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.

Solution:

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

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

Problem 2. Rounding RDA (25 points)

We consider the following modification of RDAa

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

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

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

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

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

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

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

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

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

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

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 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.

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

The straight through gradient for a rounding operation is given by

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

(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.

Solution:

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

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

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

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:

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

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.