Analysis and design of high dimensional sampling

for scientific computing

Yifan Chen

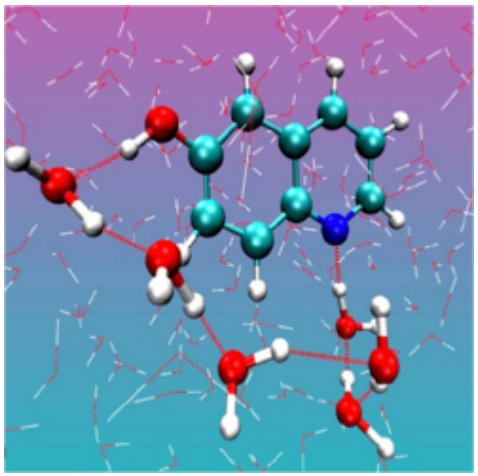
Courant Institute, New York University

January 2025

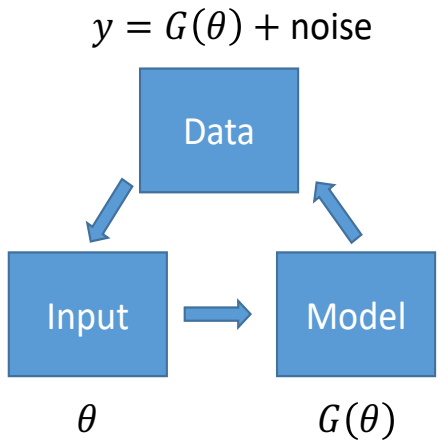
Context

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

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

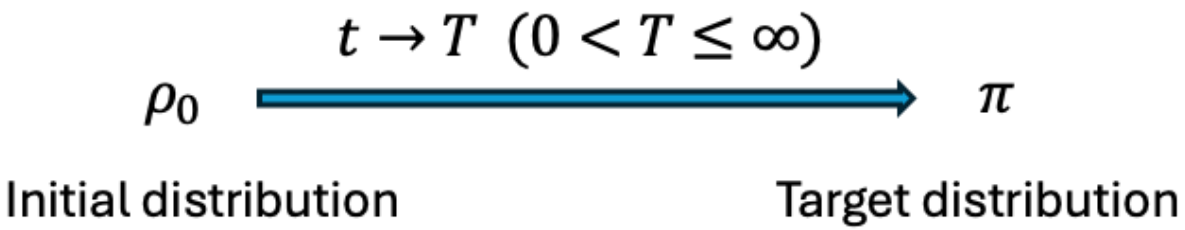
These probability distributions are very high dimensional

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



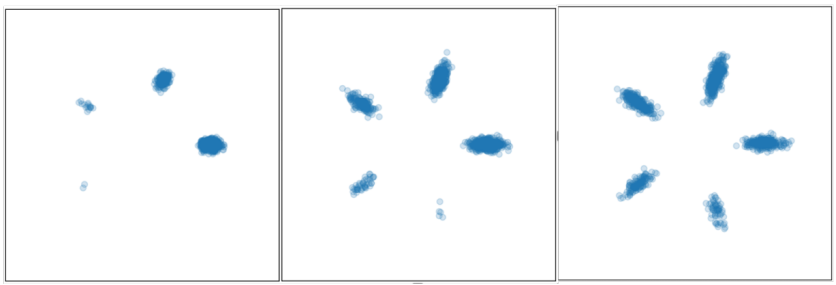
Implementation of ρ_t leads to sampling algorithms

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 (2D illustrations)

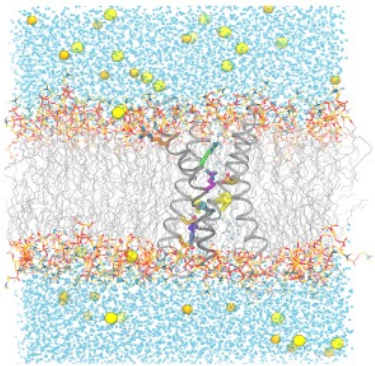
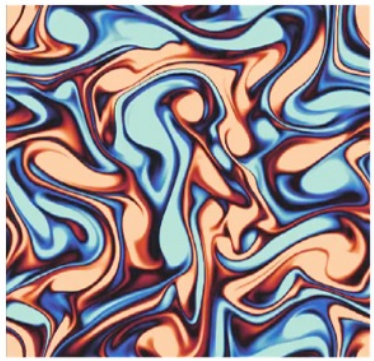
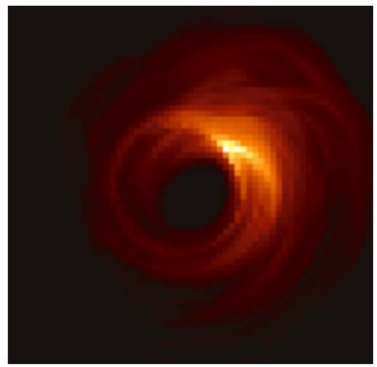
diffusion for images

Guiding questions:

- Why and how can methods work in high dimensions?
- How to design methods for targeted scientific applications?

Outline of the talk

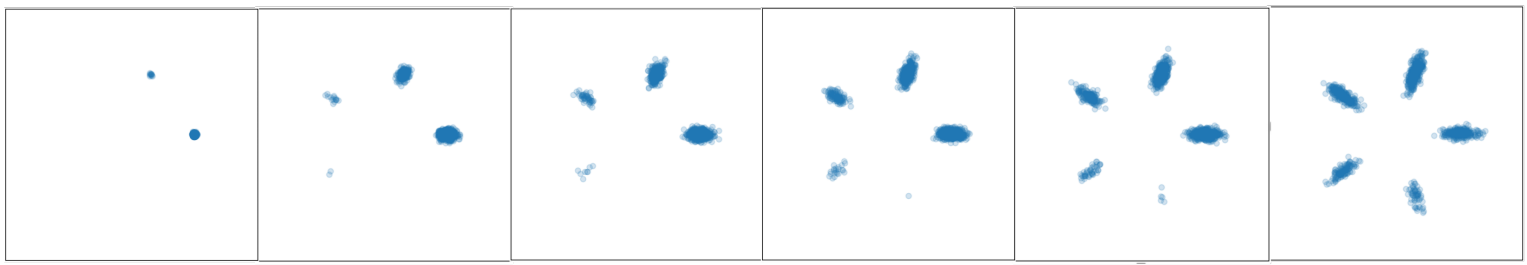
1. **Analysis of unadjusted Langevin in high dimensions**
   (analysis w/ methodological insights)
   - A new "delocalization of bias" phenomenon
   - Inspiration drawn from molecular dynamics simulation

