TTIC 31230 Fundamentals of Deep Learning
Problems For Fundamental Equations.

Assume that probability distributions $P(y)$ are discrete with $\sum_y \; P(y) = 1$.

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

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

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.

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

Assuming that the model probability $P_\Phi(y)$ can be computed for any given y, but that we have no way of computing $\mathrm{Pop}(y)$ for a given y, 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

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

Define y^* by

$$y^* = \operatorname*{argmax}_y \; Q(y)$$

Let δ_y be the distribution such that $\delta_y(y) = 1$ and $\delta_y(y') = 0$ for $y' \neq y$. Show that δ_{y^*} minimizes (1).
Next consider

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

Show that Q is the minimizer of (2).
Next consider a subset S of the possible values and let Q_S be the restriction of Q to the set S.

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

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 5. Prove the data processing inequality that for any function f with $z = f(y)$ we have $H(z) \leq H(y)$.

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 mutual information by

$$I(x, y) = KL(P, Q)$$

which I will write as

$$I(x, y) = KL(P(x, y), Q(x, y))$$

We define conditional entropy $H(y|x)$ by

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

(a) Show

$$I(x, y) = H(y) - H(y|x) = H(x) - H(x|y)$$

(b) Explain why (a) implies $H(x) \geq H(x|y)$.

(c) By stating (b) conditioned on z we have

$$H(x|z) \geq H(x|y, z).$$

Use this to show that the data process inequality applies to mutual information, i.e., that for $z = f(y)$ we have $I(x, z) \leq I(x, y)$.

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

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

(b) Show that this implies

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

(c) Now define

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

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

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

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 $P(y)$ or $Q(y)$. Instead we can use a model score $s_\Phi(y)$ where $s_\Phi(y)$ can be computed.