Transformers Learn the Optimal DDPM Denoiser for Multi-Token GMMs

Hongkang Li UPenn *Hancheng Min* SJTU *René Vidal* UPenn

ICML International Conference On Machine Learning

Overview

- **Diffusion models** generate samples by learning either
  - a denoiser $f(\Psi; X, t)$: parameter Ψ, input X, diffusion time step t;
  - the score function: $s(X,t) = \nabla_X \log p(X)$.
- **Gap between theory and practice**: existing work studies
  - Landscape: no proof of convergence to optimal denoiser.
  - Training dynamics: guarantees only for simplified architectures.
- **Contributions**: for transformer-based diffusion models trained on multi-token Gaussian mixture data with population loss, we show that gradient descent converges to the Bayes-optimal denoiser.

Data Model and Problem Formulation

Motivations: Image tokens correspond to a small number of patterns.

- **Tokens of the sky**: $\mathcal{N}(\mu_1, \rho^2 I)$.
- **Tokens of the eagle**: $\mathcal{N}(\mu_2, \rho^2 I)$.
- **Tokens of the cloud**: $\mathcal{N}(\mu_3, \rho^2 I)$.

Data Model: Multi-token Gaussian mixture (MTGM)

- $X = (x_1, \dots, x_P) \in \mathbb{R}^{d\times P} \sim \mathcal{D}(\tilde{\pi}, K, \{\mu_i\}_{i=1}^M, \rho)$. Mixture of M components.
- Select $K \le M$ components: $Z \sim Unif(z \in \{0,1\}^M : 0 < ||z||_0 = K)$.
- Normalize proportions: $\pi(Z)$ with $\pi_i(Z) = [Z]_i \tilde{\pi}_i / Z^\top \tilde{\pi}$.
- Sample P tokens: $x_p \mid Z \sim MoG(\pi(Z), \{\mu_i\}_{i=1}^M, \rho^2 I)$.

Forward Process: (DDPM) $X^t = \sqrt{\bar{\alpha}_t} X^0 + \sqrt{1-\bar{\alpha}_t} E^t$, $E^t \sim \mathcal{N}(0, I)$, $t \in [T]$.

Noise schedule — Gaussian noise

Denoising Network: $f(\Psi; X, t) = v_t(X - X\,\mathrm{softmax}(X^\top W X / d)), \Psi = \{W, \{v_t\}_{t=1}^T\}$.

- DDPM population loss $L_D(\Psi) = \sum_{t=1}^T \mathbb{E}_{X^0, E^t}\left[||f(\Psi; X^t, t) - E^t||_F^2/(2dPT)\right]$.
- Score matching $L_S(\Psi) = \sum_{t=1}^T \mathbb{E}_{X^0, E^t}\left[||g(\Psi,\ X^t, t) - s(X^t, t)||_F^2/(2dPT)\right]$.

Gradient updates: $\Psi^{(\tau+1)} = \Psi^{(\tau)} - \eta \nabla_\Psi L(\Psi^{(\tau)})$.

Main Theoretical Results

Notations

- Time-averaged SNR: $SNR = \mathbb{E}_t[\bar{\alpha}_t/(1-\bar{\alpha}_t)]$.
- Minimal pattern ratio: $\nu(K) = \min_{u\in[M]} \mathbb{P}(\text{a token is drawn from pattern } u) \le \frac{1}{M}$.
- Bayes denoising risk: $R_{Bayes} = \mathbb{E}_{X^0, E^t, t}\left[||\mathbb{E}[E^t|X^t] - E^t||_F^2/(2dP)\right]$.

Theorem 1: Convergence of DDPM Training.

For any $\epsilon \in (0, \min_{u\in[M]} \tilde{\pi}_u^{\Theta(1)})$, if the step size $\eta \le O(1)$,

- the token dimension $d \ge \epsilon^{-1} \log(\nu(K)^{-1})$
- the number of iterations $I = \Omega((\epsilon^{-1} + \nu(K)^{-3})\eta^{-1}\nu(K)^{-1} \cdot SNR^{-3})$,
- the number of tokens per data $P \ge \Omega(\nu(K)^{-1}(\rho^2+1)\epsilon^{-1}\log d)$,
- the number of diffusion steps $T \ge \Omega(\log d)$,

then with high probability, the learned model $\Psi^{(I)}$ satisfies $L(\Psi^{(I)}) \le R_{Bayes} + O(\epsilon)$.

