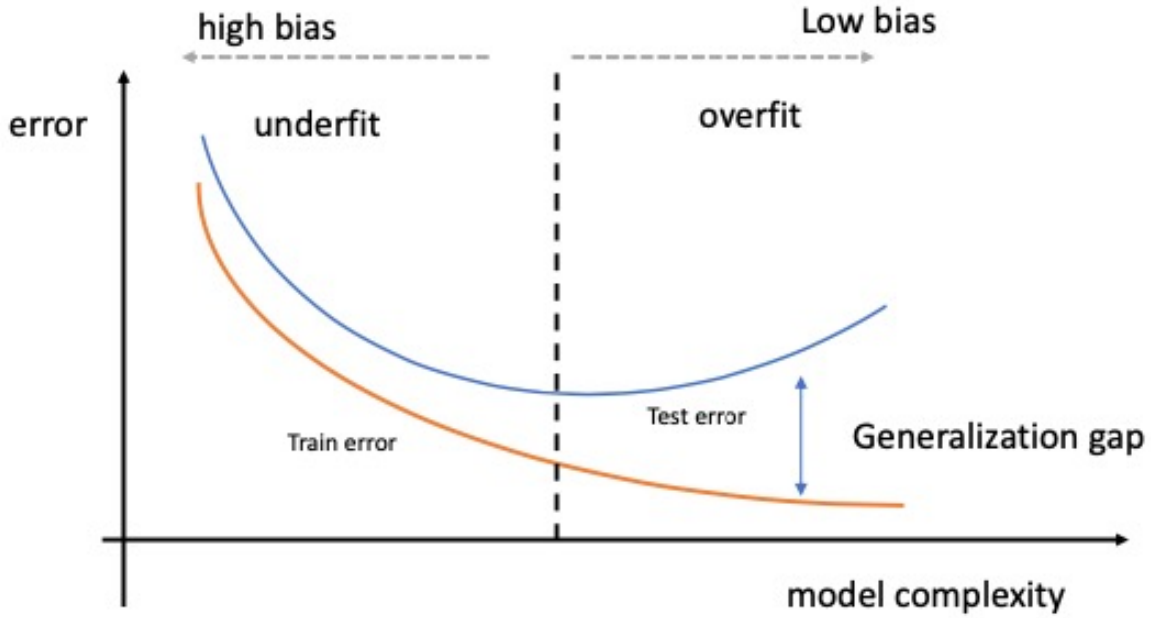
Theoretical understanding of generalization and memorization in generative models

Qi Chen
Principal Investigator, ELLIS Institute Finland
Assistant Professor, Aalto University
2026.3.12

Generalization of Discriminative Models



A supervised learning example:

$$\ell(\theta, x, y) = -\log p(y \mid x, \theta)$$

$$\mathcal{L}(\theta, P_{X,Y}) = \mathbb{E}_{x,y\sim P_{X,Y}}[\ell(\theta, x, y)]$$

$$\mathcal{L}(\theta, \hat{P}_{X,Y}) = -\sum_{i=1}^{m} \log p(y_i \mid x_i; \theta)$$

Bounding $\mathcal{L}(\theta, P_{X,Y}) - \mathcal{L}(\theta, \hat{P}_{X,Y})$ --- on expectation or high probability w.r.t dataset S

Generalization of Generative Models

An unsupervised learning example – density estimation:

$$\ell(\theta, x) = -\log p(x \mid \theta)$$

$$\mathcal{L}(\theta, P_X) = \mathbb{E}_{x \sim P_X}[\ell(\theta, x)]$$

$$\mathcal{L}(\theta, \hat{P}_X) = -\sum_{i=1}^{m} \log p(x_i \mid \theta)$$

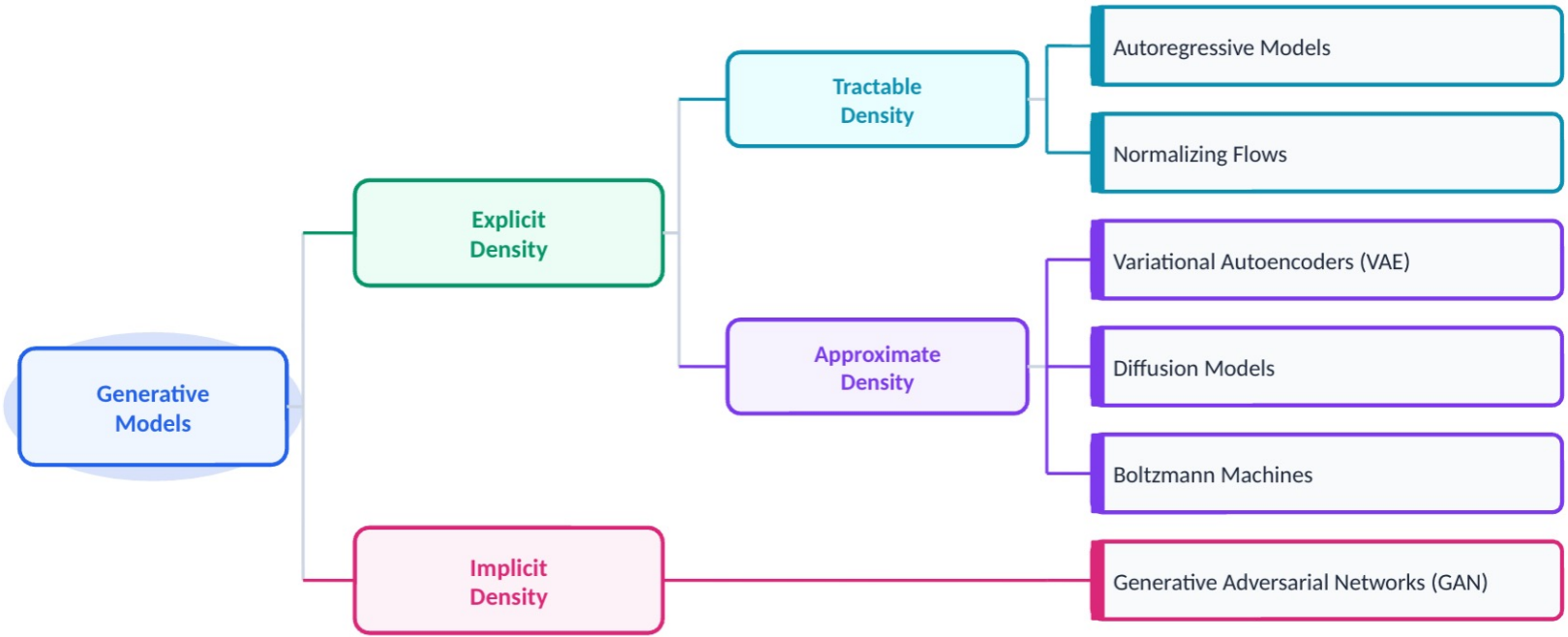
Cross entropy:

$$\begin{aligned}\mathbb{E}_{x \sim P_X}[l(\theta, x)] &= -\int p(x) \log p(x|\theta)\, dx \\ &= \underbrace{\int p(x) \log \frac{p(x)}{p(x|\theta)}\, dx}_{D_{\mathrm{KL}}(P_X \| P_{X|\theta})} + \underbrace{\int p(x) \log \left(\frac{1}{p(x)}\right) dx}_{H(P_X)}\end{aligned}$$

For measure the generalization of learning the density model:

$\text{Bounding } \mathcal{L}(\theta, P_X) - \mathcal{L}(\theta, \hat{P}_X)$ --- on expectation or high probability w.r.t dataset S

A Taxonomy of Generative Models



Overview of different types of generative models.

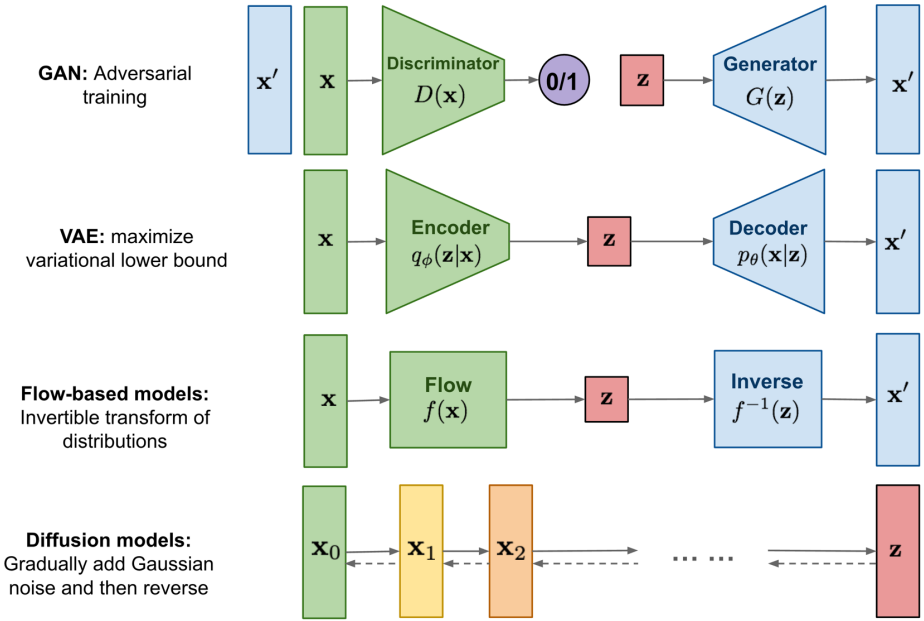


Figure source: https://lilianweng.github.io/posts/2021-07-11-diffusion-models/

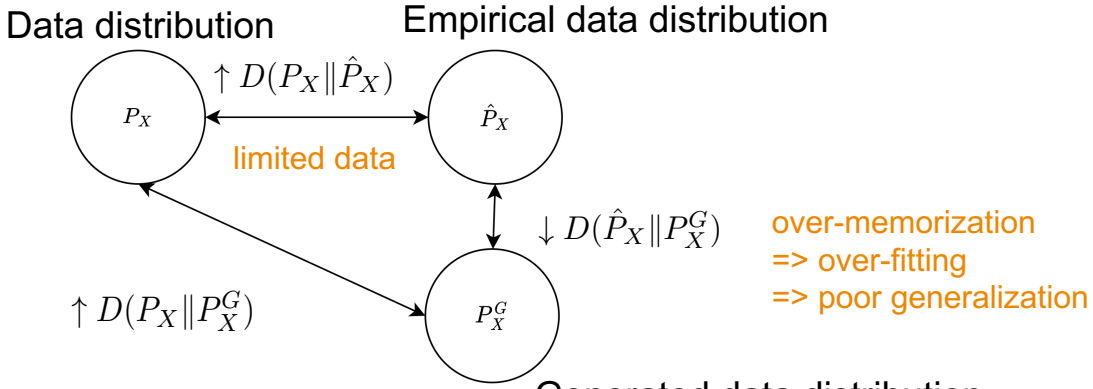
Memorization, hallucination and generalization

