EBERHARD KARLS UNIVERSITÄT TÜBINGEN

imprs-is

What do we want from a generative model and how do we get it from a VAE?

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

Talk at Imperial College London - Computational Statistics and Machine Learning reading group
2 November, 2023

About me

- BSc in Computer Science – University of Manchester
  - Thesis on Generative adversarial networks (GANs)
- MSc in Computer Science – University of Oxford
  - Thesis on measuring uncertainty in Transformer for neural machine translation
- MRes in Computational Statistics and Machine Learning – UCL
  - Thesis on active learning using semi-supervised generative model
- (Currently) PhD student – University of Tübingen (with Robert Bamler)
  - Deep probabilistic models

Fun fact!

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

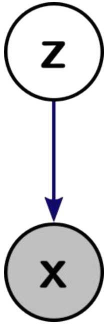
(Photo from Yingzhen's website)

Outline

- What do we want from a generative model?
- How do we get it from a VAE?
  - Part 1: Loss function
    - "*Trading Information between Latents in Hierarchical Variational Autoencoders*" (ICLR 2023)
  - Part 2: Training data
    - "*Upgrading VAE Training With Unlimited Data Plans Provided by Diffusion Models*"
    - "*The SVHN Dataset Is Deceptive for Probabilistic Generative Models Due to a Distribution Mismatch*"
- Conclusion

Deep Generative Models (DGMs)

EBERHARD KARLS UNIVERSITÄT TÜBINGEN



1. Generative path (i.e., z → x)
   - Drawing z, then what can be the corresponding x?
   - Often explicitly defined
2. Inference path (i.e., x → z)
   - Given an x, what can be the corresponding z?
   - Explicit: Normalizing Flows, VAEs, Diffusion models
   - Implicit: GANs (inference by optimization), non-amortized Bayesian inference

Applications of DGMs

1. Data generation (Gen.)
2. Data compression (Gen. + Inf.)
3. Representation learning (Inf.)
4. Image-to-Image translation (Gen. + Inf.)
5. Anomaly detection (Gen., Inf.)

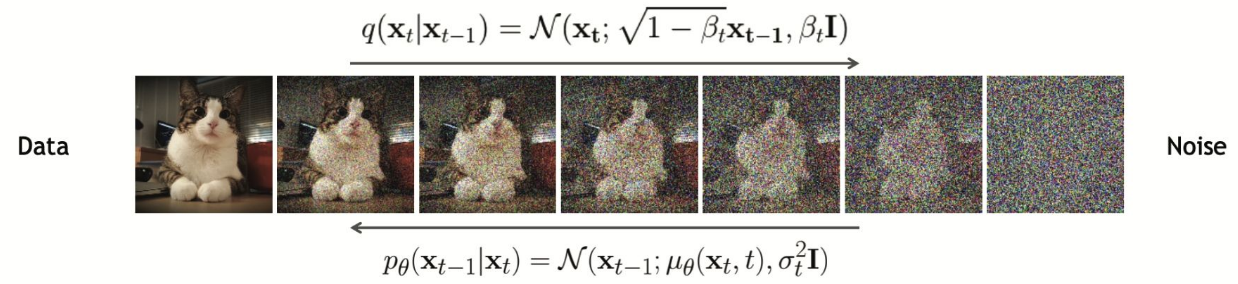
Note: Not all generative models are suitable for all above applications!

Applications of DGMs (e.g., Diffusion)

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

1. Data generation (Gen.)
2. Data compression (Gen. + Inf.)
3. Representation learning (Inf.)

- → SOTA
- → Not natural for lossy compression
- → Representation not learned and not semantically meaningful



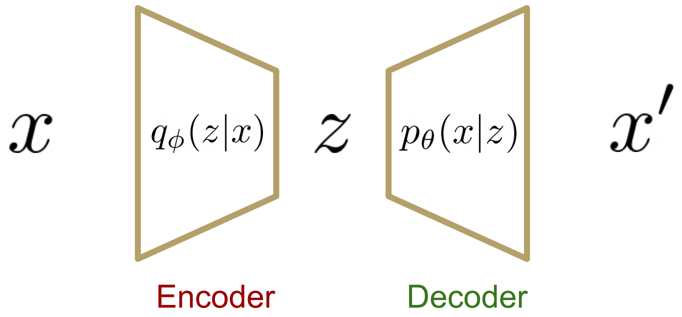
[Figure 25.1 from Murphy (2023)]

Applications of DGMs (e.g., VAEs)

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

1. Data generation (De.)
2. Data compression (De. + En.)
3. Representation learning (En.)

- → Good in Deep HVAE
- → Natural for lossy compression
- → Learned inference model

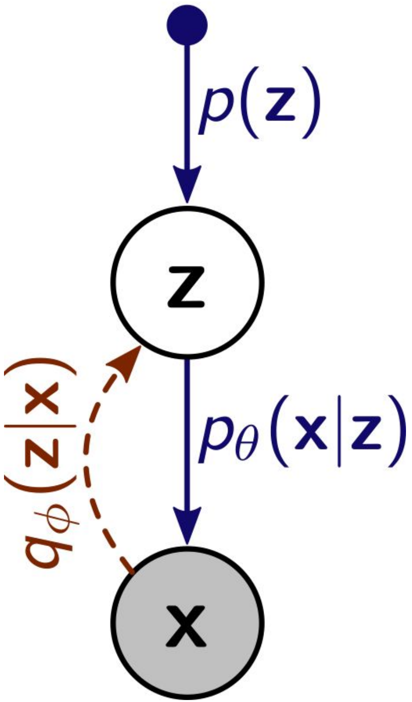


Outline

- What do we want from a generative model?
  - Different ways of using the generative and the inference path
  - VAEs seems more general in turns of applications, even though diffusion models are better in data generation

Next:

- How do we get it from a VAE?
  - Part 1: By loss function (from Info. Theory perspective)
    - How to tune VAEs towards different applications?
  - Part 2: By training data



Marginal likelihood:

$$\log p_\theta(\mathbf{x}) = \log \int p(\mathbf{z})\, p_\theta(\mathbf{x}|\mathbf{z})\, dz$$

$$\geq \underbrace{\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log p_\theta(\mathbf{x}|\mathbf{z})\right]}_{\text{– reconstruction error (“distortion”)}} - \underbrace{D_{\mathrm{KL}}\left(q_\phi(\mathbf{z}|\mathbf{x}) \,\|\, p(\mathbf{z})\right)}_{\text{information content in } \mathbf{z} \text{ (“bit rate”)}}$$

