$

An alternative way to obtain the same answer is to construct the joint multivariate normal distribution over $(\boldsymbol{z}_n, \boldsymbol{x}_n, \boldsymbol{y}_n)$, then use the rules of Gaussian marginalization and conditioning.

Problem 5: *The Bayesian Lasso*

The Lasso problem is an L_1 penalized least squares problem,

$$\mathscr{L}(\boldsymbol{w}) = \sum_{n=1}^{N} \|y_n - \boldsymbol{x}_n^\top \boldsymbol{w}\|_2^2 + \lambda_0 \sum_{d=1}^{D} |w_d|. \tag{3}$$

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

$$\begin{aligned} w_d &\overset{\text{iid}}{\sim} \text{Lap}(\lambda) \\ y_n &\overset{\text{iid}}{\sim} \mathcal{N}(\boldsymbol{x}_n^\top \boldsymbol{w}, \sigma^2), \end{aligned} \tag{4}$$

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 λ such that the MAP estimate of model (4) is the same as the minimizer of eq. 3. Your solution should be in terms of λ_0 and σ^2.

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

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

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

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

on an extended space that includes the *augmentation variables* $\boldsymbol{v} = (v_1, \ldots, v_D)$, such that the marginal distribution $p(\boldsymbol{w}, \boldsymbol{y} \mid \boldsymbol{X}; \lambda, \sigma^2)$ matches that of the generative model described in eq. (4).

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

Solution:

(a) The log joint probability of the Bayesian model as a function of $\boldsymbol{w}$ is,

$$\begin{aligned} \mathscr{J}(\boldsymbol{w}) &= \log p(\boldsymbol{w}, \boldsymbol{y} \mid \boldsymbol{X}) \\ &= \sum_{d=1}^{D} \log \text{Lap}(w_d; \lambda) + \sum_{n=1}^{N} \log \mathcal{N}(y_n; \boldsymbol{w}^\top \boldsymbol{x}_n, \sigma^2) + c \\ &= -\lambda \sum_{d=1}^{D} |w_d| - \sum_{n=1}^{N} \frac{1}{2\sigma^2} \|y_n - \boldsymbol{w}^\top \boldsymbol{x}_n\|_2^2 + c. \end{aligned}$$

Multiplying by $2\sigma^2$ does not change the arg max of this objective (i.e. MAP estimate), so define,

$$\mathscr{J}'(\boldsymbol{w}) = -2\lambda\sigma^2 \sum_{d=1}^{D} |w_d| - \sum_{n=1}^{N} \|y_n - \boldsymbol{w}^\top \boldsymbol{x}_n\|_2^2 + c.$$

Note, if $\lambda_0 = 2\lambda\sigma^2$ then $\mathscr{J}'(\boldsymbol{w}) = -\mathscr{L}(\boldsymbol{w})$. With this setting of λ_0, the Lasso estimate that minimizes $\mathscr{L}(\boldsymbol{w})$ coincides the MAP estimate that maximizes $\mathscr{J}'(\boldsymbol{w})$.

(b) With this integral representation of the Laplace distribution,

$$\begin{aligned} p(\boldsymbol{w}, \boldsymbol{y} \mid \boldsymbol{X}) &= \prod_{d=1}^{D} \left[\int_0^\infty \mathcal{N}(w_d; 0, \nu_d) \operatorname{Exp}(\nu_d; \tfrac{\lambda^2}{2}) \, \mathrm{d}\nu_d \right] \prod_{n=1}^{N} \mathcal{N}(y_n; \boldsymbol{w}^\top \boldsymbol{x}_n, \sigma^2) \\ &= \int_0^\infty \cdots \int_0^\infty \prod_{d=1}^{D} \mathcal{N}(w_d; 0, \nu_d) \operatorname{Exp}(\nu_d; \tfrac{\lambda^2}{2}) \prod_{n=1}^{N} \mathcal{N}(y_n; \boldsymbol{w}^\top \boldsymbol{x}_n, \sigma^2) \, \mathrm{d}\nu_1 \ldots \mathrm{d}\nu_D \\ &= \int p(\boldsymbol{w}, \boldsymbol{\nu}, \boldsymbol{y} \mid \boldsymbol{X}) \, \mathrm{d}\boldsymbol{\nu} \end{aligned}$$

where

$$p(\boldsymbol{w}, \boldsymbol{\nu}, \boldsymbol{y} \mid \boldsymbol{X}) = \prod_{d=1}^{D} \mathcal{N}(w_d; 0, \nu_d) \operatorname{Exp}(\nu_d; \tfrac{\lambda^2}{2}) \prod_{n=1}^{N} \mathcal{N}(y_n; \boldsymbol{w}^\top \boldsymbol{x}_n, \sigma^2)$$