Corollary 1: Special Cases of the Required Number of Iterations I.

- Min I: $K = 1$, $I = \Omega((\epsilon^{-1} + M^3)\eta^{-1} M \cdot SNR^{-3})$;
- Max I: $K = M$, $I = \Omega((\epsilon^{-1} + (\min_{u\in[M]} \tilde{\pi}_u)^{-3})\eta^{-1} (\min_{u\in[M]} \tilde{\pi}_u)^{-1} \cdot SNR^{-3})$.

Theorem 2: Score Matching.

Let $\Psi^{(I)}$ be defined as in Theorem 1. For some $\epsilon \in (0, \min_{u\in[M]} \tilde{\pi}_u^{\Theta(1)})$, we can construct the following score network with parameters $\Psi^{(I)}$

$$g(\Psi^{(I)},\ X^t, t) = -f(\Psi^{(I)}; X^t, t)/\sqrt{1-\bar{\alpha}_t},$$

such that $L_S(\Psi^{(I)}) \le L_D(\Psi^{(I)}) \cdot (SNR + 1) \le \epsilon \cdot (SNR + 1)$.

Denoising Mechanism of Transformers

Key Insights

The Transformer trained from DDPM objective achieves near-optimal denoising by learning the *oracle MMSE estimator* to approximate the *true MMSE estimator*.

Transformer

DDPM training

$\mathbb{E}[E^t|X^t]$ ≈ $\mathbb{E}[E^t|X^t, M_X]$

True MMSE estimator (Bayes optimal) — Oracle MMSE estimator

R_{Bayes} − $R_{oracle} \le O(\epsilon)$

Notations

- M_X: the patterns of all tokens of X.
- Oracle denoising risk: $R_{oracle} = \frac{1}{2dP}\mathbb{E}_{X^0, E^t, t}\left[||\mathbb{E}[E^t|X^t, M_X] - E^t||_F^2\right]$.

Proposition 1: Transformer Learns Oracle MMSE Estimator

Trained Model = $v_t^{(I)}(X^t - X^t\,\mathrm{softmax}(X^{t\top} W^{(I)} X^t/d))$

$O(\epsilon)$ $O(\epsilon)$

Oracle MMSE = $\sqrt{1-\bar{\alpha}_t}/(1-\bar{\alpha}_t+\rho^2\bar{\alpha}_t) \cdot (X^t - \sqrt{\bar{\alpha}_t} M_X)$

The trained model satisfies after I iterations, $\forall\, t \in [T]$, $v_t^{(I)}$ and $X^t\,\mathrm{softmax}(X^{t\top} W^{(I)} X^t/d)$ approximate $\sqrt{1-\bar{\alpha}_t}/(1-\bar{\alpha}_t+\rho^2\bar{\alpha}_t)$ and $\sqrt{\bar{\alpha}_t} M_X$, respectively.

How can $X^t\,softmax(\frac{X^{t\top} W^{(I)} X^t}{d})$ approximate $\sqrt{\bar{\alpha}_t} M_X$?

Proposition 2: Mean Denoising Mechanism

Attention weights are concentrated and uniformly distributed among tokens with the same pattern as the query, leading to a minimum-variance unbiased estimator (MVUE) of the mean pattern of each token for approximation.

$\mathcal{N}(\sqrt{\bar{\alpha}_t}\mu_1, (\bar{\alpha}_t\rho^2 + 1 - \bar{\alpha}_t)I)$ $\mathcal{N}(\sqrt{\bar{\alpha}_t}\mu_2, (\bar{\alpha}_t\rho^2 + 1 - \bar{\alpha}_t)I)$

x_1^t	...	$x_{P_1}^t$	$x_{P_1+1}^t$	...	x_{P+1}^t

Attention score at x_1^t

≈ | $1/P_1$ | ... | $1/P_1$ | 0 | ... | 0 |

$\approx (x_1^t + \cdots x_{P_1}^t)/P_1$

$\mathcal{N}\left(\sqrt{\bar{\alpha}_t}\mu_1, \frac{\bar{\alpha}_t\rho^2 + 1 - \bar{\alpha}_t}{P_1} I\right)$ Variance reduced!