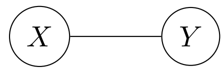
CPSC 486/586: Probabilistic Machine Learning *Out: February 1, 2023*

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 $|\beta| < \alpha$.

Solution: We can write

$$\begin{aligned}\nu(x, y) &\propto \exp\left(-\frac{\alpha}{2}\|x\|^2 - \frac{\alpha}{2}\|y\|^2 + \beta x^\top y\right) \\ &\propto \exp\left(-\frac{1}{2}\begin{pmatrix} x \\ y \end{pmatrix}^\top \begin{pmatrix} \alpha & -\beta \\ -\beta & \alpha \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix}\right)\end{aligned}$$

This shows that $\nu = \mathcal{N}(\mu, \Sigma)$ is Gaussian with

$$\mu = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

and

$$\Sigma = \begin{pmatrix} \alpha & -\beta \\ -\beta & \alpha \end{pmatrix}^{-1} = \frac{1}{\alpha^2 - \beta^2}\begin{pmatrix} \alpha & \beta \\ \beta & \alpha \end{pmatrix}$$

The last step above is valid when $\alpha^2 - \beta^2 \neq 0$. If $|\beta| < \alpha$, then $\alpha^2 - \beta^2 > 0$.

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

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

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

Solution: The X and Y marginals can be obtained from the components of μ, Σ:

$$\begin{aligned}\mu_X &= 0 \\ \mu_Y &= 0 \\ \Sigma_X &= \frac{\alpha}{\alpha^2 - \beta^2} \\ \Sigma_Y &= \frac{\alpha}{\alpha^2 - \beta^2}.\end{aligned}$$

(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

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

Solution: Note: $H(\rho) = -\mathbb{E}_\rho[\log \rho]$ is entropy. We can write:

$$\begin{aligned}\mathsf{KL}(\rho_X \otimes \rho_Y || \nu) &= \mathbb{E}_{\rho_X \otimes \rho_Y}\left[\log \frac{\rho_X(x)\rho_Y(y)}{\nu(x,y)}\right] \\ &= -H(\rho_X) - H(\rho_Y) - \mathbb{E}_{\rho_X \otimes \rho_Y}\left[\log \nu(x,y)\right],\end{aligned}$$

where the above uses the common decomposition of KL divergence into negative entropy and cross entropy. Then,

$$-\mathbb{E}_{\rho_X \otimes \rho_Y}\left[\log \nu(x,y)\right] = \frac{\alpha}{2}\mathbb{E}_{\rho_X}[x^2] + \frac{\alpha}{2}\mathbb{E}_{\rho_Y}[y^2] - \beta \mathbb{E}_{\rho_X}[x]\mathbb{E}_{\rho_Y}[y]$$

$$= \frac{\alpha}{2}(\Sigma_X + \mu_X^2) + \frac{\alpha}{2}(\Sigma_Y + \mu_Y^2) - \beta \mu_X \mu_Y$$

Altogether, we can write the KL divergence as a function of $\mu_X, \mu_Y, \Sigma_X, \Sigma_Y$ (dropping constant terms):

$$\mathcal{F}(\mu_X, \mu_Y, \Sigma_X, \Sigma_Y) = -\frac{1}{2}\log \Sigma_X + \frac{\alpha}{2}\Sigma_X + \frac{\alpha}{2}\mu_X^2 - \beta \mu_X \mu_Y - \frac{1}{2}\log \Sigma_Y + \frac{\alpha}{2}\Sigma_Y + \frac{\alpha}{2}\mu_Y^2$$

and minimize (checking that the function is convex in each variable):

$$\begin{aligned}
\frac{\partial \mathcal{F}}{\partial \Sigma_X} &= -\frac{1}{2\Sigma_X} + \frac{\alpha}{2} = 0 \iff \Sigma_X = \frac{1}{\alpha}, \qquad \frac{\partial^2 \mathcal{F}}{\partial \Sigma_X^2} = \frac{1}{2\Sigma_X^2} > 0 \\
\frac{\partial \mathcal{F}}{\partial \Sigma_Y} &= -\frac{1}{2\Sigma_Y} + \frac{\alpha}{2} = 0 \iff \Sigma_Y = \frac{1}{\alpha}, \qquad \text{(same)} \\
\nabla_{\mu_X, \mu_Y} \mathcal{F} &= \begin{pmatrix} \alpha \mu_X - \beta \mu_Y \\ \alpha \mu_Y - \beta \mu_X \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \iff \underbrace{\mu_X = \frac{\beta}{\alpha}\mu_Y, \mu_X = \frac{\alpha}{\beta}\mu_Y \Rightarrow}_{\alpha \neq \beta} \mu_X = \mu_Y = 0, \\
\nabla^2_{\mu_X, \mu_Y} \mathcal{F} &= \begin{pmatrix} \alpha & -\beta \\ -\beta & \alpha \end{pmatrix} \succ 0 \text{ if } \alpha^2 > \beta^2
\end{aligned}$$

Thus we have solved for $\rho_X^* \otimes \rho_Y^*$, with

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

(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:

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

Solution: As usual, we can write KL divergence as

$$\mathsf{KL}(\nu || \rho_X \otimes \rho_Y) = -H(\nu) + \mathbb{E}_\nu \left[-\log \rho_X - \log \rho_Y \right],$$

where the last term is

$$
\begin{aligned}
&\mathbb{E}_\nu\left[-\log \rho_X - \log \rho_Y\right] \\
&= \mathbb{E}_\nu\left[\frac{1}{2}\log(2\pi\Sigma_X) + \frac{(x-\mu_X)^2}{2\Sigma_X} + \frac{1}{2}\log(2\pi\Sigma_Y) + \frac{(y-\mu_Y)^2}{2\Sigma_Y}\right] \\
&= \frac{1}{2}\log(2\pi\Sigma_X) + \frac{1}{2}\log(2\pi\Sigma_Y) + \frac{1}{2\Sigma_X}\mathbb{E}_\nu[(x-\mu_X)^2] + \frac{1}{2\Sigma_Y}\mathbb{E}_\nu[(y-\mu_Y)^2]
\end{aligned}
$$

