ELBO surgery: yet another way to carve up the variational evidence lower bound

Matthew D. Hoffman
Adobe Research
mathoffm@adobe.com

Matthew J. Johnson
Google Brain
mattjj@google.com

Abstract

We rewrite the ELBO of VAEs in a way that highlights the role of the encoded data distribution . This perspective suggests that to improve our variational bounds we should improve our priors and not just the encoder and decoder.

1 Introduction

variational EM in models of the form

$$p_\theta(\boldsymbol{x}) = \int p_\theta(\boldsymbol{x} \mid \boldsymbol{z}) p(\boldsymbol{z}) \, \mathrm{d}\boldsymbol{z}, \tag{1}$$

where $p(\boldsymbol{z})$ is a prior on lvs $\boldsymbol{z} = \{z_n\}_n{}^N{}_{=1}$ and $p_\theta(\boldsymbol{x} \mid \boldsymbol{z})$ is a likelihood on observations $\boldsymbol{x} = \{x_n\}_{n=1}^N$ parameterized by θ . In particular, we focus on variational autoencoders (VAEs, also known as DLGMs) [1, 2], in which the prior and likelihood follow from the generative model

$$z_n \overset{\text{iid}}{\sim} \mathcal{N}(0, I), \qquad x_n \mid z_n \sim \mathcal{N}(\mu(z_n; \theta),\ \Sigma(z_n; \theta)), \qquad n = 1, 2, \ldots, N, \tag{2}$$

where the mean $\mu(z_n; \theta)$ and the covariance $\Sigma(z_n; \theta)$ depend on the latent variable z_n through a neural network with parameters θ. We write joint densities as products of independent densities,

$$p(\boldsymbol{z}) = \prod_n p(z_n), \qquad p_\theta(\boldsymbol{x} \mid \boldsymbol{z}) = \prod_n p_\theta(x_n \mid z_n). \tag{3}$$

The model is fit by maximizing the ELBO $\mathcal{L}$,

$$\log p_\theta(\boldsymbol{x}) = \log \int p_\theta(\boldsymbol{z}, \boldsymbol{x}) \, \mathrm{d}\boldsymbol{z} = \log \int q_\phi(\boldsymbol{z} \mid \boldsymbol{x}) \frac{p_\theta(\boldsymbol{z}, \boldsymbol{x})}{q_\phi(\boldsymbol{z} \mid \boldsymbol{x})} \, \mathrm{d}\boldsymbol{z} \geq \mathbb{E}_{q_\phi(\boldsymbol{z} \mid \boldsymbol{x})} \log \frac{p_\theta(\boldsymbol{z}, \boldsymbol{x})}{q_\phi(\boldsymbol{z} \mid \boldsymbol{x})} \triangleq \mathcal{L}(\theta, \phi), \tag{4}$$

where each term in the variational density $q_\phi(\boldsymbol{z} \mid \boldsymbol{x}) = \prod_n q_\phi(z_n \mid x_n)$ is a Gaussian in which the mean $\mu(x_n; \phi)$ and covariance $\Sigma(x_n; \phi)$ depend on the observation x_n through a neural network with free parameters ϕ. In these models, the variational distribution $q_\phi(z_n \mid x_n)$ acts as a stochastic "encoder" from an observation x_n to a distribution on the latent variable z_n, and the likelihood $p_\theta(x_n \mid z_n)$ acts as a stochastic "decoder" from the latent variable z_n to a distribution on the observation x_n.

There are several ways to rewrite the objective $\mathcal{L}(\theta, \phi)$, and each provides its own perspective.

Evidence minus posterior KL. One form of $\mathcal{L}(\theta, \phi)$ emphasizes that the lower bound becomes tighter as the variational distribution better approximates the posterior:

$$\mathcal{L}(\theta, \phi) = \log p_\theta(\boldsymbol{x}) - \mathrm{KL}(q_\phi(\boldsymbol{z} \mid \boldsymbol{x}) \,\|\, p_\theta(\boldsymbol{z} \mid \boldsymbol{x})). \tag{5}$$

Thus we can improve the ELBO by improving the model log evidence $\log p_\theta(\boldsymbol{x})$, through the prior $p(\boldsymbol{z})$ or the likelihood $p_\theta(\boldsymbol{x} \mid \boldsymbol{z})$, or by improving the variational posterior approximation $q_\phi(\boldsymbol{z} \mid \boldsymbol{x})$.

