Variational Autoencoders and the EM Algorithm

Dwaipayan Saha and Sunay Joshi

August 2024

1 Probability Background

We work on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Fix a pair of separable metric spaces $\mathcal{X}, \mathcal{Z}$ equipped with their Borel σ-algebras Σ_1, Σ_2 respectively. (In the case that $\mathcal{Z}$ is finite, we equip it with the discrete metric, so that Σ_2 is the power set $2^{\mathcal{Z}}$.) We consider a pair of random variables $X : \Omega \to \mathcal{X}$ and $Z : \Omega \to \mathcal{Z}$.

Let ρ denote the joint distribution of (X, Z) on $\mathcal{X} \times \mathcal{Z}$, i.e. ρ is the push-forward of $\mathbb{P}$ by the map $\omega \mapsto (X, Z)(\omega)$. Let μ denote the law of X, defined by $\mu(A) = \rho(A \times \mathcal{Z})$ for each $A \in \Sigma_1$. Similarly, let ν denote the law of Z. By the disintegration theorem [1, Section 4.4], there exists a stochastic kernel $\kappa_{Z|X}$ such that for all $(A, B) \in \Sigma_1 \times \Sigma_2$, we have

$$\rho(A \times B) = \int_A \kappa_{Z|X}(x, B)\mu(dx).$$

Similarly, there exists a stochastic kernel $\kappa_{X|Z}$ such that for all $(A, B) \in \Sigma_1 \times \Sigma_2$, we have

$$\rho(A \times B) = \int_B \kappa_{X|Z}(z, A)\nu(dz).$$

Intuitively, the maps $\kappa_{Z|X}$ and $\kappa_{X|Z}$ represent the conditional distribution of Z given X and the conditional distribution of X given Z, respectively.

Throughout this work, we will often identify distributions with their densities. Let $f_X : \mathcal{X} \to [0, +\infty)$ be the density of μ with respect to a reference measure λ on $\mathcal{X}$, i.e. the Radon-Nikodym derivative $\frac{d\mu}{d\lambda}$. The existence of this density is ensured by the Radon-Nikodym Theorem [1, Section 4.1].[1] Similarly, let $f_Z : \mathcal{Z} \to [0, +\infty)$ be the density of ν with respect to a reference measure on $\mathcal{Z}$. For each $z \in \mathcal{Z}$, we let $f_{X|Z}(\cdot|z) : \mathcal{X} \to [0, +\infty)$ denote the density of the conditional measure $\kappa_{X|Z}(z, \cdot)$. For each $x \in \mathcal{X}$, we let $f_{Z|X}(\cdot|x) : \mathcal{Z} \to [0, +\infty)$ denote the density of the conditional measure $\kappa_{Z|X}(x, \cdot)$.

Given a pair of probability measures ξ, η on a measurable space $(\mathcal{M}, \Sigma)$, the Kullback-Leibler (KL) divergence of ξ from η is defined as $D_{\mathrm{KL}}(\xi||\eta) = \mathbb{E}_{z\sim\xi}[\log \frac{d\xi}{d\eta}(z)]$ when $\xi \ll \eta$ and $+\infty$ otherwise. Given the Lebesgue measure

[1] Here λ is an appropriate dominating reference measure on $\mathcal{X}$ with $\mu \ll \lambda$. If X is a continuous random variable taking values in $\mathbb{R}^d$, then λ is the standard Lebesgue measure on $\mathbb{R}^d$. If X is discrete random variable, then λ is the counting measure on $\mathcal{X}$.

λ on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$, if ξ is a probability measure on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$, then if $\xi \ll \lambda$ with $\frac{d\xi}{d\lambda} \log \frac{d\xi}{d\lambda} \in L^1(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d), \lambda)$, the differential entropy of ξ is defined as $h(\xi) = -\mathbb{E}_{z\sim\xi}[\log \frac{d\xi}{d\lambda}(z)]$, and the differential entropy is $+\infty$ otherwise.[2]

In what follows, $\mathcal{X}$ can be interpreted as the data space and $\mathcal{Z}$ can be interpreted as the latent space. When we assume that a latent space exists, we are making a generative assumption. As we will see, the latent space can differ depending on the problem setting, but some common examples include categorical labels and embeddings in Euclidean space. To the reader unfamiliar with measure-theoretic probability, one can interpret the notions of distributions and densities in the classical sense for what follows.

2 Maximum Likelihood Derivation

We are given a set of i.i.d samples from the data distribution μ. We denote these datapoints $\{x_i\}_{i=1}^n$, where $x_i \in \mathcal{X} \subseteq \mathbb{R}^d$. We attempt to learn μ via maximum likelihood. In line with our earlier discussion, we make the following generative assumption on the data: for each $i \in [n]$,

1. Sample z_i from ν (equivalently, f_Z). This is known as the *prior distribution.*
2. Sample x_i from $\kappa_{X|Z}(z_i, \cdot)$ (equivalently, $f_{X|Z}(\cdot|z_i)$). This is known as the *sampling distribution.*
3. Reveal x_i and hide z_i.

