KL Divergence in Variational Autoencoders

Closed–form Derivation and Extension to a Gaussian Mixture Prior

Student Name: PO-JUI, CHU
Student ID: 11028141

April 22, 2025

Abstract

This report addresses two closely related questions about the Kullback–Leibler (KL) term in Variational Autoencoders (VAEs). Part 2 derives a closed–form expression for the KL divergence when the encoder posterior is a diagonal Gaussian and the prior is the standard normal distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$. Part 3 investigates the feasibility and practical considerations of replacing the standard prior by a Gaussian Mixture Model (GMM). Challenges such as the loss of analytic KL, gradient estimation for discrete mixture indices, and training instability are discussed with corresponding mitigation strategies.

1 Notation and Assumptions

Let $\mathbf{x} \in \mathbb{R}^n$ be an observed data point and $\mathbf{z} \in \mathbb{R}^d$ its latent representation. Throughout the report we assume the *encoder* (variational posterior) is a diagonal Gaussian

$$q_\phi(\mathbf{z} \mid \mathbf{x}) \;=\; \mathcal{N}\big(\,\boldsymbol{\mu}(\mathbf{x}),\, \operatorname{diag}\boldsymbol{\sigma}^2(\mathbf{x})\big), \tag{1}$$

where ϕ denotes encoder parameters and $\boldsymbol{\mu}, \boldsymbol{\sigma} : \mathbb{R}^n \to \mathbb{R}^d$ are outputs of a neural network. Unless otherwise stated the *decoder* prior is $p(\mathbf{z}) = \mathcal{N}(\mathbf{0}, \mathbf{I})$ and $p_\theta(\mathbf{x} \mid \mathbf{z})$ is an arbitrary likelihood parameterised by θ.

2 Closed–form KL for the Standard Gaussian Prior

2.1 General multivariate–Gaussian formula

For two Gaussians $q = \mathcal{N}(\boldsymbol{\mu}_q, \Sigma_q)$ and $p = \mathcal{N}(\boldsymbol{\mu}_p, \Sigma_p)$ [1],

$$D_{\mathrm{KL}}(q\|p) = \frac{1}{2}\Big[\operatorname{tr}\big(\Sigma_p^{-1}\Sigma_q\big) + (\boldsymbol{\mu}_p - \boldsymbol{\mu}_q)^\top \Sigma_p^{-1}(\boldsymbol{\mu}_p - \boldsymbol{\mu}_q) - d + \ln\frac{\det\Sigma_p}{\det\Sigma_q}\Big]. \tag{2}$$

2.2 Specialising to $p(\mathbf{z}) = \mathcal{N}(\mathbf{0}, \mathbf{I})$

Set $\Sigma_p = \mathbf{I}$ and $\boldsymbol{\mu}_p = \mathbf{0}$. Let $\Sigma_q = \operatorname{diag}(\sigma_1^2, \ldots, \sigma_d^2)$ and $\boldsymbol{\mu}_q = \boldsymbol{\mu}$ obtained from the encoder. Then (2) reduces to

$$\boxed{D_{\mathrm{KL}}\big(q_\phi(\mathbf{z} \mid \mathbf{x}) \,\|\, \mathcal{N}(\mathbf{0}, \mathbf{I})\big) = \frac{1}{2}\sum_{i=1}^{d}\Big(\sigma_i^2 + \mu_i^2 - 1 - \ln\sigma_i^2\Big)}. \tag{3}$$

Back-prop friendly. Equation (3) is a deterministic, analytic function of μ_i and σ_i^2, hence of ϕ. Its gradient is

$$\frac{\partial \mathrm{KL}}{\partial \mu_i} = \mu_i, \qquad \frac{\partial \mathrm{KL}}{\partial \sigma_i} = \sigma_i - \sigma_i^{-1}, \tag{4}$$

which modern frameworks compute automatically.

