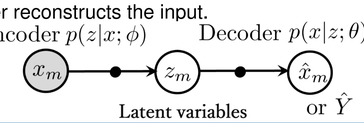
Information-Theoretic Generalization Bounds for VAEs: A Role of Encoder and Latent Variable

Futoshi Futami* (The Univ. of Osaka / The Univ. of Tokyo / RIKEN AIP)
Masahiro Fujisawa* (The Univ. of Osaka / RIKEN AIP)
*: Equal contribution

Poster session @ Wed 8 July 2.30 p.m. KST – 4.15 p.m. KST

Motivation

- **VAE = Encoder–Decoder model**: the encoder compresses data into **latent variables (LVs)**, and the decoder reconstructs the input.



$$\ell_{\mathrm{vae}}(x,\phi,\theta) = \mathbb{E}_{Z\sim q(z|\phi,x)}[-\log p(x \mid \theta, Z)] + \beta \mathrm{KL}(q(z \mid \phi, x)\|p(z)).$$

$\min_{\{\phi,\theta\}}$ Training error + Complexity of LVs

Minimizing the reconstruction loss while regularizing LVs is crucial.

Motivation. Does regularizing the LVs imply generalization? In this work, we answer this question by using the Information-theoretic generalization analysis.

Existing Work and Our Contribution

- Existing analyses focus on **all parameters**, **fixed/data-independent decoders**, or **discrete/single LVs**.
- We give a decoder-independent IT bound for **standard and hierarchical Gaussian VAEs** with **continuous LVs** under joint training of the **decoder and encoder**.

Existing	Model	Bound type	Training	Data gen.
Chérief-Abdellatif et al.	Standard VAE	All params	Joint	No
Mbacke et al.	Standard VAE	Entangled	Data-indep.	Yes
Futami & Fujisawa	VQ-VAE	Encoder & LVs	Joint	Yes
Ours	Standard & H-VAE	Encoder & LVs	Joint	Yes

Problem Setup

Leave-one-out (LOO) Conditional MI. $X^n = \{X_1,\dots,X_n\}$, $U \sim \mathrm{Unif}([n])$, $S = X^n \setminus \{X_U\}$. We use the leave-one-out index for the generalization analysis.

- **Model.** Gaussian encoder $q(z \mid \phi, x) = \mathcal{N}(\mu_\phi(x), \mathrm{diag}(\sigma^2_\phi(x)))$ and decoder $g_\theta : \mathcal{Z} \to \mathcal{X}$ are trained together, $W = (\phi, \theta)$. We assume the randomized algorithm ($W \sim q(W|S)$)
- **Loss.** We compare population and empirical reconstruction loss $\ell_0(w,x) = \mathbb{E}_{q(Z|\phi,x)}\|x - g_\theta(Z)\|^2$ under bounded data diameter Δ.

Assumptions. i.i.d. samples; bounded/clipped loss with diameter Δ; controlled encoder variances $\sigma^2_{\min} \le \sigma^2_{\phi,j}(x) \le \sigma^2_{\max}$; we jointly train the **decoder and encoder** networks.

Role of assumptions. Bounded reconstruction loss controls DV/McDiarmid exponential moments; Gaussian encoder bounds control latent tails before the continuous limit.

$$\mathrm{gen}(n,\mathcal{D}) = \left| \mathbb{E}_{X^n,U}\mathbb{E}_{q(W|S)}\ell_0(W,X_U) - \mathbb{E}_{X^n,U}\mathbb{E}_{q(W|S)}\frac{1}{n-1}\sum_{i\neq U}\ell_0(W,X_i) \right|.$$

Main Theorem: Encoder–LV Decomposition

$$\mathrm{gen}(n,\mathcal{D}) \le \frac{\Delta}{\sqrt{n-1}} + \frac{3n\Delta}{\sqrt{2}(n-1)}\sqrt{\underbrace{I(Z^n;U \mid \phi, X^n)}_{\text{LV complexity}} + \underbrace{I(\phi;U \mid X^n)}_{\text{Encoder complexity}}}.$$

Upper-bound structure:

Generalization $\le C\sqrt{\text{LV information} + \text{Encoder overfitting}} + O(n^{-1/2})$.

- $I(Z^n;U \mid \phi, X^n)$ is the LV information term.
- $I(\phi;U \mid X^n)$ is the encoder information / encoder-overfitting term.
- **Decoder complexity does not explicitly appear**; the bound isolates the LV and learned encoder channels.

Reading the LV information term. The LV term is the entropy drop

$$I(Z^n;U \mid \phi, X^n) = H(U \mid \phi, X^n) - H(U \mid \phi, X^n, Z^n).$$

It measures how much LVs reveal the LOO test index: **LV index memorization**.

