Exploring High Dimensions in Dynamical Sampling

Flattening the Scaling Curve!

Yifan Chen

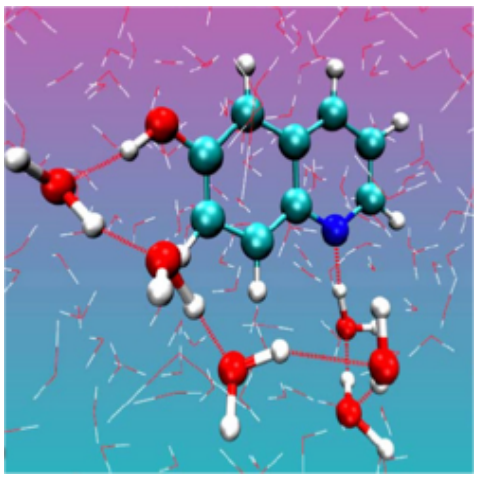
UCLA Mathematics

2025

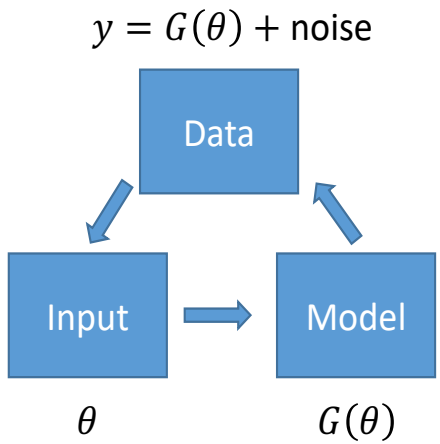
Context

Sampling from probability distributions is a classical and fundamental challenge in scientific computing, statistics, and data science

It has become even more popularized through its key role in generative AI and machine learning



molecular dynamics



Bayes inverse problem

DALL·E 3

Physical models and observed data often exhibit complex structures with natural probabilistic interpretations

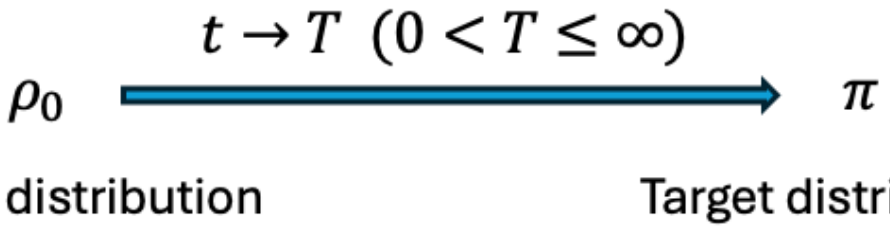
These probability distributions are very high dimensional

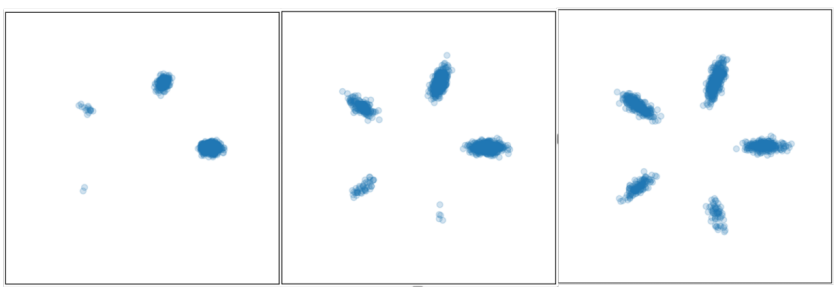
Problem setting

Goal: draw new samples from $\pi \propto \exp(-V)$ either through

- queries to the potential V
- given some sampled data $\{x_i\}_{i=1}^N \sim \pi$

Methodology: typically addressed by building dynamics of measures





MCMC for 2D mixtures

diffusion for images

Problem setting

Goal: draw new samples from $\pi \propto \exp(-V)$ either through

- queries to the potential V
- given some sampled data $\{x_i\}_{i=1}^N \sim \pi$

Methodology: typically addressed by building dynamics of measures

Guiding questions:

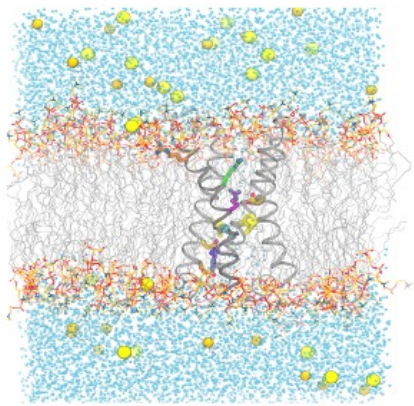
- Rationale for their success in high dimensions
- Design principles for targeted scientific applications

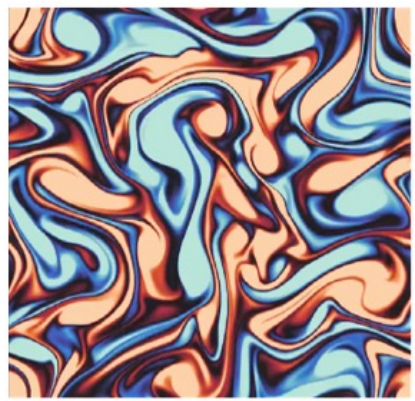
Outline of the talk

1 Analysis of unadjusted Langevin in high dimensions

(analysis w/ methodological insights)

- A new “delocalization of bias” phenomenon for understanding
- Dimension-independent behavior for low-dimensional marginals





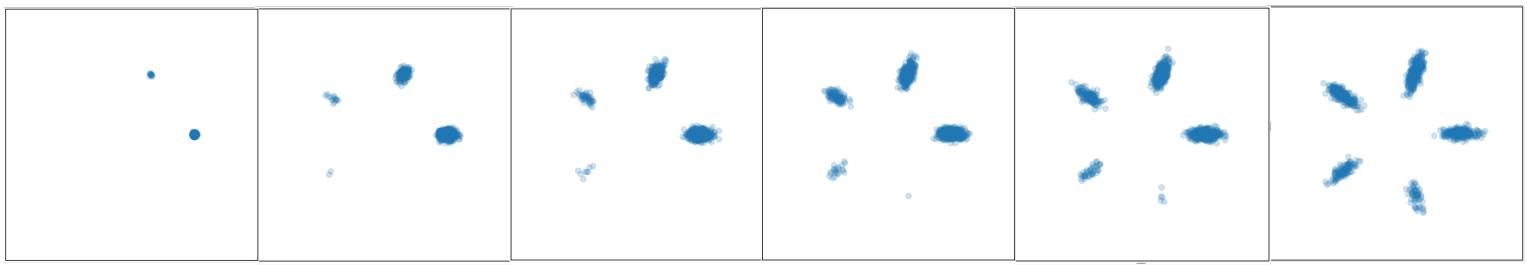
2 Design and application of generative probability flows

(methodology w/ analytical insights)

- An efficient "optimal Lipschitz energy" criteria for design
- Dimension-robust performance with respect to resolution

Goal: draw new samples from π, given queries to V

Markov Chain Monte Carlo (MCMC) provides one of the most widely used dynamics for sampling $\pi \propto \exp(-V)$



One illustration for a 2D Gaussian mixture π (multiple initializations)

A particular class is based on **(overdamped) Langevin's dynamics**

$$\mathrm{d}X_t = -\nabla V(X_t)\mathrm{d}t + \sqrt{2}\mathrm{d}W_t$$

Under mild assumptions, as $t \to \infty$, $\mathrm{Law}(X_t) \to \pi \propto \exp(-V)$

- In molecular dynamics: V is the inter-atomic potential
- In Bayes inverse problem: π is posterior distribution

MCMC algorithm with Langevin's dynamics

Overdamped Langevin's dynamics

$$\mathrm{d}X_t = -\nabla V(X_t)\mathrm{d}t + \sqrt{2}\mathrm{d}W_t$$

Under mild assumptions, as $t \to \infty$, $\mathrm{Law}(X_t) \to \pi \propto \exp(-V)$

- **Unadjusted Langevin:** Euler–Maruyama scheme

$$X_{(k+1)h} = X_{kh} - h\nabla V(X_{kh}) + \sqrt{2}(W_{(k+1)h} - W_{kh})$$

As $k \to \infty$, $\mathrm{Law}(X_{kh}) \to \pi_h$ where hopefully $\pi_h \approx \pi$ (bias)

MCMC algorithm with Langevin's dynamics

Overdamped Langevin's dynamics

$$\mathrm{d}X_t = -\nabla V(X_t)\mathrm{d}t + \sqrt{2}\mathrm{d}W_t$$

Under mild assumptions, as $t \to \infty$, Law$(X_t) \to \pi \propto \exp(-V)$

- **Unadjusted Langevin:** Euler–Maruyama scheme

$$X_{(k+1)h} = X_{kh} - h\nabla V(X_{kh}) + \sqrt{2}(W_{(k+1)h} - W_{kh})$$

As $k \to \infty$, Law$(X_{kh}) \to \pi_h$ where hopefully $\pi_h \approx \pi$ (bias)

- How large is the bias? For $V \in C^2$ with $\alpha I \preceq \nabla^2 V \preceq \beta I$:

$$W_2(\pi, \pi_h) = O(\frac{\beta}{\alpha}\sqrt{dh}) \qquad \text{[Durmus, Moulines, 2019], etc.}$$

- Implication: $h \sim 1/d$ for bounded bias in any dimension

Can be improved to $h \sim 1/d^{1/2}$ with more assumptions [Li, Zha, Tao 2022]

Bias can be completely eliminated

Metropolis-adjusted Langevin: accept $X_{(k+1)h}$ w/ probability

$$p_{\text{accept}} = \min\left\{1, \frac{\pi(X_{(k+1)h})q(X_{kh}|X_{(k+1)h})}{\pi(X_{kh})q(X_{(k+1)h}|X_{kh})}\right\}$$

where q is the transition kernel of unadjusted Langevin; otherwise reject and $X_{(k+1)h} = X_{kh}$. There will be no bias

[Rossky, Doll, Friedman 1978], [Roberts, Tweedie 1997]

Bias can be completely eliminated

Metropolis-adjusted Langevin: accept $X_{(k+1)h}$ w/ probability

$$p_{\text{accept}} = \min\left\{1, \frac{\pi(X_{(k+1)h})q(X_{kh}|X_{(k+1)h})}{\pi(X_{kh})q(X_{(k+1)h}|X_{kh})}\right\}$$

where q is the transition kernel of unadjusted Langevin; otherwise reject and $X_{(k+1)h} = X_{kh}$. There will be no bias

[Rossky, Doll, Friedman 1978], [Roberts, Tweedie 1997]

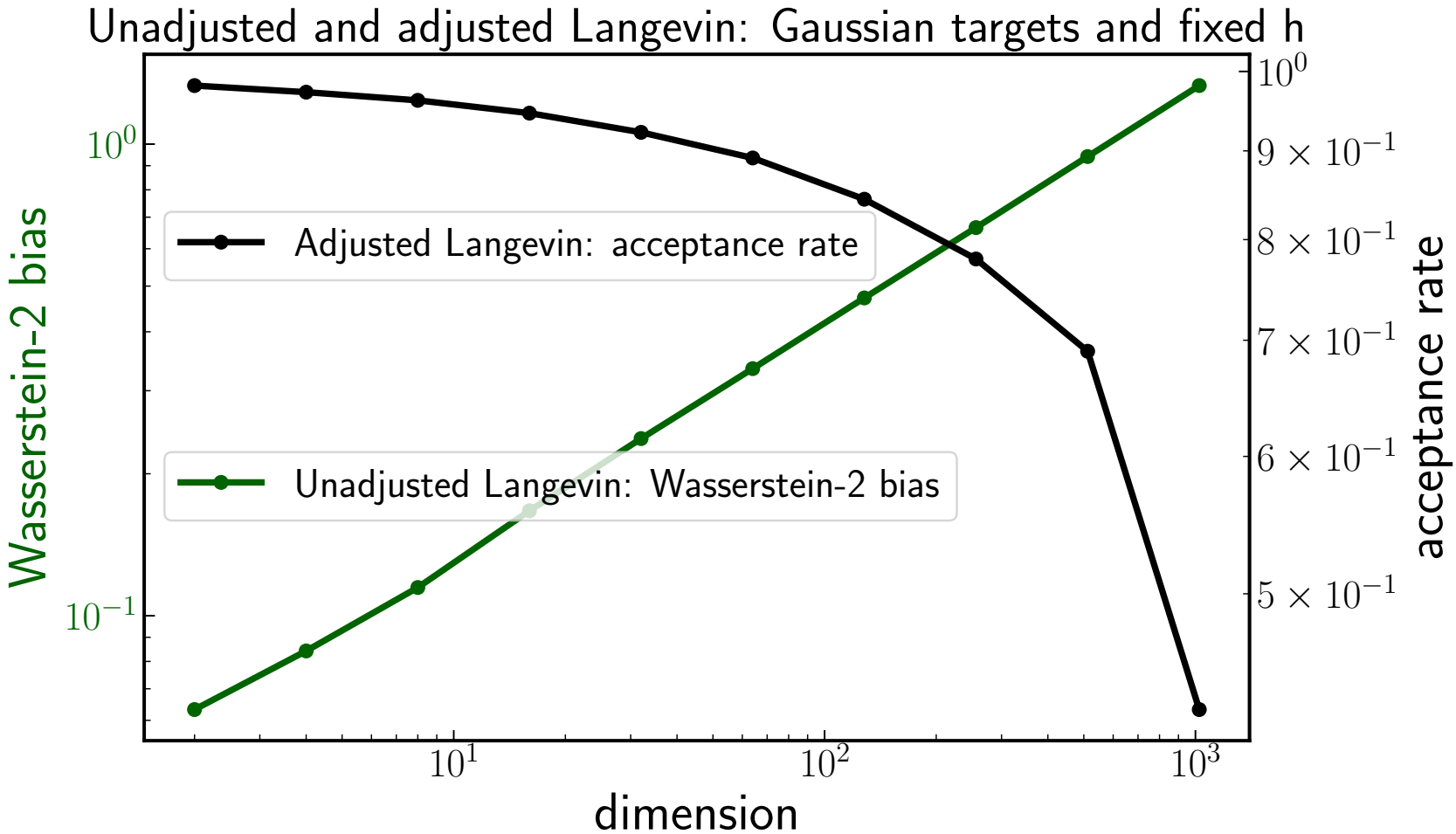
However, for this algorithm, h must be small when d is large

- Existing theory suggests $h \sim 1/d^{1/3}, 1/d^{1/2}, 1/d$ depending on notion of convergence and distribution of X_0