Average negative energy plus entropy. Another way to rewrite the ELBO is

$$\mathcal{L}(\theta,\phi) = \mathbb{E}_{q_\phi(\boldsymbol{z}\,|\,\boldsymbol{x})}[\log p_\theta(\boldsymbol{z},\boldsymbol{x})] + \mathbb{H}[q_\phi(\boldsymbol{z}\,|\,\boldsymbol{x})], \quad (6)$$

where the log joint $\log p_\theta(\boldsymbol{z},\boldsymbol{x})$ is interpreted as the negative energy in a Boltzmann distribution. Since we choose (θ,ϕ) to maximize the ELBO, this version highlights that a good posterior approximation $q_\phi(\boldsymbol{z}\,|\,\boldsymbol{x})$ must assign most of its probability mass to regions of low energy (i.e. high joint probability density) while also maximizing the entropy of $q_\phi(\boldsymbol{z}\,|\,\boldsymbol{x})$. This perspective is useful in contrasting variational EM with a MAP ; while MAP need only find a single value of $\boldsymbol{z}$ that maximizes the joint density (even if it lies in a region with very low posterior mass), the entropy term in the ELBO prevents $q_\phi(\boldsymbol{z}\,|\,\boldsymbol{x})$ from collapsing to an atom.

Average term-by-term reconstruction minus KL to prior. Finally, we can write

$$\mathcal{L}(\theta,\phi) = \frac{1}{N}\sum_{n=1}^{N} \mathbb{E}_{q_\phi(z_n\,|\,x_n)}\left[\log p_\theta(x_n\,|\,z_n)\right] - \mathrm{KL}(q_\phi(z_n\,|\,x_n)\,\|\,p(z_n)). \quad (7)$$

For each observation index n, this version has a reconstruction term for the nth observation and a KL divergence from each encoding distribution to the prior. We can interpret this KL-divergence term as a regularizer that is minimized when $q_\phi(z_n\,|\,x_n) = p(z_n)$ for all z; this perspective has been used to explain the tendency of variational EM to "prune out" many of the latent dimensions in z [e.g., 3].

Can we do more? This last decomposition is interesting, but it leaves an important question unanswered: what is a "reasonable" value for this KL-divergence term to take on? Ideally it would be small, but we do not want or expect it to approach 0, since that would imply that x_n and z_n were almost independent, whereas virtually all of our modeling power comes from strongly coupling x_n to z_n. So if the KL term is large, is that a sign of underfitting, overfitting, or neither?

In the following section, we show that this KL term can be further decomposed in terms of the *average encoding distribution*, which we define as

$$q_\phi^{\text{avg}}(z) \triangleq \frac{1}{N}\sum_{n=1}^{N} q_\phi(z\,|\,x_n). \quad (8)$$

In fact, we show that the *marginal* KL divergence $\mathrm{KL}(q_\phi^{\text{avg}}(z)\,\|\,p(z))$ is hidden in (and a major contributor to) the ELBO. This marginal KL is important because, unlike the individual terms $q_\phi(z_n\,|\,x_n)$, the average encoding distribution $q_\phi^{\text{avg}}(z)$ *can* be made arbitrarily close to the prior $p(z)$ without sacrificing model power. Indeed, if the data are drawn from the model, $x_n \sim p_\theta(x)$, and the posterior approximation is accurate, $q_\phi(z\,|\,x_n) \approx p_\theta(z\,|\,x_n)$, then for large N we would expect

$$p(z) = \int p_\theta(z\,|\,x)p_\theta(x)\,\mathrm{d}x = \mathbb{E}_{x\sim p_\theta(x)} p_\theta(z\,|\,x) \approx \frac{1}{N}\sum_n p_\theta(z|x_n) \approx \frac{1}{N}\sum_n q_\phi(z|x_n) = q_\phi^{\text{avg}}(z).$$

2 Rewriting the ELBO

In this section, we drop parameter subscripts to simplify the notation. To write the ELBO in a way that includes the average encoder distribution $q^{\text{avg}}(z)$, it is convenient to treat the index n as a random variable. While the manipulation is entirely algebraic, this treatment makes the steps simpler and the result more interpretable. In particular, define the joint densities

$$q(n,z) \triangleq q(n)q(z\,|\,n), \qquad q(z\,|\,n) \triangleq q(z\,|\,x_n), \qquad q(n) \triangleq \frac{1}{N}, \quad (9)$$

$$p(n,z) \triangleq p(n)p(z\,|\,n), \qquad p(z\,|\,n) \triangleq p(z), \qquad p(n) \triangleq \frac{1}{N}, \quad (10)$$

where $p(z)$ denotes a standard Gaussian prior density from $z \sim \mathcal{N}(0,I)$. Note that the average encoder distribution $q^{\text{avg}}(z)$ is now simply the marginal $q(z) = \sum_{n=1}^{N} q(z,n)$.

