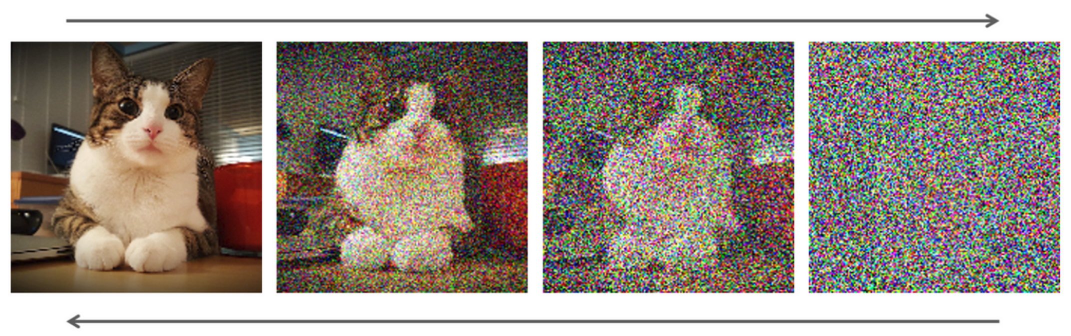
Faster Diffusion Sampling with Randomized Midpoints: Sequential and Parallel

Shivam Gupta [1] Linda Cai [2] Sitan Chen [3]

[1]UT Austin [2]UC Berkeley [3]Harvard SEAS

Sampling Process of Diffusion Models

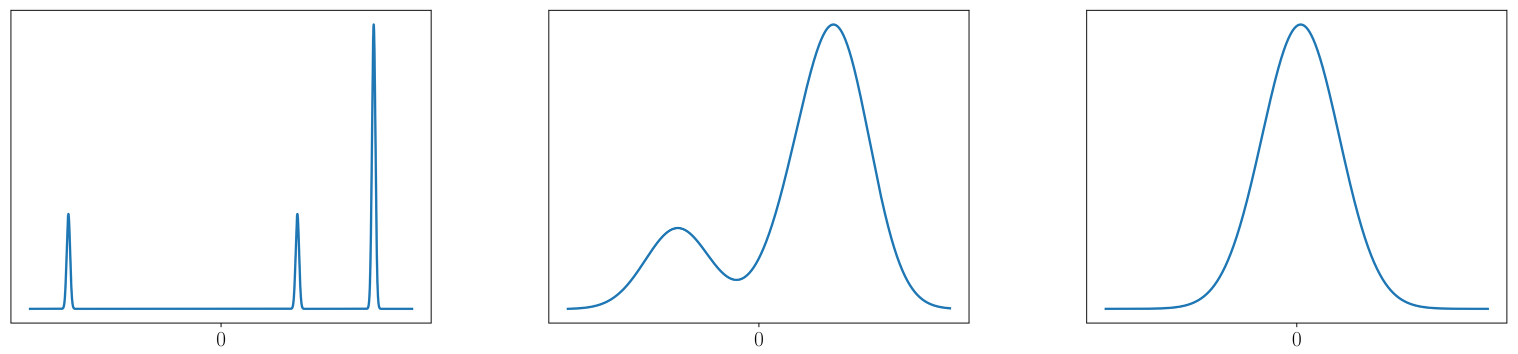
- Diffusion models learn a distribution q_0 by adding noise to training samples and learning to denoise.



- Formally, the image distribution q_0 is converted to the standard Gaussian via the *forward SDE*:

$$\mathrm{d}x_t^{\rightarrow} = -x_t^{\rightarrow}\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t \qquad x_0^{\rightarrow} \sim q_0$$

- $x_0 \sim q_0$ evolves into $x_t \sim e^{-t}x_0 + \mathcal{N}(0, \sigma_t^2 I_d)$ at time t, where $\sigma_t^2 = 1 - e^{-2t}$. As t grows, distribution converges to $\mathcal{N}(0, I_d)$.



t=0 t = 0.5 t=1

- There is a corresponding *probability flow ODE*, that can be used to *sample* from q_0 given a sample $x_0 \sim q_T \approx \mathcal{N}(0, I_d)$:

$$\mathrm{d}x_t = (x_t + \nabla \ln q_{T-t}(x_t))\,\mathrm{d}t\,.$$

- Need to learn the **score** function $\nabla \log q_t$.
- Given score approximations s_t's, diffusion models can provably sample from q_0, provided enough discretization steps are used.

Question

How many steps are necessary for accurate diffusion sampling?