  [Roberts, Rosenthal 1998], [Christensen, Roberts, Rosenthal 2005], [Dwivedi, Chen, Wainwright, Yu 2018], [Chewi, Lu, Ahn, Cheng, Gouic, Rigollet 2021], etc

- This is necessary for non-negligible acceptance rates

Performance illustration: for fixed, non-decreasing stepsize h

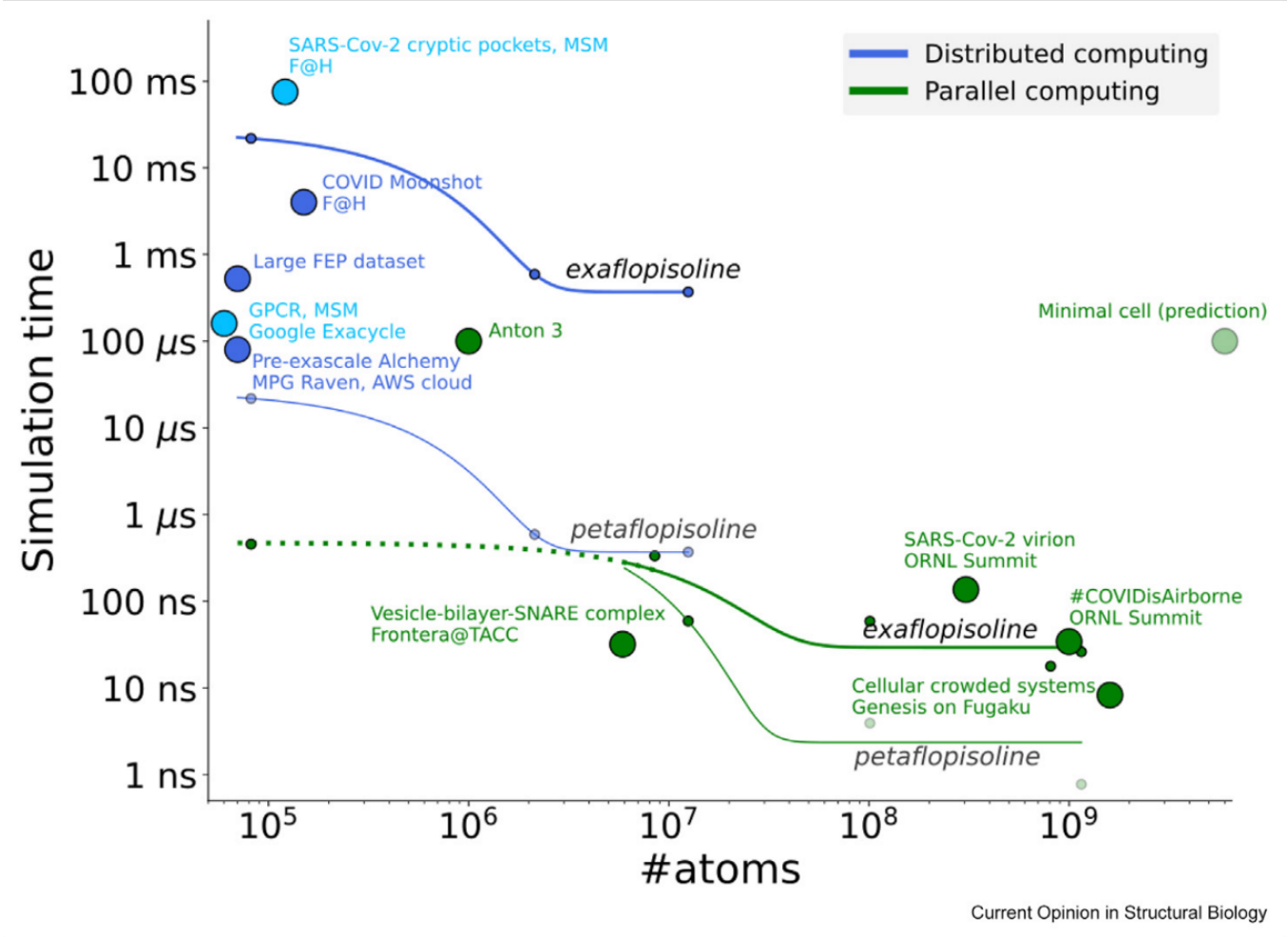


- Fixed h seems to fail when d increases
- Existing theory: power law decay is required

Looks okay for $d =$ thousands, but hard for *billions*

Empirical evidence in molecular dynamics

Variants of unadjusted Langevin routinely applied in high dims

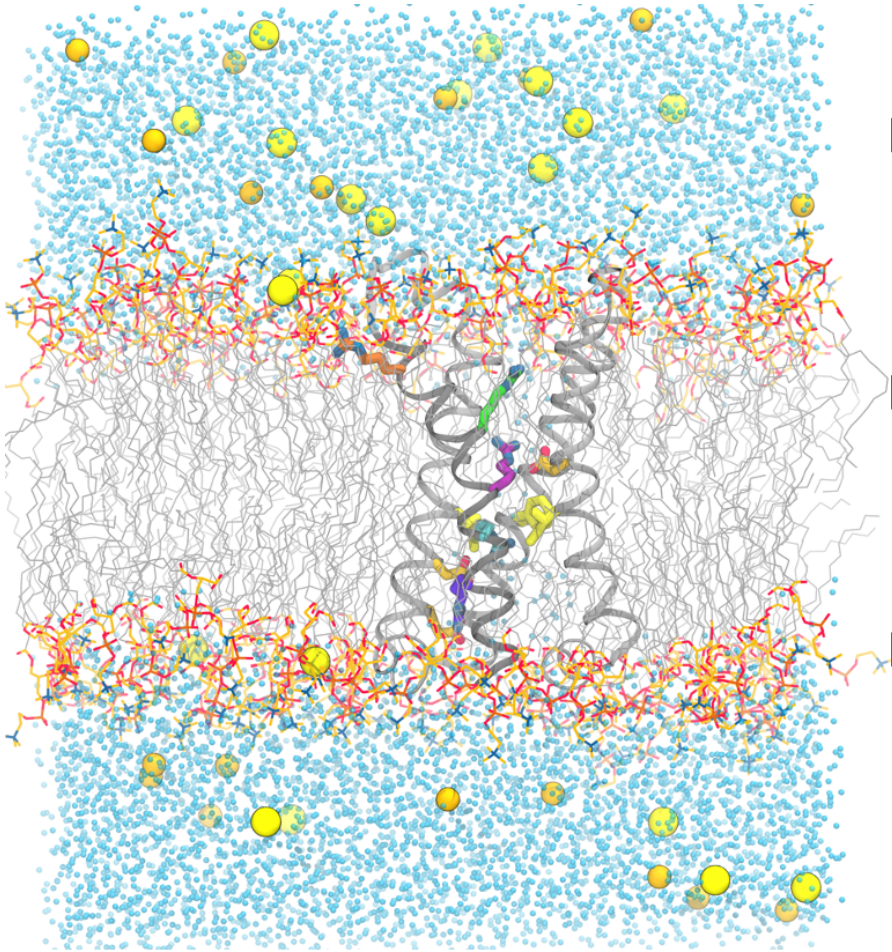


[Gapsys, Kopec, Matthes, de Groot 2024]

This is achieved using $h =$ a few fs, without reducing stepsize

Which could be the catch?

Often high dimensionality occurs when many nuisance variables are required to accurately describe the remaining variables' distribution



[Figure credits to Spencer Guo]

Molecular dynamics (MD) example

- We often care about averages with respect to a few atoms in the voltage sensing protein in the middle
- We usually do not care about averages with respect to atoms in the lipid or water molecules
- We need all the atoms to accurately describe the system

We are interested in *a small part!*

Disclaimer: the potential V in MD is more complex than considered in our analysis

Measuring errors of low dimensional marginals

Goal: measure 1D marginal error $W_2(\pi^{(j)}, \pi_h^{(j)}), 1 \leq j \leq d$

Measuring errors of low dimensional marginals

Goal: measure 1D marginal error $W_2(\pi^{(j)}, \pi_h^{(j)}), 1 \leq j \leq d$

through a metric that incorporates all coordinates

Measuring errors of low dimensional marginals

Goal: measure 1D marginal error $W_2(\pi^{(j)}, \pi_h^{(j)}), 1 \leq j \leq d$

through a metric that incorporates all coordinates

Standard W_2 metric: ℓ^2 measures full coordinates

$$W_2(\pi, \pi_h) = \left(\min_{\gamma \in \Pi(\pi, \pi_h)} \int |x - y|_2^2 \, \gamma(\mathrm{d}x, \mathrm{d}y) \right)^{1/2}$$

where $\Pi(\pi, \pi_h)$ is the set of all couplings between π and π_h

New W_{2,ℓ^∞} metric: replace ℓ^2 by ℓ^∞

$$W_{2,\ell^\infty}(\pi, \pi_h) = \left(\min_{\gamma \in \Pi(\pi, \pi_h)} \int |x - y|_\infty^2 \gamma(\mathrm{d}x, \mathrm{d}y) \right)^{1/2}$$

Property: $W_{2,\ell^\infty}(\pi, \pi_h) \geq W_2(\pi^{(j)}, \pi_h^{(j)})$ serves an upper bound

- Extends to any K marginals at the cost of a factor $\sqrt{K}$

How would bias behave under the W_{2,ℓ^∞} metric?

Example: W_{2,ℓ^∞} bias for product measures

Consider $\pi \propto \exp(-V)$ where $V(x) = \sum_{i=1}^d V_i(x^{(i)})$ where $x = (x^{(1)}, \dots, x^{(d)})$ and $\alpha \leq \nabla^2 V_i \leq \beta$. Then it holds that

$$W_{2,\ell^\infty}(\pi, \pi_h) = O(\frac{\beta}{\alpha}\sqrt{h \log(2d)})$$

Example: W_{2,ℓ^∞} bias for Gaussian measures

Consider $\pi \propto \exp(-V)$ and $V(x) = \frac{1}{2}(x-m)^T\Sigma^{-1}(x-m)$ where $m \in \mathbb{R}^d$ and $\alpha I \preceq \Sigma^{-1} \preceq \beta I$. Then it holds

$$W_{2,\ell^\infty}(\pi, \pi_h) = O\left(\sqrt{h \log(2d)}\right)$$

Both cases: W_{2,ℓ^∞} bias, and 1D W_2 bias, are nearly dimension free

Is this a universal phenomenon?

Negative example: W_{2,ℓ^∞} bias for rotated product measures

Consider $\pi = \rho^{\otimes d}$ where ρ is a 1D centered distribution, such that the mean of ρ and the biased ρ_h differs by $\delta > 0$.

Let $\tilde{\pi} = Q\#\pi$ where Q is a rotation $(Qx)^{(1)} = \frac{1}{\sqrt{d}} \sum_{i=1}^{d} x^{(i)}$. Then

$$W_{2,\ell^\infty}(\tilde{\pi}, \tilde{\pi}_h) \geq \sqrt{d}\delta$$

where $\tilde{\pi}_h$ is the corresponding biased distribution for $\tilde{\pi}$

Proof sketch: we have $\tilde{\pi}_h = Q\#\pi_h$

$$W_{2,\ell^\infty}(\tilde{\pi}, \tilde{\pi}_h) \geq W_{1,\ell^\infty}(\tilde{\pi}, \tilde{\pi}_h) \geq \left| \int x^{(1)}(\tilde{\pi} - \tilde{\pi}_h) \right|$$

$$= \left| \int (\frac{1}{\sqrt{d}} \sum_{i=1}^{d} x^{(i)})(\pi - \pi_h) \right| = \sqrt{d}\delta$$

This example exhibits dense and strong interactions

Negative example: W_{2,ℓ^∞} bias for rotated product measures

Consider $\pi = \rho^{\otimes d}$ where ρ is a 1D centered distribution, such that the mean of ρ and the biased ρ_h differs by $\delta > 0$.

Let $\tilde{\pi} = Q\#\pi$ where Q is a rotation $(Qx)^{(1)} = \frac{1}{\sqrt{d}} \sum_{i=1}^{d} x^{(i)}$. Then

$$W_{2,\ell^\infty}(\tilde{\pi}, \tilde{\pi}_h) \geq \sqrt{d}\delta$$

where $\tilde{\pi}_h$ is the corresponding biased distribution for $\tilde{\pi}$

Proof sketch: we have $\tilde{\pi}_h = Q\#\pi_h$

$$W_{2,\ell^\infty}(\tilde{\pi}, \tilde{\pi}_h) \geq W_{1,\ell^\infty}(\tilde{\pi}, \tilde{\pi}_h) \geq \left| \int x^{(1)}(\tilde{\pi} - \tilde{\pi}_h) \right|$$

$$= \left| \int (\frac{1}{\sqrt{d}} \sum_{i=1}^{d} x^{(i)})(\pi - \pi_h) \right| = \sqrt{d}\delta$$

This example exhibits dense and strong interactions

Active research: identifying boundary of delocalization phenomenon

Results: Delocalization for sparse or weak interactions

[**Chen**, Cheng, Niles-Weed, Weare 2024]

We consider $V \in C^2(\mathbb{R}^d)$ with $\alpha I \preceq \nabla^2 V \preceq \beta I$ in which $\alpha > 0$

