Empirical Bayes VAEs: Excess Risk Bound

Liam Ludington
ENS Paris Saclay
91190 Gif-sur-Yvette, France
whfludington@gmail.com

Elif Nebioglu
ENS Paris Saclay
91190 Gif-sur-Yvette, France
elifnebiogllu@gmail.com

Abstract

Variational Autoencoders (VAEs) excel at modeling high-dimensional data but have lacked deep theoretical insights until EBVAEs Tang and Yang [2021] introduced a learned latent prior. Their theory recommends optimizing the prior's covariance matrix to better capture data complexity. Following this approach, our MNIST experiments show that learning both an expressive prior and noise parameter significantly improves reconstruction and sample quality. We do not, however, explore broader datasets or more complex priors in this work.

Figure 1: Generation of MNIST digits with four types of EBVAEs.

1 Introduction

Many recent machine learning tasks deal with complex data, such as images or text. Variational Autoencoders (VAEs) help by using a single global inference network, rather than a separate inference procedure for each data point. However, when VAEs rely on a simple diagonal Gaussian prior, they may not fully capture the real structure of the data, causing underfitting. This happens because the chosen prior does not match the true complexity of the data (i.e., model misspecification).

To address these limitations, Empirical Bayes Variational Autoencoders (EBVAEs) introduce hyperparameters in the prior distribution over latent variables, allowing joint optimization of the prior, encoder, and decoder. This empirical Bayes approach enhances flexibility, leading to better density estimation and improved generative performance. The paper under discussion establishes a theoretical framework for analyzing the excess risk of EBVAEs in both parametric and nonparametric settings, building on prior results that characterize the approximation error of VAEs in simpler cases.

A notable strength of this work is how it directly informs practice. The theory presents clear ways to enhance VAEs—such as learning the prior's covariance matrix instead of relying on a fixed Gaussian. By making the prior more flexible, EBVAEs can better capture complex data patterns. Experiments on MNIST show that a more expressive prior noticeably boosts the quality of generated samples.

Building on these theoretical insights, we focus our experiments on MNIST to illustrate the benefits of learning a more expressive latent prior and noise parameter. Even in this relatively simple setting, our

Preprint. Under review.

results demonstrate that tuning the prior's covariance structure and the decoder's variance can improve both reconstruction fidelity and sample quality. This approach paves the way for extending EBVAEs to more challenging datasets and exploring additional forms of latent priors, which we plan to investigate in future work.

1.1 Notation

Here we explain the notation as well as some of the definitions used hereafter, which are mostly identical to the original paper. Data and latent variables live in $X \subseteq \mathbb{R}^{d_x}$ and $Z \subseteq \mathbb{R}^{d_z}$, respectively. The likelihood is written as $p(x|z)$, the encoder as $q(z|x)$, and the latent prior as $\pi(z)$. When these functions are parametrized as neural networks, we use $\theta \in \Theta_\theta$, $\phi \in \Theta_\phi$, and $\beta \in \Theta_\beta$ to denote the parameters that determine the functions $p_\theta(x|z)$, $q_\phi(z|x)$, and $\pi_\beta(z)$ in the decoder family $\mathcal{F}_\phi$, the encoder family $\mathcal{F}_\theta$, and the prior family $\mathcal{F}_\beta$, respectively. The distribution $p_D(x)$ defines the data distribution of x.

2 Empirical Bayes Variational Autoencoders

VAEs are a generative model that learn a model of their input data by learning and encoder that takes data to a distribution of latent codes and latent codes to a distribution in data space. Rather than learn the data distribution directly by optimizing the log evidence of their generated data, VAEs exploit the Evidence Lower Bound (ELBO) to indirectly optimize the evidence by simultaneously learning the likelihood and learning a model of the latent space that matches the posterior distribution. Thus the VAE tries to maximize the empirical ELBO

$$\frac{1}{n}\sum_{i=1}^{n}\left\{\int \log p(x_i|z)q(z|x_i)dz - D_{KL}\left(q(\cdot|x_i)\middle\|\pi(\cdot)\right)\right\}, \tag{1}$$

