Learning Mixtures of Gaussians Using the DDPM Objective

Kulin Shah, Sitan Chen, Adam Klivans

Paper

IFML Institute for Foundations of MACHINE LEARNING

Diffusion Models: A Generative Modelling framework

- Backbone of the powerful models such as DALL-E 2, Imagen, Stable Diffusion and Dream-Fusion
- **Assuming the existence of an oracle for score estimation**, several work establishes convergence guarantees under mild assumptions on the data distribution [CCL+23, LLT23]

Are there any data distributions under which gradient descent provably achieves accurate score estimation?

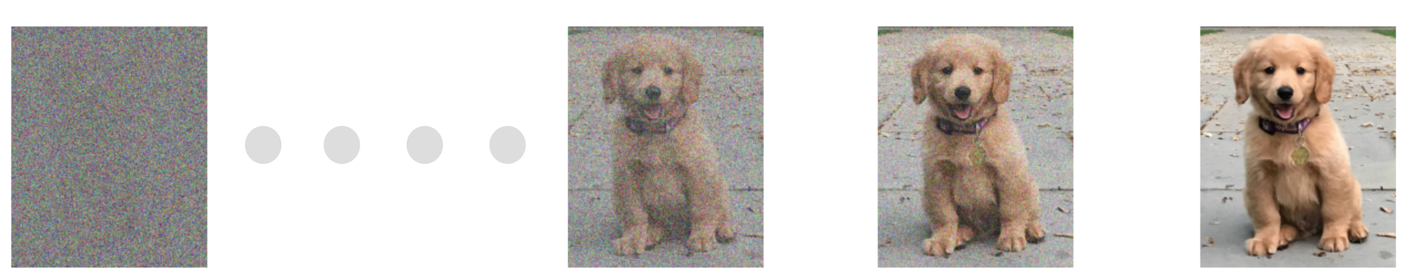
Forward and Backward Process

Forward Corruption Process

$X_0 \sim q_0$ (Data distribution) $X_t \sim q_t$ (Distribution at noise t) $X_T \sim q_T$ ($\approx$ Normal distribution)

Forward SDE: $dX_t = -X_t dt + \sqrt{2} dW_t$ (Ornstein–Uhlenbeck process)

Reverse Generative Process



Reverse SDE: $dX_t^{\leftarrow} = (X_t^{\leftarrow} + 2\nabla \ln q_{T-t}(X_t^{\leftarrow}))\, dt + \sqrt{2} dW_t$

Learning the score using score-matching at various noise scales:

$$s_{T-t} = \arg\min_s \ \mathbb{E}[\|s_{T-t}(X_t^{\leftarrow}) - \nabla \ln q_{T-t}(X_t^{\leftarrow})\|^2]$$

Denoising DDPM Objective [HJA20] and sliced score matching [SGSE19] are different ways to minimize the objective and equivalent to the score matching objective upto a data dependent constant.

Takeaways

- Gradient Descent on the DDPM Objective learns parameters of
  - mixtures of two Gaussians from random initialization
  - mixtures of K Gaussians from warm initialization

  (First efficient learning results for diffusion models.)
- Our proofs use a new connection between score-based methods and two other approaches to distribution learning: the Expectation-Maximization (EM) algorithm and spectral methods.

Mixtures of two Gaussians

Mixtures of two Gaussians:

$$q = 0.5\mathcal{N}(\mu^*, I) + 0.5\mathcal{N}(-\mu^*, I)$$

Informal Result: Gradient descent (GD) on the DDPM objective with random initialization efficiently learns the parameters of an unknown mixture of two spherical Gaussians with 1/poly(dimension)-separated centers.

