EBERHARD KARLS UNIVERSITÄT TÜBINGEN

imprs-is

ICLR 2023

Trading Information between Latents in Hierarchical Variational Autoencoders

Tim Z. Xiao, Robert Bamler
Department of Computer Science,
Cluster of Excellence "Machine Learning for Science", Tübingen AI Center,
University of Tübingen

Presentation on 11th International Conference on Learning Representations (ICLR), 2023

SCAN ME

Controlling Information in β-VAEs

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

$$\underbrace{\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log p_\theta(\mathbf{x}|\mathbf{z})\right]}_{-\text{ distortion } D} - \beta \underbrace{D_{\mathrm{KL}}\left(q_\phi(\mathbf{z}|\mathbf{x}) \,\|\, p(\mathbf{z})\right)}_{\text{rate } R}$$

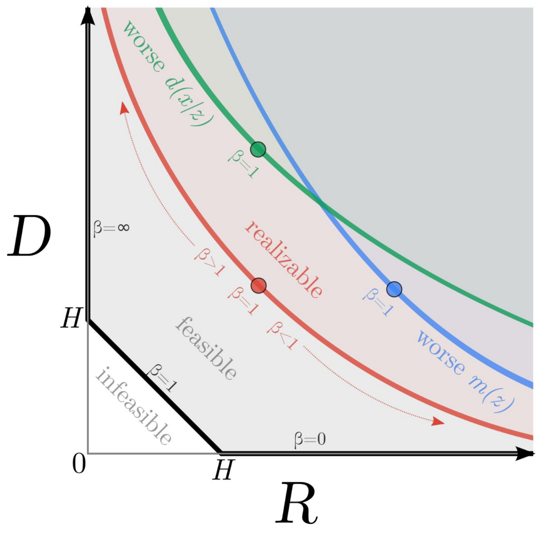
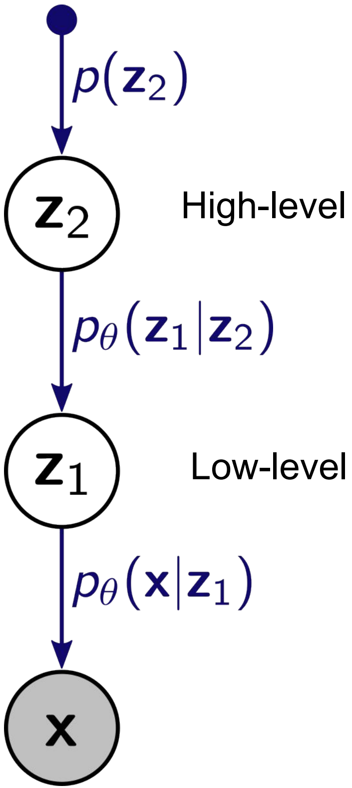


Figure taken from Alemi et al., *Fixing a Broken ELBO*, ICML 2018.



Defining Layer-Wise Bit Rates

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

For one architecture, total bit rate separates into:

$$R = R(\boldsymbol{z}_L) + R(\boldsymbol{z}_{L-1}|\boldsymbol{z}_L) + R(\boldsymbol{z}_{L-2}\,|\,\boldsymbol{z}_{L-1}, \boldsymbol{z}_L) + \,\ldots\, + R(\boldsymbol{z}_1|\boldsymbol{z}_{\geq 2})$$

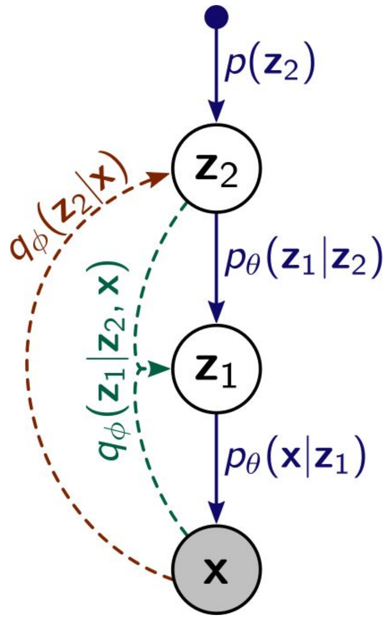
where:

$$R(\boldsymbol{z}_\ell|\boldsymbol{z}_{\geq \ell+1}) = \mathbb{E}_{q(\boldsymbol{z}_{\geq \ell+1}|\boldsymbol{x})}\left[D_{\mathrm{KL}}\left[q_\phi(\boldsymbol{z}_\ell\,|\,\boldsymbol{z}_{\geq \ell+1}, \boldsymbol{x}) \,\|\, p_\theta(\boldsymbol{z}_\ell\,|\,\boldsymbol{z}_{\geq \ell+1})\right]\right]$$

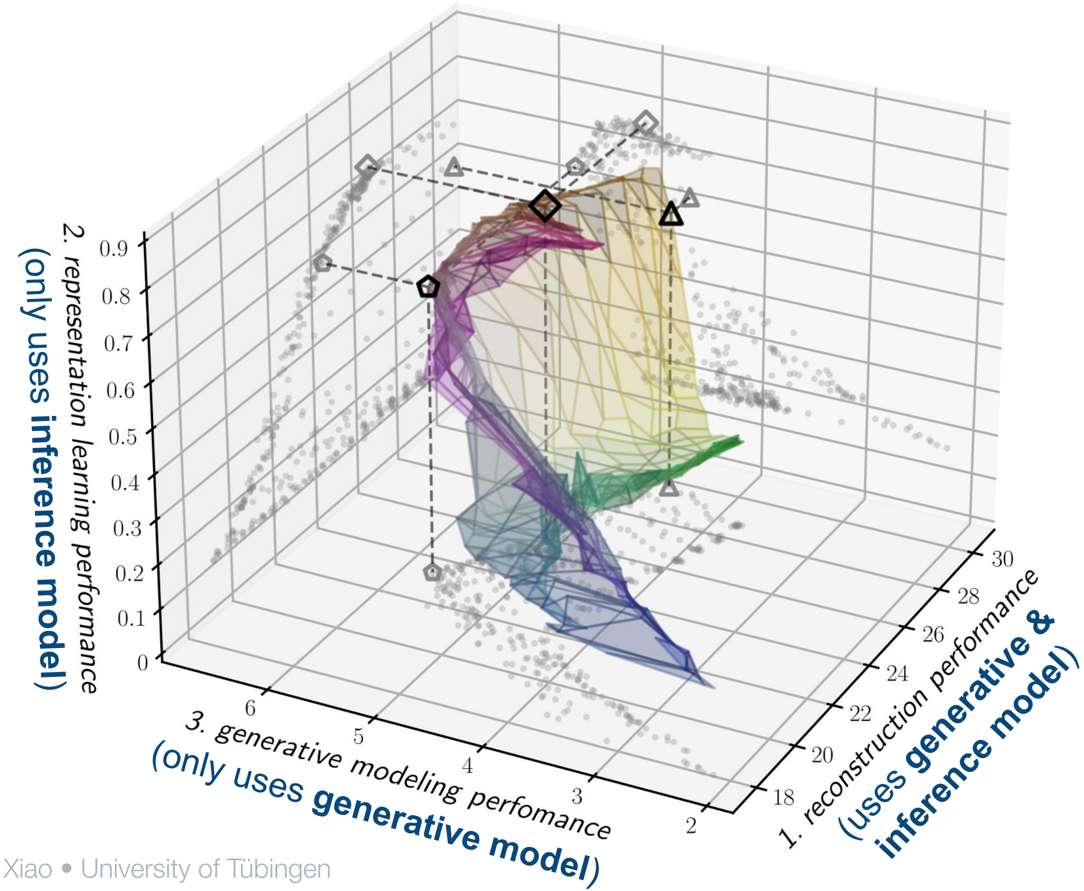
⇒ Proposed training objective:

$$\mathbb{E}_{\boldsymbol{x}\sim\mathbb{X}_{\text{train}}}\left[D + \beta_L R(\boldsymbol{z}_L) + \beta_{L-1} R(\boldsymbol{z}_{L-1}|\boldsymbol{z}_L) + \ldots + \beta_1 R(\boldsymbol{z}_1|\boldsymbol{z}_{\geq 2})\right]$$