Theorem: W_{2,ℓ^∞} bias for sparse-interaction potentials

If matrix $\Pi_{j=1}^k \nabla^2 V(x_j)$ is $(k+1)^n$-sparse for any $x_j, 1 \le j \le k$,

$$W_{2,\ell^\infty}(\pi, \pi_h) \le C_{\alpha,\beta} \sqrt{h(\log(2d))^{n+3}}$$

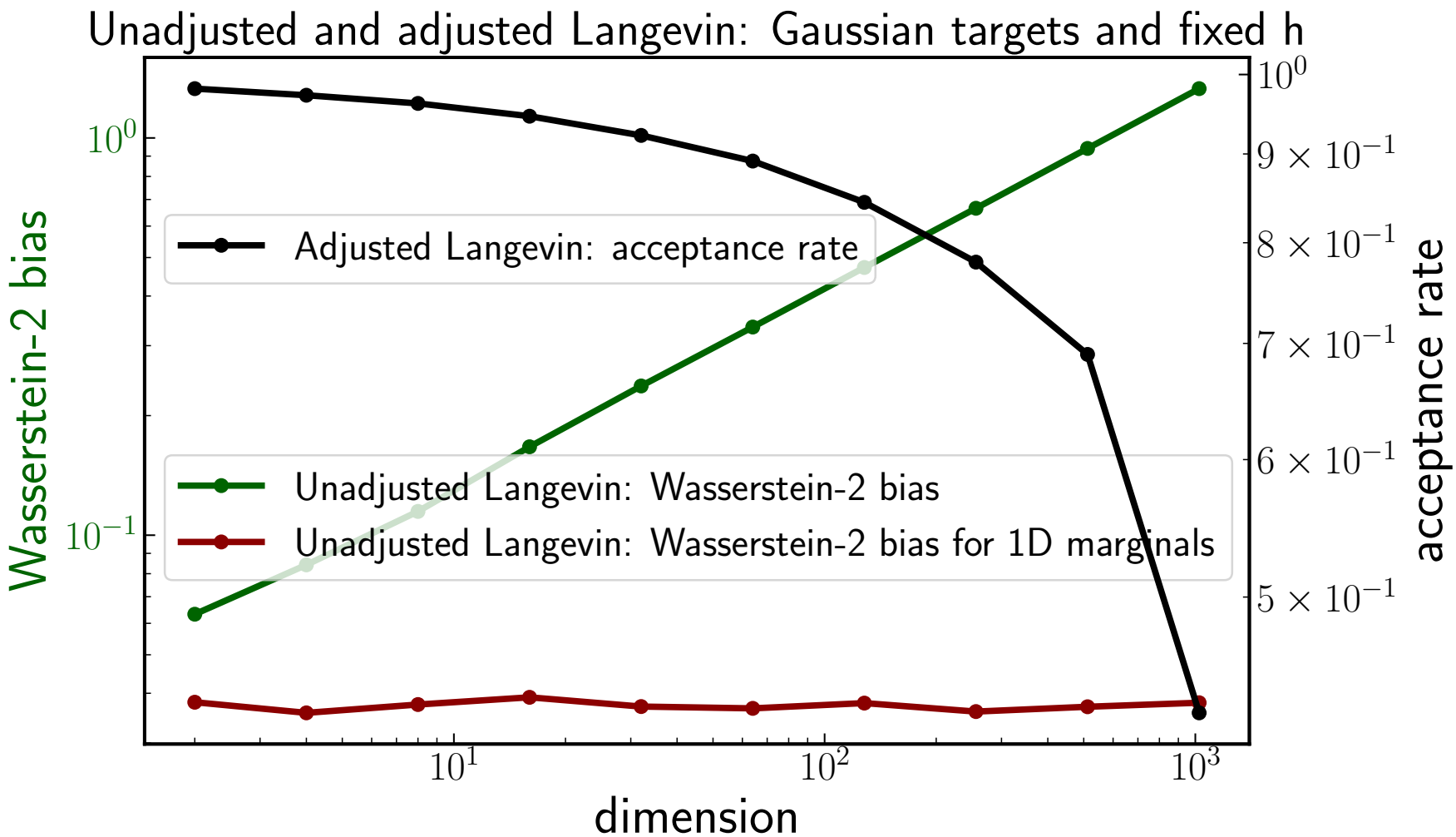
Theorem: W_{2,ℓ^∞} bias for weak-interaction potentials

If the off-diagonal part $|\nabla^2 V(x) - \mathrm{diag}(\nabla^2 V(x))|_\infty \le \delta\alpha$ for $\delta < 1$,

$$W_{2,\ell^\infty}(\pi, \pi_h) \le C \frac{\beta}{\alpha} \sqrt{h \log(2d)}$$

- Proof based on technical (multiple-step) coupling arguments and ℓ^∞ analysis for propagators of unadjusted Langevin
- Results for delocalization in KL divergence: entropy methods, propagation of chaos, and marginal hierarchy [Lacker, Zhou 2025]

Updated performance illustration: for fixed stepsize h



- Same for K-marginals, if K is independent of dimension
  (under the assumption of Gaussian or sparse/weak interactions)

Summary message: delocalization of bias

Simple insight: Even if a system is extremely high dimensional, bias of a small part of the system can be nearly dimension-free

- No curse of dims if interested in low-dim marginals!
  (under the assumption of Gaussian or sparse/weak interactions)

Algorithmic implications

- "Do not Metropolize in very high dims!"
- Support for approximate versus unbiased in very high dims

Theoretical outlook (coming soon)

- Non-log-concave measures (e.g., by reflection coupling)
- Extension to other algorithms (HMC, underdamped, etc.)
- Different settings: function space measure AC to Gaussians

  Preconditioned Crank–Nicolson [Cotter, Roberts, Stuart, White 2013]

Improving the scaling of widely used Metropolized samplers

- Metropolized samplers widely used (not as high dim as billions)
- Popular Metropolized ensemble sampler by Goodman-Weare

Ensemble samplers with **affine invariance**
J Goodman, J Weare - … in applied mathematics and computational science, 2010 - msp.org
… of a practical **sampler** that has this **affine invariance** property for any general class of densities. In this paper we propose a family of **affine invariant ensemble samplers**. An **ensemble**, X, …
☆ Save 99 Cite Cited by 3753 Related articles All 7 versions Web of Science: 2424 »

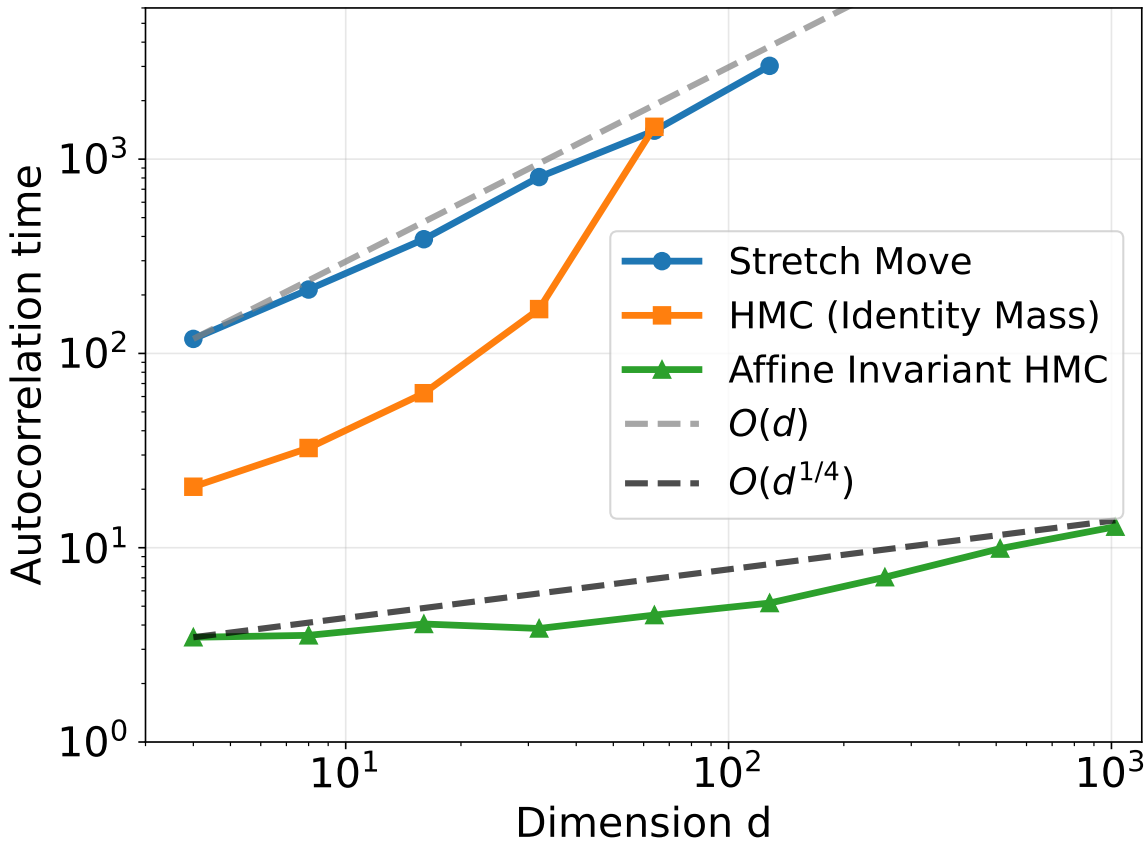
emcee: the MCMC hammer
D Foreman-Mackey, DW Hogg, D Lang… - Publications of the …, 2013 - iopscience.iop.org
We introduce a stable, well tested Python implementation of the affine-invariant ensemble sampler for Markov chain Monte Carlo (MCMC) proposed by Goodman & Weare (2010). The …
☆ Save 99 Cite Cited by 14494 Related articles All 12 versions Web of Science: 10196 »

- Implemented in `emcee` package, routinely used in applications
- Affine invariance property: automatically handle ill-conditioning
- Affine-invariant ensemble sampler is reported to behave well for moderate dimensions but suffer from higher dims (e.g. $d \geq 50$)

  [Huijser, Goodman, Brewer 2015]

New algorithm: Affine invariant ensemble Hamiltonian Monte Carlo

[**Chen** 2025]

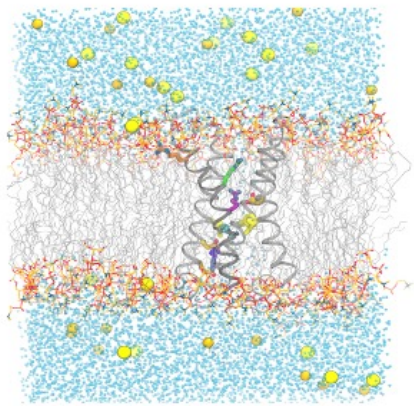


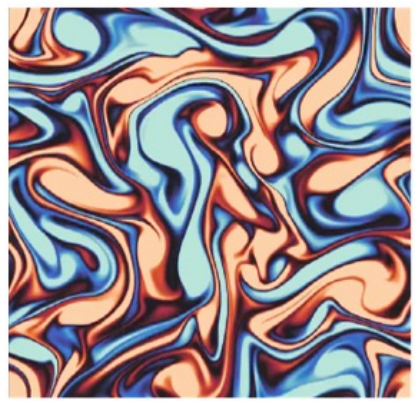
- Experiment on $\exp\left(-\int_0^1 \frac{1}{2}(\partial_x u(x))^2 + (1-u^2(x))^2 \mathrm{d}x\right)$
- Affine invariant HMC automatically "tunes mass" w/ small costs

Active research: developing algorithm & software "h-emcee"

Outline of the talk

1. Analysis of unadjusted Langevin in high dimensions
   (analysis w/ methodological insights)
   - A new “delocalization of bias” phenomenon for understanding
   - Dimension-independent behavior for low-dimensional marginals





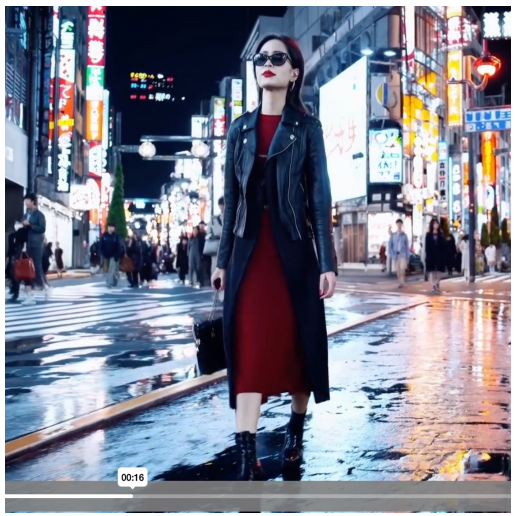
2. Design and application of generative probability flows
   (methodology w/ analytical insights)
   - An efficient "optimal Lipschitz energy" criteria for design
   - Dimension-robust performance with respect to resolution

Success of generative modeling

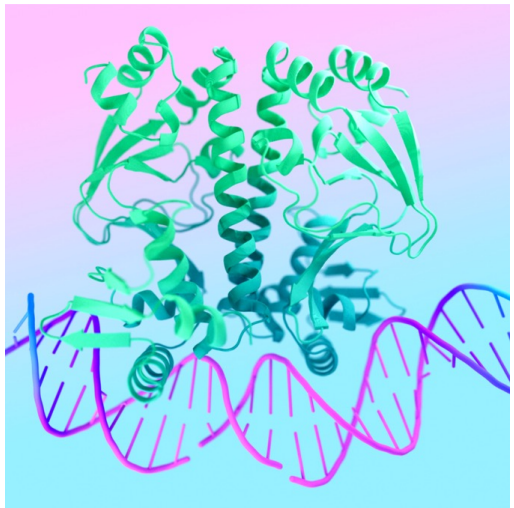
Generative modeling

Goal: draw new samples from π, given data $\{x_i\}_{i=1}^N \sim \pi$

DALL·E 3



Sora



Alpha Fold 3

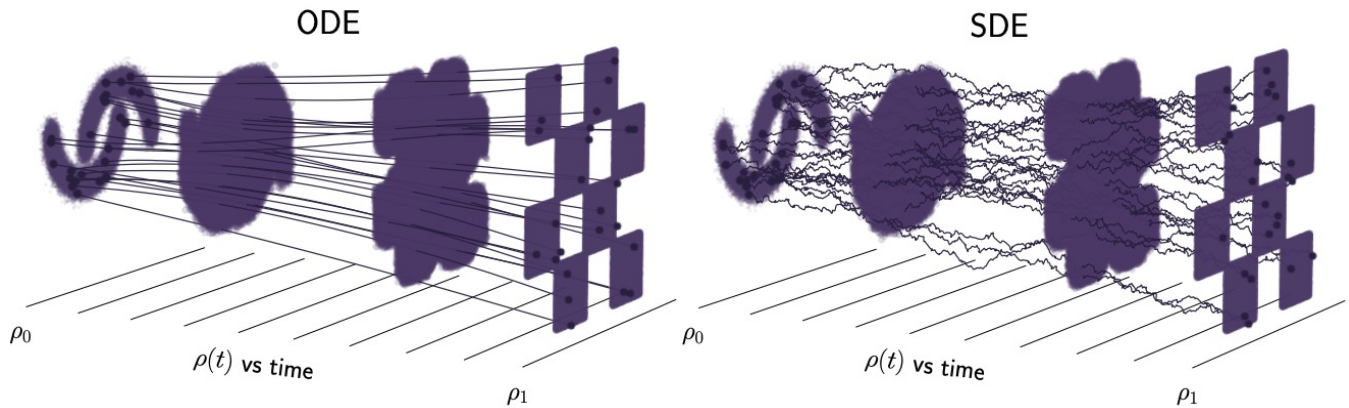
Breakthrough in computer vision and success extended to sciences

DALL·E 3: `https://openai.com/index/dall-e-3/`
Sora: `https://openai.com/sora/`
Alpha Fold 3: `https://deepmind.google/science/alphafold/`

State of the art methodology: flow and diffusion dynamics

Recent advances in generative modeling driven by building dynamics of measures that iteratively refine the generation to the desired

Diffusion models, score-based generative models



Flow matching, rectified flow, stochastic interpolants, ...