2. **Design and application of generative diffusions**
   (methodology w/ analytical insights)
   - Probabilistic forecasting (benchmarking Navier-Stokes)
   - Probabilistic imaging (real data black hole imaging)

Sampling given queries to the potential 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 $t \to \infty$, $\mathrm{Law}(X_t) \to \pi \propto \exp(-V)$

- **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)

- 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]

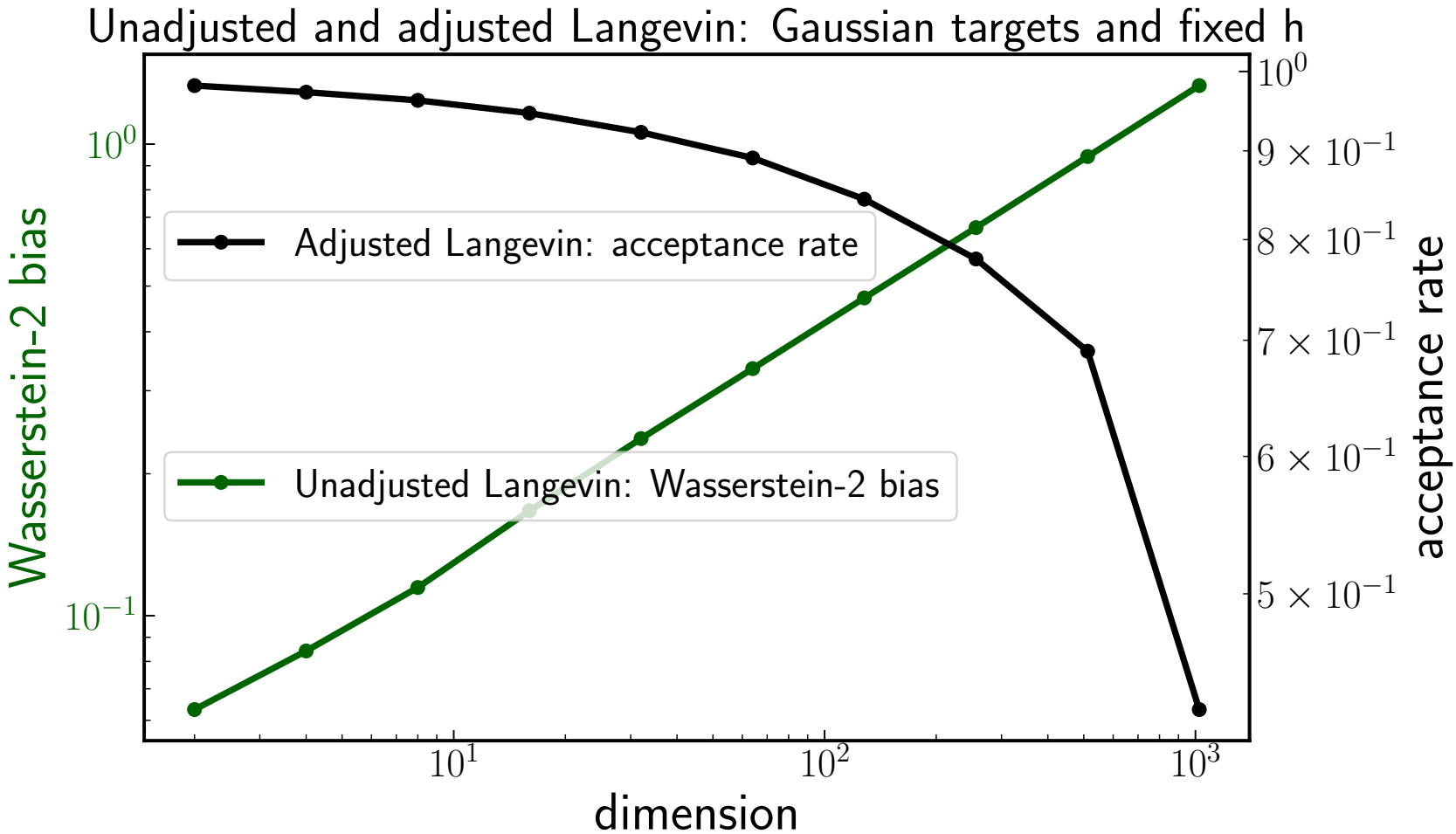
However, for this algorithm, h must be small when d is large

