HOMEWORK 5
BEYOND LIKELIHOOD[1]

10-708 PROBABILISTIC GRAPHICAL MODELS (SPRING 2022)
https://andrejristeski.github.io/10708-22/

OUT: 4/13/22
DUE: 4/27/22 at 11:59 PM
TAs: Kai-Ling, Aaron

START HERE: Instructions

- **Collaboration policy:** The purpose of student collaboration is to facilitate learning, not to circumvent it. Studying the material in groups is strongly encouraged. It is also allowed to seek help from other students in understanding the material needed to solve a particular homework problem, provided no written notes (including code) are shared, or are taken at that time, and provided learning is facilitated, not circumvented. The actual solution must be done by each student alone. The presence or absence of any form of help or collaboration, whether given or received, must be explicitly stated and disclosed in full by all involved. See the Academic Integrity Section on the course site for more information: https://andrejristeski.github.io/10708-22/#:~:text=Academic%20Integrity%20Policies
- **Late Submission Policy:** See the late submission policy here: https://andrejristeski.github.io/10708-22/#:~:text=Grace%20Day/Late%20Homework%20Policy
- **Submitting your work to Gradescope:** We use Gradescope (https://www.gradescope.com/courses/349316/assignments) to collect PDF submissions of open-ended questions on the homework (e.g. mathematical derivations, plots, short answers). The course staff will manually grade your submission, and you'll receive personalized feedback explaining your final marks. The homework template must be used and can be completed in Latex or by hand. Handwritten submissions must be legible otherwise we will not be able to give credit to your solutions. No changes should be made to the template, boxes and choices **MUST** remain the same size and in the same locations between the template and your completed submission, the document has 13 pages so your submission must contain no more and no less than 13 pages.
- For **multiple choice** or **select all that apply questions**, shade in the box or circle in the template document corresponding to the correct answer(s) for each of the questions. For LaTeX users, replace `\choice` with `\CorrectChoice` to obtain a shaded box/circle, and don't change anything else.

[1]Compiled on Wednesday 13th April, 2022 at 23:40

A Written Questions [100 pts]

1. (Exploring properties of losses when applied to exponential families)

   In this question, we will derive some properties of the losses we saw in class: maximum likelihood, noise contrastive estimation and score matching, when applied to learning exponential families.

   An *exponential family* of distributions is a set of distributions $\{p_\theta : \theta \in \mathbb{R}^k\}$, described by the *sufficient statistics* for the family $T(x) : \mathbb{R}^d \to \mathbb{R}^k$. The members of the family have the form

   $$p_\theta(x) : \mathbb{R}^d \to \mathbb{R}; p_\theta(x) \propto \exp\left(\langle \theta, T(x) \rangle\right)$$

   You can always assume in this problem that the functions $T(x)$ are twice differentiable, and such that the partition function is finite (that is, $\int_{x \in \mathbb{R}^d} \exp\left(\langle \theta, T(x) \rangle\right) < \infty$).

   (a) [9 pts] Consider first score matching. Let us parametrize the score function as $s_\theta(x) = DT(x)\theta$, where $DT(x) \in \mathbb{R}^{d \times k}$ denotes the Jacobian of function $T(x)$ and $\theta \in \mathbb{R}^k$ is the (unknown) vector of parametrize we are training. Suppose we are given samples $x_1, x_2, \ldots, x_N$ from some distribution p_{data}. Find the closed-form solution of the minimum of the score-matching loss:

   $$\text{argmin}_\theta \frac{1}{N} \sum_{i=1}^{N} \left(\|s_\theta(x_i)\|^2 + 2\text{Tr}(Ds_\theta(x_i)) \right)$$

   where Ds_θ denotes the Jacobian of s_θ.

(b) [10 pts] Proceeding to NCE, remember that if we denote $p_{\theta,c}(x) = \exp\left(\langle \theta, T(x)\rangle - c\right)$, we train a classifier $D_{\theta,c}(x) = \frac{p_{\theta,c}(x)}{p_{\theta,c}(x)+q(x)}$. It will be convenient to denote by $\tilde{T}(x)$ the vector $\tilde{T}(x) = (T(x), 1)$—that is, the vector $T(x)$ with 1 appended as an additional coordinate. Similarly, let $\tilde{\theta} = (\theta, c)$ — i.e. the vector θ with c appended as an additional coordinate. With this notation, we can write $p_{\theta,c}(x)$ as $p_{\tilde{\theta}}(x) = \exp\left(\langle \tilde{\theta}, \tilde{T}(x)\rangle\right)$ and $D_{\tilde{\theta}}(x) = \frac{p_{\tilde{\theta}}(x)}{p_{\tilde{\theta}}(x)+q(x)}$.

