The β-VAE's Implicit Prior

Matthew D. Hoffman
Google

Carlos Riquelme
Google Brain

Matthew J. Johnson
Google Brain

1 Introduction

Variational autoencoders [VAEs; 1, 2] are a popular and powerful class of deep generative models. They resemble a classical autoencoder, except that the encoding function is replaced with a *distribution* $q(z \mid x)$ over latent codes, and this distribution is regularized to have small KL divergence to a (usually pre-specified) marginal distribution $p(z)$. If the reconstruction log-likelihood $\mathbb{E}_q[\log p(x \mid z)]$ has the same weight as the KL-divergence penalty $\mathbb{E}_q[\log \frac{p(z)}{q(z|x)}]$, then the training procedure can be interpreted as maximizing a bound on the marginal likelihood $p(x)$ (sometimes called the evidence lower bound or ELBO):

$$\log p(x) = \log \int_z p(x,z)dz \geq \mathbb{E}_q[\log p(x \mid z)] - \mathrm{KL}(q(z \mid x) \,\|\, p(z)) \triangleq \mathcal{L}. \tag{1}$$

However, recent work has explored reweighting the KL-divergence term in $\mathcal{L}$, either to alleviate optimization issues during training [3] or to exert greater control over the sorts of latent spaces that get learned [4, 5]. Following Higgins et al. [4], we call VAEs fit with a reweighted KL term "β-VAEs".

In this paper, we analyze β-VAEs with the KL-divergence weight $\beta < 1$. We argue that optimizing this partially regularized ELBO is equivalent to doing approximate variational EM with an *implicit* prior $r(z)$ that depends on the marginal posterior $q(z) \triangleq \frac{1}{N}\sum_n q(z \mid x_n)$, with one main difference; it ignores the normalizing constant of this implicit distribution. We show how to estimate this missing normalizing constant.

2 Background

The β-VAE [4] rescales the KL-divergence term in the usual ELBO by a factor β:

$$\mathcal{L}_\beta \triangleq \frac{1}{N}\sum_n \mathbb{E}_q[\log p(x_n \mid z)] - \beta\mathrm{KL}(q(z \mid x_n) \,||\, p(z)). \tag{2}$$

This objective can no longer be interpreted as a bound on the log marginal likelihood $\log p(x) = \log \int_z p(x,z)dz$.

Hoffman et al. [6] show that the average KL-divergence term in (2) can be decomposed into a mutual-information term and a different KL-divergence term:

$$\frac{1}{N}\sum_n \mathrm{KL}(q(z \mid x_n) \,||\, p(z)) = \mathcal{I}(z;n) + \mathrm{KL}(q(z) \,||\, p(z)); \quad q(z) \triangleq \frac{1}{N}\sum_n q(z \mid x_n), \tag{3}$$

where $\mathcal{I}(z;n)$ denotes the mutual information between z and an index n into the training set; that is, $\mathcal{I}$ is how much information a sample from the mixture $q(z)$ gives us about which $q(z \mid x_n)$ z was sampled from. Hoffman et al. [6] argue that, in practice, the mutual-information term tends to saturate at its maximum value ($\log N$), and its impact on the optimization is therefore negligible. They also point out that one could set $\mathrm{KL}(q(z) \,||\, p(z)) = 0$ by replacing $p(z) = \mathcal{N}(0, I)$ with $p(z) = q(z)$, although this would lead to computational issues and overfitting. We will argue below that β-VAE training implicitly uses a q-dependent prior that interpolates between this extreme proposal and the default option of using a simple pre-specified prior.

Second workshop on Bayesian Deep Learning (NIPS 2017), Long Beach, CA, USA.

3 Main Result

We argue that optimizing the β-VAE objective resembles variational EM with an alternate prior

$$r(z) \triangleq \frac{q(z)^{1-\beta}p(z)^{\beta}}{C}; \quad C \triangleq \int_z q(z)^{1-\beta}p(z)^{\beta}dz, \tag{4}$$

where $q(z)$ is the marginal approximate posterior from equation (3). Plugging $r(z)$ into the usual ELBO, we get

$$\begin{aligned}\tfrac{1}{N}\textstyle\sum_n \log p(x_n) &\geq \tfrac{1}{N}\textstyle\sum_n \mathbb{E}_q[\log p(x_n \mid z)] - \mathbb{E}_q\left[\log \tfrac{r(z)}{q(z|x_n)}\right] \\ &= \tfrac{1}{N}\textstyle\sum_n \mathbb{E}_q[\log p(x_n \mid z)] - \mathcal{I}(z;n) + \mathbb{E}_q\left[\log \tfrac{q(z)^{1-\beta}p(z)^{\beta}}{q(z)}\right] - \log C \\ &= \tfrac{1}{N}\textstyle\sum_n \mathbb{E}_q[\log p(x_n \mid z)] - \mathcal{I}(z;n) + \mathbb{E}_q\left[\log \tfrac{p(z)^{\beta}}{q(z)^{\beta}}\right] - \log C \\ &= \tfrac{1}{N}\textstyle\sum_n \mathbb{E}_q[\log p(x_n \mid z)] - \mathcal{I}(z;n) - \beta\mathrm{KL}(q(z) \mid\mid p(z)) - \log C \\ &= \mathcal{L}_\beta - (1-\beta)\mathcal{I}(z;n) - \log C.\end{aligned} \tag{5}$$

