Theoretical guarantees for Deep Generative Models: A PAC-Bayesian Approach

Sokhna Diarra Mbacke

Université Laval

sokhna-diarra.mbacke.1@ulaval.ca

March 2024

Co-authors

Summary

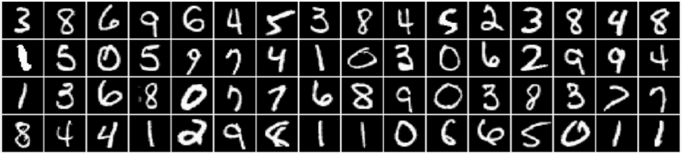
Context

- Generative models are widely used.
- Determining if a generative model generalizes well is a difficult problem.
- PAC-Bayes is a powerful tool in statistical learning theory.

Goal: Use PAC-Bayes to study the properties of generative models.

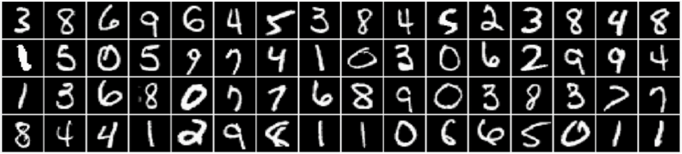
Next up

- Given finite iid samples:

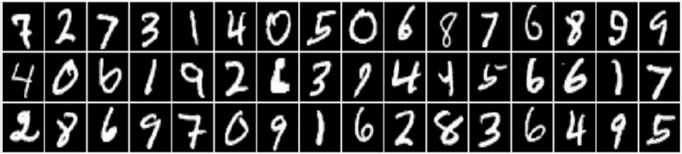


Generative Modelling

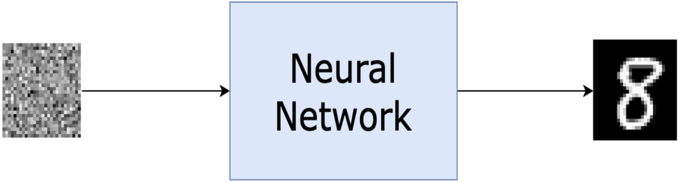
- Given finite iid samples:



- The goal is to learn to generate samples from the same distribution.



The goal is to learn a neural network that transforms noise into data.



Analyzing a Generative Model

Analyzing a generative model is a challenging because:

- The data-generating distribution is unknown;
- Unlike supervised learning, one cannot simply compute the accuracy on the test set;
- Different ways of defining the similarity between probability measures yield different results and have different interpretations.

Next up

PAC-Bayes

- PAC-Bayes provides high-probability generalization bounds for machine learning models.
- The theory requires very few assumptions, e.g. no assumption on the data-generating distribution.
- The bounds are numerically computable.

PAC-Bayes: Definitions

We consider the following concepts.

- An instance space $\mathcal{X}$ and an *unknown* distribution P^* on $\mathcal{X}$.
- A set $S = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ of observations iid sampled from P^*.
- A class $\mathcal{H}$ of models, called the hypothesis class.
- A loss function $\ell : \mathcal{H} \times \mathcal{X} \to [0, \infty)$.

Instead of individual hypotheses $h \in \mathcal{H}$, most PAC-Bayes bounds consider *aggregate* hypotheses $\rho \in \mathcal{M}_+^1(\mathcal{H})$.

PAC-Bayes: Risk

Given a loss function $\ell : \mathcal{H} \times \mathcal{X} \to [0, \infty)$, the empirical and true risks of $\rho \in \mathcal{M}_+^1(\mathcal{H})$ are defined as follows.

Empirical Risk

$$\hat{\mathcal{R}}_S(\rho) = \mathop{\mathbb{E}}_{h\sim\rho}\left[\frac{1}{n}\sum_{i=1}^{n}\ell(h, \mathbf{x}_i)\right]$$

True Risk

$$\mathcal{R}(\rho) = \mathop{\mathbb{E}}_{h\sim\rho}\left[\mathop{\mathbb{E}}_{\mathbf{x}\sim P^*}[\ell(h, \mathbf{x})]\right]$$

Definition: The KL divergence

Definition

Given probability distributions P, Q on $\mathcal{H}$ with densities p and q,

$$\mathrm{KL}(P \,||\, Q) = \int_{\mathcal{H}} p(h) \log \frac{p(h)}{q(h)} \, dh.$$