- 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 stepsize h

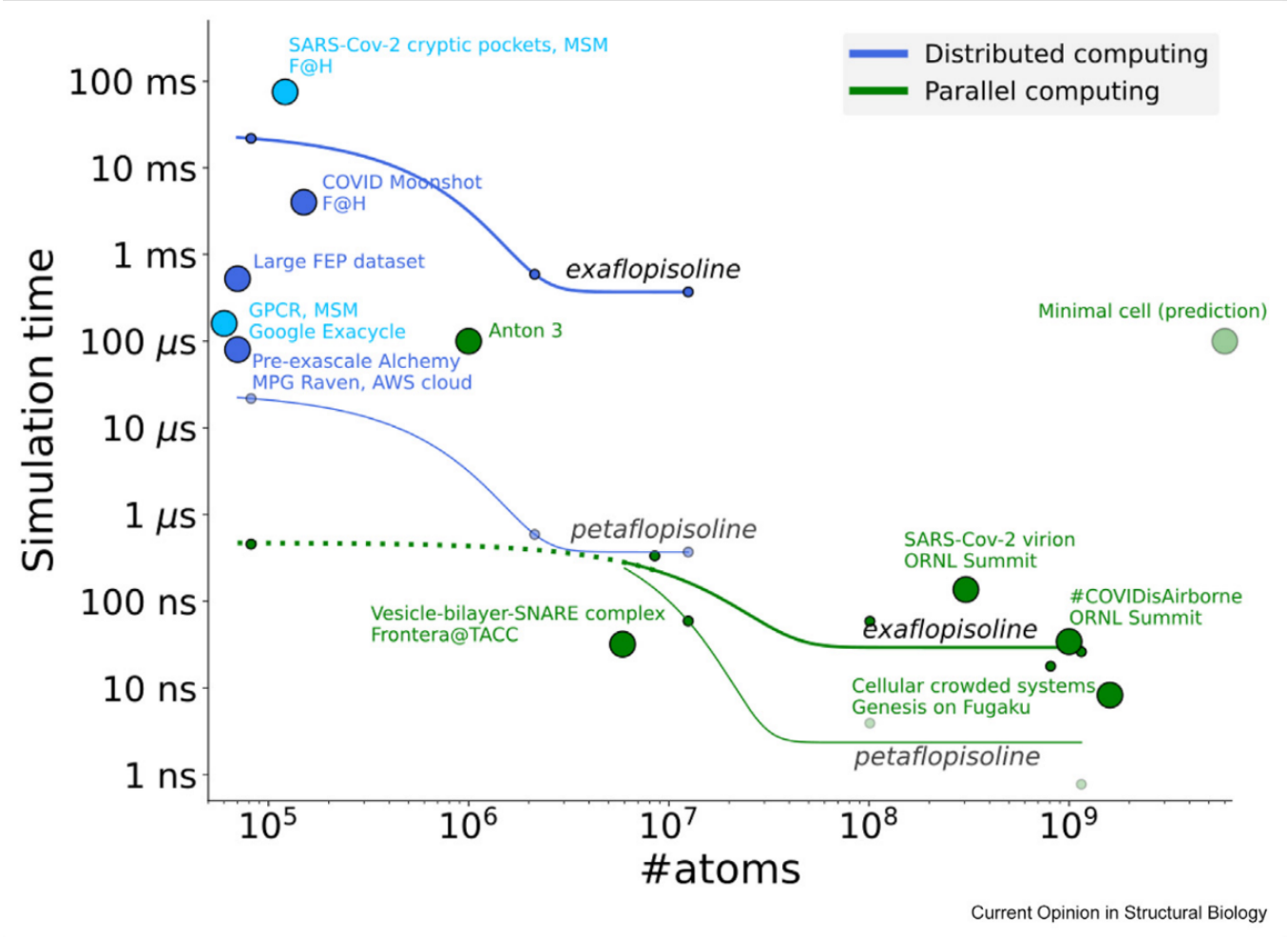


- Fixed h seems to fail when d increases
- Existing theory: $h \sim 1/d^c$ is required

Is this a full story?

Empirical evidence in molecular dynamics

Variants of unadjusted Langevin routinely applied in high dims

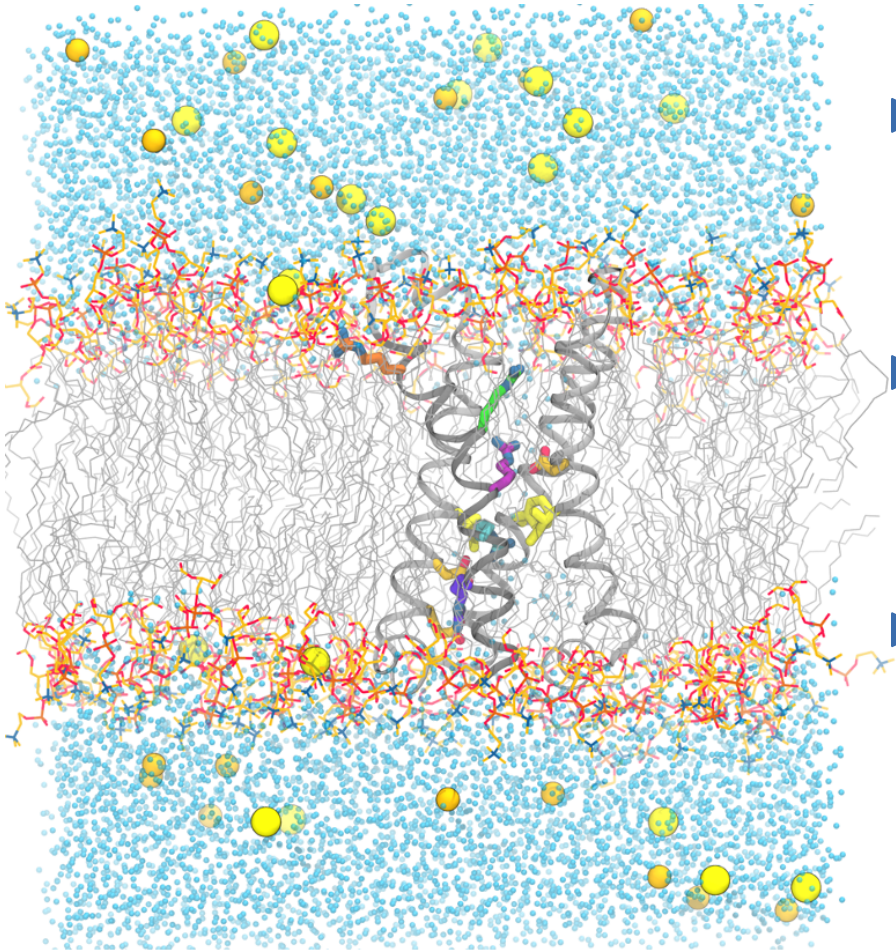


[Gapsys, Kopec, Matthes, de Groot 2024]

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

What could be the catch?

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



[Thanks to Spencer Guo]

Molecular dynamics (MD) example

- We care about averages with respect to a few atoms in the voltage sensing protein in the middle
- We 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

"What we finally measure" matters?

Disclaimer: the potential V in MD is more complicated than our current analysis would cover

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 $V(x) = \sum_{i=1}^{d} V_i(x^{(i)})$ satisfies $\alpha \leq \nabla^2 V_i \leq \beta$. Then it holds that

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

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

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

Consider $\pi \propto \exp(-V)$ where $V(x) = \sum_{i=1}^{d} V_i(x^{(i)})$ satisfies $\alpha \leq \nabla^2 V_i \leq \beta$. 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

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

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

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

Consider $\pi \propto \exp(-V)$ where $V(x) = \sum_{i=1}^d V_i(x^{(i)})$ satisfies $\alpha \leq \nabla^2 V_i \leq \beta$. 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 $\tilde{\pi}_h = Q\#\pi_h$

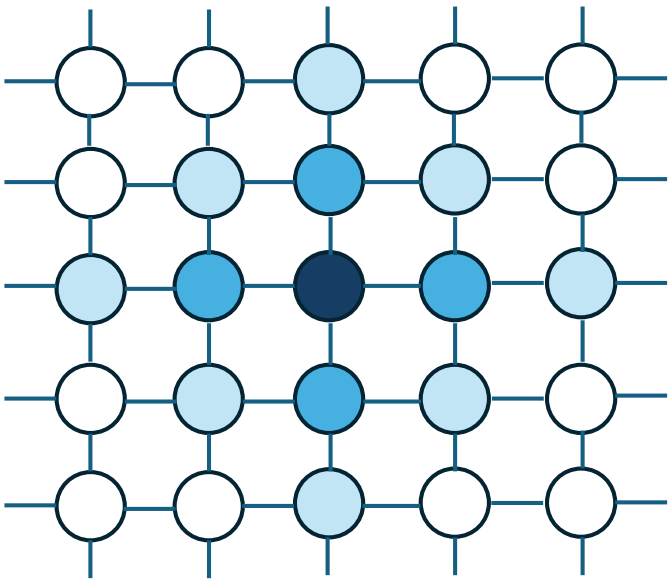
$$\begin{aligned} W_{2,\ell^\infty}(\tilde{\pi}, \tilde{\pi}_h) &\geq W_{1,\ell^\infty}(\tilde{\pi}, \tilde{\pi}_h) \\ &\geq \left| \int x^{(1)} (\tilde{\pi} - \tilde{\pi}_h) \right| \\ &= \left| \int (\frac{1}{\sqrt{d}} \sum_{i=1}^{d} x^{(i)})(\pi - \pi_h) \right| = \sqrt{d}\delta \end{aligned}$$

