上海交通大学
约翰·霍普克罗夫特
计算机科学中心
John Hopcroft Center for Computer Science

NEURAL INFORMATION PROCESSING SYSTEMS

Leveraging Drift to Improve Sample Complexity of Variance Exploding Diffusion Models

Ruofeng Yang[1], Zhijie Wang[1], Bo jiang[1] ,Shuai Li[1,*]

1. Shanghai Jiao Tong University

The Paradigm of Diffusion Models

- Diffusion model: Forward and reverse process.
- The general forward process:

$$dX_t = f(X_t, t)dt + g(t)dB_t, X_0 \sim q_0 \in \mathbb{R}^d$$

- Two common forward processes:

(1) Variance Preserving (VP): $f(X_t, t) = -\frac{1}{2}X_t, g(t) = 1$

(2) Variance Exploding (VE): $f(X_t, t) = 0, g(t) = \sqrt{d\sigma_t^2/dt}$ $(\sigma_t^2 = t \text{ or } t^2)$

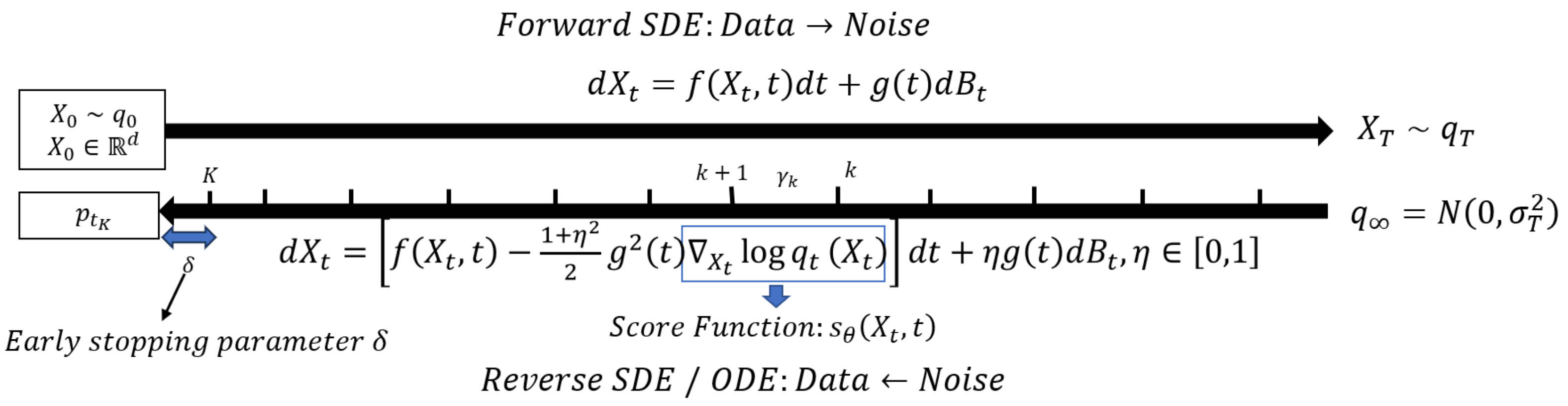
The Reverse Process

- Reverse the forward process → Reverse process

$$X_t = \left[f(X_t, t) - \frac{1+\eta^2}{2} g^2(t) \nabla_x \log q_t (X_t) \right] dt + \eta g(t) dB_t, \eta \in [0,1]$$

- $\eta = 1$ → Reverse SDE (Stochastic sampler)

  $\eta = 0$ → Reverse probability flow ODE (PFODE, deterministic sampler)



The Current Sample Complexity Results

Many works assume an accurate enough score function

$$\|\log q_t(X,t) - s_\theta(X,t)\|_2^2 \leq \epsilon_{score}^2$$

and analyze the sample complexity $K = (T-\delta)/\gamma_K$ to guarantee $Dis\left(p_{t_K}, q_0\right) \leq \epsilon$.

- VP-based models is well-studies and require weakly bounded support assumption

(a)Reverse SDE: $\frac{1}{\epsilon_{W_2}^8 \epsilon_{TV}^2}$ result [1]

(b) Reverse PFODE: $\frac{1}{\epsilon_{W_2}^8 \epsilon_{TV}}$ [2]

- VE-based models lacks of analysis and require strong assumption

(a)Reverse SDE: $\frac{1}{\epsilon_{W_2}^8 \epsilon_{TV}^4}$ result under the log-Sobelev inequality (LSI) [3]

(b) Reverse PFODE: Lack

Motivations

- Variance exploding (VE)-based diffusion model has great performance.
- The sample complexity of VE-based models is larger than variance preserving (VP) based models.

What is the source of large sample complexity of VE-based Models?

