ρ-VAE: Autoregressive parametrization of the VAE encoder

Sohrab Ferdowsi, **Shideh Rezaeifar**, **Maurits Diephuis**, and **Slava Voloshynovkiy**

Department of Computer Science, University of Geneva, Switzerland
{sohrab.ferdowsi, shideh.rezaeifar, maurits.diephuis, svolos}@unige.ch

Abstract

We make a simple, but very effective alteration to the VAE model. This is about a drop-in replacement for the (sample-dependent) approximate posterior to change it from the standard white Gaussian with diagonal covariance to the first-order autoregressive Gaussian. We argue that this is a more reasonable choice to adopt for natural signals like images, as it does not force the existing correlation in the data to disappear in the posterior. Moreover, it allows more freedom for the approximate posterior to match the true posterior. This allows for the reparametrization trick, as well as the divergence term to still have closed-form expressions, obviating the need for its sample-based estimation. Although providing more freedom to adapt to correlated distributions, our parametrization has even less number of parameters than the diagonal covariance, as it requires only two scalars, ρ and s, to characterize correlation and a scaling factor, respectively. As validated by the experiments, our proposition noticeably and consistently improves the quality of image generation in a plug-and-play manner, needing no further parameter tuning, and across all setups. The code to reproduce our experiments is available at `https://github.com/sssohrab/rho_VAE/`.

1 Introduction

Arguably, one of the most successful approaches to generative and representation learning is that of "Auto-encoding variational Bayes" [1]. Considering a latent-based model for the data, where an underlying but hidden variations are assumed to be responsible for the creation of the observed data, this approach realizes the standard variational Bayes in the form of a neural network and offers a practical recipe for end-to-end learning of its parameters, while providing effective approximation of the intractable posterior. This has then given rise to the very popular Variational AutoEncoder (VAE), a set of models successful at generating images (e.g., see [] and [] among others), as well as learning useful representation without supervision(e.g., as in [] and []).

Essentially, the VAE bridges the tasks of generation of the data from latent codes, with that of inferring the latent codes from the data as two parts of the same body, the decoder and the encoder of an autoencoder architecture, respectively. This requires an implicit statistical model for each of these parts.

As for the decoder, the standard model is a white Gaussian distribution, centered on the latent codes when passed through the decoder. To provide higher capacity and hence matching better with natural images, this is then generalized to autoregressive models which bring about better performance, e.g., as in [], albeit adding to the computational burden.

The encoder part, however, is more delicate to treat. From one hand, it has to be realistic enough to match the true posterior and hence tighten the variational bound. From the other hand it should be

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

kept simple to make the gradient-based optimization feasible. This latter requirement has multiple sides: Firstly, it is preferred to have a closed-form expression for the regularization of the posterior to push it closer to the prior. Secondly, since the link between the encoder and the decoder cannot be direct, as explicit generation of latent codes breaks the differentiability, the encoder should be parametrized such that it can be injected to the latent space through reparametrization trick.

2 The VAE models

Here we briefly review the standard VAE model, highlighting aspects relevant to our work, as well as some of its further developments.

In a typical probabilistic model where a latent variable $\mathbf{z} \in \Re^d$ is the underlying factor to generate the observable samples $\mathbf{x}$'s $\in \Re^n$, the standard variational Bayes [?] paradigm is concerned with finding an approximation $q(\mathbf{z})$ for the intractable posterior $p(\mathbf{z}|\mathbf{x})$. This is achieved by minimizing the Kullback-Leibler divergence between these two distributions , i.e., $D_{\text{KL}}\Big[q(z)||p(z|x)\Big]$. Standard treatments of this quantity, along with its non-negativity property will then amount to the following inequality:

$$\log(p(\mathbf{x})) \leqslant \mathbb{E}_{q(\mathbf{z})}\Big[\log(p(\mathbf{x}|\mathbf{z}))\Big] - D_{\text{KL}}\Big[q(\mathbf{z})||p(\mathbf{z})\Big]. \tag{1}$$

2.1 The standard VAE

Autoencoding variational Bayes [1] is then constructing an explicit dependence of the latent variables to the i^{th} training sample by considering a parametrized distribution $q_\phi(\mathbf{z}^{(i)}|\mathbf{x}^{(i)})$ for the approximate posterior, whose construction resembles the encoder part of an autoencoder network with a set of learnable weights ϕ. Furthermore the training samples can be decoded with $p_\theta(\mathbf{x}^{(i)}|\mathbf{z}^{(i)})$, another network with parameters symbolized as θ.