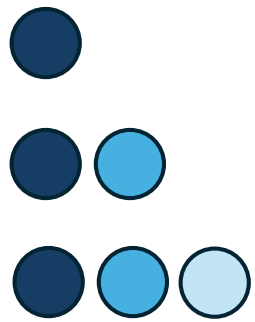
This example exhibits global and strong interactions (due to rotation)

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

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





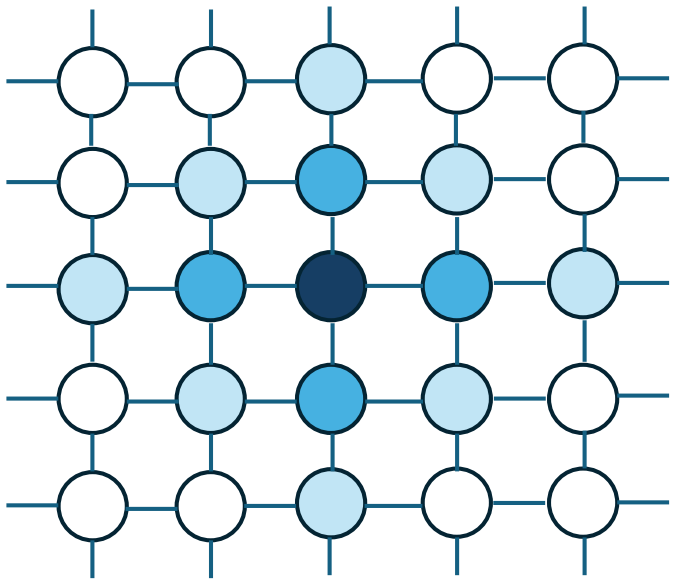
Potential $V(x) = \sum_{i=1}^{d} V_i(x)$
and V_i only depends on $N_1(x^{(i)})$

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

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

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



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

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

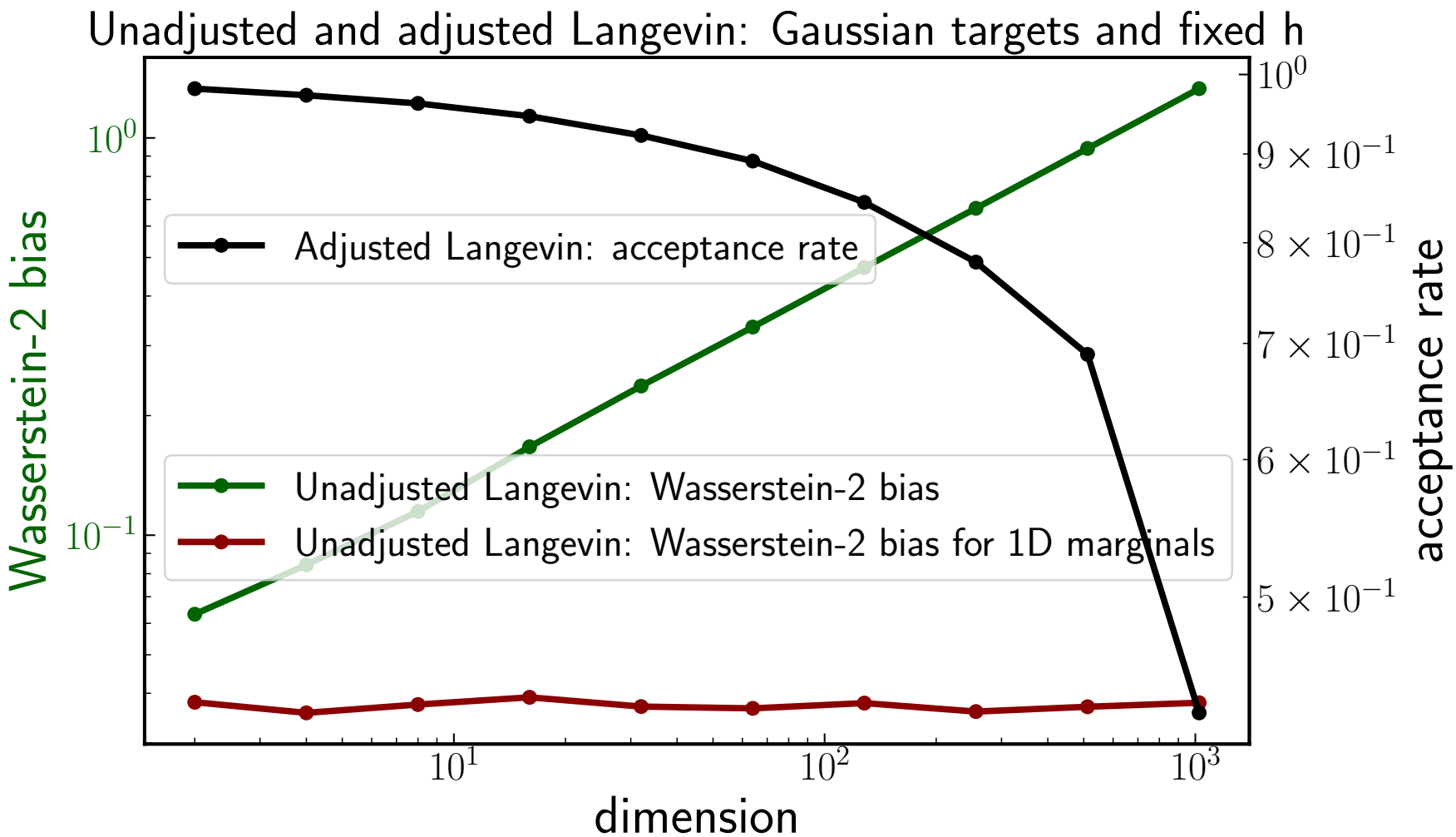
2nd layer: $N_2(x^{(i)})$

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
- Weak global mean field interaction works too (see others in paper)

