The Probability Flow ODE is Provably Fast

Sitan Chen (Harvard) Sinho Chewi (Yale) **Holden Lee** (Johns Hopkins)
Yuanzhi Li (CMU) Jianfeng Lu (Duke) Adil Salim (Microsoft)

NeurIPS 2023

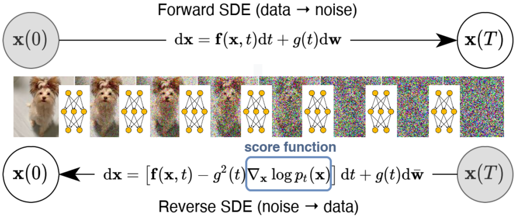
Diffusion models

Problem (Generative Modeling)

Learn a probability distribution from samples, and generate additional samples.

Diffusion models are a modern paradigm for generative modeling with state-of-the-art performance on image, audio, video generation, with applications to inverse problems, molecular modeling, etc.

Picture from Y. Song, Sohl-Dickstein, Kingma, et al. 2020.



What **theoretical guarantees** can we obtain for diffusion models? Show convergence

- given L^2-accurate score estimate,
- for general data distributions.

Expensive to evaluate; care about dependence on dimension d.

SDE vs. ODE formulation

<table>
<tr><th>Denoising Diffusion Probabilistic Modeling (SDE)</th><th>Probability Flow (ODE)</th></tr>
<tr><td>

$$dx_t^{\rightarrow} = -x_t^{\rightarrow}\,dt + \sqrt{2}\,dW_t$$
$$dx_t^{\leftarrow} = x_t^{\leftarrow}\,dt + 2\underbrace{\nabla \log p_{T-t}(x_t^{\leftarrow})}_{\approx s_{T-t}(x_t^{\leftarrow})}\,dt + \sqrt{2}\,dW_t.$$

- Convergence guarantees with $O(d)$ steps.
  S. Chen, Chewi, Li, et al. 2023; H. Chen, Lee, and Lu 2023; Benton, De Bortoli, Doucet, et al. 2023
- Lower bound $\Omega(d)$ for trajectory-wise analysis, even for critically damped Langevin diffusion (S. Chen, Chewi, Li, et al. 2023).

</td><td>

$$dx_t^{\rightarrow} = -x_t^{\rightarrow}\,dt - \nabla \log p_t(x_t^{\leftarrow})\,dt$$
$$dx_t^{\leftarrow} = x_t^{\leftarrow}\,dt + \underbrace{\nabla \log p_{T-t}(x_t^{\leftarrow})}_{\approx s_{T-t}(x_t^{\leftarrow})}\,dt.$$

- Much faster (10x–50x) in practice (J. Song, Meng, and Ermon 2020)...
- ...but can sometimes be less stable.
- **This work:** $O(\sqrt{d})$ steps using corrector steps.

</td></tr>
</table>

The trouble with SDE's

DDPM:

$$dx_t^{\leftarrow} = [x_t^{\leftarrow} + 2\nabla \log p_{T-t}(x_t^{\leftarrow})]\, dt + \sqrt{2}\, dw_t$$
$$x_{t+h}^{\leftarrow} \approx x_t^{\leftarrow} + h\,[x_t^{\leftarrow} + 2\nabla \log p_{T-t}(x_t^{\leftarrow})] + \sqrt{2h}\,\xi,\ \xi \sim N(0, I_d).$$

Discretization error from...

- Drift term (order 1): $O(Lh\sqrt{d}) \to$ can take $h = O\left(\frac{1}{L\sqrt{d}}\right)$.
- Diffusion term (order 1/2): $O(L\sqrt{hd}) \to$ need to take $h = O\left(\frac{1}{L^2 d}\right)$.
  Trajectories of Brownian motion are not smooth!

Probability flow ODE:

$$dx_t^{\leftarrow} = [x_t^{\leftarrow} + \nabla \log p_{T-t}(x_t^{\leftarrow})]\, dt.$$

Assumptions

Assumption

1. p_0 has second moment $\mathbb{E}_{p_0} \|x\|^2 = \mathfrak{m}_2^2$.
2. For each t_k, the score estimate s_{t_k} has error

$$\|\nabla \log p_{t_k} - s_{t_k}\|^2_{L^2(p_{t_k})} \leq \varepsilon_{\mathrm{sc}}^2.$$

3. $\nabla \log p_t$ is L-Lipschitz for every t.
4. The score estimate s_{t_k} is L-Lipschitz for every t_k.

