VAE lecture notes

Chin-Wei Huang

April 5, 2020

The goal of this lecture notes is to summarize the technical details of (vanilla) variational autoencoders. I will not motivate things like I did during the lecture; see [3] for more details.

Set-up: we assume a data generating process: $z \sim p_Z(z) := \mathcal{N}(z; 0, I)$ (fixed prior, but in general can be a parametric family that is trainable), $x \sim p_{X|Z}(x|z) := \mathcal{N}(x; \mu_\theta(x), \sigma^2_\theta(x))$. The latter is a convention since σ^2 is a vector (so is μ): we mean the covariance of the multivariate normal distribution is $\mathrm{diag}(\sigma^2_\theta(z))$, *i.e.* a diagonal matrix. I will also drop the subscript of the density function for simplicity, and write $p(x|z)$ and $p(z)$ (they are different functions!). We will be calling $p(x,z) = p(x|z)p(z)$ the generative model (omitting "generative" sometimes), and we will introduce an "inference" model later.

1 Variational Lower Bound

The marginal density of x under the data generating process we just defined is $p(x) = \int p(x,z)dz = \int p(x|z)p(z)dz$, which is in general intractable (try to write out the conditional density $p(x|z)$, and you'll see it's impossible to integrate). So to approximately compute the log likelihood, we will derive a variational lower bound. Fixing some $q(z)$, we have

$$\log p(x) = \log \int p(x|z)p(z)dz \tag{1}$$

$$= \log \int q(z) \frac{p(x|z)p(z)}{q(z)} dz \tag{2}$$

$$= \log \mathbb{E}_{z\sim q(z)} \left[\frac{p(x|z)p(z)}{q(z)} \right] \tag{3}$$

$$\geq^* \mathbb{E}_{z\sim q(z)} \left[\log \frac{p(x|z)p(z)}{q(z)} \right] =: L[q] \tag{4}$$

where $*$ is due to Jensen's inequality (see Appendix). $L[q]$ is called the variational lower bound, or the *Evidence Lower BOund* (ELBO) of the log marginal likelihood. This is a *variational form*[1] of the log marginal density since it can be represented as $\log p(x) = \sup_{q:\int q=1} L[q]$, where the supremum (maximum) is taken over the set of valid probability density functions. In other words, $\log p(x)$ can be approximated by maximizing this lower bound by tuning q. We know that the maximum is attainable since Jensen's inequality becomes an equality *iff*

[1] The term "variational form" comes from the theory of variational calculus; calculus for infinitesimal changes in functions.

$p(x|z)p(z)/q(z) = C$ for some C (since log is strictly concave). This means $q(z) = \frac{p(x|z)p(z)}{C}$, and since $q(z)$ needs to be a probability density function (integrating to 1), the maximizer of the lower bound is $q^*(z) = \frac{p(x|z)p(z)}{\int p(x|z)p(z)dz} = \frac{p(x|z)p(z)}{p(x)} = p(z|x)$. By this we know the maximum value of the lower bound is $C = \int p(x|z)p(z)dz = p(x)$, *i.e.* the marginal likelihood!

We'll see how different rearrangements of $L[q]$ will give us different intuitions of what the lower bound is actually doing (hopefully). First,

$$L[q] = \mathbb{E}_{z\sim q(z)}[\log p(x,z)] + H_q(z) \tag{5}$$

where the first term is the *Expected Complete Data (log-)Likelihood* (ECDL), and the second term is the *entropy* of the random variable z distributed by $q(z)$. This allows us to view $-\log p(x,z) =: E(z)$ as some energy function we want to minimize (flipping the sign of the maximization problem), and the entropy term makes sure q will not collapse to a point mass at the minimizer of energy function $E(z)$. Writing it this way makes it somewhat easier to see the solution of this problem is the Gibbs density $q^*(z) \propto \exp(-E(z))$.