[Sohl-Dickstein et al 2015], [Ho, Jain, Abbeel 2020], [Song et al 2021], [Peluchetti 2021], [De Bortoli et al. 2021], [Liu, Gong, Liu 2022], [Albergo, Vanden-Eijnden, 2022], [Lipman et al 2022], [Albergo, Boffi, Vanden-Eijnden 2023], [Shi et al 2023], etc.

Challenge: field data with a wide range of Fourier spectra

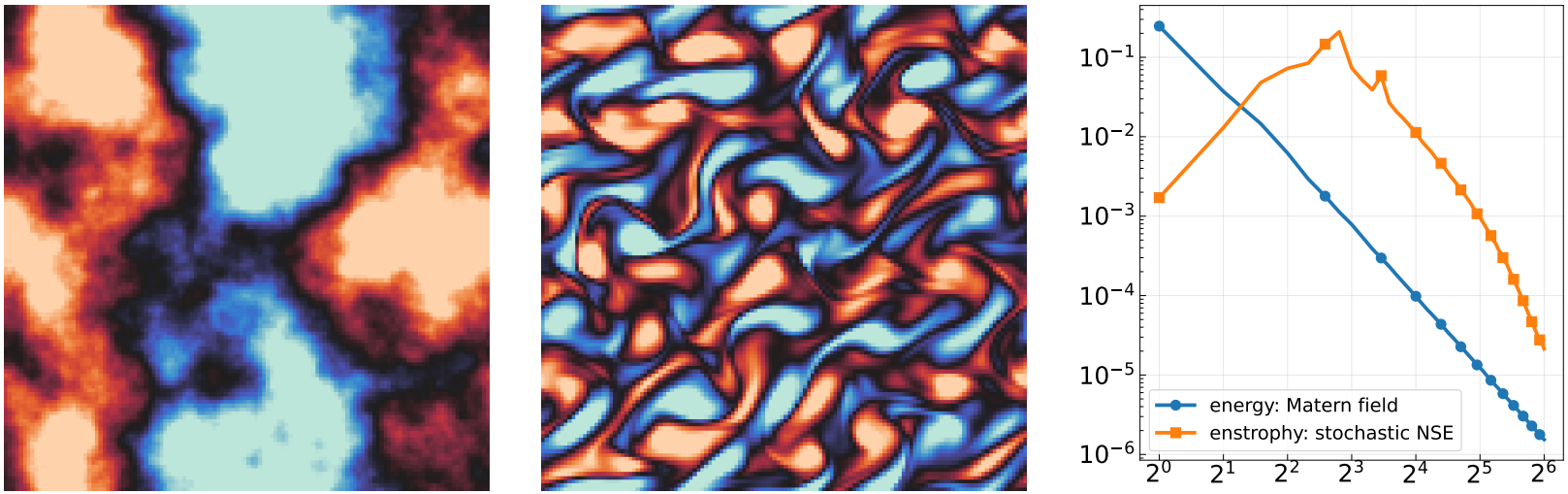


Figure: Examples of data samples from Matérn Gaussian processes (left panel) and invariant measure of stochastically forced Navier-Stokes (middle panel). The right panel shows their energy and enstrophy spectra

- Precise fine scale accuracy is numerically challenging
- Existing function space framework often aims for a different goal of coarse scale stability (stable under resolution refinement)

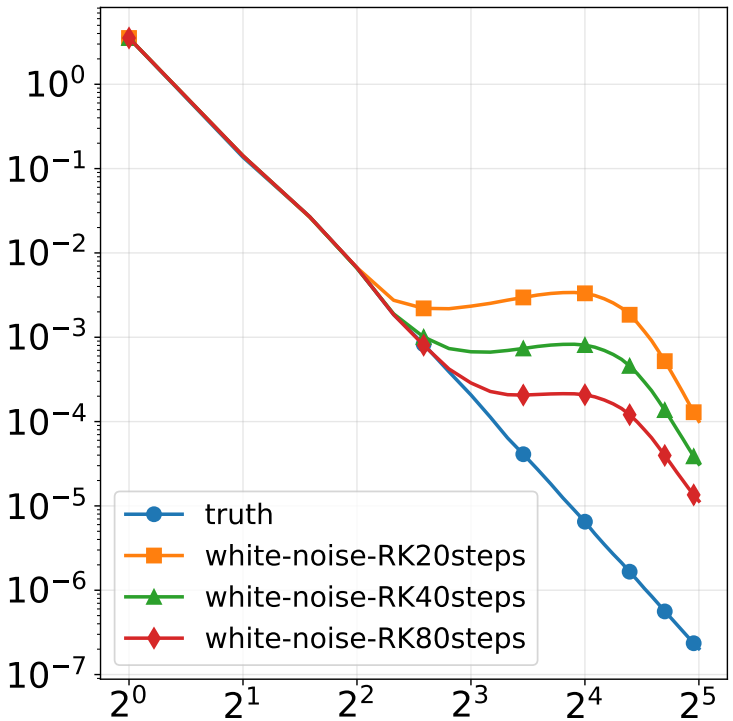
Function space generative models [Lim et al 2023], [Hagemann, Ruthotto, Steidl, Yang 2023], [Pidstrigach, Marzouk, Reich, Wang 2023], [Kerrigan, Migliorini, Smyth 2023], etc.

Wavelets and multiscale generative models [Guth, Coste, Bortoli, Mallat 2022], etc.

Numerical challenge for approximating generative dynamics

Dynamical approaches require approximating ODEs or SDEs

Case study: z white noise and $x_1 \sim \mathrm{N}(0, C_1)$ with $C_1 = (-\Delta + I)^{-3}$



- Much more costs when resolution (or dimension) increases
- Many advanced integration methods can help. Fundamentally, the challenge remains when resolution is very fine

Optimal transport approach for design of dynamics

Minimal kinetic energy in optimal transport approaches

$$\min_{b_t} \ \mathbb{E}[\|b_t(X_t)\|_2^2]$$
$$\text{s.t.} \ \ \dot{X}_t = b_t(X_t), X_0 \sim \mathrm{N}(0, \mathrm{I}), X_1 \sim \rho^*$$

- Benamou-Brenier formula [Benamou, Brenier 2000]
- Trajectories are straight lines: one step integration is exact
- However, $b_t(x)$ can be spatially highly irregular

  [Tsimpos, Ren, Zech, Marzouk 2025]

Widely discussed and pursued in generative models [Liu, Gong, Liu 2022], [Albergo, Vanden-Eijnden, 2022], etc.

Entropy regularized OT (a.k.a. Schrödinger's bridges) [Léonard 2014]

Efficient algorithm in generative modeling: [Bortoli, Thornton, Heng, Doucet 2021] [Shi, Bortoli, Campbell, Doucet 2023], [**Chen**, Goldstein, Hua, Albergo, Boffi, Vanden-Eijnden 2024], [Pooladian, Niles-Weed 2024], etc.

Minimal Lipschitz energy [Chen, Vanden-Eijnden, Xu 2025]

$$\min_{b_t} \ \mathbb{E}[\|\nabla b_t(X_t)\|_2^2]$$

$$\text{s.t.} \ \ \dot{X}_t = b_t(X_t), X_0 \sim \mathsf{N}(0, \mathrm{I}), X_1 \sim \rho^*$$

Minimal Lipschitz energy [Chen, Vanden-Eijnden, Xu 2025]

$$\min_{b_t} \ \mathbb{E}[\|\nabla b_t(X_t)\|_2^2]$$
$$\text{s.t. } \dot{X}_t = b_t(X_t), X_0 \sim \mathrm{N}(0, \mathrm{I}), X_1 \sim \rho^*$$

Practical strategy: constrained optimization in the class of dynamics

$$b_t(x) = \mathbb{E}[\dot{I}_t | I_t = x], \quad I_t = \alpha_t z + \beta_t x_1$$

- $\alpha_0 = \beta_1 = 1, \alpha_1 = \beta_0 = 1$ to be optimized
- Noise $z \sim \mathrm{N}(0, \mathrm{I}) \perp x_1 \sim \rho^*$ the data distribution
- Given α_t, β_t, the b_t is learned from data using the objective

$$\min_{\hat{b}} L(\hat{b}) = \int_0^1 \mathbb{E}[\|\hat{b}_t(I_t) - \dot{I}_t\|_2^2] \, \mathrm{d}t$$

  where the expectation is replaced by empirical averages
- For any such α_t, β_t, using the drift b_t yields $X_1 \sim \rho^*$ [Gyöngy 1986]

[Liu, Gong, Liu 2022], [Albergo, Vanden-Eijnden, 2022], [Lipman, Chen, Ben-Hamu, Nickel, Le 2022], [Albergo, Boffi, Vanden-Eijnden 2023], etc.

Analytically optimized schedules

Gaussian targets: $x_1 \sim \mathrm{N}(0, C) \perp z \sim \mathrm{N}(0, \mathrm{I})$ in d dims. Let eigenvalues of C be $1 \geq \lambda^{(1)} \geq \ldots \geq \lambda^{(d)} > 0$. Denote $M^\star = 1/\lambda^{(d)}$

Theorem: For the common linear schedule $\alpha_t = 1 - t, \beta_t = t$

$$\int_0^1 \mathbb{E}[\|\nabla b_t(X_t)\|_2^2]\mathrm{d}t = \Omega(\sqrt{M^\star}), \quad \max_{t,x} \|\nabla b_t(x)\|_2 = \Omega(M^\star)$$

If we optimize Lipschitz energy over all possible linear stochastic interpolants I_t with scalar schedules, then

$$\alpha_t = \sqrt{\frac{(M^\star)^{1-t} - 1}{M^\star - 1}}, \beta_t = \sqrt{\frac{M^\star - (M^\star)^{1-t}}{M^\star - 1}}$$

For the optimal solution, $\|\nabla b_t(x)\|_2 = \frac{1}{2} \log M^\star$ for any t, x

- Other analytic results on Gaussian mixtures using Euler-Lagrange equation and general distributions using Beltrami Identity

Optimized schedule: performance for Gaussian measures

Target $\rho^* = \mathrm{N}(0, C_1)$, where $C_1 = (-\Delta + I)^{-3}$

- Discretize on $N \times N$ grid points
- Compare to standard schedule $\alpha_t = 1 - t, \beta_t = t$

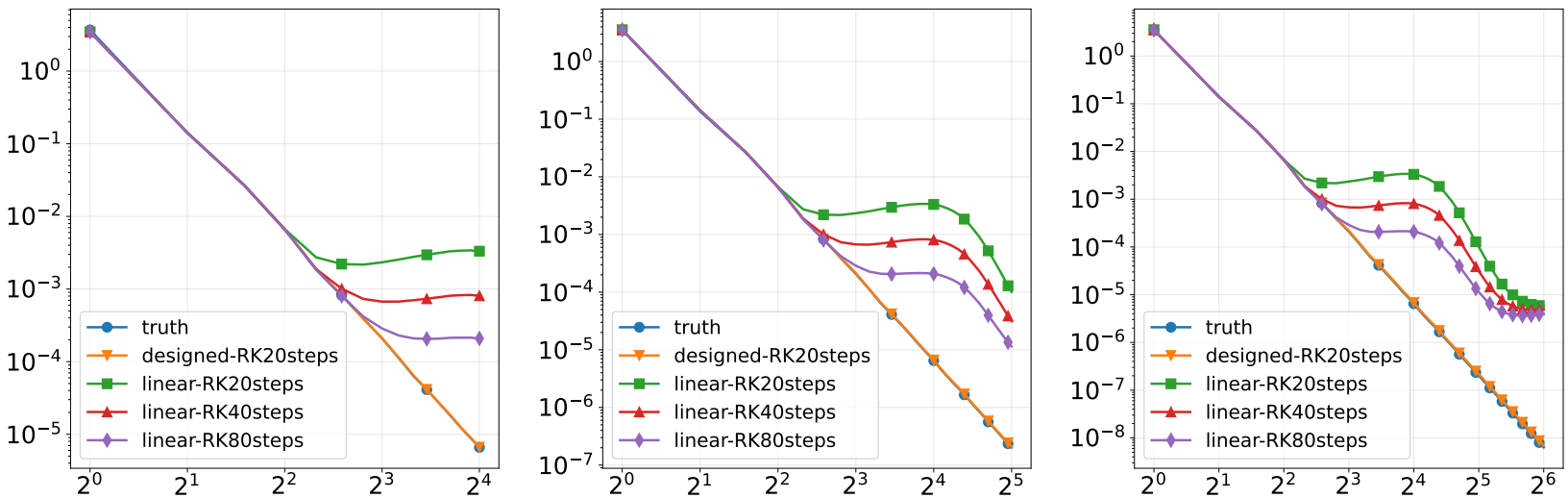


Figure: Gaussian measure example. Linear schedule versus optimized schedules. Left: 32×32; middle: 64×64; right: 128×128

Resolution robust performance with the same integration steps

Performance for invariant distribution to stochastic Allen-Cahn

Target $\rho^*(u) \propto \exp\left(-\int_0^1 \frac{1}{2}(\partial_x u(x))^2 + (1-u^2(x))^2 \mathrm{d}x\right)$

- Invariant distribution to stochastic Allen-Cahn
- Discretize on N grid points

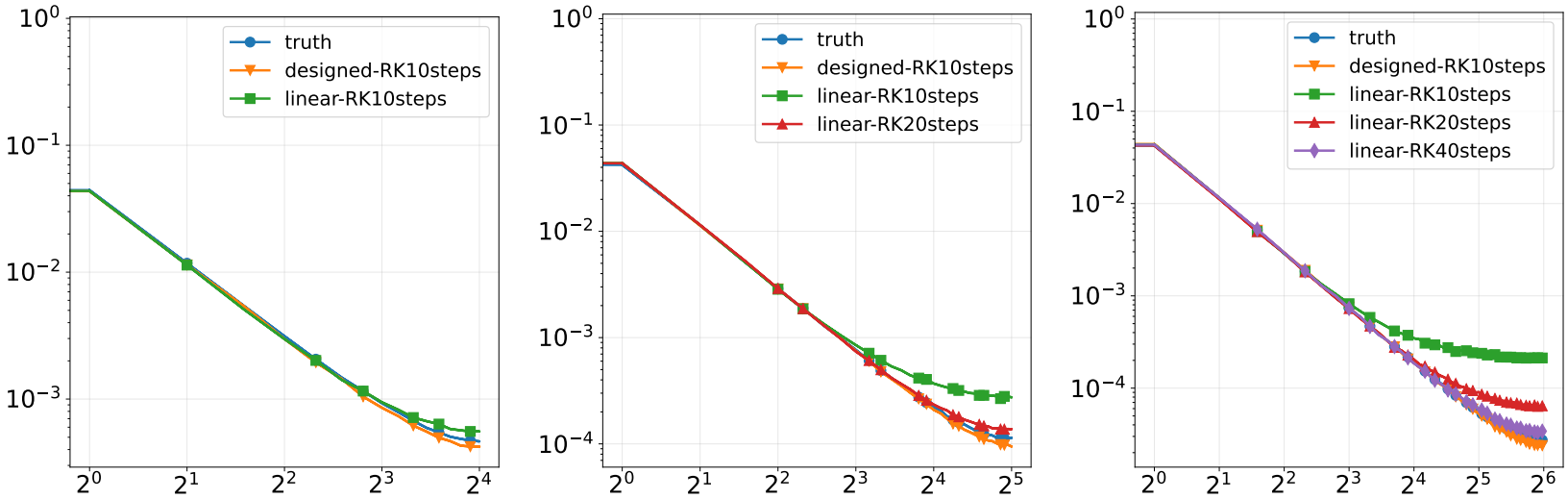


Figure: Stochastic Allen-Cahn example. Linear schedule versus optimized schedules. Left: $N = 32$; middle: $N = 64$; right: $N = 128$