The conditional distribution $f_{Z|X}(\cdot|x)$ will be relevant in what follows, and this is known as the *posterior distribution* over the latent space.

Throughout, we assume the prior f_Z is known. One way to learn the marginal distribution of X is to postulate a parametric family $(f^\theta_{X|Z}(\cdot|z))_{\theta\in\Theta}$ of densities for each $z \in \mathcal{Z}$. This induces a parametric family $(f^\theta_X(\cdot))_{\theta\in\Theta}$ of marginal distributions on X, where we define

$$f^\theta_X(x) := \int_{\mathcal{Z}} f^\theta_{X|Z}(x|z) f_Z(z) dz.$$

We learn the marginal distribution of X by solving for the maximum likelihood estimator $\hat{\theta}_{ML}$. Since our samples x_i are independent, the maximum likelihood estimator is given by

$$\hat{\theta}_{ML} = \operatorname*{argmax}_\theta \ell(\theta) = \operatorname*{argmax}_\theta \sum_{i=1}^n \log f^\theta_X(x_i). \tag{1}$$

[2]Notice that the differential entropy of a random variable is exactly its negative KL divergence to the Lebesgue measure on the same measurable space, despite the latter not necessarily being a probability measure.

Using our generative assumption and the law of total probability, Equation 1 can be written as

$$\hat{\theta}_{ML} = \operatorname*{argmax}_{\theta} \ell(\theta) = \operatorname*{argmax}_{\theta} \sum_{i=1}^{n} \log \left(\int_{\mathcal{Z}} f^{\theta}_{X|Z}(x_i|z) f_Z(z) dz \right).$$

Unfortunately, the objective function requires the computation of a potentially high-dimensional integral, and in certain settings this is a *non-convex* optimization problem. As such, we manipulate the objective to obtain a more tractable optimization problem. For simplicity, we work with a *single* sample x below, and we extend the derivation to our full dataset at the end.

We seek a proxy for the log likelihood $\log f^{\theta}_X(x)$. We derive a lower bound using the following trick. We introduce a parametrized family of posteriors $(f^{\phi}_{Z|X}(\cdot|x))_{\phi \in \Phi}$, and we manipulate the KL divergence $D_{\mathrm{KL}}(f^{\phi}_{Z|X}(\cdot|x)||f^{\theta}_{Z|X}(\cdot|x))$ until $\log f^{\theta}_X(x)$ appears.

We rewrite this KL divergence as follows. By Bayes' rule, $f^{\theta}_{Z|X}(z|x) = \frac{f^{\theta}_{X|Z}(x|z) f_Z(z)}{f^{\theta}_X(x)}$, so that

$$\frac{f^{\phi}_{Z|X}(z|x)}{f^{\theta}_{Z|X}(z|x)} = \frac{f^{\phi}_{Z|X}(z|x)}{f_Z(z)} \cdot \frac{f^{\theta}_X(x)}{f^{\theta}_{X|Z}(x|z)}.$$

We manipulate the equation to make the prior f_Z appear. This technique leads to a *prior matching* term. The KL objective becomes

$$\begin{aligned}
&\mathbb{E}_{z \sim f^{\phi}_{Z|X}(\cdot|x)} \left[\log \frac{f^{\phi}_{Z|X}(z|x)}{f^{\theta}_{Z|X}(z|x)} \right] \\
&= \mathbb{E}_{z \sim f^{\phi}_{Z|X}(\cdot|x)} \left[\log \frac{f^{\phi}_{Z|X}(z|x)}{f_Z(z)} + \log f^{\theta}_X(x) - \log f^{\theta}_{X|Z}(x|z) \right] \\
&= (D_{\mathrm{KL}}(f^{\phi}_{Z|X}(\cdot|x)||f_Z(\cdot)) - \mathbb{E}_{z \sim f^{\phi}_{Z|X}(\cdot|x)}[\log f^{\theta}_{X|Z}(x|z)]) + \log f^{\theta}_X(x).
\end{aligned}$$

Notice that the log likelihood appears in this manipulation. Isolating this log likelihood, we obtain

$$\begin{aligned}
\log f^{\theta}_X(x) = &D_{\mathrm{KL}}(f^{\phi}_{Z|X}(\cdot|x)||f^{\theta}_{Z|X}(\cdot|x)) \\
&+ (\mathbb{E}_{z \sim f^{\phi}_{Z|X}(\cdot|x)}[\log f^{\theta}_{X|Z}(x|z)] - D_{\mathrm{KL}}(f^{\phi}_{Z|X}(\cdot|x)||f_Z(\cdot))).
\end{aligned}$$