Working out the two expectation terms:

$$
\begin{aligned}
\mathbb{E}_\nu[(x-\mu_X)^2] &= \mathbb{E}_\nu[x^2 - 2x\mu_X + \mu_X^2] \\
&= \underbrace{\mathrm{Var}_\nu(x)}_{=\frac{\alpha}{\alpha^2-\beta^2}} + \underbrace{\mathbb{E}_\nu[x]^2}_{=0} - 2\underbrace{\mathbb{E}_\nu[x]}_{=0}\mu_2 + \mu_X^2 \\
&= \frac{\alpha}{\alpha^2-\beta^2} + \mu_X^2, \\
\text{Similarly,} \qquad \mathbb{E}_\nu[(y-\mu_Y)^2] &= \frac{\alpha}{\alpha^2-\beta^2} + \mu_Y^2
\end{aligned}
$$

Altogether, our objective (in $\mu_X, \mu_Y, \Sigma_X, \Sigma_Y$) is:

$$
\begin{aligned}
\mathcal{F}(\mu_X, \mu_Y, \Sigma_X, \Sigma_Y) \propto \frac{1}{2}\log(\Sigma_X) + \frac{\mu_X^2}{2\Sigma_X} + \frac{1}{2\Sigma_X}\left(\frac{\alpha}{\alpha^2-\beta^2}\right) \\
+ \frac{1}{2}\log(\Sigma_Y) + \frac{\mu_Y^2}{2\Sigma_Y} + \frac{1}{2\Sigma_Y}\left(\frac{\alpha}{\alpha^2-\beta^2}\right)
\end{aligned}
$$

We can consider the problem in the X variables only, since the problem in the Y variables is symmetric. Solving for critical points (first for μ_X):

$$
\frac{\partial \mathcal{F}}{\partial \mu_X} = \frac{\mu_X}{\Sigma_X} = 0 \iff \mu_X = 0
$$

Plug this into $\mathcal{F}$,

$$
\begin{aligned}
\mathcal{F}(\mu_X = 0, \Sigma_X) &= \frac{1}{2}\log(\Sigma_X) + \frac{1}{2\Sigma_X}\left(\frac{\alpha}{\alpha^2-\beta^2}\right) \\
\frac{\partial \mathcal{F}}{\partial \Sigma_X} &= \frac{1}{2\Sigma_X} - \frac{\alpha}{2(\alpha^2-\beta^2)\Sigma_X^2} = 0 \Rightarrow \Sigma_X = \frac{\alpha}{\alpha^2-\beta^2} \\
\frac{\partial^2 \mathcal{F}}{\partial \Sigma_X^2} &= -\frac{1}{2\Sigma_X^2} + \frac{\alpha}{(\alpha^2-\beta^2)\Sigma_X^3} = \frac{-(\alpha^2-\beta^2)\Sigma_X + 2\alpha}{2(\alpha^2-\beta^2)\Sigma_X^3} \geq 0 \iff \Sigma_X \leq \frac{2\alpha}{\alpha^2-\beta^2}
\end{aligned}
$$

The above says we are convex (in Σ_X) in the region $\Sigma_X \leq \frac{2\alpha}{\alpha^2-\beta^2}$, which our critical point from the first order condition satisfies. There are no other critical points, so this must be the minimizer. Similarly, we should get that

$$
\mu_Y = 0, \qquad \Sigma_Y = \frac{\alpha}{\alpha^2-\beta^2}.
$$

Thus, we get

$$\tilde{\rho}_X^* = \mathcal{N}\left(0, \frac{\alpha}{\alpha^2 - \beta^2}\right) = \nu_X$$
$$\tilde{\rho}_Y^* = \mathcal{N}\left(0, \frac{\alpha}{\alpha^2 - \beta^2}\right) = \nu_Y$$

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

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

for all $(x_1, \ldots, x_n) \in \{-1, 1\}^n$, 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 $X_{\setminus i} = x_{\setminus i} = (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)$ is given by:

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

Solution: By definition, we can compute

$$\begin{aligned}
\nu(X_i = 1 \mid X_{\setminus i} = x_{\setminus i}) &= \frac{\nu(X_i = 1, X_{\setminus i} = x_{\setminus i})}{\nu(X_{\setminus i} = x_{\setminus i})} \\
&= \frac{\nu(X_i = 1, X_{\setminus i} = x_{\setminus i})}{\nu(X_i = 1, X_{\setminus i} = x_{\setminus i}) + \nu(X_i = -1, X_{\setminus i} = x_{\setminus i})} \\
&\propto \frac{\exp\left(\beta \sum_{j\in N(i)} x_j\right)}{\exp\left(\beta \sum_{j\in N(i)} x_j\right) + \exp\left(-\beta \sum_{j\in N(i)} x_j\right)} \\
&= \frac{1}{1 + \exp\left(-2\beta \sum_{j\in N(i)} x_j\right)}
\end{aligned}$$

where the last equality is obtained by dividing the numerator and denominator by $\exp\left(\beta \sum_{j\in N(i)} x_j\right)$.