That is, the β-VAE is optimizing the standard ELBO with the implicit prior $r(z)$, plus a term that effectively down-weights the mutual-information, minus the log-normalizer $\log C$ for $r(z)$.

4 The Log-Normalizer C

We now look in more detail at the log-normalizer C. Assume that $p(z) = \mathcal{N}(0, I)$, and

$$q(z \mid x_n) = \mathcal{N}(z; \mu(x_n), \Sigma(x_n)) \triangleq \mathcal{N}(z; \mu_n, \Sigma_n), \tag{6}$$

so that the marginal approximate posterior $q(z)$ is a mixture of N Gaussians. For example, we might have a VAE whose encoder outputs μ_n and Σ_n. Plugging this into the formula for C, we have

$$\begin{aligned}C &= \int_z \mathcal{N}(z;0,I)^{\beta}\left(\tfrac{1}{N}\textstyle\sum_n \mathcal{N}(z;\mu_n,\Sigma_n)\right)^{1-\beta}dz \\ &= \int_z \mathcal{N}(z;0,I)^{\beta}\left(\tfrac{1}{N}\textstyle\sum_n \tfrac{\phi(n;z)}{\phi(n;z)}\mathcal{N}(z;\mu_n,\Sigma_n)\right)^{1-\beta}dz \\ &\geq N^{\beta-1}\int_z \mathcal{N}(z;0,I)^{\beta}\textstyle\sum_n \phi(n;z)^{\beta}\mathcal{N}(z;\mu_n,\Sigma_n)^{1-\beta}dz,\end{aligned} \tag{7}$$

where $\phi(n;z)$ is an arbitrary function satisfying $\sum_n \phi(n;z) = 1$ and $\phi(n;z) > 0$, and the inequality follows from Jensen's inequality and the concavity of the function $x^{1-\beta}$ for $\beta < 1$. Forming a Lagrangian and taking derivatives with respect to $\phi(n;z)$ shows that the bound is tightest when $\phi^\star(n;z) \propto \mathcal{N}(z;\mu_n,\Sigma_n)$. That is, the optimal $\phi^\star(n;z)$ is the posterior over n for which $q(z \mid x_n)$ might have generated z. Plugging this result back into (7) makes the bound perfectly tight, i.e.,

$$C \equiv N^{\beta-1}\int_z \mathcal{N}(z;0,I)^{\beta}\textstyle\sum_n \phi^\star(n;z)^{\beta}\mathcal{N}(z;\mu_n,\Sigma_n)^{1-\beta}dz. \tag{8}$$

Now, suppose that $\mathcal{I}(z;n)$ is saturated at $\log N$. This implies that there is essentially no overlap between $\mathcal{N}(\mu_n,\Sigma_n)$ and $\mathcal{N}(\mu_{n'},\Sigma_{n'})$ for any $n \neq n'$. This implies in turn that, for virtually all values of z, $\phi^\star(n;z)$ has zero entropy. If this is true, then we can simplify (8):

$$\begin{aligned}C &= N^{\beta-1}\int_z \mathcal{N}(z;0,I)^{\beta}\textstyle\sum_n \phi^\star(n;z)^{\beta}\mathcal{N}(z;\mu_n,\Sigma_n)^{1-\beta}dz \\ &\geq N^{\beta-1}\int_z \mathcal{N}(z;0,I)^{\beta}\mathcal{N}(z;\mu_{n(z)},\Sigma_{n(z)})^{1-\beta}dz,\end{aligned} \tag{9}$$

where $n(z) \triangleq \arg\max_n \phi^\star(n;z)$. The inequality follows because we are substituting an indicator vector for the optimal $\phi^\star(n;z)$; if $\phi^\star(n;z)$ does in fact have approximately zero entropy, then the bound will be nearly tight. Taking this a bit further, we have

$$\begin{aligned}C &\geq N^{\beta-1}\int_z \mathcal{N}(z;0,I)^{\beta}\mathcal{N}(z;\mu_{n(z)},\Sigma_{n(z)})^{1-\beta}dz \\ &= N^{\beta-1}\textstyle\sum_n \int_{z|n(z)=n}\mathcal{N}(z;0,I)^{\beta}\mathcal{N}(z;\mu_n,\Sigma_n)^{1-\beta}dz \\ &\leq N^{\beta-1}\textstyle\sum_n \int_z \mathcal{N}(z;0,I)^{\beta}\mathcal{N}(z;\mu_n,\Sigma_n)^{1-\beta}dz.\end{aligned} \tag{10}$$

This new estimate will be relatively tight if every z with significant mass under the unnormalized distribution $\mathcal{N}(z;\mu_n,\Sigma_n)^{1-\beta}$ satisfies $n(z) = n$. Smaller values of β will encourage this in two ways: the exponent on the unnormalized distribution will be larger, and the lower weight on the KL term during training will encourage more concentrated posteriors.

