

TTIC 31230 Fundamentals of Deep Learning

Problems for RDAs and VAEs

Problem 1. Mutual Information as Channel Capacity

The mutual information between two random variables x and y is defined by

$$I(x, y) = E_{x,y} \ln \frac{P(x, y)}{P(x)P(y)} = KL(P(x, y), P(x)P(y))$$

Mutual information has an interpretation as a channel capacity.

Suppose that we draw a random bit $y \in \{0, 1\}$ with $P(0) = P(1) = 1/2$ and send it across a noisy channel to a receiver who gets $y' = y \oplus \epsilon$ where ϵ is an independent “noise variable” with $\epsilon \in \{0, 1\}$, where \oplus is exclusive or (y gets flipped when $\epsilon = 1$), and where the “noise” ϵ has a probability P of being 1.

(a) Solve for the channel capacity $I(y, y')$ as a function of P in units of bits. When measured in bits, this channel capacity has units of bits received per message sent.

Solution:

$$\begin{aligned} I(y, y') &= H(y) - H(y|y') \\ H(y) &= 1 \text{ bit} \end{aligned}$$

$$\begin{aligned} H(y|y') &= P(y = y')(-\log_2 P(y = y')) + P(y \neq y')(-\log_2 P(y \neq y')) \\ &= P(\epsilon = 0)(-\log_2 P(\epsilon = 0)) + P(\epsilon = 1)(-\log_2 P(\epsilon = 1)) \\ &= (1 - P)\log_2 1/(1 - P) + P\log_2 1/P \\ &= H(P) \end{aligned}$$

(b) Explain why your answer to part (a) makes sense in terms of what the receiver knows for $P = 1/2$ and when $P = 1$.

Solution: For $P = 1/2$ we have $H(P) = 1$ bit and $I(y, y') = H(y) - H(P) = 0$ and the receiver knows nothing about y . For $P = 1$ we have $H(P) = 0$ and $I(y', y) = 1$ bit. Note that in this case y' is $1 - y$ so y' carries full information about y .

Problem 2. Rate-Distortion Autoencoders

(a) Consider an arbitrary distribution $P(z, y)$. Show the variational equation

$$I(y, z) = \inf_q E_{y \sim \text{pop}} KL(P_\Phi(z|y), Q(z))$$

where Q ranges over distributions on z . Hint: It suffices to show

$$I(y, z) \leq E_y KL(P_\Phi(z|y), Q(z))$$

and that there exists a Q achieving equality.

Solution:

$$\begin{aligned} I(y, z) &= E_{y \sim \text{pop}} KL(P(z|y), P(z)) \\ &= E_{y, z \sim P(z|y)} \left(\ln \frac{P(z|y)}{Q(z)} + \ln \frac{Q(z)}{P(z)} \right) \\ &= E_{y \sim \text{pop}} KL(P(z|y), Q(z)) + \left(E_{y \sim \text{pop}, z \sim P(z|y)} \ln \frac{Q(z)}{P(z)} \right) \\ &= E_y KL(P(z|y), Q(z)) + E_{z \sim P(z)} \ln \frac{Q(z)}{P(z)} \\ &= E_y KL(P(z|y), Q(z)) - KL(P(z), Q(z)) \\ &\leq E_{y \sim \text{pop}} KL(P(z|y), Q(z)) \end{aligned}$$

$P(z)$ not computable
 \Downarrow
 use $Q(z)$ or $P_\Phi(z)$
 model to get the
 upper bound

Equality is achieved when $Q(z) = P(z)$.

(b) Consider a rate-distortion autoencoder.

$$\Phi^* = \underset{\Phi}{\operatorname{argmin}} I_\Phi(y, z) + \lambda E_{y \sim \text{pop}, z \sim P_\Phi(z|y)} \text{Dist}(y, y_\Phi(z)).$$

Here $I_\Phi(y, z)$ is defined by the distribution where we draw y from pop and z from $P_\Phi(z|y)$. We will write $P_{\text{pop}}(z)$ for the marginal on z under this distribution.

$$P_{\text{pop}}(z) = E_{y \sim \text{Pop}} P_\Phi(z|y)$$

Based on the result from part (b) rewrite the above definition of rate-distortion autoencoder to be a minimization over three independent models $P_\Phi(z)$ and

$P_\Phi(y|z)$ and $P_\Phi(z|y)$ (although these models share parameters we will assume that Φ is sufficiently rich that the models are independently optimizable).