Using this notation, we can write the second term in the VAE objective (4) as

$$\frac{1}{N}\sum_{n=1}^{N} \mathrm{KL}(q(z_n\,|\,x_n)\,\|\,p(z_n)) = \mathrm{KL}(q(z)\,\|\,p(z)) + \left(\log N - \mathbb{E}_{q(z)}[\mathbb{H}[q(n\,|\,z)]]\right) \quad (11)$$

$$= \mathrm{KL}(q(z)\,\|\,p(z)) + \mathbb{I}_{q(n,z)}[n,z], \quad (12)$$

where $\mathbb{I}_{q(n,z)}[n,z] = \mathbb{E}_{q(n,z)}[\log \frac{q(n,z)}{q(n)q(z)}]$ denotes the mutual information of n and z in $q(n,z)$. To check this expression, write

$$\frac{1}{N}\sum_{n=1}^{N} \mathrm{KL}(q(z_n \mid x_n) \,\|\, p(z_n)) = \sum_n q(n,z) \log \frac{q(n,z)}{p(n,z)} \tag{13}$$

$$= \mathrm{KL}(q(z) \,\|\, p(z)) + \mathbb{E}_{q(z)}[\mathrm{KL}(q(n \mid z) \,\|\, p(n))] \tag{14}$$

$$= \mathrm{KL}(q(z) \,\|\, p(z)) + \left(\log N - \mathbb{E}_{q(z)}[\mathbb{H}[q(n \mid z)]]\right), \tag{15}$$

where the first equality can be checked by expanding $p(n,z)$ and $q(n,z)$ and canceling the $p(n)$ and $q(n)$ factors, the second equality follows from the chain rule and splitting the log, and the last line follows from using $p(n) = \frac{1}{N}$. To check the mutual information expression, write

$$\mathbb{I}_{q(n,z)}[n,z] = \mathbb{E}_{q(z)}\left[\mathbb{E}_{q(n \mid z)}\left[\log \frac{q(n \mid z)}{q(n)}\right]\right] = \log N - \mathbb{E}_{q(z)}\left[\mathbb{H}\left[q(n \mid z)\right]\right]. \tag{16}$$

Thus substituting the KL expression (12) into the ELBO (4), we can write the ELBO in three terms,

$$\mathcal{L}(\theta,\phi) = \underbrace{\left[\frac{1}{N}\sum_{n=1}^{N} \mathbb{E}_{q(z_n \mid x_n)}[\log p(x_n \mid z_n)]\right]}_{\text{① average reconstruction}} - \underbrace{\left(\log N - \mathbb{E}_{q(z)}[\mathbb{H}[q(n \mid z)]]\right)}_{\text{② index-code mutual info.}} - \underbrace{\mathrm{KL}(q(z) \,\|\, p(z))}_{\text{③ marginal KL to prior}}. \tag{17}$$

3 Qualitative perspectives

We can make several observations about the ELBO expression given in (17). First, the two terms ① and ② are in tension with each other because to get a good average reconstruction score for ①, we typically need each encoding z_n to be specific to its corresponding observation x_n and hence $q(n \mid z)$ should have low entropy. Term ② acts as a regularizer, in that it encourages the encodings $q(z \mid x_n)$ to overlap for distinct observations n, but this effect is likely to be weak relative to the reconstruction term ①. Interestingly, ② is bounded above and below, because

$$0 \leq \log N - \mathbb{E}_{q(z)}\mathbb{H}[q(n \mid z)] \leq \log N. \tag{18}$$

Empirically, we have found that reconstructions are very precise and, correspondingly, $q(z \mid n)$ is very concentrated relative to $q(z)$, resulting in ② is close to its maximum value of $\log N$.

Second, while $q(z)$ appears in all terms, $p(z)$ only appears in ③. Thus when considering choosing priors $p(z)$ to optimize the ELBO, only this term is affected. Observe that we could set ③ to zero without sacrificing model power by simply defining the prior to be $q(z)$. This choice would not be amenable to scalable computation because it is difficult to evaluate ② in isolation: to normalize $q(n \mid z)$ at each evaluation requires accessing all N observations (and the normalization also precludes us from making unbiased Monte Carlo estimates). Setting ③ to zero may also be undesirable due to the potential for overfitting or the inability to use the prior to sculpt the latent representation [4]. Nevertheless, because ③ can in principle be set to zero, whenever it is large it indicates a very strong and potentially unwanted regularization effect from the prior.