(c) The simple answer is you could do HMC, but you could have done that on the original model anyway. (There's a slight complication with the prior being non-differentiable at $w_d = 0$, but you could throw HMC at it anyway.) However, this augmentation scheme presents an opportunity to introduce a pretty cool trick. It turns out that in the augmented model, the conditional distributions of $\boldsymbol{w}$ and $\boldsymbol{\nu}$ are tractable, and we can use them to perform Gibbs sampling. We have,

$$p(\boldsymbol{w} \mid \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{\nu}) = \mathcal{N}(\boldsymbol{w}; \boldsymbol{J}^{-1}\boldsymbol{h}, \boldsymbol{J}^{-1})$$

where

$$\begin{aligned} \boldsymbol{J} &= \operatorname{diag}(\boldsymbol{\nu})^{-1} + \frac{1}{\sigma^2} \sum_{n=1}^{N} \boldsymbol{x}_n \boldsymbol{x}_n^\top \\ \boldsymbol{h} &= \frac{1}{\sigma^2} \sum_{n=1}^{N} \boldsymbol{x}_n y_n. \end{aligned}$$

The conditional distribution of ν_d is,

$$\begin{aligned} p(\nu_d \mid w_d) &\propto \mathcal{N}(w_d; 0, \nu_d) \operatorname{Exp}(\nu_d; \tfrac{\lambda^2}{2}) \\ &\propto \frac{1}{\sqrt{\nu_d}} \exp\left\{ -\frac{w_d^2}{2\nu_d} - \frac{\lambda^2 \nu_d}{2} \right\} \end{aligned}$$

This is a strange looking conditional... in exponential family form, its sufficient statistics are $\log \nu_d$, ν_d^{-1}, and ν_d.

You wouldn't be expected to know this for the actual final, but it turns out that this conditional distribution can be rewritten as a (somewhat) standard density. Instead of the variance ν_d, let's do

a change of variables to consider the conditional distribution of the precision $\tau_d = \nu_d^{-1}$,

$$\begin{aligned} p(\tau_d \mid w_d) &\propto \left| \frac{\mathrm{d}\tau_d^{-1}}{\mathrm{d}\tau_d} \right| \mathcal{N}(w_d; 0, \tau_d^{-1}) \operatorname{Exp}(\tau_d^{-1}; \tfrac{\lambda^2}{2}) \\ &\propto \frac{1}{\tau_d^2} \sqrt{\tau_d} \exp\left\{ -\frac{1}{2} \left(\tau_d w_d^2 + \frac{\lambda^2}{\tau_d} \right) \right\} \\ &\propto \tau_d^{-\frac{3}{2}} \exp\left\{ -\frac{w_d^2}{2\tau_d} \left(\tau_d^2 + \frac{\lambda^2}{w_d^2} \right) \right\} \\ &\propto \operatorname{IG}\left(\tau_d; \frac{\lambda}{w_d}, \lambda^2 \right) \end{aligned}$$

where $\operatorname{IG}(x; \mu, \kappa)$ denotes the *inverse Gaussian* distribution with mean μ and shape κ. This isn't a distribution that came up in class, but it appears in the context of stochastic processes. For example, the time a Brownian motion with positive drift takes to reach a fixed positive level is inverse-Gaussian distributed.

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 $z_n \in \{0, 1\}$,

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

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

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

(c) Compute the expected log probability,

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

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

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

Solution:

(a) The marignal distribution is,

$$\begin{aligned} p(y_n \mid wx_n = 5; \gamma) &= \sum_{z_n} \text{Po}(y_n; w x_n z_n)\, p(z_n; \gamma) \\ &= \gamma \text{Po}(y_n; 5) + (1-\gamma)\text{Po}(y_n; 0) \\ &= \gamma \text{Po}(y_n; 5) + (1-\gamma)\mathbb{I}[y_n = 0] \end{aligned}$$

It looks like a re-scaled Poisson pmf with an extra mass of $(1-\gamma)$ added to $p(y_n = 0 \mid -)$.