Objective: $\inf_G D(P_X \| G\#\pi)$

For some easy –to-sample distribution $\pi = \mathcal{N}(0, \sigma^2)$

The generated data distribution $P_X^G \stackrel{\text{def}}{=} G\#\pi$

The training data and original data distribution $\hat{P}_X, P_X$



Privacy and copyright issue caused by memorization

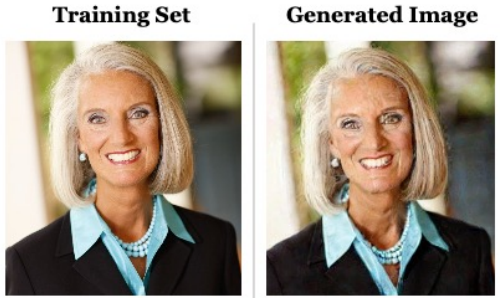


Figure 1: Diffusion models memorize individual training examples and generate them at test time. **Left:** an image from Stable Diffusion's training set (licensed CC BY-SA 3.0, see [49]). **Right:** a Stable Diffusion generation when prompted with "Ann Graham Lotz". The reconstruction is nearly identical (ℓ_2 distance = 0.031).

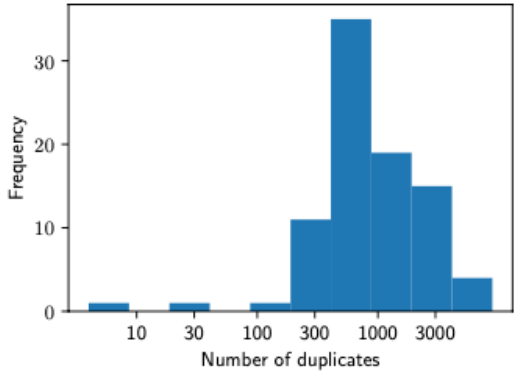
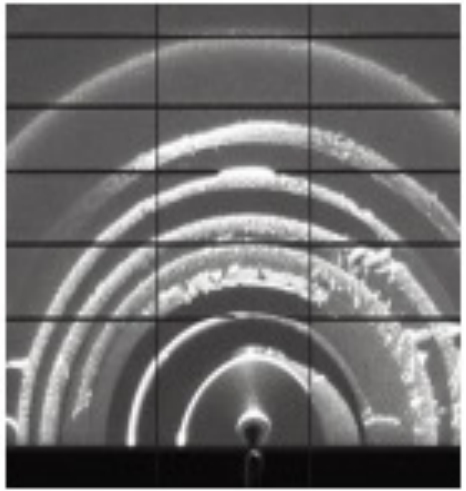


Figure 5: Our attack extracts images from Stable Diffusion most often when they have been duplicated at least $k = 100$ times; although this should be taken as an upper bound because our methodology explicitly searches for memorization of duplicated images.

"Diffusion models leak more than twice train data than GAN"

[1] Carlini, Nicolas, et al. "Extracting training data from diffusion models." *32nd USENIX Security Symposium (USENIX Security 23)*. 2023.

Hallucination



X-ray data: Real Fake

Generalization theory may provide implicit regularizations to:

- reduce memorization
- improve in-distribution generalization, may reduce hallucination as well

Generative Adversarial Networks (GAN)

- Directly learn G from data $\inf_G D_f(P_X \| G_\# P_Z)$

Definition 6 (**f-GAN (Nowozin et al., 2016)**) *Let $f : \mathbb{R} \to (-\infty, \infty]$ denote a convex function with property $f(1) = 0$ and $\mathcal{D} \subset \mathscr{F}(\mathcal{X}, \mathbb{R})$ a set of discriminators. The f-GAN model minimizes the following objective for a generator $G : \mathcal{Z} \to \mathcal{X}$*

$$\mathrm{GAN}_f(P_X, G; \mathcal{D}) := \sup_{d \in \mathcal{D}} \left\{ \mathbb{E}_{x \sim P_X}[d(x)] - \mathbb{E}_{z \sim P_Z}[f^*(d(G(z)))] \right\}, \tag{3}$$

where $f^\star(x) = \sup_y \{x \cdot y - f(y)\}$ is the convex conjugate of f.

Generative Adversarial Networks (GAN)

- Directly learn G from data $\inf_G W_1(P_X \| G_\# P_Z)$

Wasserstein GAN (WGAN). The Wasserstein-1 distance between P_X and $G_\# P_Z$ is defined via the Kantorovich–Rubinstein duality as

$$W_1(P_X, G_\# P_Z) = \sup_{\|f\|_L \le 1} \mathbb{E}_{x \sim P_X}[f(x)] - \mathbb{E}_{x \sim G_\# P_Z}[f(x)],$$

where the supremum is taken over all 1-Lipschitz functions f.

