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.)

Solution: Note that since i and j are not edges in the graph, we can write

$$\begin{aligned}
\rho(x_i, x_j, x_{\setminus\{i,j\}}) &= \frac{1}{Z} \exp\left(-\alpha_i x_i^2 - \alpha_j x_j^2 + \sum_{(i',j') \in E,\ (i',j') \neq (i,j)} \beta_{i'j'} x_{i'} x_{j'} \right) \\
&= \exp(-\alpha_i x_i^2) \exp(-\alpha_j x_j^2) \cdot \frac{1}{Z} \exp\left(\sum_{(i',j') \in E,\ (i',j') \neq (i,j)} \beta_{i'j'} x_{i'} x_{j'} \right).
\end{aligned}$$

Then we have that

$$\begin{aligned}
\rho(x_i, x_{\setminus\{i,j\}}) &= \int_{\mathbb{R}} \exp(-\alpha_i x_i^2) \exp(-\alpha_j x_j^2) \cdot \frac{1}{Z} \exp\left(\sum_{(i',j') \in E,\ (i',j') \neq (i,j)} \beta_{i'j'} x_{i'} x_{j'} \right) dx_j \\
&= \exp(-\alpha_i x_i^2) \cdot \frac{1}{Z} \exp\left(\sum_{(i',j') \in E,\ (i',j') \neq (i,j)} \beta_{i'j'} x_{i'} x_{j'} \right) \cdot \int_{\mathbb{R}} \exp(-\alpha_j x_j^2) dx_j,
\end{aligned}$$

And similarly,

$$\rho(x_j, x_{\setminus\{i,j\}}) = \exp(-\alpha_j x_j^2) \cdot \frac{1}{Z} \exp\left(\sum_{(i',j') \in E,\ (i',j') \neq (i,j)} \beta_{i'j'} x_{i'} x_{j'} \right) \cdot \int_{\mathbb{R}} \exp(-\alpha_i x_i^2) dx_i,$$

$$\rho(x_{\setminus\{i,j\}}) = \frac{1}{Z} \exp\left(\sum_{(i',j') \in E,\ (i',j') \neq (i,j)} \beta_{i'j'} x_{i'} x_{j'} \right) \cdot \left(\int_{\mathbb{R}} \exp(-\alpha_i x_i^2) dx_i \right) \left(\int_{\mathbb{R}} \exp(-\alpha_j x_j^2) dx_j \right).$$

Let $A = \exp(-\alpha_i x_i^2)$, $B = \exp(-\alpha_j x_j^2)$, and $C = \frac{1}{Z} \exp\left(\sum_{(i',j') \in E,\ (i',j') \neq (i,j)} \beta_{i'j'} x_{i'} x_{j'} \right)$. Then we can rewrite everything succinctly as

$$\begin{aligned} \rho(x_i, x_j, x_{\setminus\{i,j\}}) &= A \cdot B \cdot C \\ \rho(x_i, x_{\setminus\{i,j\}}) &= A \cdot C \cdot \int_{\mathbb{R}} B dx_j \\ \rho(x_j, x_{\setminus\{i,j\}}) &= B \cdot C \cdot \int_{\mathbb{R}} A dx_i \\ \rho(x_{\setminus\{i,j\}}) &= C \cdot \int_{\mathbb{R}} A dx_i \int_{\mathbb{R}} B dx_j. \end{aligned}$$

Now, the LHS of the expression is

$$\rho(x_i, x_j \mid x_{\setminus\{i,j\}}) = \frac{\rho(x_i, x_j, x_{\setminus\{i,j\}})}{\rho(x_{\setminus\{i,j\}})} = \frac{ABC}{C \int A dx_i \int B dx_j} = \frac{AB}{\int A dx_i \int B dx_j}.$$

The RHS of the expression is

$$\begin{aligned} \rho(x_i \mid x_{\setminus\{i,j\}}) \cdot \rho(x_j \mid x_{\setminus\{i,j\}}) &= \frac{\rho(x_i, x_{\setminus\{i,j\}})}{\rho(x_{\setminus\{i,j\}})} \cdot \frac{\rho(x_j, x_{\setminus\{i,j\}})}{\rho(x_{\setminus\{i,j\}})} \\ &= \frac{AC \int B dx_j}{C \int A dx_i \int B dx_j} \cdot \frac{BC \int A dx_i}{C \int A dx_i \int B dx_j} \\ &= \frac{AB}{\int A dx_i \int B dx_j}. \end{aligned}$$

Thus, the two expressions are equal.