Solution:

$$\Phi^* = \underset{\Phi}{\operatorname{argmin}} E_{y \sim \text{pop}, z \sim P_\Phi(z|y)} \ln \frac{P_\Phi(z|y)}{P_\Phi(z)} + \lambda \operatorname{Dist}(y, y_\Phi(z)).$$

Problem 3. Modeling Rounding with Continuous Noise.

Consider a rate-distortion autoencoder with y and z continuous.

$$\Phi^* = \underset{\Phi}{\operatorname{argmin}} E_{y \sim \text{Pop}} KL(p_\Phi(z|y), p_\Phi(z)) + \lambda E_{y \sim \text{Pop}, z \sim P(z|y)} \operatorname{Dist}(y, y_\Phi(z)).$$

Define $p_\Phi(z|y)$ by $z = z_\Phi(y) + \epsilon$ with $z_\Phi(y) \in \mathbb{R}^d$ and ϵ drawn uniformly from $[0, 1]^d$. In other words, we add noise drawn uniformly from $[0, 1]$ to each component of $z_\Phi(y)$.

Define $p_\Phi(z)$ to be log-uniform in each dimension. More specifically $p_\Phi(z)$ is defined by drawing $s[i]$ uniformly from the interval $[0, s_{\max}]$ and then setting $z[i] = e^s$ so that $\ln z[i]$ is uniformly distributed over the interval $[0, s_{\max}]$. This gives

$$\begin{aligned} dz &= e^s ds = z ds \\ dp &= \frac{1}{s_{\max}} ds \\ p_\Phi(z[i]) &= \frac{dp}{dz} = \frac{1}{s_{\max} z[i]} \end{aligned} \quad \begin{array}{l} \text{"dy"} \\ \downarrow x \\ \text{probability density} \end{array}$$

Assume That we have that $z_\Phi(y) \in [1, e^{s_{\max}} - 1]^d$ so that with probability 1 over the draw of ϵ we have $\ln(z_\Phi(y) + \epsilon) \in [0, s_{\max}]$.

(a) For $z \in [z_\Phi(y), z_\Phi(y) + 1]$ what is $p_\Phi(z|y)$?

Solution: 1

(b) Solve for $KL(p_\Phi(z|y), p_\Phi(z))$ in terms of $z_\Phi(y)$ under the above specifications and simplify your answer for the case of $z_\Phi(y)[i] \gg 1$.

density is just normal $[z, z+1]$

Solution:

$$\begin{aligned}
& KL(p_{\Phi}(z|y), p_{\Phi}(z)) \\
&= E_{z \sim P_{\Phi}(z|y)} \ln \frac{p_{\Phi}(z_{\Phi}(y))}{p_{\Phi}(z)} \\
&= E_{z \sim P_{\Phi}(z|y)} \sum_i \ln \frac{1}{1/(s_{\max} z[i])} \\
&= \sum_i E_{z[i]} \ln(s_{\max} z[i]) \\
&= \left(\sum_i \int_{z_{\Phi}(y)[i]}^{z_{\Phi}(y)[i]+1} \ln z \, dz \right) + d \ln s_{\max} \\
&= \left(\sum_i [z \ln z - z]_{z_{\Phi}(y)[i]}^{z_{\Phi}(y)[i]+1} \right) + d \ln s_{\max} \\
&= \left(\sum_i [z \ln z]_{z_{\Phi}(y)[i]}^{z_{\Phi}(y)[i]+1} \right) + d \ln s_{\max} - 1 \\
&= \left(\sum_i \ln(z_{\Phi}(y)[i] + 1) + z_{\Phi}(y)[i] (\ln(z_{\Phi}(y)[i] + 1) - \ln z_{\Phi}(y)[i]) \right) + d \ln s_{\max} - 1 \\
&= \left(\sum_i \ln(z_{\Phi}(y)[i] + 1) + z_{\Phi}(y)[i] \ln \left(1 + \frac{1}{z_{\Phi}(y)[i]} \right) \right) + d \ln s_{\max} - 1 \\
&\approx \left(\sum_i \ln z_{\Phi}(y)[i] \right) + d \ln s_{\max} \quad \text{for } z_{\Phi}(y)[i] \gg 1
\end{aligned}$$