TTIC 31230 Fundamentals of Deep Learning
Problems For Information Theory

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

Problem 1. This problem is on softmax, entropy and cross entropy. Let the variable i range over the non-negative integers (0 to ∞). Consider the following softmax distribution where $\beta > 0$ is an inverse temperature parameter.

$$P(i) = \underset{i}{\text{softmax}}\left[-\beta i\right]$$

(a) Give an expression for $P(i)$ in terms of i without using a softmax or infinite sum. You can use the equality that for $0 < \lambda < 1$ we have

$$\sum_{n=0}^{\infty} \lambda^n = \frac{1}{1-\lambda}$$

(b) Solve for the entropy of this distribution. You can use the equality that for $0 < \lambda < 1$ we have

$$\sum_{n=0}^{\infty} n\lambda^n = \frac{\lambda}{(1-\lambda)^2}$$

(c) Consider a one-hot population distribution defined by $\text{Pop}(k) = 1$ and $\text{Pop}(i) = 0$ for $i \neq k$. What is the cross-entropy $H(\text{Pop}, P)$ and KL-divergence $KL(\text{Pop}, P)$? What is the cross-entropy $H(P, \text{Pop})$ and the KL divergence $KL(P, \text{Pop})$?

Problem 2. 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

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

Problem 3. Unmeasurability of KL divergence and Population Entropy The problem of population density estimation is defined by the following equation.

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

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.

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

(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).

Problem 5. Asymmetry of cross entropy and KL-divergene 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 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 6. 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.

Problem 7. 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

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

(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 8. 20 pts Consider the distribution on non-negative integers given by

$$P(i) = \frac{1}{2^{i+1}}.$$

(a) Using $\sum_{i=0}^{\infty} ar^i = \frac{a}{1-r}$ show that $\sum_{i=0}^{\infty} \ P(i) = 1$.

(b) Using $\sum_{i=0}^{\infty} \ ir^i = \frac{r}{(1-r)^2}$ compute the numerical value of the entropy $H_2(P)$ for this distribution (with your answer in bits).

(c) Give a code word (a bit string) $c(i)$ for each non-negative integer i such that the code is prefix-free (no code word is a proper prefix of any other code word)

and such that expected code length $E_{i\sim P}|c(i)|$ equals the entropy in bits you calculated in part (b).

Problem 9. 30 pts Problem 2 was on converting probabilities to codes. This problem is on converting codes to probabilities. Consider any prefx-free code $c(i)$ for the non-negative integers i. Give a sampling procedure that either returns an integer i or fails to terminate and where the probability of returning i is $2^{-|c(i)|}$.

Problem 10. 30 pts Shannon's source coding theorem states that for any prefix-free code we have

$$E_{x\sim P}|c(x)| \geq H_2(P)$$

and for any P there exists a prefix-free code such that

$$E_{x\sim P}|c(x)| \leq H_2(P) + 1.$$

In this problem we will prove the second inequality. We consider the case of a countably infinite set where each element has nonzero probability and consider the following procedure for constructing a code.

> Enumerate the elements of $\mathcal{X}$ as x_1, x_2, x_3, ... in order of decreasing probability.
>
> Initialize the code to be empty (no x_i is asigned any code)
>
> For $i = 1, 2, 3, \ldots$ assign an unused code $c(x_i)$ to x_i such that $|c(x_i)| = \lceil -\log_2 P(x_i) \rceil$ and such that no preifx of that code word has been previously assigned.

Suppose we have defined code words $c(x_1)$, ..., $c(x_i)$ and are trying to find a code word for x_{i+1}.

(a) Explain why no unassigned code word of length $\lceil -\log_2 P(x_{i+1}) \rceil$ can be a prefix of any previously assigned code word.

(b) Explain why there must exist an unallocated code word $c(x_{i+1})$ satisfying the specified conditions. Hint: Show that the probability of non-termination for the procedure of problem 7 is nonzero.

problem 11.

Problem 12. 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.

b. Now rewrite the ELBO by separating it into one the term for $Q_\Phi(y)$ 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 $\ln Q_\Phi(y)$.

Problem 13. The Donsker-Varadhan Bound (a) For three distributions P, Q and G show the following equality.

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

(b) Show that this implies

$$KL(P, Q) = \sup_G E_{y\sim P} \ln \frac{G(y)}{Q(y)} \tag{3}$$

(c) Now define

$$G(y) = \frac{1}{Z} Q(y) e^{s(y)} \tag{4}$$

$$Z = \sum_y Q(y) e^{s(y)} \tag{5}$$

Show that if Q has full support (is nonzero everywhere) then any distribution G with full supprt can be represented by a score $s(y)$ satisfying (4) and that under this change of variables we have

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

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.