Adaptive Discretization for Consistency Models

Jiayu Bai [1] Zhanbo Feng [2] Zhijie Deng [2] Tianqi Hou [3] Robert C. Qiu [1]
Zenan Ling [†1]

[1]Huazhong University of Science and Technology [2]Shanghai Jiao Tong University
[3]Huawei
[†]Correspondence to: lingzenan@hust.edu.cn

2025/11

Introduction

- **Diffusion Models:** Diffusion Models (DMs) achieve high quality data generation by solving a reverse SDE or a PF-ODE.

Introduction

- **Diffusion Models:** Diffusion Models (DMs) achieve high quality data generation by solving a reverse SDE or a PF-ODE.
- **Consistency Models:** Consistency Models (CMs) enable efficient one-step generation by mapping each point on the PF-ODE trajectories of DMs to their corresponding endpoint.

Introduction

- **Diffusion Models:** Diffusion Models (DMs) achieve high quality data generation by solving a reverse SDE or a PF-ODE.
- **Consistency Models:** Consistency Models (CMs) enable efficient one-step generation by mapping each point on the PF-ODE trajectories of DMs to their corresponding endpoint.
- **Discretization for Consistency Models:** CMs minimize the distance between adjacent points on the trajectories, *i.e.*,

$$\min_{\theta} \mathbb{E}_{\boldsymbol{x}_0, \boldsymbol{z}, t}[w(t) \cdot \|f_\theta(\boldsymbol{x}_t) - f_{\theta^-}(\boldsymbol{x}_{t-\Delta t})\|_2^2],$$

where θ^- stands for $\mathrm{stopgrad}(\theta)$ and Δt is the time step corresponding to t. We refer to the selection of Δt as the **discretization** for CMs.

Introduction

- **Diffusion Models:** Diffusion Models (DMs) achieve high quality data generation by solving a reverse SDE or a PF-ODE.
- **Consistency Models:** Consistency Models (CMs) enable efficient one-step generation by mapping each point on the PF-ODE trajectories of DMs to their corresponding endpoint.
- **Discretization for Consistency Models:** CMs minimize the distance between adjacent points on the trajectories, *i.e.*,

$$\min_{\theta} \mathbb{E}_{\boldsymbol{x}_0,\boldsymbol{z},t}[w(t) \cdot \|f_\theta(\boldsymbol{x}_t) - f_{\theta^-}(\boldsymbol{x}_{t-\Delta t})\|_2^2],$$

  where θ^- stands for $\mathrm{stopgrad}(\theta)$ and Δt is the time step corresponding to t. We refer to the selection of Δt as the **discretization** for CMs.
- **Previous discretization strategies:** Previous works adopt different empirical discretization strategies. However, It remains unclear that:

Introduction

- **Diffusion Models:** Diffusion Models (DMs) achieve high quality data generation by solving a reverse SDE or a PF-ODE.
- **Consistency Models:** Consistency Models (CMs) enable efficient one-step generation by mapping each point on the PF-ODE trajectories of DMs to their corresponding endpoint.
- **Discretization for Consistency Models:** CMs minimize the distance between adjacent points on the trajectories, *i.e.*,

$$\min_{\theta} \mathbb{E}_{\boldsymbol{x}_0,\boldsymbol{z},t}[w(t) \cdot \|f_{\theta}(\boldsymbol{x}_t) - f_{\theta^-}(\boldsymbol{x}_{t-\Delta t})\|_2^2],$$

  where θ^- stands for $\mathrm{stopgrad}(\theta)$ and Δt is the time step corresponding to t. We refer to the selection of Δt as the **discretization** for CMs.
- **Previous discretization strategies:** Previous works adopt different empirical discretization strategies. However, It remains unclear that:
  - how different discretization strategies affect CMs' training, and
  - how to adaptively determine the optimal discretization strategy.

Introduction

Contributions

- We provide a unified framework for the discretization for CMs. Guided by local consistency and global consistency, we formulate a constrained optimization problem for selecting the discretization step.
- We propose Adaptive Discretization for Consistency Models (ADCMs). We first relax the optimization problem using the Lagrangian method. Then we derive an analytical solution using the Gauss-Newton method, enabling adaptive discretization.
- Our experiments demonstrate that ADCMs significantly improve the training efficiency, while achieving competitive performance in one-step generation. Furthermore, ADCMs adapt to advanced variants of DMs such as Flow Matching without manual adjustments.