Rearranging another way gives

$$\begin{aligned} L[q] &= \mathbb{E}_{z\sim q(z)}[\log p(x|z)] - \mathbb{E}_{z\sim q(z)}[\log q(z) - \log p(z)] \\ &= \mathbb{E}_{z\sim q(z)}[\log p(x|z)] - D_{\mathrm{KL}}(q(z)||p(z)) \end{aligned} \tag{6}$$

so if we maximize the lower bound by tuning q (to make it a better approximation of $\log p(x)$), it amounts to making $z \sim q(z)$ a more likely z that generates x while making $q(z)$ close to $p(z)$. Recall we will be sampling from $z \sim p(z)$ as per the definition of the generative model, so this is the kind of regularization we want for the variational distribution.

2 Maximizing the ELBO

Recall that we use θ to parameterize the generative network (the mean and diagonal variance of the conditional Gaussian $p(x|z)$). We will use ϕ to denote the parameters of the variational distribution. For simplicity, we assume q is also a multivariate Gaussian with diagonal covariance (mean-field assumption). That is, $\phi = (\mu_q, \sigma_q^2)$. Then the ELBO is a function of θ and ϕ, so we write it as $L(\theta, \phi)$.

Before we start, note that we are trying to differentiate an expected value (which involves integrating a complicated, nonlinear function) so that we can perform gradient-based optimization. This usually does not have an analytic solution (i.e. a convenient expression). It is possible, however, to rearrange the gradient of the expectation as an expected value of some other random variable $\nabla L = \mathbb{E}[\eta]$ where η is a noisy yet unbiased estimate for the gradient. This way, we can use η to perform stochastic gradient optimization.

2.1 Tuning parameters of the "model"

Since the ELBO is still an expected value, it is still not tractable. Luckily we can have an "unbiased estimate" for the gradient wrt the model parameters. Since $H_{q_\phi}(z)$ is a constant wrt θ, differentiating (5) yields

$$\begin{aligned}\nabla_\theta L(\theta,\phi) &= \nabla_\theta \mathbb{E}_z[\log p_\theta(x|z)p(z)] && (7)\\ &= \nabla_\theta \int q_\phi(z)\log p_\theta(x|z)p(z)dz && (8)\\ &= \int q_\phi(z)\nabla_\theta \log p_\theta(x|z)p(z)dz && (9)\\ &= \mathbb{E}_z[\nabla_\theta \log p_\theta(x|z)p(z)] && (10)\end{aligned}$$

Writing the gradient as an expectation allows us to draw a Monte-Carlo sample of it as an estimate of the gradient itself. That is, with $z \sim q(z)$, we simply update the model by taking the stochastic gradient direction

$$\nabla_\theta \log p_\theta(x|z)p(z)$$

When we choose not to learn the prior (e.g. when we fix it to be a standard Gaussian), then it is equal to

$$\nabla_\theta \log p_\theta(x|z)$$

This gradient estimator usually has a low variance in practice as long as the variational distribution $q(z)$ is good enough.

2.2 Tuning the variational parameters

In this section, we will try to make the variational distribution a better approximation to the posterior $p(z|x)$ by tuning its parameters ϕ. Unfortunately, when taking the gradient wrt ϕ, unlike how it was done in Equation (8), we do not have the luxury to simply push the gradient sign inside of the expectation, since the density function q_ϕ now depends on ϕ

$$\nabla_\phi L(\theta,\phi) = \nabla_\phi \int q_{\phi}(z)[\log p_\theta(x|z)p(z) - \log q_\phi(z)]dz \tag{11}$$

There are many tricks[2] one can apply to derive a Monte Carlo estimator for the gradient, of which the well-celebrated REINFORCE algorithm (aka the *score function estimator*) is a classic example [5]:

$$\nabla_\phi L(\theta,\phi) = \mathbb{E}[\eta_{sf}]$$

where

$$\eta_{sf} = \underbrace{[\log p_\theta(x|z)p(z) - \log q_\phi(z)]}_{reward:R(z)} \cdot \overbrace{\nabla_\phi \log q_\phi(z)}^{score}$$

with $z \sim q_\phi(z)$. The *score* itself is "uninformative", since one can show $\mathbb{E}[\nabla_\phi \log q_\phi(z)] = 0$, but when weighted by the "reward function", the weighted average points towards the correct gradient direction of the ELBO. Intuitively, if we draw multiple z's from q, it tells us to place more importance on the gradients that correspond to higher reward $R(z)$, as our goal is to maximize the average reward $\mathbb{E}[R(z)]$.

