TTIC 31230 Fundamentals of Deep Learning
Quiz 1

Problem 1: Optimizing Cross Entropy. For this problem we consider a population distribution Pop on the non-negative natural numbers $k \geq 0$. We will work with the population mean $\mu = E_{k \sim \mathrm{Pop}}\ k$. We consider model distributions defined by the single parameter λ with $0 \leq \lambda < 1$ and defined by the distribution

$$Q_\lambda(k) \quad = \quad (1-\lambda)\lambda^k$$

(a) Given an expression for the cross-entropy $H(\mathrm{Pop}, Q_\lambda)$ in terms of μ and λ.

(b) Solve for the optimal value λ^* minimizing $H(\mathrm{Pop}, Q_\lambda)$ as a function of μ.

(c) Solve for mean value of the distribution Q_{λ^*} in terms of μ.

Problem 2. Maximum Mutual Information Training. Consider a population distribution Pop on pairs $\langle x, y\rangle$ and a model distribution $Q_\Phi(\hat{y}|x)$. Consider a distribution P_Φ on triples $x, y, \hat{y}$ where $\langle x, y\rangle$ is drawn from Pop and $\hat{y}$ is drawn from $Q_\Phi(\hat{y}|x)$. Under the distribution P_Φ the mutual information between y and $\hat{y}$ is defined by

$$I_\Phi(y,\hat{y}) \quad = \quad KL(P_\Phi(y,\hat{y}),\ \ \mathrm{Pop}(y)P_\Phi(\hat{y}))$$

$$\begin{aligned}
P_\Phi(y,\hat{y}) &= \sum_x \mathrm{Pop}(x)\ \mathrm{Pop}(y|x)\ Q_\Phi(\hat{y}|x) \\
&= E_{x\sim \mathrm{Pop}}\ \mathrm{Pop}(y|x) Q_\Phi(\hat{y}|x) \\
\mathrm{Pop}(y) &= E_{x\sim \mathrm{Pop}}\ \mathrm{Pop}(y|x) \\
P_\Phi(\hat{y}) &= E_{x\sim \mathrm{Pop}}\ Q_\Phi(y|x)
\end{aligned}$$

Here we are interested in comparing the fundamental cross entropy objective to the objective of maximizing the mutual information $P_\Phi(y,\hat{y})$.

$$\begin{aligned}
\Phi_1^* &= \operatorname*{argmin}_\Phi E_{\langle x, y\rangle \sim \mathrm{Pop}}\ -\ln Q_\Phi(y|x) \\
\Phi_2^* &= \operatorname*{argmax}_\Phi I_\Phi(y,\hat{y})
\end{aligned}$$

(a) Suppose that there exists a perfect predictor – a parameter setting Φ^* such that $P_\Phi(\hat{y}|x) = 1$ for $\hat{y} = y$ and zero otherwise. Show using an explicit calculation and standard information theoretic inequalities that a perfect predictor is an optimum of both the cross-entropy objective and the maximum mutual information objective.

(b) Consider binary classification where we have $y, \hat{y} \in \{-1, 1\}$. Show using an explicit calculation and standard information-theoretic inequalities that a perfect anti-predictor with $P_\Phi(hayy|x) = 1$ for $\hat{y} = -y$ is also optimal for the maximum mutual information objective.

Problem 3. Backpropagation for Layer Normalization. Layer normalization is an alternative to batch normalization and is used in the transformer to handle "covariate shift". In the transformer each a layer has positions in the text that I will index by t and neurons at each position that I will index by i. We can think of this as a sequence of vectors $L[t, I]$. Layer normalization is defined by the following equations where the vectors $A_{\ell+1}[I]$ and $B_{\ell+1}[I]$ are trained parameters and σ is an arbitrary activation function, typically ReLU.

$$\begin{aligned}
\mu_\ell &= \frac{1}{TI} \sum_{t,i} L_\ell[t,i] \\
\sigma_\ell &= \sqrt{\frac{1}{TI} \sum_{t,i} (L_\ell[t,i] - \mu_\ell)^2} \\
\tilde{L}_{\ell+1}[t,i] &= \sigma\left(\frac{A_{\ell+1}[i]}{\sigma_\ell}(L_\ell[t,i] - \mu_\ell) + B_{\ell+1}[i]\right)
\end{aligned}$$

Write backpropagation equations for these three assignments.