All experiments are done using 2M-parameter-Unet to train b_t

Again robust performance with the same integration steps

Case study: 2d NSE with stochastic forcing

$$\mathrm{d}\omega + v \cdot \nabla \omega \mathrm{d}t = \nu \Delta \omega \mathrm{d}t - \alpha \omega \mathrm{d}t + \epsilon \mathrm{d}\eta \quad \text{on } \mathbb{T}^2$$

- vorticity ω, velocity v, and $\mathrm{d}\eta$ forcing Ergodicity: [Hairer, Mattingly, 2006]
- $\nu = 10^3$, $\mathrm{d}\eta$ random forcing acts on a few Fourier modes

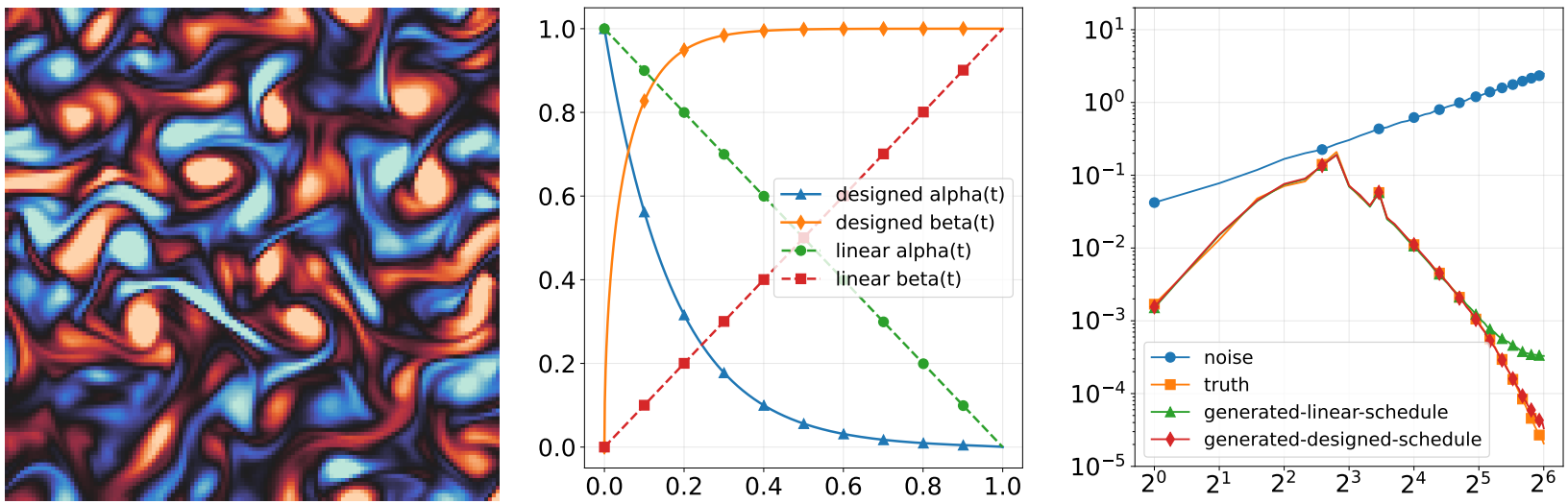
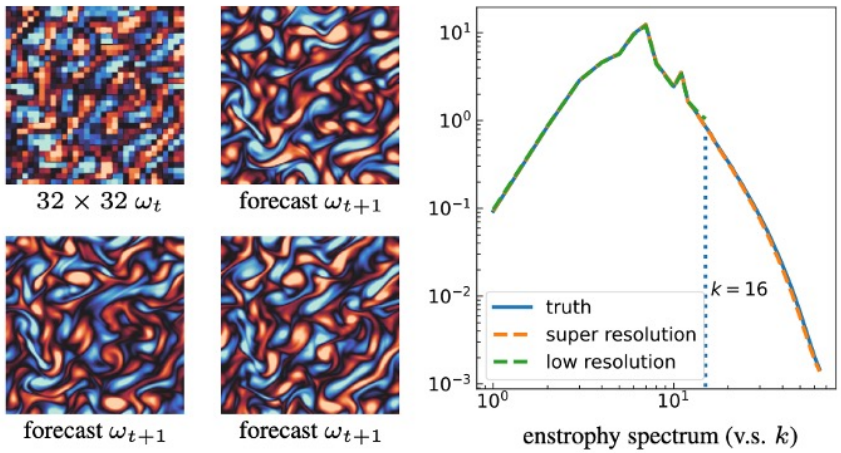


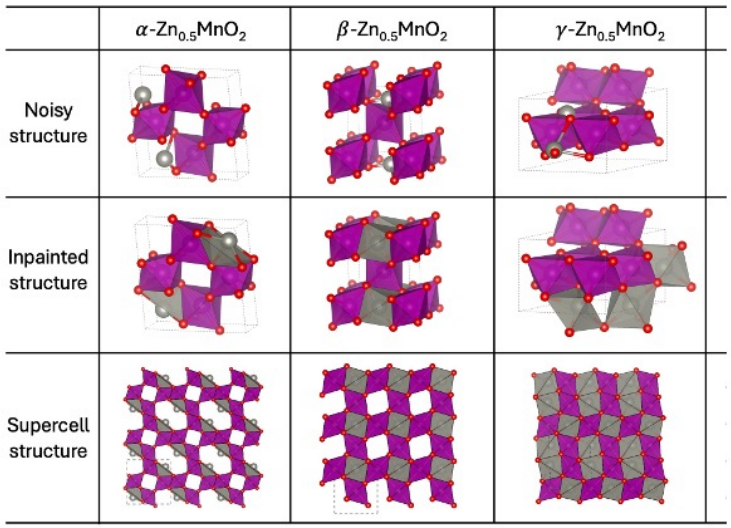
Figure: We use white noise and 10 RK4 integration steps. 128×128

- Left: generated samples w/ optimized schedule
- Middle: linear versus optimized schedule ($M^* = 10^5$)
- Right: enstrophy spectra of truth, noise, and generations

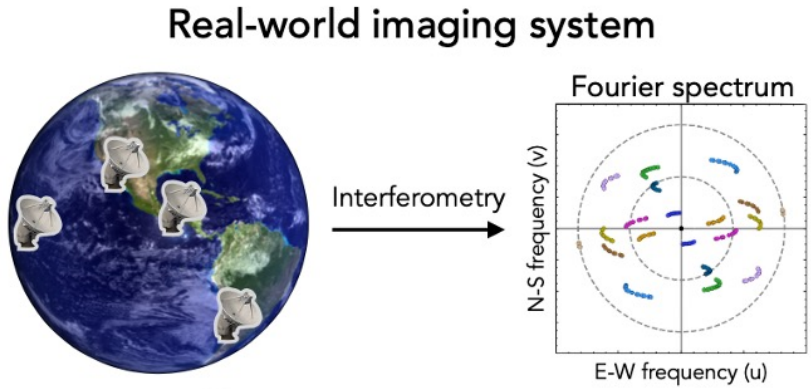
Towards more numerical insights for other scientific tasks

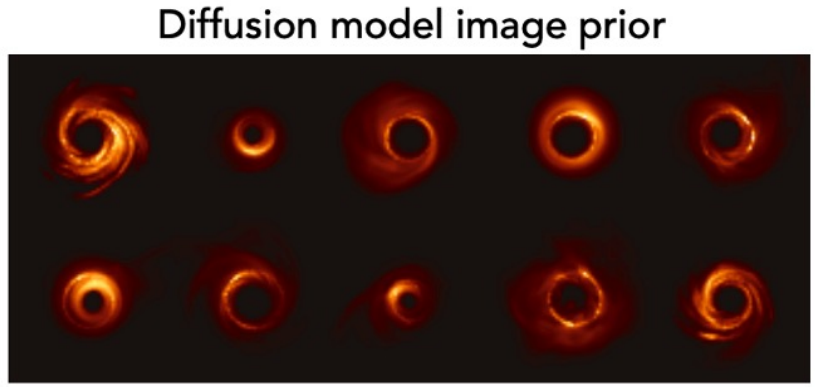


[Chen, Goldstein, Hua, Albergo, Boffi, Vanden-Eijnden 2024]



[Dai, Zhong, Deng, Chen, Ceder 2024]





[Sun, Wu, Chen, Feng, Bouman 2023]
[Wu, Sun, Chen, Zhang, Yue, Bouman 2024]

Conclusion

Understand/alleviate high-dimensional curse in stochastic dynamical methods

- For approximate samplers, bias of low-dimensional marginals can exhibit dimension independent scaling
- Unbiased affine invariant samplers achieve much better dimensional scaling by carefully using Hamiltonian dynamics
- Lipschitz-optimal design of generative dynamics can achieve dimension robust performance with respect to resolution

Towards more practical algorithms and theoretical insights

Flattening the high-dimensional scaling curve!

References

- Y. Chen, X. Cheng, J. Niles-Weed, J. Weare. Convergence of Unadjusted Langevin in High Dimensions: Delocalization of Bias. arXiv:2408.13115, 2024
- Y. Chen. New Affine Invariant Ensemble Samplers and Their Dimensional Scaling. arXiv:2505.02987, 2025
- Y. Chen, E. Vanden-Eijnden. Scale-Adaptive Generative Flows for Multiscale Scientific Data. arXiv:2509.02971, 2025
- Y. Chen, E. Vanden-Eijnden, J Xu. Lipschitz-Guided Design of Interpolation Schedules in Generative Models. arXiv:2509.01629, 2025
- Y. Chen, M. Goldstein, M. Hua, M. Albergo, N. Boffi, E. Vanden-Eijnden. Probabilistic Forecasting with Stochastic Interpolants and Föllmer Processes. ICML 2024

Thank you!

Back-Up Slides

Simple summary of methodology in one slide

- Corruption path via interpolation between data and noise

$$I_t = \alpha_t z + \beta_t x_1, \quad \alpha_0 = \beta_1 = 1, \alpha_1 = \beta_0 = 1$$

where, noise $z \sim \mathrm{N}(0, \mathrm{I}) \perp x_1 \sim \rho^*$ the data distribution

- Generation dynamics via numerically solving

$$\mathrm{d}X_t = b_t(X_t)\mathrm{d}t, \quad b_t(x) = \mathbb{E}[\dot{I}_t | I_t = x]$$

Thm: for such b_t, it holds $X_1 \sim \rho^*$ the target [Gyöngy 1986]

- b_t can be learned from data using the objective

$$\min_{\hat{b}} L(\hat{b}) = \int_0^1 \mathbb{E}[\|\hat{b}_t(I_t) - \dot{I}_t\|_2^2] \, \mathrm{d}t$$

where the expectation is replaced by empirical averages

[Liu, Gong, Liu 2022], [Albergo, Vanden-Eijnden, 2022], [Lipman, Chen, Ben-Hamu, Nickel, Le 2022], [Albergo, Boffi, Vanden-Eijnden 2023], etc.

Delocalization of Bias: Technical Details

Sketch of arguments through coupling with same Brownion motion

Continuous time $Y_t, t \in [kh, (k+1)h]$ and unadjusted X_{kh}

$$X_{(k+1)h} = X_{kh} - h\nabla V(X_{kh}) + \sqrt{2}(B_{(k+1)h} - B_{kh})$$

One step bound:

$$\sqrt{\mathbb{E}[|X_{(k+1)h} - Y_{(k+1)h}|_\infty^2]}$$
$$\leq \underbrace{\sqrt{\mathbb{E}[|X_{(k+1)h} - \overline{Y}_{(k+1)h}|_\infty^2]}}_{\text{(a)}} + \underbrace{\sqrt{\mathbb{E}[|\overline{Y}_{(k+1)h} - Y_{(k+1)h}|_\infty^2]}}_{\text{“discretization error” = } O(\beta h^{3/2}\sqrt{\log(2d)})}$$

where $\overline{Y}_{(k+1)h} = Y_{kh} - h\nabla V(Y_{kh}) + \sqrt{2}(B_{(k+1)h} - B_{kh})$

$$\text{(a)} = \sqrt{\mathbb{E}[|X_{kh} - Y_{kh} - h(\nabla V(X_{kh}) - \nabla V(Y_{kh}))|_\infty^2]}$$
$$= \sqrt{\mathbb{E}[|H_k(X_{kh} - Y_{kh})|_\infty^2]}$$

where $H_k = I - h\int_0^1 \nabla^2 V(uX_{kh} + (1-u)Y_{kh})\mathrm{d}u$

Sketch of arguments: multiple-step coupling

- One-step iteration

$$\sqrt{\mathbb{E}[|X_{(k+1)h} - Y_{(k+1)h}|_\infty^2]} \le \sqrt{\mathbb{E}[|H_k(X_{kh} - Y_{kh})|_\infty^2]} + \mathrm{error}(1)$$

- Moving back and two-step iterations

$$\begin{aligned}
&\sqrt{\mathbb{E}[|H_k(X_{kh} - Y_{kh})|_\infty^2]} + \mathrm{error}(1) \\
\le &\sqrt{\mathbb{E}[|H_k(X_{kh} - \overline{Y}_{kh})|_\infty^2]} + \sqrt{\mathbb{E}[|H_k(\overline{Y}_{kh} - Y_{kh})|_\infty^2]} + \mathrm{error}(1) \\
= &\sqrt{\mathbb{E}[|H_k H_{k-1}(X_{(k-1)h} - Y_{(k-1)h})|_\infty^2]} + \mathrm{error}(2)
\end{aligned}$$

- N-step iterations

$$\begin{aligned}
&\sqrt{\mathbb{E}[|X_{(k+N)h} - Y_{(k+N)h}|_\infty^2]} \\
\le &\sqrt{\mathbb{E}[|H_{k+N-1} H_{k+N-2} \cdots H_k(X_{kh} - Y_{kh})|_\infty^2]} + \mathrm{error}(N) \\
\le &\exp(-\alpha N h)\sqrt{d}\sqrt{\mathbb{E}[|X_{kh} - Y_{kh}|_\infty^2]} + \mathrm{error}(N)
\end{aligned}$$