PAC-Bayes: Example

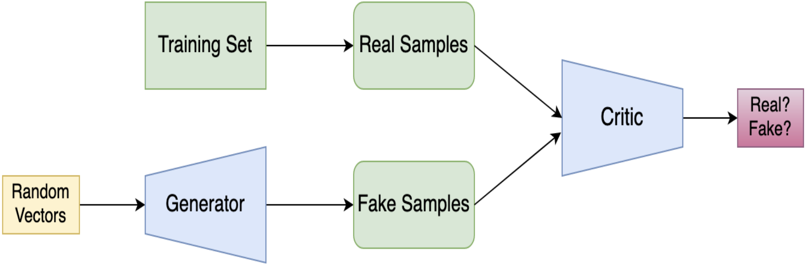
Theorem (Catoni (2003))

Given a distribution P^* over $\mathcal{X}$, a hypothesis class $\mathcal{H}$, a loss function $\ell : \mathcal{H} \times \mathcal{X} \to [0,1]$, a prior distribution π over $\mathcal{H}$, a real number $\delta \in (0,1)$, and a real number $\lambda > 0$, with probability at least $1-\delta$ over the choice of $S \overset{iid}{\sim} P^{*\otimes n}$, the following holds for any posterior distribution $\rho \in \mathcal{M}_+^1(\mathcal{H})$:

$$\mathcal{R}(\rho) \leq \hat{\mathcal{R}}_S(\rho) + \frac{\lambda}{8n} + \frac{\mathrm{KL}(\rho \,||\, \pi) + \log \frac{1}{\delta}}{\lambda}.$$

Next up

Generative Adversarial Networks (GANs) (Goodfellow et al., 2014)



We consider the following concepts.

- Instance Space: $\mathcal{X}$, data-generating distribution $P^* \in \mathcal{M}_+^1(\mathcal{X})$, and training set
$$S = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\} \stackrel{\text{iid}}{\sim} P^*.$$
- Generator Family $\mathcal{G}$: Each generator $g \in \mathcal{G}$ induces a distribution $P^g \in \mathcal{M}_+^1(\mathcal{X})$.
- Critic Family $\mathcal{F}$: A family $\mathcal{F}$ of functions $f : \mathcal{X} \to \mathbb{R}$.

Definition

Let $P, Q \in \mathcal{M}_+^1(\mathcal{X})$. The Wasserstein distance between P and Q is defined as

$$W_1(P, Q) = \sup_{f \in \mathrm{Lip}_1(\mathcal{X})} \left[\mathop{\mathbb{E}}_{\mathbf{x} \sim P} f(\mathbf{x}) - \mathop{\mathbb{E}}_{\mathbf{x} \sim Q} f(\mathbf{x}) \right],$$

where

$$\mathrm{Lip}_1(\mathcal{X}) = \{ f : \mathcal{X} \to \mathbb{R} \text{ s.t. } |f(\mathbf{x}) - f(\mathbf{y})| \leq d(\mathbf{x}, \mathbf{y}) \}.$$

The Wasserstein GAN

- The goal is to minimize the Wasserstein distance $W_1(P^*, P^g)$.
- $\mathrm{Lip}_1(\mathcal{X})$ is approximated by a subset $\mathcal{F} \subseteq \mathrm{Lip}_1(\mathcal{X})$ parameterized by a neural network.
- The optimization objective is

$$\min_{g \in \mathcal{G}} \max_{f \in \mathcal{F}} \left\{ \mathop{\mathbb{E}}_{\mathbf{x} \sim P^*} [f(\mathbf{x})] - \mathop{\mathbb{E}}_{\hat{\mathbf{x}} \sim P^g} [f(\hat{\mathbf{x}})] \right\}.$$

- In practice, these expectations are approximated using finite samples.