(b) The conditional is,

$$\begin{aligned} p(z_n = 1 \mid y_n, x_n, w; \gamma) &\propto p(z_n = 1; \gamma)\, p(y_n \mid z_n = 1, x_n, w) \\ &\propto \gamma \text{Po}(y_n; w x_n) \end{aligned}$$

Likewise,

$$\begin{aligned} p(z_n = 0 \mid y_n, x_n, w; \gamma) &\propto p(z_n = 0; \gamma)\, p(y_n \mid z_n = 0, x_n, w) \\ &\propto (1-\gamma)\text{Po}(y_n; 0) \\ &\propto (1-\gamma)\mathbb{I}[y_n = 0]. \end{aligned}$$

Normalizing yields,

$$p(z_n = 1 \mid y_n, x_n, w; \gamma) = \frac{\gamma \text{Po}(y_n; w x_n)}{\gamma \text{Po}(y_n; w x_n) + (1-\gamma)\mathbb{I}[y_n = 0]}$$

(c) The expected log probability is,

$$\begin{aligned}
\mathscr{L}(w) &= \mathbb{E}_{p(z|y,x,w',\gamma)}\left[\log p(\{y_n, x_n, z_n\}_{n=1}^N, w \mid \alpha, \beta, \gamma)\right] \\
&= \mathbb{E}_{p(z|y,x,w',\gamma)}\left[\log p(w;\alpha,\beta) + \sum_{n=1}^N \log p(y_n \mid w, x_n, z_n)\right] + c \\
&= \log \mathrm{Ga}(w;\alpha,\beta) + \sum_{n=1}^N \mathbb{E}_{p(z_n|-)}\left[\log \mathrm{Po}(y_n \mid w x_n z_n)\right] + c \\
&= (\alpha-1)\log w - \beta w + \sum_{n=1}^N \mathbb{E}_{p(z_n|-)}\left[y_n \log(w x_n z_n) - w x_n z_n\right] + c \\
&= (\alpha-1)\log w - \beta w + \sum_{n=1}^N y_n \log w - w x_n q_n + c \\
&= \log \mathrm{Ga}(w;\alpha',\beta')
\end{aligned}$$

where

$$\begin{aligned}
\alpha' &= \alpha + \sum_{n=1}^N y_n \\
\beta' &= \beta + \sum_{n=1}^N x_n q_n.
\end{aligned}$$

(d) The mode of the gamma is at

$$w^\star = \frac{\alpha'-1}{\beta'} = \frac{\alpha + \sum_{n=1}^N y_n - 1}{\beta + \sum_{n=1}^N x_n q_n}$$

As a sanity check, consider the case where $x_n = 1$ for all n. Then the maximum likelihood estimate of w under a standard Poisson model (equivalently, with $z_n = 1$ for all n), would be $w_{\mathsf{MLE}} = \frac{1}{N}\sum_n y_n$. With the zero-inflated model, some of the observations where $y_n = 0$ are attributed to z_n being zero. Thus, we should expect our estimate of w to be a bit higher. Indeed, under an uninformative prior ($\alpha = 1, \beta = 0$), we would have $w^* = (\sum_n y_n)/(\sum_n q_n) \geq w_{\mathsf{MLE}}$, since the denominator is at most N.

Problem 7: *Mixed Membership Models*

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

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

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,

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

where $\boldsymbol{u}_n \in \mathbb{R}^{K-1}$. The correlations in $\boldsymbol{u}_n$ 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?