β-VAE objective:

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

Controlling Information in β-VAEs

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

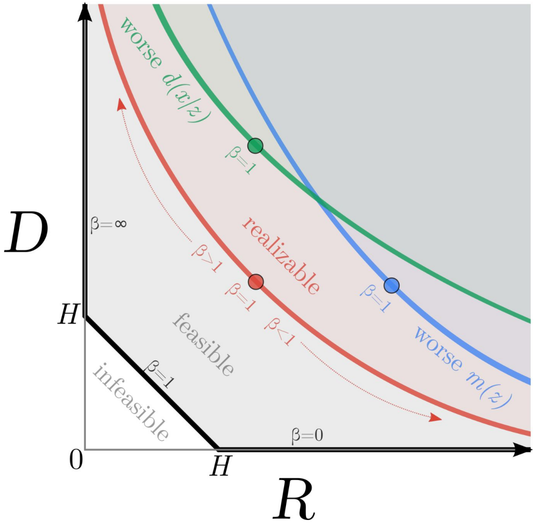
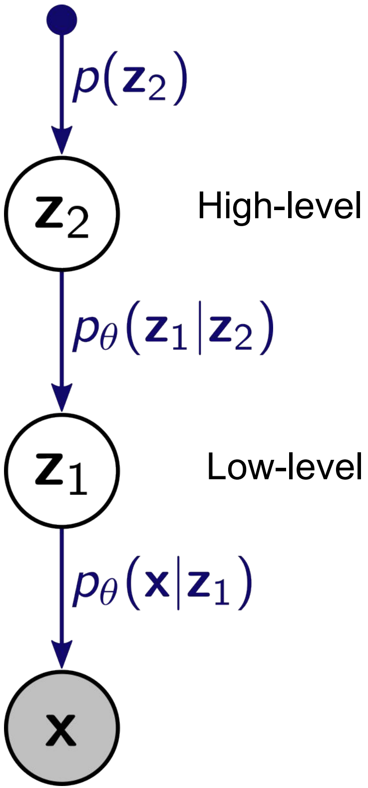
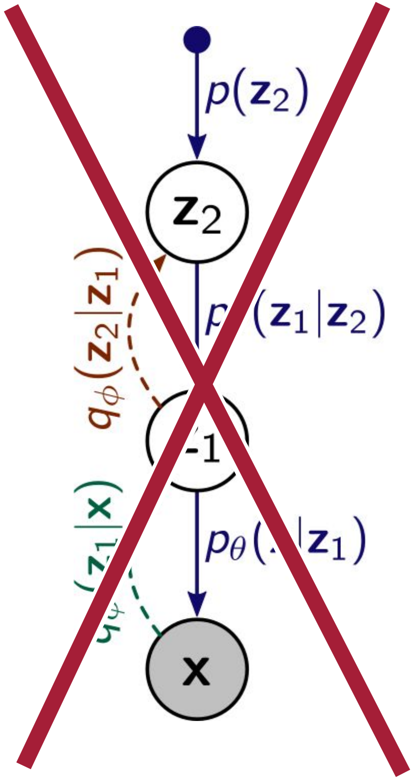
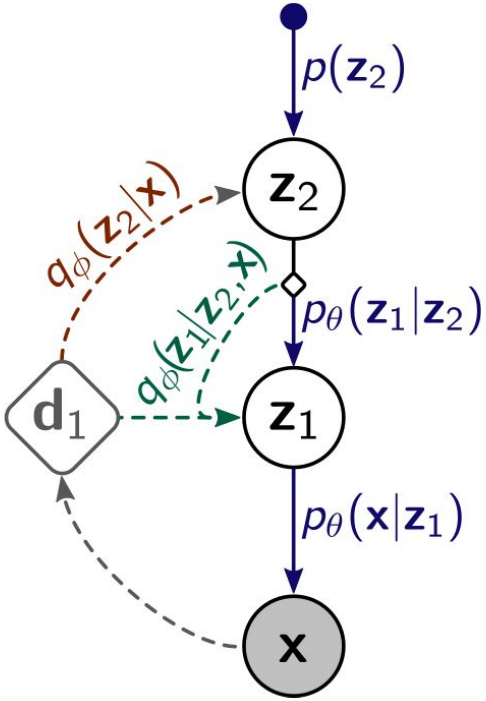


Figure taken from Alemi et al.,
Fixing a Broken ELBO, ICML 2018.



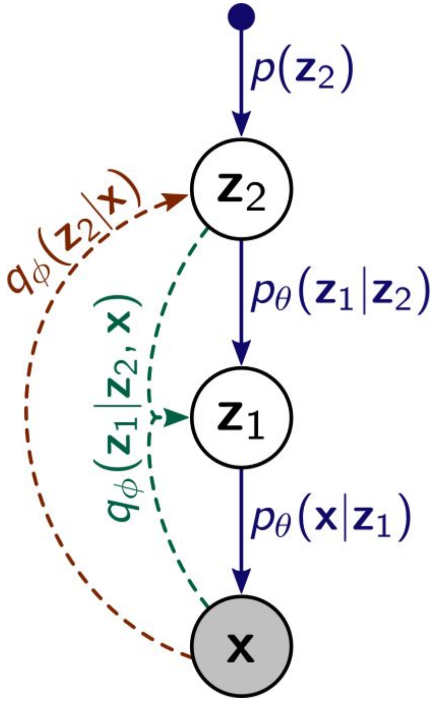
Defining Layer-Wise Bit Rates





"Ladder VAE"

[Sønderby et al., 2016]



most general architecture that admits definition of layer-wise rates

Defining Layer-Wise Bit Rates

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

For one architecture, total bit rate separates into:

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

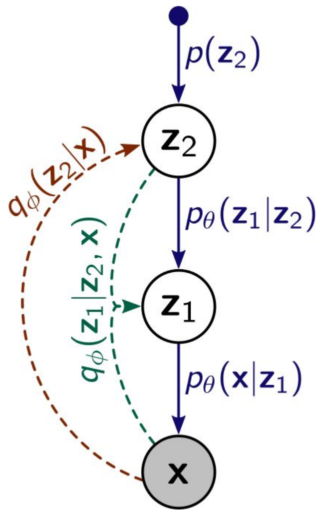
where:

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

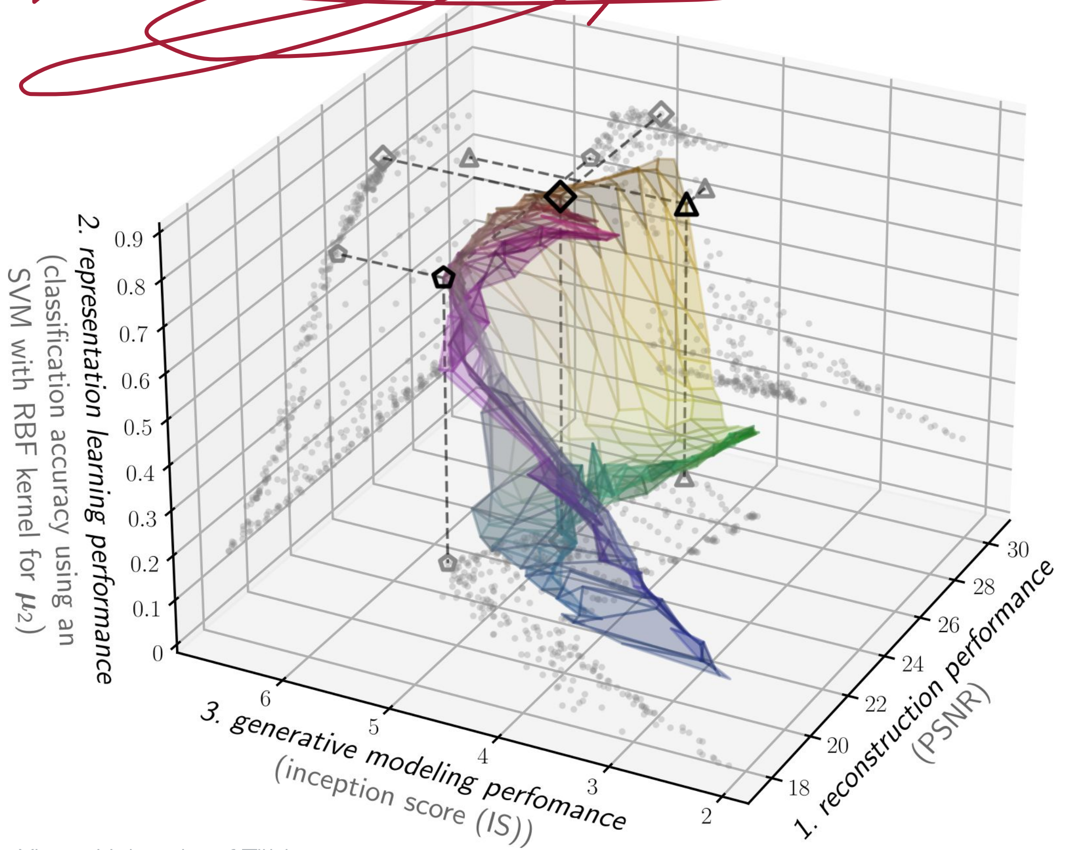
⇒ Proposed training objective:

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

L independent Lagrange multipliers



One VAE to rule them all?