Why the Standard LOO Bound is Not Enough

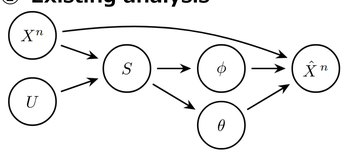
- Direct LOO-CMI treats the entire learned model as one block:

$$\mathrm{gen}(n,\mathcal{D}) \le \frac{n}{\sqrt{2}(n-1)}\sqrt{I(W;U \mid X^n)}, \qquad W = (\phi,\theta).$$

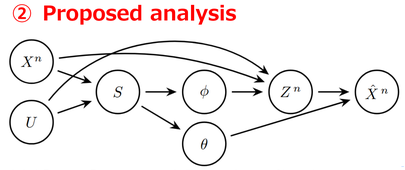
This **entangles encoder, decoder, and LVs**.

Key idea. Use a data-dependent prior of **LVs that averages over U** (shuffling over the LOO randomness), while keeping the same learned encoder–decoder training rule.

① Existing analysis



② Proposed analysis



The figure illustrates the dependency of U (See Step 4 of Proof outline).

- Existing: The index U specifies the train/test split for the entire prediction block.
- Proposed: The marginalization over U is performed on Z^n.

Proof Outline: 5 Steps

1. **Truncate** Gaussian LVs to $B(R)$ so the latent domain is bounded (Z_R).
2. **Discretize** Z_R by a fixed, data-independent codebook index $J \in [K]$.
3. **Bias control** from Gaussian tails and quantization.
4. **Apply Donsker-Valadhan (DV)** lemma to $(\mathbf{Q},\mathbf{P})$ on the finite index space J^n.
5. **Return to continuous LVs** by data processing and limits.

Steps 1–3: Define the Discrete Latent Variable

Purpose. We apply the VQ-VAE analysis teachnieque (Futami & Fujisawa (2025)), which leads to the decoder-independent bound. For that purpose, we discretize the Gaussian LVs.

$$U_{\mathrm{loo}} \to Z \xrightarrow{\text{1. truncate}} Z_R \xrightarrow{\text{2. discretize}} J \xrightarrow{\text{3. bias control}} \widehat{Z}$$

1. Truncate Choose $B_R = \{z : \|z\|_2 \le R\}$ and set $Z_R = Z$ inside B_R; outside B_R, replace Z by an auxiliary draw in B_R.

2. Discretize Fix a data-independent τ-net of B_R with Voronoi cells $A_1,\dots,A_K$. Define the finite LV index by $J = k \iff Z_R \in A_k$.

3. Bias control Given $J = k$, sample $\widehat{Z} \sim \mathrm{Unif}(A_k)$ and compare continuous expectations after applying DV on J^n.

$$\widehat{Z} \mid J = k \sim \mathrm{Unif}(A_k), \qquad |\mathbb{E}L(Z) - \mathbb{E}L(\widehat{Z})| \lesssim \Pr(Z \notin B_R) + \tau \to 0.$$

Step 4: Discrete DV Bound

$$\mathbf{Q} := p(U)\,q(\phi,\theta \mid X_U^{n-1})\,q(J^n \mid \phi, X_U, X_U^{n-1}),$$
$$\mathbf{P} := p(U)\,q(\phi,\theta \mid X_U^{n-1})\,\mathbb{E}_{p(U')}q(J^n \mid \phi, X_{U'}, X_{U'}^{n-1}).$$

Why no decoder-complexity term? The learned decoder appears in both $\mathbf{Q}$ and $\mathbf{P}$ through the same $q(\phi,\theta \mid X_U^{n-1})$; the KL measures mismatch in the **LV index distribution** and encoder dependence on U.

$$\mathbb{E}_{\mathbf{Q}}F \le \frac{1}{\lambda}\mathrm{KL}(\mathbf{Q}\|\mathbf{P}) + \frac{1}{\lambda}\log\mathbb{E}_{\mathbf{P}}e^{\lambda F}, \qquad \mathrm{KL}(\mathbf{Q}\|\mathbf{P}) \le I(J^n;U \mid \phi, X^n) + I(\phi;U \mid X^n).$$

Step 5: Continuous Limit

- The truncation and quantization are data-independent post-processing steps.
- They form the Markov chain $U \to Z^n \to Z_R^n \to J^n$ conditional on (ϕ, X^n); hence **data processing** gives:

$$I(J^n;U \mid \phi, X^n) \le I(Z_R^n;U \mid \phi, X^n) \le I(Z^n;U \mid \phi, X^n).$$

Finally, take the limits $R \to \infty$ and $K \to \infty$ (equivalently $\tau \to 0$) so the truncation and quantization errors vanish while the information-theoretic bound remains valid.

Encoder Flatness

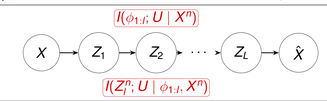
Encoder term as stability. $I(\phi;U \mid X^n)$ measures how much the learned encoder changes when the held-out index changes. We approximate this quantity using the influence function.

$$g_i := \nabla_\phi \ell_{\mathrm{vae}}(x_i,\phi^*,\theta^*) \qquad \phi_i^* - \phi^* \approx -\frac{1}{n}(H_{w^*}^{\phi})^{-1}g_i \qquad I(\phi;U \mid X^n) \lesssim \frac{1}{2n}\widehat{\mathbb{E}}_{i,j}(g_i-g_j)^\top(H_{w^*}^{\phi})^{-1}(g_i-g_j).$$

Reading. Stable leave-one-out perturbations and flatter encoder geometry reduce the encoder-complexity term.