Solution:

(a) See the following figure from from Blei, David and John Lafferty. "Correlated topic models." *Advances in Neural Information Processing Systems* 18 (2005). When the first two coordinates are

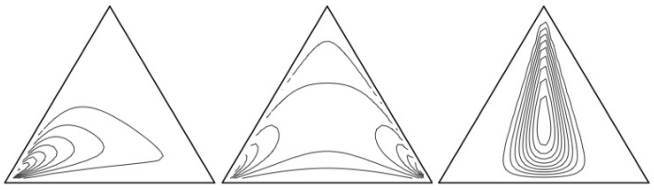


Figure 1: Top: Graphical model representation of the correlated topic model. The logistic normal distribution, used to model the latent topic proportions of a document, can represent correlations between topics that are impossible to capture using a single Dirichlet. Bottom: Example densities of the logistic normal on the 2-simplex. From left: diagonal covariance and nonzero-mean, negative correlation between components 1 and 2, positive correlation between components 1 and 2.

uncorrelated, the distribution is essentially symmetric. (The figure here shows what happens when the mean is nonzero — it's pulled toward one of the simplex vertices.) When the two coordinates of the Gaussian are anti-correlated, you end up with a multi-modal distribution on the simplex with modes toward one vertex or another. When the two coordinates are positively correlated, the mass is pulled toward the midline. That's because both Gaussian coordinates are large and positive, it maps to $\boldsymbol{\pi}_n = [0.5, 0.5, 0]$; when both coordinates are negative, it maps to $\boldsymbol{\pi}_n = [0, 0, 1]$.

(b) We can work with either $\boldsymbol{u}_n$ or $\boldsymbol{\pi}_n$. Since the prior is specified on $\boldsymbol{u}_n$, let's go with that. The problem is that we have a Gaussian prior and a categorical likelihood. The optimal CAVI updates are still,

$$\begin{aligned} q(\boldsymbol{u}_n) &\propto \exp\left\{\mathbb{E}_{q(\boldsymbol{z}_n)}[\log p(\boldsymbol{u}_n) + \log p(\boldsymbol{z}_n \mid \boldsymbol{u}_n)]\right\} \\ &\propto \mathcal{N}(\boldsymbol{u}_n; \boldsymbol{\mu}, \boldsymbol{\Sigma}) \exp\left\{\sum_{l=1}^{L} \mathbb{E}_{q(z_{n,l})}[\log \mathrm{Cat}(z_{n,l}; \mathrm{softmax}(\boldsymbol{u}_n))]\right\} \\ &\propto \mathcal{N}(\boldsymbol{u}_n; \boldsymbol{\mu}, \boldsymbol{\Sigma}) \exp\left\{\sum_{l=1}^{L}\sum_{v=1}^{V} q_{n,l,v}\left[u_{n,v} - \log(1 + \sum_{v'=1}^{V} e^{u_{n,v'}})\right]\right\} \end{aligned}$$

where $q_{n,l,v} = q(z_{n,l} = v)$. The problem is that the log-sum-exp of $\boldsymbol{u}_n$ in the exponent is not conjugate with the Gaussian prior.

Problem 8: *Variational autoencoders*

Consider the following *deep mixture model*,

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

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

(a) Suppose you wanted to perform fixed form variational inference to approximate the posterior, $p(z_n, \boldsymbol{x}_n \mid \boldsymbol{y}_n; \boldsymbol{\theta}) \approx q(z_n, \boldsymbol{x}_n; \boldsymbol{\phi})$, 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 $p(\boldsymbol{x}_n, \boldsymbol{y}_n; \boldsymbol{\theta})$, and assume a variational posterior $q(\boldsymbol{x}_n; \boldsymbol{\phi})$. Can you use the reparameterization trick now?

Solution:

(a) Since z_n is a discrete variable, we cannot reparameterize it as a differentiable transformation of "noise." That prevents us from using the standard pathwise gradient estimator/reparameterization tricks.

(b) We could use the score-function gradient estimator, since that works for discrete latent variables. However, without good control variates it can lead to higher variance estimates. Alternatively, we could use a continuous relaxation of the discrete variables, as in the Concrete or Gumbel-softmax approximations for discrete VAEs.

