ELBO Surgery: Yet another way to carve up the evidence lower bound

Matthew D. Hoffman (Adobe Research); Matthew J. Johnson (Google Brain)

Setup

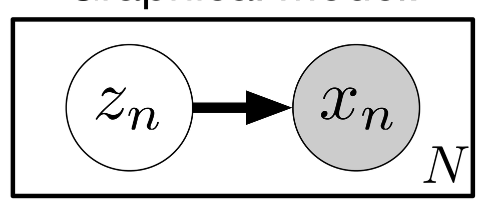
We're interested in variational EM in models of the form

$$p_\theta(\boldsymbol{x}) = \int p_\theta(\boldsymbol{x} \,|\, \boldsymbol{z}) p(\boldsymbol{z}) \, \mathrm{d}\boldsymbol{z}.$$

Fit by maximizing log evidence lower bound (ELBO) $\mathcal{L}$,

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

Graphical Model:



$$p(\boldsymbol{z}) = \prod_{n=1}^{N} p(z_n), \qquad p_\theta(\boldsymbol{x} \,|\, \boldsymbol{z}) = \prod_{n=1}^{N} p_\theta(x_n \,|\, z_n),$$

$$q_\phi(\boldsymbol{z} \,|\, \boldsymbol{x}) = \prod_{n=1}^{N} q_\phi(z_n \,|\, x_n).$$

Existing Perspectives on the ELBO

Evidence minus posterior KL

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

Emphasizes that ELBO is a lower bound that becomes tighter as the variational distribution better approximates the posterior.

Average negative energy plus entropy

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

Emphasizes that, unlike maximum a posteriori (MAP), a good posterior approximation must not only assign its probability mass to regions of low energy (high joint probability) but also try to maximize the entropy of $q_\phi(\boldsymbol{z} \,|\, \boldsymbol{x})$.

Average term-by-term reconstruction minus KL to prior

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

Emphasizes that the ELBO has an autoencoder's average reconstruction term as well as a KL divergence from each encoding distribution to the prior.

The Average Encoding Distribution

Consider the average encoding distribution,

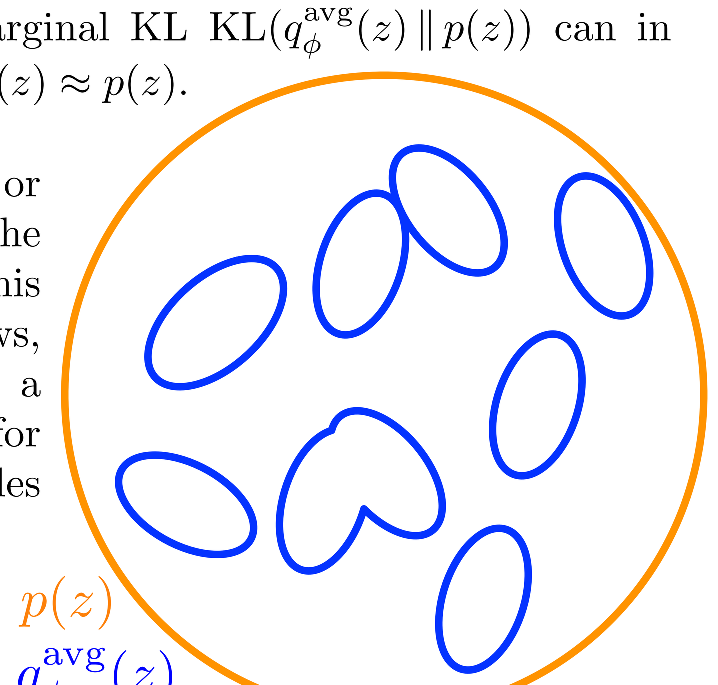
$$q_\phi^{\mathrm{avg}}(z) \triangleq \frac{1}{N} \sum_{n=1}^{N} q_\phi(z \,|\, x_n).$$

We would expect that if $x_n \sim p_\theta(x)$ and $q_\phi(z \,|\, x_n) \approx p_\theta(z \,|\, x_n)$, then for large N,

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

Indeed, unlike $\mathrm{KL}(q(z_n \,|\, x_n) \,\|\, p(z_n))$, the marginal KL $\mathrm{KL}(q_\phi^{\mathrm{avg}}(z) \,\|\, p(z))$ can in principle be made arbitrarily small, with $q_\phi^{\mathrm{avg}}(z) \approx p(z)$.

In practice, however, there may be large gaps or holes in the average encoding distribution. The cartoon at right shows one scenario where this could happen. As the latent dimension grows, we should expect it to get harder to fill up a large spherical space with many small blobs, for the same reason kernel density estimation scales poorly with dimension.



A New Rewrite of the ELBO

The average encoding distribution is hidden in the ELBO. To simplify notation, treat the index n as a random variable and define

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

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

Note that $q^{\mathrm{avg}}(z) = \sum_{n=1}^{N} q(z, n)$.

We can rewrite the KL to the prior in (1) as

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

where $\mathbb{I}_{q(n,z)}[n, z]$ denotes the mutual information of n and z in $q(n, z)$.

To check this expression, write

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

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 $p(n) = \frac{1}{N}$.