The quantity in parentheses is called the *Evidence Lower Bound* (ELBO):

$$\mathrm{ELBO} := (\mathbb{E}_{z \sim f^{\phi}_{Z|X}(\cdot|x)}[\log f^{\theta}_{X|Z}(x|z)] - D_{\mathrm{KL}}(f^{\phi}_{Z|X}(\cdot|x)||f_Z(\cdot))).$$

Indeed, since the KL divergence appearing in the preceding equation is non-negative, the ELBO is a lower bound for the "evidence" $\log f^{\theta}_X(x)$.[3] The ELBO

[3]The non-negativity of the KL Divergence follows from Jensen's Inequality.

will serve as our desired proxy. Instead of maximizing the log likelihood, we maximize the ELBO over (θ, ϕ):

$$\underset{\theta,\phi}{\operatorname{argmax}}\, \mathbb{E}_{z\sim f^{\phi}_{Z|X}(\cdot|x)}[\log f^{\theta}_{X|Z}(x|z)] - D_{\mathrm{KL}}(f^{\phi}_{Z|X}(\cdot|x)||f_Z(\cdot)) \tag{2}$$

The first term is called the reconstruction term. This term forces the likelihood of a latent variable z sampled from the posterior to be large. The second term is the prior matching term. This term forces our learned posterior to be close to the prior. The prior matching term is also known as entropy regularization.

Note that in general, if we know the prior $f_Z(\cdot)$ and the sampling distribution $f_{X|Z}(\cdot|z)$, then this completely specifies the joint distribution $f_{X,Z}(x, z)$ via the formula $f_{X,Z}(x, z) = f_{X|Z}(x|z)f_Z(z)$. In particular, this completely specifies the posterior $f_{Z|X}(\cdot|x)$. Thus, if we learn $f^{\theta}_{X|Z}(\cdot|z)$, we induce a posterior $f^{\theta}_{Z|X}(\cdot|x)$ via the formula

$$f^{\theta}_{Z|X}(z|x) = \frac{f^{\theta}_{X|Z}(x|z)f_Z(z)}{\int_{\mathcal{Z}} f^{\theta}_{X|Z}(x|z)f_Z(z)dz}.$$

However, computing the denominator involves a complicated integral. Therefore it's simpler to separately parametrize the sampling and posterior distributions by θ and ϕ respectively when optimizing the ELBO.

Returning to the setting of n samples, we sum the ELBO for each x_i to obtain the objective

$$(\theta^*, \phi^*) = \underset{\theta,\phi}{\operatorname{argmax}} \sum_{i=1}^{n} (\mathbb{E}_{z\sim f^{\phi}_{Z|X}(\cdot|x_i)}[\log f^{\theta}_{X|Z}(x_i|z)] - D_{\mathrm{KL}}(f^{\phi}_{Z|X}(\cdot|x_i)||f_Z(\cdot))).$$

Notice that $\theta^* \neq \hat{\theta}_{ML}$. However, since we optimized a lower bound on the original problem 1, the hope is that the two quantities are close. Indeed, if ϕ^* is such that $f^{\phi^*}_{Z|X}(\cdot|x)$ agrees with the posterior $f^{\theta^*}_{Z|X}(\cdot|x)$, then the log likelihood coincides with the ELBO. In this case, $f^{\theta^*}_X$ coincides with our maximum likelihood estimator of the data distribution.

As a sidenote, the ELBO is related to the field of variational inference (VI). In VI, the goal is to approximate the true posterior $f_{Z|X}(\cdot|x)$ using a parametric family $(f^{\phi}_{Z|X}(\cdot|x))_{\phi\in\Phi}$ of densities for each $x \in \mathcal{X}$. In VI, the optimal ϕ is computed by minimizing the KL divergence, which is a type of variational problem.

3 Gibbs Variational Principle Derivation

Theorem 1 (Gibbs Variational Principle, Lemma 4.10 in [7]). *Given any measure η on a measurable space $(\mathcal{Z}, \mathcal{B}(\mathcal{Z}))$ and a measurable function $g : \mathbb{R}^m \to \mathbb{R}$, we have*

$$\log \mathbb{E}_{\eta}[e^g] = \sup_{\xi\in\mathcal{P}(\mathcal{Z})} \left\{\mathbb{E}_{\xi}[g] - D_{KL}(\xi||\eta)\right\}.$$

The ELBO looks very much like the right-hand side of the Gibbs variational principle.[4] To make the connection precise, we let η be the prior f_Z and we let $g(z) = \log f^\theta_{X|Z}(x|z)$. Then Theorem 1 implies

$$\log \mathbb{E}_{z\sim f_Z(\cdot)}[f^\theta_{X|Z}(x|z)] = \sup_{\xi\in\mathcal{P}(\mathcal{Z})} \left\{ \mathbb{E}_{z\sim\xi}[\log f^\theta_{X|Z}(x|z)] - D_{\mathrm{KL}}(\xi||f_Z(\cdot)) \right\}.$$

The left-hand side can be expanded as

$$\int_{\mathcal{Z}} f^\theta_{X|Z}(x|z) f_Z(z) dz = f^\theta_X(x).$$

Next, note that

$$\mathrm{RHS} \geq \max_\phi \mathbb{E}_{z\sim f^\phi_{Z|X}(\cdot|x)}[\log f^\theta_{X|Z}(x|z)] - D_{\mathrm{KL}}(f^\phi_{Z|X}(\cdot|x)||f_Z(\cdot)).$$

Plugging in, we find

$$\log f^\theta_X(x) \geq \mathrm{ELBO},$$

which is exactly the statement that the ELBO is a lower bound on the log likelihood for any pair (θ, ϕ). We have equality in the case that the parametric family $f^\phi_{Z|X}(\cdot|x)$ consists of all probability distributions on $\mathcal{Z}$.

