The Variational Autoencoder

John Thickstun

We want to estimate an unknown distribution $p(x)$ given i.i.d. samples $x_i \in \mathcal{X} \sim p$. Given a parameterized family of densities p_θ, the maximum likelihood estimator is:

$$\hat{\theta}_{\text{mle}} \equiv \arg\max_\theta \mathop{\mathbb{E}}_{x\sim p} \log p_\theta(x). \tag{1}$$

One way to model the distribution $p(x)$ is to introduce a latent variable $z \sim r$ on an auxiliary space $\mathcal{Z}$ and a likelihood $p_\theta(x|z)$. Together, $r(z)$ and $p_\theta(x|z)$ define a joint distribution over $(x, z) \in \mathcal{X} \times \mathcal{Z}$. The marginal distribution $p_\theta(x)$ is given by

$$p_\theta(x) = \int_{\mathcal{Z}} p_\theta(x|z) r(z)\, dz. \tag{2}$$

Introducing a latent variable z can be convenient. We can sample from $p_\theta(x)$ by first sampling a latent $z \sim r$, and then sampling $x \sim p_\theta(\cdot|z)$. The idea is to use a simple prior distribution $r(z)$, e.g. $\mathcal{N}(0, I)$, and parameterize the conditional distribution $p_\theta(x|z)$ with an expressive model. In contrast to an autoregressive model, which requires d serial evaluations of the model to generate a sample $x \in \mathbb{R}^d$, a latent variable model can produce a sample with just one evaluation of the model that parameterizes $p_\theta(z|x)$. And whereas autoregressive models struggle to capture global structure over high dimension dimensional spaces $\mathcal{X}$, we hope that conditioning on a global code z might help promote global coherence.

Training Latent Variable Models

How do we train the conditional model $p_\theta(x|z)$ so that $p_\theta(x) \approx p(x)$? Direct evaluation of the marginal requires computation of the integral in Equation (2), which does not seem promising. Recall from our discussion of EM that we can construct a variational lower bound on $\log p_\theta(x)$ using a proposal distribution $q(z|x)$ as an importance sampler. By Jensen's inequality,

$$\log p_\theta(x) = \log \mathop{\mathbb{E}}_{z\sim q(\cdot|x)} \left[\frac{p_\theta(x,z)}{q(z|x)}\right] \geq \mathop{\mathbb{E}}_{z\sim q(\cdot|x)} \left[\log \frac{p_\theta(x,z)}{q(z|x)}\right] \equiv \mathop{\mathbb{E}}_{z\sim q(\cdot|x)} \mathcal{L}(x, z; \theta, q). \tag{3}$$

Equality holds exactly when $q(z|x)$ equals the posterior $p_\theta(z|x)$. The expectation on the right-hand side is referred to as the evidence lower-bound (ELBO) on the marginal likelihood of x.

The ELBO can be sharpened by optimizing over a minibatch of samples [Burda et al., 2016]:

$$\log p_\theta(x) = \log \mathop{\mathbb{E}}_{z_i\sim q(\cdot|x)} \left[\sum_{i=1}^{M} \frac{p_\theta(x,z)}{q(z|x)}\right] \geq \mathop{\mathbb{E}}_{z_i\sim q(\cdot|x)} \left[\log \sum_{i=1}^{M} \frac{p_\theta(x,z)}{q(z|x)}\right]. \tag{4}$$

The bound improves monotonically in M, and as $M \to \infty$, the bound becomes tight. The choice of M can represent a tradeoff between accuracy of the lower bound and the computational cost of evaluating it; see Cremer et al. [2017] for another interesting interpretation of Equation (4), and Rainforth et al. [2018] for some cautionary discussion about using tighter bounds.