Updated performance illustration: for fixed stepsize h



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

Take-home messages: delocalization of bias

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

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/local/weak interactions)

Take-home messages: delocalization of bias

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

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/local/weak interactions)

Algorithmic insights (ongoing and future work)

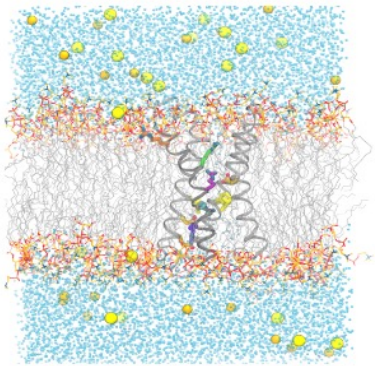
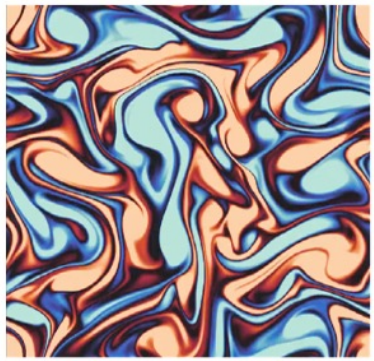
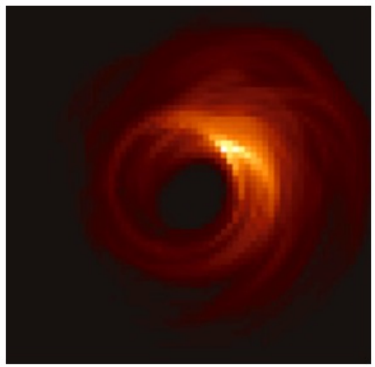
- "Do not Metropolize in very high dims!"
- Development of multilevel unadjusted schemes [Giles 2015], [Ruzayqat, Chada, Jasra 2023], [Chada, Leimkuhler, Paulin, Whalley 2024]

Theoretical outlook (ongoing and future work)

- Non-log-concave measures (e.g., by reflection coupling)
- General approximation of general dynamics: metric is key!
- Relative bias of observables corresponding to rare events

Outline of the talk

1. Analysis of unadjusted Langevin in high dimensions
   (analysis w/ methodological insights)
   - A new “delocalization of bias” phenomenon
   - Inspiration drawn from molecular dynamics simulation

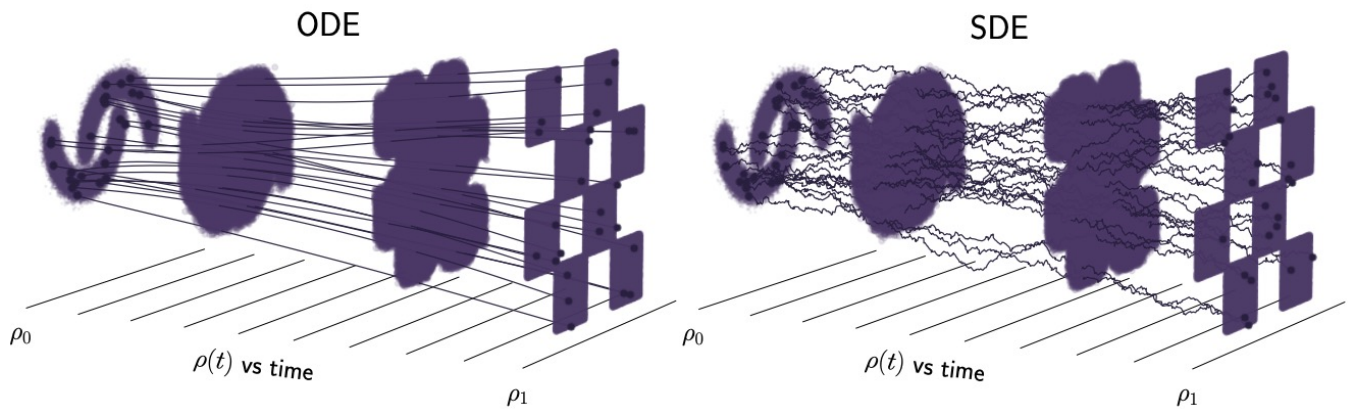




2. Design and application of generative diffusions
   (methodology w/ analytical insights)
   - Probabilistic forecasting (benchmarking Navier-Stokes)
   - Probabilistic imaging (real data black hole imaging)

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

Recent advances in generative modeling driven by building dynamics of measures that can be learned from data efficiently

Diffusion models, score based generative models



Stochastic interpolants, rectified flow, flow matching, bridge matching

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

Probabilistic forecasting through generative modeling

A benchmark case study: 2d NSE with stochastic forcing

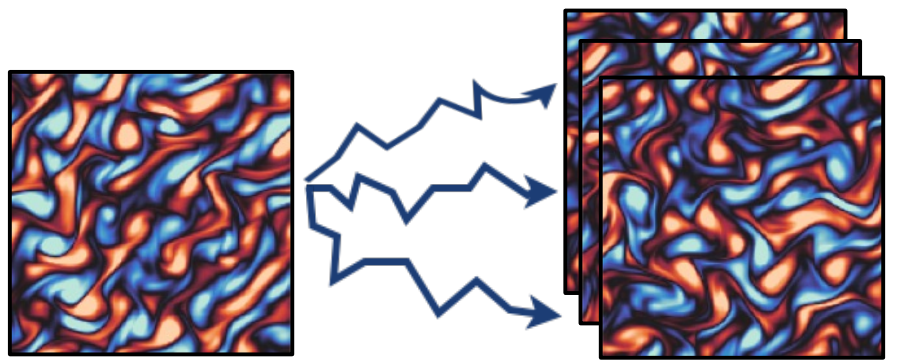
$$\mathrm{d}\omega + v \cdot \nabla\omega \mathrm{d}t = \nu\Delta\omega \mathrm{d}t - \alpha\omega \mathrm{d}t + \epsilon \mathrm{d}\eta \quad \text{on } \mathbb{T}^2$$