The loss we saw in class can be written as:

$$L(D_{\tilde{\theta}}) = -\frac{1}{k+1}\mathbb{E}_{p_{\text{data}}} \log(D_{\tilde{\theta}}(x)) - \frac{k}{k+1}\mathbb{E}_q \log(1 - D_{\tilde{\theta}}(x))$$

Prove that the loss is convex in $\tilde{\theta}$ (which is the same as the loss being convex in θ, c).
Hint: a twice differentiable function is convex iff its Hessian is positive-definite.

(c) [10 pts] Finally, we'll see that when $q = p_{\text{data}}$, asymptotically the efficiency of NCE and MLE is the same. Recall that for a loss L, asymptotically, the statistical efficiency is governed by the Hessian at the optimum. (Intuitively, as we get more samples, the empirical loss at the optimum θ^* starts looking like a Gaussian centered at the optimum with a covariance matrix given by the Hessian at θ^*.)

First, consider the maximum likelihood loss

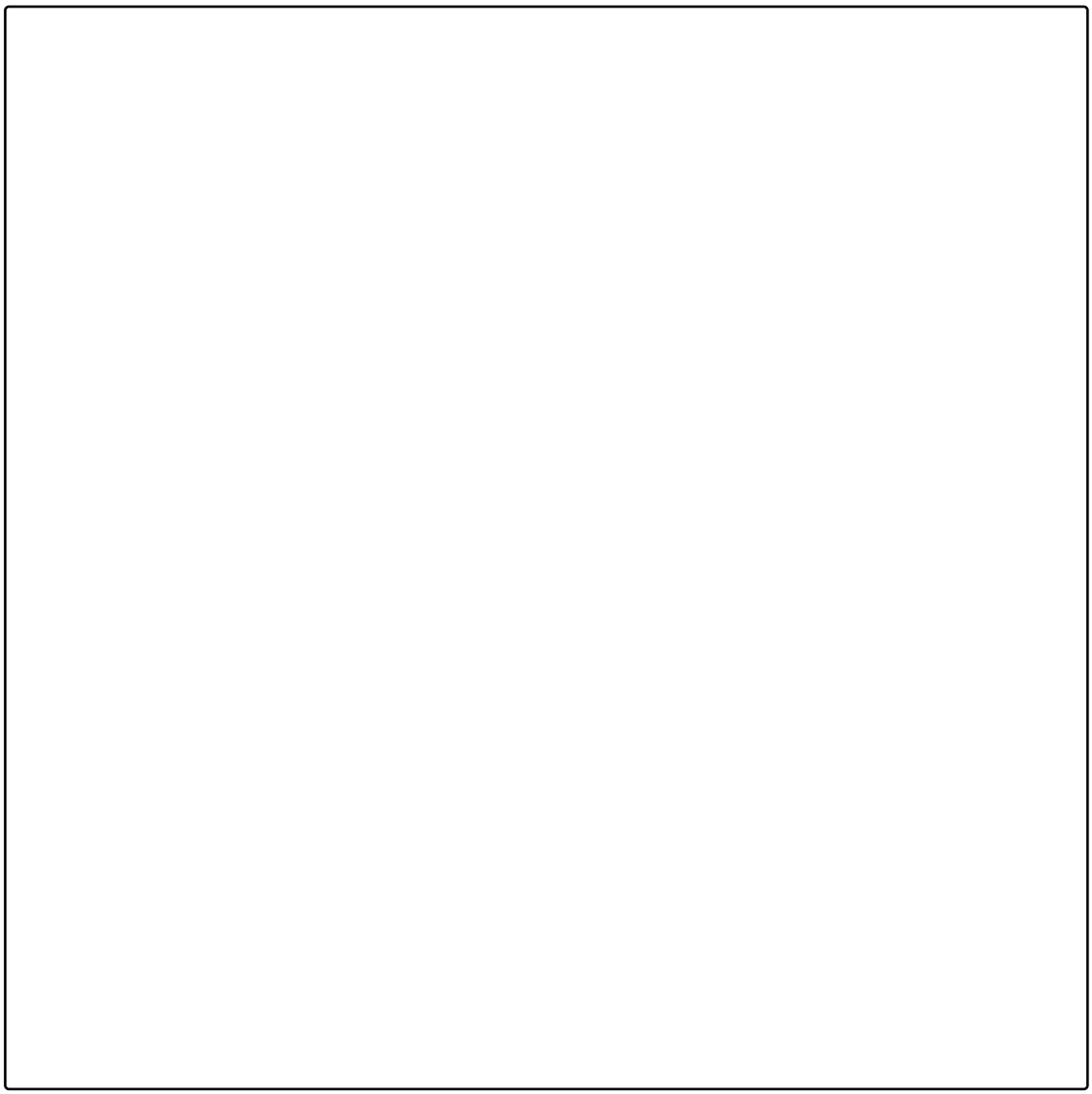
$$L_{\text{MLE}}(\theta) = -\mathbb{E}_{p_{\text{data}}} \log p_\theta(x)$$

If $p_{\text{data}} = p_{\theta^*}$, the minimum of this loss, as well as the minimum of the NCE loss will be reached at θ^*.

