

# **TTIC 31230, Fundamentals of Deep Learning**

David McAllester, Winter 2019

Expectation Maximization (EM)

The Evidence Lower Bound (the ELBO)

Variational Autoencoders (VAEs)

## Latent Variable Models

We are often interested in models of the form

$$P_{\Phi}(y) = \sum_z P_{\Phi}(z) P_{\Phi}(y|z).$$

$$P_{\Phi}(y|x) = \sum_z P_{\Phi}(z|x) P_{\Phi}(y|z).$$

For example, CTC and probabilistic grammar models.

# Expectation Maximization (EM)

## Mixture of Gaussian Modeling

$$\Phi = (\pi_1, \mu_1, \Sigma_1, \dots, \pi_k, \mu_k, \Sigma_k)$$

$$\begin{aligned} p_{\Phi}(y) &= \sum_i P(i) p(y|i) \\ &= \sum_i \pi_i \frac{1}{Z_i} \exp \left( -\frac{1}{2} (y - \mu_i)^{\top} \Sigma_i^{-1} (y - \mu_i) \right) \end{aligned}$$

$i$  is the latent variable.

# Expectation Maximization (EM)

## Mixture of Gaussian Modeling

$$\Phi = (\pi_1, \mu_1, \Sigma_1, \dots, \pi_k, \mu_k, \Sigma_k)$$

$$\text{Train} = \{y_1, \dots, y_N\}$$

Until Convergence:

$$P_{\Phi}(i|y_j) = \frac{\pi_i P(y_j|i)}{\sum_i \pi_i P(y_j|i)} \quad \text{Inference (E step)}$$

$$\left. \begin{aligned} \pi_i^{t+1} &= \frac{1}{N} \sum_j P_{\Phi^t}(i|y_j) \\ \mu_i^{t+1} &= \frac{1}{N} \sum_j P_{\Phi^t}(i|y_j) y_j \\ \Sigma_i^{t+1} &= \frac{1}{N} \sum_j P_{\Phi^t}(i|y_j) y_j y_j^{\top} \end{aligned} \right\} \quad \text{Model Update (M step)}$$

## General EM

$$\Phi^* = \underset{\Phi}{\operatorname{argmin}} E_{y \sim \text{Train}} - \ln P_{\Phi}(y)$$

$$P_{\Phi}(y) = \sum_z P_{\Phi}(z) P_{\Phi}(y|z).$$

$$\Phi^{t+1} = \underset{\Phi}{\operatorname{argmin}} E_{y \sim \text{Train}} E_{z \sim P_{\Phi^t}(z|y)} - \ln P_{\Phi}(z, y)$$

Update  
(M Step)

Inference  
(E Step)

## Colorization



**Input**

**Our Method**

**Ground-truth**

$x$

$\hat{y}$

$y$

Larsson et al., 2016

$x$  is a black and white image.

$y$  is a color image drawn from  $\text{Pop}(y|x)$ .

$\hat{y}$  is an arbitrary color image.

$P_{\Phi}(\hat{y}|x)$  is the probability that model  $\Phi$  assigns to the color image  $\hat{y}$  given black and white image  $x$ .

# Colorization with Latent Semantic Segmentation (TZ)



Input

Our Method

Ground-truth

$x$

$\hat{y}$

$y$

$$P_{\Phi}(\hat{y}|x) = \sum_z P_{\Phi}(z|x) P_{\Phi}(\hat{y}|z, x).$$

input  $x$

$P_{\Phi}(z|x) = \dots$  semantic segmentation

$P_{\Phi}(\hat{y}|z, x) = \dots$  segment colorization

# Maybe EM?

$$P_{\Phi}(y) = \sum_z P_{\Phi}(z)P_{\Phi}(y|z).$$

$$\Phi^{t+1} = \underset{\Phi}{\operatorname{argmin}} \underbrace{E_{y \sim \text{Train}}}_{\text{Update}} \underbrace{E_{z \sim P_{\Phi^t}(z|y)} - \ln P_{\Phi}(z, y)}_{\text{Inference}}$$

In most cases the inference is intractable!



