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.

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

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.

(a) Show the following.

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

Solution:

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

Since $H(\mathrm{Pop})$ does not depend on Φ the minima are the same.

(b) Explain why we can measure $H(\mathrm{Pop}, Q_\Phi)$ but cannot measure $KL(\mathrm{Pop}, Q_\Phi)$ for the structured object unconditional case (language modeling) and for the the conditional (labeling) case (imagenet).

Solution: We assume that the model is such that $Q_\Phi(y)$ can be computed. For example, an auto-regressive language model allows us to compute $Q_\Phi(y)$ for a sentence y as a product of next-word probabilities.
Assuming $Q_\Phi(y)$ can be computed, we can compute (a good approximation to) $E_{y\sim \mathrm{Pop}} \; -\ln Q_\Phi(y)$ by sampling sentences $y_1, \ldots\; y_n$ from Pop and computing

$$\hat{H}(\mathrm{Pop}, Q_\Phi) = \frac{1}{N}\sum_i -\ln Q_\Phi(y_i).$$

The confidence interval for this estimate shrinks as $1/\sqrt{N}$.

However, in the case of strcutured objects, such as sentences, while we can sample from Pop, we cannot compute $\mathrm{Pop}(y)$. Therefore we have no way of computing or even approximating, $H(\mathrm{Pop})$. So we cannot compute

$$KL(\mathrm{Pop}, Q_\Phi) = H(\mathrm{Pop}, Q_\Phi) - H(\mathrm{Pop}).$$

For the conditional case we have

$$
\begin{aligned}
KL(\mathrm{Pop}(y|x), Q_\Phi(y|x)) &= E_{x,y\sim \mathrm{Pop}}\ \ln \frac{\mathrm{Pop}(y|x)}{Q_\Phi(y|x)} \\
H(\mathrm{Pop}(y|x), Q_\Phi(y|x)) &= E_{x,y\sim \mathrm{Pop}}\ -\ln Q_\Phi(y|x)
\end{aligned}
$$

We assume that $Q_\Phi(y|x)$ can be computed and that allows $H(\mathrm{Pop}(y|x), Q_\Phi(y|x))$ to be computed (to a good approximation) by taking the average of a sample. However, we cannot compute $\mathrm{Pop}(y|x)$, even for binary classification, because (in most applications) we will never sample the same x twice.

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 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).

Solution: Consider an arbitrary distribution P. We must show that $H(P, Q) \geq H(\delta_{y^*}, Q)$.

$$
\begin{aligned}
Q(y) &\leq Q(y^*) \\
-\ln Q(y) &\geq -\ln Q(y^*) \\
E_{y\sim P} - \ln Q(y) &\geq -\ln Q(y^*) \\
H(P, Q) &\geq -\ln Q(y^*) = H(\delta_{y^*}, Q)
\end{aligned}
$$

Next consider

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

Show that Q is the minimizer of (2).

Solution: This follows from

$$
\begin{aligned}
KL(P, P) &= E_{y\sim P} \ln \frac{P(x)}{P(x)} = 0 \\
KL(P, Q) &\geq 0
\end{aligned}
$$

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.

Solution:

$$
\begin{aligned}
KL(Q_S, Q) &= E_{y \sim Q_S} \ \ln \frac{Q_S(y)}{Q(y)} \\
&= E_{y \sim Q_S} \ \ln \frac{Q(y)/Q(S)}{Q(y)} \\
&= E_{y \sim Q_S} \ -\ln Q(S) \\
&= -\ln Q(S)
\end{aligned}
$$

Show that, in contrast, $KL(Q, Q_S)$ is infinite unless S covers all values with non-zero propability.

Solution: If there exists a value $\tilde{y}$ not in S with $P(\tilde{y}) > 0$ then

$$E_{y \sim P} \ -\ln P_S(y) \geq P(\tilde{y}) - \ln 0 = \infty$$

When we optimize a model Q_Φ under the objective $KL(Q_\Phi, Q)$ we can get that Q_Φ covers only one high probability region (a mode) of Q (a problem called mode collapse) while optimizing Q_Φ under the objective $KL(Q, Q_\Phi)$ we will tend to get that Q_Φ covers all of Q. The two directions are very different even though both are minimized at $P = Q$.

Problem 3. Prove the data processing inequality that for any function f with $z = f(y)$ we have $H(z) \leq H(y)$.

Warning: This data processing inequality does not apply to contiuous densities. A function on a continuous density can either expand or shrink the distribution which increases or decrease its differential entropy respectively.

Solution:

$$
\begin{aligned}
H(y, z) &= H(y) + H(z|y) = H(y) \\
&= H(z) + H(y|z)
\end{aligned}
$$

The result now follows from the fact that $H(y|z) \geq 0$

Problem 4: 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

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

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

Solution:

$$\begin{aligned}
I(x,y) &= E_{x,y\sim P(x,y)} \;\ln \frac{P(x,y)}{P(x)P(y)} \\
&= E_{x,y\sim P(x,y)} \;\ln \frac{P(x)P(y|x)}{P(x)P(y)} \\
&= E_{x,y\sim P(x,y)} \;\ln \frac{P(y|x)}{P(y)} \\
&= \left(E_{y\sim P(y)} \;\; -\ln P(y)\right) - \left(E_{x,y\sim P(x,y)} \;\; -\ln P(y|x)\right) \\
&= H(y) - H(y|x)
\end{aligned}$$

The other equality is similar.

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

Solution: This is because the information $I(x,y)$ is a KL divergence which is always non-negative.

(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 $I(x,z) \leq I(x,y)$.

Warning: This data processing equality does not apply to continuous density functions.

Solution: We first note that for discrete distributions where z is a function of y we have $P(x|y,z) = P(x|y)$ which implies that $H(x|y,z) = H(x|y)$. so the above inequality can be written as

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

The result then follows from

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

and

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

Problem 5: (a) For three distributions P, Q and G show the following equality.

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

Solution:

$$\begin{aligned}
KL(P,Q) &= E_{y\sim P}\ \ln\frac{P(y)}{Q(y)} \\
&= E_{y\sim P}\ \ln\frac{P(y)G(y)}{Q(y)G(y)} \\
&= \left(E_{y\sim P}\ \ln\frac{G(y)}{Q(y)}\right) + \left(E_{y\sim P}\ \ln\frac{P(y)}{G(y)}\right) \\
&= \left(E_{y\sim P}\ \ln\frac{G(y)}{Q(y)}\right) + \left(E_{y\sim P}\ \ln\frac{P(y)}{G(y)}\right) \\
&= \left(E_{y\sim P}\ \ln\frac{G(y)}{Q(y)}\right) + KL(P,G)
\end{aligned}$$

(b) Show that this implies

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

Solution: Part (a) implies that

$$KL(P,Q) \leq\ E_{y\sim P}\ \ln\frac{G(y)}{Q(y)}$$

and also implies that for $G = Q$ we have equality.

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

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

Solution: Given any G which does not assign zero probability to any point we can take $s(y) = \ln \frac{G(y)}{Q(y)}$ which gives $Z = 1$ and satisfies the above equation. Plugging this expression for G into part (b) gives the result.

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.