Let's further assume that $\mathbb{E}_{p_{\theta^*}}[T(x)] = 0$ (zero vector). Show that the Hessian at θ^* matches the Hessian of the NCE loss when $q = p_{\text{data}}$, up to a constant scaling, that is, there exists $c > 0$ such that:

$$\nabla^2 L_{\text{MLE}}(\theta^*) = c \cdot L_{\text{NCE}}(\theta^*)$$

Hint: For MLE, the partition function is ***not*** *a parameter we are optimizing over like in NCE. In other words, when calculating derivatives and Hessians, you need to take into account the dependence of Z_θ on θ.*



2. (Generative Adversarial Network) Consider the setup for W-GAN s.t. the class of generators $\mathcal{G}$ is parametrized by a matrix $W \in \mathbb{R}^{d\times d}$, s.t. $G_W(z) = Wz$, and the class of discriminators $\mathcal{F}$ is the set of all quadratic functions parametrized by a matrix $V \in \mathbb{R}^{d\times d}$,s.t.$F_V(x) = x^\top Vx$. Furthermore, let us assume the distribution of the input data x has standard Gaussian distribution. Remember, the W-GAN loss in the limit of infinite training data has the form:

$$\min_{W\in\mathbb{R}^d} \max_{V\in\mathbb{R}^d} \mathbb{E}_{x\sim\mathcal{N}(0,I_d)} F_V(x) - \mathbb{E}_{x'\sim\mathcal{P}_W} F_V(x')$$

where $\mathcal{P}_W$ is the pushforward of the standard Gaussian through G_W, i.e. it is the distribution of $G_W(z)$ for $z \sim \mathcal{N}(0, I_d)$. For notational convenience, we will denote the loss we are min-maxing

$$L(V, W) := \mathbb{E}_{x\sim\mathcal{N}(0,I_d)} F_V(x) - \mathbb{E}_{x'\sim\mathcal{P}_W} F_V(x')$$

(a) [8 pts] Show that the loss has the form

$$L(V, W) = \text{Tr}(V(I - WW^\top))$$

Hint: The trace is cyclically invariant, that is Tr(ABC) = Tr(BCA) = Tr(CAB)

(b) [6 pts] Show that if $WW^\top$ is such that $WW^\top \neq I$, then there exists a discriminator V, s.t. $L(V, W) \neq 0$. Conversely, if $WW^\top = I$, show that for any V, $L(V, W) = 0$. Conclude that any W achieving $\max_V L(V, W) = 0$ corresponds to recovering a generator G_W, s.t. the pushforward of the Gaussian through G_W is a standard Gaussian (i.e. the input distribution).

(c) [5 pts] Consider the case of $d = 1$. Then, V, W are scalars, and the loss is

$$L(v, w) = v(1 - w^2)$$

Suppose that we further restrict the generator/discriminator, s.t. $|w| \leq C$ and $|v| \leq 1$, for some constant $C > 1$. For a fixed $w, |w| \leq C$, what is the "best response " v, that is $\text{argmax}_v v(1 - w^2)$? Justify your answer.

(d) [4 pts] In the same setting as in (c), for a fixed $v, |v| \leq 1$, what is the "best response " w, that is $\text{argmin}_w v(1 - w^2)$? Justify your answer.

(e) [6 pts] In the same setting in (c), (d), show that if we initialize the generator with $w = 0$, the alternating the "best response dynamics" in (c) and (d) will not converge, but rather produce an infinite cycle in which v alternates between 1 and -1, and w alternates between 0 and c.

3. (Discriminator metrics for exponential families) In this problem, we will see more properties of discriminator metrics, and in particular, we will see how a "small family" of discriminators can be constructed for exponential families. (Recall, in class we saw this for invertible neural networks.)

   We mentioned in passing in class that

   $$d_{\mathcal{F}}(p, q) = \sup_{f \in \mathcal{F}} \left| \mathbb{E}_p[f(x)] - \mathbb{E}_q[f(x)] \right| \tag{A.1}$$

   is a "distance metric" – though we did not precisely define what a metric is.

   A *metric* is a function which maps pairs of distributions to $\mathbb{R}$ and has the following properties:

   - non-negativity: $d(p, q) \geq 0 \ \forall p, q$
   - symmetry: $d(p, q) = d(q, p) \ \forall p, q$
   - subadditivity: $d(p, q) \leq d(p, r) + d(r, q) \ \forall p, q, r$
   - $d(p, q) = 0 \iff p = q$

   (a) [6 pts] Prove that KL divergence is not a metric.