## Variational Inference:

### The Evidence Lower Bound (The ELBO)

We introduce a friendly model  $P_{\Psi}(z|y)$  to approximate  $P_{\Phi}(z|y)$ .

$$\begin{aligned}\ln P_{\Phi}(y) &= E_{z \sim P_{\Psi}(z|y)} \ln P_{\Phi}(y) \\&= E_{z \sim P_{\Psi}(z|y)} \left( \ln P_{\Phi}(y) \frac{P_{\Phi}(z|y)}{P_{\Psi}(z|y)} + \ln \frac{P_{\Psi}(z|y)}{P_{\Phi}(z|y)} \right) \\&= \left( E_{z \sim P_{\Psi}(z|y)} \ln \frac{P_{\Phi}(z, y)}{P_{\Psi}(z|y)} \right) + KL(P_{\Psi}(z|y), P_{\Phi}(z|y)) \\&= \text{ELBO} + KL(P_{\Psi}(z|y), P_{\Phi}(z|y))\end{aligned}$$

## EM is Alternating Maximization of the ELBO

$$\text{ELBO} = E_{z \sim P_{\Psi}(z|y)} \ln \frac{P_{\Phi}(z, y)}{P_{\Psi}(z|y)} \quad (1)$$

$$= \ln P_{\Phi}(y) - KL(P_{\Psi}(z|y), P_{\Phi}(z|y)) \quad (2)$$

$$\text{by (2)} \quad \Psi^{t+1} = \underset{\Psi}{\operatorname{argmin}} E_{y \sim \text{Train}} KL(P_{\Psi}(z|y), P_{\Phi^t}(z|y)) = \Phi^t$$

$$\text{by (1)} \quad \Phi^{t+1} = \underset{\Phi}{\operatorname{argmax}} E_{y \sim \text{Train}} E_{z \sim P_{\Phi^t}(z|y)} \ln P_{\Phi}(z, y)$$

## Different Ways of Writing the ELBO

$$\begin{aligned}\text{ELBO} &= E_{z \sim P_{\Psi}(z|y)} \ln \frac{P_{\Phi}(z, y)}{P_{\Psi}(z|y)} \\ &= \ln P_{\Phi}(y) - KL(P_{\Psi}(z|y), P_{\Phi}(z|y)) \\ &= \left( E_{z \sim P_{\Psi}(z|y)} \ln P(y|z) \right) - KL(P_{\Psi}(z|x), P_{\Phi}(z)) \\ &= \left( E_{z \sim P_{\Psi}(z|y)} P_{\Phi}(z, y) \right) + H(P_{\Psi}(z|y))\end{aligned}$$

## Hard ELBO

Hard ELBO is to ELBO as hard EM is to EM.

$$\mathcal{L}_{\text{ELBO}}(y, \Phi, \Psi) = KL(P_{\Psi}(z|y), P_{\Phi}(z|y)) - \ln P_{\Phi}(y)$$

$$\mathcal{L}_{\text{ELBO}}(y, \Phi, \Psi) = E_{z \sim P_{\Psi}(z|y)} - \ln P_{\Phi}(z, y) + \ln P_{\Psi}(z|y)$$

$$\mathcal{L}_{\text{HELBO}}(y, \Phi, \Psi) = E_{z \sim P_{\Psi}(z|y)} - \ln P_{\Phi}(z, y)$$

## Measuring the ELBO

$$\text{ELBO} = E_{z \sim P_{\Psi}(z|y)} \ln \frac{P_{\Phi}(z, y)}{P_{\Psi}(z|y)}$$

If  $P_{\Phi}(z)$ ,  $P_{\Phi}(y|z)$ , and  $P_{\Psi}(z|y)$  are friendly (even when  $P_{\Phi}(y)$  is not friendly) we can measure ELBO loss through sampling.

If we can measure it, we can do gradient descent on it (but perhaps with difficulty).

**We want  $\Psi$  to adapt to  $\Phi$**

$$\mathcal{L}_{\text{ELBO}}(y, \Phi, \Psi) = KL(P_{\Psi}(z|y), P_{\Phi}(z|y)) - \ln P_{\Phi}(y)$$

$$Q^*(z|y) = P_{\Phi}(z|y)$$

$$E_{y \sim \text{Pop}} \mathcal{L}_{\text{ELBO}}(y, \Phi, Q^*) = H(\text{Pop}, P_{\Phi})$$

**However,  $\Phi$  can ignore  $\Psi$**

$$\mathcal{L}_{\text{ELBO}}(y, \Phi, \Psi) = KL(P_{\Psi}(z|y), P_{\Phi}(z|y)) - \ln P_{\Phi}(y)$$