(b) Let $C = \mathsf{Cov}_\nu(X) \in \mathbb{R}^{n \times n}$ be the covariance matrix of $X = (X_1, \dots, X_n) \in \mathbb{R}^n$. 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.)

Solution: Let B be the $n \times n$ matrix

$$[B]_{ij} = \begin{cases} 2\beta_{ij} & \text{if } (i,j) \in E \\ 0 & \text{otherwise} \end{cases}$$

and let A be the diagonal matrix $\mathsf{diag}(2\alpha_1, \ldots, 2\alpha_n)$. We can rewrite $\rho(x)$ in the form

$$\rho(x) = \frac{1}{Z} \exp\left(-\frac{1}{2} x^\top (A - B) x\right).$$

This is a multivariate Gaussian distribution with $C^{-1} = A - B$. Note that $(C^{-1})_{ii} = A_{ii}$ for the diagonal elements, but for all $i \neq j$, we have $(C^{-1})_{ij} = B_{ij}$ and the components of B correspond exactly with the adjacency matrix of the graph (i.e., $B_{ij} \neq 0 \iff (i,j) \in E$ and vice versa).

(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:

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

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

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

[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.

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:

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

Include a snippet of your code or calculations.

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

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

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.

Solution: Refer to this colab notebook: `https://colab.research.google.com/drive/1encPv7tlDpB5fjezQGl31K-HMo3u1XTl?usp=sharing`.

(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 $\gamma = \mathcal{N}(0, I)$.

Solution: The marginal ρ_1 is given by a convolution with Gaussian:

$$\rho_1 = \rho_0 * \mathcal{N}(0, I) = \int_{\mathbb{R}^d} \rho_0(x)(2\pi)^{-d/2} \exp\left(-\frac{1}{2}\|y - x\|^2\right) dx.$$

(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*.)

Solution: We'll use the fact that $\nabla \log \rho_1(y) = \frac{\nabla \rho_1}{\rho_1}$. Computing the numerator:

$$\nabla \rho_1 = \nabla_y \int_{\mathbb{R}^d} \rho_0(x)(2\pi)^{-d/2} \exp\left(-\frac{1}{2}\|y - x\|^2\right) dx = \int_{\mathbb{R}^d} \rho_0(x)(2\pi)^{-d/2}(x-y) \exp\left(-\frac{1}{2}\|y - x\|^2\right) dx$$

Then we have that

$$\nabla \log \rho_1 = \frac{\nabla \rho_1}{\rho_1} = \int_{\mathbb{R}^d} (x - y) \frac{\rho_0(x)(2\pi)^{-d/2} \exp\left(-\frac{1}{2}\|y - x\|^2\right)}{\rho_1(y)} dx = \mathbb{E}_{\rho_{0|1=y}}[X] - y.$$

(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 $\mu = \mathbb{E}_{\rho_{0|1=y}}[X]$.

Solution: Consider the just the first term from the identity in part (b), and just its i^{th} component. Then