Risk

Given iid samples $S = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\} \overset{\text{iid}}{\sim} P^*$ and $S_g = \{\hat{\mathbf{x}}_1, \ldots, \hat{\mathbf{x}}_n\} \overset{\text{iid}}{\sim} P^g$, let P_n^* and P_n^g denote the corresponding empirical distributions:

$$P_n^* = \frac{1}{n}\sum_{i=1}^{n} \delta_{\mathbf{x}_i} \quad \text{and} \quad P_n^g = \frac{1}{n}\sum_{i=1}^{n} \delta_{\hat{\mathbf{x}}_i}.$$

Risk

Given iid samples $S = \{\mathbf{x}_1, \dots, \mathbf{x}_n\} \overset{\text{iid}}{\sim} P^*$ and $S_g = \{\hat{\mathbf{x}}_1, \dots, \hat{\mathbf{x}}_n\} \overset{\text{iid}}{\sim} P^g$, let P_n^* and P_n^g denote the corresponding empirical distributions:

$$P_n^* = \frac{1}{n}\sum_{i=1}^{n} \delta_{\mathbf{x}_i} \quad \text{and} \quad P_n^g = \frac{1}{n}\sum_{i=1}^{n} \delta_{\hat{\mathbf{x}}_i}.$$

We define the empirical risk of a hypothesis $g \in \mathcal{G}$ as :

$$\mathcal{W}_{\mathcal{F}}(P_n^*, P^g) = \mathop{\mathbb{E}}_{S_g}\left[d_{\mathcal{F}}(P_n^*, P_n^g)\right]$$

where

$$d_{\mathcal{F}}(P, Q) = \sup_{f \in \mathcal{F}} \left[\mathop{\mathbb{E}}_{\mathbf{x}\sim P}[f(\mathbf{x})] - \mathop{\mathbb{E}}_{\mathbf{x}\sim Q}[f(\mathbf{x})]\right].$$

Theorem for Bounded Instance Spaces

- The *n*-sized training set S is iid sampled from a distribution P^* on $\mathcal{X}$.
- Each generator $g \in \mathcal{G}$ induces a distribution $P^g \in \mathcal{M}_+^1(\mathcal{X})$.
- The prior distribution $\pi \in \mathcal{M}_+^1(\mathcal{G})$ is independent of S.
- $\lambda > 0$ and $\delta \in (0, 1)$ are some given real numbers.
- The critic family $\mathcal{F} \subseteq \mathrm{Lip}_1$ is symmetric.
- $(\mathcal{X}, d)$ is a metric space with finite diameter $\Delta = \sup_{\mathbf{x}, \mathbf{x}' \in \mathcal{X}} d(\mathbf{x}, \mathbf{x}')$.

Theorem for WGANs

Theorem (Mbacke et al. (2023a))

The following holds with probability $\geq 1-\delta$ over the random draw of S, for any $\rho \in \mathcal{M}_+^1(\mathcal{G})$:

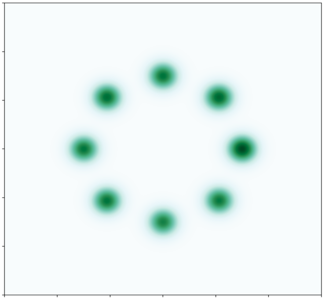
$$\mathop{\mathbb{E}}_{g\sim\rho}\mathop{\mathbb{E}}_{S}\left[\mathcal{W}_{\mathcal{F}}(P_n^*, P^g)\right] \leq \mathop{\mathbb{E}}_{g\sim\rho}\left[\mathcal{W}_{\mathcal{F}}(P_n^*, P^g)\right] + \frac{1}{\lambda}\left[\mathrm{KL}(\rho\,||\,\pi) + \log\frac{1}{\delta}\right] + \frac{\lambda\Delta^2}{4n},$$

where

$$\mathcal{W}_{\mathcal{F}}(P_n^*, P^g) = \mathop{\mathbb{E}}_{S_g}\left[d_{\mathcal{F}}(P_n^*, P_n^g)\right].$$

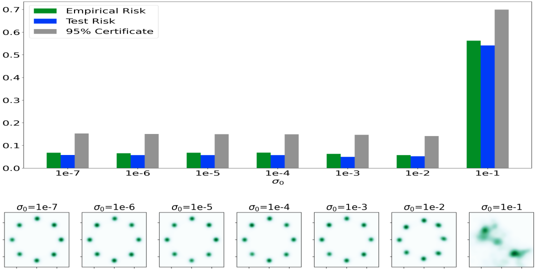
Next up

We performed experiments with a WGAN on the following Gaussian Mixtures:

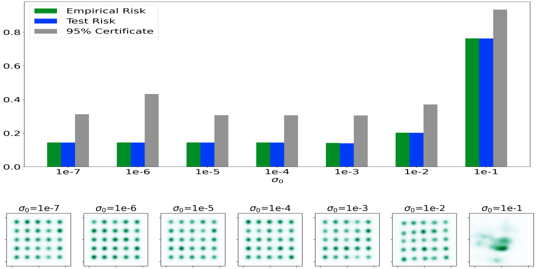


Objective: Determine the order of magnitude of the numerical values of the bounds.