4 How to use learned sampling and posterior?

Suppose we optimize the ELBO to obtain (θ^*, ϕ^*). Denote our learned posterior distribution by $f^{\phi^*}_{Z|X}(\cdot|x)$ and sampling distribution by $f^{\theta^*}_{X|Z}(\cdot|z)$. A typical use of a learned VAE is the generation of new samples from $\mathcal{X}$. We refer to this as the generation task.

4.1 Reconstruction

For this task we are given some input data $x \in \mathcal{X}$. We encode it to get the corresponding latent representation by modelling the distribution $f^{\phi^*}_{Z|X}(\cdot|x)$ over $\mathcal{Z}$ space as $N(\mu_{\phi^*}(x), \mathrm{diag}(\sigma^2_{\phi^*}(x))$ where the mean and covariance functions are typically Feedforward Neural Networks $\mu_{\phi^*} : \mathcal{X} \to \mathbb{R}^m$ and $\sigma^2_{\phi^*} : \mathcal{X} \to \mathbb{R}^m_+$.

Sampling from $N(\mu_{\phi^*}(x), \mathrm{diag}(\sigma^2_{\phi^*}(x))$ yields the relevant latent z. From here one can decode it by sampling $\tilde{x}$ from $f^{\theta^*}_{X|Z}(\cdot|z)$, which is a distribution over $\mathcal{X}$ space. Once again this is modelled as $N(\mu_{\theta^*}(z), \mathrm{diag}(\sigma^2_{\theta^*}(z))$ where the mean and covariance functions are typically Feedforward Neural Networks $\mu_{\theta^*} : \mathcal{Z} \to \mathbb{R}^d$ and $\sigma^2_{\phi^*} : \mathcal{Z} \to \mathbb{R}^d_+$. Overall, the forward pass yields $\tilde{x}$ as the reconstruction of x.

[4] It is known that the maximizer is the measure ξ^* given by

$$\xi^*(dz) = \frac{e^g \eta(dz)}{\mathbb{E}_\eta[e^g]},$$

which is a "tilted" version of η that places extra weight where g is large.

4.2 Generation

Given a learned VAE, there are two steps to generate a new sample. First, sample a latent z from the prior distribution $f_Z(\cdot)$ on the space $\mathcal{Z}$. Then, decode this latent z by sampling x from the learned sampling distribution $f^{\theta^*}_{X|Z}(\cdot|z)$ on the space $\mathcal{X}$. This sampling distribution is modeled as the Gaussian $N(\mu_{\theta^*}(z), \text{diag}(\sigma^2_{\theta^*}(z)))$, where the mean and covariance functions are typically Feed-Forward Neural Networks $\mu_{\theta^*} : \mathcal{Z} \to \mathbb{R}^d$ and $\sigma^2_{\phi^*} : \mathcal{Z} \to \mathbb{R}^d_+$.

4.3 The use of Gaussians

It might seem surprising that we model the posterior and sampling distributions as Gaussians. To see why this works, consider the generation task in the special case that $\sigma^2_{\theta^*} = 0$. In this setting, we first sample z from the prior, a standard Gaussian, and then we apply the deterministic map μ_{θ^*} to obtain the sample $x = \mu_{\theta^*}(z)$. The distribution of x is the pushforward of the standard Gaussian by the map μ_{θ^*}. Since μ_{θ^*} is modeled as a neural network, it can be a very complicated function. Thus the distribution of the sample x can also be very complex.

5 A Simple Setting: Image Generation

We are given a set of images $\{x_i\}_{i=1}^n \subseteq \mathcal{X}$, where $\mathcal{X} \subseteq \mathbb{R}^d$. This could be a set of images of animals, digits, etc. We wish to generate new images which did not exist in our dataset and ensure that they still look accurate. Suppose $\mathcal{X} \subseteq \mathbb{R}^d$ is the space of all animal images. In the categorical setting, the latent space $\mathcal{Z}$ is a finite set $[m]$, and each category represents a different type of animal. Alternatively, we could simply consider $\mathcal{Z} \subseteq \mathbb{R}^m$. In our dataset, we do not have access to the latent variable z_i corresponding to each x_i. In real-world datasets, $\mathcal{X}$ concentrates around a low dimensional subspace of the ambient $\mathbb{R}^d$.