Making this double-sided data dependence more explicit, and by summing over all N training samples results to the following inequality:

$$\frac{1}{N}\sum_{i=1}^{N}\log(p(\mathbf{x}^{(i)})) \leqslant \frac{1}{N}\sum_{i=1}^{N}\Big[\log(p_\theta(\mathbf{x}^{(i)}|\mathbf{z}^{(i)})) - D_{\text{KL}}\big[q_\phi(\mathbf{z}^{(i)}|\mathbf{x}^{(i)})||p(\mathbf{z})\big]\Big]. \tag{2}$$

This, in fact, is highly relevant for generative modeling as the marginal log-likelihood of the training samples will be upper bounded by two terms, both of which amenable to mini-batch optimization with stochastic gradient descent.

During optimization, the first term of the LHS can be considered as a data fidelity term, minimized e.g., in the ℓ_2 sense, since a natural choice for the decoder is $p_\theta(\mathbf{x}^{(i)}|\mathbf{z}^{(i)}) = \mathcal{N}\Big(\big(\mathbf{x}^{(i)}|\mathbf{z}^{(i)}\big), \sigma^2\mathrm{I}_n\Big)$, where I_n is the n-dimensional unity matrix.

The second term, from the other hand, can be interpreted as a regularization term, pushing the approximate posterior to a prior imposed on the latent space, most conveniently a simple $p(z) = \mathcal{N}(\mathbf{0}, \mathrm{I}_d)$. Provided that the optimization is successful, and the inequality (**??**) is tight, one can generate random samples from this prior, pass it through the learned decoder and generate samples (non-trivially) similar to the underlying data.

However, the above scenario comes with a major caution: the fact that sampling $\mathbf{z}$ from $q_\phi(\mathbf{z}|\mathbf{x})$ is a non-differentiable operation. The work-around for this issue is the wise "reparametrization trick", as proposed in [1].

The idea is to create the required randomness from a fixed distribution $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathrm{I}_d)$. The samples of the appropriate distribution can then be generated by injecting the learnable moments, e.g., using $\mathbf{z}^{(i)} = \boldsymbol{\mu}^{(i)} + \tilde{\mathrm{C}}^{(i)}\boldsymbol{\epsilon}$, where $\boldsymbol{\mu}^{(i)}$ is the mean vector of the posterior learned for the i^{th} sample and $\tilde{\mathrm{C}}^{(i)}$ is the Choleskiy decomposition of the corresponding covariance matrix $\mathrm{C}^{(i)}$.

This then limits the practical choices for $\mathrm{C}^{(i)}$ to have analytical Choleskiy decomposition forms, since both $\mathrm{C}^{(i)}$ and $\tilde{\mathrm{C}}^{(i)}$ participate in the optimization simultaneously.

Another issue to address is the calculation of $D_{\mathrm{KL}}\big[q_\phi(\mathbf{z}^{(i)}|\mathbf{x}^{(i)})||p(\mathbf{z})\big]$, which we elaborate slightly more in section 2.2. In order to avoid many practical difficulties, the standard choice is to pick a closed-form expression for it, hence further limiting the choices of $\mathrm{C}^{(i)}$.