- 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

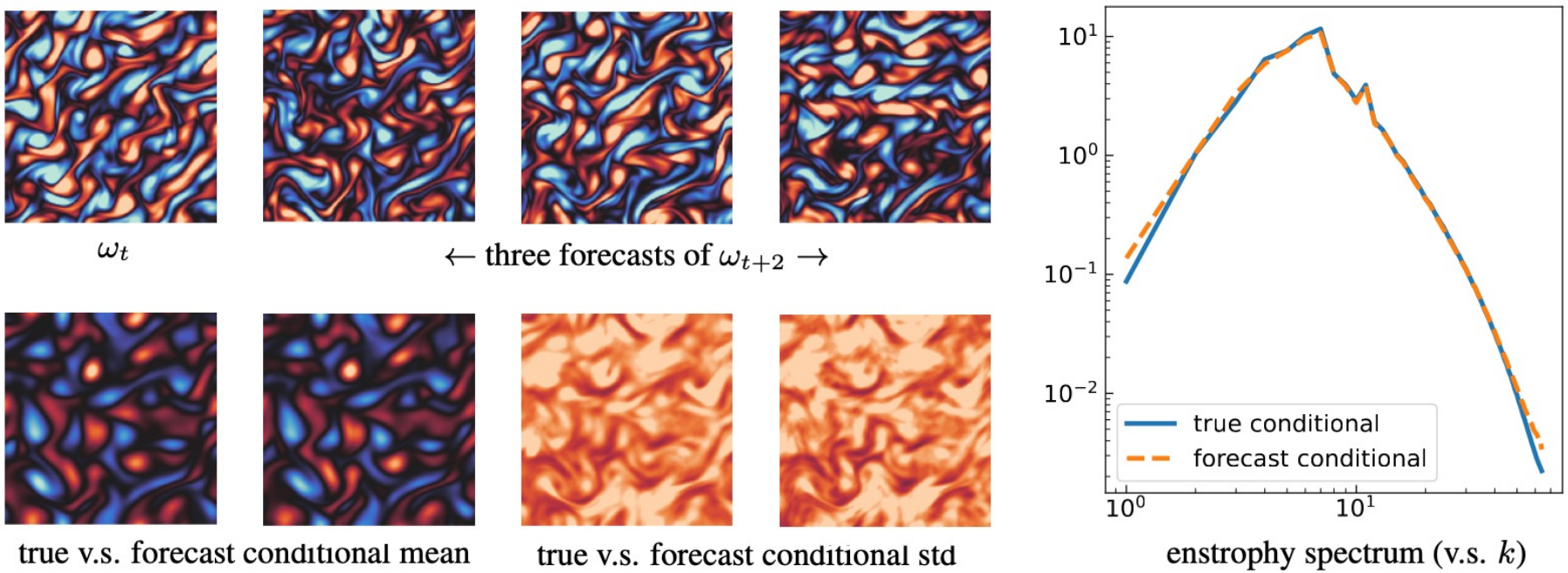


Figure: Lag $\tau = 2$ (autocorrelation 10%). Resolution 128×128, using $200K$ data pairs for training 2M-parameter-Unet

- 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

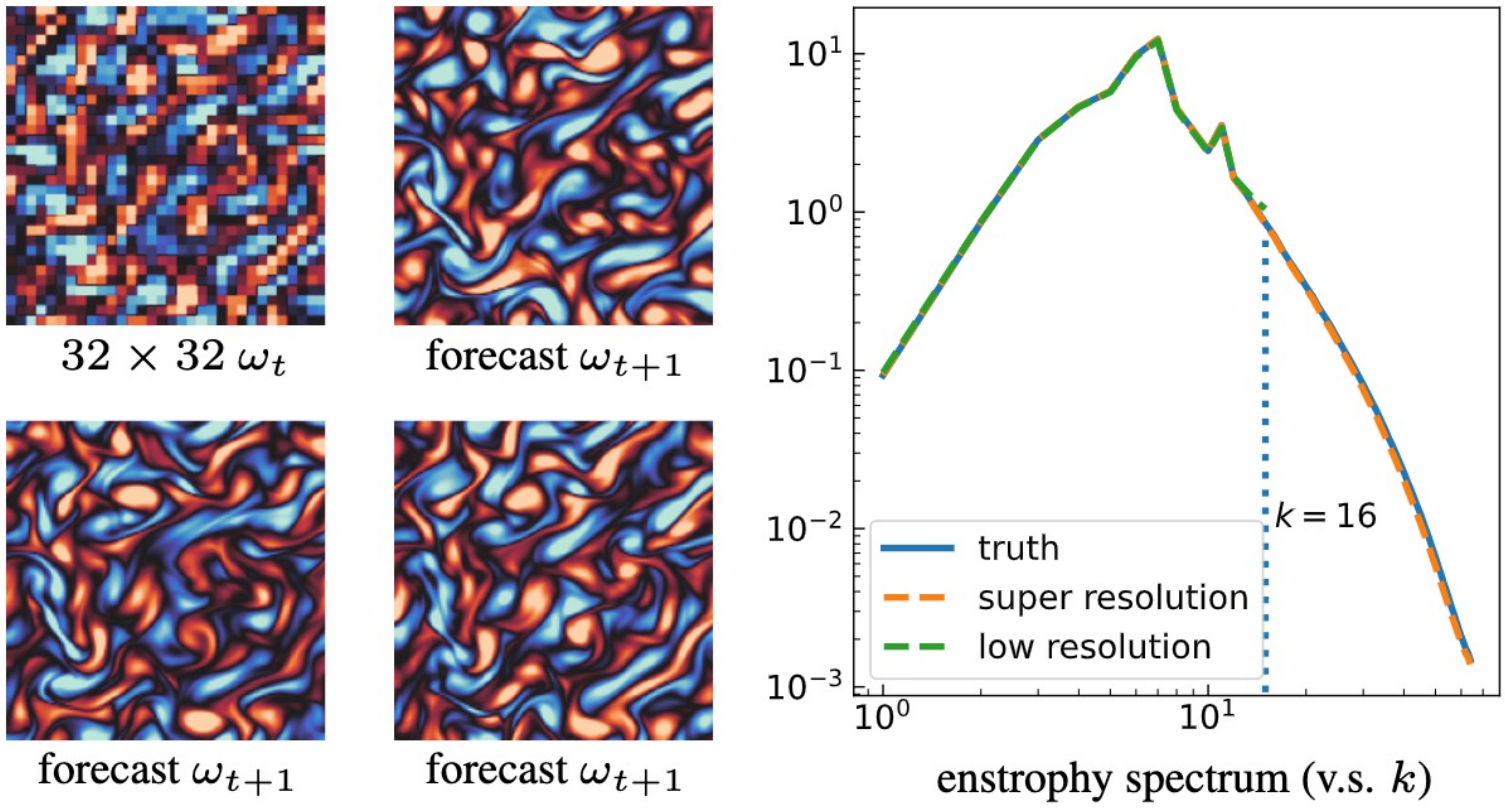


Figure: Probabilistic forecasting with low resolution input, using $200K$ data pairs for training 2M-parameter-Unet

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)

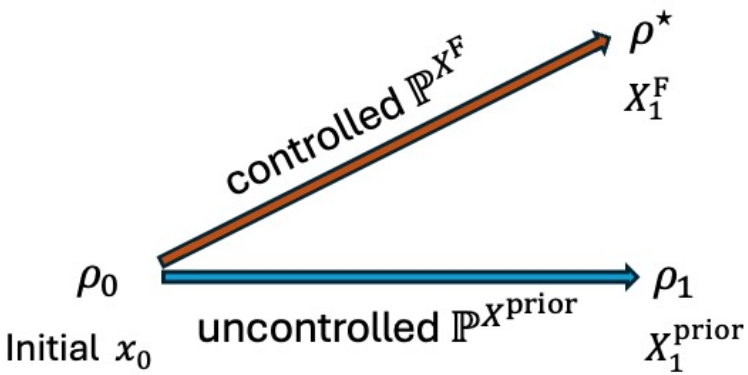
Further insights: What is special about this g^{F}?