Each of these N integrals is tractable. Suppressing the n subscript, we have

$$D(\mu, \Sigma) \triangleq \int_z \mathcal{N}(z; 0, I)^\beta \mathcal{N}(z; \mu, \Sigma)^{1-\beta} dz = \frac{|\Sigma'|^{\frac{1}{2}}}{|\Sigma|^{\frac{1-\beta}{2}}} \exp\left\{-\tfrac{1-\beta}{2}\mu^\top \Sigma^{-1}\mu + \tfrac{1}{2}\mu'^\top \Sigma'^{-1}\mu'\right\},$$

$$\Sigma' \triangleq (\beta I + (1-\beta)\Sigma^{-1})^{-1}; \quad \mu' \triangleq (1-\beta)\Sigma'\Sigma^{-1}\mu = \left(\tfrac{\beta}{1-\beta}\Sigma + I\right)^{-1} \mu. \tag{11}$$

This yields an approximation to C as $C \approx N^{\beta-1} \sum_n D(\mu_n, \Sigma_n)$. This is tractable to compute, but for intuition one can further simplify things if β and Σ are small, in which case

$$\Sigma' \approx \tfrac{1}{1-\beta}\Sigma; \quad \mu' \approx \mu; \quad D(\mu, \Sigma) \approx |\Sigma|^{\frac{\beta}{2}}; \quad C \approx N^{\beta-1} \textstyle\sum_n |\Sigma_n|^{\frac{\beta}{2}}. \tag{12}$$

4.1 Implications

Now we will argue that replacing $p(z)$ with $r(z)$ and optimizing the resulting ELBO leads to a degenerate solution. We will also argue that ignoring the log-normalizer term $\log C$ in this ELBO (as β-VAE training effectively does), is approximately equivalent to regularizing this problematic ELBO to prevent this degeneracy.

If the approximation in equation 12 holds, then the derivative of $\log C$ is approximately

$$\nabla_{\Sigma_n} \log C \approx \frac{\beta}{2} \frac{|\Sigma_n|^{\frac{\beta}{2}}}{\sum_m |\Sigma_m|^{\frac{\beta}{2}}} \Sigma_n^{-1} \triangleq w_n \tfrac{1}{N} \tfrac{\beta}{2} \Sigma_n^{-1} = \nabla_{\Sigma_n} \beta \tfrac{1}{N} \textstyle\sum_m -w_m \mathbb{E}_q[\log q(z \mid x_m)], \tag{13}$$

where $w_n \triangleq \frac{|\Sigma_n|^{\frac{\beta}{2}}}{\frac{1}{N}\sum_m |\Sigma_m|^{\frac{\beta}{2}}}$. That is, the derivative of $\log C$ with respect to Σ_n is equal to the derivative of β times a weighted average of the entropies of the N variational distributions. If all of the determinants are equal, then $w_n = 1$ for all n, and the gradient of $-\log C$ exactly cancels out the gradient of the entropy term in the KL divergence in equation 5. If they are not equal, then the gradient will push the determinants towards equality.

The log-normalizer has thus effectively removed any incentive for the optimizer to use variational distributions $q(z \mid x)$ that have nonzero variance. Since $\mathbb{E}_q[\log p(x, z)]$ is maximized by $q(z \mid x) = \delta(z - \arg\max_{z'} p(x, z'))$, the optimizer should set all variances to zero. In this case, the approximation in equation 12 becomes exact, implying that there are degenerate stable maxima of the implicit ELBO.

On the other hand, the analysis above implies that ignoring $\log C$ (i.e., doing simple β-VAE training) is roughly equivalent to adding a $-\frac{\beta}{N}\sum_n \mathbb{E}_q[\log q(z \mid x)]$ term to the implicit ELBO, regularizing the model towards solutions with higher-variance approximate posteriors. Insofar as one believes a priori that the marginal distribution on z should not be degenerate, this seems like a reasonable constraint to enforce.

We can therefore interpret β-VAE training as approximately optimizing the implicit ELBO obtained by replacing $p(z)$ with $r(z)$, and adding an entropy regularization term to avoid degenerate solutions.