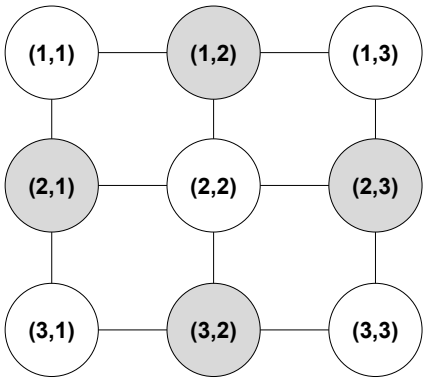
(c) Marginalizing out z_n yields,

$$p(\boldsymbol{x}_n, \boldsymbol{y}_n; \boldsymbol{\theta}) = \left(\sum_{k=1}^K \pi_k \mathcal{N}(\boldsymbol{x}_n; \boldsymbol{u}_k, \boldsymbol{\Sigma}_k) \right) \mathcal{N}(\boldsymbol{y}_n \mid f(\boldsymbol{x}_n; \boldsymbol{w}), \sigma^2 \boldsymbol{I}).$$

If we then choose a Gaussian variational posterior for $q(\boldsymbol{x}_n; \boldsymbol{\phi})$, then we **can** use the reparameterization trick since there are no longer any discrete variables to worry about.

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 a image $\boldsymbol{y} \in \mathbb{R}^{H \times W}$ which we believe to be a noisy version of an underlying binary image $\boldsymbol{x} \in \{0,1\}^{H \times W}$. Given $\boldsymbol{y}$, we wish to recover the true image $\boldsymbol{x}$ 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 $\boldsymbol{y}$ and $\boldsymbol{x}$.



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

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

Here, $\mathscr{E}$ is the edge set of the pixel graph, $Z(\theta)$ is a normalizing constant, and $\psi_\theta : \{0,1\} \times \{0,1\} \to \mathbb{R}_{++}$ 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 $\boldsymbol{y}$ given $\boldsymbol{x}$ as:

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

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

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

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

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

where

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

(b) Suppose we also incorporate a prior $p(\theta)$ on θ, e.g. $p(\theta) = \text{Gamma}(\theta; \alpha, \beta)$. It is not possible to derive a closed form for θ's complete conditional $p(\theta \mid \boldsymbol{x}, \boldsymbol{y})$. 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 the random variables $\{x_{ij} : (i, j) \in S\}$ are independent given $\boldsymbol{x}_{-\mathcal{S}}, \boldsymbol{y}, \theta$ and how we can exploit this for an efficient parallel block Gibbs update.

Solution:

(a) We have,

$$\begin{aligned} p(x_{ij} \mid \boldsymbol{x}_{\neg ij}, \boldsymbol{y}) &\propto \left(\prod_{k,l \in \mathcal{E}(i,j)} \psi_\theta(x_{ij}, x_{kl}) \right) \mathcal{N}(y_{ij} \mid x_{ij}, \sigma^2) \\ &\propto \left(\prod_{k,l \in \mathcal{E}(i,j)} e^{\theta \mathbb{I}[x_{ij} = x_{kl}]} \right) \mathcal{N}(y_{ij} \mid x_{ij}, \sigma^2) \\ &\propto e^{\theta \sum_{k,l \in \mathcal{E}(i,j)} \mathbb{I}[x_{ij} = x_{kl}]} \mathcal{N}(y_{ij} \mid x_{ij}, \sigma^2) \end{aligned}$$

For the two cases we have,

$$p(x_{ij} = 1 \mid \boldsymbol{x}_{\neg ij}, \boldsymbol{y}) \propto e^{\theta N_1(i,j,\boldsymbol{x}_{\neg i,j})} \mathcal{N}(y_{ij} \mid 1, \sigma^2)$$
$$p(x_{ij} = 0 \mid \boldsymbol{x}_{\neg ij}, \boldsymbol{y}) \propto e^{\theta N_0(i,j,\boldsymbol{x}_{\neg i,j})} \mathcal{N}(y_{ij} \mid 0, \sigma^2)$$

Normalizing and writing in terms of the logs yields the desired answer.

(b) The hard part about inferring θ is that it appears in the log normalizer,

$$Z(\theta) = \sum_{\boldsymbol{x} \in \{0,1\}^{H \times W}} \prod_{(ij,kl) \in \mathcal{E}} \psi_\theta(x_{ij}, x_{kl}),$$

and the log normalizer consists of a sum with 2^{HW} terms. For small lattices we can enumerate all these terms, evaluate $Z(\theta)$, and do inference with algorithms like HMC. However, this won't scale to even moderately sized grids.