CPSC 486/586: Probabilistic Machine Learning	*Out: February 15, 2023*

Problem Set 3

Instructor: Andre Wibisono *Due: March 1, 2023*

(P1) Consider a Gaussian graphical model on an undirected graph $G = (V, E)$ on vertices $V = \{1, 2, \ldots, n\}$. This means $X = (X_1, X_2, \ldots, X_n) \in \mathbb{R}^n$ has joint probability distribution $\rho \colon \mathbb{R}^n \to \mathbb{R}$ with density:

$$\rho(x) = \frac{1}{Z} \exp\left(-\sum_{i \in V} \alpha_i x_i^2 + \sum_{(i,j) \in E} \beta_{ij} x_i x_j \right) \qquad \forall\, x = (x_1, \ldots, x_n) \in \mathbb{R}^n$$

for some $\alpha_i > 0$, $\beta_{ij} \in \mathbb{R}$, where Z is the normalizing constant. Assume $Z < \infty$ and $\beta_{ij} \neq 0$ for all $(i, j) \in E$.

(a) Show that for all $i, j \in V$ with $(i, j) \notin E$ and $i \neq j$, we have:

$$X_i \perp X_j \mid X_{\setminus\{i,j\}}.$$

That is, show that the density of (X_i, X_j) given $X_{\{\setminus i,j\}} = (X_k \colon k \neq i, j)$ factorizes:

$$\rho(x_i, x_j \mid x_{\setminus\{i,j\}}) = \rho(x_i \mid x_{\setminus\{i,j\}}) \cdot \rho(x_j \mid x_{\setminus\{i,j\}})$$

for all $x_i, x_j \in \mathbb{R}$, and $x_{\setminus\{i,j\}} \in \mathbb{R}^{n-2}$.
(This means the *absence* of edges in the graph encodes the conditional independence between the random variables.)

(b) Let $C = \mathsf{Cov}_\nu(X) \in \mathbb{R}^{n \times n}$ be the covariance matrix of $X = (X_1, \ldots, X_n) \in \mathbb{R}^n$. Show that the nonzero pattern of C^{-1} matches the edge pattern of G, i.e., for all $i, j \in V$, $i \neq j$:

$$(C^{-1})_{ij} = 0 \qquad \Leftrightarrow \qquad (i, j) \notin E.$$

(*Hint:* Note that ρ is a Gaussian distribution.)

(P2) (Bayesian logistic regression) Suppose we have a hidden parameter $X \in \mathbb{R}$ with a Gaussian prior: $X \sim \rho_0 = \mathcal{N}(0, 1)$. For $i = 1, \ldots, n$, suppose we are given the covariates $W_1, \ldots, W_n \in \mathbb{R}$.[1] We observe the labels $Y_i, \ldots, Y_n \in \{0, 1\}$ following the Bernoulli distribution:

$$Y_i \mid \{X = x, W_i = w_i\} \sim \mathrm{Ber}(\sigma(x w_i)) \qquad \text{for } i = 1, \ldots, n \text{ iid.}$$

[1] Note that usually the notation is x_i for the covariates and w for the hidden parameter, but the notation is changed here. This is to be consistent with the other problems, which describe the distribution of interest in x variables.

This means $\Pr(Y_i = 1 \mid X = x, W_i = w_i) = \sigma(xw_i) = \frac{1}{1+e^{-xw_i}} = \frac{e^{xw_i}}{e^{xw_i}+1}$ where $\sigma(z) = \frac{1}{1+e^{-z}}$ is the sigmoid function.

Let $\rho_n(x) = \rho_n(x \mid y_1, \dots, y_n)$ be the posterior distribution of X after seeing n observations $Y = (y_1, \dots, y_n) \in \{0,1\}^n$. Recall (or check) that we can write $\rho_n(x) \propto \exp(-f_n(x))$ where

$$f_n(x) = \frac{1}{2}x^2 - \sum_{i=1}^{n} y_i w_i x + \sum_{i=1}^{n} \log(1 + \exp(w_i x)).$$

In this problem, suppose concretely we observe the following $n = 10$ observations:

$$\begin{aligned}
(w_1, y_1) &= (1, 1) \\
(w_2, y_2) &= (-2, 0) \\
(w_3, y_3) &= (3, 1) \\
(w_4, y_4) &= (5, 1) \\
(w_5, y_5) &= (-5, 0) \\
(w_6, y_6) &= (7, 1) \\
(w_7, y_7) &= (-1, 1) \\
(w_8, y_8) &= (-3, 0) \\
(w_9, y_9) &= (4, 1) \\
(w_{10}, y_{10}) &= (-10, 0)
\end{aligned}$$

We want to approximate ρ_n by a Gaussian distribution $\rho^* = \mathcal{N}(m^*, C^*)$ for some $m^* \in \mathbb{R}$ and $C^* \geq 0$. For each method below, compute the approximation explicitly.

(You can use any numerical method to solve the resulting (1-dimensional) computational problem, e.g. implementing an optimization algorithm, or integrating via numerical method.)

(a) Compute the Laplace approximation:

$$\rho^*_{\text{Lap}} = \mathcal{N}(m_{\text{Lap}}, C_{\text{Lap}})$$

Include a snippet of your code or calculations.