Numerical values for the ring dataset:



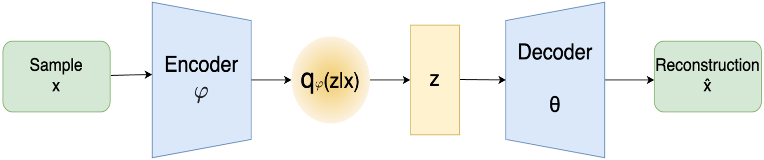
Numerical values for the grid dataset:



Next up

1. Context
2. Preliminary Concepts
   - Generative models
   - PAC-Bayesian Theory
3. Results for GANs
   - Bound for Wasserstein GANs
   - Numerical Experiments
4. Results for VAEs
   - Variational Autoencoders
   - Bound for the Generative Model
5. Conclusion

Variational Autoencoders (VAEs) (Kingma and Welling, 2014)

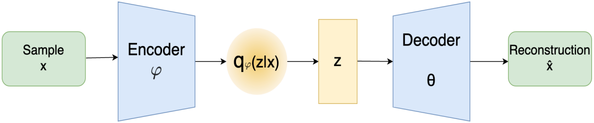


Variational Autoencoders: Definitions

We consider the following concepts.

- An instance space $\mathcal{X} \subseteq \mathbb{R}^D$, and a data-generating distribution $\mu \in \mathcal{M}_+^1(\mathcal{X})$.
- A latent space $\mathcal{Z} = \mathbb{R}^{d_{\mathcal{Z}}}$.
- A prior distribution $p(\mathbf{z}) = \mathcal{N}(\mathbf{0}, \mathbf{I})$ on the latent space.
- A posterior $q_\phi(\mathbf{z}|\mathbf{x})$ is parameterized by the encoder.

The Encoder

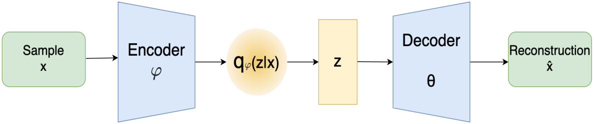


The encoder is a function

$$Q_\phi : \mathcal{X} \to \mathbb{R}^{2d_\mathcal{Z}}, \qquad Q_\phi(\mathbf{x}) = \begin{bmatrix} \mu_\phi(\mathbf{x}) \\ \sigma_\phi(\mathbf{x}) \end{bmatrix},$$

where the distribution $q_\phi(\mathbf{z}|\mathbf{x}) = \mathcal{N}\left(\mu_\phi(\mathbf{x}), \operatorname{diag}(\sigma^2_\phi(\mathbf{x}))\right)$.

The Decoder



The decoder is a function

$$g_\theta : \mathcal{Z} \to \mathcal{X}.$$

We assume g_θ is K_θ-Lipschitz:

$$\left\| g_\theta(\mathbf{z}_1) - g_\theta(\mathbf{z}_2) \right\| \leq K_\theta \left\| \mathbf{z}_1 - \mathbf{z}_2 \right\|.$$

The Optimization Objective

Given a training set $S = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$, minimize:

$$\mathcal{L}_{\mathsf{VAE}}(\phi, \theta) = \frac{1}{n} \sum_{i=1}^{n} \left[\underbrace{\mathop{\mathbb{E}}_{\mathbf{z} \sim q_\phi(\mathbf{z}|\mathbf{x}_i)} \ell_{\mathsf{rec}}^{\theta}(\mathbf{z}, \mathbf{x}_i)}_{\text{Reconstruction loss}} + \beta \underbrace{\mathrm{KL}(q_\phi(\mathbf{z}|\mathbf{x}_i) \,||\, p(\mathbf{z}))}_{\text{KL loss}} \right].$$

The Optimization Objective

Given a training set $S = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$, minimize:

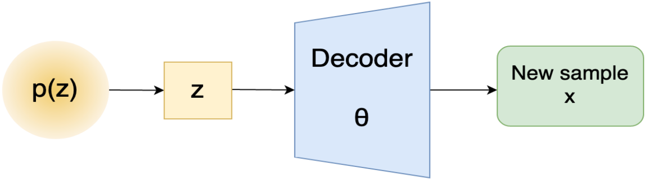
$$\mathcal{L}_{\mathsf{VAE}}(\phi, \theta) = \frac{1}{n} \sum_{i=1}^{n} \left[\underbrace{\mathop{\mathbb{E}}_{\mathbf{z} \sim q_\phi(\mathbf{z}|\mathbf{x}_i)} \ell^\theta_{\mathsf{rec}}(\mathbf{z}, \mathbf{x}_i)}_{\text{Reconstruction loss}} + \beta \underbrace{\mathrm{KL}(q_\phi(\mathbf{z}|\mathbf{x}_i) \,||\, p(\mathbf{z}))}_{\text{KL loss}} \right].$$

We define the reconstruction loss as: $\ell^\theta_{\mathsf{rec}} : \mathcal{Z} \times \mathcal{X} \to [0, \infty)$,

$$\ell^\theta_{\mathsf{rec}}(\mathbf{z}, \mathbf{x}) = \|\mathbf{x} - g_\theta(\mathbf{z})\|.$$

Next up

The VAE's generative model



- Once trained, the VAE defines the following generative model:

$$g_\theta \sharp p(\mathbf{z}).$$

- Our goal is to bound the distance:

$$W_1(\mu, g_\theta \sharp p(\mathbf{z})).$$

Generation Guarantees for Bounded Instance Spaces

- $\mu \in \mathcal{M}_+^1(\mathcal{X})$ is the data-generating distribution;
- $S = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\} \overset{\text{iid}}{\sim} \mu$ is a set of observed samples;
- $p(\mathbf{z}) = \mathcal{N}(\mathbf{0}, \mathbf{I})$ is the prior distribution on $\mathcal{Z}$;
- $\lambda > 0$ and $\delta \in (0, 1)$;
- $\mathcal{X}$ has finite diameter: $\Delta = \sup_{\mathbf{x}, \mathbf{x}' \in \mathcal{X}} d(\mathbf{x}, \mathbf{x}') < \infty$.

Generation Guarantees for Bounded Instance Spaces

Theorem (Mbacke et al. (2023b))

With probability at least $1-\delta$ *over the random draw of* $S \sim \mu^{\otimes n}$, *the following holds for any posterior* $q_\phi(\mathbf{z}|\mathbf{x})$:

$$W_1(\mu, g_\theta \sharp p(\mathbf{z})) \leq \frac{1}{n}\sum_{i=1}^{n}\left\{\mathop{\mathbb{E}}_{q_\phi(\mathbf{z}|\mathbf{x}_i)} \ell_{rec}^{\theta}(\mathbf{z}, \mathbf{x}_i)\right\} + \frac{1}{\lambda}\left(\sum_{i=1}^{n}\mathrm{KL}(q_\phi(\mathbf{z}|\mathbf{x}_i) \,||\, p(\mathbf{z})) + \log\frac{1}{\delta} + \frac{\lambda^2\Delta^2}{8n}\right) + \frac{K_\theta}{n}\sum_{i=1}^{n}\sqrt{\left\|\mu_\phi(\mathbf{x}_i)\right\|^2 + \left\|\sigma_\phi(\mathbf{x}_i) - \vec{1}\right\|^2}.$$

Conclusion

- Generative models are widely used in machine learning and difficult to analyze.
- PAC-Bayes is a powerful tool of statistical learning theory that can be used to analyze generative models (GANs, VAEs, diffusion models (Mbacke and Rivasplata, 2023)).
- PAC-Bayes bounds for generative models are empirical, hence they may enable new applications in practice.

References

Anil, C., Lucas, J., and Grosse, R. (2019). Sorting out lipschitz function approximation. In *International Conference on Machine Learning*, pages 291–301. PMLR.

Björck, Å. and Bowie, C. (1971). An iterative algorithm for computing the best estimate of an orthogonal matrix. *SIAM Journal on Numerical Analysis*, 8(2):358–364.

Catoni, O. (2003). A PAC-Bayesian approach to adaptive classification. *preprint*, LPMA 840.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative Adversarial Nets. In *Advances in Neural Information Processing Systems*, volume 27.

Kingma, D. P. and Welling, M. (2014). Auto-encoding Variational Bayes. In *International Conference on Learning Representations*.