Proof Idea

Population update $\approx \mathbb{E}[\ \tanh(\mu^\top x)x$ + Other terms]

EM Update

Large noise scale training:

- Population GD on DDPM objective $\approx$ Power method on $I + \mu^* \mu^{*\top}$ matrix (without normalization)
- Convergence of power method in angular distance $\Longrightarrow$ Warm start for the DDPM training

Small noise scale training:

- Population GD on DDPM objective $\approx$ Expectation-Maximization (EM) algorithm update
- Convergence of EM update $\Longrightarrow$ Convergence of DDPM training

Mixtures of K Gaussians

Mixtures of K Gaussians with $\Omega(\sqrt{\log(K)})$-separated centers

$$q = \frac{1}{K}\sum_{i=1}^{K} \mathcal{N}(\mu_i^*, I) \quad \text{where } \mu_i^* = \text{mean of } i^{th} \text{ component}$$

Informal Result: When there is a warm start of the centers, gradient descent on the DDPM objective learns the parameters of the mixture of K Gaussians in polynomial time and sample complexity.

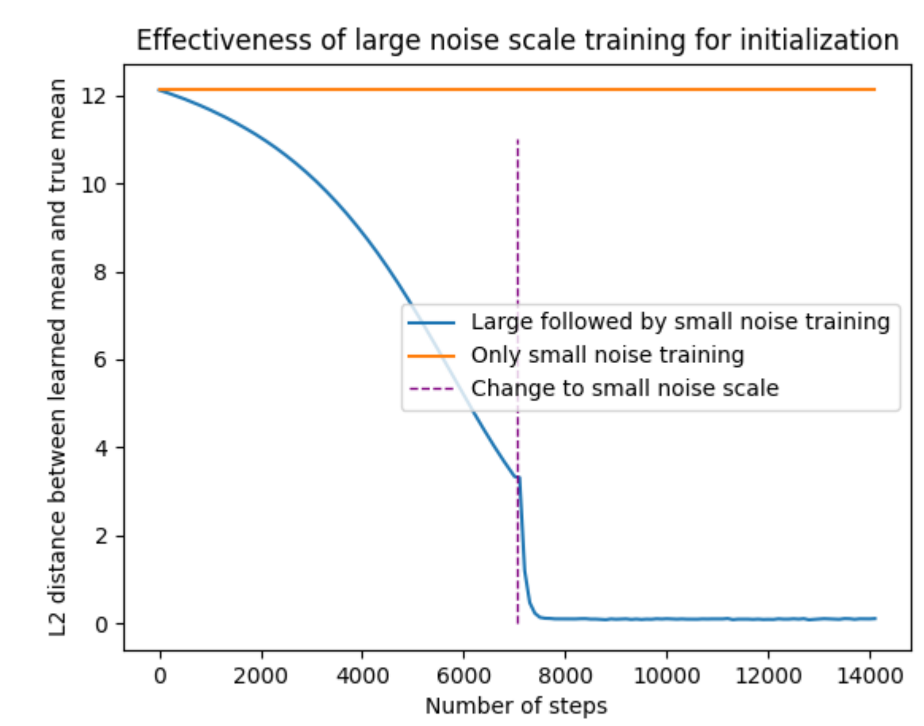
- Connection between Population GD on DDPM objective and EM algorithm extends to mixtures of K Gaussians.
- Known results on local convergence of EM [KC20, SN21] implies the learning.

Experiments

Task: Learning mixtures of two Gaussians

Observation:

- Large noise scale training provide warm start of the training
- Only small noise scale training does not converge to ground-truth parameters



References

[CCL+23] Sitan Chen, Sinho Chewi, Holden Lee, Yuanzhi Li, Jianfeng Lu, and Adil Salim. Sampling is as easy as learning the score: theory for diffusion models with minimal data assumptions. ICLR 2023.

[LLT23] Holden Lee, Jianfeng Lu, and Yixin Tan. Convergence of score-based generative modeling for general data distributions. ALT 2023.

[SGSE19] Yang Song, Sahaj Garg, Jiaxin Shi, and Stefano Ermon. Sliced score matching: A scalable approach to density and score estimation. UAI 2019.

[HJA20] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. NeurIPS 2020.

[KC20] Jeongyeol Kwon and Constantine Caramanis. The em algorithm gives sample optimality for learning mixtures of well-separated gaussians. COLT 2020.

[SN21] Nimrod Segol and Boaz Nadler. Improved convergence guarantees for learning gaussian mixture models by em and gradient em. Electronic journal of statistics. 2021.