DPUM (Diffusion Predictor + Underdamped Modeling)

Theorem (DPUM, S. Chen, Chewi, Lee, et al. 2023)

Suppose that the Assumptions hold. If the score error satisfies $\varepsilon_{\mathrm{sc}} \le \widetilde{O}(\frac{\varepsilon}{\sqrt{L}})$, *then the output of DPUM gives TV error* ε *with number of steps* $N = \widetilde{\Theta}\left(\frac{L^2 d^{1/2}}{\varepsilon}\right)$.

Algorithm (simplified)

Draw $\widehat{x}_0 \sim N(0, I_d)$. For $n = 0, \ldots, LT-1$:

- **Predictor:** Starting from $\widehat{x}_{n/L}$, run the discretized probability flow ODE from time $\frac{n}{L}$ to $\frac{n+1}{L}$ with step size h_{pred} to obtain $\widehat{x}'_{\frac{n+1}{L}}$.

$$x^{\leftarrow}_{t+h} = e^h x^{\leftarrow}_t + (e^h - 1) s_{T-t}(x^{\leftarrow}_t).$$

- **Corrector:** Starting from $\widehat{x}'_{\frac{n+1}{L}}$, run underdamped LMC for time $\frac{1}{\sqrt{L}}$ with step size h_{corr} to obtain $\widehat{x}_{\frac{n+1}{L}}$.

Challenges

Problem: Cannot use Girsanov's Theorem with ODE's.
Solution: Use **Wasserstein analysis** with coupling.

- **Score perturbation lemma**: Bound the time derivative of score.

$$\mathbb{E}[\|\partial_t \nabla \log q_t^{\rightarrow}(y_t)\|^2] \lesssim L^2 d \left(L + \frac{1}{t}\right).$$

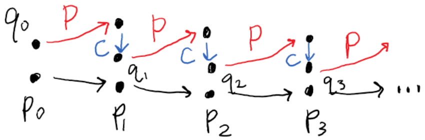
- By Grönwall, get error bounds within $\frac{1}{L}$ time.

Challenges

Problem: Cannot use Girsanov's Theorem with ODE's.
Solution: Use **Wasserstein analysis** with coupling.

Problem: Distance grows exponentially with rate L; can only run for time $O(1/L)$.
Solution: Convert Wasserstein to TV error with a **corrector** step (short-time regularization).
Using data processing inequality for TV distance, we can restart coupling.

- **Predictor (P):** Simulate the reverse SDE/ODE to track a *time-varying* distribution.
- **Corrector (C):** Run MCMC (e.g., Langevin Monte Carlo) to converge towards a *stationary* distribution.
- **Predictor-corrector (PC):** Intersperse P & C steps.



Challenges

Problem: Cannot use Girsanov's Theorem with ODE's.
Solution: Use **Wasserstein analysis** with coupling.

Problem: Distance grows exponentially with rate L; can only run for time $O(1/L)$.
Solution: Convert Wasserstein to TV error with a **corrector** step (short-time regularization).
Using data processing inequality for TV distance, we can restart coupling.

Problem: Overdamped Langevin needs $O(d)$ steps.
Solution: Use **underdamped Langevin** (Langevin "with acceleration"), which needs $O(\sqrt{d})$ steps.

$$\begin{aligned} dx_t &= v_t \, dt \\ dv_t &= -\nabla f(x_t) \, dt - \gamma v_t \, dt + \sqrt{2\gamma} \, dB_t \end{aligned}$$

Conclusion

- Using an ODE instead of SDE, in conjunction with underdamped corrector, reduces dimension dependence from $O(d)$ to $O(\sqrt{d})$.
- Questions:
  - Can we relax smoothness assumptions?
  - Is the corrector necessary?
  - Is the higher error necessary?
  - Other ways to improve parameter dependence and stability?

Bibliography I

- Benton, Joe et al. (2023). "Linear convergence bounds for diffusion models via stochastic localization". In: *arXiv preprint arXiv:2308.03686*.
- Chen, Hongrui, Holden Lee, and Jianfeng Lu (2023). "Improved Analysis of Score-based Generative Modeling: User-Friendly Bounds under Minimal Smoothness Assumptions". In: *arXiv preprint arXiv:2211.01916*.
- Chen, Sitan et al. (2023). "Sampling is as easy as learning the score: theory for diffusion models with minimal data assumptions". In: *arXiv preprint arXiv:2209.11215*.
- Chen, Sitan et al. (2023). *The probability flow ODE is provably fast*. arXiv: 2305.11798 [cs.LG].
- Song, Jiaming, Chenlin Meng, and Stefano Ermon (2020). "Denoising diffusion implicit models". In: *arXiv preprint arXiv:2010.02502*.
- Song, Yang et al. (2020). "Score-Based Generative Modeling through Stochastic Differential Equations". In: *International Conference on Learning Representations*.