Using a variational lower bound like Equation (3), we can rephrase the maximum likelihood objective (1) as

$$\hat{\theta}_{\text{MLE}} = \arg\max_{\theta} \sup_{q} \mathop{\mathbb{E}}_{\substack{x\sim p \\ z\sim q(\cdot|x)}} \mathcal{L}(x, z; \theta, q). \tag{5}$$

For simple latent variable models like GMM's, we can directly solve the inner optimization problem in Equation (5) by computing the posterior $p_\theta(z|x)$ for fixed parameters θ. In such settings, alternating maximization algorithms like EM are enticing. But in the present setting, we are interested in modeling rich distributions $p(x)$; modeling these distributions requires an expressive likelihood $p_\theta(x|z)$, and this comes at the expense of an analytically tractable posterior $p_\theta(z|x)$.

Instead, we estimate the posterior using another expressive, parameterized model $q_\varphi(z|x)$ and our optimization problem becomes [Kingma and Welling, 2014]

$$\hat{\theta}_{\text{MLE}} = \arg\max_{\theta} \sup_{\varphi} \mathop{\mathbb{E}}_{\substack{x\sim p \\ z\sim q_\varphi(\cdot|x)}} \mathcal{L}(x, z; \theta, \varphi). \tag{6}$$

This approach is sometimes called "amortized inference," because we replace an expensive computation of the posterior at each value of x with a function that is learned to approximate the posterior across all values of x. While this emphasizes the computational advantages of amortized inference, there is also evidence that amortization can have a positive regularizing effect on the maximum likelihood objective [Shu et al., 2018].

The parameterizations θ and φ are typically high-dimensional (e.g. the parameter spaces of large neural networks) and we are therefore interested in first-order optimization techniques for jointly optimizing Equation (6). We can directly compute gradients of the objective with respect to θ but, due to the expectation, gradients with respect to φ require more care.

Monte Carlo Gradient Estimators

Optimizing Equation (6) requires estimation of gradients with respect to φ. For a comprehensive survey of approaches to problems of this form, see Mohamed et al. [2019]. One generic approach to constructing these estimates uses the score function estimator [Rubinstein and Kreimer, 1983] a.k.a. REINFORCE [Williams, 1992]:

$$\nabla_\varphi \mathop{\mathbb{E}}_{z\sim q_\varphi(\cdot|x)} \big[h(z)\big] = \mathop{\mathbb{E}}_{z\sim q_\varphi(\cdot|x)} \big[h(z)\nabla_\varphi \log q_\varphi(z|x)\big]. \tag{7}$$

Taking $h(z) = \log p_\theta(x, z) - \log q_\varphi(z|x)$ in Equation (7) and replacing the expectation with a finite sum over samples yields an unbiased monte carlo estimate of the ELBO (3). In practice, the score function estimator exhibits prohibitively high variance that prevents its use in this setting.

We can construct an estimator with much lower variance by exploiting structure of the distribution $q_\varphi(z|x)$. When the parameterization of the density $q_\varphi(z|x)$ is given by a pushforward distribution $f_\varphi(x, \epsilon)$, we can rewrite the expectation over z as an expectation over ϵ. Because ϵ isn't a function of φ, the derivative passes through the expectation, allowing us to construct straightforward monte carlo estimates of the gradient:

$$\nabla_\varphi \mathop{\mathbb{E}}_{z\sim q_\varphi(\cdot|x)} \big[h(z)\big] = \mathop{\mathbb{E}}_{\epsilon} \big[\nabla_\varphi h(f_\varphi(x, \epsilon))\big]. \tag{8}$$

This simple change of variables goes by many names, in the recent literature it is commonly referred to as the pathwise estimator [Glasserman, 2013], the "re-parameterization trick" [Kingma and Welling, 2014], and stochastic backpropagation [Rezende et al., 2014].