Extension to Hierarchical VAEs (H-VAEs)

- H-VAEs use stochastic layers $Z_1,\dots,Z_L$; each layer can serve as the representation.
- For layer l, the encoder part is cumulative: $\phi_{1:l}$ (We denote the cumulative parameters up to layer l by $\phi_{1:l} = (\phi_1,\dots,\phi_l)$) contains the encoders used up to that layer.



$$\mathrm{gen}_{\mathrm{H}}(n,\mathcal{D}) \le \frac{\Delta}{\sqrt{n-1}} + \frac{3n\Delta}{\sqrt{2}(n-1)}\sqrt{I(Z_l^n;U \mid \phi_{1:l},X^n) + I(\phi_{1:l};U \mid X^n)}.$$

$$D_l := \sqrt{I(Z_l^n;U \mid \phi_{1:l},X^n) + I(\phi_{1:l};U \mid X^n)}, \qquad \mathrm{gen}_{\mathrm{H}}(n,\mathcal{D}) \le \min_{l\in[L]}\{\text{right-hand side for layer } l\}.$$

Interpretation. The bound is a layer-wise trade-off: deeper Z_l may compress latent information, while $I(\phi_{1:l};U \mid X^n)$ can accumulate through the encoder stack.

Experiments

Settings. Target: $|L_{\mathrm{test}}^{\mathrm{recon}} - L_{\mathrm{train}}^{\mathrm{recon}}|$. Each dataset uses a fixed architecture and $n_{\mathrm{train}} = 30{,}000$. We sweep $\beta \in \{0.01, 0.03, 0.3, 1, 3, 10\}$ over 10 splits, giving 60 pairs per dataset. Single-layer diagnostic: $D := \sqrt{I_Z + I_\phi}$; H-VAE uses layer-wise D_l for $l = 1,\dots,4$.

H-VAE model. MNIST: hierarchical MLP, hidden dims [512, 256, 128, 64]. F-MNIST: hierarchical CNN, channels [32, 64, 64, 64]. Both use latent dims [32, 16, 8, 4].

Single-layer beta-VAE.

Data	Diag.	Pear.	Spear.	Kend.
MNIST	I_ϕ	0.434	0.567	0.374
MNIST	I_Z	**0.930**	0.775	0.529
MNIST	D	0.928	**0.780**	**0.531**
F-MNIST	I_ϕ	-0.559	-0.800	-0.628
F-MNIST	I_Z	0.904	**0.932**	**0.767**
F-MNIST	D	**0.961**	0.914	0.738
CIFAR-10	I_ϕ	-0.332	-0.857	-0.679
CIFAR-10	I_Z	0.923	**0.883**	**0.714**
CIFAR-10	D	**0.929**	0.846	0.667

Hierarchical beta-VAE.

Data	Diag.	Pear.	Spear.	Kend.
MNIST	D_1	0.479	0.366	0.256
MNIST	D_2	0.435	0.228	0.150
MNIST	D_3	0.406	**0.553**	**0.385**
MNIST	D_4	**0.529**	0.365	0.242
MNIST	$\min_l D_l$	**0.529**	0.365	0.242
F-MNIST	D_1	**0.952**	**0.761**	**0.560**
F-MNIST	D_2	0.837	0.738	0.513
F-MNIST	D_3	0.873	0.701	0.476
F-MNIST	D_4	0.876	0.708	0.486
F-MNIST	$\min_l D_l$	0.876	0.708	0.486

Results from the beta sweeps.

- Varying β changes the reconstruction gap through the latent-regularization channel.
- I_Z is strongly positive on MNIST, F-MNIST, and CIFAR-10. (the LV term plays a central role in explaining the observed variation.)
- The combined quantity (D) gives comparable or stronger correlations.

Data Generation Guarantee

- After training, the model generates new data by sampling a LV z from the prior $p(z)$ and transforming it through the decoder network g_θ.
- The resulting data distribution is the pushforward measure $\hat{\mu} := g_\theta \# p(z)$.
- The same encoder–LV control bounds the quality of the generated data.

$$\mathbb{E}W_2^2(\mathcal{D},\hat{\mu}) \le \frac{2\Delta}{\sqrt{n-1}} + \mathbb{E}\left[\frac{2}{n-1}\sum_{i=1}^{n-1}\ell_0(W,X_i) + 3\Delta\sqrt{\frac{2}{n-1}\sum_{i=1}^{n-1}\mathrm{KL}(q(Z_i \mid \phi,X_i)\|p(Z_i))}\right].$$

Interpretation. Generation quality is tied to empirical reconstruction and latent KL, again without an explicit decoder parameter-count term.