Vanilla VAEs make the assumption that the latent space is normally distributed with a diagonal covariance matrix, and is data independent. EBVAEs generalize VAEs by learning the parameters of a prior distributions $\pi_\beta(z)$ in a family of prior functions $\mathcal{F}_\beta$. The authors add a $\log(p_D)$ term for the sake of analysis to define the EBVAE loss as

$$m(p,q,\pi,x) = \log\frac{p_D(x)}{\int p(x|z)\pi(z)dz} + D_{KL}\left(q(\cdot|x)\middle\|\frac{p(x|\cdot)\pi(\cdot)}{\int p(x|z)\pi(z)dz}\right). \tag{2}$$

The data-dependent optimization problem can then be framed as finding the tuple

$$(\hat{p},\hat{q},\hat{\pi}) = \arg\min_{p\in\mathcal{F}_\phi, q\in\mathcal{F}_\theta, \pi\in\mathcal{F}_\beta}\left\{\frac{1}{n}\sum_{i=1}^{n} m(p,q,\pi,x_i)\right\}. \tag{3}$$

In line with the typical risk analysis, we care not only about finding the the minimum empirical risk but also the population-level risk given by

$$\Psi^* = \arg\min_{p\in\mathcal{F}_\phi, q\in\mathcal{F}_\theta, \pi\in\mathcal{F}_\beta}\mathbb{E}_{p_D(x)}\left[m(p,q,\pi,x)\right]. \tag{4}$$

The goal of the theoretical section of the paper is to provide bounds on the empirical risk of this optimization problem, that is, the difference between the population-level loss evaluated at the empirical minimizers of the EBVAE loss $m(\hat{p},\hat{q},\hat{p},x)$ and the population-level loss evaluated at the true minimizers of the EBVAE loss. Simply, the goal is to understand how well the optimization of the empirical loss can approximate the optimization of the loss over the whole dataset.

3 Excess Risk Bounds

In order to relate the expected loss of the empirical minimizers and the expected loss of the true minimizers, the authors use a tail bound on the distribution of the loss function given $p_D(x)$.

Assumption 1. *For a random variable X with density $p_D(x)$, there exist positive constants (α, D) such that:*

$$\sup_{p\in\mathcal{F}_\phi, q\in\mathcal{F}_\theta, \pi\in\mathcal{F}_\beta}\left\{\left|\log\frac{\int p(X|z)\pi(z)dz}{p_D(X)}\right| + D_{KL}\left(q(\cdot|X)\middle\|\frac{p(X|\cdot)\pi(\cdot)}{\int p(X|z)\pi(z)dz}\right)\right\} \leq D. \tag{5}$$

In essence, this assumption states that there exists α such that the distribution of the EBVAE loss is sub-exponential at level α for all permitted encoder, decoder, and prior functions. This bound is similar to other bounds in the literature on empirical risk bounds Grünwald and Mehta [2020] and valid for commonly used encoder and decoder families.

To bound the estimation error of the empirical minimizers, the authors use a data-dependent estimate of the complexity of the star hull $\bar{G}^*$ of G^*, the class of EBVAE loss functions shifted to have minimum value 0, defined formally for any $(p^*, q^*, \pi^*) \in \Psi^*$ as

$$G^* = \{g(x) = m(p, q, \pi, x) - m(p^*, q^*, \pi^*, x) \mid p \in F_{dd}, q \in F_{ed}, \pi \in F_{prior}\}.$$

and G^* as all nonnegative scalar multiples of those functions.

The definitions allow the authors to obtain the crucial estimate of the local Rademacher complexity Bartlett et al. [2005]

$$R_n(\delta, \bar{G}^*) = \mathbb{E}_{p_D(x),\, \varepsilon}\left[\sup_{g \in \bar{G}^*, \|g\|_2 \leq \delta} \left|\frac{1}{n}\sum_{i=1}^{n} \varepsilon_i g(x_i)\right|\right],$$

where ε_i are independent Rademacher random variables. Assuming that the Rademacher complexity of the start hull of G^* does not grow too rapidly with the ℓ^2 norm of its elements, as well as assumption 1, the authors show a high-probability bound on the empirical risk.

Theorem 1. *Consider the EBVAE estimator $\hat{p}, \hat{q}, \hat{\pi}$ defined in equation 3. **(Under Assumption 1)** Let δ_n be suitably small, and suppose there exists $(p^*, q^*, \pi^*) \in \Psi^*$. Then for the EBVAE estimator $(\hat{p}, \hat{q}, \hat{\pi})$, with sufficiently large n, the following inequality holds with high probability:*

$$\mathbb{E}_{p_D}\big[m(\hat{p}, \hat{q}, \hat{\pi}, x)\big] \;\leq\; \inf_{\gamma > 0}\Big\{ (1+\gamma) \min_{p,q,\pi} \mathbb{E}_{p_D}\big[m(p, q, \pi, x)\big] \;+\; c_2\Big(1 + \tfrac{1}{\gamma}\Big)\Delta_n\Big\},$$

where Δ_n is an error term depending on parameters such as n, D, and α. The EBVAE loss function $m(\cdot)$ is defined in terms of KL-based terms.