In WGAN, the discriminator is replaced by a critic f_w constrained to be 1-Lipschitz, and the objective is

$$\min_G \max_{f_w \in \mathcal{F}} \mathbb{E}_{x \sim P_X}[f_w(x)] - \mathbb{E}_{z \sim P_Z}[f_w(G(z))].$$

VAE and WAE

- Indirectly learn G from data

$$\inf_G D_{KL}(P_X \| G\#P_Z) \leq \inf_{G\in\mathcal{G}, E\in\mathcal{E}} D_{KL}(P_X \times E(X) \| P_Z \times G(Z))\,.$$

Definition 5 (Wasserstein Autoencoder (Tolstikhin et al., 2017)) *Let* $c : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_{\geq 0}$, $\lambda > 0$ *and* $\Omega : \mathscr{P}(\mathcal{Z}) \times \mathscr{P}(\mathcal{Z}) \to \mathbb{R}_{\geq 0}$ *with* $\Omega(P, P) = 0$ *for all* $P \in \mathscr{P}(\mathcal{Z})$. *The Wasserstein Autoencoder objective is*

$$\mathrm{WAE}_{c,\lambda\cdot\Omega}(P_X, G) = \inf_{E\in\mathscr{F}(\mathcal{X},\mathscr{P}(\mathcal{Z}))} \left\{ \int_{\mathcal{X}} \mathbb{E}_{z\sim E(x)}[c(x, G(z))] dP_X(x) + \lambda \cdot \Omega(E\#P_X, P_Z) \right\}$$

We remark that there are various choices of c and $\lambda \cdot \Omega$. Tolstikhin et al. (2017) select these by tuning λ and selecting different probability distortions for Ω.

Generalization Bounds for GANs

Theorem 14 *Let $(\mathcal{X}, c)$ be a metric space and suppose $\Delta := \max\{\Delta_{c,P_X}, \Delta_{c,P_G}\} < \infty$. For any $n \in \mathbb{N}_*$, let $\hat{P}_X$ and $\hat{P}_G$ denote the empirical distribution with n samples drawn i.i.d from P_X and P_G respectively. Let $s_X > d^*(P_X)$ and $s_G > d^*(P_G)$. For all $f : \mathbb{R} \to (-\infty, \infty]$ convex functions, $f(1) = 0$ and $\lambda > 0$, we have*

$$\mathrm{GAN}_{\lambda f}(P_X, G; \mathcal{H}_c) \leq \overline{W}_{c,\lambda\cdot f}(\hat{P}_X, P_G) + O\left(n^{-1/s_X} + \Delta\sqrt{\frac{1}{n}\ln\left(\frac{1}{\delta}\right)}\right), \tag{9}$$

with probability at least $1 - \delta$ for any $\delta \in (0, 1)$ and if $f(x) = |x - 1|$ is chosen then we have for all $\lambda > 0$

$$\mathrm{GAN}_{\lambda f}(P_X, G; \mathcal{H}_c) \leq \overline{W}_{c,\lambda\cdot f}(\hat{P}_X, \hat{P}_G) + O\left(n^{-1/s_X} + n^{-1/s_G} + \Delta\sqrt{\frac{1}{n}\ln\left(\frac{1}{\delta}\right)}\right), \tag{10}$$

with probability at least $1 - \delta$ for any $\delta \in (0, 1)$.

Husain et al. 2019

Husain, Hisham, Richard Nock, and Robert C. Williamson. "Adversarial networks and autoencoders: The Primal-Dual relationship and generalization bounds." *arXiv preprint arXiv:1902.00985* (2019)
Weed, J. and Bach, F. Sharp asymptotic and finite-sample rates of convergence of empirical measures in Wasserstein distance. Bernoulli, 25(4A):2620–2648, 2019.

Generalization Bounds for GANs

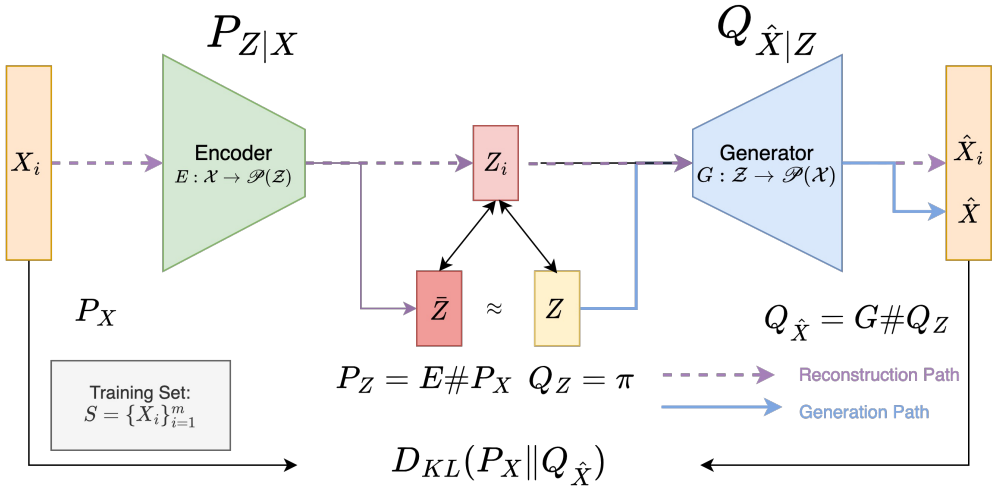
(**Generalization Bounds for WGAN**) For any probability measure ρ over $\mathcal{G}$ such that $\rho \ll \pi$, with probability at least $1-\delta$ over the draw of S,

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

Mbacke et al. 2023

Mbacke, Sokhna Diarra, Florence Clerc, and Pascal Germain. "PAC-Bayesian generalization bounds for adversarial generative models." *International Conference on Machine Learning*. PMLR, 2023.
Weed, J. and Bach, F. Sharp asymptotic and finite-sample rates of convergence of empirical measures in Wasserstein distance. Bernoulli, 25(4A):2620–2648, 2019.