At this point one can postulate the same data generation algorithm and optimization problem as in Section 2. Solving that problem would yield a learned sampling distribution that could be used to generate new images of animals in a given latent z. However, we must select an appropriate prior distribution. Two options are:

1. Using the uniform distribution over $[m]$.
2. Considering the standard Gaussian measure on $\mathbb{R}^m$.

Even if our latent space is discrete in this case, the computations above work so long as the posterior has the same discrete support as the prior..

Alternatively, this task can be approached using the Vector Quantized-VAE (VQ-VAE) introduced in [6]. Given a specified number of categories m, the VQ-VAE discretizes a continuous latent space $\mathcal{Z}$ by learning a codebook of m vectors $(e_j)_{j\in[m]}$ in $\mathcal{Z}$. Unlike a typical VAE, the encoder is a deterministic map parametrized by ϕ whose output is rounded to the nearest vector in the

codebook. Thus the posterior is a distribution supported on a single vector e_j. As in a VAE, the sampling distribution is parametrized by θ. The VQ-VAE learns the parameters θ and ϕ, as well as the codebook $(e_j)_{j\in[m]}$. The VQ-VAE objective includes a reconstruction term, as well as additional quantization and commitment terms that depend on the codebook.

6 Training

In this section, we discuss how to train a VAE, following [5]. As can be recalled from Section 4, implementing a VAE requires a pair of Feed-Forward Neural Nets. The first neural net is the encoder, which learns ϕ and thus the posterior $f^{\phi}_{Z|X}(\cdot|x)$. The second neural net is the decoder, which learns θ and thus the sampling distribution $f^{\theta}_{X|Z}(\cdot|z)$. We compute the pair (θ^*, ϕ^*) that maximizes the ELBO using gradient descent.

Suppose we want to perform a single step of gradient descent on the ELBO. First, we replace the expectation $\mathbb{E}_{z\sim f^{\phi}_{Z|X}(\cdot|x)}$ with an empirical average. For each $i \in [n]$, sample $\{z_i^{(\ell)}\}_{\ell\in[L]}$ i.i.d. from the posterior $f^{\phi}_{Z|X}(\cdot|x_i)$. The empirical objective is given by

$$\operatorname*{argmax}_{\theta,\phi} \sum_{i=1}^{n} \frac{1}{L} \sum_{\ell=1}^{L} \log f^{\theta}_{X|Z}(x_i|z_i^{(\ell)}) - \sum_{i=1}^{n} D_{\mathrm{KL}}(f^{\phi}_{Z|X}(\cdot|x_i)||f_Z(\cdot))$$

(The law of large numbers implies that the empirical objective converges almost surely to the population objective as $L \to \infty$.) Suppose we compute the gradient with respect to ϕ. If we treat the sampled $z_i^{(\ell)}$'s as constants, then we fail to take into account the dependence of the first sum on ϕ, and our gradient will be inaccurate. In order to make the dependence of the $z_i^{(\ell)}$'s on ϕ explicit, we must *reparametrize* $f^{\phi}_{Z|X}(\cdot|x_i)$. Since $f^{\phi}_{Z|X}(\cdot|x)$ is the Gaussian $N(\mu_\phi(x), \mathrm{diag}(\sigma^2_\phi(x)))$, a sample from this distribution can be written as $z = \mu_\phi(x) + \sigma_\phi(x) \odot \varepsilon$, where $\varepsilon \sim N(0, I)$ and where $\odot$ is the element-wise product. Crucially, the distribution of ε has no dependence on ϕ. Thus if we rewrite $z_i^{(\ell)}$ as $z_i^{(\ell)} = \mu_\phi(x_i) + \sigma_\phi(x_i) \odot \varepsilon_i^{(\ell)}$ where $\varepsilon_i^{(\ell)}$ are i.i.d. $N(0, I)$, our objective becomes

$$\operatorname*{argmax}_{\theta,\phi} \sum_{i=1}^{n} \left[\frac{1}{L} \sum_{\ell=1}^{L} \log f^{\theta}_{X|Z}(x_i|\mu_\phi(x_i) + \sigma_\phi(x_i) \odot \varepsilon_i^{(\ell)}) - D_{\mathrm{KL}}(f^{\phi}_{Z|X}(\cdot|x_i)||f_Z(\cdot)) \right].$$

Since the dependence on ϕ is made explicit, we can treat the $\varepsilon_i^{(\ell)}$'s as constant and compute the gradient with respect to θ, ϕ.

When performing gradient descent, at each iteration t, given current parameter estimates (θ^t, ϕ^t), we sample a fresh set of noise variables $\{\varepsilon_{i;t}^{(\ell)}\}_{\ell\in[L]}$ for each $i \in [n]$, and we compute the gradient of the resulting objective.