(b) (Mean field.) Suppose we want to approximate $\nu(x_1, \ldots, x_n)$ by a product distribution $\hat{\nu}(x_1, \ldots, x_n) = \bigotimes_{i \in V} \hat{\nu}_i(x_i)$ 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:

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

Solution: Note: $H(\rho) = -\mathbb{E}_\rho[\log \rho]$ is entropy.

$$\begin{aligned}
\mathsf{KL}(\hat{\nu} \parallel \nu) &= -H(\hat{\nu}) + \mathbb{E}_{\hat{\nu}}\left[\log(\nu)\right] \\
&= -\sum_i H(p_i) - \mathbb{E}_{\hat{\nu}}\left[\beta \sum_{(i,j)} x_i x_j\right] + \text{const.} \\
&= \sum_i (p_i \log p_i + (1 - p_i)\log(1 - p_i)) - \beta \sum_{(i,j)} \mathbb{E}_{\hat{\nu}}[x_i x_j] \\
&= \sum_i (p_i \log p_i + (1 - p_i)\log(1 - p_i)) - \beta \sum_{(i,j)} (2p_i - 1)(2p_j - 1)
\end{aligned}$$

Differentiate w.r.t. each p_i:

$$\begin{aligned}
\frac{\partial \mathsf{KL}(\hat{\nu} \parallel \nu)}{\partial p_i} &= p_i \cdot \frac{1}{p_i} + \log p_i + (1 - p_i) \cdot \frac{-1}{1 - p_i} + (-1)\log(1 - p_i) - \beta \sum_{j \in \mathcal{N}(i)} 2(2p_j - 1) \\
&= \log p_i - \log(1 - p_i) - 2\beta \sum_{(i,j)} (2p_j - 1)
\end{aligned}$$

Since $\mathsf{KL}(\hat{\nu} \parallel \nu)$ is convex with respect to each p_i, the optimality condition is $\frac{\partial \mathsf{KL}(\hat{\nu} \parallel \nu)}{\partial p_i} = 0$:

$$\log p_i - \log(1 - p_i) - 2\beta \sum_{(i,j)} (2p_j - 1) = 0 \Rightarrow \log\left(\frac{p_i}{1 - p_i}\right) = 2\beta \sum_{(i,j)} (2p_j - 1)$$

This is the logit function, whose inverse is the sigmoid function. That is, $\mathrm{logit}(p) = \sigma^{-1}(p)$ and $\sigma(\alpha) = \mathrm{logit}^{-1}(\alpha)$. So to solve for p_i we have

$$p_i = \sigma\left(2\beta \sum_{(i,j)} (2p_j - 1)\right) = \frac{1}{1 + \exp\left(-2\beta \sum_{(i,j)} (2p_j - 1)\right)}.$$

(P3) Let $T\colon \mathbb{R}^d \to \mathbb{R}^m$ 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 $\Theta = \{\theta \in \mathbb{R}^m \colon A(\theta) < \infty\}$.

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

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

Solution: You can also refer to the recitation notes from 02/03.

$$\begin{aligned}
\nabla A(\theta) &= \nabla \log \left(\int \exp(\theta^\top T(x)) dx \right) \\
&= \frac{\nabla_\theta \int \exp(\theta^\top T(x)) dx}{\int \exp(\theta^\top T(x)) dx} \\
&= \frac{\int \nabla_\theta \exp(\theta^\top T(x)) dx}{\int \exp(\theta^\top T(x)) dx} && \text{Use DCT} \\
&= \frac{\int T(x) \exp(\theta^\top T(x)) dx}{Z(\theta)} && Z(\theta) = \int \exp(\theta^\top T(x)) dx \\
&= \int T(x) \exp(\theta^\top T(x) - A(\theta)) dx && A(\theta) = \log Z(\theta) \\
&= \mathbb{E}_{p_\theta}[T(X)].
\end{aligned}$$

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

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

Solution: You can also refer to the recitation notes from 02/03.

$$\begin{aligned}
\nabla^2 A(\theta) &= \nabla_\theta \frac{\int T(x) \exp(\theta^\top T(x)) dx}{Z(\theta)} \\
&= \frac{\left(\int \nabla_\theta T(x) \exp(\theta^\top T(x)) dx\right) \cdot Z(\theta)}{Z(\theta)^2} - \frac{\left(\int T(x) \exp(\theta^\top T(x)) dx\right) \left(\int \nabla_\theta \exp(\theta^\top T(x)) dx\right)^\top}{Z(\theta)^2} \\
&= \frac{\int T(x) T(x)^\top \exp(\theta^\top T(x)) dx}{Z(\theta)} - \frac{\int T(x) \exp(\theta^\top T(x)) dx}{Z(\theta)} \cdot \frac{\left(\int T(x) \exp(\theta^\top T(x)) dx\right)^\top}{Z(\theta)} \\
&= \mathbb{E}_{p_\theta}[T(X)T(X)^\top] - \mathbb{E}_{p_\theta}[T(X)]\mathbb{E}_{p_\theta}[T(X)]^\top
\end{aligned}$$

$= \mathsf{Cov}_{p_\theta}(T(X))$

(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:

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

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 $\mathbb{E}_p[T(X)] = \mu(\theta)$.)

Solution: You can also refer to the recitation notes from 02/03. Here is another way to prove it. First, we'll prove a few small facts that we will use for the main proof using Lagrangian duality.