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

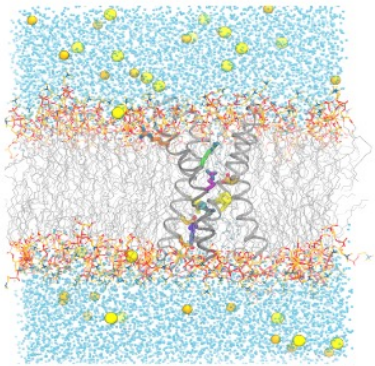
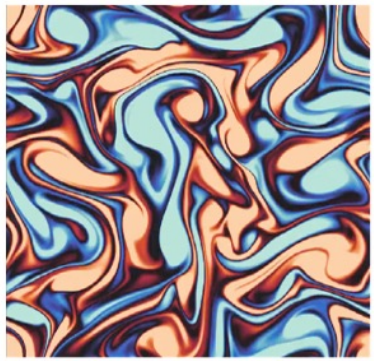
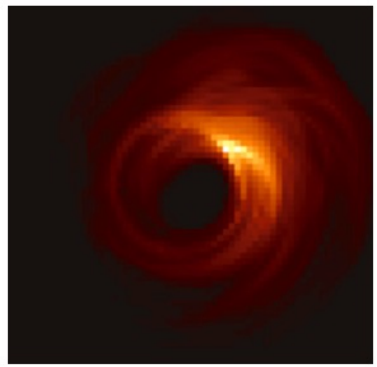
  e.g., [Bauerschmidt, Bodineau, Dagallier 2023]
- Function space noise with spectrum decay

  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]

Outline of the talk

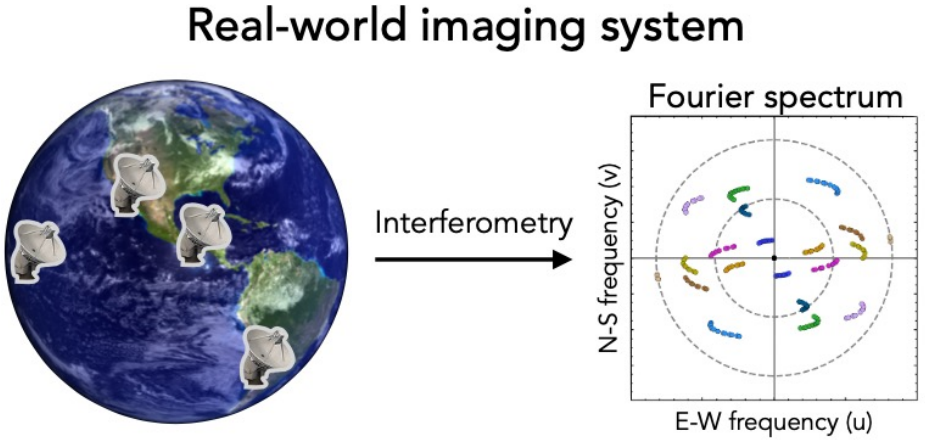
1. Analysis of unadjusted Langevin in high dimensions
   (analysis w/ methodological insights)
   - A new “delocalization of bias” phenomenon
   - Inspiration drawn from molecular dynamics simulation

2. Design and application of generative diffusions
   (methodology w/ analytical insights)
   - Probabilistic forecasting (benchmarking Navier-Stokes)
   - 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

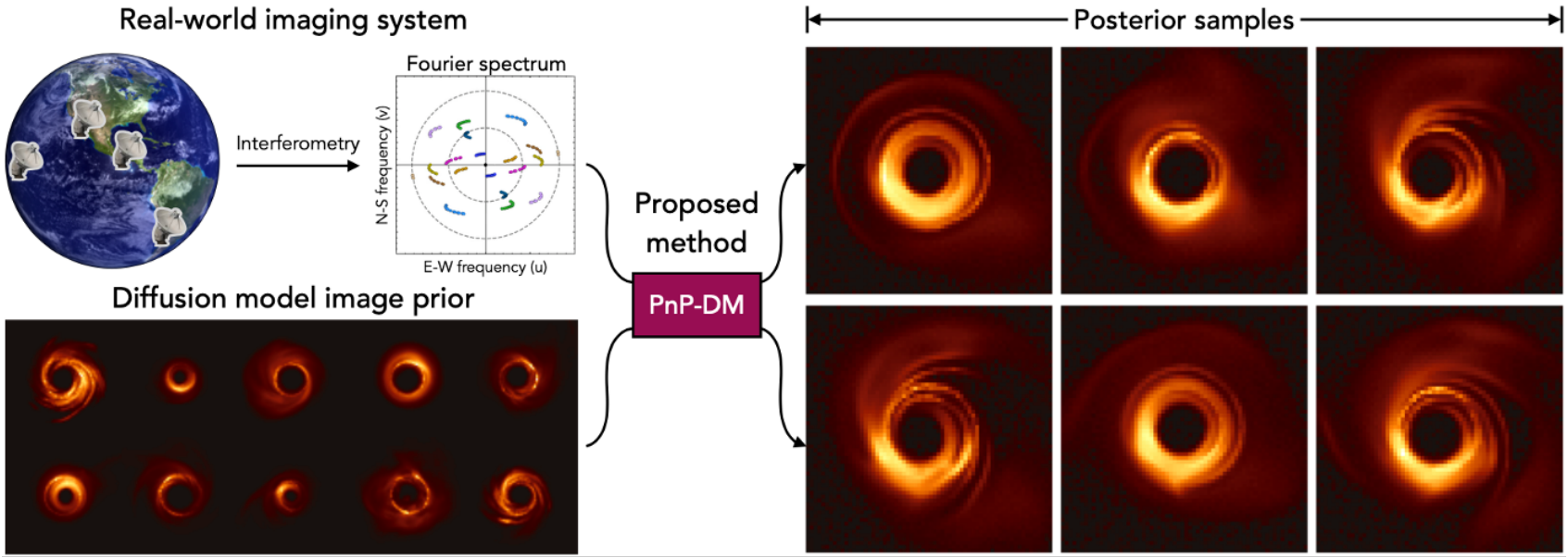
- **Data:** nonlinear functions of Fourier components of the image (very sparse and with strong noise)
- **Prior:** black holes simulated based on General Relativistic Magnetohydrodynamics (GRMHD)

