VAE with a Stein-Learned Hyperprior

1 Generative Model

We introduce a global latent variable $\mathbf{m} \in \mathbb{R}^{d_m}$ that indexes latent-space coordinate systems.

Hyperprior.

$$\mathbf{m} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}). \tag{1}$$

Conditional latent prior. For each datapoint $i = 1, \ldots, N$,

$$\boldsymbol{\epsilon}_i \mid \mathbf{m} \sim \mathcal{N}(\mathbf{m}, \mathbf{I}), \tag{2}$$

$$\mathbf{z}_i = f_\lambda(\boldsymbol{\epsilon}_i), \tag{3}$$

where f_λ is an invertible normalizing flow shared across all datapoints.

Decoder.

$$\mathbf{x}_i \mid \mathbf{z}_i \sim p_\theta(\mathbf{x} \mid \mathbf{z}_i). \tag{4}$$

The conditional prior $p_\lambda(\mathbf{z} \mid \mathbf{m})$ is therefore the pushforward of $\mathcal{N}(\mathbf{m}, \mathbf{I})$ through f_λ.

Important: The covariance of $\boldsymbol{\epsilon}_i$ is fixed to $\mathbf{I}$. All geometric flexibility is handled by the shared flow f_λ. This avoids redundancy and stabilizes inference.

2 Variational Guide

We use a structured variational family

$$q(\mathbf{m}, \mathbf{z}_{1:N} \mid \mathbf{x}_{1:N}) = q(\mathbf{m}) \prod_{i=1}^{N} q_\phi(\mathbf{z}_i \mid \mathbf{x}_i, \mathbf{m}). \tag{5}$$

2.1 Global Posterior via Stein Mixture Inference

The global latent posterior is represented as a mixture

$$q(\mathbf{m}) = \frac{1}{M} \sum_{k=1}^{M} \mathcal{N}(\mathbf{m}; \mathbf{a}_k, \mathbf{B}_k), \tag{6}$$

where $\{(\mathbf{a}_k, \mathbf{B}_k)\}_{k=1}^{M}$ (keep covariance diagonal)) are updated using Stein Mixture Inference with repulsive interactions between components.

SMI is applied *only* to these parameters. All neural network parameters remain shared.

2.2 Shared Amortized Encoder

A single encoder network produces parameters of a Gaussian conditioned on both the datapoint $\mathbf{x}_i$ and the global latent $\mathbf{m}$:

$$(\boldsymbol{\mu}_\phi(\mathbf{x}_i, \mathbf{m}), \boldsymbol{\Sigma}_\phi(\mathbf{x}_i, \mathbf{m})) = \mathrm{Enc}_\phi(\mathbf{x}_i, \mathbf{m}), \tag{7}$$

$$\mathbf{z}_i \sim \mathcal{N}\big(\boldsymbol{\mu}_\phi(\mathbf{x}_i, \mathbf{m}), \boldsymbol{\Sigma}_\phi(\mathbf{x}_i, \mathbf{m})\big). \tag{8}$$

All datapoints share the same encoder parameters ϕ; conditioning on $\mathbf{m}$ allows different mixture components to define different latent charts.

3 Recommended Starting Configuration (Best ROI)

The following configuration is recommended as the default:

- Use a moderate-dimensional global latent ($d_m \approx d_z$ or smaller).
- Fix the base covariance to identity: $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{m}, \mathbf{I})$.
- Use a single shared normalizing flow f_λ on $\mathbf{z}$.
- Condition the encoder on $\mathbf{m}$ via concatenation or FiLM modulation.
- Apply SMI only to $q(\mathbf{m})$ (do not replicate encoder/decoder parameters).

This yields a mixture of transported base measures with shared geometry, providing flexibility where needed while keeping optimization well-conditioned.