(b) [4 pts] Give an example of a set of functions $\mathcal{F}$ such that $d_{\mathcal{F}}$ is not a metric.

(c) [8 pts] Suppose the class $\mathcal{F}$ is sufficiently expressive such that $d_{\mathcal{F}}(p, q) = 0 \iff p = q$. Show that $d_{\mathcal{F}}$ is now a metric.

Finally, we move on to demonstrating that we can design a "small" family of discriminators when the generator lies in an *exponential family*.

This will parallel what we saw in class that if the family of distributions $\mathcal{G}$ is those generated by an invertible neural network, then for any $p_\theta \in \mathcal{G}$ a "similarly sized" neural network can evaluate $\log p_\theta$. If $\mathcal{F}$ contains all such neural networks, it's easy to construct the function f for any p, q to show that $d_{\mathcal{F}}(p, q) \geq \mathrm{KL}(p||q) + \mathrm{KL}(q||p)$. In the remaining questions, we will consider the case where $\mathcal{G}$ is instead an exponential family over $\mathbb{R}^d$.

As in Problem 1, an *exponential family* is a parameterized set of distributions $\{p_\theta \mid \theta \in \Theta \subseteq \mathbb{R}^k\}$ over $\mathbb{R}^d$ whose density can be written as

$$p_\theta(x) = \exp\{\theta^T T(x) - A(\theta)\}$$

for a fixed function $T(x) : \mathbb{R}^d \to \mathbb{R}^k$.

$A(\theta)$ is the log-partition function which ensures the distribution integrates to 1. Observe that the log-partition function is uniquely defined by θ:

$$A(\theta) = \log\left(\int_{\mathbb{R}^d} \exp\{\theta^T T(x)\}\, dx\right).$$

A key property of the log-partition function is that it can be used to express the means of the sufficient statistics of the given distribution.

(d) [6 pts] Prove that $\frac{dA(\theta)}{d\theta} = \mathbb{E}_{p_\theta}[T(x)]$.

The *Bregman divergence* is a notion of distance between two functions, defined in terms of a "measuring" function F. For such a function F, the Bregman divergence is written as

$$B_F(q, p) = F(q) - F(p) - \langle \nabla F(p), q - p \rangle.$$

This generalizes several other distance functions, including KL divergence.

(e) [8 pts] Let $p = p_{\theta_1}, q = q_{\theta_2}$ be arbitrary distributions from the same exponential family. Prove that

$$\mathrm{KL}(p_{\theta_1} || q_{\theta_2}) = B_F(\theta_2, \theta_1),$$

where $F = A$ is the log-partition function.

With these results, we are now ready to move on to the final question of relating the KL divergence to the $d_{\mathcal{F}}$ distance for exponential families. (In class, we did this for the Jensen-Shannon distance instead, when talking about invertible nets.)

For a given exponential family parameterized by Θ, define $\mathcal{F}$ as the set of all norm-bounded linear functionals over the sufficient statistics of p_θ:

$$\mathcal{F} = \{x \to \langle v, T(x) \rangle : \|v\|_2 \leq 1\}.$$

(f) [10 pts] Assume that the log-partition function satisfies $\gamma I \preceq \nabla^2 A(\theta) \preceq \beta I$, for $0 < \gamma \leq \beta$. Let $\theta_1, \theta_2 \in \Theta$, and write $p = p_{\theta_1},\ q = q_{\theta_2}$. Prove that

$$\frac{\gamma}{\sqrt{\beta}}\sqrt{\mathrm{KL}(p||q)} \leq d_{\mathcal{F}}(p, q) \leq \frac{\beta}{\sqrt{\gamma}}\sqrt{\mathrm{KL}(p||q)}.$$

Hint 1: You should start by working towards expressing both distance measures in terms of the derivatives of the log-partition function.
Hint 2: Recall the variational form of the ℓ_2 norm of a vector: $\|x\|_2 = \sup_{\|v\|_2 \leq 1} v^T x$
Hint 3: The Bregman divergence can also be written

$$D_F(\theta_2, \theta_1) = \int_0^1 \rho^T \nabla^2 F(\theta_2 + t\rho)\rho \, dt$$

where $\rho = \theta_1 - \theta_2$.

A.1 Collaboration Policy

After you have completed all other components of this assignment, report your answers to the collaboration policy questions detailed in the Academic Integrity Policies for this course.

1. Did you receive any help whatsoever from anyone in solving this assignment? If so, include full details including names of people who helped you and the exact nature of help you received.

2. Did you give any help whatsoever to anyone in solving this assignment? If so, include full details including names of people you helped and the exact nature of help you offered.

3. Did you find or come across code that implements any part of this assignment? If so, include full details including the source of the code and how you used it in the assignment.