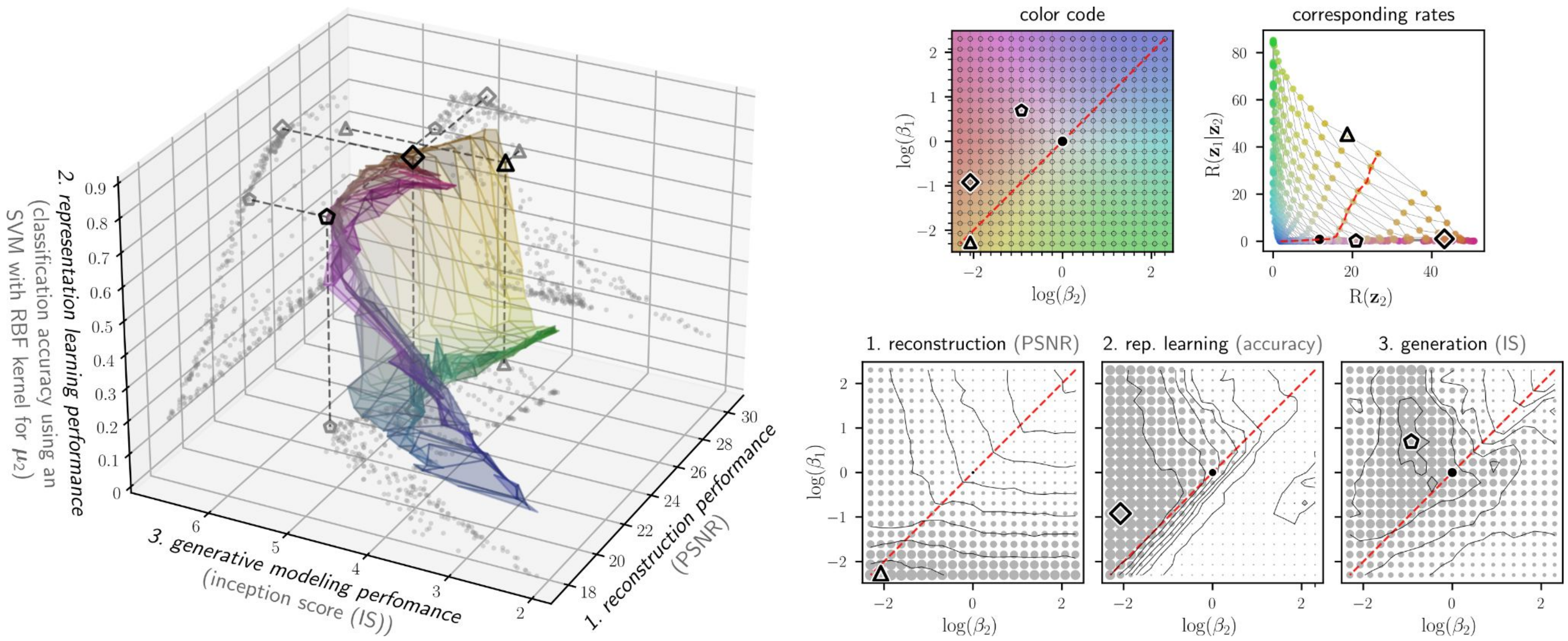
diverse application domains

⇓

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

Trading Information between Layers

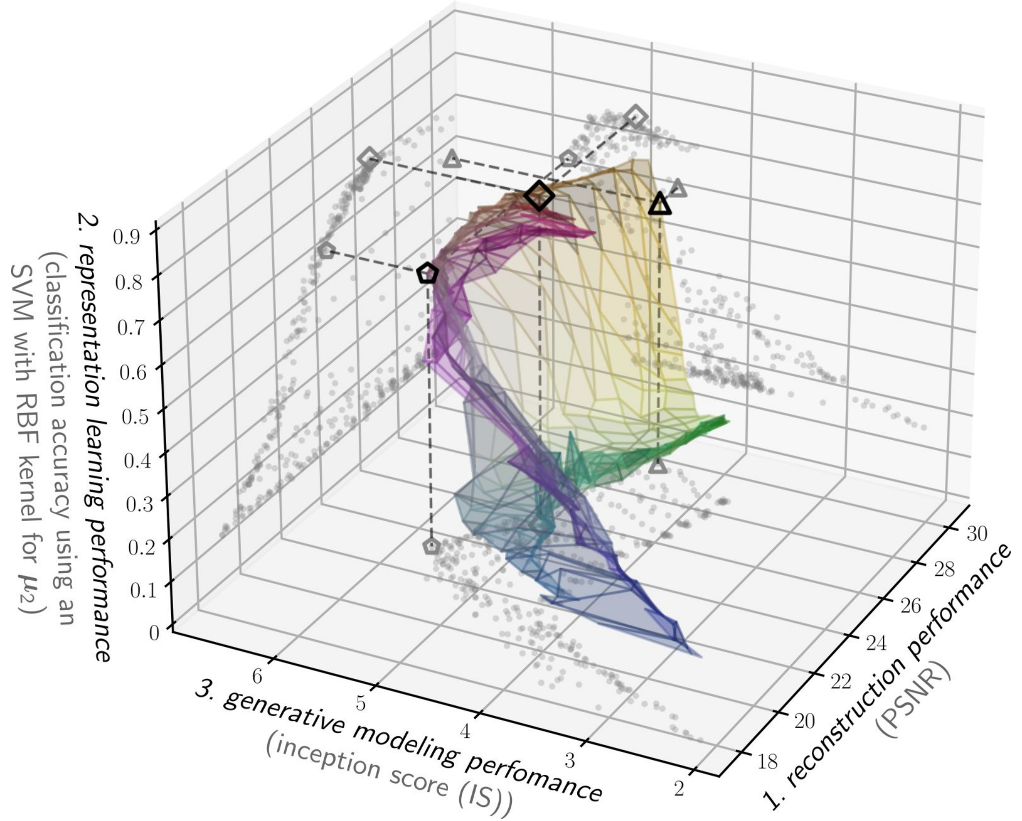
EBERHARD KARLS UNIVERSITÄT TÜBINGEN



(example: 2-layer HVAE trained on SVHN data set; similar trends for other models & data sets, see our paper)

Part 1 – Summary

EBERHARD KARLS UNIVERSITÄT TÜBINGEN



diverse application domains

⇓

need fine-grained control

⇓

control layer-wise rates

Outline

- What do we want from a generative model?
- How do we get it from a VAE?
  - Part 1: By loss function
    - Optimizing VAEs towards different applications by trading off information across layers
    - There is no one-size-fits-all VAE

Next:

- Part 2: By training data
  - Improving generalization in VAEs by exploiting pre-trained diffusion models
  - Remark: Distribution mismatch in SVHN results in false evaluation of generative models
  - "*Upgrading VAE Training With Unlimited Data Plans Provided by Diffusion Models*"
  - "*The SVHN Dataset Is Deceptive for Probabilistic Generative Models Due to a Distribution Mismatch*"

Overfitting in VAEs (Cremer et al., 2018)

$$\log p_\theta(\boldsymbol{x}) \geq \mathbb{E}_{\boldsymbol{z}\sim q_\phi(\boldsymbol{z}\,|\,\boldsymbol{x})}\left[\log p_\theta(\boldsymbol{x}\,|\,\boldsymbol{z}) + \log p(\boldsymbol{z}) - \log q_\phi(\boldsymbol{z}\,|\,\boldsymbol{x})\right] =: \mathrm{ELBO}_\Theta(\boldsymbol{x})$$

Ideally: $\mathcal{L} = \mathbb{E}_{\boldsymbol{x}\sim p_{\mathrm{data}}(\boldsymbol{x})}[\mathrm{ELBO}_\Theta(\boldsymbol{x})]$ In practice: $\mathcal{L} = \mathbb{E}_{\boldsymbol{x}\sim \mathcal{D}_{\mathrm{train}}}[\mathrm{ELBO}_\Theta(\boldsymbol{x})]$

A. ***model architecture and training algorithm***

B. ***training data*** (us)

What will be a better approximation for $p_{\mathrm{data}}(\boldsymbol{x})$ than $\mathcal{D}_{\mathrm{train}}$?

1. **a continuous distribution**;
2. **an accurate approximation of** $p_{\mathrm{data}}(\boldsymbol{x})$.

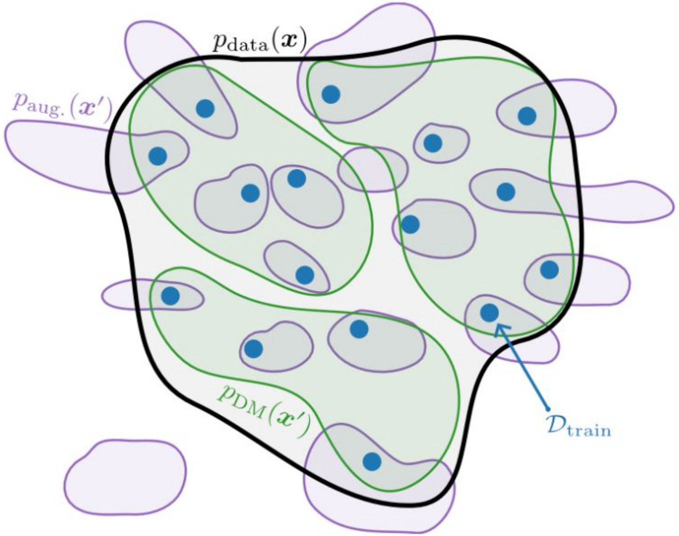
approx. by	$\mathcal{D}_{\mathrm{train}}$	$p_{\mathrm{aug}}(\boldsymbol{x}')$	$p_{\mathrm{DM}}(\boldsymbol{x}')$
(1) continuous	✗	✓	✓
(2) accurate	✓	✗	✓