5 Experiments

Figure 1 shows some samples from a β-VAE trained on statically binarized MNIST with $\beta = 0.75$, where the z vectors used to generate the samples were sampled either from $p(z)$ or the implicit prior $r(z)$. The samples from $r(z)$ look much better than those from $p(z)$. The bottom row demonstrates that $r(z)$ is not simply memorizing training examples.

Figure 2 shows the ELBO, KL divergence, and likelihood terms (evaluated using $r(z)$, not $p(z)$) of β-VAEs trained with values of $\beta \in [0.5, 1]$. The log-normalizer was estimated with the simple approximation in equation 9. The training ELBO is best for small β, but the generalization gap is very large. This is because $r(z)$ has overfit to the training data, and assigns low probability to z vectors that could generate held-out data. As β gets larger and $r(z)$ approaches $p(z)$, the gap narrows.

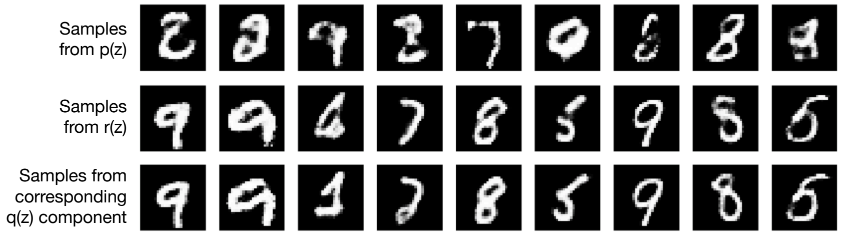

Figure 1: Samples and reconstructions from a β-VAE trained with $\beta = 0.75$. Top: Samples generated by drawing from $p(z) = \mathcal{N}(0, I)$. Middle: Samples generated by drawing from $r(z) \propto q(z)^{0.25} p(z)^{0.75}$. Bottom: Samples from an approximate posterior $q(z|x_n)$, where n is chosen to maximize the likelihood of the corresponding $z \sim r(z)$ above.

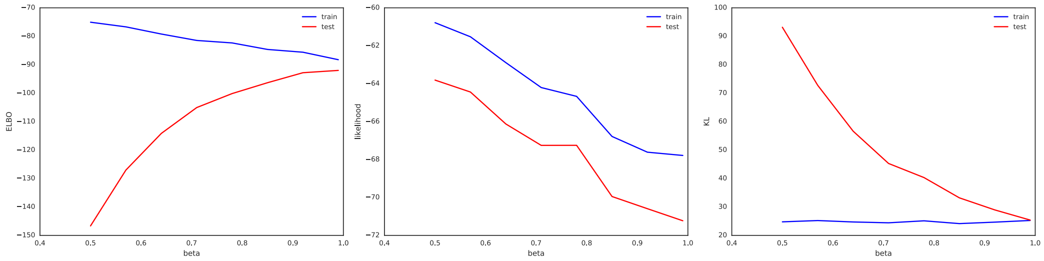

Figure 2: ELBO, likelihood, and KL of various β-VAEs, evaluated with implicit prior $r(z)$.

6 Conclusion

We have argued that training a β-VAE with $\beta < 1$ can be interpreted as optimizing an approximate log marginal likelihood bound under an alternative prior $r(z)$, regularized to prevent degeneracy. Even though this $r(z)$ is difficult to work with directly, we have derived some approximations to examine it.

There are two reasons we might be interested in $r(z)$. First, we may want to sample from it to do unconditional generation. This may be computationally challenging. But in analysis or conditional generation applications, we may be interested in $r(z)$ mostly for the *posterior* $p(z \mid x) \propto r(z)p(x \mid z)$ it induces. Our analysis suggests that, even if we never work with $r(z)$ explicitly, the encoder distribution $q(z \mid x)$ of a β-VAE can still learn to approximate this posterior.

It would be interesting to consider other weights on the entropy regularizer than β, since there is nothing in our analysis that suggests that it is a natural choice. We leave this investigation to future work.

We suspect this interpretation of β-VAE models and inference networks may enable new inference techniques and help us understand alternative training objectives.

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] Samuel R Bowman, Luke Vilnis, Oriol Vinyals, Andrew M Dai, Rafal Jozefowicz, and Samy Bengio. “Generating sentences from a continuous space”. In: *International Conference on Learning Representations*. 2016.

[4] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. “beta-VAE: Learning basic visual concepts with a constrained variational framework”. In: *International Conference on Learning Representations*. 2017.

[5] Anonymous. “An information-theoretic analysis of deep latent-variable models”. In: *International Conference on Learning Representations*. 2018.

[6] Matthew D Hoffman and Matthew J Johnson. “ELBO surgery: yet another way to carve up the variational evidence lower bound”. In: *Advances in Approximate Bayesian Inference (NIPS 2016 Workshop)*. 2016.