7 Relationship with Expectation-Maximization Algorithm

The VAE is related to the Expectation-Maximization (EM) algorithm, which can be used for model-based clustering of the dataset $\{x_i\}_{i=1}^n$ where $x_i \in \mathcal{X} \subseteq \mathbb{R}^d$. The EM algorithm is an unsupervised learning algorithm that uses the same generative assumptions as above. As above, we derive EM through a variational route. We start with a single sample x.

Suppose that we parametrize the prior and sampling distributions as f_Z^θ and $f_{X|Z}^\theta$ for some parameter θ. Since the joint distribution is determined by the prior and sampling distributions, this induces a parametrization of the joint and the posterior distributions. In what follows, we write $f_{X,Z}^\theta$ and $f_{Z|X}^\theta$ as the induced parametrizations.

Notice that the marginal log likelihood $\log f_X^\theta(x)$ is actually equal to the ELBO objective from Equation 2, since we use the induced θ posterior.[5] Specifically,

$$\begin{aligned}
\log f_X^\theta(x) &= D_{\mathrm{KL}}(f_{Z|X}^\theta(\cdot|x)||f_{Z|X}^\theta(\cdot|x)) \\
&\quad + (\mathbb{E}_{z\sim f_{Z|X}^\theta(\cdot|x)}[\log f_{X|Z}^\theta(x|z)] - D_{\mathrm{KL}}(f_{Z|X}^\theta(\cdot|x)||f_Z^\theta(\cdot))) \\
&= \mathbb{E}_{z\sim f_{Z|X}^\theta(\cdot|x)}[\log f_{X|Z}^\theta(x|z)] - D_{\mathrm{KL}}(f_{Z|X}^\theta(\cdot|x)||f_Z(^\theta\cdot)) \\
&= \mathbb{E}_{z\sim f_{Z|X}^\theta(\cdot|x)}\left[\log\left(\frac{f_{X|Z}^\theta(x|z)f_Z^\theta(z)}{f_{Z|X}^\theta(z|x)}\right)\right].
\end{aligned}$$

Rewriting the expectation as an integral, and applying Bayes' rule we find

$$= \int_{\mathcal{Z}} \log\left(\frac{f_{X|Z}^\theta(x|z)f_Z^\theta(z)}{f_{Z|X}^\theta(z|x)}\right) f_{Z|X}^\theta(z|x)dz = \int_{\mathcal{Z}} \log\left(\frac{f_{X,Z}^\theta(x,z)}{f_{Z|X}^\theta(z|x)}\right) f_{Z|X}^\theta(z|x)dz.$$

Splitting the log once more, the right-hand side becomes

$$= \mathbb{E}_{z\sim f_{Z|X}^\theta(\cdot|x)}[\log f_{X,Z}^\theta(x,z)] + h(f_{Z|X}^\theta(\cdot|x)).$$

The first term is an expected full-data log likelihood, and the second term is the differential entropy of the posterior. Since the differential entropy is non-negative, it's natural to instead optimize the lower bound

$$\operatorname*{argmax}_{\theta,\phi} \mathbb{E}_{z\sim f_{Z|X}^\theta(\cdot|x)}[\log f_{X,Z}^\theta(x,z)],$$

This is known as the EM objective function derived via the variational method. Returning to the full sample, we find the objective

$$\operatorname*{argmax}_{\theta,\phi} \sum_{i=1}^n \mathbb{E}_{z\sim f_{Z|X}^\theta(\cdot|x_i)}[\log f_{X,Z}^\theta(x_i,z)]. \tag{3}$$

[5] This follows since the KL divergence $D_{\mathrm{KL}}(\xi,\eta) \geq 0$ and is 0 if and only if $\xi = \eta$.

The difficulty in solving this optimization problem is that for each term, both the integrand and the distribution of z depend on θ. Instead of directly optimizing the objective, the EM algorithm "decouples" the two occurrences of θ through an iterative approach, alternating between two steps. It falls within the much broader class of Majorization-Minimization (MM) algorithms. At iteration t, we keep a current estimate θ^t of the parameter θ. In the first step, the *Expectation Step*, we compute the expectation

$$\ell(\theta; \theta^t) := \sum_{i=1}^{n} \mathbb{E}_{z \sim f_{Z|X}^{\theta^t}(z|x_i)}[\log f_{X,Z}^{\theta}(x_i, z)].$$

This serves as our current approximation of the expected full-data log likelihood objective. In the second step, the *Maximization Step*, we maximize this proxy to obtain the new estimate $\theta^{t+1} := \text{argmax}_\theta \ell(\theta; \theta^t)$. We repeat until the iterates $\{\theta^t\}$ converge. In some cases, as shown later the function $\ell(\theta; \theta^t)$ has a closed form.

