TTIC 31230 Fundamentals of Deep Learning, winter 2020

Quiz 4

Problem 1. The Variational Upper Bound on Mutual Information (25 points)

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.

Problem 2. Rounding RDA (25 points)

We consider the following modification of RDAa

$$\text{RDA}: \; \Phi^* \quad = \quad \operatorname*{argmin}_{\Phi} \; E_{y \sim \text{Pop},\, z \sim P_\Phi(z|y)} \; - \ln \frac{P_\Phi(z)}{P_\Phi(z|y)} + \lambda \text{Dist}(y, y_\Phi(z))$$

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

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 $-\log_2 P_\Phi(\text{round}(z_\Phi(y)))$ is (approximately) the number of bits in the compressed file.

(a) What is $\nabla_\Psi \ln P_\Phi(\text{round}(z_\Psi(y))$?

(b) What is $\nabla_\Psi \text{Dist}(y, y_\Phi(\text{round}(z_\Psi(y))))$?

To optimize Ψ Balle et al. used two tricks. They replaced $P_\Phi(\text{round}(z_\Phi(y)))$ with $p_\Phi(z_\Phi(y))$ 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

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

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

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

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

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

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.

3. VQ-VAEs (50 points)

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.

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

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

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

This has well defined gradients for Φ and Θ 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.

$$\begin{aligned}
y &\sim \text{Pop} \\
L &= L_\Psi(y) \\
k[p] &= \underset{k}{\text{argmin}} \; ||C[k, I] - L[p, I]|| \\
\hat{L}[p, I] &= C[k[p], I] \\
\hat{y} &= y_\Phi(\hat{L}) \\
\text{Loss} &= \frac{1}{2} ||y - \hat{y}||^2
\end{aligned}$$

The straight through gradient for a rounding operation is given by

$$L.\text{grad} \texttt{ += } \hat{L}.\text{grad}$$

(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.

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

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

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.

(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

(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?