(b) Compute the EP (expectation propagation) approximation:

$$\rho^*_{\text{EP}} = \mathcal{N}(m_{\text{EP}}, C_{\text{EP}}) = \arg \min_{\rho = \mathcal{N}(m,c)} \mathsf{KL}(\rho_n \parallel \rho)$$

Include a snippet of your code or calculations.

(c) Compute the VB (variational Bayes) approximation:

$$\rho^*_{\mathrm{VB}} = \mathcal{N}(m_{\mathrm{VB}}, C_{\mathrm{VB}}) = \arg\min_{\rho = \mathcal{N}(m,c)} \mathsf{KL}(\rho \parallel \rho_n)$$

Include a snippet of your code.

(d) Provide a table to summarize the different values of m and C above. Plot the density of the posterior ρ_n and the three Gaussian approximations above, and also plot the log-density.

(P3) Consider a Bayesian model where $X \in \mathbb{R}^d$ has a prior probability distribution ρ_0, and we observe $Y = X + Z$ where $Z \sim \mathcal{N}(0, I)$ is an independent Gaussian random variable in $\mathbb{R}^d$.

For each $y \in \mathbb{R}^d$, let $\rho_{0|1}(x \mid y)$ denote the posterior distribution of X given $Y = y$, which is:

$$\rho_{0|1}(x \mid y) = \frac{\rho_0(x) \cdot (2\pi)^{-\frac{d}{2}} \exp(-\frac{1}{2}\|y - x\|^2)}{\rho_1(y)}.$$

We also write $\rho_{0|1=y} \equiv \rho_{0|1}(\cdot \mid y)$ for the posterior distribution of X given $Y = y$.

Let $\rho_1(y)$ denote the marginal distribution of Y at y according to the process above.

(a) Write down what is ρ_1 in terms of ρ_0 and $\gamma = \mathcal{N}(0, I)$.

(b) Recall the *score function* of ρ_1 is the gradient of log-density $\nabla \log \rho_1(y)$. Show that the score function of ρ_1 can be written in terms of the expectation under the posterior distribution:

$$\nabla \log \rho_1(y) = \mathbb{E}_{\rho_{0|1=y}}[X] - y.$$

(This is also known as *Tweedie's formula.*)

(c) Show that the Jacobian (derivative) of the score function, which is the second derivative of $\log \rho_1$, can be written in terms of the covariance of the posterior:

$$\nabla^2 \log \rho_1(y) = \mathsf{Cov}_{\rho_{0|1=y}}[X] - I.$$

Above, $\mathsf{Cov}_{\rho_{0|1=y}}[X] = \mathbb{E}_{\rho_{0|1=y}}[(X - \mu)(X - \mu)^\top]$ is the covariance where $\mu = \mathbb{E}_{\rho_{0|1=y}}[X]$.

(P4) Start thinking about how to relate the problem or topic that you proposed in PS2, to techniques and ideas we've learned in the class. Concretely,

(a) State a question that you are interested in answering (or a new question, if you found that it was answered in PS2).

(b) Find relevant papers in topics related to the course (e.g., references given from the class or a paper from `https://scorebasedgenerativemodeling.github.io` that is relevant to your research or is the most interesting to you) that may help you with your problem.

(c) Pick one and describe the result, as well as how it relates to your chosen topic/problem. Does it answer your question?

Additional questions for 586

(Q1) Let $\nu \propto e^{-f}$ be a probability distribution on $\mathbb{R}^d$ where $f\colon \mathbb{R}^d \to \mathbb{R}$ is differentiable and α-strongly convex, which means ν is α-strongly log-concave (SLC). Let $x^* = \arg\min_{x \in \mathbb{R}^d} f(x)$. Show that $X \sim \nu$ is not too far from x^* on average:

$$\mathbb{E}_\nu[\|X - x^*\|^2] \le \frac{d}{\alpha}.$$

(Q2) Recall if ν is α-SLC, then it satisfies *α-Poincaré inequality*, which means for any $\phi\colon \mathbb{R}^d \to \mathbb{R}$:

$$\mathrm{Var}_\nu(\phi(X)) \le \frac{1}{\alpha}\mathbb{E}_\nu[\|\nabla\phi(X)\|^2].$$

Use this fact to show that if $Z \sim \mathcal{N}(0, I)$ is a standard Gaussian random variable in $\mathbb{R}^d$, then

$$\sqrt{d-1} \le \mathbb{E}[\|Z\|] \le \sqrt{d}.$$

(This means on average Z lies on a thin shell of radius $O(\sqrt{d})$ with shell width $O(1)$.)

(Q3) Let $X \in \mathbb{R}^d$ with a prior distribution p_0, and observation $Y \mid \{X = x\} \sim p(Y \mid x)$. Assume $p_0 \propto e^{-f_0}$ is a log-concave distribution, i.e. $f_0\colon \mathbb{R}^d \to \mathbb{R}$ is a convex function. Assume $p(y \mid x) \propto e^{-\ell(x,y)}$ satisfies the following property with some $\alpha > 0$: for all $y \in \mathbb{R}^d$, the negative log-likelihood $x \mapsto \ell(x, y) = -\log p(y \mid x)$ is an α-strongly convex function.

Let $p_n(x) = p(x \mid y_1, \dots, y_n)$ be the posterior distribution of X after seeing observations $y_1, \dots, y_n \in \mathbb{R}^d$. Show that the posterior variance decreases with the number of observations:

$$\mathrm{Var}_{p_n}(\theta) \le \frac{d}{n\alpha}.$$