Claim 1 (Space of densities is convex.)**.** *Let $\mathcal{P} := \{p(\cdot) : \int p(x)dx = 1, p(x) \geq 0\}$. Then $\mathcal{P}$ is a convex set.*

Proof. Let $p, q \in \mathcal{P}$ arbitrary, and let $t \in (0, 1)$. Then

$$\begin{aligned}
\int_{\mathbb{R}^d} tp(x) + (1-t)q(x)dx &= t\int p(x)dx + (1-t)\int q(x)dx && \text{Linearity of integral} \\
&= t + (1-t) && p, q \in \mathcal{P} \\
&= 1.
\end{aligned}$$

Further, since $t, 1-t > 0$, and $p(x), q(x) \geq 0$ for all x, it also holds that $tp(x)+(1-t)q(x) \geq 0$ for all x. Thus $tp + (1-t)q \in \mathcal{P}$. ☐

Claim 2 (Negative Entropy functional is convex over $\mathcal{P}$)**.** *Let $-H(p) = \int p(x)\log p(x)dx$ for $p \in \mathcal{P}$. This functional $-H(p)$ is convex over $\mathcal{P}$.*

Proof. Pointwise, $u \mapsto u \log u$ is convex. Let $t \in (0, 1)$ and $p, q \in \mathcal{P}$. For each x,

$$\left(\underbrace{t\underbrace{p(x)}_{u} + (1-t)\underbrace{q(x)}_{v}}_{tu+(1-t)v}\right)\underbrace{\log(tp(x) + (1-t)q(x))}_{\log(tu+(1-t)v)} \leq tp(x)\log p(x) + (1-t)q(x)\log q(x)$$

This holds for all x. Taking integral over x (and using monotonicity of the integral):

$$\mathbb{E}_{tp+(1-t)q}[\log(tp + (1-t)q)] \leq t\mathbb{E}_p[\log p] + (1-t)\mathbb{E}_q[\log q].$$

☐

Claim 3 (Functional derivative of $-H(p)$ with respect to p)**.** *The functional derivative of $F(p) := -H(p)$ with respect to p is given by $\frac{\delta F}{\delta p} = 1 + \log p$.*

Proof. The functional differential of F in the direction of a function f is:

$$\begin{aligned}
\delta F(p)[f] &= \lim_{h\to 0} \frac{F(p+hf) - F(p)}{h} \\
&= \lim_{h\to 0} \frac{\int (p(x) + hf(x)) \log(p(x) + hf(x))dx - \int p(x) \log p(x) dx}{h} \\
&= \int \lim_{h\to 0} \left(\frac{(p(x) + hf(x)) \log(p(x) + hf(x)) - p(x) \log p(x)}{h} \right) dx && \text{DCT} \\
&= \int (1 + \log p(x)) f(x) dx && \text{Dir. deriv. pointwise}
\end{aligned}$$

This exists provided that f is a test function which satisfies the conditions needed to use dominated convergence theorem to exchange the limit and integral, for all h small enough. Then the functional derivative is $\frac{\delta F}{\delta p}$ that satisfies:

$$\delta F(p)[f] = \int \frac{\delta F}{\delta p}(x) f(x) dx.$$

We see that $\frac{\delta F}{\delta p}(x) = 1 + \log p(x)$ satisfies this. (Similarly, we can show that if the functional is given by $F(p) = -H(p) - \int \lambda^\top T(x) p(x)$, the corresponding functional derivative is $1 + \log p(x) - \lambda^\top T(x)$, which will show up later.) □

Main proof. Now, we have a convex functional with equality constraint which is linear in p (the expectation over p is a linear functional with respect to p). There exists a p, namely p_θ, which satisfies this constraint and with $p_\theta(x) > 0$ for all x (Slater's condition) so strong duality should hold. So it suffices to solve the dual problem optimally.

Since maximizing entropy is equivalent to minimizing the negative entropy, we can write the optimization problem as:

$$\begin{aligned}
&\min_{p\in\mathcal{P}} \ \mathbb{E}_p[\log p] \\
&\text{s.t. } \mathbb{E}_p[T(x)] = \mu(\theta).
\end{aligned}$$

The Lagrangian for the max entropy problem ($p \in \mathcal{P}, \lambda \in \mathbb{R}^m$):

$$\begin{aligned}
\mathcal{L}(p, \lambda) &= \int p(x) \log p(x) dx - \int \lambda^\top T(x) p(x) dx + \lambda^\top \mu(\theta) \\
&= \left(\int (p(x) \log p(x) - \lambda^\top T(x) p(x)) dx \right) + \lambda^\top \mu(\theta).
\end{aligned}$$

Using our earlier claim with $F(p) = \int p(x)\log p(x) - \lambda^\top T(x)p(x)dx$, we get $\frac{\delta F}{\delta p}(x) = 1 + \log p(x) - \lambda^\top T(x)$. We said this functional was convex (the negative entropy part is convex in p and the second term is linear in p), so the minimizing density function should be when $1 + \log p(x) - \lambda^\top T(x) = 0 \Rightarrow p(x) \propto \exp(\lambda^\top T(x))$. Since it is a density, we have $p(x) = \exp\left(\lambda^\top T(x) - A(\lambda)\right)$. Let's call this p_λ.[1]

The dual function for this problem is

$$\begin{aligned} g(\lambda) &= \inf_p \mathcal{L}(p,\lambda) = \mathcal{L}(p_\lambda, \lambda) \\ &= \mathbb{E}_{p_\lambda}[\log p_\lambda] - \lambda^\top \mathbb{E}_{p_\lambda}[T(x)] + \lambda^\top \mu(\theta) \\ &= \mathbb{E}_{p_\lambda}[\lambda^\top T(x) - A(\lambda)] - \lambda^\top \mathbb{E}_{p_\lambda}[T(x)] + \lambda^\top \mu(\theta) \\ &= \lambda^\top \mu(\theta) - A(\lambda). \end{aligned}$$

The dual problem becomes:

$$\max_{\lambda \in \mathbb{R}^m} \lambda^\top \mu(\theta) - A(\lambda).$$

Remember that $A(\lambda)$ is convex in λ, so its negative is concave. The first term is linear in λ. Solve by taking gradient (with respect to λ) equal to 0:

$$\mu(\theta) - \mathbb{E}_{p_\lambda}[T(x)] = 0 \Rightarrow \mathbb{E}_{p_\theta}[T(x)] = \mathbb{E}_{p_\lambda}[T(x)]$$

These are exactly equal if $p_\lambda = p_\theta$ with $\theta = \lambda$.

