**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.

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

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.

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

Assuming that the model probability $P_\Phi(y)$ can be computed for any given $y$, but that wehave 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

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

Define $y^*$ by

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

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

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

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 $S$.

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

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 $P_\Phi$ under the objective $KL(P_\Phi, Q)$ we can get that $P_\Phi$ covers only one high probability region (a mode) of $Q$ (a problem called mode collapse) while optimizing $P_\Phi$ under the objective $KL(Q, P_\Phi)$ we will tend to

get that $P_\Phi$ 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 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

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

Show

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

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

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

Show that this implies

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

Next 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.

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

**Problem 6.** This problem uses the definition of $H(x|y)$ and $I(x,y)$ from problem 3. For $z = f(y)$ prove $P(x|z) = E_{y\sim P(y|z)} P(x|y)$. Use this fact and Jensen's inequality to prove $H(x|z) \geq H(x|y)$. Using the results of problem 3, conclude that the mutual information $I(x,z) \leq I(x,y)$.