Generalization in encoder-decoder generative models



Generation Error <= Reconstruction Error +

$$\frac{\sqrt{2}R}{m}\sum_{i=1}^{m}\sqrt{\mathbb{E}_{X_i}[\mathbb{D}_{KL}(E(X_i)\|\pi)]} + \frac{\sqrt{2}R}{m}\sum_{i=1}^{m}\sqrt{I(\hat{X}_i;X_i|Z_i)}$$

Generalization of Encoder **Generalization of Generator**

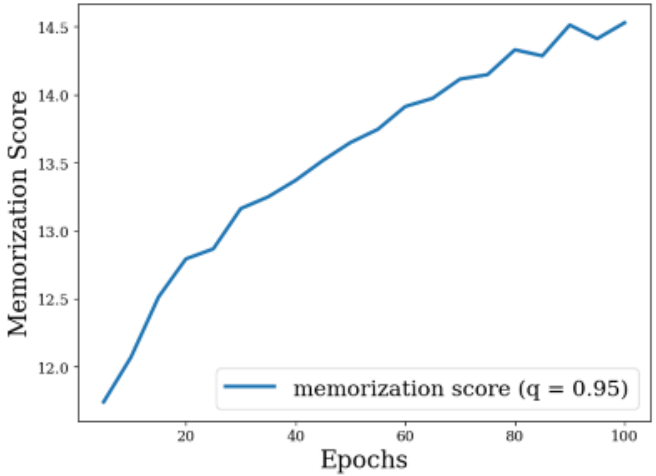
Qi Chen, Jerry Zhu, Florian Shkurti. Generalization in VAE and Diffusion Models: A Unified Information-Theoretic Analysis. ICLR 2025

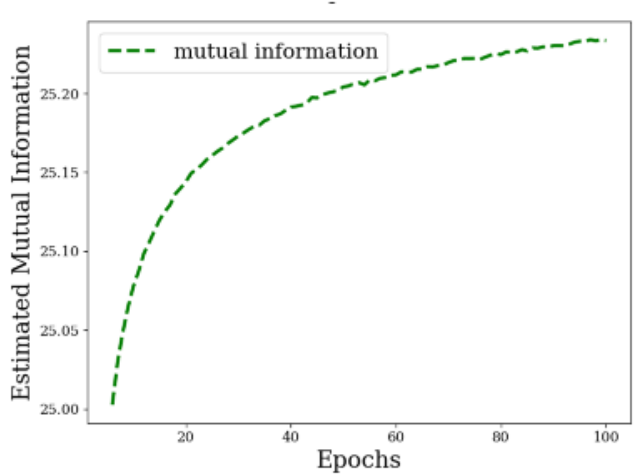
Generalization Bounds for VAEs

$$\mathbb{D}_{W_1}(P_X \| Q^{\pi}_{G_\theta}) \le \mathbb{E}_S \mathcal{L}_{\hat{P}_X}(E_\phi, G_\theta) + \frac{\sqrt{2}R}{m}\sum_{i=1}^{m}\sqrt{\mathbb{E}_{X_i}[\mathbb{D}_{KL}(E(X_i)\|\pi)]} + \frac{\sqrt{2}R}{m}\sum_{i=1}^{m}\sqrt{I(\hat{X}_i; X_i|Z_i)}$$

Distortion + Rate = VAE Objective

generalization for generator

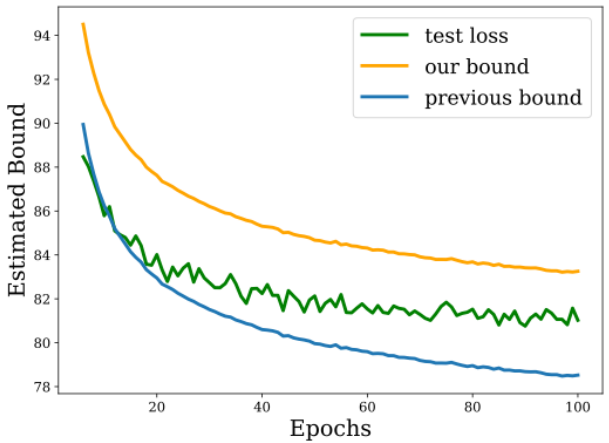




Memorization score: $M^{\mathrm{LOO}}(\mathcal{A}, S, i) = \log P_{\mathcal{A}}(\mathbf{x}_i|S) - \log P_{\mathcal{A}}(\mathbf{x}_i|S_{[n]\setminus\{i\}})$. [(Van den Burg & Williams, 2021]

Qi Chen, Jerry Zhu, Florian Shkurti. Generalization in VAE and Diffusion Models: A Unified Information-Theoretic Analysis. ICLR 2025

Generalization Bounds for VAEs



Optimizing the empirical VAE objective over epochs

Rewrite the bound of Mbake et al. :