[2] See [4] for a comprehensive review.

One major criticism of the score function estimator is that it usually has a high variance in practice, which motivates the quest for a more efficient estimator. One such an example which is used to train a variational autoencoder is the *pathwise estimator* (aka the *reparameterization trick*). The pathwise estimator is found to have smaller variance in practice[3]. The intuition is that it makes better use of the reward function $R(z)$ by looking at its first order gradient.

To derive the pathwise estimator, one needs a "reparameterization" of the random variable $z \sim q_\phi(z)$. A reparameterization is a pair (ϵ, g_ϕ) where ϵ is a "source" random variable whose density function does not depend on ϕ, and g_ϕ is a mapping parameterized by ϕ such that $g_\phi(\epsilon)$ has q_ϕ as its density. This allows us to change the variable $z = g_\phi(\epsilon)$ and get

$$\begin{aligned}
\nabla_\phi L(\theta,\phi) &= \nabla_\phi \mathbb{E}_{z\sim q_\phi(z)}[\log p_\theta(x|z)p(z) - \log q_\phi(z)] && (12)\\
&= \nabla_\phi \mathbb{E}_{\epsilon\sim q(\epsilon)}[\log p_\theta(x|g_\phi(\epsilon))p(g_\phi(\epsilon)) - \log q_\phi(g_\phi(\epsilon))] && (13)\\
&= \nabla_\phi \int q(\epsilon)[\log p_\theta(x|g_\phi(\epsilon))p(g_\phi(\epsilon)) - \log q_\phi(g_\phi(\epsilon))]d\epsilon && (14)\\
&= \int q(\epsilon)\nabla_\phi[\log p_\theta(x|g_\phi(\epsilon))p(g_\phi(\epsilon)) - \log q_\phi(g_\phi(\epsilon))]d\epsilon && (15)\\
&= \mathbb{E}_{\epsilon\sim q(\epsilon)}\left[\nabla_\phi[\log p_\theta(x|g_\phi(\epsilon))p(g_\phi(\epsilon)) - \log q_\phi(g_\phi(\epsilon))]\right] && (16)
\end{aligned}$$

where the second line is due to the *Law of the unconscious statistician* (LOTUS). This allows us to obtain an unbiased gradient estimate by differentiating through the reparameterized random variable $z = g_\phi(\epsilon)$.

A classic example is Gaussian reparameterization. If $z \sim \mathcal{N}(\mu, \sigma^2)$, then $(\epsilon, \mu + \sigma \odot \epsilon)$ where $\epsilon \sim \mathcal{N}(0, I)$ is a reparameterization of z. This can be easily verified by looking at the characteristic function of $\mu + \sigma \odot \epsilon$ or by using using the change-of-variable formula to derive its density. A wider family of random variables can also be reparameterized by inverting the cumulative distribution function, a technique known as the *inverse transform sampling*.

3 Variational Autoencoders

The last component of VAE is *amortized inference*. Note that the marginal likelihood $\log p(x)$ we seek to approximate using the ELBO is wrt a particular data point x. For each data point x_i where i iterates through $1, ..., n$ of a training set of n data points, we need a variational distribution $q_i(z)$ to approximate the posterior $p(z|x_i)$ for the corresponding ELBO. This means the memory requirement for the variational parameters scales linearly. The idea of amortization is to introduce an encoder (a recognition network) which predicts the variational parameters of the variational distribution. In the case of Gaussian variational distribution, this amounts to

$$q_\phi(z|x) = \mathcal{N}(z; \mu_\phi(x), \sigma^2_\phi(x))$$

where μ_ϕ and σ^2_ϕ are outputs of the encoder and ϕ is the parameters of the encoder network.