Mbacke, S. D., Clerc, F., and Germain, P. (2023a). PAC-Bayesian Generalization Bounds for Adversarial Generative Models. In *Proceedings of the 40th International Conference on Machine Learning*, Proceedings of Machine Learning Research. PMLR.

Mbacke, S. D., Clerc, F., and Germain, P. (2023b). Statistical Guarantees for Variational Autoencoders using PAC-Bayesian Theory. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Mbacke, S. D. and Rivasplata, O. (2023). A Note on the Convergence of Denoising Diffusion Probabilistic Models. *arXiv preprint arXiv:2312.05989*.

Questions?

Experimental Details for WGANs

$$\mathop{\mathbb{E}}_{g\sim\rho}\mathop{\mathbb{E}}_{S}[\mathcal{W}_{\mathcal{F}}(P_n^*, P^g)] \leq \mathop{\mathbb{E}}_{g\sim\rho}[\mathcal{W}_{\mathcal{F}}(P_n^*, P^g)] + \frac{1}{\lambda}\left[\mathrm{KL}(\rho\,||\,\pi) + \log\frac{1}{\delta}\right] + \frac{\lambda\Delta^2}{4n}.$$

- WGAN with probabilistic layers for the generator.
- Lipschitz constraint with Björck Orthonormalization(Björck and Bowie, 1971) and GroupSort activations (Anil et al., 2019).
- We used part of the training set to learn the prior π, and the remaining part to compute the bound.
- The standard deviation of the prior's parameters $\sigma_0 \in \{10^{-7}, 10^{-6}, 10^{-5}, 10^{-4}, 0.001, 0.01, 0.1\}$.

Reconstruction Guarantees for Bounded Instance Spaces

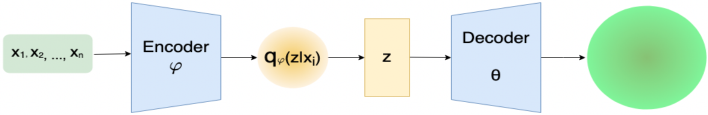
- $\mu \in \mathcal{M}_+^1(\mathcal{X})$ is the data-generating distribution;
- $S = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\} \overset{\text{iid}}{\sim} \mu$ is a set of observed samples;
- $p(\mathbf{z}) \in \mathcal{M}_+^1(\mathcal{Z})$ is the prior distribution on $\mathcal{Z}$;
- $\lambda > 0$ and $\delta \in (0, 1)$;
- $\mathcal{X}$ has finite diameter: $\Delta = \sup_{\mathbf{x}, \mathbf{x}' \in \mathcal{X}} d(\mathbf{x}, \mathbf{x}') < \infty$.

Theorem

Given a decoder θ, with probability at least $1-\delta$ over the random draw of $S \sim \mu^{\otimes n}$, the following holds for any posterior $q_\phi(\mathbf{z}|\mathbf{x})$:

$$\mathop{\mathbb{E}}_{\mathbf{x}\sim\mu} \mathop{\mathbb{E}}_{q_\phi(\mathbf{z}|\mathbf{x})} \ell^\theta_{rec}(\mathbf{z},\mathbf{x}) \leq \frac{1}{n}\sum_{i=1}^{n}\left\{ \mathop{\mathbb{E}}_{q_\phi(\mathbf{z}|\mathbf{x}_i)} \ell^\theta_{rec}(\mathbf{z},\mathbf{x}_i)\right\} + \frac{1}{\lambda}\sum_{i=1}^{n}\mathrm{KL}(q_\phi(\mathbf{z}|\mathbf{x}_i)\,||\,p(\mathbf{z})) + \frac{1}{\lambda}\log\frac{1}{\delta} + K_\phi K_\theta \Delta + \frac{\lambda\Delta^2}{8n}.$$

Regenerated Distribution



Define

$$\hat{\mu}_{\phi,\theta} = \frac{1}{n} \sum_{i=1}^{n} g_\theta \sharp q_\phi(\mathbf{z}|\mathbf{x}_i).$$

The triangle inequality implies

$$W_1(\mu, g_\theta \sharp p(\mathbf{z})) \leq W_1(\mu, \hat{\mu}_{\phi,\theta}) + W_1(\hat{\mu}_{\phi,\theta}, g_\theta \sharp p(\mathbf{z})).$$