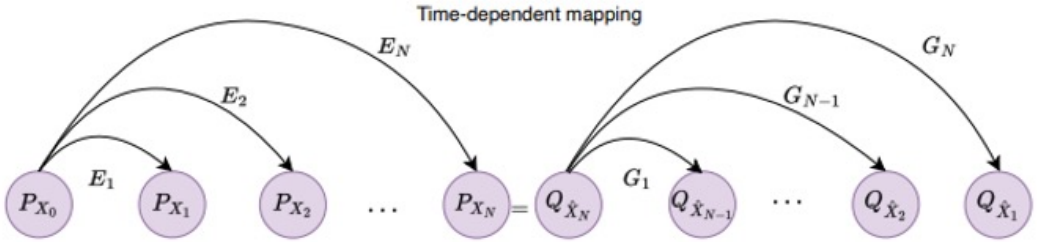
$$D_{W_1}(P_X \| Q^{\pi}_{G_\theta}) \leq \mathbb{E}_S[\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}_{Z\sim E_\phi(X_i)}\mathbb{E}_{\hat{X}\sim G_\theta(Z)}\|\hat{X}-X_i\|] + \frac{1}{\lambda}\mathbb{E}_S[\sum_{i=1}^{m}\mathbb{D}_{KL}(E_\phi(X_i)\|\pi)] + \frac{\lambda\Delta^2}{8m} + \frac{K_\theta}{m}\mathbb{E}_S\sum_{i=1}^{m}\mathbb{D}_{W_2}(E_\phi(X_i)\|\pi)\,,$$

$\Delta := sup_{x,x'}\|x-x'\|$ is the diameter of the bounded input space, K_θ is the Lipchitz constant of encoder.

Mbacke et al. (2024) utilize the triangle inequality for the Wasserstein distance, separately bounding $D_{W_1}(P_X \| G_\theta \# \hat{P}_Z^{E_\phi})$ and $D_{W_1}(G_\theta \# \hat{P}_Z^{E_\phi} \| G_\theta \# \pi)$.

Sokhna Diarra Mbacke, Florence Clerc, and Pascal Germain. Statistical guarantees for variational autoencoders using pac-bayesian theory. Advances in Neural Information Processing Systems, 36, 2024.

Generalization for Diffusion Models



$N = \frac{T}{\tau}$, τ is the step size.

$$D_{KL}(P_X \| G_T^\theta \# \pi) \le \mathbb{E}_S \Big(\underbrace{-\frac{1}{m}\sum_{i=1}^{m} D_{KL}(E_T(X_i) \| E_T \# \hat{P}_X)}_{T_1} + \hat{\mathcal{L}}_{ESM}(\theta, \lambda(\cdot)) \Big) + \underbrace{\frac{\sqrt{2}R}{m}\sum_{i=1}^{m}\sqrt{\mathbb{E}_{X_i}[D_{KL}(E_T(X_i)\|\pi)]}}_{T_2} + \underbrace{\frac{\sqrt{2}R}{m}\sum_{i=1}^{m}\sqrt{I(\hat{X}_0; X_i|\hat{X}_T)}}_{T_3}.$$

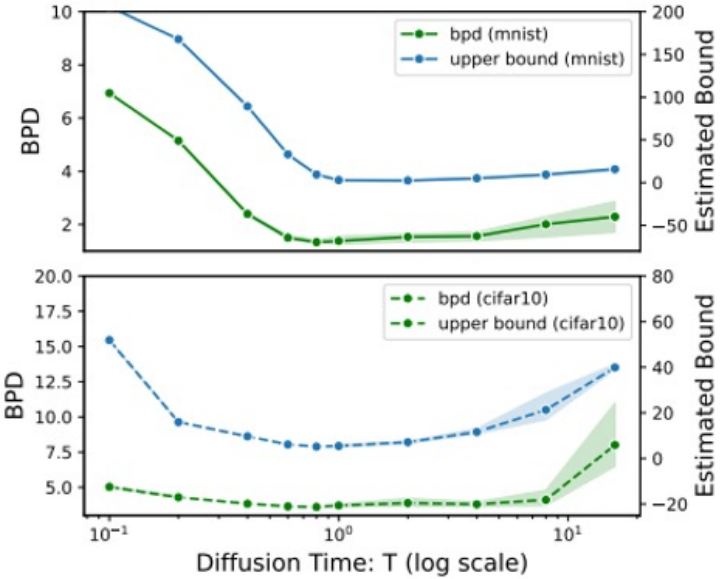
Score matching loss

Trade-off on diffusion time T: $T_1 \to 0, T_2 \to 0, T_3 \to \infty$ as $T \to \infty$.

Qi Chen, Jerry Zhu, Florian Shkurti. Generalization in VAE and Diffusion Models: A Unified Information-Theoretic Analysis. ICLR 2025

Generalization for Diffusion Models

Co-existence of
Memorization &
Hallucination



To achieve given performance,
choosing optimal T can implicitly improve data efficiency