We can also check that it is the (unique) minimizer. Suppose there exists another density q such that $\mathbb{E}_q[T(X)] = \mu(\theta)$. Then

$$\begin{aligned} -H(q) &= \int q \log q dx \\ &= \int q \log \frac{q}{p_\theta} dx - \int q \log p_\theta dx \\ &= KL(q||p_\theta) - \int q(x) \left(\theta^\top T(x) - A(\theta)\right) dx \\ &= KL(q||p_\theta) - \int p_\theta(x) \left(\theta^\top T(x) - A(\theta)\right) dx \qquad (*) \\ &= KL(q||p_\theta) + H(p_\theta), \end{aligned}$$

where in (*) we used the fact that $\int q(x)T(x)dx = \int p_\theta(x)T(x)dx = \mu(\theta)$.

Since $KL(q||p_\theta) \geq 0$ is uniquely minimized when $q = p_\theta$, the negative entropy is indeed uniquely minimized by p_θ, given the constraint.

[1] Note: one could set up additional Lagrange multipliers for the constraints needed for p to be in $\mathcal{P}$, but the $p(x) \geq 0$ constraints will be unnecessary since we always have $p(x) > 0$ in this form, and due to complementary slackness we can set its multipliers to 0. The remaining multiplier corresponding to the normalization constraint will end up accounting for the $-A(\lambda)$ term.

Note: Finite state space. If we assume that the state space is finite, i.e., $|\mathcal{X}| = n$, then we can give a simpler proof as follows: we can write the negative entropy

$$-H(p) = \sum_{i=1}^{n} p(x_i) \log p(x_i),$$

and solve for the optimal (minimum) p_i for each x_i, using the fact that $u \mapsto u \log u$ is convex. The solution should be each $p_i \propto \exp(\lambda^\top T(x_i))$ and we get $p(x) = \exp(\lambda^\top T(x) - A(\lambda))$. The remaining dual problem should look like the continuous version, as well as the optimality condition.

(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

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

Solution: We use integration by parts. Let $Z = \int_{\mathbb{R}^d} e^{-f(x)} dx$, so $\nu(x) = e^{-f(x)}/Z$. For each component $i = 1, \ldots, d$:

$$\begin{aligned}
\mathbb{E}_\nu[\nabla f]_i &= \int_{\mathbb{R}^d} \nu(x) \frac{\partial}{\partial x_i} f(x) dx \\
&= \frac{1}{Z} \int_{\mathbb{R}^d} e^{-f(x)} \frac{\partial}{\partial x_i} f(x) dx \\
&= \frac{1}{Z} \int_{\mathbb{R}^{d-1}} \int_{\mathbb{R}} e^{-f(x)} \frac{\partial}{\partial x_i} f(x) dx_i \, dx_{\setminus i} \\
&= \frac{1}{Z} \int_{\mathbb{R}^{d-1}} \int_{\mathbb{R}} e^{-u} du dx_{\setminus i} && \text{Fubini's, } u = f(x), du = \frac{\partial}{\partial x_i} f(x) dx_i \\
&= \frac{1}{Z} \int_{\mathbb{R}^{d-1}} -e^{-f(x)} \Big|_{x_i=-\infty}^{x_i=+\infty} dx_{\setminus i} \\
&= \frac{1}{Z} \int_{\mathbb{R}^{d-1}} (0 - 0) dx_{\setminus i} && \lim_{x_i \to \pm\infty} e^{-f(x)} = 0 \\
&= 0
\end{aligned}$$

(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: $\Delta f = \mathsf{Tr}(\nabla^2 f)$.)

Solution: Use integration by parts.

$$
\begin{aligned}
J(\nu) &= \mathbb{E}_\nu[\|\nabla f\|^2] \\
&\propto \int_{\mathbb{R}^d} \|\nabla f(x)\|^2 e^{-f(x)} dx \\
&= \int_{\mathbb{R}^d} \langle \nabla f(x), \nabla f(x) \rangle \, e^{-f(x)} dx \\
&= \int_{\mathbb{R}^d} \sum_{i=1}^d \frac{\partial}{\partial x_i} f(x) \cdot \frac{\partial}{\partial x_i} f(x) e^{-f(x)} dx
\end{aligned}
$$

Consider one of the terms in the summation:

$$
\begin{aligned}
&\int_{\mathbb{R}^d} -\frac{\partial}{\partial x_i} f(x) \cdot \underbrace{-\frac{\partial}{\partial x_i} f(x) e^{-f(x)}}_{dv} dx \\
&= \int_{\mathbb{R}^{d-1}} -\frac{\partial}{\partial x_i} f(x) e^{-f(x)} \Big|_{x_i=-\infty}^{x_i=+\infty} + \int_{\mathbb{R}} \frac{\partial^2}{\partial x_i^2} f(x) e^{-f(x)} dx_i dx_{\setminus i} && \text{Fubini's, IBP} \\
&= \int_{\mathbb{R}^d} \frac{\partial^2}{\partial x_i^2} f(x) e^{-f(x)} dx && (*)
\end{aligned}
$$