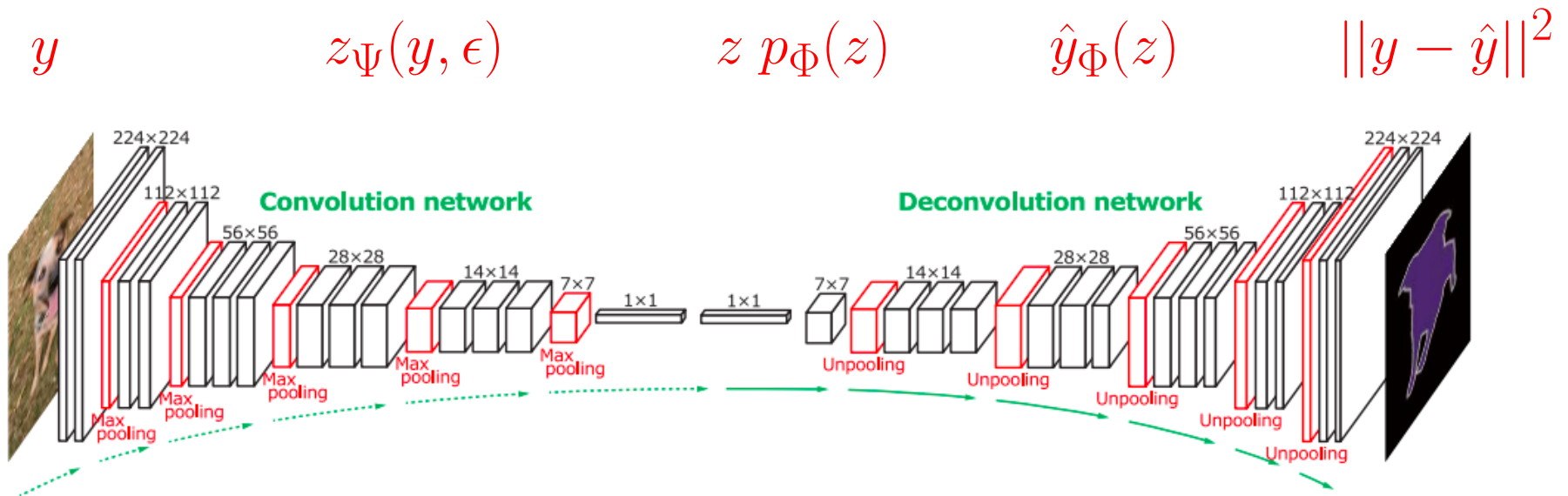
$$\begin{aligned} P^*(z) &= P_{\Psi}(z) \\ P^*(y|z) &= P_{\Phi}(y) \end{aligned}$$

$$E_{y \sim \text{Pop}} \mathcal{L}_{\text{ELBO}}(y, P^*, \Psi) = H(\text{Pop}, P_{\Phi})$$

It seems important that  $P_{\Phi}(y|z)$  have limited expressive power.

# A VAE for Images

Auto-Encoding Variational Bayes, Diederik P Kingma, Max Welling, 2013.



[Hyeonwoo Noh et al.]



## Gaussian Distributions

$$p_{\Phi}(z) \propto \exp \left( \sum_i (z[i] - \mu[i])^2 / (2\sigma[i]^2) \right)$$

$$p_{\Phi}(y|z) \propto \exp \left( \sum_j (y[j] - y_{\Phi}(z)[j])^2 / (2\gamma[j]^2) \right)$$

$$p_{\Psi}(z|y) \propto \exp \left( \sum_i (z[i] - z_{\Psi}(y)[i])^2 / (2\sigma_{\Psi}(y)[i]^2) \right)$$

## KL-Divergence Form for the ELBO

$$\begin{aligned} & E_{z \in p_{\Psi}(z|y)} \ln p_{\Psi}(z|y) - \ln p_{\Phi}(z)p_{\Phi}(y|z) \quad \mathcal{L}_{\text{ELBO}} \\ &= KL(p_{\Psi}(z|y), p_{\Phi}(z)) + E_{z \in P_{\Psi}(z|y)} - \ln p_{\Phi}(y|z) \end{aligned}$$

The ELBO is a KL-divergence + a cross entropy

Continuous KL-divergence is ok.

Continuous cross-entropy has issues — we will come back to that later.