$p_{\mathrm{aug}}(\boldsymbol{x}') = \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}_{\mathrm{train}}}[p_{\mathrm{aug}}(\boldsymbol{x}'\,|\,\boldsymbol{x})]$

A good diffusion model that has been pre-trained on $\mathcal{D}_{\mathrm{train}}$ *satisfies these two criteria!*

Unlimited Data Plans Provided by DMs

Idea: Diffusion Model as a $p_{\text{data}}(\boldsymbol{x})$ (short: DMaaPx)



Ideal:

$$\mathcal{L} = \mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}(\boldsymbol{x})}\left[\text{ELBO}_\Theta(\boldsymbol{x})\right] \quad (1)$$

Normal Training:

$$\mathcal{L} = \mathbb{E}_{\boldsymbol{x}\sim \mathcal{D}_{\text{train}}}\left[\text{ELBO}_\Theta(\boldsymbol{x})\right] \quad (2)$$

Augmentation:*

$$\mathcal{L} = \mathbb{E}_{\boldsymbol{x}\sim \mathcal{D}_{\text{train}}}\left[\mathbb{E}_{p_{\text{aug}}(\boldsymbol{x}' \mid \boldsymbol{x})}\left[\text{ELBO}_\Theta(\boldsymbol{x}')\right]\right] \quad (3)$$

DMaaPx (proposed):

$$\mathcal{L} = \mathbb{E}_{\boldsymbol{x}'\sim p_{\text{DM}}(\boldsymbol{x}')}\left[\text{ELBO}_\Theta(\boldsymbol{x}')\right] \quad (4)$$

Three Gaps to Evaluate the Impact

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

- **Generalization gap:** (*Generalization*)

$$\mathcal{G}_{\mathrm{g}} = \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}_{\mathrm{train}}}\left[\mathrm{ELBO}_{\Theta}(\boldsymbol{x})\right] - \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}_{\mathrm{test}}}\left[\mathrm{ELBO}_{\Theta}(\boldsymbol{x})\right]$$

- **Amortization gap:** (*Inference*)

$$\mathcal{G}_{\mathrm{a}} = \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}_{\mathrm{test}}}\left[\mathrm{ELBO}^{*}_{\theta}(\boldsymbol{x})\right] - \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}_{\mathrm{test}}}\left[\mathrm{ELBO}_{\Theta}(\boldsymbol{x})\right]$$

where $\mathrm{ELBO}^{*}_{\theta}(\boldsymbol{x}) = \mathbb{E}_{\boldsymbol{z}\sim q^{*}(\boldsymbol{z}\,|\,\boldsymbol{x})}\left[\log p_{\theta}(\boldsymbol{x}\,|\,\boldsymbol{z}) + \log p(\boldsymbol{z}) - \log q^{*}(\boldsymbol{z}\,|\,\boldsymbol{x})\right]$

- **Robustness gap:** (*Robustness*)

$$\mathcal{G}_{\mathrm{r}} = \mathbb{E}_{\boldsymbol{x}^{\mathrm{a}}\sim p(\boldsymbol{x}^{\mathrm{a}}\,|\,\boldsymbol{x}^{\mathrm{r}})}\,\mathbb{E}_{\boldsymbol{x}^{\mathrm{r}}\sim\mathcal{D}_{\mathrm{test}}}\left[\text{MS-SSIM}\left[\boldsymbol{x}^{\mathrm{r}},\boldsymbol{x}^{\mathrm{a}}\right] - \text{MS-SSIM}\left[\tilde{\boldsymbol{x}}^{\mathrm{r}},\tilde{\boldsymbol{x}}^{\mathrm{a}}\right]\right]$$

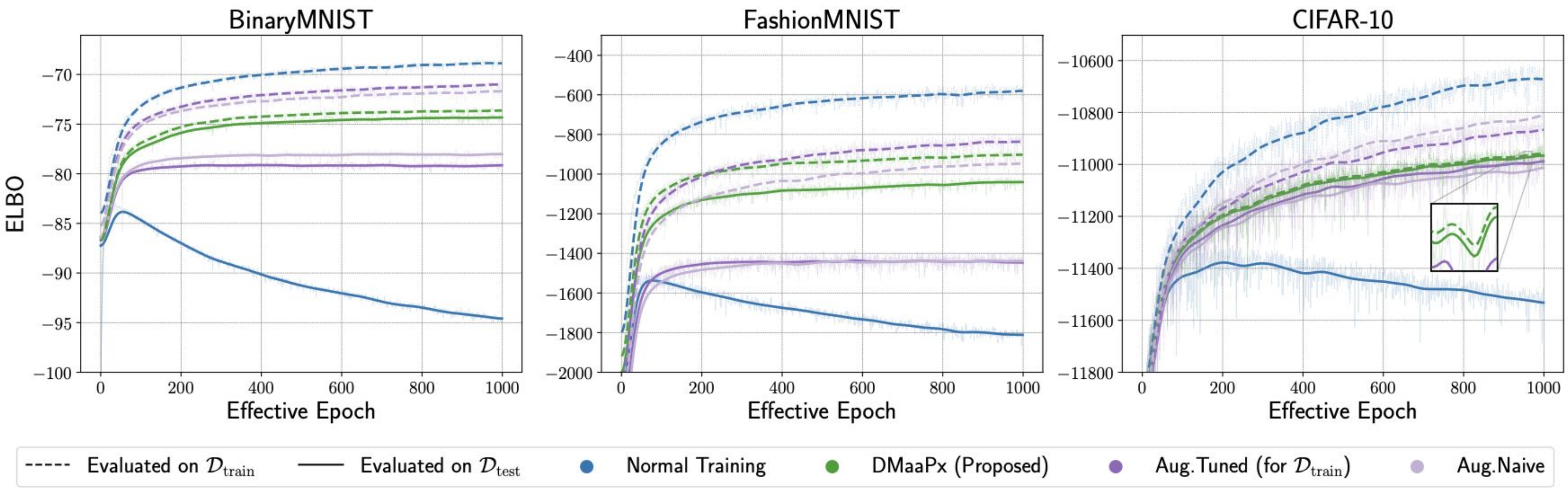
where $\boldsymbol{x}^{\mathrm{a}} = \boldsymbol{x}^{\mathrm{r}} + \boldsymbol{\epsilon}$ (s.t. $\|\boldsymbol{\epsilon}\| \leq \delta$) and $\boldsymbol{\epsilon} = \underset{\|\boldsymbol{\epsilon}\|_{\infty}\leq\delta}{\arg\max}\ \mathrm{SKL}\left[q_{\phi}(\boldsymbol{z}\,|\,\boldsymbol{x}^{\mathrm{r}}+\boldsymbol{\epsilon})\,\|\,q_{\phi}(\boldsymbol{z}\,|\,\boldsymbol{x}^{\mathrm{r}})\right]$