(*) If $\lim_{x_i \to \infty} \frac{\partial}{\partial x_i} f(x) e^{-f(x)} = \lim_{x_i \to -\infty} \frac{\partial}{\partial x_i} f(x) e^{-f(x)} = 0$ or the same finite constant.
If we sum over all the terms, we get

$$
\int_{\mathbb{R}^d} \left(\sum_{i=1}^d \frac{\partial^2}{\partial x_i^2} f(x) \right) e^{-f(x)} dx = \mathbb{E}_\nu[\Delta f]
$$

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

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

Solution: By the earlier parts, $J(\nu) = \mathbb{E}_\nu[\|\nabla f\|^2] = \mathbb{E}_\nu[\Delta f]$.
Note that $\nabla^2 f(x)$ is a symmetric, real matrix. The spectral theorem says it has real eignevalues. Let $\lambda_1, \ldots, \lambda_d$ be the eigenvalues of $\nabla^2 f(x)$. Since f is L-smooth, we know $|\lambda_i| \le L$ for $i = 1, \ldots, d$. Then for all $x \in \mathbb{R}^d$,

$$
\Delta f(x) = \mathsf{Tr}(\nabla^2 f(x)) = \sum_{i=1}^d \lambda_i \le dL.
$$

Therefore,

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

(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:

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

Solution: Using integration by parts,

$$\begin{aligned}
J_\nu(\rho) &= \mathbb{E}_\rho\left[\left\|\nabla \log \frac{\rho}{\nu}\right\|^2\right] \\
&= \int_{\mathbb{R}^d} \|\nabla \log \rho(x) - \nabla \log \nu(x)\|^2 \rho(x) dx \\
&= \int_{\mathbb{R}^d} \left(\|\nabla \log \rho(x)\|^2 - 2\left\langle \nabla \log \rho(x), \nabla \log \nu(x)\right\rangle + \|\nabla \log \nu(x)\|^2\right) \rho(x) dx \\
&= \underbrace{\mathbb{E}_\rho\left[\|\nabla \log \rho\|^2\right]}_{J(\rho)} + \mathbb{E}_\rho\left[\|\nabla f\|^2\right] - 2\int_{\mathbb{R}^d} \left\langle \nabla \log \rho(x), \nabla \log \nu(x)\right\rangle \rho(x) dx \\
&= J(\rho) + \mathbb{E}_\rho\left[\|\nabla f\|^2\right] - 2\int_{\mathbb{R}^d} \left\langle \frac{\nabla \rho(x)}{\rho(x)}, -\nabla f(x)\right\rangle \rho(x) dx \\
&= J(\rho) + \mathbb{E}_\rho\left[\|\nabla f\|^2\right] + 2\int_{\mathbb{R}^d} \left\langle \nabla \rho(x), \nabla f(x)\right\rangle dx && \text{IBP (from PS1)}
\end{aligned}$$

$$= J(\rho) + \mathbb{E}_\rho\left[\|\nabla f\|^2\right] - 2\int_{\mathbb{R}^d} \rho(x) \underbrace{\nabla \cdot (\nabla f(x))}_{\Delta f} dx$$

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

(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 $\mathbb{R}^d$.

Solution: Note that $\log \rho_2 = -\frac{1}{2}\|x-\mu_2\|^2_{\Sigma_2^{-1}}$ and

$$\nabla(\log \rho_2) = -\Sigma_2^{-1}(x-\mu_2) \qquad \nabla^2(\log \rho_2) = -\Sigma_2^{-1}, \qquad \Delta(\log \rho_2) = -\mathsf{Tr}(\Sigma_2^{-1}).$$

The same identities hold for ρ_1 but with μ_1 and Σ_1. Using the identity from part (a) (and noting that $f = -\log \rho_2$), with the results from P4,

$$\begin{aligned}
J_{\rho_2}(\rho_1) &= -\mathbb{E}_{\rho_1}[\Delta \log \rho_1] + \mathbb{E}_{\rho_1}[2\Delta(\log \rho_2) + \|\nabla \log \rho_2\|^2] \\
&= \mathsf{Tr}(\Sigma_1^{-1}) - 2\mathsf{Tr}(\Sigma_2^{-1}) + \mathbb{E}_{\rho_1}\left[\mathsf{Tr}\left(\Sigma_2^{-1}\underbrace{(x-\mu_2)(x-\mu_2)^\top}\Sigma_2^{-1}\right)\right] \\
&= \mathsf{Tr}(\Sigma_1^{-1}) - 2\mathsf{Tr}(\Sigma_2^{-1}) + \mathbb{E}_{\rho_1}\left[\mathsf{Tr}\left(\Sigma_2^{-1}(x-\mu_1+\mu_1-\mu_2)(x-\mu_1+\mu_1-\mu_2)^\top\Sigma_2^{-1}\right)\right] \\
&= \mathsf{Tr}(\Sigma_1^{-1}) - 2\mathsf{Tr}(\Sigma_2^{-1}) + \mathbb{E}_{\rho_1}\left[\mathsf{Tr}\left(\Sigma_2^{-1}\Sigma_1\Sigma_2^{-1}\right)\right] + \mathbb{E}_{\rho_1}\left[\mathsf{Tr}\left(\Sigma_2^{-1}(\mu_1-\mu_2)(\mu_1-\mu_2)^\top\Sigma_2^{-1}\right)\right] \\
&= \mathsf{Tr}\left(\Sigma_2^{-2}\Sigma_1 - 2\Sigma_2^{-1} + \Sigma_1^{-1}\right) + \|\Sigma_2^{-1}(\mu_1-\mu_2)\|^2.
\end{aligned}$$

The last line used the cyclic property of trace, the fact that the covariance matrix is symmetric $\Sigma_2^{-1} = \Sigma_2^{-\top}$, and the fact that $\mathsf{Tr}(vv^\top) = \mathsf{Tr}(v^\top v) = \|v\|^2$ for a vector v.