[3] In theory, though, this is not always true. See section 3.1.2 of [2] for a counter example.

Putting things together, our goal is to maximize the ELBO by approximately following the gradient

$$\nabla_{\theta,\phi}L(\theta,\phi) = \nabla_{\theta,\phi}\left\{\mathbb{E}_z[\log p_\theta(x|z)] - D_{\mathrm{KL}}(q_\phi(z|x)||p(z))\right\} \tag{17}$$

$$= \nabla_{\theta,\phi}\left\{\mathbb{E}_\epsilon[\log p_\theta(x|\mu_\phi(x) + \sigma_\phi(x)\odot\epsilon)] - D_{\mathrm{KL}}(q_\phi(z|x)||p(z))\right\} \tag{18}$$

$$= \mathbb{E}_\epsilon[\nabla_{\theta,\phi}\log p_\theta(x|\mu_\phi(x) + \sigma_\phi(x)\odot\epsilon)] - \nabla_{\theta,\phi}D_{\mathrm{KL}}(q_\phi(z|x)||p(z)) \tag{19}$$

The first part can be estimated and is essentially a stochastic quadratic loss for reconstruction (since $p_\theta(x|z)$ is Gaussian). The second part can be computed analytically (since both $q_\phi(z|x)$ and $p(z)$ are Gaussian) which penalizes q for differing from the prior.

A Jensen's Inequality

In this appendix, I will provide a formal proof of Jensen's inequality. Technical details can be skipped but the main results are important and deserve more attention, because they are useful tools for deriving bounds. Recall the definition of convexity and strict convexity of a function:

Definition 1. *$f : X \to \mathbb{R}$ is convex if X is convex and:*

$$\forall x_1, x_2 \in X, \forall t \in [0,1]: \qquad f(tx_1 + (1-t)x_2) \le tf(x_1) + (1-t)f(x_2)$$

The following two properties are known as the 1st and 2nd order conditions for convexity.

Property 1. *A differentiable function f is convex iff $f(y) \ge f(x) + f'(x)(y-x)$ for all $x, y \in \mathrm{dom} f$.*

Property 2. *A twice differentiable function f is convex iff $f''(x) \ge 0$ for all $x \in \mathrm{dom} f$.*

Theorem 2. *(Jensen's inequality) Suppose $f : U \to \mathbb{R}$ is a convex function on an open interval U, $\mathbb{P}(X \in U) = 1$, $\mathbb{E}[|X|] < \infty$ and $\mathbb{E}[|f(X)|] < \infty$. Then $f(\mathbb{E}[X]) \le \mathbb{E}[f(X)]$.*

Proof. (Assuming differentiability) In general f does not need to be differentiable, but for simplicity I'll assume f is differentiable everywhere (for non-differentiable f we need to take the sub-derivative, which is a generalized notion of derivative).

Let $\mu = \mathbb{E}[X]$ and $m = f'(\mu)$. By the 1st order condition, we know $f(x) \ge f(\mu) + m(x - \mu)$ for all $x \in U$. Taking the expectation on both sides concludes the proof. □

Definition 3. *$f : X \to \mathbb{R}$ is **strictly** convex if:*

$$\forall x_1, x_2 \in X, x_1 \ne x_2, \forall t \in (0,1): \qquad f(tx_1 + (1-t)x_2) < tf(x_1) + (1-t)f(x_2)$$

Proposition 4. *If f is strictly convex, the equality in Jensen's inequality holds if and only if $X = \mathbb{E}[X]$ with probability one (*i.e. *X is essentially a constant).*

Proof. Again, assume f is differentiable for simplicity (otherwise take the sub-derivative) and let $m = f'(\mu)$ where $\mu = \mathbb{E}[X]$. Assume the equality in Jensen's holds. Let $l_\mu(X) = f(\mu) + m(X - \mu)$. We have $\mathbb{E}[f(X)] = \mathbb{E}[l_\mu(X)]$, and thus $\mathbb{E}[f(X) - l_\mu(X)] = 0$. The non-negativity of the integrand implies[4] $f(X) - l_\mu(X) = 0$ and thus $f(X) = f(\mu) - m(X - \mu) = l_\mu(X)$ almost surely. Note that $f(\mu) = l_\mu(\mu)$.

