TTIC 31230 Fundamentals of Deep Learning, Autumn 2021

Exam 3

Problem 1: 25 pts. Consider a probability distribution on structured labels $\mathcal{Y}[N]$ where $\mathcal{Y}[n]$ is either -1, 0 or 1. Consider a score function $s(\mathcal{Y})$ defined by

$$s(\mathcal{Y}) = \left(\sum_{n=0}^{N-2} \mathcal{Y}[n]\ \mathcal{Y}[n+1]\right) + \mathcal{Y}[N-1]\mathcal{Y}[0]$$

We can think of this as a ring of edge potentials with no node potentials. We are interested in the probability defined by the exponential softmax

$$P_s(\mathcal{Y}) = \frac{1}{Z_s} e^{s(\mathcal{Y})}$$
$$Z_s = \sum_{\mathcal{Y}} e^{s(\mathcal{Y})}$$

(a) Given an expression for the negative log pseud-likelihood $-\ln \tilde{P}_s(\mathcal{Y})$ where $\mathcal{Y}$ is the constant assignment defined by $\mathcal{Y}[n] = 0$ for all n. Your expression should be a simple function of N.

(b) Repeat part (a) but for the constant structured label defined by $\mathcal{Y}[n] = 1$.

Problem 2. 25 pts This problem is on GAN language modeling. A GAN takes noise as input and transforms it to xsan output. We consider the case where the output is a string of symbols $w_1, \ldots, w_T$ where for simplicity we always generate a string of exactly length T and where the words are integers with $w_t \in \{0, \ldots, I-1\}$ where I is the size of the vocabulary. The GAN parameters are just the parameters of a bigram model, i.e., the parameters are probability tables

$$P[i] \quad = \quad P(w_0 = i)$$

$$Q[i,j] \quad = \quad P(w_{t+1} = j \mid w_t = i)$$

We take the noise input to the GAN to be a sequence of random real numbers $\epsilon_1, \ldots, \epsilon_T$ where each ϵ_t is drawn uniformly from the interval $[0, 1]$.

(a) Write a function $\hat{w}(P[I], \epsilon_1)$ which deterministically returns the first word given the noise value ϵ_1 such that the probability over the draw of ϵ_1 that $\hat{w}(P[I], \epsilon_1) = i$ is P_i.

(b) Write a function $\hat{w}(Q[I,I], w_t, \epsilon_t)$ which deterministically returns the word w_{t+1} given w_t such that the probability over the draw of ϵ_t that $\hat{w}(Q[I,I], w_t, \epsilon_t) = w_j$ is $Q[w_t, j]$.

(c) There is a problem with this GAN. For string generated by the GAN we need to back-propagate the discriminator loss into the GAN generator parameters. Explain why this is problematic. Is this always problematic when the generator output is discrete?

Problem 3. 25 pts This problem is on VAE language modeling (in contrast to GAN language modeling). Consider a VAE where the signal s is a word string $w_1, \ldots, w_T$ (as in problem 2). In the VAE we can have a continuous latent variable z. The VAE optimization problem is then

$$\Phi^*, \Theta^*, \Psi^* = \operatorname*{argmin}_{\Phi,\Theta,\Psi} \; E_{s\sim\mathrm{Pop},\; z\sim p_\Psi(z|s)} \; \ln \frac{p_\Psi(z|s)}{p_\Phi(z)} \; - \; \ln P_\Theta(s|z) \qquad (1)$$

Here the first "rate term" is defined on densities and the final "distortion term" is defined for a discrete sentence s. To explicitly handle the reparameterization trick will take the encoder density to be a Gaussian. For a Gaussian encoder we compute a mean vector $\hat{z}_\Psi(s)$ and a variance $\sigma^2_\Psi(s)[i]$ for each component $z[i]$ of z. The Gaussian density for the encoder is then.

$$p_\Psi(z[i]|s) \propto \exp(-(z[i] - \hat{z}_\Psi(s)[i])^2/(2\sigma^2_\Psi(s)[i])$$

(a) For a noise value $\epsilon \in \mathbb{R}$ drawn from $\mathcal{N}(0,1)$, and for given values $\hat{z} \in \mathbb{R}$ and $\sigma^2 \in \mathbb{R}$, define a deterministic function $z(\hat{z}, \sigma^2, \epsilon)$ such that over the draw of the noise ϵ we have that $z(\hat{z}, \sigma^2, \epsilon)$ has the density

$$p(z) \propto \exp(-(z - \hat{z})^2/(2\sigma^2)).$$

(b) Applying your solution to part (a) to the individual components of z equation (1) can be rewritten as

$$\Phi^*, \Theta^*, \Psi^* = \operatorname*{argmin}_{\Phi,\Theta,\Psi} \; E_{s\sim\mathrm{Pop},\; \epsilon\sim\mathcal{N}(0,I)} \; \ln \frac{p_\Psi(z|s)}{p_\Phi(z)} \; - \; \ln P_\Theta(s|z) \qquad (2)$$

Are there any problems with doing SGD on the optimization defined by (2) due to the use of continuous z and discrete s? Explain your answer.

(c) It can be shown that if we hold the encoder Ψ fixed then the optimal value of the prior density $p_\Phi(z)$ is just the marginal on z of the distribution defined by sampling $s \sim$ Pop and $z \sim p_\Psi(z|s)$. We can write this marginal as $p_{\mathrm{Pop},\Psi}(z)$. Now consider the rate term when $p_\Phi(z) = p_{\mathrm{Pop},\Psi}(z)$.

$$\mathrm{rate} = E_{s\sim\mathrm{Pop},\; z\sim P_\Psi(z|s)} \; \ln \frac{p_\Psi(z|s)}{p_{\mathrm{Pop},\Psi}(z)}$$

Write this rate term as a differential mutual information.

Problem 4. 25 pts This problem is on VAEs when both z and s are discrete. This happens in the second layer of a progressive VAE as defined in the slides. Is the discreteness of z an issue in this case? Explain your answer.