L independent Lagrange multipliers



There is no "One VAE Fits All"



diverse application domains

⇓

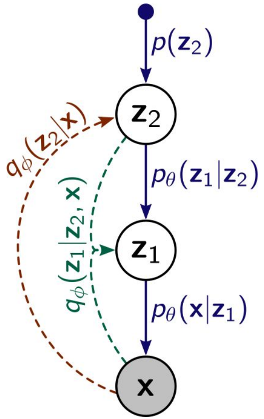
need fine-grained control
(no one-size-fits-all hierarchical VAE)

Application Type 1: Reconstruction

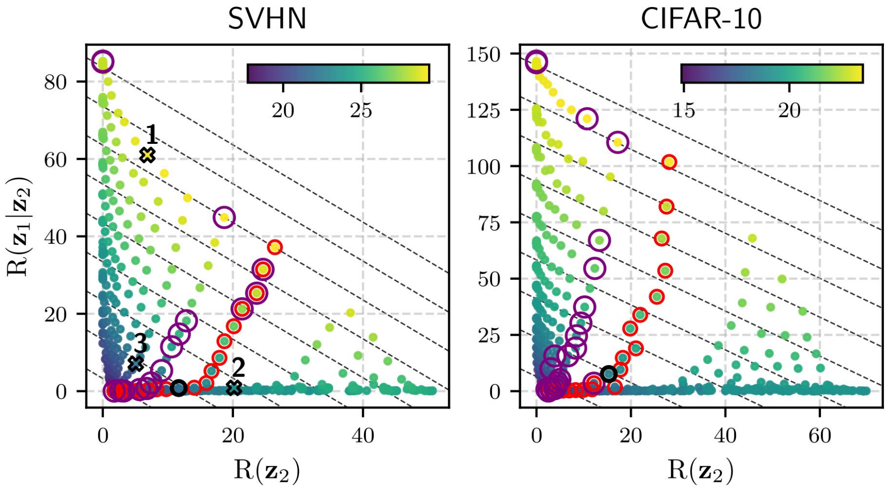
Theory:

$$\mathbb{E}_{\mathbf{x}\in p_{\text{data}}}[D] \geq H[p_{\text{data}}] - E_{\mathbf{x}\in p_{\text{data}}}\left[R(\mathbf{z}_L) + R(\mathbf{z}_{L-1}|\mathbf{z}_L) + \ldots + R(\mathbf{z}_1|\mathbf{z}_{\geq 2})\right]$$

distortion

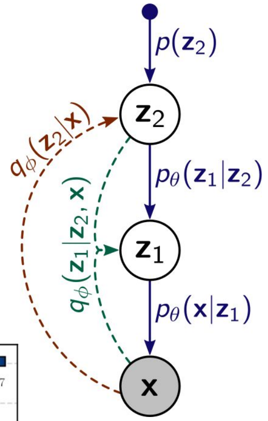


Experiment:



Application Type 2: Rep. Learning

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

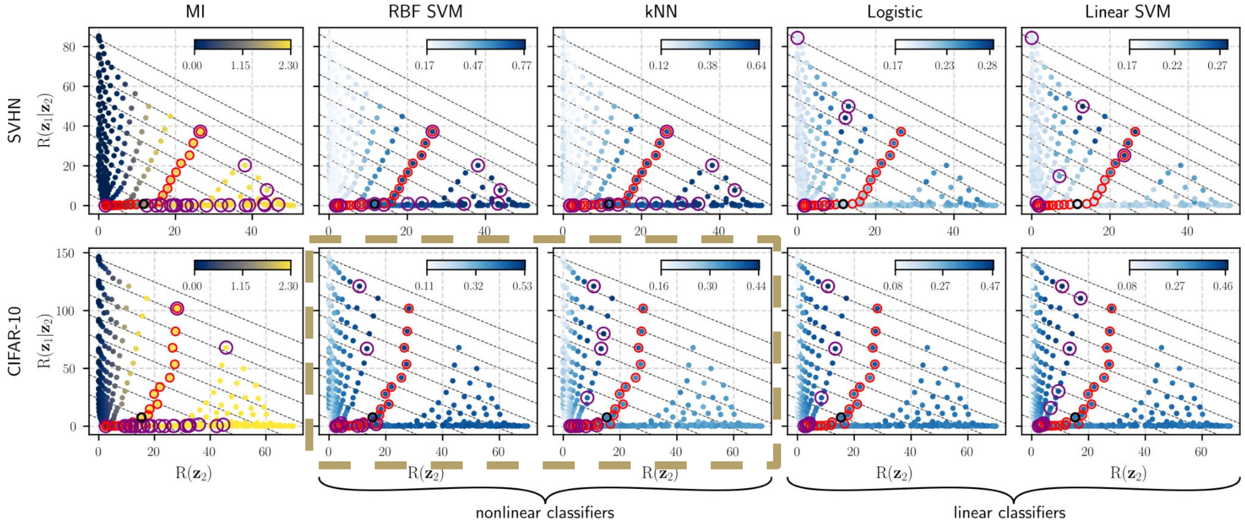


Theory: consider classifier operating on $\mathbf{z}_2$

$$\Rightarrow \text{accuracy} \leq f^{-1}\big(I_q(\text{label}; \mathbf{z}_2)\big) \leq f^{-1}\big(\mathbb{E}_{p_\text{data}}[R(\mathbf{z}_2)]\big)$$

$$f(\alpha) = H[p_\text{data}(y)] + \alpha \log \alpha + (1-\alpha) \log \frac{1-\alpha}{M-1}$$ [analogous to Meyen, 2016]

Experiment:



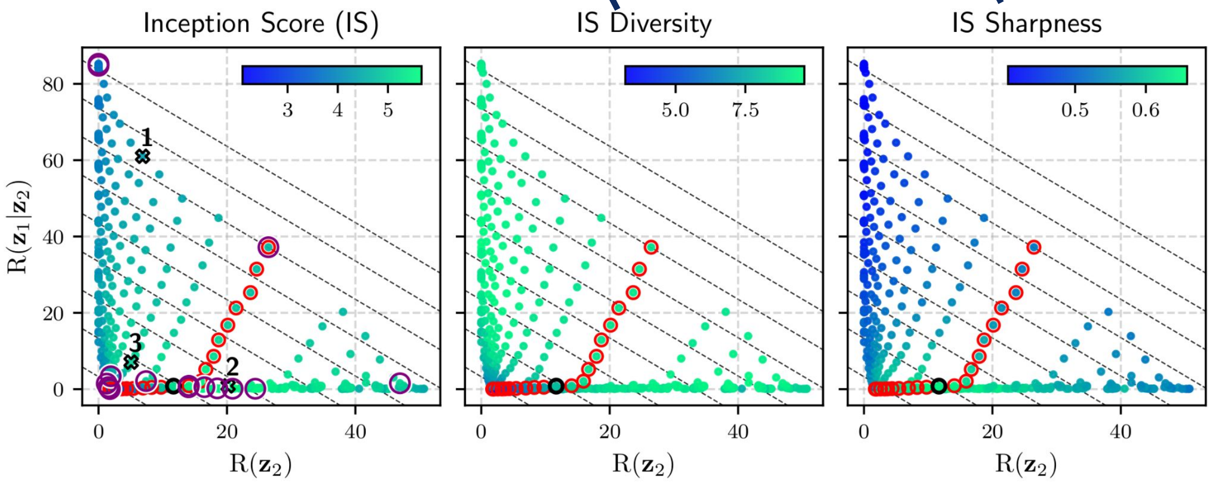
Application Type 3: Generation

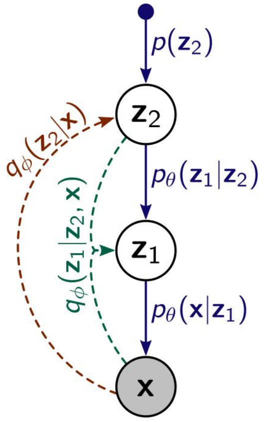
Theory: expect best generative performance when all $\beta = 1$