While the prior distribution is chosen as $p(z) = \mathcal{N}(\mathbf{0}, \mathrm{I}_d)$, considering the above two constraints, the standard choice widely adopted in many further variants for the sample-wise approximate posterior is to set $\mathrm{C}^{(i)}_{(\mathbf{s})} = \mathrm{diag}\big(\mathbf{s}^{(i)}\big)$. In other words, $q_\phi(\mathbf{z}^{(i)}|\mathbf{x}^{(i)}) = \mathcal{N}\Big(\boldsymbol{\mu}^{(i)}, \mathrm{diag}\big(\mathbf{s}^{(i)}\big)\Big)$, a diagonal Gaussian distribution parametrized by the pair$(\boldsymbol{\mu}^{(i)}, \mathbf{s}^{(i)})$.

Note that now, the reparametrization trick can run smoothly, since the Choleskiy decomposition has a closed expression as $\tilde{\mathrm{C}}^{(i)}_{(\mathbf{s})} = \mathrm{diag}\big(\sqrt{\mathbf{s}^{(i)}}\big)$. Furthermore, the regularization term $D_{\mathrm{KL}}\Big[q_\phi(\mathbf{z}^{(i)}|\mathbf{x}^{(i)})||p(\mathbf{z})\Big]$ is also calculated analytically as:

$$D_{\mathrm{KL}}\Big[\mathcal{N}\Big(\boldsymbol{\mu}^{(i)}, \mathrm{diag}\big(\mathbf{s}^{(i)}\big)\Big)\Big|\Big|\mathcal{N}\big(\mathbf{0}, \mathrm{I}_d\big)\Big] = \frac{1}{2}\Big[\mathbf{1}_d^T\mathbf{s}^{(i)} + ||\boldsymbol{\mu}^{(i)}||_2^2 - d - \mathbf{1}_d^T \log\big(\mathbf{s}^{(i)}\big)\Big], \quad (3)$$

where $\mathbf{1}_d$ is the unity vector of dimension d, $||\cdot||_2^2$ is the squared ℓ_2-norm, and $\log\big(\mathbf{s}^{(i)}\big)$ is applied element-wise.

While this is a very practical choice, we argue in section 3 that it is too simplistic, as it disregards any correlation within dimensions.

2.2 Further variations

The literature around VAE is immense and still very active. Without aiming for any comprehensive literature review, here we still point out several of its variants.

As categorized in [2], VAE variants come in 3 categories.

3 The ρ-VAE

We saw in section 2.1 that two considerations limit the choices of approximate posterior: the need for a parametric Choleskiy factorization of its covariance matrix, as well as closed-form expression for the regularization term of (2), which basically requires the expression of log-determinant of the covariance.

In spite of the general consensus to pick $\mathrm{C}^{(i)}_{(\mathbf{s})} = \mathrm{diag}\big(\mathbf{s}^{(i)}\big)$, which does not allow any correlation between the dimensions of the approximate posterior, this work proposes another parametrization that grants such freedom, satisfies the above-mentioned restrictions, and yet has less number of parameters.

In particular, we chose a first-order autoregressive covariance which is characterized by a scaling factor s, and another scalar ρ to control the level of correlation, hence the term ρ-VAE. This has the form of a simple symmetric Toeplitz matrix as the following:

$$\mathrm{C}_{(\rho,s)} = s \times \mathrm{Toeplitz}\Big([1, \rho, \rho^2, \cdots, \rho^{d-1}]\Big) = s \begin{bmatrix} 1 & \rho & \rho^2 & \rho^3 & \cdots & \rho^{d-1} \\ \rho & 1 & \rho & \rho^2 & \cdots & \rho^{d-2} \\ \rho^2 & \rho & 1 & \rho & \cdots & \rho^{d-3} \\ \rho^3 & \rho^2 & \rho & 1 & \cdots & \rho^{d-4} \\ \vdots & & & \ddots & \ddots & \vdots \\ \rho^{d-1} & \cdots & \rho^3 & \rho^2 & \rho & 1 \end{bmatrix}, \quad (4)$$

where s is a positive scalar, and the correlation parameter is bounded as $-1 < \rho < +1$.

