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. A Variational Upper Bound on Mutual Information

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

$$I(y,z) = \inf_Q \; E_{y\sim P(y)} KL(P(z|y), Q(z))$$

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

$$I(y,z) \leq \; E_y KL(P(z|y), Q(z))$$

and that there exists a Q achieving equality.

Solution:

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

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

(b) Consider a rate-distortion autoencoder.

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

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 $P_{\mathrm{pop}}(z)$ for the marginal on z under this distribution.

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

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:

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

Problem 3. Modeling Rounding with Continuous Noise.

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

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

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

$$\begin{aligned}
dz &= e^s ds = z ds \\
dp &= \frac{1}{s_{\max}}\, ds \\
p_\Phi(z[i]) &= \frac{dp}{dz} = \frac{1}{s_{\max} z[i]}
\end{aligned}$$

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 $z_\Phi(y)[i] >> 1$.

Solution:

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

Problem 4. Rounding RDA

We consider the following modification of RDAa

$$
\begin{aligned}
\text{RDA}: \;\; \Phi^* &= \operatorname*{argmin}_\Phi \; E_{y\sim\text{Pop},\, z\sim P_\Phi(z|y)} \; -\ln\frac{P_\Phi(z)}{P_\Phi(z|y)} + \lambda\text{Dist}(y, y_\Phi(z)) \\
\text{Rounding RDA}: \;\; \Phi^*, \Psi^* &= \operatorname*{argmin}_\Phi \; E_{y\sim\text{Pop},\, z:=\text{round}(z_\Psi(y))} \; -\ln P_\Phi(z) + \lambda\text{Dist}(y, y_\Phi(z))
\end{aligned}
$$

Here $\text{round}(z) \in \mathcal{Z}$ where $\mathcal{Z}$ is a discrete set of vectors defined independent of the choice of y. For example, rounding might map each real number in z to the nearest integer as was done in Balle et al. 2017. Or rounding might map the vector z to the nearest center vector resulting from K-means vector quantization as in VQ-VAE. Other roundings are possible. The Rounding RDA corresponds to practical image compression where $-\log_2 P_\Phi(\mathrm{round}(z_\Psi(y)))$ is (approximately) the number of bits in the compressed file.

(a) What is $\nabla_\Psi \ln P_\Phi(\mathrm{round}(z_\Psi(y))$? **Solution**: zero

(b) What is $\nabla_\Psi \mathrm{Dist}(y, y_\Phi(\mathrm{round}(z_\Psi(y))))$? **Solution**: zero

To optimize Ψ Balle et al. used two tricks. They replaced $P_\Phi(\mathrm{round}(z_\Psi(y)))$ with $p_\Phi(z_\Psi(y))$ where $p_\Phi(z)$ is a continuous density, and they replace the rounding operation with additive noise. Although rounding will be used for image compression, gradient descent is then done on

$$\Phi^*, \Psi^* = \operatorname*{argmin}_{\Phi,\Psi} E_{y,\epsilon} \; - \ln p_\Phi(z_\Psi(y)) + \lambda \mathrm{Dist}(y_\Phi(z_\Psi(y) + \epsilon))$$

To model rounding to the nearest integer we take each dimension of ϵ to be drawn uniformly over the interval $(-1/2, 1/2)$.

(c) The density $p_\Phi(\tilde{z})$ defines a discrete distribution on the discrete values $\tilde{z} \in Z$ defined by

$$P_\Phi(\tilde{z}) = P_{z \sim p_\Phi}(\mathrm{round}(z) = \tilde{z})$$

Consider the case where $\mathcal{Z}$ is the discrete set of vectors with integer coordinates. Assume that the density $p_\Phi(z)$ is locally approximated by its first order Taylor expansion

$$p_\Phi(z + \Delta z) = p_\Phi(z) + (\nabla_z p_\Phi(z))^\top \Delta z$$

Assuming the first order Taylor expansion is exact, give a closed-form expression for the discrete distribution $P_\Phi(\tilde{z})$ in terms of the continuous density $p_\Phi(z)$. Hint: write $P_\Phi(\tilde{z})$ as an expectation over ϵ drawn from the uniform distribution on $[-1/2, 1/2]^d$ where d is the dimension of z.