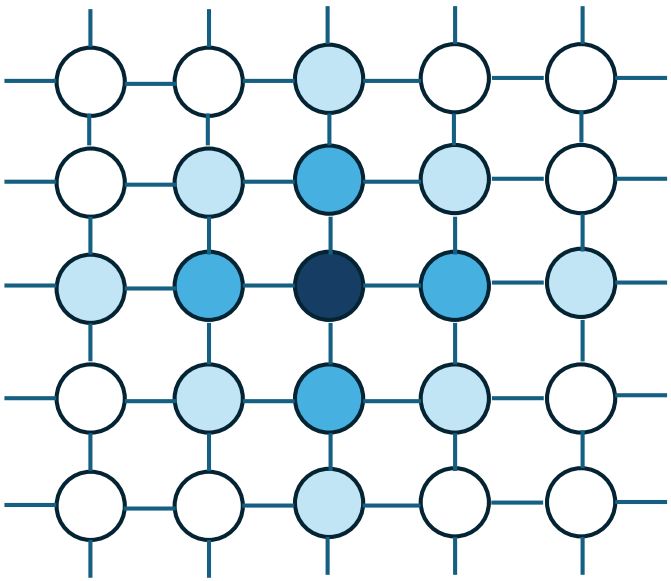
Here $N \sim (\log d)/h$ leads to a contraction

Result 1: Delocalization for sparse potentials

Theorem: W_{2,ℓ^∞} bias for sparse potentials

For $V \in C^2$ with $\alpha I \preceq \nabla^2 V \preceq \beta I$ that satisfies the sparsity condition illustrated in the figure with $s_k \leq C(k+1)^n$, then

$$W_{2,\ell^\infty}(\pi, \pi_h) \leq \sqrt{h \log(2d)} \left(O\left(\frac{\beta}{\alpha} \log(2d)\right) \right)^{\frac{n}{2}+1}$$



Some ith variable $x^{(i)}$

1st layer: $N_1(x^{(i)})$

2nd layer: $N_2(x^{(i)})$

Potential $V(x) = \sum_{i=1}^{d} V_i(x)$
and V_i only depends on $N_1(x^{(i)})$

Sparsity parameter $s_k = \max_i \left| N_k\left(x^{(i)}\right) \right|$. This example: $s_k = O(k^2)$

- Proof based on sparsity analysis for propagators of unadjusted Langevin to control ℓ^∞ errors; and coupling arguments

Result 2: Delocalization for weak potentials

Previous example: interaction is strong, but sparse

Theorem: W_{2,ℓ^∞} bias for weak potentials

Consider $V \in C^2$ such that $V = V_1 + V_2$ with $\alpha I \preceq \nabla^2 V_1 \preceq \beta I$, and $|\nabla^2 V_2|_\infty \leq \delta\alpha$ for some $\delta < 1$. Here V_1 corresponds to a product measure. Then

$$W_{2,\ell^\infty}(\pi, \pi_h) = O(\frac{\beta}{\alpha}\sqrt{h \log(2d)})$$

- Interaction is dense, but weak
- Entropy methods and KL divergence
  [Lacker, Zhou 2025]

Sketch of arguments: Bound discretization errors

- For general N

$$\text{error}(N) \leq \left(\sum_{i=1}^{N} \exp(-\alpha h(i-1)) \sqrt{s_{r_i}} \right) \cdot O\left(\beta h^{3/2} \sqrt{\log(2d)} \right)$$

with $r_i = O(e^2 ih\beta + \log d)$, due to a technical bound on sparsity of the propagator of unadjusted Langevin

- Recall the recursive bound

$$W_{2,\ell^\infty}(\rho_{(k+N)h}, \pi) \leq \exp(-\alpha N h) \sqrt{d} W_{2,\ell^\infty}(\rho_{kh}, \pi) + \text{error}(N)$$

- Using $s_k = O((k+1)^n)$ and taking $N = \lceil \frac{\log(2\sqrt{d})}{h\alpha} \rceil$

$$W_{2,\ell^\infty}(\rho_{(k+N)h}, \pi) \leq \frac{1}{2} W_{2,\ell^\infty}(\rho_{kh}, \pi) + \sqrt{h \log(2d)} \left(O\left(\frac{\beta}{\alpha} \log(2d) \right) \right)^{\frac{n}{2}+1}$$

- Finally $W_{2,\ell^\infty}(\pi_h, \pi) \leq \sqrt{h \log(2d)} \left(O\left(\frac{\beta}{\alpha} \log(2d) \right) \right)^{\frac{n}{2}+1}$

Unadjusted Langevin bias: asymptotic perspective using PDEs

Bias of observables: asymptotic expansion

Assume f is sufficiently regular and $\int f\pi = 0$. Then, it holds

$$\int f\pi - \int f\pi_h = -\frac{1}{4}h\left(\int (\Delta f + f\Delta \log \pi)\pi\right) + o(h)$$

- Obtained by comparing the generators of π and π_h

$$\mathcal{L}u(x) = \nabla \log \pi(x) \cdot \nabla u(x) + \Delta u(x)$$

$$\mathcal{L}_h u(x) = \frac{1}{h}(\mathbb{E}[u(x + h\nabla \log \pi(x) + \sqrt{2h}\xi)] - u(x))$$

- For Gaussian π, $\int f(\Delta \log \pi)\pi = 0$. The first order term $\int \pi \Delta f$ only depends on the coordinates that f takes
- Delocalization of observable bias: hold for perturbation of Gaussians too, up to $o(h)$

Poisson argument [Mattingly, Stuart, Tretyakov 2010]. Related discussion on averaged observables [Bou-Rabee, Schuh 2023], [Durmus, Eberle 2024]

Probabilistic forecasting (benchmarking Navier-Stokes)

Probabilistic forecasting through generative modeling

A benchmark case study: 2d NSE with stochastic forcing

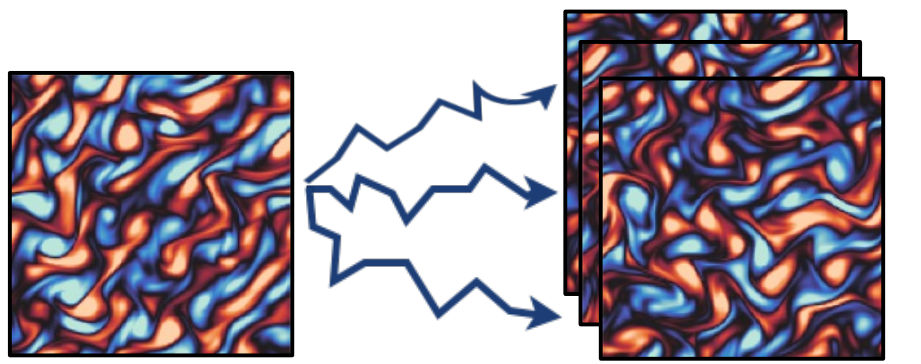
$$\mathrm{d}\omega + v \cdot \nabla\omega \mathrm{d}t = \nu\Delta\omega \mathrm{d}t - \alpha\omega \mathrm{d}t + \epsilon \mathrm{d}\eta \quad \text{on } \mathbb{T}^2$$

- vorticity ω, velocity v, and $\mathrm{d}\eta$ is white-in-time random forcing

  Ergodicity: [Hairer, Mattingly, 2006]

Set-up: given data pairs $(\omega_t, \omega_{t+\tau})$ at many t under stationarity

Task: build a generative model that takes a state ω_t as input and samples the conditional distribution $\rho^\star(\cdot|\omega_t)$ of $\omega_{t+\tau}|\omega_t$



x_0 $\qquad$ $x_1 \sim \rho^\star(\cdot|x_0)$

where we use $x_0 = \omega_t$ and $x_1 = \omega_{t+\tau}$ in the notation

Goal: Build a generative dynamics $X_{0\leq s\leq 1}$ from x_0 to $x_1 \sim \rho^\star(\cdot|x_0)$

[**Chen**, Goldstein, Hua, Albergo, Boffi, Vanden-Eijnden 2024]

Methodology: Construct the stochastic process

$$I_s = \alpha_s x_0 + \beta_s x_1 + \sigma_s W_s$$

- $\alpha_0 = \beta_1 = 1$ and $\alpha_1 = \beta_0 = \sigma_1 = 0$ so that $I_0 = x_0, I_1 = x_1$
- W is a Brownian motion with $W \perp (x_0, x_1)$

Define $b_s(x, x_0) = \mathbb{E}[\dot\alpha_s x_0 + \dot\beta_s x_1 + \dot\sigma_s W_s | I_s = x, x_0]$ and

$$\mathrm{d}X_s = b_s(X_s, x_0)\mathrm{d}s + \sigma_s \mathrm{d}W_s,\ X_{s=0} = x_0$$

It holds $\mathrm{Law}(X_s) = \mathrm{Law}(I_s|x_0)$. In particular $X_{s=1} \sim \rho^\star(\cdot|x_0)$

- Why? Itô's formula: $\mathrm{d}I_s = (\dot\alpha_s x_0 + \dot\beta_s x_1 + \dot\sigma_s W_s)\mathrm{d}s + \sigma_s \mathrm{d}W_s$
- Replacing drift by $\mathbb{E}[\dot\alpha_s x_0 + \dot\beta_s x_1 + \dot\sigma_s W_s | I_s, x_0]$ makes the SDE Markovian while keeping time-marginals unchanged

  Mimicking lemma, Markov projection [Gyöngy 1986]

Learning the generative dynamics from data

The drift $b_s(x, x_0) = \mathbb{E}[\dot{\alpha}_s x_0 + \dot{\beta}_s x_1 + \dot{\sigma}_s W_s | I_s = x, x_0]$

- **Fact:** the drift $b_s(x, x_0)$ is the unique minimizer of

$$L_b[\hat{b}_s] = \int_0^1 \mathbb{E}\big[|\hat{b}_s(I_s, x_0) - \dot{\alpha}_s x_0 - \dot{\beta}_s x_1 - \dot{\sigma}_s W_s|^2\big]\mathrm{d}s$$

  with sampled data (x_0, x_1) we can evaluate L_b

- **Algorithm:** parametrize $\hat{b}_s$ by neural nets, optimize L_b

- **Generative model:** for any x_0, integrate to $s = 1$ the SDE

$$\mathrm{d}\hat{X}_s = \hat{b}_s(\hat{X}_s, x_0)\mathrm{d}s + \sigma_s \mathrm{d}W_s, \ \hat{X}_{s=0} = x_0$$

  This will approximately sample $\rho^\star(\cdot|x_0)$ if $\hat{b}_s \approx b_s$

Experiments: Forecasting 2D stochastically forced NSE

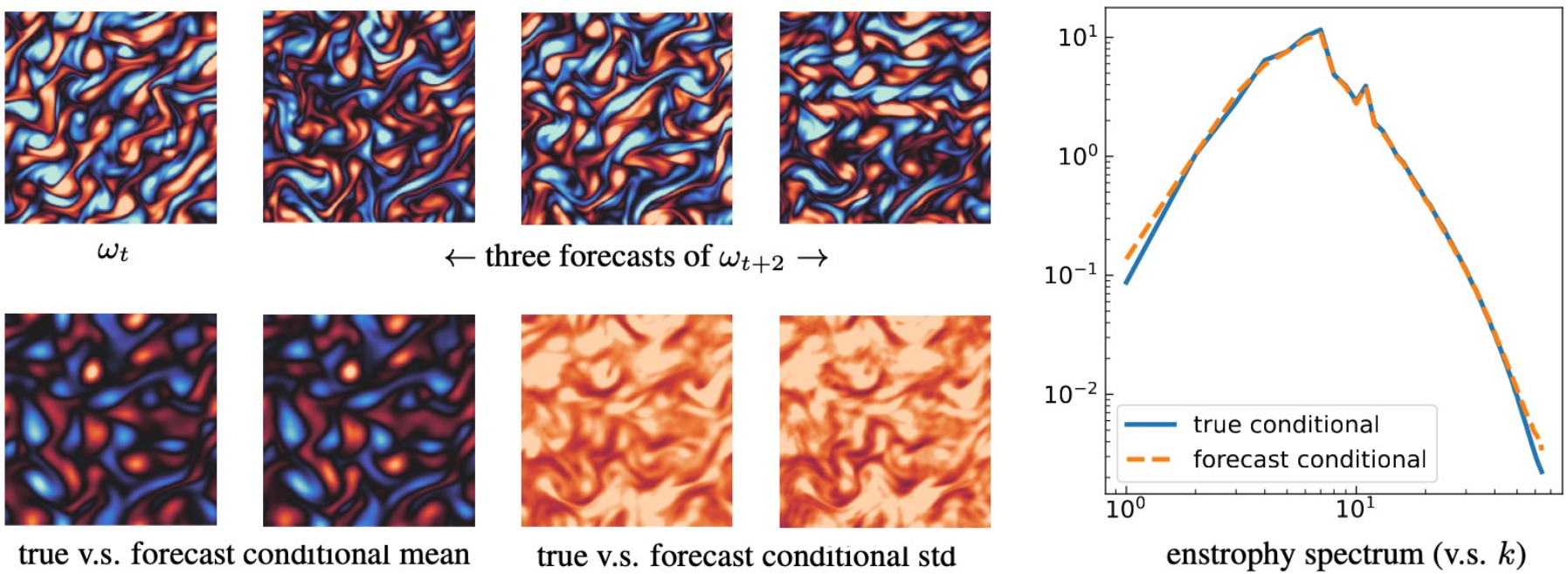


Figure: Lag $\tau = 2$ (autocorrelation 10%). Resolution 128×128, using $200K$ data pairs for training 2M-parameter-Unet

- As a surrogate model: for this example 100 times faster than running the stochastic PDE simulation