MS-SSIM is a image similarity metric, higher → more similar

(Kuzina et al., 2022)

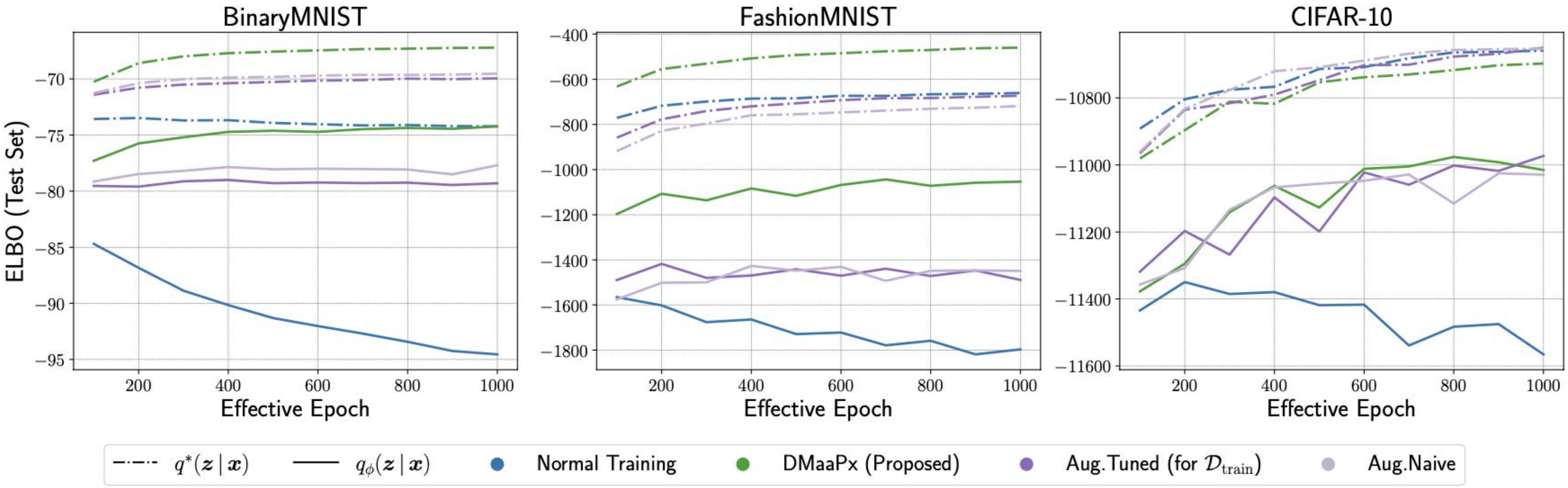
Exp. – Generalization Gap

EBERHARD KARLS UNIVERSITÄT TÜBINGEN



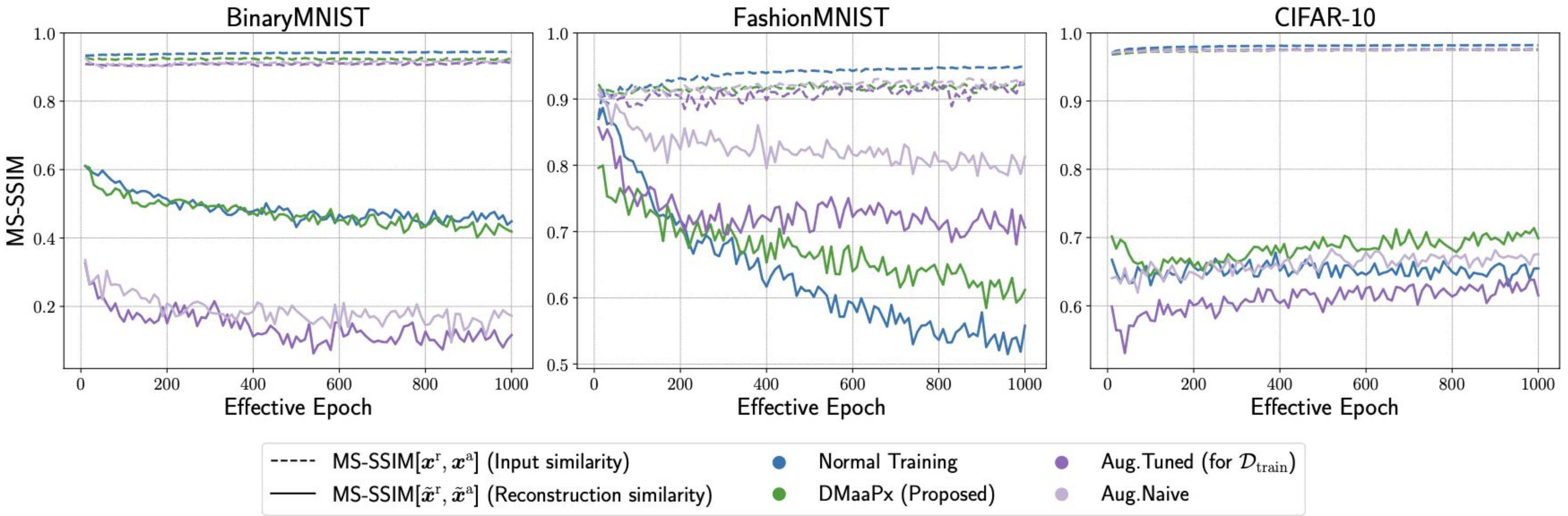
- DMaaPx has the highest ELBO on test set and smallest gap → training ELBOs can be used as accurate predictions for final performance.
- Overfitting in VAEs is more detrimental than using a somewhat distorted, but larger and more diverse, dataset.