The bound on the empirical risk is a balance between the first term corresponding to the approximation error associated with the choice of encoder, decoder, and prior families, and the second term corresponding to the estimation error associated with the complexity of the function class G. The first term generally decreases with the complexity of the family of encoders, decoders, and priors, while the second term generally increases. Minimizing the empirical risk, as is common in statistical learn, thus requires balancing the expressivity of the function classes with their complexity.

4 Applications to VAEs

Having proved their main theoretical result, the authors then apply their theorem to specific cases of VAEs by calculating the approximation error and δ_n in Theorem 1 and upper-bounding their growth according to each model's characteristics. Specifically, they investigate the bound in the widely used case of VAEs with a Gaussian encoder and decoder model, that is, where $p_\theta(x|z) = \mathcal{N}(G_\theta(z), \sigma^I_{d_x})$ and $q_\phi(z|x) = N(\mu_\phi(x), \Sigma_\phi(x))$, where $G_\theta(z)$, $\mu_\phi(x)$ and $\Sigma_\phi(x)$ are functions parametrized by θ and ϕ, and σ is a trainable scalar value. In this case, no matter how expressive the class of encoder families, the approximation error is non-zero, since the posterior latent distribution may not be Gaussian. However, assuming that the data distribution can be written as the push-forward of the prior distribution plus some Gaussian noise $(G_{D\#}\pi_D) * N(0, \sigma \times 2I_{d_x})$, with the data-generating function G_D deterministic and invertible, the approximation error can be shown to vanish. The authors also condition on the regularity of the gradients of the family of priors.

Condition 1. *The family of prior $F_\beta = \{\pi_\beta(z)|\beta \in \Theta_\beta \subset \mathbb{R}^{d_\beta}\}$ has a compact parameter space Θ_β. In addition, there exist some constants (b_2, b_3) such that for any $\beta, \beta' \in \Theta_\beta$ and $z \in \mathbb{R}^{d_z}$:*

$$\|\beta\|_2 \leq b_2,\ |\log \pi_\beta(0)| \leq b_2,\ \|\nabla_z \log \pi_\beta(z)\|_2 \leq b_2(\|z\|_2 + 1),\ \textit{and}$$
$$|\log \pi_\beta(z) - \log \pi_{\beta'}(z)| \leq b(z)\|\beta - \beta'\|_2 \textit{ with } \|b(z)\|_2 \leq b_2(\|z\|_2^{b_3} + 1).$$

This condition requires that all priors in $\mathcal{F}_\beta$ behave like a mixture of Gaussians distribution. They also make the above-mentioned assumptions on the data distribution as well as an additional one the regularity of the generating function, which allow for parametrization by ReLU networks.

Assumption 2. ***B.1**: The data distribution $p_D = (G_{D\#}\pi_D) * N(0, \sigma^{*2} I_{d_x})$ where $\sigma_1 \leq \sigma^* \leq \frac{1}{2e}$, and $\pi_D(z)$ belongs to F_β. **B.2**: There exists an integer $k \geq 2$, so that $G_D(z) : \mathbb{R}^{d_z} \to \mathbb{R}^{d_x}$ is a C^k map, and there exists a C^k map $Q_D(x) : \mathbb{R}^{d_x} \to \mathbb{R}^{d_z}$ such that $Q_D \circ G_D(z) = z$.*

These assumptions and conditions allow the approximation error to vanish except for a term induced by the Gaussian noise and permit the bounding of the estimation error in terms of $\log(n)$.

