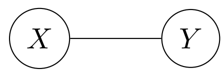
CPSC 486/586: Probabilistic Machine Learning *Out: February 1, 2023*

Problem Set 2

Instructor: Andre Wibisono *Due: February 15, 2023*

(P1) Consider a Gaussian graphical model on a two-node graph.



This means we have a joint distribution on $(x, y) \in \mathbb{R} \times \mathbb{R}$:

$$\nu(x,y) = \frac{1}{Z} \exp\left(-\frac{\alpha}{2}x^2 - \frac{\alpha}{2}y^2 + \beta xy\right)$$

for some parameters $\alpha > 0$ and $\beta \in \mathbb{R}$. Assume $|\beta| < \alpha$. Here

$$Z = \int_{\mathbb{R}\times\mathbb{R}} \exp\left(-\frac{\alpha}{2}\|x\|^2 - \frac{\alpha}{2}\|y\|^2 + \beta x^\top y\right) dx\, dy$$

is the normalizing constant.

(a) Note that $\nu = \mathcal{N}(\mu, \Sigma)$ is a joint Gaussian distribution on $\mathbb{R}^2$. Compute $\mu \in \mathbb{R}^2$ and $\Sigma \in \mathbb{R}^{2\times 2}$ in terms of α, β. Explain why we need the assumption $|\beta| < \alpha$.

(b) Note that the marginal distributions of X, Y are Gaussian:

$$\nu_X = \mathcal{N}(\mu_X, \Sigma_X)$$
$$\nu_Y = \mathcal{N}(\mu_Y, \Sigma_Y).$$

Compute $\mu_X, \mu_Y \in \mathbb{R}$ and $\Sigma_X, \Sigma_Y > 0$ in terms of α, β.

(c) We want to approximate ν with an independent Gaussian distribution $\rho = \rho_X \otimes \rho_Y$ (this means $\rho(x,y) = \rho_X(x)\rho_Y(y)$ where $\rho_X = \mathcal{N}(\mu_X, \Sigma_X)$ and $\rho_Y = \mathcal{N}(\mu_Y, \Sigma_Y)$ for some $\mu_X, \mu_Y \in \mathbb{R}$ and $\Sigma_X, \Sigma_Y > 0$; equivalently, $\rho = \rho_X \otimes \rho_Y = \mathcal{N}\left(\begin{pmatrix}\mu_X \\ \mu_Y\end{pmatrix}, \begin{pmatrix}\Sigma_X & 0 \\ 0 & \Sigma_Y\end{pmatrix}\right)$).
We choose the best approximation by minimizing the KL divergence:

$$\rho^* = \arg\min_{\rho = \rho_X \otimes \rho_Y} \mathsf{KL}(\rho \| \nu)$$

where the minimization is over Gaussian distributions ρ_X, ρ_Y on $\mathbb{R}$. Show that the minimizer $\rho^* = \rho_X^* \otimes \rho_Y^*$ is given by

$$\rho_X^* = \mathcal{N}\left(0, \frac{1}{\alpha}\right)$$
$$\rho_Y^* = \mathcal{N}\left(0, \frac{1}{\alpha}\right).$$

(d) Suppose now we minimize the KL divergence in the opposite order:

$$\tilde{\rho}^* = \arg \min_{\rho = \rho_X \otimes \rho_Y} \mathsf{KL}(\nu \| \rho)$$

where we are minimizing over Gaussian distributions ρ_X, ρ_Y on $\mathbb{R}$. Show that the minimizer $\tilde{\rho}^* = \tilde{\rho}^*_X \otimes \tilde{\rho}^*_Y$ is given by the marginal distributions:

$$\begin{aligned} \tilde{\rho}^*_X &= \nu_X \\ \tilde{\rho}^*_Y &= \nu_Y. \end{aligned}$$

(P2) Let $G = (V, E)$ be a connected, undirected graph on n vertices $V = \{1, \dots, n\}$. Consider the Ising model, which models the joint distribution of random variables $X_i \in \{-1, 1\}$, $i \in V$, as

$$\nu(x_1, \dots, x_n) = \frac{1}{Z} \exp\left(\beta \sum_{(i,j) \in E} x_i x_j\right)$$

for all $(x_1, \dots, x_n) \in \{-1, 1\}^n$, for some $\beta \in \mathbb{R}$, where $Z = \sum_{\{-1,1\}^n} \exp\left(\beta \sum_{(i,j)\in E} x_i x_j\right)$ is the normalization constant. Let $N(i) = \{j \in V : (i, j) \in E\}$ be the set of neighbors of i.