Generalization and sample complexity

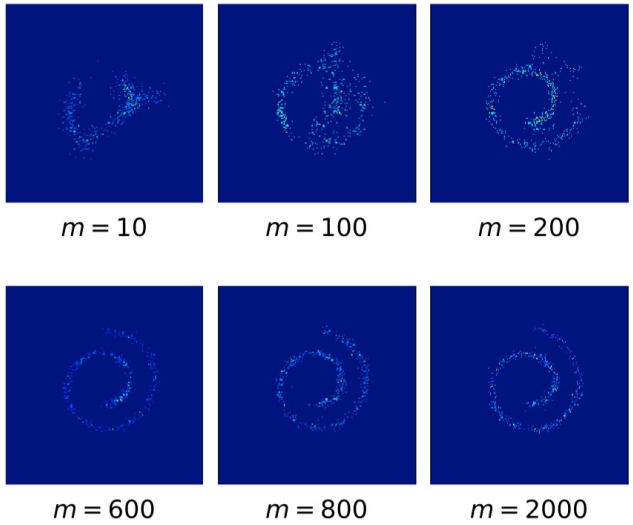
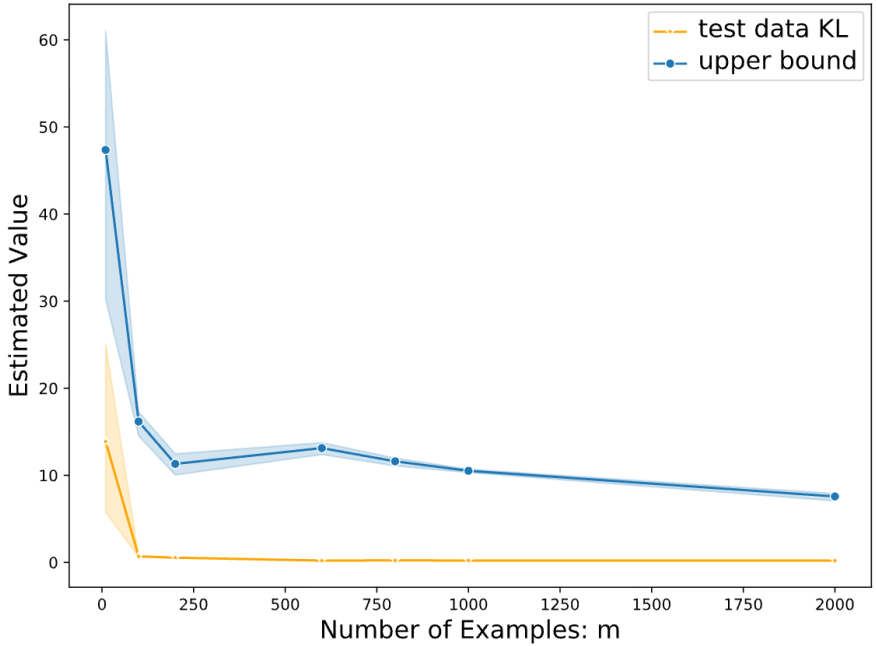


Figure 5: Sampling results w.r.t. different train data size m: 1000 data points generated by a score-based model trained with 10000 gradient iterations and diffusion time $T = 1$. The sampling is conducted after 1000 steps when solving the discretized backward SDE.



Theoretical Results that Consider Training Dynamics

- Our work (Chen et al. 2025)
  - assumes a sufficient small score matching loss -> enough training epochs
  - does not analyze the convergence for optimization error

- Puheng Li, Zhong Li, Huishuai Zhang, and Jiang Bian. On the generalization properties of diffusion models. Advances in Neural Information Processing Systems, 36, 2024.

- Bonnaire, Tony, et al. "Why Diffusion Models Don't Memorize: The Role of Implicit Dynamical Regularization in Training." Advances in Neural Information Processing Systems (2025).

Theoretical Results that Consider Training Dynamics

Theorem 1. *Suppose that the target distribution p_0 is continuously differentiable and has a compact support set, i.e., $||\boldsymbol{x}||_\infty$ is uniformly bounded, and there exists a reproducing kernel Hilbert space (RKHS) $\mathcal{H}$ $(:=\mathcal{H}_{k_{\rho_0}})$ such that $\bar{\boldsymbol{s}}_{0,\bar{\boldsymbol{\theta}}^*} \in \mathcal{H}$. Assume that the initial loss, trainable parameters, the embedding function $\boldsymbol{e}(t)$ and weighting function $\lambda(t)$ are all bounded. Then for any $\delta > 0$, $\delta \ll 1$, with the probability of at least $1-\delta$, we have*

$$D_{\mathrm{KL}}\left(p_0 \| p_{0,\hat{\boldsymbol{\theta}}_n(\tau)}\right) \lesssim \left[\frac{\tau^4}{mn} + \frac{\tau^3}{m^2} + \frac{1}{\tau}\right] + \left[\frac{1}{m} + \bar{\tilde{\mathcal{L}}}\left(\bar{\boldsymbol{\theta}}^*\right) + \tilde{\mathcal{L}}\left(\boldsymbol{\theta}^*\right)\right] + D_{\mathrm{KL}}\left(p_T \| \pi\right), \quad \tau \geq 1,$$

where $\lesssim$ hides the term $d\log(d+1)$, the polynomials of $\log(1/\delta^2)$, finite RKHS norms and universal positive constants only depending on T.

Puheng Li, Zhong Li, Huishuai Zhang, and Jiang Bian. On the generalization properties of diffusion models. Advances in Neural Information Processing Systems, 36, 2024.