- [CCL+23a] showed that a standard discretization of the ODE interleaved with "corrector" steps can sample in $\widetilde{O}(\sqrt{d})$ steps with small **Total Variation (TV)** error.
- For *log-concave sampling* with small **2-Wasserstein** error, [SL19] showed that a randomized midpoint discretization can sample in $\widetilde{O}(d^{1/3})$ steps.

Our Assumptions

- Have access to score approximations s_t's with

$$\mathop{\mathbb{E}}_{x \sim q_t}\left[\|s_t(x_t) - \nabla q_t(x_t)\|^2\right] < \varepsilon_{\mathrm{sc}}^2$$

- True scores $\nabla q_t(\cdot)$ and score approximations $s_t(\cdot)$ are L-Lipschitz

Iteration Complexity of Diffusion Sampling

Our Results

- We show that a randomized midpoint discretization of the ODE interleaved with corrector steps can sample from diffusion models in $\widetilde{O}(d^{5/12})$ steps with small TV error.
- Our algorithm can be parallelized to sample in only $\widetilde{O}(\log^2 d)$ parallel rounds – first provable guarantees for parallel diffusion sampling
- We also obtain an improvement for log-concave sampling in **Total Variation** – show that $\widetilde{O}(d^{5/12})$ steps suffice compared to previous $\widetilde{O}(\sqrt{d})$

Work	Number of Iterations	Notes
[CCL+23b]	$\widetilde{O}(d)$	Standard SDE Discretization
[CCL+23a]	$\widetilde{O}(\sqrt{d})$	Discretized ODE + Corrector steps
Ours	$\widetilde{O}(d^{5/12})$	Randomized Midpoint Discretization of ODE

Sequential Sampling

- The standard discretization of the ODE is given by:

$$\mathrm{d}x_t = \left(x_0 + s_{T-\lfloor t/h \rfloor}(x_{\lfloor t/h \rfloor h})\right)\mathrm{d}t \quad \text{for } x_0 \sim \mathcal{N}(0, I_d)$$

$s_T(x_0)$ $s_{T-h}(x_h)$ $s_{T-2h}(x_{2h})$

- **Idea:** Use the score at some point in the middle of the interval instead

$s_{T-\alpha_1 h}(x_{\alpha_1 h})$ $s_{T-\alpha_2 h}(x_{h(1+\alpha_2)})$ $s_{T-h(2+\alpha_3)}(x_{h(2+\alpha_3)})$

- The integral formulation of the reverse ODE is given by

$$x_{t_0+h} = e^h x_{t_0} + \int_{t_0}^{t_0+h} e^{t_0+h-t} s_{T-t}(x_t)dt$$

- The standard discretization (DDIM) gives the approximation

$$x_{t_0+h} \approx e^h x_{t_0} + (e^h - 1)s_{T-t_0}(x_{t_0})$$

Issue: Inherently biased

- Instead, use an unbiased estimate:

$$\int_{t_0}^{t_0+h} e^{t_0+h-t} s_{T-t}(x_t)dt \approx he^{(1-\alpha)h} s_{T-(t_0+\alpha h)}(x_{t_0+\alpha h})$$

where $\alpha \sim [0, 1]$. How to obtain $x_{t_0+\alpha h}$?

- First approximate $x_{t_0+\alpha h}$:

$$x_{t_0+\alpha h} \approx e^{\alpha h} x_{t_0} + (e^{\alpha h} - 1)s_{T-t_0}(x_{t_0})$$

- Then use the above to approximate x_{t_0+h}:

$$x_{t_0+h} \approx e^h x_{t_0} + he^{(1-\alpha)h} s_{T-(t_0+\alpha h)}(x_{t_0+\alpha h})$$

Parallel Sampling

- Instead of choosing a single randomized midpoint α, we break up our window $[t_0, t_0 + h]$ into R sub-windows and select randomized midpoints $\alpha_1, \ldots, \alpha_R$ for these sub-windows.
- For $\delta = h/R$ we note that:

$$x_{t_0+\alpha_i h} \approx e^{\alpha_i h} x_{t_0} + \sum_{j=1}^{i}\left(e^{\alpha_i h-(j-1)\delta} - \max(e^{\alpha_i h - j\delta}, 1)\right) \cdot s_{T-(t_0+\alpha_j h)}(x_{t_0+\alpha_j h})\,.$$

Tends to equality as $R \to \infty$.

