

**TTIC 31230 Fundamentals of Deep Learning**

**Exam 1: 10% of class grade**

In all problems we assume that all probability distributions  $P(x)$  are discrete so that we have  $\sum_x P(x) = 1$ .

**Problem 1 (25 pts):** 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:**

$$\begin{aligned} H(P) &= E_{(x,y) \sim P} - \ln P(x, y) \\ &= E_{(x,y) \sim P} - \ln P(x)P(y|x) \\ &= E_{(x,y) \sim P} (-\ln P(x) - \ln P(y|x)) \\ &= (E_{(x,y) \sim P} - \ln P(x)) + (E_{(x,y) \sim P} - \ln P(y|x)) \\ &= H(x) + H(y|x) \end{aligned}$$

**Problem 2 (25 pts)** 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.

$$Q(x, y) = P(x)P(y).$$

Derive the following equalities.

$$KL(P(x, y), Q(x, y)) = H(y) - H(y|x) = H(x) - H(x|y)$$

The above quantity is called the mutual information between  $x$  and  $y$ , written  $I(x, y)$ . Explain why this quantity is always non-negative.

**Solution:**

$$\begin{aligned}
I(x, y) &= KL(P(x, y), Q(x, y)) \\
&= E_{(x, y) \sim P(x, y)} \ln \frac{P(x, y)}{P(x)P(y)} \\
&= E_{(x, y) \sim P(x, y)} \ln \frac{P(x)P(y|x)}{P(x)P(y)} \\
&= E_{(x, y) \sim P(x, y)} \ln \frac{P(y|x)}{P(y)} \\
&= E_{(x, y) \sim P(x, y)} (-\ln P(y) + \ln P(y|x)) \\
&= (E_{(x, y) \sim P(x, y)} (-\ln P(y))) - (E_{(x, y) \sim P(x, y)} -\ln P(y|x)) \\
&= H(y) - H(y|x)
\end{aligned}$$

The derivation of  $I(x, y) = H(x) - H(x|y)$  is similar.

$I(x, y)$  is non-negative because KL divergence is always non-negative.

**Problem 3 (25 pts):** Consider two (possibly unrelated) distributions  $P(z, x)$  and  $Q(z|x)$ .

(a) Show that for any specific value of  $x$  we have

$$E_{z \sim Q(z|x)} \ln P(z, x) = \ln P(x) - H(Q(z|x)) - KL(Q(z|x), P(z|x)).$$

Hint: Introduce a factor of  $1 = Q(z|x)/Q(z|x)$ .

**Solution:**

$$\begin{aligned}
E_{z \sim Q(z|x)} \ln P(z, x) &= E_{z \sim Q(z|x)} \ln \frac{P(z, x)Q(z|x)}{Q(z|x)} \\
&= E_{z \sim Q(z|x)} \ln \frac{P(x)P(z|x)Q(z|x)}{Q(z|x)} \\
&= E_{z \sim Q(z|x)} \left( \ln P(x) + \ln Q(z|x) + \ln \frac{P(z|x)}{Q(z|x)} \right) \\
&= (E_{z \sim Q(z|x)} \ln P(x)) + (E_{z \sim Q(z|x)} \ln Q(z|x)) + \left( E_{z \sim Q(z|x)} \ln \frac{P(z|x)}{Q(z|x)} \right) \\
&= \ln P(x) - H(Q(z|x)) - KL(Q(z|x), P(z|x))
\end{aligned}$$

(b) Explain why this implies

$$\ln P(x) \geq \left( E_{z \sim Q(z|x)} \ln P(z, x) \right) + H(Q(z|x))$$

**Solution:** This follows from the previous part and the the fact that KL-divergence is non-negative.

This last inequality is called the evidence lower bound (the ELBO). This terminology comes from viewing an observed variable  $x$  as evidence for a latent variable  $z$ . The ELBO is the core of expectation maximization (EM) and variational auto encoders (VAEs).

**Problem 4 (25 pts)** (a) For three distributions  $P$ ,  $Q$  and  $G$  show the following equality.

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

**Solution:**

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

(b) Explain why this implies

$$KL(P, Q) \geq E_{x \sim P} \log \frac{G(x)}{Q(x)}$$

**Solution:** This again follows from the fact that KL-divergence is non-negative

(c) Define

$$\begin{aligned} G(x) &= \frac{1}{Z} Q(x) e^{s(x)} \\ Z &= \sum_x Q(x) e^{s(x)} \end{aligned}$$

Show that this definition of  $G(x)$  gives

$$KL(P, Q) \geq E_{x \sim P} s(x) - \ln E_{x \sim Q} e^{s(x)}$$

**Solution:**

$$\begin{aligned} KL(P, Q) &\geq E_{x \sim P} \ln \frac{G(x)}{Q(x)} \\ &= E_{x \sim P} \ln \frac{Q(x)e^{s(x)}}{ZQ(x)} \\ &= E_{x \sim P} \ln \frac{e^{s(x)}}{Z} \\ &= E_{x \sim P} (s(x) - \ln Z) \\ &= (E_{x \sim P} s(x)) - (E_{x \sim P} \ln Z) \\ &= (E_{x \sim P} s(x)) - \ln Z \\ &= (E_{x \sim P} s(x)) - \ln \sum_x Q(x)e^{s(x)} \\ &= (E_{x \sim P} s(x)) - \ln E_{x \sim Q} e^{s(x)} \end{aligned}$$

This is the Donsker-Varadhan lower bound on KL-divergence.