7.1 Why does the EM algorithm work?

We must redo our ELBO derivation in order to justify the EM algorithm; we follow [4]. We would like to relate the marginal log likelihood $\log f_X^\theta(x)$ to our function $\ell(\theta; \theta^t)$, which is an expectation with respect to $f_{Z|X}^{\theta^t}(\cdot|x)$. Fix an arbitrary z. Then

$$\log f_X^\theta(x) = \log \frac{f_{X,Z}^\theta(x, z)}{f_{Z|X}^\theta(z|x)} = \log f_{X,Z}^\theta(x, z) - \log f_{Z|X}^\theta(z|x).$$

Taking an expectation with respect to $z \sim f_{Z|X}^{\theta^t}(\cdot|x)$, we obtain

$$\log f_X^\theta(x) = \ell(\theta; \theta^t) - \mathbb{E}_{z \sim f_{Z|X}^{\theta^t}(\cdot|x)}[\log f_{Z|X}^\theta(\cdot|x)].$$

If we evaluate this identity at $\theta = \theta^t$, we obtain

$$\log f_X^{\theta^t}(x) = \ell(\theta^t; \theta^t) + h(f_{Z|X}^{\theta^t}(\cdot|x)).$$

Taking the difference between these two equations, we obtain

$$\begin{aligned}(\log f_X^\theta(x) - \log f_X^{\theta^t}(x)) = &\left(\ell(\theta; \theta^t) - \ell(\theta^t; \theta^t)\right) \\ &+ \left(-\mathbb{E}_{z \sim f_{Z|X}^{\theta^t}(\cdot|x)}[\log f_{Z|X}^\theta(\cdot|x)] - h(f_{Z|X}^{\theta^t}(\cdot|x))\right).\end{aligned}$$

We claim that the second term on the right-hand side is non-negative, with equality when $\theta = \theta^t$. Indeed, this follows from the non-negativity of the KL

divergence[6]

$$\begin{aligned}-\mathbb{E}_{z\sim f^{\theta^t}_{Z|X}(\cdot|x)}[\log f^{\theta}_{Z|X}(\cdot|x)] - h(f^{\theta^t}_{Z|X}(\cdot|x)) &= \mathbb{E}_{z\sim f^{\theta^t}_{Z|X}(\cdot|x)}\left[\log \frac{f^{\theta^t}_{Z|X}(\cdot|x)}{f^{\theta}_{Z|X}(\cdot|x)}\right] \\ &= D_{\mathrm{KL}}(f^{\theta^t}_{Z|X}(\cdot|x)||f^{\theta}_{Z|X}(\cdot|x)) \\ &\geq 0.\end{aligned}$$

In other words, we obtain

$$\log f^{\theta}_X(x) - \log f^{\theta^t}_X(x) \geq \ell(\theta;\theta^t) - \ell(\theta^t;\theta^t).$$

Thus if the iteration at step t increases the function $\ell(\cdot;\theta^t)$, it also increases the value of the marginal log likelihood by that same amount.

7.2 Example: Gaussian Mixture Model

We make the above explicit in the special case of a Gaussian Mixture Model (GMM). Specifically, we assume that the label z is sampled from a categorical distribution $\pi = (\pi_1, \ldots, \pi_m)$ on $[m]$, where $\sum_{j=1}^m \pi_j = 1$ and $\pi \geq 0$. Next, we sample the datapoint x from the sampling distribution $f^{\theta}_{X|Z}(x|z)$, which we model as the Gaussian $N(\mu_z, \Sigma_z)$.[7] In other words, we seek to learn the parameter $\theta = \{\mu_j, \Sigma_j, \pi_j\}_{j\in[m]}$. For each $x \in \mathcal{X}$, the posterior has the following explicit form:

$$\kappa_{Z|X}(x, \{z\}) = \frac{\pi_z f^{\theta}_{X|Z}(x|z)}{\sum_{j=1}^m \pi_j f^{\theta}_{X|Z}(x|j)} = \frac{\pi_z \varphi(x;\mu_z,\Sigma_z)}{\sum_{j=1}^m \pi_j \varphi(x;\mu_j,\Sigma_j)} = f^{\theta}_{Z|X}(z|x).$$

This makes it clear that the posterior and the joint distribution can indeed be parametrized by θ. We initialize our parameters at step 0 randomly and denote the vector as θ^0. Our EM update at iteration t consists of two steps:

1. *Expectation Step:* We wish to compute the expectation as a function of θ given our current best estimate of the parameters θ^t. We obtain

$$\begin{aligned}\ell(\theta;\theta^t) &= \sum_{i=1}^n \mathbb{E}_{z\sim f^{\theta^t}_{Z|X}(z|x_i)}[\log f^{\theta}_{X,Z}(x_i, z)] \\ &= \sum_{i=1}^n \sum_{j=1}^m \log f^{\theta}_{X,Z}(x_i, j) f^{\theta^t}_{Z|X}(j|x_i) \\ &= \sum_{i=1}^n \sum_{j=1}^m (\log \pi_j + \log \varphi(x_i;\mu_j,\Sigma_j)) \cdot \frac{\pi^t_j \varphi(x_i;\mu^t_j,\Sigma^t_j)}{\sum_{l=1}^m \pi^t_l \varphi(x_i;\mu^t_l,\Sigma^t_l)}.\end{aligned}$$

[6] There are multiple methods to prove this, in the information theoretic context one might rely on using Gibbs Inequality which is (as expected) proven via a classical application of Jensen's Inequality.

