On the Generalization Properties of Diffusion Models

Puheng Li*, Zhong Li*, Huishuai Zhang, Jiang Bian

Stanford & MSRA

Nov 10, 2023

Background

- **Diffusion model:** A generative model establishing a stochastic transport map between an empirically observed, yet unknown, target distribution and a known prior.

Background

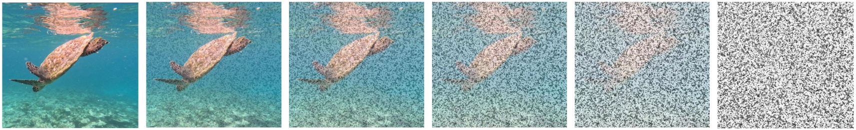
- **Diffusion model:** A generative model establishing a stochastic transport map between an empirically observed, yet unknown, target distribution and a known prior.
- Real-world applications: DALL·E, Imagen, Stable Diffusion...

Background

- **Diffusion model:** A generative model establishing a stochastic transport map between an empirically observed, yet unknown, target distribution and a known prior.
- Real-world applications: DALL·E, Imagen, Stable Diffusion...
- Theoretical foundations of diffusion models remain under-explored, particularly, the fundamental **generalization** problem

Formulation

$$\boldsymbol{x}(0) \longrightarrow \quad d\boldsymbol{x} = \boldsymbol{f}(\boldsymbol{x}, t)dt + g(t)d\boldsymbol{W}_t \quad \longrightarrow \boldsymbol{x}(T)$$



$$\boldsymbol{x}(0) \longleftarrow \quad d\boldsymbol{x} = \left[\boldsymbol{f}(\boldsymbol{x}, t) - g^2(t)\underbrace{\nabla_{\boldsymbol{x}} \log p_t(\boldsymbol{x})}_{\approx\, \boldsymbol{s}_{t,\boldsymbol{\theta}}(\boldsymbol{x}) := \frac{1}{m}\boldsymbol{A}\sigma(\boldsymbol{W}\boldsymbol{x} + \boldsymbol{U}\boldsymbol{e}(t)),\, \boldsymbol{\theta} = \boldsymbol{A}}\right]dt + g(t)d\bar{\boldsymbol{W}}_t \quad \longleftarrow \boldsymbol{x}(T)$$

Target: Finitely-supported prob. & Gaussian mixtures

Loss: Time-dependent score matching (Eq. (7))

Algorithm: Gradient flow

Notations:

t: SDE time $\quad$ T: maximal SDE time

$p_T \approx \pi$: a known prior

τ : training time

Main Result I

Theorem (Data-Independent Generalization Gap)

Suppose p_0 is continuously differentiable with a **compact support** *set, 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 the initial loss, trainable parameters, the embedding function $\boldsymbol{e}(t)$ and weighting function $\lambda(t)$ are all bounded. Then with high probability, we have*

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

- Early-stopping generalization gap:

$$\tau_{\mathrm{es}} = \Theta\left(n^{\frac{2}{5}}\right) \Rightarrow D_{\mathrm{KL}}\left(p_0 \| p_{0,\hat{\boldsymbol{\theta}}_n(\tau_{\mathrm{es}})}\right) \lesssim (1/n)^{\frac{2}{5}} + (1/m)^{\frac{4}{5}}.$$

Early-Stopping Generalization

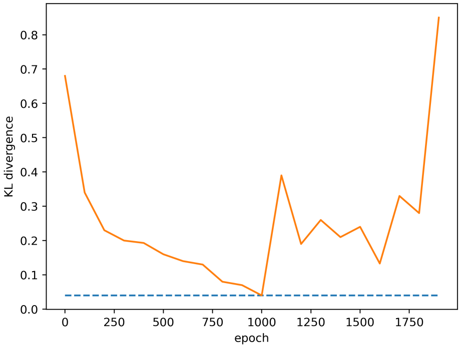


Figure: The KL divergence dynamics.

Main Result II

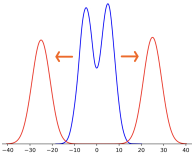


Figure: Illustration of modes shift.

Theorem (Data-Dependent Generalization Gap)

Suppose $p_0(x) = q_1\mathcal{N}(x; -\mu, 1) + q_2\mathcal{N}(x; \mu, 1)$, *where* $q_1, q_2 > 0$ *with* $q_1 + q_2 = 1$. *Under the same conditions, with high probability we have*

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

Modes Shift Effect

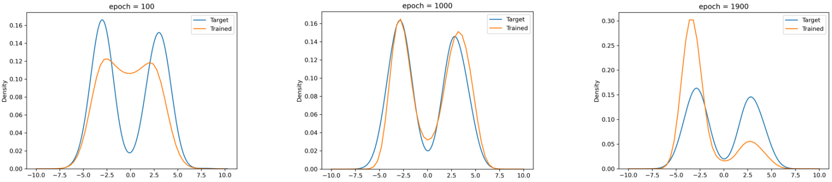


Figure: Training dynamics when the distance between two modes is 6 ($\mu = 3$).

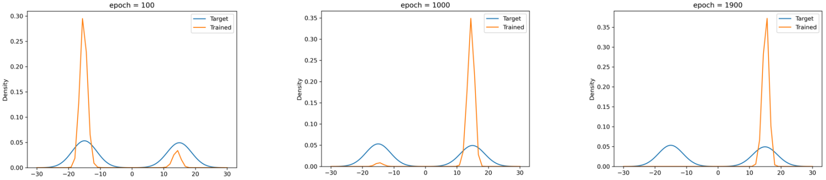


Figure: Training dynamics when the distance between two modes is 30 ($\mu = 15$).

Real-World Experiment

U-net + MNIST:

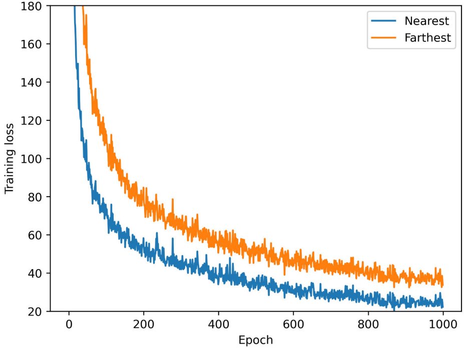


Figure: The training loss dynamics.

Real-World Experiment


Figure: Sampling from the farthest (left) and nearest (right) clusters.

Conclusion

- Data-independent regime: For target distributions with finite supports, diffusion models have polynomially small generalization errors in n (sample size) and m (model capacity) with early-stopping.

Conclusion

- Data-independent regime: For target distributions with finite supports, diffusion models have polynomially small generalization errors in n (sample size) and m (model capacity) with early-stopping.
- Data-dependent regime: For target distributions with increasing modes distances, the generalization performance of diffusion models becomes significantly worse.

Thank you!