2.3 ELBO with the closed-form KL

The evidence lower bound (ELBO) for a single data point is

$$\mathcal{L}(\mathbf{x}; \theta, \phi) = \mathbb{E}_{q_\phi}\big[\log p_\theta(\mathbf{x} \mid \mathbf{z})\big] - \frac{1}{2}\sum_{i=1}^{d}\big(\sigma_i^2 + \mu_i^2 - 1 - \ln \sigma_i^2\big). \tag{5}$$

This expression is the objective maximised during training. The first term is the *reconstruction* term, and the second term acts as a *regulariser* that aligns the aggregated posterior with the unit Gaussian prior.

3 Replacing the Prior with a Gaussian Mixture Model

3.1 Motivation

A Gaussian Mixture prior,

$$p(\mathbf{z}) = \sum_{k=1}^{K} \pi_k \, \mathcal{N}\big(\mathbf{z} \mid \boldsymbol{\mu}_k, \Sigma_k\big), \qquad \sum_{k=1}^{K} \pi_k = 1, \tag{6}$$

can capture multi-modal latent structure that a single isotropic Gaussian cannot. Examples include clustered data, discrete semantics, or style/identity separation.

3.2 What breaks?

1. No analytic KL.

$$\mathrm{KL}\big(q \| \mathrm{GMM}\big) = \mathbb{E}_q\big[\log q(\mathbf{z}) - \log \textstyle\sum_k \pi_k \mathcal{N}(\mathbf{z} \mid \boldsymbol{\mu}_k, \Sigma_k)\big],$$

the log–sum term precludes a closed form.

2. Discrete component index c. Sampling from a mixture involves a categorical step $c \sim \mathrm{Cat}(\boldsymbol{\pi})$. The discontinuity blocks gradient flow.

3. Training instability. High-variance gradient estimates and mode attraction may cause the encoder to collapse onto a single component.

3.3 Practical remedies

R1. Monte-Carlo KL estimation.
Draw L latent samples $\mathbf{z}^{(\ell)} \sim q_\phi(\mathbf{z} \mid \mathbf{x})$ via the reparameterisation trick, approximate

$$\mathrm{KL} \approx \frac{1}{L}\sum_{\ell=1}^{L}\Big[\log q_\phi\big(\mathbf{z}^{(\ell)} \mid \mathbf{x}\big) - \log p\big(\mathbf{z}^{(\ell)}\big)\Big].$$

R2. Gumbel–Softmax for the component index.
Relax the categorical sampling with a continuous, differentiable approximation $\tilde{c} = \text{softmax}\big((\log \boldsymbol{\pi} + \mathbf{g})/\tau\big)$, where $\mathbf{g}$ are i.i.d. Gumbel variables; anneal temperature $\tau \to 0$.

R3. Alternative divergences.
Replace the KL by Maximum Mean Discrepancy (MMD), Wasserstein distance, or an adversarial matching loss (e.g. Adversarial Auto-Encoder).

R4. Learnable priors.
VampPrior [2] or flow-based priors retain analytic KL yet remain flexible.

R5. Tighter bounds.
Importance-weighted ELBO (IWAE) reduces the bias introduced by Monte-Carlo estimation and mitigates mode collapse.

3.4 Summary

- **Feasibility:** Using a GMM prior *is* possible and can improve modelling of multi-modal data.
- **Main hurdles:** lack of closed-form KL, discrete mixture index, high gradient variance, and potential mode collapse.
- **Solutions:** stochastic KL estimation, differentiable discrete sampling, alternative divergence measures, or adopting learnable/flow priors.

Implementation Note (PyTorch)

The code implementation can be based on the framework used reference standard. For example, in PyTorch:

```
# one-line closed-form KL against N(0,I)
kl = 0.5 * (logvar.exp() + mu**2 - 1.0 - logvar).sum(dim=1).mean()
```

References

[1] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.

[2] J. M. Tomczak and M. Welling, "VAE with a VampPrior," in *AISTATS*, 2018.