Let $A = \{\omega : f(X(\omega)) = l_\mu(X(\omega))\}$. We know that $\mathbb{P}(A) = 1$. Let $B = \{\omega : X(\omega) \neq \mu\}$. We want to show $\mathbb{P}(B) = 0$. Assume instead $\mathbb{P}(B) > 0$. Then

$$\mathbb{P}(B^c \cup A^c) \leq \mathbb{P}(B^c) + \mathbb{P}(A^c) = 1 - \mathbb{P}(B) < 1,$$

which implies $\mathbb{P}(B \cap A) = 1 - \mathbb{P}(B^c \cup A^c) > 0$. Thus $B \cap A$ must not be empty. For $\omega \in B \cap A$, we have

$$\begin{aligned} f\left(\frac{X(\omega) + \mu}{2}\right) &< \frac{f(X(\omega)) + f(\mu)}{2} \\ &= \frac{l_m(X(\omega)) + l_m(\mu)}{2} \\ &= l_m\left(\frac{X(\omega) + \mu}{2}\right) \\ &\leq f\left(\frac{X(\omega) + \mu}{2}\right) \end{aligned}$$

where the first line is due to the distinctiveness implied by $\omega \in B$ and strict convexity, the second line is because $\omega \in A$, the third line is because l_m is linear, and the last line is due to convexity. This is clearly a contradiction. Therefore $\mathbb{P}(B) = 0$, i.e. $X = \mu$ almost surely. The other direction is trivial. □

B Kullback–Leibler divergence

Definition 5. *Let q and p be probability density functions. The Kullback–Leibler (KL) divergence of q from p is defined as*

$$D_{\mathrm{KL}}(p||q) = \int_{-\infty}^{\infty} p(x) \log\left(\frac{p(x)}{q(x)}\right) dx \;=\; \mathbb{E}_{x \sim p(x)}\left[\log \frac{p(x)}{q(x)}\right]$$

Property 3. *(Asymmetry) In general, $D_{\mathrm{KL}}(p||q) \neq D_{\mathrm{KL}}(q||p)$.*

Property 4. *(Non-subadditivity) In general, $D_{\mathrm{KL}}(p||q) \not\leq D_{\mathrm{KL}}(q||r) + D_{\mathrm{KL}}(r||p)$ for arbitrary density r.*

Due to the above, KL divergence is not a metric (a distance function), but it satisfies the requirement to be a *statistical divergence* (which is a weaker notion of distance). This is due to the next property, which is a corollary of Jensen's inequality (since log is strictly concave).

Property 5. *(Positive-definitenss) $D_{\mathrm{KL}}(p||q) \geq 0$ in general, and $D_{\mathrm{KL}}(p||q) = 0$ if and only if $p = q$.*

[4] This is a standard measure-theoretic arugment. See exercise 1.4.1. of [1].

References

[1] Rick Durrett. *Probability: theory and examples*, volume 49. Cambridge university press, 2019.

[2] Yarin Gal. Uncertainty in deep learning. *University of Cambridge*, 1:3, 2016.

[3] Diederik P Kingma, Max Welling, et al. An introduction to variational autoencoders. *Foundations and Trends® in Machine Learning*, 12(4):307–392, 2019.

[4] Shakir Mohamed, Mihaela Rosca, Michael Figurnov, and Andriy Mnih. Monte carlo gradient estimation in machine learning. *arXiv preprint arXiv:1906.10652*, 2019.

[5] Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3-4):229–256, 1992.