Experiments: Forecasting and superresolution

Let ω_t be of 32×32 while $\omega_{t+\tau}$ is of 128×128

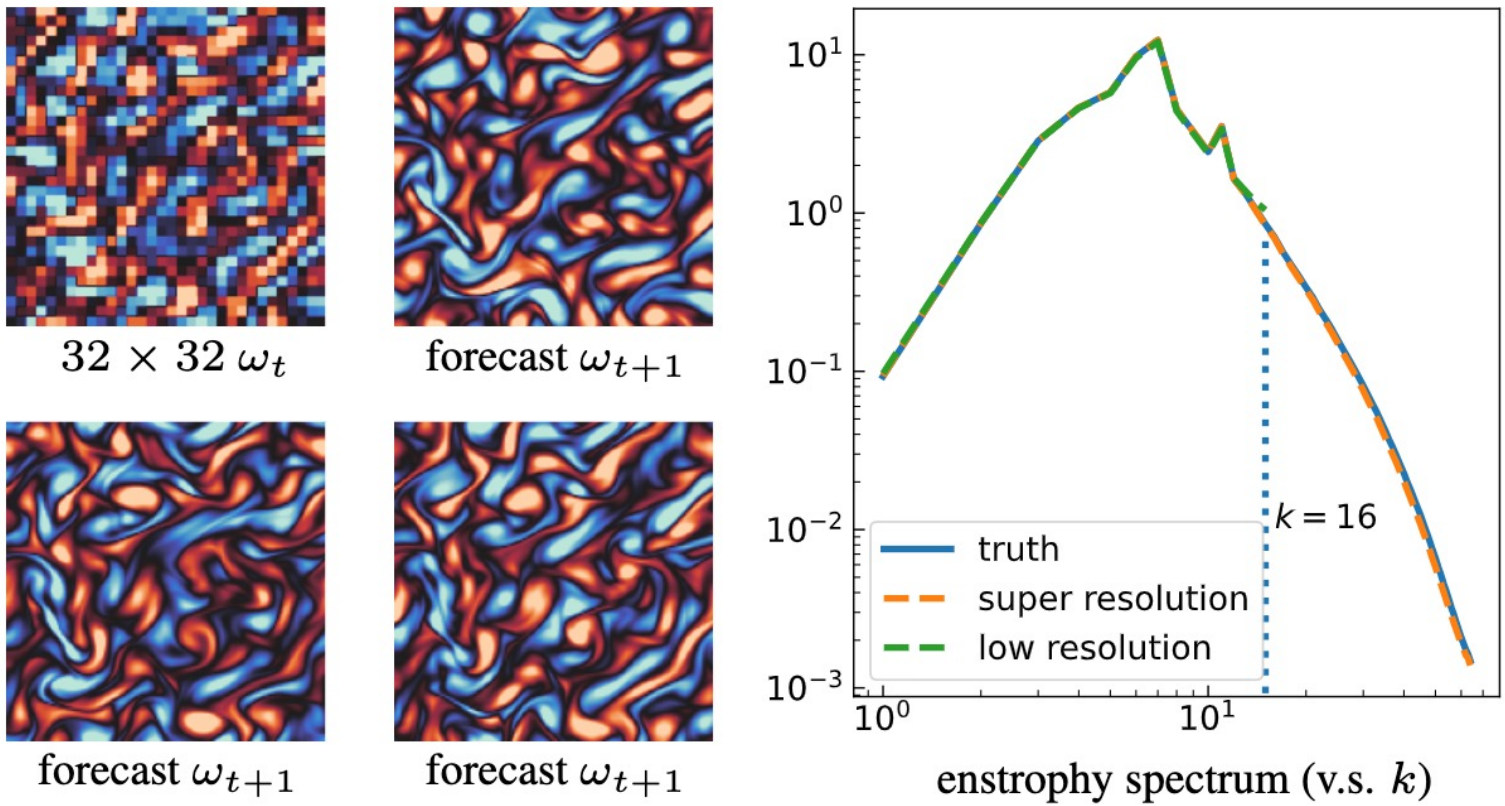


Figure: Probabilistic forecasting with low resolution input, using $200K$ data pairs for training 2M-parameter-Unet

A family of SDEs can be used. Which to choose?

Fact: It holds that Law(X_s) = Law(X_s^g) for

$$\mathrm{d}X_s^g = b_s^g(X_s^g, x_0)\mathrm{d}s + g_s \mathrm{d}W_s$$

with $b_s^g(x, x_0) = b_s(x, x_0) + \frac{1}{2}(g_s^2 - \sigma_s^2)\nabla \log \rho_s(x|x_0)$

- Fact due to Fokker-Planck equations and $\nabla \cdot (\rho \nabla \log \rho) = \Delta \rho$
- $\nabla \log \rho_s(x|x_0)$ is the score, with $\widehat{\text{score}}$ an estimator

New "learned" drift: $\hat{b}_s^g = \hat{b}_s + \frac{1}{2}(g_s^2 - \sigma_s^2)\widehat{\text{score}}$

A family of SDEs can be used. Which to choose?

Fact: It holds that Law(X_s) = Law(X_s^g) for

$$\mathrm{d}X_s^g = b_s^g(X_s^g, x_0)\mathrm{d}s + g_s \mathrm{d}W_s$$

with $b_s^g(x, x_0) = b_s(x, x_0) + \frac{1}{2}(g_s^2 - \sigma_s^2)\nabla \log \rho_s(x|x_0)$

- Fact due to Fokker-Planck equations and $\nabla \cdot (\rho \nabla \log \rho) = \Delta \rho$
- $\nabla \log \rho_s(x|x_0)$ is the score, with $\widehat{\text{score}}$ an estimator

New "learned" drift: $\hat{b}_s^g = \hat{b}_s + \frac{1}{2}(g_s^2 - \sigma_s^2)\widehat{\text{score}}$

Many existing studies on how to choose g in generative models

- ODEs versus SDEs, numerical schemes, perturbation analysis

  [Song et al 2021], [Song, Meng, Ermon 2021], [Karras, Aittala, Aila, Laine 2022], [Zhang, Tao, Chen 2023], [Albergo, Boffi, Vanden-Eijnden 2023], [Cao, Chen, Luo, Zhou 2024]

Answer to this question would depend on the choice of "metric"

KL divergence over path measures as the "metric": theory and practice

Theorem: Let $\mathbb{P}^{X^g}$ and $\mathbb{P}^{\hat{X}^g}$ denote the path measures of

- the truth SDE solution $X^g = (X_s^g)_{s\in[0,1]}$ with drift b^g
- the approximation $\hat{X}^g = (\hat{X}_s^g)_{s\in[0,1]}$ with learned $\hat{b}^g$

Then, the path-level KL optimization

$$\min_g \mathrm{KL}[\mathbb{P}^{X^g} \| \mathbb{P}^{\hat{X}^g}]$$

has an explicit solution $g = g^{\mathsf{F}}$ with

$$g_s^{\mathsf{F}} = \left| 2s\sigma_s^2 \frac{\mathrm{d}}{\mathrm{d}s} \log \frac{\beta_s}{\sqrt{s}\sigma_s} \right|^{1/2}$$

Interpretation: $\frac{\beta_s}{\sqrt{s}\sigma_s}$ is

$\sim$ "signal-to-noise ratio"

since by definition

$$I_s = \alpha_s x_0 + \beta_s x_1 + \sigma_s W_s$$

KL divergence over path measures as the "metric": theory and practice

Theorem: Let $\mathbb{P}^{X^g}$ and $\mathbb{P}^{\hat{X}^g}$ denote the path measures of

- the truth SDE solution $X^g = (X_s^g)_{s\in[0,1]}$ with drift b^g
- the approximation $\hat{X}^g = (\hat{X}_s^g)_{s\in[0,1]}$ with learned $\hat{b}^g$

Then, the path-level KL optimization

$$\min_g \mathrm{KL}[\mathbb{P}^{X^g} \| \mathbb{P}^{\hat{X}^g}]$$

has an explicit solution $g = g^{\mathsf{F}}$ with

$$g_s^{\mathsf{F}} = \left| 2s\sigma_s^2 \frac{\mathrm{d}}{\mathrm{d}s} \log \frac{\beta_s}{\sqrt{s}\sigma_s} \right|^{1/2}$$

Interpretation: $\frac{\beta_s}{\sqrt{s}\sigma_s}$ is

$\sim$ "signal-to-noise ratio"

since by definition

$$I_s = \alpha_s x_0 + \beta_s x_1 + \sigma_s W_s$$

SDE with $\sigma_s \mathrm{d}W_s$	SDE with $g_s^{\mathrm{F}} \mathrm{d}W_s$	ODE with Gaussian base
8.49e-3±1.57e-3	2.79e-3±9.19e-4	4.63e-3±9.63e-4

Empirical end-point KL err (total enstrophy of truth v.s. generated samples)

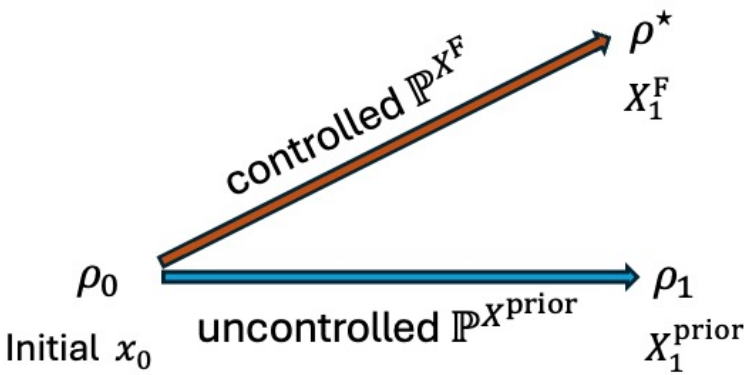
Further insights: What is special about this g^{F}?

Theorem: The optimal $X^{\mathrm{F}} := X^{g^{\mathrm{F}}}$ is an Föllmer process

- Solution to Schrödinger bridge when one endpoint is point mass

$$X^{\mathrm{F}} = \operatorname*{argmin}_{X} \mathrm{KL}[\mathbb{P}^{X} \| \mathbb{P}^{X^{\mathrm{prior}}}] \ \text{ s.t. } \ X_1 \sim \rho^{\star}(\cdot|x_0)$$

Standard Föllmer: X^{prior} is a Brownian motion
In our algorithm: X^{prior} is induced by the choices of $\alpha_s, \beta_s, \sigma_s$




Schrödinger


Föllmer

Interpretation: such optimal g^{F} is a "Bayes"/control solution!

[Schrödinger 1932]. Föllmer process [Föllmer 1986] wide applications in functional inequality [Lehec 2013] and in sampling [Zhang, Chen 2021], [Huang et al 2021], [Vargas et al 2023], etc

Further insights: What is special about this g^{F}?

Theorem: The optimal $X^{\mathrm{F}} := X^{g^{\mathrm{F}}}$ is an Föllmer process

- Solution to Schrödinger bridge when one endpoint is point mass

$$X^{\mathrm{F}} = \operatorname*{argmin}_{X} \mathrm{KL}[\mathbb{P}^{X} \| \mathbb{P}^{X^{\mathrm{prior}}}] \text{ s.t. } X_1 \sim \rho^{\star}(\cdot|x_0)$$

Standard Föllmer: X^{prior} is a Brownian motion
In our algorithm: X^{prior} is induced by the choices of $\alpha_s, \beta_s, \sigma_s$

Outlook: Design physically motivated X^{prior} (ongoing and future work)

- Multiscale interpolation I_s, connected to renormalization group

  e.g., [Bauerschmidt, Bodineau, Dagallier 2023]
- Function space noise with spectrum decay

  e.g., [Lim et al 2023], [Pidstrigach, Marzouk, Reich, and Wang 2023]
- Improved design choices for better numerical performance

  e.g., [Lim, Wang, Yu, Hart, Mahoney, Li, Erichson 2024]

Forecasting videos: CLEVER datasets

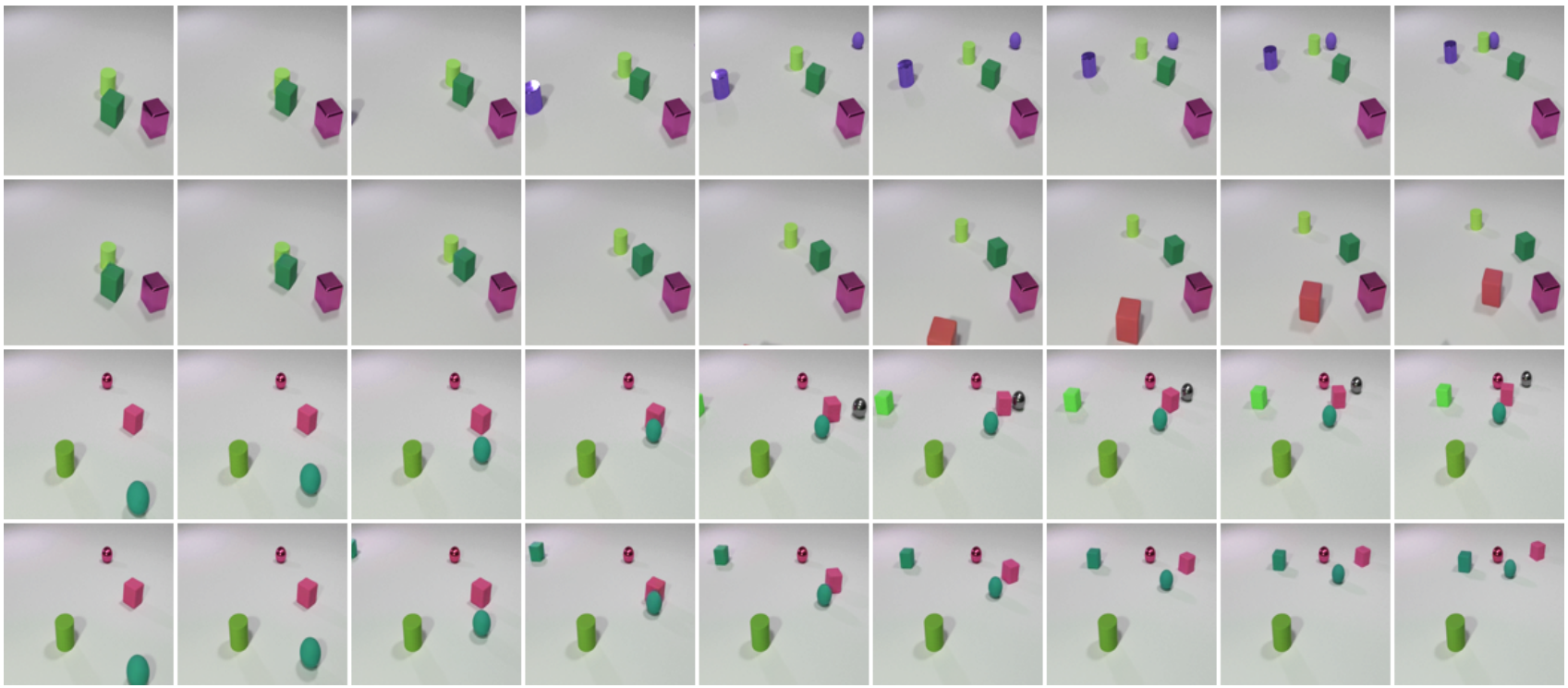


Figure: **Top row:** Real trajectory. **Second row:** Generated trajectory. A new, red cube enters the scene. **Third row:** Real trajectory. **Fourth row:** Generated trajectory. A new green cube enters the scene, and collision physics is respected (green ball hits red cube).