(a) (Gibbs sampling.) For each $i \in V$, show that the conditional distribution of X_i given the other values $X_{\setminus i} = x_{\setminus i} = (x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n)$ is given by:

$$\nu(X_i = 1 \mid X_{\setminus i} = x_{\setminus i}) = \frac{1}{1 + \exp(-2\beta \sum_{j \in N(i)} x_j)}.$$

(b) (Mean field.) Suppose we want to approximate $\nu(x_1, \dots, x_n)$ by a product distribution $\hat{\nu}(x_1, \dots, x_n) = \bigotimes_{i \in V} \hat{\nu}_i(x_i)$ where $\hat{\nu}_i$ is a Bernoulli distribution on $\{-1, +1\}$ with parameter $p_i = \hat{\nu}_i(x_i = 1) \in [0, 1]$. We choose the best approximation by minimizing the KL divergence:

$$\min_{\hat{\nu} = \bigotimes_{i \in V} \hat{\nu}_i} \mathsf{KL}(\hat{\nu} \parallel \nu).$$

Show that the minimizer $\nu_i^* = \mathsf{Ber}(p_i^*)$ is characterized by $p_i^* = \Pr_{\nu_i^*}(x_i = 1)$ which satisfies the fixed point equations:

$$p_i^* = \frac{1}{1 + \exp(-2\beta \sum_{j \in N(i)} (2p_j^* - 1))} \qquad \forall\, i \in V.$$

(P3) Let $T : \mathbb{R}^d \to \mathbb{R}^m$ be a given function (the sufficient statistics). For $\theta \in \mathbb{R}^m$, consider the exponential family distribution

$$p_\theta(x) = \exp(\langle \theta, T(x) \rangle - A(\theta))$$

where $A(\theta) = \log \int_{\mathbb{R}^d} \exp(\langle \theta, T(x) \rangle)\, dx$ is the log-partition function, which is a function of the parameter θ with domain $\Theta = \{\theta \in \mathbb{R}^m : A(\theta) < \infty\}$.

(a) Show that the gradient of A with respect to θ gives the expected sufficient statistics: For all $\theta \in \Theta$,

$$\nabla A(\theta) = \mathbb{E}_{p_\theta}[T(X)].$$

(b) Show that the Hessian of A with respect to θ gives the covariance matrix of the sufficient statistics: For all $\theta \in \Theta$,

$$\nabla^2 A(\theta) = \mathsf{Cov}_{p_\theta}(T(X)).$$

(c) Show that p_θ is the maximum entropy distribution given the expected sufficient statistic. Concretely, for any $\theta \in \Theta$, let $\mu(\theta) = \mathbb{E}_{p_\theta}[T(X)] \in \mathbb{R}^m$. Show that:

$$p_\theta = \arg \max_{p\colon \mathbb{E}_p[T(X)] = \mu(\theta)} H(p)$$

where the maximization is over all probability distributions $p(x)$ on $\mathbb{R}^d$ with $\mathbb{E}_p[T(X)] = \mu(\theta)$. Here $H(p) = -\mathbb{E}_p[\log p]$ is the entropy of distribution p.

(*Hint:* Write down the Langrange multiplier for the constraint $\mathbb{E}_p[T(X)] = \mu(\theta)$.)

(P4) Let $\nu \propto e^{-f}$ be a probability distribution on $\mathbb{R}^d$ where $f\colon \mathbb{R}^d \to \mathbb{R}$ is twice differentiable. Recall the Fisher information of ν is defined as $J(\nu) = \mathbb{E}_\nu[\|\nabla f\|^2]$.

(a) Show that

$$\mathbb{E}_\nu[\nabla f] = 0.$$

(b) Show that we can also write the Fisher information as

$$J(\nu) = \mathbb{E}_\nu[\Delta f].$$

(Note that Δ is the Laplacian operator: $\Delta f = \mathsf{Tr}(\nabla^2 f)$.)