4 Basic empirical results

To get a sense for the new terms in (17), we fit a basic variational autoencoder to a binarized MNIST dataset. The encoder and decoder each had two hidden layers with 500 units each and used softplus nonlinearities, and we fit them using the Adam optimizer [5]. For more details, see the code.

After optimization, we estimated the marginal KL term ③ via Monte Carlo:

$$\mathrm{KL}(q(z) \,\|\, p(z)) \approx \frac{1}{S}\sum_{s=1}^{S} \log \frac{q(\hat{z}_s)}{p(\hat{z}_s)}, \quad \hat{z}_s \mid \hat{n}_s \overset{\text{iid}}{\sim} q(z \mid x_{\hat{n}_s}), \quad \hat{n}_s \overset{\text{iid}}{\sim} \mathrm{Unif}(\{1,2,\ldots,N\}), \tag{19}$$

	ELBO	Avg. KL	Mutual info. ②	Marg. KL ③
2D latents	-129.63	7.41	7.20	0.21
10D latents	-88.95	19.17	10.82	8.35
20D latents	-87.45	20.2	10.67	9.53

Table 1: Estimated values for ELBO terms on binarized MNIST. Note that the values in the average KL column, which are computed as $\frac{1}{N}\sum_n \mathrm{KL}(q(z_n \mid x_n) \,\|\, p(z_n))$, equal the sum of the corresponding mutual information and marginal KL terms.

for sample indices $s = 1, 2, \ldots, S$, which requires total time proportional to NS to compute. We also computed the average KL $\frac{1}{N}\sum_{n=1}^{N} \mathrm{KL}(q(z_n) \,\|\, p(z_n))$ analytically in the usual way and hence estimated the mutual information term ② by subtraction. As shown in Table 1, while the marginal KL term ③ could in principle be set to be very small, it still contributes to and significantly reduces the ELBO value for nontrivial dimension sizes. We also see that for nontrivial dimension sizes the mutual information term ② is near its maximum value of $\log N \approx \log(60000) < 11.0021$, indicating that the individual encoding distributions $q(z \mid x_n)$ do not have significant overlap.

These results confirm that our current encoder and decoder models (and optimizers) find it difficult to match $q(z)$ and $p(z)$. This issue has also been observed by Makhzani et al. [6], who address it by replacing the $\mathrm{KL}(q(z_n) \,\|\, p(z_n))$ term in the ELBO with an adversarial loss. But our theoretical analysis suggests that we need not abandon the principle of maximum (marginal) likelihood; if DLGMs find it difficult to produce unimodal Gaussian marginal posteriors, then perhaps we should investigate multimodal priors that can meet $q(z)$ halfway.

5 Conclusion

This new decomposition of the ELBO provides some new perspectives on the role of the prior and the encoded data distribution. In particular, we split the average KL term of (7) into an index-code mutual information term and a marginal KL term from the encoded data distribution to the prior, as in (17). Evaluating these terms separately, we found that for nontrivial latent dimension sizes the marginal KL term, while it could in principle could be made very small, has large detrimental impact on the ELBO. In addition, we found that the mutual information term seems to be maximized, which is consistent with intuition and suggests that to improve the ELBO value we should focus on improving the marginal KL term. This new ELBO decomposition also provides a computational diagnostic to evaluate when underfitting may be caused by a rigid prior that the encoder and decoder are unable to match. In future work it may prove fruitful to investigate alternative, multimodal priors that can "meet in the middle" with the encoder and decoder networks.

References

[1] Diederik P. Kingma and Max Welling. "Auto-Encoding Variational Bayes". In: *International Conference on Learning Representations* (2014).

[2] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. "Stochastic backpropagation and approximate inference in deep generative models". In: *International Conference on Machine Learning*. 2014.

[3] Yuri Burda, Roger Grosse, and Ruslan Salakhutdinov. "Importance weighted autoencoders". In: *International Conference on Learning Representations* (2016).

[4] Matthew Johnson, David Duvenaud, Alex Wiltschko, Sandeep Datta, and Ryan Adams. "Composing graphical models and neural networks for structured representations and fast inference". In: *Advances in Neural Information Processing Systems 29*. 2016.

[5] Diederik Kingma and Jimmy Ba. "Adam: A Method for Stochastic Optimization". In: *International Conference on Learning Representations*. 2015.

[6] Alireza Makhzani, Jonathon Shlens, Navdeep Jaitly, and Ian Goodfellow. "Adversarial Autoencoders". In: *International Conference on Learning Representations*. 2016.