(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.)

Solution: Let $\mu = \mathbb{E}_\nu[X]$. Note $\gamma \propto \exp\left(-\frac{1}{2}\|x-\mu\|^2_{C^{-1}}\right)$, so $\log\gamma = -\frac{1}{2}\|x-\mu\|^2_{C^{-1}}$, and

$$\nabla(\log\gamma) = -C^{-1}(x-\mu), \qquad \nabla^2(\log\gamma) = -C^{-1}, \qquad \Delta(\log\gamma) = -\mathsf{Tr}(C^{-1}).$$

Using the identity from Q1 (and noting that $f = -\log\gamma$),

$$J_\gamma(\nu) = J(\nu) + \mathbb{E}_\gamma\left[2\Delta(\log\gamma) + \|\nabla\log\gamma\|^2\right]$$

$$
\begin{aligned}
&= J(\nu) - 2\mathsf{Tr}(C^{-1}) + \mathbb{E}_\gamma \left[\mathsf{Tr}\left(C^{-1} \underbrace{(x-\mu)(x-\mu)^\top}_{C} C^{-1} \right) \right] \\
&= J(\nu) - \mathsf{Tr}(C^{-1}).
\end{aligned}
$$

We get the inequality $J(\nu) \geq \mathsf{Tr}(C^{-1})$ by noting that $J_\gamma(\nu) \geq 0$.

(b) Show that

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

Solution: Let $\lambda_1, \ldots, \lambda_d$ be the eigenvalues of C. Then C^{-1} has eigenvalues $\frac{1}{\lambda_1}, \ldots, \frac{1}{\lambda_d}$. By part (a),

$$
\begin{aligned}
J(\nu) &\geq \mathsf{Tr}(C^{-1}) \\
&= \sum_{i=1}^d \frac{1}{\lambda_i} \\
&\geq \frac{d^2}{\sum_{i=1}^d \lambda_i} \qquad \text{HM-AM Inequality: } \frac{d}{\frac{1}{x_1} + \cdots + \frac{1}{x_d}} \leq \frac{x_1 + \cdots + x_d}{d} \\
&= \frac{d^2}{\mathsf{Tr}(C)} \\
&= \frac{d^2}{\mathrm{Var}_\nu(X)}.
\end{aligned}
$$

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

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

Solution: Use P4(c), which says that $J(\nu) \leq dL$. Part (b) says that $\mathrm{Var}_\nu(X) \geq \frac{d^2}{J(\nu)} \Rightarrow \mathrm{Var}_\nu(X) \geq \frac{d^2}{dL} = \frac{d}{L}$.

(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.)

Solution: We compute both sides explicitly and verify they are equal.

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

Simplify the first term separately,

$$\begin{aligned}
&\left(\frac{\partial}{\partial t}\frac{1}{(2\pi t)^{\frac{d}{2}}}\right)\left(\int_{\mathbb{R}^d} \rho_0(x_0) e^{-\frac{1}{2t}\|x-x_0\|^2}\, dx_0\right)\\
&= -\frac{d}{2}(2\pi t)^{-d/2-1}\cdot 2\pi \int_{\mathbb{R}^d} \rho_0(x_0) e^{-\frac{1}{2t}\|x-x_0\|^2}\, dx_0\\
&= -\frac{d}{2t}\cdot \frac{1}{(2\pi t)^{\frac{d}{2}}}\int_{\mathbb{R}^d} \rho_0(x_0) e^{-\frac{1}{2t}\|x-x_0\|^2}\, dx_0\\
&= -\frac{d}{2t}\cdot \rho_t(x)
\end{aligned}$$

Then the next term,

$$\begin{aligned}
&\frac{1}{(2\pi t)^{\frac{d}{2}}}\left(\frac{\partial}{\partial t}\int_{\mathbb{R}^d} \rho_0(x_0) e^{-\frac{1}{2t}\|x-x_0\|^2}\, dx_0\right)\\
&= \frac{1}{(2\pi t)^{\frac{d}{2}}}\int_{\mathbb{R}^d} \frac{\partial}{\partial t}\left(\rho_0(x_0) e^{-\frac{1}{2t}\|x-x_0\|^2}\right) dx_0 && \text{DCT}\\
&= \frac{1}{(2\pi t)^{\frac{d}{2}}}\int_{\mathbb{R}^d} \rho_0(x_0) e^{-\frac{1}{2t}\|x-x_0\|^2}\frac{1}{2t^2}\|x-x_0\|^2 dx_0
\end{aligned}$$

And altogether, we have

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

Now calculating the right hand side, first note the gradient (in one component i):

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

$$= \frac{1}{(2\pi t)^{\frac{d}{2}}} \int_{\mathbb{R}^d} \rho_0(x_0) e^{-\frac{1}{2t}\|x-x_0\|^2} \cdot \left(-\frac{1}{t}(x-x_0)_i\right) dx_0 \qquad \text{DCT}$$