Theorem 2. *Choose $\sigma_1 \in (0, \frac{1}{2e}]$, and consider decoder family $\mathcal{F}_\phi$ as described above with generating functions in class $\mathcal{F}_G$, Gaussian encoders $\mathcal{F}_\theta$, and priors $\mathcal{F}_\beta$ satisfying condition 1. Then there exists a choice of $\mathcal{F}_G$ and $\mathcal{F}_{\mu,\Sigma}$ so that for any target distribution p_D satisfying Assumption 2, the EBVAE estimator satisfies:*

$$\mathbb{E}_{p_D(x)}[m(\hat{p}, \hat{q}, \hat{\pi}, x)] \leq c_1 \sigma^{*2} \log^{\tilde{\alpha}_1} \frac{1}{\sigma^*} + c_2 \left(\frac{d_\beta + \sigma_1^{-\frac{2d_z}{k}}}{n\sigma_1^2} \right) \log^{2/\alpha} n \log^{\tilde{\alpha}_2} \frac{1}{\sigma_1}. \tag{6}$$

where $\tilde{\alpha}_1 = \frac{28+10\alpha+3\alpha^2}{\alpha^2}$ and $\tilde{\alpha}_2 = 2/\alpha + d_z/(\alpha(k-1)) + d_z/2 + 6$.

As the noise σ^* goes to zero, the invertibility of G_D ensures that the posterior distribution $p(z|x)$ is concentrated around $Q_D(x)$. By applying the Taylor expansion of G_D at $Q_D(x)$, one finds that the posterior distribution is approximately Gaussian with covariance:

$$\sigma^{*2} \left(\nabla G_D(z)^T \nabla G_D(z) \right)^{-1} \Big|_{z=Q_D(x)}$$

which may not be diagonal. Note that in theorem 2 we have only assumed that the encoder family parametrizes a multivariate Gaussian with general covariance matrix $\Sigma_\phi(x)$. Ignoring the off-diagonal elements of the posterior covariance matrix introduces additional error to the approximation error and decreases the quality of the EBVAE estimator. Thus it is beneficial for the empirical risk to also model the entire covariance matrix as $\tilde{\Sigma}_\theta(x)^T \tilde{\Sigma}_\theta(x) + \epsilon I_{d_z}$. The theory here also indicates the importance of correctly modeling the decoder parameter σ during the optimization process.

5 Experiments

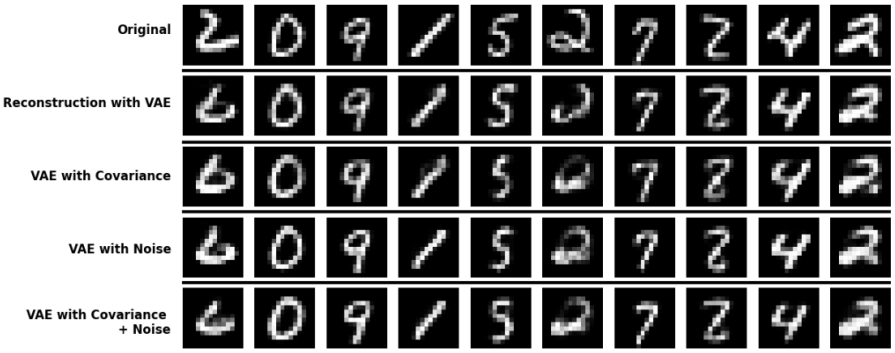

Figure 2: Reconstruction of MNIST digits with four types of EBVAEs.

In this section, we evaluate the ability of VAEs to reconstruct and generate data when modeling the full covariance matrix of the latent distribution with the encoder as well as the variance σ of the generative process with the decoder. A one-hidden-layer encoder and inverse decoder are trained with a latent dimension of eight using the classic VAE loss. While standard VAEs for MNIST often parameterize their output distribution with a Bernoulli parameter per pixel, we instead use a quantized Gaussian with support on the integers 0 to 255, where the model output represents the mean for each pixel, and optionally, a learned variance for each image. The latent distribution is assumed to be Gaussian, ensuring that the EBVAE loss formula remains consistent with the standard VAE loss formula. To improve training efficiency, the dataset is downsampled to 1000 images, reducing the resolution from 28×28 to 14×14 pixels.

Four types of EBVAEs are trained, all of which resemble standard VAEs since their latent prior is Gaussian. The first is a standard VAE. The second learns a covariance matrix $\tilde{\Sigma}_\theta(x)^T \tilde{\Sigma}_\theta(x) + \epsilon I_{d_z}$ for the latent dimension. The third learns the noising parameter σ_ϕ of the generative process, such that each decoded pixel (i, j) follows the distribution $\tilde{\mathcal{N}}(G_D(z)_{i,j}, \sigma_\phi(z))$, where $\tilde{\mathcal{N}}$ is the quantized normal distribution mentioned above. The fourth VAE learns both the covariance matrix and the noising parameter.