Exp. – Amortization Gap



- Decreasing test set performance signals overfitting.
- Optimizing individual variational parameters rather than relying on the encoder, test performance stabilizes → overfitting in VAEs is mainly due to the encoder. Aligns with Cremer et al. (2018).
- DMaaPx outperforms others. The increase of ELBOs with q* suggests it also improves the decoder.

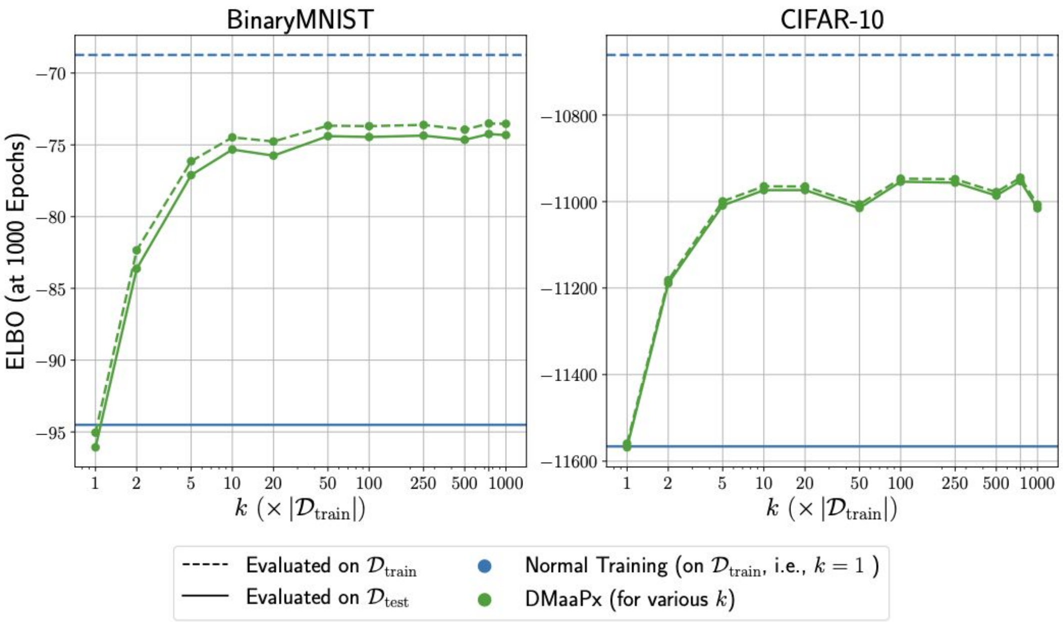
Exp. – Adversarial Robustness



- DMaaPx consistently matches or surpasses normal training.
- Augmentation displays inconsistent results
  → augmentation is more difficult to tune (requires more manual effort)

Is “Unlimited Data Plans” A Ripoff?

EBERHARD KARLS UNIVERSITÄT TÜBINGEN



Do we really need infinite number of samples? No! 10 times more seems to suffice.

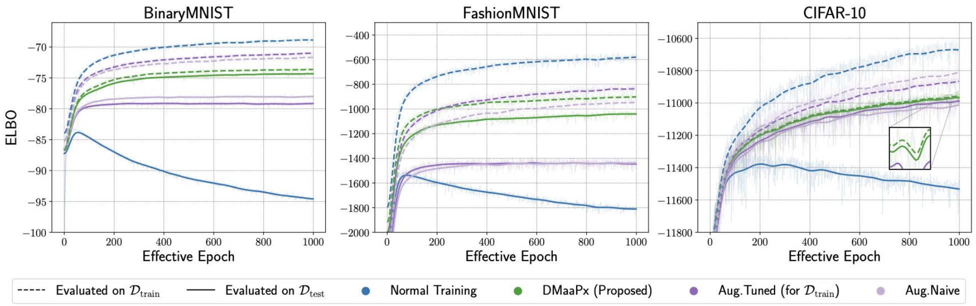
- For k = 1, DMaaPx slightly underperforms for MNIST but matches normal training for CIFAR.
- Performance increases with increasing k but plateaus for k ≥ 10.

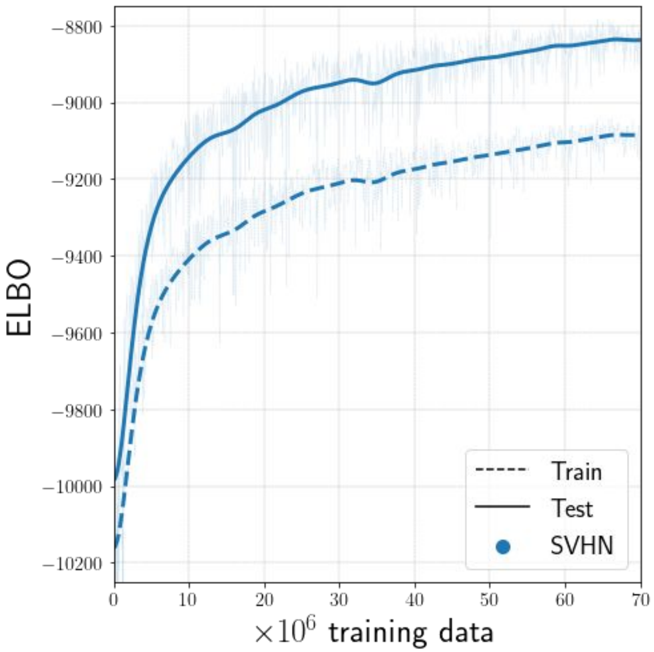
Part 2.1 – Summary

- Reduce overfitting of VAEs by training them on samples from pre-trained diffusion models
  - Improve generalization, amortized inference, and robustness
  - Don’t need to use unlimited data
- Remark: A Cross-Model-Class Distillation Perspective
  - Training data → Diffusion model → VAE
  - Useful despite data processing inequality
  - Some models have been designed with useful structures but cannot be fully exploited if trained naively.