Local and Global Consistency

Local Consistency Ensures Trainability

We need that the optimization objective of CMs is trainable. To achieve this, we need to choose an appropriate Δt *such that the objective is as small as possible, thereby satisfying local consistency, namely:*

$$\min_{\Delta t} \mathbb{E}_{\boldsymbol{x}_0, \boldsymbol{z}} \left[\| f_{\theta^-}(\boldsymbol{x}_t) - f_{\theta^-}(\boldsymbol{x}_{t-\Delta t}) \|_2^2 \right].$$

Local and Global Consistency

Local Consistency Ensures Trainability

We need that the optimization objective of CMs is trainable. To achieve this, we need to choose an appropriate Δt such that the objective is as small as possible, thereby satisfying local consistency, namely:

$$\min_{\Delta t} \mathbb{E}_{\boldsymbol{x}_0, \boldsymbol{z}} \left[\| f_{\theta^-}(\boldsymbol{x}_t) - f_{\theta^-}(\boldsymbol{x}_{t-\Delta t}) \|_2^2 \right].$$

Global Consistency Eusures Stability

The global denoising error in the training target $f_{\theta^-}(\boldsymbol{x}_{t-\Delta t})$ leads to instability. To mitigate this, we need to choose an appropriate Δt such that the error is constrained, thereby satisfying global consistency, namely:

$$\text{Find} \quad \Delta t, \quad \text{s.t.} \quad \mathbb{E}_{\boldsymbol{x}_0, \boldsymbol{z}} \left[\| f_{\theta^-}(\boldsymbol{x}_{t-\Delta t}) - \boldsymbol{x}_0 \|_2^2 \right] \leq \delta$$

A Unified Framework

Observation

How does the choice of Δt affect local consistency and global consistency?

A Unified Framework

Observation

How does the choice of Δt affect local consistency and global consistency?

(1) When $\Delta t \to 0$, the local consistency is ensured. However, the training becomes highly unstable due to the loss of global consistency.

A Unified Framework

Observation

How does the choice of Δt affect local consistency and global consistency?

(1) When $\Delta t \to 0$, the local consistency is ensured. However, the training becomes highly unstable due to the loss of global consistency.

(2) When $\Delta t = t$, the global consistency is ensured. However, the performance degrades due to the loss of local consistency.

A Unified Framework

Observation

How does the choice of Δt affect local consistency and global consistency?

(1) When $\Delta t \to 0$, the local consistency is ensured. However, the training becomes highly unstable due to the loss of global consistency.

(2) When $\Delta t = t$, the global consistency is ensured. However, the performance degrades due to the loss of local consistency.

Key idea

Given time t, find the optimal discretization step Δt *s.t.* the local consistency and global consistency are simultaneously satisfied, *i.e.*, solve the constrained optimization problem

$$\min_{\Delta t} \quad \mathbb{E}_{\boldsymbol{x}_0,\boldsymbol{z}}\left[\|f_{\theta^-}(\boldsymbol{x}_t) - f_{\theta^-}(\boldsymbol{x}_{t-\Delta t})\|_2^2\right], \quad \text{s.t.}\, \mathbb{E}_{\boldsymbol{x}_0,\boldsymbol{z}}\left[\|f_{\theta^-}(\boldsymbol{x}_{t-\Delta t}) - \boldsymbol{x}_0\|_2^2\right] \le \delta.$$

A Unified Framework

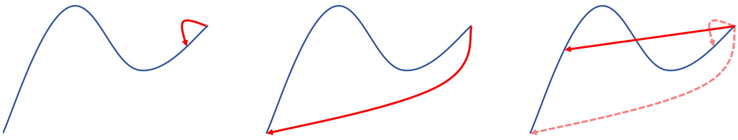


Figure: Discretization strategies of different models.

Remark

- *CMs consider only local consistency during discretization.*
- *DMs consider only global consistency during discretization.*
- *ADCMs balance local and global consistency and adaptively adjust the discretization step size based on the information from both.*

NEURAL INFORMATION PROCESSING SYSTEMS

Adaptive Discretization

Adaptive discretization through Gauss-Newton.

We first apply the Lagrange multiplier method to relax the optimization problem, yielding:

$$\Delta t^* = \arg\min_{\Delta t} \mathbb{E}_{\boldsymbol{x}_0, \boldsymbol{z}} \left[\| f_{\theta^-}(\boldsymbol{x}_t) - f_{\theta^-}(\boldsymbol{x}_{t-\Delta t}) \|_2^2 + \lambda \| f_{\theta^-}(\boldsymbol{x}_{t-\Delta t}) - \boldsymbol{x}_0 \|_2^2 \right].$$

We then apply the Gauss-Newton method to approximate $f_{\theta^-}(\boldsymbol{x}_{t-\Delta t})$ using its first-order Taylor expansion, which is:

$$f_{\theta^-}(\boldsymbol{x}_{t-\Delta t}) \approx f_{\theta^-}(\boldsymbol{x}_t) - \boldsymbol{v}\Delta t, \quad \boldsymbol{v} = \nabla_{\boldsymbol{x}_t} f_{\theta^-} \cdot \frac{\mathrm{d}\boldsymbol{x}_t}{\mathrm{d}t} + \partial_t f_{\theta^-}.$$