6 Results

All models achieve qualitatively good reconstruction abilities, as shown in Figure 2. Learning the covariance matrix decreases the reconstruction loss of both the standard VAE and the VAE that learns the noising coefficient. Learning the noising parameter improves the reconstruction loss of both the standard model and the model with learned covariance. It is possible that learning the covariance matrix deregularizes the latent space, thus increasing the reconstruction error. Meanwhile, learning the noising coefficient for each image helps control the width of the Gaussian distribution, leading to improved reconstruction loss.

Learning the covariance matrix and noise have opposite effects on KL divergence. Models with learned covariance exhibit better KL divergence, meaning their latent space is more Gaussian, than their counterparts without learned covariance. In contrast, models with learned noise show less regularized latent spaces compared to those without learned noise. While learning the covariance matrix might help enforce a more Gaussian latent space, the effect of learning the noising parameter on latent regularization remains unclear.

The model that learns the noising parameter in the decoder achieves the best EBVAE loss. Qualitatively, the standard VAE performs best at generating samples from normal latent input in terms of accuracy, diversity, and interpolation, with the noise-learning VAE performing similarly. The two models that learn covariance appear to suffer from an inadequately regularized latent space, leading to generated outputs that are overly similar.

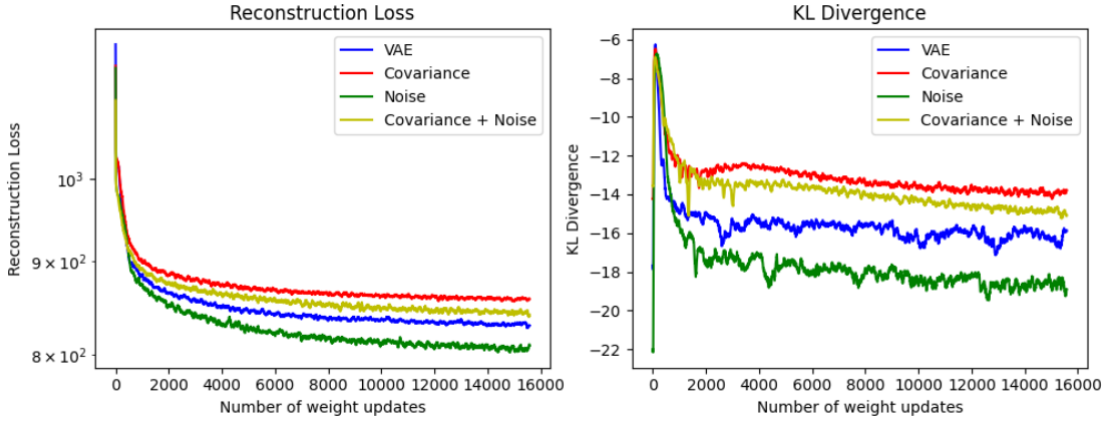

Figure 3: Reconstruction Loss and KL Divergence for four types of EBVAEs.

7 Discussion

This work investigates a paper that establishes oracle-style bounds on the empirical risk of EBVAEs, a flexible Bayesian generalization of VAEs. The theoretical recommendations for minimizing population risk between the empirical minimizer of the VAE loss and the true minimizer within a function class are implemented in four types of VAEs. While learning the noising coefficient in the decoder improves the EBVAE loss, learning a covariance matrix for the latent space significantly worsens it, providing only a marginal improvement in KL divergence. Possible explanations include that the training regime did not allow models to reach empirical optimizers or that the theory does not generalize well to the quantized Gaussians used in our decoding model.

Future directions include applying this theory to a fully Gaussian decoder on a different dataset and extending it to scenarios where the prior is non-Gaussian, such as a mixture of Gaussians for MNIST. These findings highlight the role of statistical learning theory in guiding practical design choices for VAEs, and further research into statistical guarantees for modern generative models remains an exciting avenue of exploration.

References

Peter L Bartlett, Olivier Bousquet, and Shahar Mendelson. Local rademacher complexities. 2005.

Peter D Grünwald and Nishant A Mehta. Fast rates for general unbounded loss functions: From erm to generalized bayes. *Journal of Machine Learning Research*, 21(56):1–80, 2020.

Rong Tang and Yun Yang. On empirical bayes variational autoencoder: An excess risk bound. In *Conference on Learning Theory*, pages 4068–4125. PMLR, 2021.