## Closed Form KL-Divergence

$$KL(p_{\Psi}(z|y), p_{\Phi}(z))$$

$$= \sum_i \frac{\sigma_{\Psi}(y)[i]^2 + (z_{\Psi}(y)[i] - \mu[i])^2}{2\sigma[i]^2} + \ln \frac{\sigma[i]}{\sigma_{\Psi}(y)[i]} - \frac{1}{2}$$

## Standardizing $p_\Phi(z)$

The KL-divergence term is

$$\sum_i \frac{\sigma_\Psi(y)[i]^2 + (z_\Psi(y)[i] - \mu[i])^2}{2\sigma[i]^2} + \ln \frac{\sigma[i]}{\sigma_\Psi(y)[i]} - \frac{1}{2}$$

We can adjust  $\Psi$  to  $\Psi'$  such that

$$\begin{aligned} z_{\Psi'}(y)[i] &= z_\Psi(y)[i]/\sigma[i] + \mu[i] \\ \sigma_{\Psi'}(y)[i] &= \sigma_\Psi(y)/\sigma[i] \end{aligned}$$

We then get  $KL(p_\Psi(z|y), p_\Phi(z)) = KL(p_{\Psi'}(z|y), \mathcal{N}(0, I))$ .

## Standardizing $p_\Phi(z)$

Without loss of generality the VAE becomes.

$$\min_{\Phi, \Psi} E_y KL(P_\Psi(z|y), \mathcal{N}(0, I)) + E_{z \in P_\Psi(z|y)} - \ln p_\Phi(y|z)$$

## Reparameterization Trick for the Cross-Entropy

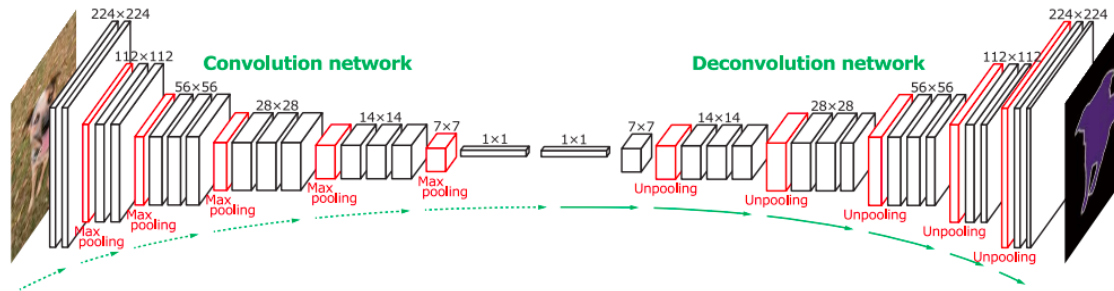
$$p_{\Psi}(z|y) \propto \exp \left( \sum_i (z[i] - \mathbf{z}_{\Psi}(y)[i])^2 / (2\sigma_{\Psi}(y)[i]^2) \right)$$

$$E_{z \in p_{\Psi}(z|y)} \ln p_{\Phi}(y|z)$$

$$= E_{\epsilon \sim \mathcal{N}(0, I)} z[i] = \mathbf{z}_{\Psi}(y)[i] + \sigma_{\Psi}(y)[i]\epsilon[i]; \quad \ln p_{\Phi}(y|z)$$

# Sampling

$$P_{\Psi}(z|y) \quad z \quad P_{\Phi}(z, y)$$



[Hyeonwoo Noh et al.]

Sampling uses just the second half  $P_{\Phi}(z, y)$ .

# Sampling



[Alec Radford]



## Why Blurry?

A common explanation for the blurriness of images generated from VAEs is the use of  $L_2$  as the distortion measure.

It does seem that  $L_1$  works better.

However, training on  $L_2$  distortion can produce sharp images in rate-distortion autoencoders.

# Noisy-Channel Rate-Distortion Autoencoders



The twilight zone is material for which I do not know of a reference.

## Differential Entropy and Cross-Entropy are Ill-Defined

$$\mathcal{L}_{\text{VAE}} = \sum_j \frac{E_{z \sim P_{\Psi}(z|y)} (\textcolor{red}{y}[j] - \hat{\textcolor{red}{y}}_{\Phi}(z)[j])^2}{2\textcolor{red}{\gamma}[j]^2} + \ln \textcolor{red}{\gamma}[j] \\ + KL(p_{\Psi}(z|y), p_{\Phi}(z))$$

Consider a probability density on light intensity.

While the first term is dimensionless,  $\textcolor{red}{\gamma}[j]$  is an intensity.

The cross-entropy term can be assigned any numerical value depending on the choice units (metric, English, or martian).

