# TTIC 31230, Fundamentals of Deep Learning

David McAllester, Winter 2021

# Gaussian Models

# and The Parels of Differential Entropy

# Gaussian VAEs

$$\text{VAE:} \quad \text{pri}^*, \text{dec}^*, \text{enc}^* = \underset{\text{pri,dec,enc}}{\text{argmin}} \quad E_{y,z} \quad -\ln \frac{p_{\text{pri}}(z)p_{\text{dec}}(y|z)}{p_{\text{enc}}(z|y)}$$

All models are Gaussian densities.

$$p_{\text{pri}}(z[i]) \propto \exp((z[i] - \mu_{\text{pri}}[i])^2/2\sigma_{\text{pri}}^2[i])$$

$$p_{\text{enc}}(z[i]|y) \propto \exp((z[i] - \hat{z}_{\text{enc}}(y)[i])^2/2\sigma_{\text{enc}}^2(y)[i])$$

$$p_{\text{dec}}(y[i]|z) \propto \exp((y[i] - \hat{y}_{\text{dec}}(z)[i])^2/2\sigma_{\text{dec}}^2(z)[i])$$

# Differential Entropy

In the case of a continuous density (as opposed to a discrete probability) we have the notion of differential entropy.

For a density $p(x)$ on a real value $x$ we have

$$H(p) = E_{x \sim p} \left[ -\ln p(x) \right]$$

$$= \int_{-\infty}^{\infty} p(x) \left( -\ln p(x) \right) dx$$

# Differential Entropy can Diverge to $-\infty$

For a uniform distribution over an interval on the real line of width $\Delta$ we have

$$H = E_{x \sim p} \left[ - \ln p(x) \right]$$

$$= E_{x \sim p} \left[ - \ln \frac{1}{\Delta} \right]$$

$$= \ln \Delta$$

As $\Delta \to 0$ we have $H \to -\infty$

# Differential Entropy can Diverge to $-\infty$

$$
\begin{aligned}
H(\mathcal{N}(0, \sigma^2)) &= E_{x \sim \mathcal{N}(0, \sigma^2)} \left[ -\ln \left( \frac{1}{\sigma\sqrt{2\pi}} \exp \frac{-x^2}{2\sigma^2} \right) \right] \\[2mm]
&= E_{x \sim \mathcal{N}(0, \sigma^2)} \left[ \ln(\sigma\sqrt{2\pi}) + \frac{-x^2}{2\sigma^2} \right] \\[2mm]
&= (\ln \sigma) + \ln(\sqrt{2\pi}) + E_x \left[ \frac{x^2}{2\sigma^2} \right] \\[2mm]
&= (\ln \sigma) + \ln(\sqrt{2\pi}) + \frac{1}{2}
\end{aligned}
$$

As $\sigma \to 0$ we have $H \to -\infty$

# Sensitivity to the Choice of Units

$$H(N(0, \sigma)) = C + \ln \sigma$$

Differential entropy depends on the choice of units — a distributions on lengths will have a different entropy when measuring in inches than when measuring in feet.

# Differential Cross-Entropy can Diverge to $-\infty$

Consider the unsupervised training objective.

$$\text{pri}^*, \text{dec}^* = \underset{\text{pri},\text{dec}}{\text{argmin}}\ E_{y\sim\text{train}}\ -\ln p_{\text{pri},dec}(y)$$

The training set is finite (discrete).

For each $y$ the density $p_{\text{pri},dec}(y)$ can go to infinity.

This will drive the cross-entropy training loss to $-\infty$.

# Differential Entropy Can Be Considered Infinite

An actual real number carries an infinite number of bits.

Consider quantizing the real numbers into bins.

A continuous probability densisty $p$ assigns a probability $p(B)$ to each bin.

As the bin size decreases toward zero the entropy of the bin distribution increases toward $\infty$.

A meaningful convention is that $H(p) = +\infty$ for any continuous density $p$.

# Differential KL-divergence is Meaningful

$$KL(p, q) = \int \left( \ln \frac{p(x)}{q(x)} \right) p(x) dx$$

Unlike differential entropy, differential KL divergence is always non-negative (but can be infinite).

Note that $KL(p, p) = 0$ independent of $H(p)$.

# Mutual Information

$I(x, y)$ is the reduction in the number of bits we need to name $y$ as a result of observing $x$ (on average).

$$I(x, y) = E \ln \frac{P(x, y)}{P(x)P(y)}$$

$$= E \ln \frac{P(x, y)}{P(x)} - \ln P(y)$$

$$= H(y) - H(y|x)$$

Intuitively, how much does $x$ know about $y$?

# Differential Mutual Information

$$I(x, y) = KL(p(x, y), p(x)p(y))$$

$$= E_{x,y} \ln \frac{p(x, y)}{p(x)p(y)}$$

Mutual information is a KL divergence and hence differential mutual information is always non-negative.

# The Data Processing Inequality

For continuous $y$ and $z$ with $z = f(y)$ we get that $H(z)$ can be either larger or smaller than $H(y)$ (consider $z = ay$ for $a > 1$ vs. $a < 1$).

However, mutual information is a KL divergence and is more meaningful than entropy and for $z = f(y)$ we do have

$$I(x, z) \leq I(x, y)$$

# Continuous Cross-Entropy as Distortion

For a Gaussian VAE on images we typically have

$$\text{pri}^*, \text{dec}^*, \text{enc}^* = \underset{\text{pri,dec,enc}}{\text{argmin}} \; E_{y,z} \; -\ln \frac{p_{\text{pri}}(z)p_{\text{dec}}(y|z)}{p_{\text{enc}}(z|y)}$$

$$p_{\text{dec}}(y|z) = \frac{1}{Z} \; \exp(-||y - \hat{y}_{\text{dec}}(z)||^2/2\sigma^2)$$

$$\text{pri}^*, \text{dec}^*, \text{enc}^* = \underset{\text{pri,dec,enc}}{\text{argmin}} \; E_{y,z} \; KL(p_{\text{enc}}(z|y), p_{\text{pri}}(z)) \; {\color{red} + \frac{1}{2\sigma^2}||y - \hat{y}_{\text{dec}}(z)||^2 + d\ln\sigma}$$

The KL divergence term is not problematic. The problematic differential cross-entropy term can just be thought of as a weighted $L_2$ distortion. We typically take $\sigma$ to be a hyperparameter rather than optimize it in SGD.

# Continuous Cross-Entropy as Distortion

$$\text{pri}^*, \text{dec}^*, \text{enc}^* = \underset{\text{pri,dec,enc}}{\text{argmin}} \ E_{y,z} \ KL(p_{\text{enc}}(z|y), p_{\text{pri}}(z)) + \lambda \, \text{Dist}(y, \hat{y}_{\text{dec}}(z))$$

Various choices for distortion are possible including $L_2$ and $L_1$ distortion measures.

$$\text{Dist}(y, \hat{y}) = ||y - \hat{y}||^2 \qquad (L_2)$$

$$\text{or} \ \ \text{Dist}(y, \hat{y}) = ||y - \hat{y}||_1 = \sum_i |y[i] - \hat{y}[i]| \qquad (L_1)$$

END