(c) Assume that f is L-smooth ($-LI \preceq \nabla^2 f(x) \preceq LI$ for all $x \in \mathbb{R}^d$). Show that

$$J(\nu) \le dL.$$

(P5) Choose a paper related to probabilistic machine learning that you find interesting. (The paper can be from your research, or see recent best papers from NeurIPS, ICLR, ICML, COLT, or `https://scorebasedgenerativemodeling.github.io`.).

(a) Write down what is the question that the paper is trying to answer.

(b) Write down what are the main results of the paper. Does it answer the question?

(c) Write down a question regarding something that you did not understand from the paper, or which was not addressed. For that question, either: (1) Answer the question by reading more related materials; or (2) Find out that the question has not been answered, in which case it would be an interesting question to study.

Additional questions for 586

(Q1) Let ρ, ν be probability distributions on $\mathbb{R}^d$ with twice-differentiable density functions. Recall the relative Fisher information of ρ with respect to ν is defined by

$$J_\nu(\rho) = \mathbb{E}_\rho\left[\left\|\nabla \log \frac{\rho}{\nu}\right\|^2\right].$$

(a) Let $\nu \propto e^{-f}$. Show that we can also write the relative Fisher information as:

$$J_\nu(\rho) = J(\rho) + \mathbb{E}_\rho[-2\Delta f + \|\nabla f\|^2].$$

(b) Compute the relative Fisher information between Gaussian distributions $\rho_1 = \mathcal{N}(\mu_1, \Sigma_1)$ and $\rho_2 = \mathcal{N}(\mu_2, \Sigma_2)$ on $\mathbb{R}^d$.

(Q2) Let $\nu \propto e^{-f}$ be a probability distribution on $\mathbb{R}^d$. Assume $f\colon \mathbb{R}^d \to \mathbb{R}$ is differentiable. Let $C = \mathsf{Cov}_\nu(X) \in \mathbb{R}^{d\times d}$ be the covariance matrix of ν.

(a) Show that

$$J(\nu) \geq \mathsf{Tr}(C^{-1}).$$

(*Hint:* Consider $J_\gamma(\nu)$ where γ is a Gaussian with the same mean and covariance as ν.)

(b) Show that

$$J(\nu) \geq \frac{d^2}{\mathrm{Var}_\nu(X)}.$$

(c) Assume f is L-smooth. Conclude that

$$\mathrm{Var}_\nu(X) \geq \frac{d}{L}.$$

(Q3) Let ρ_0 be a probability distribution on $\mathbb{R}^d$. Let $X_0 \sim \rho_0$ and $Z \sim \mathcal{N}(0, I)$ be independent. Let $X_t = X_0 + \sqrt{t}Z \in \mathbb{R}^d$ with density $\rho_t\colon \mathbb{R}^d \to \mathbb{R}$. Recall that ρ_t is given by the convolution:

$$\rho_t = \rho_0 \star \mathcal{N}(0, tI).$$

Concretely, for all $x \in \mathbb{R}^d$, the density value $\rho_t(x)$ is given by the formula:

$$\rho_t(x) = \frac{1}{(2\pi t)^{\frac{d}{2}}} \int_{\mathbb{R}^d} \rho_0(x_0) e^{-\frac{1}{2t}\|x - x_0\|^2}\, dx_0.$$

(a) Show that the formula $\rho_t(x)$ above satisfies the *heat equation*:

$$\frac{\partial \rho_t}{\partial t}(x) = \frac{1}{2}\Delta \rho_t(x).$$

(*Hint:* Compute both sides explicitly and check they are equal.)

(b) Let $f\colon \mathbb{R}^d \to \mathbb{R}$ be convex and twice differentiable. Show that

$$\mathbb{E}[f(X_t)] \geq \mathbb{E}[f(X_0)] \qquad \forall t \geq 0.$$

(c) Let $H(\rho) = -\mathbb{E}_\rho[\log \rho]$ be entropy. Show that

$$H(\rho_t) \geq H(\rho_0) \qquad \forall t \geq 0.$$