$$
\begin{aligned}
&\nabla_y \mathbb{E}_{\rho_{0|1=y}}[X]_i \\
&= \nabla_y \int_{\mathbb{R}^d} x_i \cdot \frac{\rho_0(x)\cdot(2\pi)^{-\frac{d}{2}}\exp(-\frac{1}{2}\|y-x\|^2)}{\rho_1(y)} dx \\
&= \int_{\mathbb{R}^d} x_i \rho_0(x)(2\pi)^{-\frac{d}{2}} \cdot \nabla_y \frac{\exp(-\frac{1}{2}\|y-x\|^2)}{\rho_1(y)} dx \\
&= \int_{\mathbb{R}^d} x_i \rho_0(x)(2\pi)^{-\frac{d}{2}} \cdot \left(\frac{(x-y)\exp(-\frac{1}{2}\|y-x\|^2)}{\rho_1(y)} - \frac{\exp(-\frac{1}{2}\|y-x\|^2)\nabla\rho_1(y)}{\rho_1(y)} \right) dx \\
&= \int_{\mathbb{R}^d} x_i \rho_0(x)(2\pi)^{-\frac{d}{2}} \cdot \left(\frac{(x-y)\exp(-\frac{1}{2}\|y-x\|^2)}{\rho_1(y)} - \frac{\exp(-\frac{1}{2}\|y-x\|^2)\rho_1(y)\,(\mu-y)}{\rho_1(y)^2} \right) dx \quad (*) \\
&= \int_{\mathbb{R}^d} x_i \rho_0(x)(2\pi)^{-\frac{d}{2}} \cdot \left(\frac{(x-\cancel{y})\exp(-\frac{1}{2}\|y-x\|^2)}{\rho_1(y)} - \frac{\exp(-\frac{1}{2}\|y-x\|^2)\,(\mu-\cancel{y})}{\rho_1(y)} \right) dx \\
&= \int_{\mathbb{R}^d} x_i \rho_0(x)(2\pi)^{-\frac{d}{2}} \cdot \left(\frac{(x)\exp(-\frac{1}{2}\|y-x\|^2)}{\rho_1(y)} - \frac{\exp(-\frac{1}{2}\|y-x\|^2)\,(\mu)}{\rho_1(y)} \right) dx
\end{aligned}
$$

where in (*) we used the identity from part (b) and the fact that $\nabla \log \rho = \frac{\nabla \rho}{\rho}$.
Consider the first term in the difference:

$$
\int_{\mathbb{R}^d} x_i \cdot \rho_0(x) \cdot (2\pi)^{-\frac{d}{2}} \cdot \left(\frac{x\exp(-\frac{1}{2}\|y-x\|^2)}{\rho_1(y)} \right) dx = \int_{\mathbb{R}^d} x_i x \cdot \rho_{0|1=y}(x)dx = \mathbb{E}_{\rho_{0|1=y}}[X_i X]
$$

The second term is

$$
-\mu \cdot \int_{\mathbb{R}^d} x_i \frac{\rho_0(x)\cdot(2\pi)^{-\frac{d}{2}}\exp(-\frac{1}{2}\|y-x\|^2)}{\rho_1(y)} dx = -\mu_i \cdot \mu.
$$

So, the $(i,j)^{th}$ component of $\nabla \mathbb{E}_{\rho_{0|1=y}}[X]$ is given by $\mathbb{E}_{\rho_{0|1=y}}[X_i X_j] - \mu_i \mu_j$, so the full matrix is given by

$$
\nabla \mathbb{E}_{\rho_{0|1=y}}[X] = \mathbb{E}[XX^\top] - \mu\mu^\top = \mathsf{Cov}_{\rho_{0|1=y}}[X].
$$

Now see that second term of the i^{th} component of $\nabla \log \rho_1(y)$ is y_i, whose partial derivative with respect to y_j is 1 if $j = i$ and 0 otherwise, giving the I term. Thus we have altogether that

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

(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:

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

Solution: Since f is α-strongly convex, we have that

$$\langle \nabla f(x), x - x^* \rangle \ge \alpha \|x - x^*\|^2 \Rightarrow \|x - x^*\|^2 \le \frac{1}{\alpha}\langle \nabla f(x), x - x^* \rangle$$

for all x. It remains to show that $\mathbb{E}_\nu[\langle \nabla f(X), X - x^* \rangle] \le d$.