Forecasting videos: quantitative results

	KTH		*CLEVRER*	
Method	100k	250k	100k	250k
RIVER	46.69	41.88	60.40	48.96
PFI (ours)	44.38	39.13	54.7	39.31
Auto-enc.	33.45	33.45	2.79	2.79

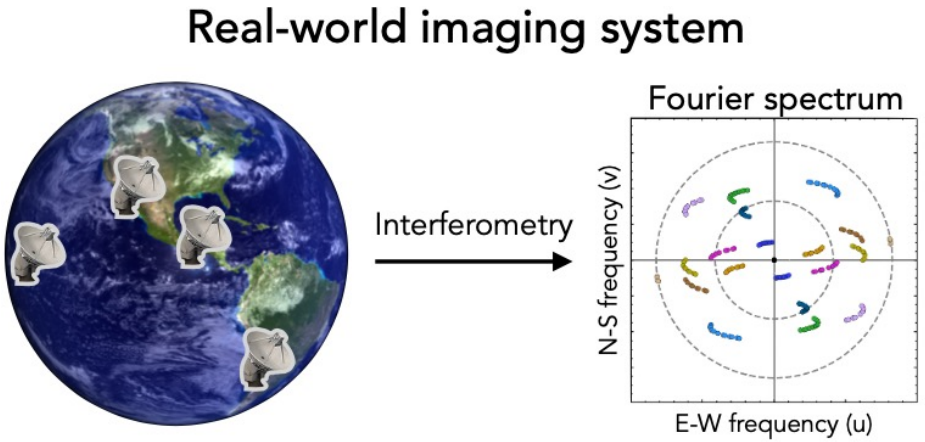
Table: FVD computed on 256 test set videos, with the model generating 100 completions for each video. Results are reported for 100k grad steps and 250k. The auto-enc represents the FVD of the pretrained encoder-decoder vs the real data. It serves as a bound on the possible model performance, as the modeling is done in the latent space of a pre-trained VQGAN.

RIVER [Davtyan, Sameni, Favaro 2023]

Probabilistic imaging
(real data black hole imaging)

Black hole imaging: Combining generative models and MCMC

[Sun, Wu, **Chen**, Feng, Bouman 2023], [Wu, Sun, **Chen**, Zhang, Yue, Bouman 2024]



As a Bayes inverse problem

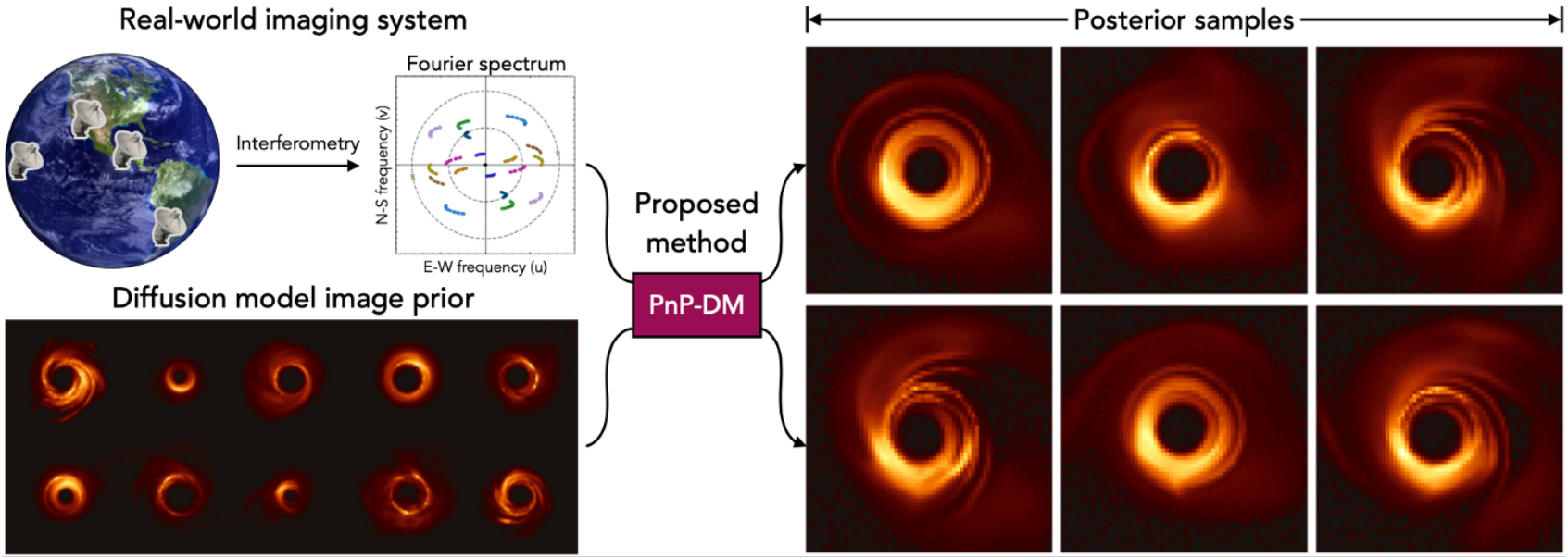
- **Data:** nonlinear functions of Fourier components of the image (very sparse and with strong noise)
- **Prior:** black holes simulated based on General Relativistic Magnetohydrodynamics (GRMHD)

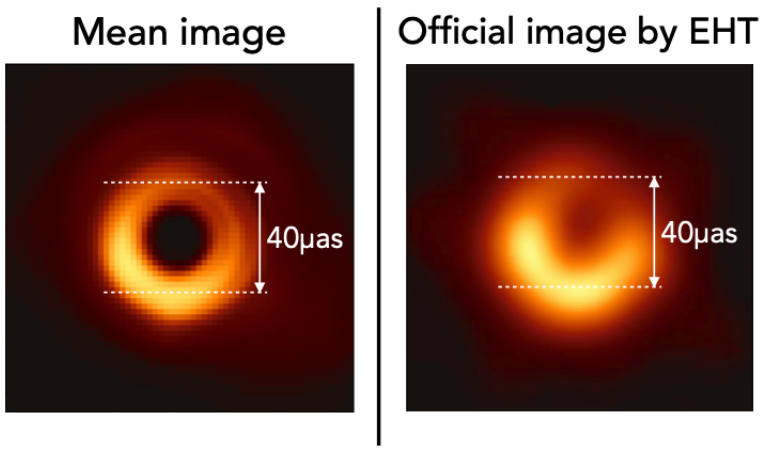
Goal: sample $\rho_{\text{post}} \propto \rho_{\text{prior}} \times L_{\text{likelihood}}$

Approach: learn ρ_{prior} using generative dynamics and combine with designed MCMC dynamics to sample ρ_{post}

Experiments with real data: PnP-DM (plug-and-play diffusion models)

PnP-DM uses split-Gibbs (alternating prior and likelihood update)





** Experiment is performed with real data for the M87 black hole*

Black hole imaging We adopted the same BHI setup as in [59, 61]. The relationship between the black hole image and each interferometric measurement, or so-called *visibility*, is given by

$$V_{a,b}^t = g_a^t g_b^t \cdot e^{-i(\phi_a^t - \phi_b^t)} \cdot \boldsymbol{F}_{a,b}^t(\boldsymbol{x}) + \eta_{a,b} \in \mathbb{C}, \tag{14}$$

where a and b denote a pair of telescopes, t represents the time of measurement acquisition, and $\boldsymbol{F}_{a,b}^t(\boldsymbol{x})$ is the Fourier component of the image $\boldsymbol{x}$ corresponding to the baseline between telescopes a and b at time t. In practice, there are three main sources of noise in (14): gain error g_a and g_b at the telescopes, phase error ϕ_a^t and ϕ_b^t, and baseline-based additive white Gaussian noise $\eta_{a,b}$. The gain and phase errors stem from atmospheric turbulence and instrument miscalibration and often cannot be ignored. To correct for these two errors, multiple noisy visibilities can be combined into data products that are invariant to these errors, which are called *closure phase* and *log closure amplitude* measurements [11]

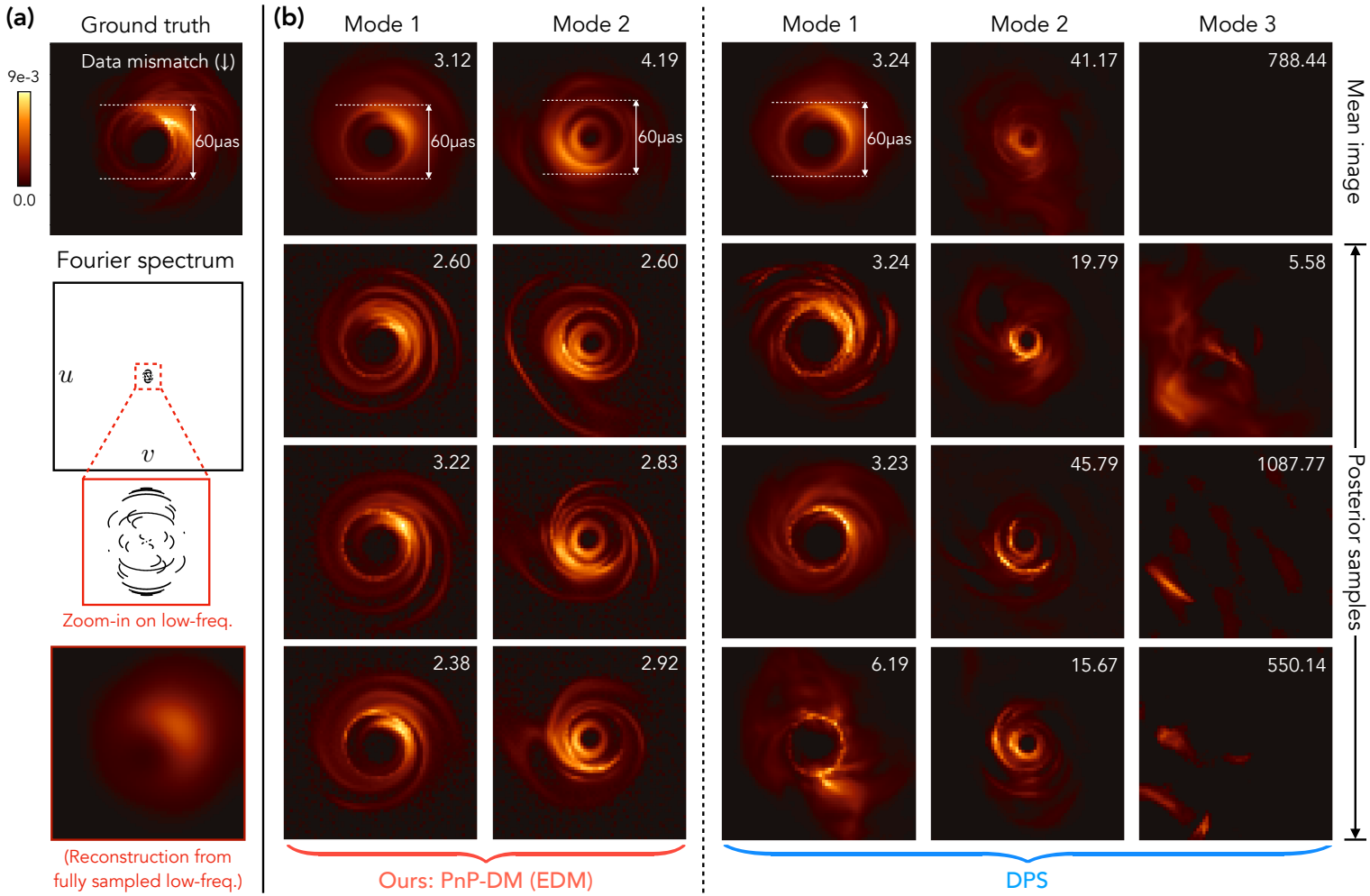
$$\begin{aligned}
\boldsymbol{y}_{t,(a,b,c)}^{\mathsf{cph}} &= \angle(V_{a,b} V_{b,c} V_{a,c}) := \mathcal{A}_{t,(a,b,c)}^{\mathsf{cph}}(\boldsymbol{x}), \\
\boldsymbol{y}_{t,(a,b,c,d)}^{\mathsf{logcamp}} &= \log\left(\frac{|V_{a,b}^t||V_{c,d}^t|}{|V_{a,c}||V_{b,d}^t|}\right) := \mathcal{A}_{t,(a,b,c,d)}^{\mathsf{logcamp}}(\boldsymbol{x}),
\end{aligned}$$

where $\angle$ computes the angle of a complex number. Given a total of M telescopes, there are in total $\frac{(M-1)(M-2)}{2}$ closure phase and $\frac{M(M-3)}{2}$ log closure amplitude measurements at time t, after eliminating repetitive measurements. In our experiments, we used a 9-telescope array ($M = 9$) from the Event Horizon Telescope (EHT) and constructed the data likelihood term based on these nonlinear closure quantities. Additionally, because the closure quantities do not constrain the total flux (i.e. summation of the pixel values) of the underlying black hole image, we added a constraint on the total flux in the likelihood term. The overall potential function of the likelihood is given by

$$f(\boldsymbol{x}; \boldsymbol{y}) = \sum_{t,\mathsf{c}} \frac{\|\mathcal{A}_{t,\mathsf{c}}^{\mathsf{cph}}(\boldsymbol{x}) - \boldsymbol{y}_{t,\mathsf{c}}^{\mathsf{cph}}\|_2^2}{2\sigma_{\mathsf{cph}}^2} + \sum_{t,\mathsf{d}} \frac{\|\mathcal{A}_{t,\mathsf{d}}^{\mathsf{logcamp}}(\boldsymbol{x}) - \boldsymbol{y}_{t,\mathsf{d}}^{\mathsf{logcamp}}\|_2^2}{2\sigma_{\mathsf{logcamp}}^2} + \frac{\left\|\sum_i \boldsymbol{x}_i - \boldsymbol{y}^{\mathsf{flux}}\right\|_2^2}{2\sigma_{\mathsf{flux}}^2}. \tag{15}$$

In this equation, $\boldsymbol{y}^{\mathsf{flux}}$ is the total flux of the underlying black hole, which can be accurately measured. We use $\boldsymbol{y} := (\boldsymbol{y}^{\mathsf{cph}}, \boldsymbol{y}^{\mathsf{logcamp}}, \boldsymbol{y}^{\mathsf{flux}})$ to denote all the measurements and c, d as the indices for

Black hole imaging: experiments with two modal synthetic data



- DPS: existing benchmark [Chung et al 2022]
- Ours: PnP-DM (plug-and-play diffusion models) using split Gibbs, with mathematical consistency guarantee