[7] We denote the density of the Gaussian measure with mean μ and covariance Σ at x by $\varphi(x;\mu,\Sigma)$.

where we used Bayes' rule and plugged in our earlier results in the last line.

2. *Maximization Step:* We compute our next best estimate of the parameters by optimizing the function $\ell(\theta, \theta^t)$ derived above:

$$\theta^{t+1} = \underset{\theta=\{\pi_j,\mu_j,\Sigma_j\}_{j=1}^m}{\operatorname{argmax}} \ell(\theta; \theta^t).$$

It is easy to verify by optimizing the relevant Lagrangian that θ^{t+1} has components

$$\begin{aligned}
\pi_j^{t+1} &= \frac{1}{n}\sum_{i=1}^n f_{Z|X}^{\theta^t}(j|x_i), \\
\mu_j^{t+1} &= \frac{\sum_{i=1}^n f_{Z|X}^{\theta^t}(j|x_i)x_i}{\sum_{i=1}^n f_{Z|X}^{\theta^t}(j|x_i)}, \\
\Sigma_j^{t+1} &= \frac{\sum_{i=1}^n f_{Z|X}^{\theta^t}(j|x_i)(x_i-\mu^{t+1})(x_i-\mu^{t+1})^\top}{\sum_{i=1}^n f_{Z|X}^{\theta^t}(j|x_i)}
\end{aligned}$$

for all $j \in [m]$.

We repeat these iterations until the iterates $\{\theta_t\}_{t\geq 0}$ satisfy satisfy some preimposed convergence condition. By our majorization guarantee, each iteration increases the log likelihood.

If one wishes to label the dataset $\{x_i\}_{i=1}^n$, one can simply consider the final parameter vector θ^* and for each $i \in [n]$ set the label to be

$$z_i = \underset{j\in[m]}{\operatorname{argmax}} f_{Z|X}^{\theta^*}(j|x_i).$$

Furthermore, the *K-Means Algorithm* is a special case of this when all the covariances are isotropic, i.e., $\Sigma_j = \sigma^2 I_d$ for some known parameter σ^2 [2].

We have applied EM to the GMM, but it applies to other distributions as well. In [8], it is shown that if the marginal log likelihood function $\ell(\theta) = \log f_X^\theta(x)$ has a unique local maximum at $\hat{\theta}_{ML}$, if $\hat{\theta}_{ML}$ is the only stationary point, and if the gradient $\nabla_\theta \ell(\theta; \tilde{\theta})$ is continuous as a function of $(\theta, \tilde{\theta})$, then EM converges to the maximum likelihood estimator.

8 Issues with VAEs

In this section, we discuss a few issues with VAEs and how to fix them, following [9].

One issue is that at times, VAEs can learn "uninformative latent codes." In some situations, the learned posterior $f_{Z|X}^\phi(\cdot|x)$ is very similar for different values of x. According to [9], the KL regularizer as the source of the problem.

Since the KL term forces $f^{\phi}_{Z|X}(\cdot|x)$ to be close to the prior $f_Z(\cdot)$ for *all* x, it is natural that the learned code doesn't depend much on x. To alleviate this, the InfoVAE of [10] uses the Maximum Mean Discrepancy regularizer from [3]. Other issues include uninterpretable latent representations and variance explosion; for more, see [9].

References

[1] A. Dembo. Probability theory: Stat310/math230; september 12, 2010, 2010.

[2] J. Fan, R. Li, C.-H. Zhang, and H. Zou. *Statistical foundations of data science*. Chapman and Hall/CRC, 2020.

[3] A. Gretton, K. Borgwardt, M. Rasch, B. Schölkopf, and A. Smola. A kernel method for the two-sample-problem. *Advances in neural information processing systems*, 19, 2006.

[4] G. Gundersen. Expectation–maximization. 2019. `https://gregorygundersen.com/blog/2019/11/10/em/`.

[5] C. Luo. Understanding diffusion models: A unified perspective. *arXiv preprint arXiv:2208.11970*, 2022.

[6] A. Van Den Oord, O. Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.

[7] R. Van Handel. Probability in high dimension. *Lecture Notes (Princeton University)*, 2(3):2–3, 2014.

[8] C. J. Wu. On the convergence properties of the em algorithm. *The Annals of statistics*, pages 95–103, 1983.

[9] S. Zhao. A tutorial on information maximizing variational autoencoders (infovae). 2017. `https://ermongroup.github.io/blog/a-tutorial-on-mmd-variational-autoencoders/`.

[10] S. Zhao, J. Song, and S. Ermon. Infovae: Information maximizing variational autoencoders. *arXiv preprint arXiv:1706.02262*, 2017.