Experiment:

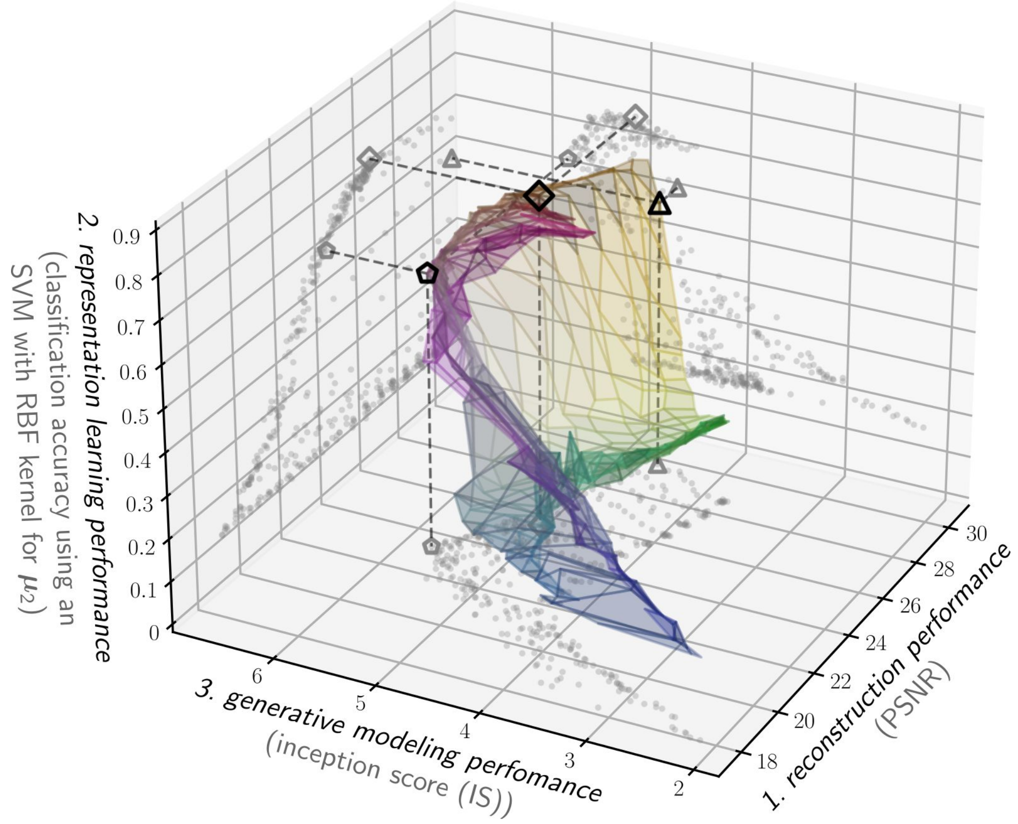
$$\mathrm{IS} = \exp\left\{\mathbb{E}_{p_\theta(\boldsymbol{x})}\left[D_{\mathrm{KL}}[p_{\mathrm{cls.}}(y|\boldsymbol{x}) \,\|\, p_{\mathrm{cls.}}(y)]\right]\right\}$$
$$= e^{H[p_{\mathrm{cls.}}(y)]} \times e^{-\mathbb{E}_{p_\theta(\boldsymbol{x})}[H[p_{\mathrm{cls.}}(y|\boldsymbol{x})]]}$$

[Salimans et al., 2016]





Summary



diverse application domains

⇓

need fine-grained control

⇓

control layer-wise rates