Substituting this new KL expression into the ELBO (1), we have

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

Observations

The three terms above encode three desiderata:

Term 1: This is a traditional autoencoder objective, which encourages accurate reconstructions of x given z.
Term 2: This is the (negative) mutual information between z and the index n. It penalizes models in which we can determine which observations x are consistent with which z vectors.
Term 3: This is the (negative) KL divergence between the average encoding distribution and the prior.

This decomposition sheds some new light on what the ELBO cares about:

• Terms 1 and 2 are in tension. To make term 1 large, we want z to tell us almost everything there is to know about x. But that often requires that n and z have high mutual information.

• Term 2 is bounded above and below: $0 \leq \log N - \mathbb{E}_{q(z)} \mathbb{H}[q(n \,|\, z)] \leq \log N$
In the case where reconstructions are very precise, we should expect term 2 to be near its maximum value of $\log N$.

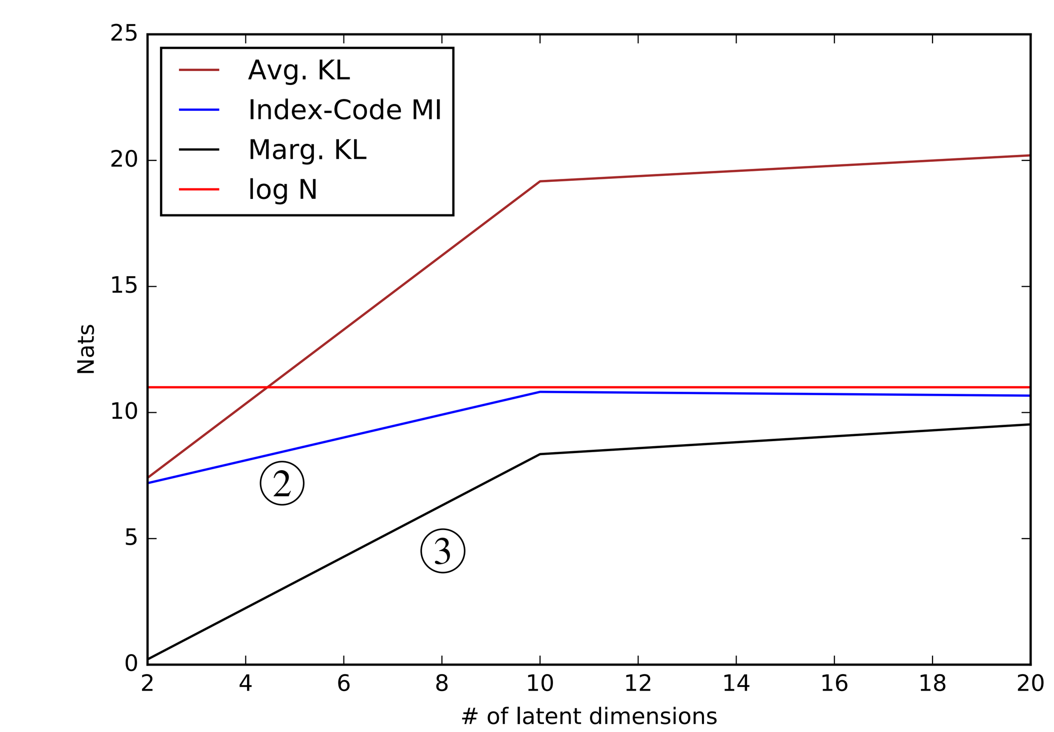
• $p(z)$ only appears in term 3, and term 3 can in principle be set to 0 for any model by setting $p(z) = q(z)$. This may be impractical or unwise, but it does imply that…

• When term 3 is large, our choice of prior $p(z)$ is regularizing our model (whether or not we want it to).

An Experiment

We fit basic variational autoencoder models with 2, 10, and 20 latent dimensions to a binarized 60000-image MNIST dataset, computed the average KL divergence from eq. 7, estimated term 3 by Monte Carlo, and estimated term 2 by subtracting our estimate of term 2 from the average KL. The results are plotted below.

We see that, for non-trivial latent spaces, term 2 (the mutual information between z and n) approaches its maximum value of $\log N$. We also see that term 3 (the marginal KL) makes a significant contribution to the ELBO, confirming that a simple encoder-decoder model has a hard time matching the marginal $q(z)$ to $p(z)$.



Food for Thought

• The results above suggest that deep latent Gaussian models have a hard time producing unimodal marginal posteriors. Perhaps we should investigate learning multimodal priors for $p(z)$ that meet $q(z)$ halfway?

• We could set $p(z) = q(z)$, but this choice is computationally impractical for large datasets, and may overfit badly. What's the right level of power for $p(z)$?

• DLGMs are powerful density estimators. Shouldn't a deeper DLGM be able to match $q(z)$?

• This analysis also applies to the non-variational case where $q(z \mid n) = p(z \mid x_n)$. What can this analysis tell us about latent-variable density estimation in general?

• Do the marginal posteriors of classical models (e.g., factor analysis) and more powerful flat models (e.g., mixtures of factor analyzers, latent Dirichlet allocation) do a better or worse job of matching their priors?

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.