A General Convergence Guarantee (Reverse SDE)

Theorem 1. Under the bounded support assumption (weaker than LSI), for VP and VE-based models

$$TV\left(p_{t_K}, q_0\right) \leq \frac{\bar{D}\sqrt{m_T}}{\sigma_T} + \frac{R^2\sqrt{d}}{\sigma_\delta^4}\sqrt{\bar{\gamma}_K \sigma_T^2 T g^2(T)} + \epsilon_{score}\sqrt{g^2(T)T} \overset{?}{\leq} \tilde{O}(\epsilon_{TV})$$

Reverse Beginning Error

Forward Convergence Rate

$TV(N(0,\sigma_T^2), q_T)$

Discretization

Approximated Score

- Balance: (a) T determined by the first term and (b) discretization part depends on T
- VP enjoy an exponential-decay first term $m_T = e^{-T}$ and $\sigma_T = 1 \rightarrow$

  A logarithmic $T = \log(1/\epsilon_{TV})$

- VE has a polynomial-decay one $m_T = 1$ and $\sigma_T^2 = poly(T) \rightarrow$

  Large sample complexity

Core Contribution 1: Drifted VESDE

Intuition: Lacks of the drift term $f(X_t, t) \rightarrow$ Slow forward convergence rate$\rightarrow$

Large sample complexity

Solution Introduce a drift term to VESDE: Drifted VESDE

$$dX_t = -\frac{1}{\tau}\beta_t X_t dt + \sqrt{2\beta_t} dB_t, where\ \tau \in [1, T^2], \beta_t \in [1, t^2]$$

- Drifted VESDE covers class forward processes

  (a) $\tau = 1, \beta_t = 1 \rightarrow$ VP; (b) $\tau = T, \beta_t = 1 \rightarrow$ VE ($\sigma_t^2 = t$); (c) $\tau = T^2, \beta_t = t \rightarrow$VE ($\sigma_t^2 = t^2$)

- Go beyond: With an aggressive β_t (e.g. $\tau = T^2$ and $\beta_t = t^2$),

Drifted VESDE balances different error terms

Drifted VESDE Balances Different Error Term

Corollary 1. For drifted VESDE ($\tau = T^2$) with $\beta_t = t^2$, it enjoys e^{-T} forward convergence guarantee. Assume $\epsilon_{score} \leq \tilde{O}(\epsilon_{TV})$, the sample complexity is

$$K \leq \tilde{O}\left(1/(\epsilon_{W_2}^8 \epsilon_{TV}^2)\right)$$

- This result is the same with VP-based models.
- Due to the logarithmic T, different from the high order requirement $\epsilon_{score} \leq \epsilon_{TV}^2$ of pure VESDE, ϵ_{score} has the same order with ϵ_{TV}.

Contribution 2: The Guarantee for VE with PFODE

- The unified tangent-based framework (Control of high order of score)

$$\|\nabla Y_{0,t_K}\| \leq \exp\left(\frac{R^2}{\delta^2} + \frac{1-\eta^2}{2}\int_0^{t_K}\frac{g^2(u)}{\sigma_T^2}du\right)$$

- For VP forward process, $\int_0^{t_K}\frac{g^2(u)}{\sigma_T^2}du = T \rightarrow \exp(T)$ term
- For VE forward process, $g^2(t) = t$ and $\sigma_T^2 = T^2 \rightarrow$ Constant term

Theorem 2. Under the bounded support and ground-truth score assumption, for VE with PFODE

$$W_1(p_{t_K}, q_0) \leq \frac{\bar{D}\sqrt{m_T}}{\sigma_T} + \exp\left(\frac{1}{\delta^2}\right) Poly(T)\sqrt{\bar{\gamma}_K}$$

Real-world Experiments

(a) Pure VESDE (More Examples)

(b) Drifted VESDE (More Examples)

- Setting: $\tau = T,\ \beta_t = 1$

- Conservative Drifted VESDE Benefits from VESDE without Training:

More detail such as hair and beard details

Conclusion

- (Reverse SDE) Drifted VESDE: balance error terms and improve the results
- (Reverse PFODE) The Exploding property of VE: The first quantitative convergence guarantee without $\exp(T)$
- Furture work
  - Polynomial Sample Complexity for VE with PFODE

Thanks!

Q&A

References

- [1] Chen, S., Chewi, S., Li, J., Li, Y., Salim, A., & Zhang, A. R. (2022). Sampling is as easy as learning the score: theory for diffusion models with minimal data assumptions. arXiv preprint arXiv:2209.11215.
- [2] Chen, S., Chewi, S., Lee, H., Li, Y., Lu, J., & Salim, A. (2024). The probability flow ode is provably fast. Advances in Neural Information Processing Systems, 36.
- [3] Lee, H., Lu, J., & Tan, Y. (2022). Convergence for score-based generative modeling with polynomial complexity. Advances in Neural Information Processing Systems, 35, 22870-22882.