## Differential Entropy and Cross-Entropy are Ill-Defined

There are also other problems with continuous entropy and cross-entropy.

- Finite continuous entropy violates the source coding theorem — it takes an infinite number of bits to code a real number.
- Finite continuous entropy violates the data processing inequality that  $H(f(x)) \leq H(x)$ . For a continuous random variable  $x$  under finite continuous entropy we can have  $H(f(x)) > H(x)$ .

For these reasons it seems best to avoid using finite continuous entropy and finite continuous cross entropy.

## Distortion

A stochastic encoder  $p_{\Phi}(z|y)$ , a decoder  $y_{\Phi}(z)$ , and distortion function  $D$  define a quantity of distortion.

$$E_{y \sim \text{Pop}, z \sim p_{\Phi}(z|y)} D(y, y_{\Phi}(z))$$

For  $L_2$  distortion we can use

$$D(y, y') = ||y - y'||_2$$

Distortion can typically be given the same units as  $y$ .

## Rate

A stochastic encoder defines a rate.

$$p_{\Phi}(z) \doteq \sum_y \text{Pop}(y) p_{\Phi}(z|y)$$

$$I_{\Phi}(y, z) = E_y KL(p_{\Phi}(z|y), p_{\Phi}(z))$$

By Shannon's channel capacity theorem,  $I_{\Phi}(y, z)$  is the channel capacity when sending  $y$  across the noisy channel  $z$ .

For  $z$  continuous, a deterministic encoder has an infinite rate.

Here  $p_{\Phi}(z)$  is not friendly.

## Bounding the Rate

$$\begin{aligned} I_{\Phi}(y, z) &= E_{y \sim \text{Pop}} KL(p_{\Phi}(z|y), p_{\Phi}(z)) \\ &= E_{y, z} \ln p_{\Phi}(z|y) - \ln p_{\Psi}(z) + \ln p_{\Psi}(z) - \ln p_{\Phi}(z) \\ &= E_y KL(p_{\Phi}(z|y), p_{\Psi}(z)) - KL(p_{\Phi}(z), p_{\Psi}(z)) \\ &\leq E_y KL(p_{\Phi}(z|y), p_{\Psi}(z)) \end{aligned}$$

We can take  $p_{\Psi}(z)$  to be friendly, and WLOG, fixed at  $\mathcal{N}(0, I)$ .

## The Noisy-Channel Rate-Distortion Autoencoder

$$\Phi^* = \operatorname{argmin}_{\Phi} E_y KL(p_{\Phi}(z|y), \mathcal{N}(0, I)) + \frac{1}{\gamma} E_{z \sim p_{\Phi}(z|y)} D(y, y_{\Phi}(z))$$

Here  $\gamma$  has the same units as distortion and controls the trade-off between rate and distortion.



## Summary: Rate-Distortion

Rate-Distortion:  $y$ , continuous,  $\tilde{z}$  a bit string,

$$\Phi^* = \operatorname{argmin}_{\Phi} E_y |\tilde{z}_{\Phi}(y)| + \lambda D(y, y_{\Phi}(\tilde{z}_{\Phi}(y)))$$

Noisy Channel:  $\tilde{z} = z_{\Phi}(y) + \sigma_{\Phi}(y) \odot \epsilon$ ,  $\epsilon \sim \mathcal{N}(0, I)$

$$\Phi^* = \operatorname{argmin}_{\Phi} E_y KL(p_{\Phi}(\tilde{z}|y), \mathcal{N}(0, I)) + E_{\tilde{z} \sim p_{\Phi}(\tilde{z}|y)} \lambda D(y, y_{\Phi}(\tilde{z}))$$

## Summary: ELBO and VAE

ELBO:  $P_\Phi(z)$ ,  $P_\Phi(y|z)$ ,  $P_\Psi(z|y)$  friendly graphical models:

$$\Phi^*, \Psi^* = \operatorname{argmin}_{\Phi, \Psi} E_{y \sim P_{\text{op}}, z \sim P_\Psi(z|y)} \ln P_\Psi(z|y) - \ln P_\Phi(z) P_\Phi(y|z)$$

VAE:  $p_\Phi(z|y)$ ,  $p_\Phi(y|z)$  Gaussian:

$$\Phi^* = \operatorname{argmin}_{\Phi} E_{y \sim P_{\text{op}}} KL(p_\Phi(z|y), \mathcal{N}(0, I)) - E_{z \sim p_\Phi(z|y)} \ln p_\Phi(y|z)$$

**END**