Solution: For an vector $\tilde{z}$ with integer coordinates we have

$$\begin{aligned}
P_\Phi(\tilde{z}) &= P_{z\sim p_\Phi}(\mathrm{round}(z) = \tilde{z}) \\
&= \int_{\epsilon \in [-1/2,1/2]^d} p_\Phi(\tilde{z} + \epsilon)\, d\epsilon \\
&= E_{\epsilon \sim \mathrm{uniform}[-1/2,1/2]^d}\; p_\Phi(\tilde{z} + \epsilon) \\
&= E_{\epsilon \sim \mathrm{uniform}[-1/2,1/2]^d}\; p_\Phi(\tilde{z}) + (\nabla_{\tilde{z}} p_\Phi(\tilde{z}))^\top \epsilon \\
&= p_\Phi(\tilde{z}) + E_{\epsilon \sim \mathrm{uniform}[-1/2,1/2]^d}\; (\nabla_{\tilde{z}} p_\Phi(\tilde{z}))^\top \epsilon \\
&= p_\Phi(\tilde{z}) + (\nabla_{\tilde{z}} p_\Phi(\tilde{z}))^\top E_{\epsilon \sim \mathrm{uniform}[-1/2,1/2]^d}\; \epsilon \\
&= p_\Phi(\tilde{z})
\end{aligned}$$

Problem 5. VQ-VAEs

In a VQ-VAE the rounding operation is parameterized by a tensor $C[K, I]$ giving K center vectors of the form $C[k, I]$. We now consider rounding-RDAs defined by the following objective.

$$\Phi^*, \Psi^*, C^* = \operatorname*{argmin}_{\Phi,\Psi,C} \; E_{y\sim\mathrm{Pop},\; \hat{L}:=\mathrm{round}_C(L_\Psi(y))} \; -\ln P_\Phi(\hat{L}) + \lambda \mathrm{Dist}(y, y_\Phi(\hat{L}))$$

In the VQ-VAE we are controlling the rate with the parameter K giving the number of clusters. In the optimization problem the prior term $P_\Phi(\hat{L})$ is being held as uniform over all $\hat{L}$ and can be ignored. Assuming L_2 distortion we are then left with

$$\Phi^*, \Psi^*, C^* = \operatorname*{argmin}_{\Psi,\Psi,C} E_y \frac{1}{2} ||y - y_\Phi(\mathrm{round}_C(L_\Psi(y)))||^2$$

This has well defined gradients for Φ and C but, because of rounding, not for Ψ. We are now trying to minimize the expected loss of the following forward calculation where $L[P, I]$ is a sequence of vectors.

$$\begin{array}{rcl}
y & \sim & \mathrm{Pop} \\
L & = & L_\Psi(y) \\
k[p] & = & \operatorname*{argmin}_k \; ||C[k, I] - L[p, I]|| \\
\hat{L}[p, I] & = & C[k[p], I] \\
\hat{y} & = & y_\Phi(\hat{L}) \\
\mathrm{Loss} & = & \frac{1}{2} ||y - \hat{y}||^2
\end{array}$$

The straight through gradient for a rounding operation is given by

$$L.\mathrm{grad} \mathrel{+}= \hat{L}.\mathrm{grad}$$

(a) 10 points. Give a for loop for computing $C[K, I]$.grad from $\hat{L}$.grad as defined by backpropagation on the above computation.

Solution:

$$\text{for } p \quad C[k[p], I].\mathrm{grad} \mathrel{+}= \hat{L}[p, I].\mathrm{grad}$$

(b) 15 points. The published formulation of VQ-VAE uses the following gradient updates.

$$\begin{array}{rcl}
L.\mathrm{grad} & += & \hat{L}.\mathrm{grad} \\
L.\mathrm{grad} & += & \beta(L - \hat{L}) \\
\text{for } p \;\; C[k[p], I].\mathrm{grad} & += & \tilde{\eta}(C[k[p], I] - L[p, I])
\end{array}$$

Actually, this has been modified from the published form to add a learning rate adjustment parameter $\tilde{\eta}$.

Give an additional loss term so that the published version is equivalent to taking the gradient of $C[K, I]$.grad from the new loss term only and $L[P, I]$.grad from both the straight-through gradient and the gradient of the new loss term.

Solution: The additional loss is

$$\frac{1}{2}\beta||L[P, I] - \hat{L}[P, I]||^2 = \sum_p \frac{1}{2}\beta||L[p, I] - C[k[p], I]||^2$$

(c) 15 points. Give a complete set of backpropagation updates defined by backpropagation on both loss terms and using straight-through backpropagation to $L[P, I]$.grad

Solution:

$$\begin{array}{rrcl}
 & L.\mathrm{grad} & \mathtt{+=} & \hat{L}.\mathrm{grad} \\
\text{for } p & C[k[p], I].\mathrm{grad} & \mathtt{+=} & \hat{L}[p, I].\mathrm{grad} \\
 & L.\mathrm{grad} & \mathtt{+=} & \beta(L - \hat{L}) \\
\text{for } p & C[k(t), I].\mathrm{grad} & \mathtt{+=} & \beta(C[k(t), I] - L[p, I])
\end{array}$$

Here any hyper-parameter for the learning rate for $C[K, I]$ must be handled elsewhere (in the optimizer).

(d) 10 points. We now have three versions of training — end-to-end with straight through as in part (a), the published version as in part (b), and the backpropagation on the both loss terms with straight-through as defined in part (c). For which of these three training algorithms is it true that at a stationary point $C[k, I]$ is mean of the vectors assigned to class k?

Solution: Of the three, this is only true for the published version.