We also tried SVHN, but …

EBERHARD KARLS UNIVERSITÄT TÜBINGEN





Distribution Mismatch in SVHN

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

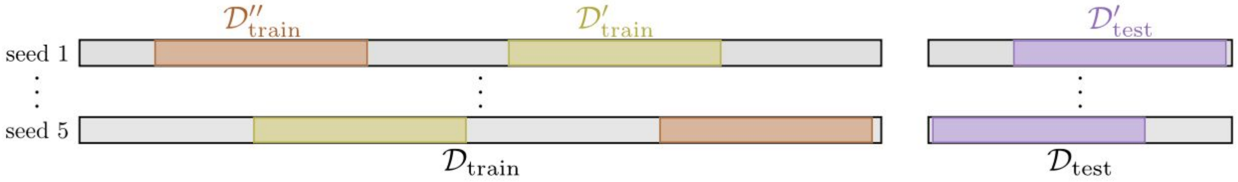
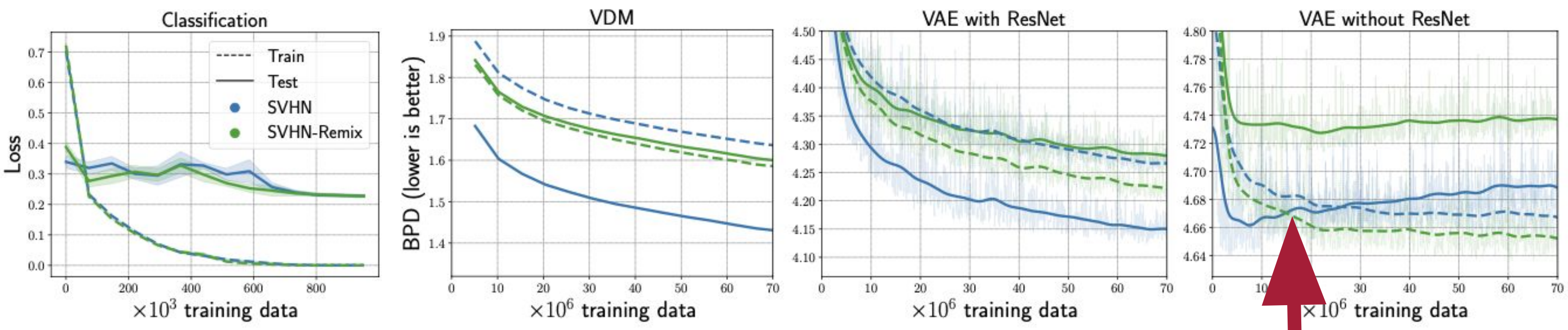


Figure 1: Five random splits (with reshuffle) of $\mathcal{D}_{\text{train}}$ and $\mathcal{D}_{\text{test}}$ into $\mathcal{D}'_{\text{train}}$, $\mathcal{D}''_{\text{train}}$, and $\mathcal{D}'_{\text{test}}$.

FID (↓), IS (↑)	SVHN	SVHN-Remix	CIFAR-10
$\text{FID}(\mathcal{D}''_{\text{train}}, \mathcal{D}'_{\text{train}})$	3.309 ± 0.029	3.334 ± 0.018	5.196 ± 0.040
$\text{FID}(\mathcal{D}''_{\text{train}}, \mathcal{D}'_{\text{test}})$	$\mathbf{16.687 \pm 0.325}$	3.326 ± 0.015	5.206 ± 0.031
$\text{IS}(\mathcal{D}'_{\text{train}} \mid \bar{\mathcal{D}}_{\text{train}})$	8.507 ± 0.114	8.348 ± 0.568	7.700 ± 0.043
$\text{IS}(\mathcal{D}'_{\text{test}} \mid \bar{\mathcal{D}}_{\text{train}})$	8.142 ± 0.501	8.269 ± 0.549	7.692 ± 0.023

Implications

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN



(a) Classification (b) VDM (c) VAEs

- Bits per dimension in proportion to negative log likelihood; lower is better
- The solid blue line first goes below the dashed blue line, then goes above it → overfitting!

Part 2.2 – Summary

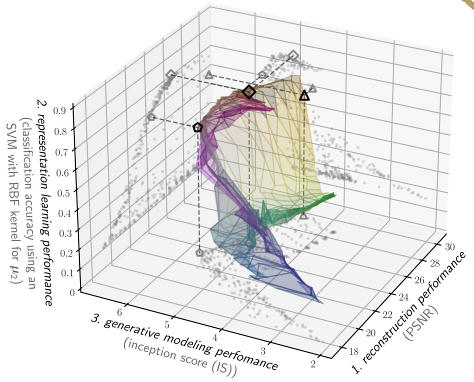
- There is a distribution mismatch in SVHN!

  (i.e., the training and test set is not from the same distribution)
- It affects the evaluation of probabilistic generative models, but not classification
- Lesson: When benchmarking generative models, we need to be mindful of distribution mismatch!

Conclusion

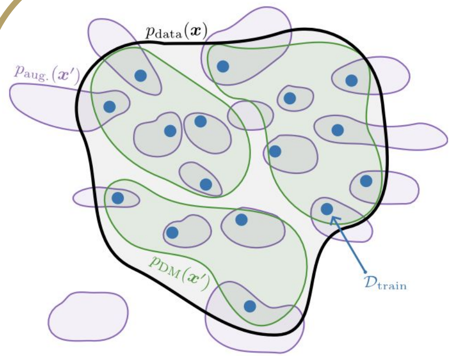
EBERHARD KARLS UNIVERSITÄT TÜBINGEN

What do we want from a generative model and how do we get it from a VAE?



No “One VAE that is optimal for all applications”, but we can optimize a VAE towards one application.

Loss function



Alleviate overfitting and improve generalization, amortized inference, and robustness by training with data generated from DMs.

SVHN	SVHN-Remix
3.309 ± 0.029	3.334 ± 0.018
16.687 ± 0.325	3.326 ± 0.015

Be mindful of training data distribution and distribution mismatch when benchmarking generative models.

Training Data

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

imprs-is

Reference:

- Part 1
  - Tim Z. Xiao, Robert Bamler.
    Trading Information between Latents in Hierarchical Variational Autoencoders.
    ICLR 2023. https://arxiv.org/abs/2302.04855
- Part 2.1
  - Tim Z. Xiao*, Johannes Zenn*, Robert Bamler
    Upgrading VAE Training With Unlimited Data Plans Provided by Diffusion Models.
    Preprint. https://arxiv.org/abs/2310.19653
- Part 2.2
  - Tim Z. Xiao*, Johannes Zenn*, Robert Bamler
    The SVHN Dataset Is Deceptive for Probabilistic Generative Models Due to a Distribution Mismatch.
    NeurIPS 2023 Workshop on Distribution Shifts. https://jzenn.github.io/svhn-remix

Personal website: https://timx.me/