The determinant for this matrix can be calculated as [3]:

$$\det\big(\mathrm{C}_{(\rho,s)}\big) = s^d(1-\rho^2)^{d-1}, \tag{5}$$

based on which we can derive the regularization term of the loss function as:

$$D_{\mathrm{KL}}\Big[\mathcal{N}\Big(\boldsymbol{\mu}^{(i)}, \mathrm{C}_{(\rho,s)}\Big)\Big|\Big|\mathcal{N}\big(\mathbf{0}, \mathrm{I}_d\big)\Big] = \frac{1}{2}\Big[||\boldsymbol{\mu}^{(i)}||_2^2 + d(s-1-\log{(s)}) - (d-1)\log{(1-\rho^2)}\Big]. \tag{6}$$

As far as the reparametrization trick is concerned, the Choleskiy decomposition of our choice of covariance matrix has the following lower triangular form:

$$\tilde{\mathrm{C}}_{(\rho,s)} = \frac{1}{\sqrt{s}}\begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ \rho & \sqrt{1-\rho 2} & 0 & 0 & \cdots & 0 \\ \rho^2 & \rho\sqrt{1-\rho 2} & \sqrt{1-\rho 2} & 0 & \cdots & 0 \\ \rho^3 & \rho^2\sqrt{1-\rho 2} & \rho\sqrt{1-\rho 2} & \sqrt{1-\rho 2} & \cdots & 0 \\ \vdots & & & \ddots & \ddots & \vdots \\ \rho^d & \cdots & \rho^3\sqrt{1-\rho 2} & \rho^2\sqrt{1-\rho 2} & \rho\sqrt{1-\rho 2} & \sqrt{1-\rho 2} \end{bmatrix}, \tag{7}$$

which can be used to generate the latent codes as $\mathbf{z}^{(i)} = \boldsymbol{\mu}^{(i)} + \tilde{\mathrm{C}}^{(i)}_{(\rho,s)}\boldsymbol{\epsilon}$, which can be constructed also as the element-wise product of $\mathrm{C}_{(\rho,s)}$ with another highly structured matrix.

Otherwise, if depending on the choice of the deep learning framework used, the realization of Toeplitz matrices is not straightforward, one can generate AR(1) samples directly from their definition, i.e., $\mathbf{z}^{(i)}[j] = \boldsymbol{\mu}^{(i)}[j] + \sqrt{s}\boldsymbol{\epsilon}[j] + \rho\mathbf{z}^{(i)}[j-1]$, for $1 < j \leqslant d$.

Although it has less number of parameters than the standard choice and is hence more resilient towards over-fitting, this structure for the approximate posterior is more natural to consider, since correlation will somehow be represented.

Note that the fact that the prior is chosen as a white Gaussian by design, i.e., $p(\mathbf{z}) = \mathcal{N}(\mathbf{0}, \mathrm{I}_d)$, does not obviate the need for the per-sample approximate posterior to account for correlation. In fact, the per-sample posterior can be correlated, yet the aggregation of all samples can be a white Gaussian matching the prior.

Furthermore, the need for correlation does not solely stem from the natural signals like images being correlated. As a matter of fact, another requirement for the success of the VAE-based generative modeling is the tightness of the bound in (2), which is controlled by $D_{\mathrm{KL}}\big[q_\phi(\mathbf{z}^{(i)}|\mathbf{x}^{(i)})||p_\theta(\mathbf{z}^{(i)}|\mathbf{x}^{(i)})\big]$.

In other words, to guarantee a successful training, the approximate posterior should have enough capacity to match the unknown and intractable posterior. In VAE models, however, it is usually only "hoped" that this will be the case. We believe (albeit without providing quantitative evidence), that accounting for correlation may help reduce this gap.

Next we will show the effectiveness of our proposition. We show that the simple alterations to the standard approach, without the need for any sort of hyper-parameter tuning, will noticeably and consistently improve the performance under all variations considered and for all setups.

4 Experiments

References

[1] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[2] Michael Tschannen, Olivier Bachem, and Mario Lucic. Recent advances in autoencoder-based representation learning. *arXiv preprint arXiv:1812.05069*, 2018.

[3] Harold Widom. On the eigenvalues of certain hermitian operators. *Transactions of the American Mathematical Society*, 88(2):491–522, 1958.