Now consider a diagonal element of the Hessian:

$$\begin{aligned}
\frac{\partial^2}{\partial x_i^2}\rho_t(x) &= \frac{\partial}{\partial x_i}\left(\frac{1}{(2\pi t)^{\frac{d}{2}}} \int_{\mathbb{R}^d} \rho_0(x_0) e^{-\frac{1}{2t}\|x-x_0\|^2} \cdot \left(-\frac{1}{t}(x-x_0)_i\right) dx_0\right) \\
&= \frac{1}{(2\pi t)^{\frac{d}{2}}} \int_{\mathbb{R}^d} \rho_0(x_0) e^{-\frac{1}{2t}\|x-x_0\|^2} \cdot \left(-\frac{1}{t}\right) dx_0 \\
&\qquad + \frac{1}{(2\pi t)^{\frac{d}{2}}} \int_{\mathbb{R}^d} \rho_0(x_0) e^{-\frac{1}{2t}\|x-x_0\|^2} \cdot \left(\frac{1}{t^2}(x-x_0)_i^2\right) dx_0 \\
&= -\frac{1}{t}\rho_t(x) + \frac{1}{(2\pi t)^{\frac{d}{2}}} \int_{\mathbb{R}^d} \rho_0(x_0) e^{-\frac{1}{2t}\|x-x_0\|^2} \cdot \left(\frac{1}{t^2}(x-x_0)_i^2\right) dx_0
\end{aligned}$$

Then we compute the Laplacian by taking the trace of the Hessian:

$$\begin{aligned}
\Delta\rho_t(x) &= \sum_{i=1}^{d}\left(-\frac{1}{t}\rho_t(x) + \frac{1}{(2\pi t)^{\frac{d}{2}}} \int_{\mathbb{R}^d} \rho_0(x_0) e^{-\frac{1}{2t}\|x-x_0\|^2} \cdot \left(\frac{1}{t^2}(x-x_0)_i^2\right) dx_0\right) \\
&= -\frac{d}{t}\rho_t(x) + \frac{1}{(2\pi t)^{\frac{d}{2}}} \int_{\mathbb{R}^d} \rho_0(x_0) e^{-\frac{1}{2t}\|x-x_0\|^2} \cdot \left(\frac{1}{t^2}\|x-x_0\|^2\right) dx_0.
\end{aligned}$$

Thus, we get the equality with $\partial\rho_t(x)/\partial t$ by multiplying this expression by $1/2$.

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

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

Solution: We can compute the time derivative:

$$\begin{aligned}
\frac{d}{dt}\mathbb{E}[f(X_t)] &= \frac{d}{dt}\int_{\mathbb{R}^d} \rho_t(x) f(x) dx \\
&= \int_{\mathbb{R}^d} \frac{\partial \rho_t}{\partial t}(x)\, f(x) dx \\
&= \int_{\mathbb{R}^d} f(x) \frac{1}{2}\Delta\rho_t(x) dx \qquad \text{(part (a))} \\
&= \frac{1}{2}\int_{\mathbb{R}^d} \Delta f(x)\, \rho_t(x) dx \\
&= \frac{1}{2}\mathbb{E}[\Delta f(X_t)].
\end{aligned}$$

In the above, we have used the integration by parts formula and the identity $\Delta f = \nabla\cdot\nabla f$:

$$\int_{\mathbb{R}^d} f(x)\Delta\rho(x)\, dx = -\int_{\mathbb{R}^d} \langle \nabla f(x), \nabla\rho(x)\rangle dx = \int_{\mathbb{R}^d} \Delta f(x)\rho(x)\, dx.$$

Since f is convex, $\nabla^2 f(x) \succeq 0$, so $\Delta f(x) \geq 0$. Therefore,

$$\frac{d}{dt}\mathbb{E}[f(X_t)] = \frac{1}{2}\mathbb{E}[\Delta f(X_t)] \geq 0.$$

This means $\mathbb{E}[f(X_t)] \geq \mathbb{E}[f(X_0)]$ for all $t \geq 0$.

(Alternatively, we can also use Jensen's inequality by noting that $\mathbb{E}[X_t \mid X_0] = X_0$.)

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

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

Solution: We can compute the time derivative of entropy and show it is nonnegative:

$$\begin{aligned}
\frac{\partial}{\partial t} H(\rho_t) &= -\frac{\partial}{\partial t}\int_{\mathbb{R}^d} \rho_t(x)\log\rho_t(x)dx \\
&= -\int_{\mathbb{R}^d} \frac{\partial \rho_t}{\partial t}(x)\left(\log\rho_t(x) + 1\right)dx && \text{DCT} \\
&= -\int_{\mathbb{R}^d} \frac{1}{2}\Delta\rho_t(x)(\log\rho_t(x)+1)dx && \text{Part (a)} \\
&= -\frac{1}{2}\int_{\mathbb{R}^d} \nabla\cdot(\nabla\rho_t(x))(\log\rho_t(x)+1)dx \\
&= \frac{1}{2}\int_{\mathbb{R}^d} \langle \nabla \log\rho_t(x), \nabla\rho_t(x)\rangle\, dx && \text{IBP} \\
&= \frac{1}{2}\int_{\mathbb{R}^d} \left\langle \frac{\nabla\rho_t(x)}{\rho_t(x)}, \nabla\rho_t(x)\right\rangle dx \\
&= \frac{1}{2}\int_{\mathbb{R}^d} \frac{\|\nabla\rho_t(x)\|^2}{\rho_t(x)}dx \\
&\geq 0.
\end{aligned}$$

Thus, we have shown that the time derivative of entropy along the heat flow is given by the Fisher information:

$$\frac{\partial}{\partial t}H(\rho_t) = \frac{1}{2}J(\rho_t)$$

where

$$J(\rho_t) = \int_{\mathbb{R}^d} \frac{\|\nabla\rho_t(x)\|^2}{\rho_t(x)}dx = \mathbb{E}_{\rho_t}[\|\nabla\log\rho_t\|^2]$$

is the Fisher information.