- To approximate $x_{t_0+\alpha_j h}$ we maintain a sequence of estimates:

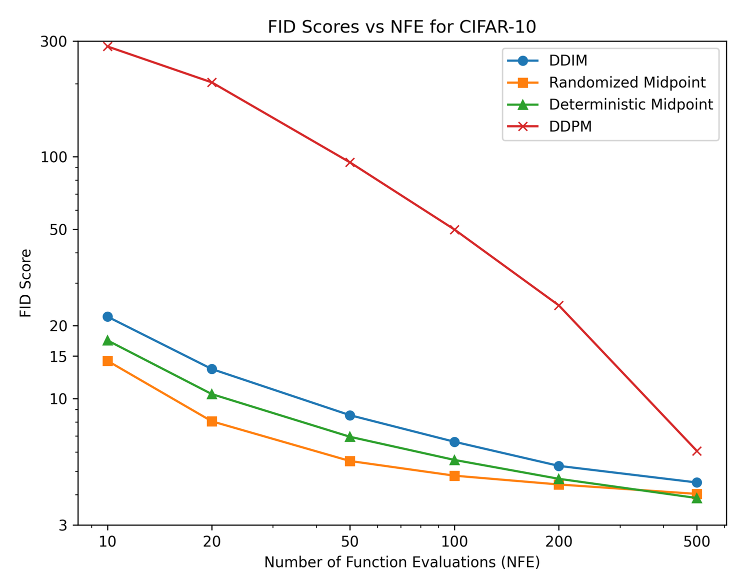
$$\widehat{x}_{t_0+\alpha_i h}^{(k)} \leftarrow e^{\alpha_i h}\widehat{x}_{t_0}^{(k-1)} + \sum_{j=1}^{i}\left(e^{\alpha_i h-(j-1)\delta} - \max(e^{\alpha_i h - j\delta}, 1)\right) \cdot s_{T-(t_0+\alpha_j h)}(\widehat{x}_{t_0+\alpha_j h}^{(k-1)})\,,$$

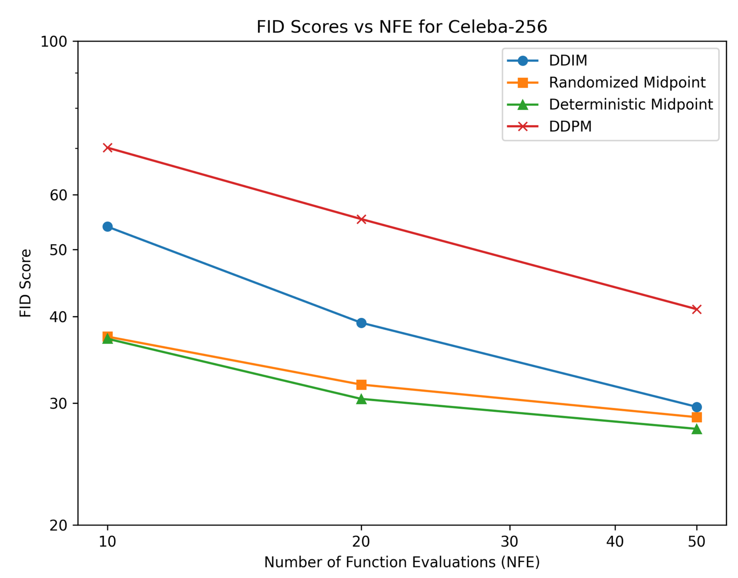
for $k = 1$ to some large enough K ($\widetilde{O}(\log^2 d)$ in our case). Estimates can be computed in parallel.

- Finally we approximate x_{t_0+h}:

$$x_{t_0+h} \approx e^h \widehat{x}_{t_0}^{(K)} + \delta \sum_{i=1}^{R} e^{(1-\alpha_i)h} s_{T-(t_0+\alpha_i h)}(\widehat{x}_{t_0+\alpha_i h}^{(K)})\,.$$

Experiments





References

[CCL+23a] Sitan Chen, Sinho Chewi, Holden Lee, Yuanzhi Li, Jianfeng Lu, and Adil Salim, *The probability flow ODE is provably fast*, Advances in Neural Information Processing Systems, 2023, pp. 68552–68575.

[CCL+23b] Sitan Chen, Sinho Chewi, Jerry Li, Yuanzhi Li, Adil Salim, and Anru Zhang, *Sampling is as easy as learning the score: theory for diffusion models with minimal data assumptions*, International Conference on Learning Representations, 2023.

[SL19] Ruoqi Shen and Yin Tat Lee, *The randomized midpoint method for log-concave sampling*, Advances in Neural Information Processing Systems, vol. 32, 2019.