Let $Z = \int e^{-f} dx$ so that $\nu = \frac{1}{Z}e^{-f}$. Now since $\mathbb{E}_\nu[\langle \nabla f(X), X - x^* \rangle] = \mathbb{E}_\nu\left[\sum_{i=1}^d \frac{\partial}{\partial x_i} f(X)(X_i - x_i^*)\right]$, consider one term of the summation:

$$\begin{aligned}
\mathbb{E}_\nu\left[\frac{\partial}{\partial x_i} f(X)(X_i - x_i^*)\right] &= \frac{1}{Z}\int_{\mathbb{R}^{d-1}}\int_{\mathbb{R}} \frac{\partial}{\partial x_i} f(x)(x_i - x_i^*) \cdot e^{-f(x)} dx_i dx_{\setminus i} && \text{Fubini} \\
&= \frac{1}{Z}\int_{\mathbb{R}^{d-1}}\int_{\mathbb{R}} e^{-f(x)} \cdot \frac{\partial}{\partial x_i}(x_i - x_i^*) dx_i dx_{\setminus i} && \text{IBP} \\
&= \frac{1}{Z}\int_{\mathbb{R}^{d-1}}\int_{\mathbb{R}} e^{-f(x)} dx_i dx_{\setminus i} \\
&= 1
\end{aligned}$$

Thus we have that

$$\mathbb{E}_\nu\left[\langle \nabla f(X), X - x^* \rangle\right] = d,$$

so we have the result

$$\mathbb{E}_\nu\left[\|X - x^*\|^2\right] \le \frac{1}{\alpha}\mathbb{E}_\nu\left[\langle \nabla f(X), X - x^* \rangle\right] = \frac{d}{\alpha}.$$

(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 $O(1)$.)

Solution: Note that $\nu = \mathcal{N}(0, I)$ is 1-SLC and 1-smooth. Thus, it satisfies the α-Poincaré inequality for $\phi(x) = \|x\|$. This ϕ is also 1-Lipschitz, since $\nabla\phi(x) = x/\|x\| \Rightarrow \|\nabla\phi(x)\| = 1$.

Now, by the Poincaré inequality and the fact that ϕ is 1-Lipschitz, we have

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

Also,

$$\mathbb{E}\left(\|Z\|^2\right) = d,$$

since this is also the sum of the individual, independent scalar variances of the components of Z. Finally, the upper bound on the variance gives

$$\mathbb{E}(\|Z\|)^2 = \mathbb{E}\left(\|Z\|^2\right) - \mathrm{Var}_\nu(\|Z\|) \Rightarrow \mathbb{E}(\|Z\|)^2 \ge d - 1 \Rightarrow \mathbb{E}(\|Z\|) \ge \sqrt{d-1}.$$

The variance is also lower bounded by 0, which yields

$$\mathbb{E}(\|Z\|)^2 \le d - 0 \Rightarrow \mathbb{E}(\|Z\|) \le \sqrt{d}.$$

(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 for 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 of $x \in \mathbb{R}^d$ (e.g., this holds when $y \mid x \sim \mathcal{N}(x, \alpha^{-1}I)$).

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:

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

Solution: The posterior distribution is proportional to the prior times the likelihood:

$$
\begin{aligned}
& p_n(x) \propto p_0(x) \cdot p(y \mid x) \\
\Rightarrow \qquad & -\log p_n(x) = f_0(x) + \sum_{i=1}^{n} \ell(x, y_i) + \text{constant}.
\end{aligned}
$$

Since f_0 is convex and $x \mapsto \ell(x, y)$ is α-strongly convex, we have that $\nabla^2 f_0(x) \succeq 0$ and $\nabla^2 \ell(x, y_i) \succeq \alpha I$ for $i = 1, \ldots, n$.

Thus we have that $-\nabla^2 \log p_n(x) \succeq n\alpha I$, which means that $-\log p_n(x)$ is $(n\alpha)$-strongly convex, or that p_n is $(n\alpha)$-strongly log-concave. Thus, p_n satisfies the $(n\alpha)$-Poincaré inequality. (for $\phi(X) = X$):

$$
\mathrm{Var}_{p_n}(X) \leq \frac{1}{n\alpha} \mathbb{E}_{p_n}[\| \underbrace{\nabla X}_{1-\text{vector}} \|^2] = \frac{d}{n\alpha}.
$$