The Gaussian VAE

The Gaussian Variational Autoencoder (VAE) proposed in Kingma and Welling [2014] sets a Gaussian prior $r(z) = \mathcal{N}(z; 0, I)$ and an additive Gaussian likelihood model $p_\theta(x|z) = \mathcal{N}(x; g_\theta(z), \sigma_\theta^2(z)I)$, where $g_\theta : \mathcal{Z} \to \mathcal{X}$ and $\sigma_\theta^2 : \mathcal{Z} \to \mathbb{R}$ are expressive (typically neural) parameterizations. You can think of this parameterization using Gaussians as an infinite Gaussian mixture model [Rasmussen, 2000], where the categorical class probabilities π_i are replaced with a continuous density $p(z)$, and the Gaussian at index z is parameterized by $\mu = g_\theta(z)$ and $\sigma^2 = \sigma_\theta^2(z)$. If $\sigma = \sigma_\theta^2(z)$ is constant in z, then the Gaussian VAE is reminiscent of a kernel density estimator with bandwidth σ [Rosenblatt et al., 1956, Parzen, 1962]; this should provide some intuition that the Gaussian VAE is sufficiently expressive to model richly structured distributions over data.

Unlike (finite) Gaussian mixture models, the posterior $p_\theta(z|x)$ of the VAE is intractable. Therefore, we will use a posterior estimate $q_\varphi(z|x)$ and optimize the lower bound (3) as presented in Equation (6). The Gaussian VAE uses $q_\varphi(z|x) = \mathcal{N}(z; f_\varphi(x), \Sigma_\varphi(x))$, where $f_\varphi : \mathcal{X} \to \mathcal{Z}$ and $\Sigma_\varphi : \mathcal{X} \to \mathcal{Z}^{\otimes 2}$ are expressive parameterizations of the conditional mean and variance of $q_\varphi(z|x)$. This is a simplifying assumption: there is no reason to suppose that the posterior $p_\theta(z|x)$ is realizable as a Gaussian, and this may contribute to common criticisms of VAE's including "bluriness."

Using Gaussian parameterizations $r(z) = \mathcal{N}(z; 0, I)$, likelihood $p_\theta(x|z) = \mathcal{N}(x; g_\theta(z), \sigma^2 I)$, and approximate posterior $q_\varphi(z|x) = \mathcal{N}(f_\varphi(x), \Sigma_\varphi(x))$, we can re-express the ELBO (3) in several ways:

$$\begin{aligned}
\mathcal{L}(x; \theta, q) &= \mathop{\mathbb{E}}_{z \sim q(\cdot|x)} [\log p_\theta(x|z) + \log p(z) - \log q_\varphi(z|x)] && (9) \\
&= \mathop{\mathbb{E}}_{z \sim q_\varphi(\cdot|x)} \left[\log p_\theta(x|z)\right] - D(q_\varphi(z|x) \parallel r(z)) && (10) \\
&= -\frac{\dim(\mathcal{X})}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \mathop{\mathbb{E}}_{z \sim q_\varphi(\cdot|x)} \|x - g_\theta(z)\|^2 - D(q_\varphi(z|x) \parallel r(z)). && (11)
\end{aligned}$$

And using the form given by Equation (11), we can re-express the optimization problem (6) as

$$\hat{\theta}_{\text{mle}} = \arg\min_\theta \inf_\varphi \mathop{\mathbb{E}}_{\substack{x \sim p \\ z \sim q_\varphi(\cdot|x)}} \left[\frac{1}{2\sigma^2} \|x - g_\theta(z)\|^2 + D(q_\varphi(z|x) \parallel r(z))\right]. \qquad (12)$$

The two terms in Equation (12) are referred to as the "reconstruction" and "divergence" terms respectively. In this form, there is a clear connection to auto-encoders [Vincent et al., 2010], where $f_\varphi(x)$ and $g_\theta(z)$ take the roles of "encoder" and "decoder" networks and the KL-divergence term acts as a regularizer. Conveniently, when both $q_\varphi(z|x)$ and $p(z)$ are parameterized by Gaussians, the divergence term can be computed analytically, although Roeder et al. [2017] argue that the monte carlo estimate used in Equation (9) could be better in some cases.