Theoretical Results that Consider Training Dynamics

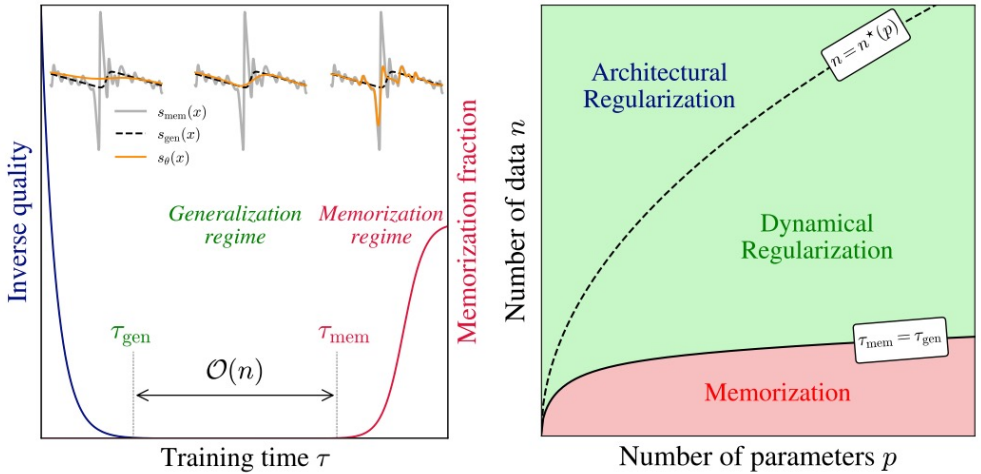


Figure 1: **Qualitative summary of our contributions.** *(Left)* Illustration of the training dynamics of a diffusion model. Depending on the training time τ, we identify three regimes measured by the inverse quality of the generated samples (blue curve) and their memorization fraction (red curve). The generalization regime extends over a large window of training times which increases with the training set size n. On top, we show a one dimensional example of the learned score function during training (orange). The gray line gives the exact empirical score, at a given noise level, while the black dashed line corresponds to the true (population) score. *(Right)* Phase diagram in the (n, p) plane illustrating three regimes of diffusion models: Memorization when n is sufficiently small at fixed p, Architectural Regularization for $n > n^\star(p)$ (which is model and dataset dependent, as discussed in [15, 25]), and Dynamical Regularization, corresponding to a large intermediate generalization regime obtained when the training dynamics is stopped early, i.e. $\tau \in [\tau_{\text{gen}}, \tau_{\text{mem}}]$.

Early stopping can improve generalization

Memorization normally happens when n << p
Model too simple => underfit

Bonnaire, Tony, et al. "Why Diffusion Models Don't Memorize: The Role of Implicit Dynamical Regularization in Training." Advances in Neural Information Processing Systems (2025).

Theoretical Results that Consider Training Dynamics

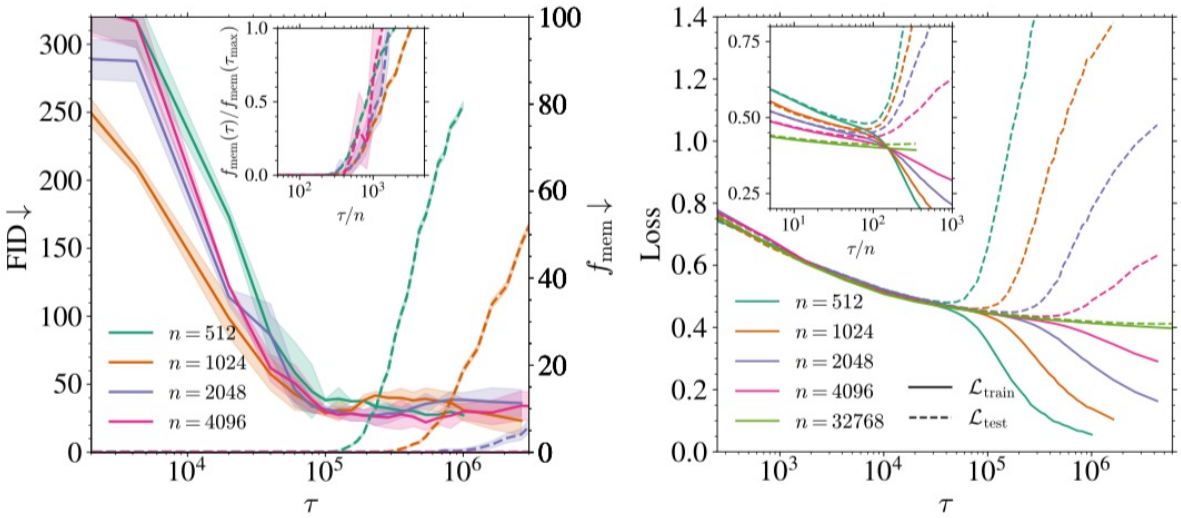


Figure 2: Memorization transition as a function of the training set size n for U-Net score models on CelebA. *(Left)* FID (solid lines, left axis) and memorization fraction f_{mem} (in %, dashed lines, right axis) against training time τ for various n. Inset: normalized memorization fraction $f_{\text{mem}}(\tau)/f_{\text{mem}}(\tau_{\max})$ with the rescaled time τ/n. *(Right)* Training (solid lines) and test (dashed lines) loss with τ for several n at fixed $t = 0.01$. Inset: both losses plotted against τ/n. Error bars on the losses are imperceptible.

Bonnaire, Tony, et al. "Why Diffusion Models Don't Memorize: The Role of Implicit Dynamical Regularization in Training." Advances in Neural Information Processing Systems (2025).

Thanks for your attention!