Then the optimal discretization step is given by:

$$\Delta t^* = \frac{\lambda}{1+\lambda} \frac{\mathbb{E}_{\boldsymbol{x}_0, \boldsymbol{z}}[\boldsymbol{v}^\top (f_{\theta^-}(\boldsymbol{x}_t) - \boldsymbol{x}_0)]}{\mathbb{E}_{\boldsymbol{x}_0, \boldsymbol{z}}[\boldsymbol{v}^\top \boldsymbol{v}]}.$$

Putting ADCMs into Practice

Technique 1:

We perform simulation-based optimization and record the optimization process as $\mathbb{T} = \{t_1^*, \ldots, t_N^*\}$, from which t and $t - \Delta t$ are sampled.

Technique 2:

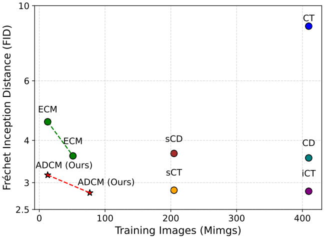
We alternately optimize the time segmentation $\mathbb{T}$ and the NN's parameters θ during training.

Technique 3:

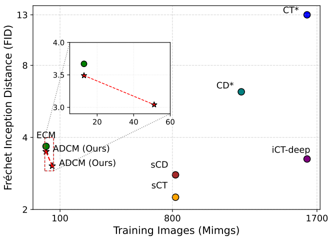
We propose an adaptive weighting function and use the Pseudo-Huber metric to further enhance performance.

NEURAL INFORMATION PROCESSING SYSTEMS

Experiments



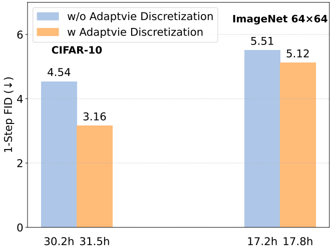
CIFAR-10



ImageNet 64×64

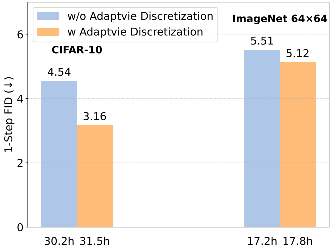
ADCMs significantly improve the **training efficiency** on both unconditional CIFAR-10 and class-conditional ImageNet 64×64.

Experiments

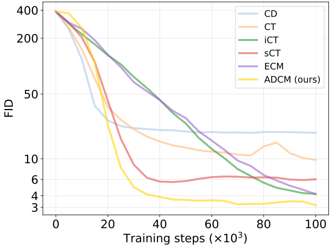


ADCMs introduce only about 4% **additional time cost** while improving the final performance.

Experiments

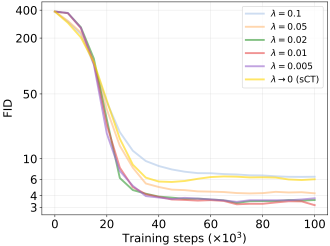


ADCMs introduce only about 4% **additional time cost** while improving the final performance.



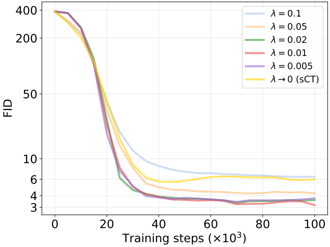
Compared to other CMs' approaches, ADCMs have a **faster convergence rate** and better final performance.

Experiments

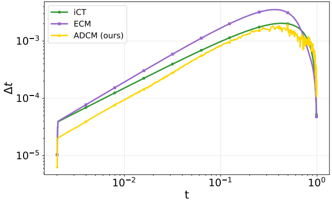


A large λ improves stability but hurts final performance, while a too-small λ reduces stability and hinders convergence.

Experiments



A large λ improves stability but hurts final performance, while a too-small λ reduces stability and hinders convergence.



Adaptive discretization on Flow Matching. ADCMs are able to **adaptively learn discretization strategies** that are similar in trend to empirical ones without manual adjustments.

Reference and Acknowledgement

Reference:

1. Song, Y., Dhariwal, P., Chen, M. & Sutskever, I. *Consistency Models.* in *ICML* (2023).
2. Song, Y. & Dhariwal, P. *Improved Techniques for Training Consistency Models.* in *ICLR* (2024).
3. Geng, Z., Pokle, A., Luo, W., Lin, J. & Kolter, J. Z. *Consistency Models Made Easy.* in *ICLR* (2025).
4. Lu, C. & Song, Y. *Simplifying, Stabilizing and Scaling Continuous-time Consistency Models.* in *ICLR* (2025).

Thank you!