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: Joint Entropy and Conditional Entropy We define conditional entropy $H(y|x)$ as follows

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

Given a distribution $P(x, y)$ show

$$H(P) = H(x) + H(y|x).$$

Solution:

$$
\begin{aligned}
H(P) &= E_{(x,y)\sim P} \; -\ln P(x, y) \\[2ex]
&= E_{(x,y)\sim P} \; -\ln P(x)P(y|x) \\[2ex]
&= E_{(x,y)\sim P} \; (-\ln P(x) - \ln P(y|x)) \\[2ex]
&= \left(E_{(x,y)\sim P} \; -\ln P(x)\right) + \left(E_{(x,y)\sim P} \; -\ln P(y|x)\right) \\[2ex]
&= H(x) + H(y|x)
\end{aligned}
$$

Problem 2: Unmeasurability of KL divergence and Population Entropy The problem of population density estimation is defined by the following equation.

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

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.

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

Solution:

$$\underset{\Phi}{\operatorname{argmin}} \; KL(\text{Pop}, Q_\Phi) = \underset{\Phi}{\operatorname{argmin}} \; H(\text{Pop}, Q_\Phi) - H(\text{Pop})$$

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

(b) Explain why we can measure $H(\text{Pop}, Q_\Phi)$ but cannot measure $KL(\text{Pop}, Q_\Phi)$ 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) $E_{y \sim \text{Pop}} \ -\ln Q_\Phi(y)$ by sampling sentences $y_1, \ldots y_n$ from Pop and computing

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

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 $\text{Pop}(y)$. Therefore we have no way of computing or even approximating, $H(\text{Pop})$. So we cannot compute

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

For the conditional case we have

$$KL(\text{Pop}(y|x), Q_\Phi(y|x)) \quad = \quad E_{x,y \sim \text{Pop}} \ \ln \frac{\text{Pop}(y|x)}{Q_\Phi(y|x)}$$

$$H(\text{Pop}(y|x), Q_\Phi(y|x)) \quad = \quad E_{x,y \sim \text{Pop}} \ -\ln Q_\Phi(y|x)$$

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

Problem 3: Asymmetry of cross entropy and KL-divergene Consider the objective

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

Define y^* by

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

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 (**??**).

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

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

Show that Q is the minimizer of (**??**).

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.

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 4. Data Processing Inequality 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 5: Mutual Information 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 $I(x,z) \leq I(x,y)$.

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 6: The ELBO We consider a model distribution $Q_\Phi(z,y)$ with marginal distribution

$$Q_\Phi(y) = \sum_z Q_\Phi(z,y).$$

We are interested in minimizing the unconditional (or unsupervised) cross-entropy of this model.

$$\Phi^* = \operatorname*{argmin}_{\Phi} E_{y \sim \text{Train}} - \ln Q_\Phi(y)$$

For many models of interest $Q_\Phi(z, y)$ can be efficiently computed as $Q_\Phi(z)Q_\Phi(y|z)$ but $Q_\Phi(y)$ is intractable to compute. In a variational auto-encoder we train a second model $\tilde{Q}_\Psi(z|y)$ and use the following inequality

$$
\begin{aligned}
\ln Q_\Phi(y) &\geq \text{ELBO} \\
&= E_{z \sim \tilde{Q}(z|y)} \ \ln \frac{Q_\Phi(z, y)}{\tilde{Q}_\Psi(z|y)}
\end{aligned}
$$

Rather than minimize the cross entropy we can maximize the ELBO (the Evidence Lower BOund) which corresponds to minimizing an upper bound on the cross entropy. Maximization of the ELBO with respect to model parameters Φ and Ψ define a variational auto encoder (VAE). We will consider this in much more detail later in the class. For now we just consider the formal equations.

a. The ELBO can be written as

$$\text{ELBO} = E_{z \sim \tilde{Q}(z|y)} \ \ln \frac{Q_\Phi(y)Q_\Phi(z|y)}{\tilde{Q}_\Psi(z|y)}.$$

Here we have that the ELBO is the expectation of a log of a product of three terms. Separate all three terms and express the terms other than $\ln Q_\Phi(y)$ as entropies or cross entropies.

Solution:

$$
\begin{aligned}
ELBO &= E_{z \sim \tilde{Q}_\Psi(z|y)} \ \ln \frac{Q_\Phi(y)Q_\Phi(z|y)}{\tilde{Q}_\Psi(z|y)} \\
&= \left(E_{z \sim \tilde{Q}_\Psi(z|y)} \ \ln Q_\Phi(y) \right) + \left(E_{z \sim \tilde{Q}_\Psi(z|y)} \ \ln Q_\Phi(z|y) \right) + \left(E_{z \sim \tilde{Q}_\Psi(z|y)} \ \ln \frac{1}{\tilde{Q}_\Psi(z|y)} \right) \\
&= \ln Q_\Phi(y) - H(\tilde{Q}_\Psi(z|y), Q_\Phi(z|y)) + H(\tilde{Q}(z|y))
\end{aligned}
$$

b. Now rewrite the ELBO by separating it into one the term for $P_\Phi(y)$ and one term for the other two combined and write the combined term as a KL divergence. Explain why your expression implies that the ELBO is a lower bound on $\ln Q_\Phi(y)$.

Solution:

$$
\begin{aligned}
ELBO &= E_{z \sim \tilde{Q}_\Psi(z|y)} \ln \frac{Q_\Phi(y) Q_\Phi(z|y)}{\tilde{Q}_\Psi(z|y)} \\
&= \left(E_{z \sim \tilde{Q}_\Psi(z|y)} \ln Q_\Phi(y) \right) + \left(E_{z \sim \tilde{Q}_\Psi(z|y)} \ln \frac{Q_\Phi(z|y)}{\tilde{Q}_\Psi(z|y)} \right) \\
&= \ln Q_\Phi(y) - KL(\tilde{Q}_\Psi(z|y), Q_\Phi(z|y))
\end{aligned}
$$

The lower bound property follows from the fact that KL divergence is non-negative.

Problem 7: The Donsker-Varadhan Bound (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

$$G(y) = \frac{1}{Z} Q(y)e^{s(y)}$$

$$Z = \sum_y Q(y)e^{s(y)}$$

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.