TTIC 31230 Fundamentals of Deep Learning
Problems For Fundamental Equations.

In all problems we assume that all probability distributions $P(x)$ are discrete so that we have $\sum_x \; P(x) = 1$.

Problem 1: The problem of population density estimation is defined by the following equation.

$$\Phi^* = \underset{\Phi}{\mathrm{argmin}} \;\; H(\mathrm{Pop}, P_\Phi) = E_{x\sim\mathrm{Pop}} \;\; -\log \; P_\Phi(x)$$

This equation is used for language modeling — estimating the probability distribution over the population of English sentences that appear, say, in the New York Times. Show the following.

$$\Phi^* = \underset{\Phi}{\mathrm{argmin}} \;\; H(\mathrm{Pop}, P_\Phi) = \underset{\Phi}{\mathrm{argmin}} \;\; KL(\mathrm{Pop}, P_\Phi)$$

Assuming that the model probability $P_\Phi(x)$ can be computed for any given x, but that wehave no way of computing $\mathrm{Pop}(x)$ for a given x, explain why gradient descent on the cross-entropy objective can be done while gradient descent on the KL-divergence form is problematic.

Problem 2: Consider the objective

$$P^* = \underset{P}{\mathrm{argmin}} \;\; H(P, Q) \tag{1}$$

Define x^* by

$$x^* = \underset{x}{\mathrm{argmax}} \;\; Q(x)$$

Let δ_x be the distribution such that $\delta_x(x) = 1$ and $\delta_x(x') = 0$ for $x' \neq x$. Show that δ_{x^*} minimizes (**??**).
Next consider

$$P^* = \underset{P}{\mathrm{argmin}} \;\; KL(P, Q) \tag{2}$$

Show that Q is the minimizer of (**??**).
Next consider a subset S of the possible values and let Q_S be the restriction of Q to the set S.

$$Q_S(x) \quad = \quad \frac{1}{Q(S)} \begin{cases} Q(x) & \text{for } x \in S \\ 0 & \text{otherwise} \end{cases}$$

Show that that $KL(Q_S, Q) = -\ln Q(S)$, which will be quite small if S covers much of the mass. Show that, in contrast, $KL(Q, Q_S)$ is infinite unless $Q_S = Q$.
When we optimize a model P_Φ under the objective $KL(P_\Phi, Q)$ we can get that P_Φ covers only one high probability region (a mode) of Q (a problem called mode

collapse) while optimizing P_Φ under the objective $KL(Q, P_\Phi)$ we will tend to get that P_Φ covers all of Q. The two directions are very different even though both are minimized at $P = Q$.

Problem 3: Consider a joint distribution $P(x, y)$ on discrete random variables x and y. We define the marginal distributions $P(x)$ and $P(y)$ as follows.

$$\begin{aligned} P(x) &= \sum_y P(x,y) \\ P(y) &= \sum_x P(x,y) \end{aligned}$$

Let $Q(x, y)$ be defined to be the product of marginals.

$$Q(x,y) = P(x)P(y).$$

We define conditional entropy $H(y|x)$ as follows

$$H(y|x) = E_{x,y}\; - \log P(y|x).$$

Derive the following equalities.

$$KL(P(x,y), Q(x,y)) = H(y) - H(y|x) = H(x) - H(x|y)$$

The above quantity is called the mutual information between x and y, written $I(x, y)$. Explain why this quantity is always non-negative.

Problem 4: For three distributions P, Q and G show the following equality.

$$KL(P,Q) = \left(E_{x\sim P}\; \log \frac{G(x)}{Q(x)}\right) + KL(P,G)$$

Show that this implies

$$KL(P,Q) = \sup_G\; E_{x\sim P}\; \log \frac{G(x)}{Q(x)}$$

Next define

$$\begin{aligned} G(x) &= \frac{1}{Z}\, Q(x)e^{s(x)} \\ Z &= \sum_x Q(x)e^{s(x)} \end{aligned}$$

Show that a distribution $G(x)$ which does not assign zero to any point can be represented by a score $s(x)$ and that under this change of variables we have

$$KL(P,Q) = \sup_s\; E_{x\sim P}\; s(x) - \log E_{x\sim Q}\; e^{s(x)}$$

This is the Donsker-Varadhan variational representation of KL-divergence. This can be used in cases where we can sample from P and Q but cannot compute $P(x)$ or $Q(x)$. Instead we can use a model score $s_\Phi(x)$ where $s_\Phi(x)$ can be computed.

Problem 3.

a) Calculate the differential entropy of a Gaussian distribution

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-x^2}{2\sigma^2}}.$$

Use the natural logarithm in your definition of entropy.

b) Let the "signal" x be a Gaussian random variable with variance σ_x and let the "noise" ϵ be an independent Gaussian random variable with variance σ_ϵ. Let $z = x + \epsilon$. Use the fact that a sum of independent Gaussians is Gaussian with $\sigma_z^2 = \sigma_x^2 + \sigma_\epsilon^2$ to compute the differential mutual information $I(x, z)$. Express your answer in terms of the signal to noise ratio $\sigma_x^2/\sigma_\epsilon^2$. Hint: select a convenient expression for mutual information and use part (a).

c) For both the differential entropy in (a) and the mutual information in (b) say whether the numerical value depends on the choice of units.

Problem: Calculate the KL divergence between Gaussians

Problem: Calculate mutual information between gaussian x and x plus noise.

Problem: Prove the data processing inequality.

Problem: Show the various definitions of mutual inforamtion are equivalent.