Evaluation

A direct way to evaluate a VAE is to report the lower bound:

$$\mathop{\mathbb{E}}_{z \sim q(\cdot|x)} \mathcal{L}(x, z; \theta, q) \approx \sum_{i=1}^{M} \log \frac{p_\theta(x, z_i)}{q(z_i|x)} \text{ where } z_i \sim q(z|x). \qquad (13)$$

If q effectively approximates the posterior $p_\theta(z|x)$ then Equation (13) will be a relatively accurate estimate of the log-likelihood. But it is conservative metric, in the sense that it is a lower bound the true log-likelihood, and will be an overly pessimistic evaluation of the quality of the model $p_\theta(x)$.

We can build a monte carlo estimate of the marginal likelihood (2) by importance sampling:

$$\log p_\theta(x) = \log \mathop{\mathbb{E}}_{z\sim q(\cdot|x)} \left[\frac{p_\theta(x,z)}{q(z|x)}\right] \approx \log \frac{1}{M}\sum_{i=1}^{M} \frac{p_\theta(x|z_i)r(z_i)}{q(z_i|x)}, \text{ where } z_i \sim q(z|x). \tag{14}$$

Note that this is the same approach as (4), using the one-point empirical estimate of the expectation. If $M = 1$ then (14) equals the one-point empirical estimate of the evidence lower bound (3) that we use to optimize the VAE. For any finite M (14) is an unbiased estimate of an upper bound on the log-likelihood. As $M \to \infty$, (14) converges to $\log p_\theta(x)$. How large does M need to be to get a good estimate? Somewhere around $M = 5,000$ is common, although this may be model-dependent Burda et al. [2016], Tomczak and Welling [2018].

In practice, the sum (14) is numerically unstable because the probabilities are all vanishingly small. To handle the instability, we replace the terms with their log, and evaluate the resulting stable log-sum-exp:

$$\log p_\theta(x) \approx \log \sum_{i=1}^{M} \exp\left(\log p_\theta(x|z_i) + \log r(z_i) - \log q_\varphi(z_i|x)\right) - \log M.$$

References

Yuri Burda, Roger Grosse, and Ruslan Salakhutdinov. Importance weighted autoencoders. *International Conference on Learning Representations (ICLR)*, 2016. (document)

Chris Cremer, Quaid Morris, and David Duvenaud. Reinterpreting importance-weighted autoencoders. *International Conference on Learning Representations Workshop Track*, 2017. (document)

Chris Cremer, Xuechen Li, and David Duvenaud. Inference suboptimality in variational autoencoders. *arXiv preprint arXiv:1801.03558*, 2018.

Paul Glasserman. *Monte Carlo methods in financial engineering*. Springer Science & Business Media, 2013. (document)

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *International Conference on Learning Representations (ICLR)*, 2014. (document)

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

Emanuel Parzen. On estimation of a probability density function and mode. *The annals of mathematical statistics*, 1962. (document)

Tom Rainforth, Adam R Kosiorek, Tuan Anh Le, Chris J Maddison, Maximilian Igl, Frank Wood, and Yee Whye Teh. Tighter variational bounds are not necessarily better. *International Conference on Machine Learning*, 2018. (document)

Carl Edward Rasmussen. The infinite gaussian mixture model. In *Advances in Neural Information Processing Systems*, 2000. (document)

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. *arXiv preprint arXiv:1401.4082*, 2014. (document)

Geoffrey Roeder, Yuhuai Wu, and David K Duvenaud. Sticking the landing: Simple, lower-variance gradient estimators for variational inference. In *Advances in Neural Information Processing Systems*, 2017. (document)

Murray Rosenblatt et al. Remarks on some nonparametric estimates of a density function. *The Annals of Mathematical Statistics*, 1956. (document)

YR Rubinstein and J Kreimer. About one monte carlo method for solving linear equations. *Mathematics and Computers in Simulation*, 1983. (document)

Rui Shu, Hung H Bui, Shengjia Zhao, Mykel J Kochenderfer, and Stefano Ermon. Amortized inference regularization. In *Advances in Neural Information Processing Systems*, 2018. (document)

Jakub Tomczak and Max Welling. Vae with a vampprior. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2018. (document)

Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, Pierre-Antoine Manzagol, and Léon Bottou. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of machine learning research*, 2010. (document)

Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 1992. (document)