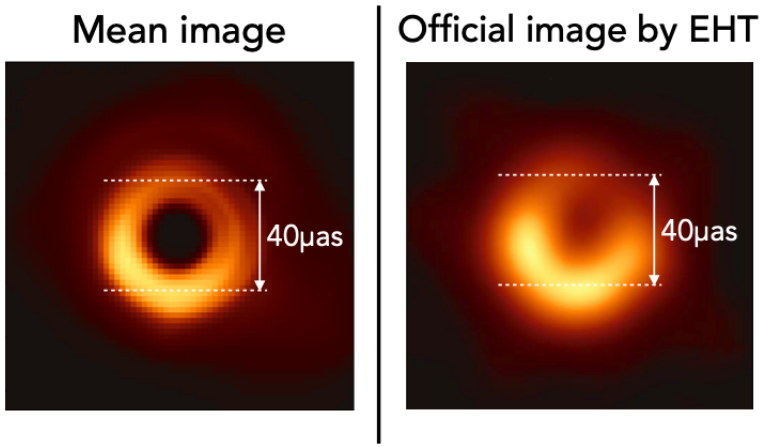
Goal: sample $\rho_{\text{post}} \propto \rho_{\text{prior}} \times L_{\text{likelihood}}$

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*

Thank you!

Back-Up Slides

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]

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]

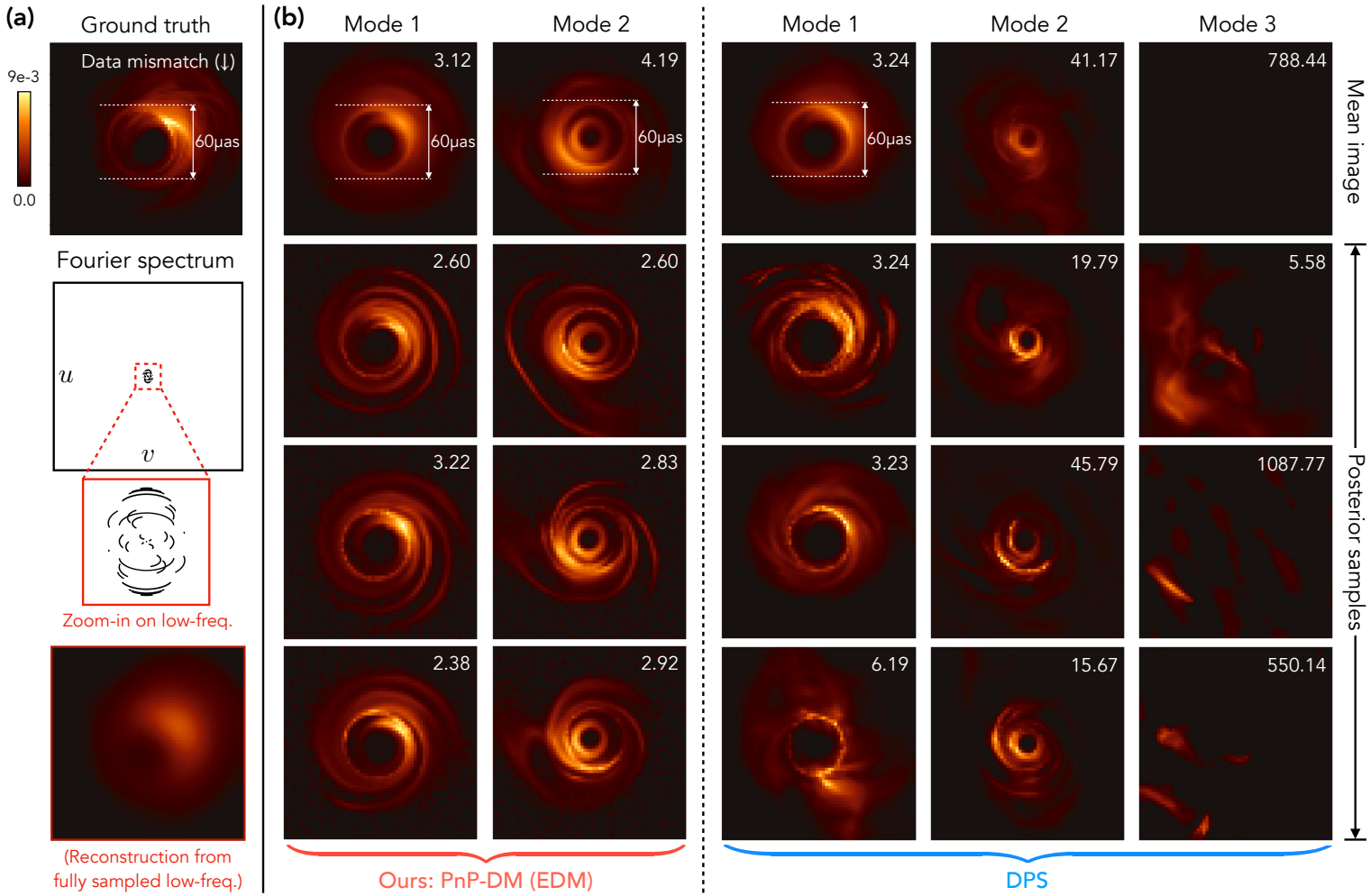
$$\begin{aligned}
\boldsymbol{y}_{t,(a,b,c)}^{\mathsf{cph}} &= \angle(V_{a,b} V_{b,c} V_{a,c}) := \mathcal{A}_{t,(a,b,c)}^{\mathsf{cph}}(\boldsymbol{x}), \\
\boldsymbol{y}_{t,(a,b,c,d)}^{\mathsf{logcamp}} &= \log\left(\frac{|V_{a,b}^t||V_{c,d}^t|}{|V_{a,c}||V_{b,d}^t|}\right) := \mathcal{A}_{t,(a,b,c,d)}^{\mathsf{logcamp}}(\boldsymbol{x}),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

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

Forecasting videos: CLEVER datasets

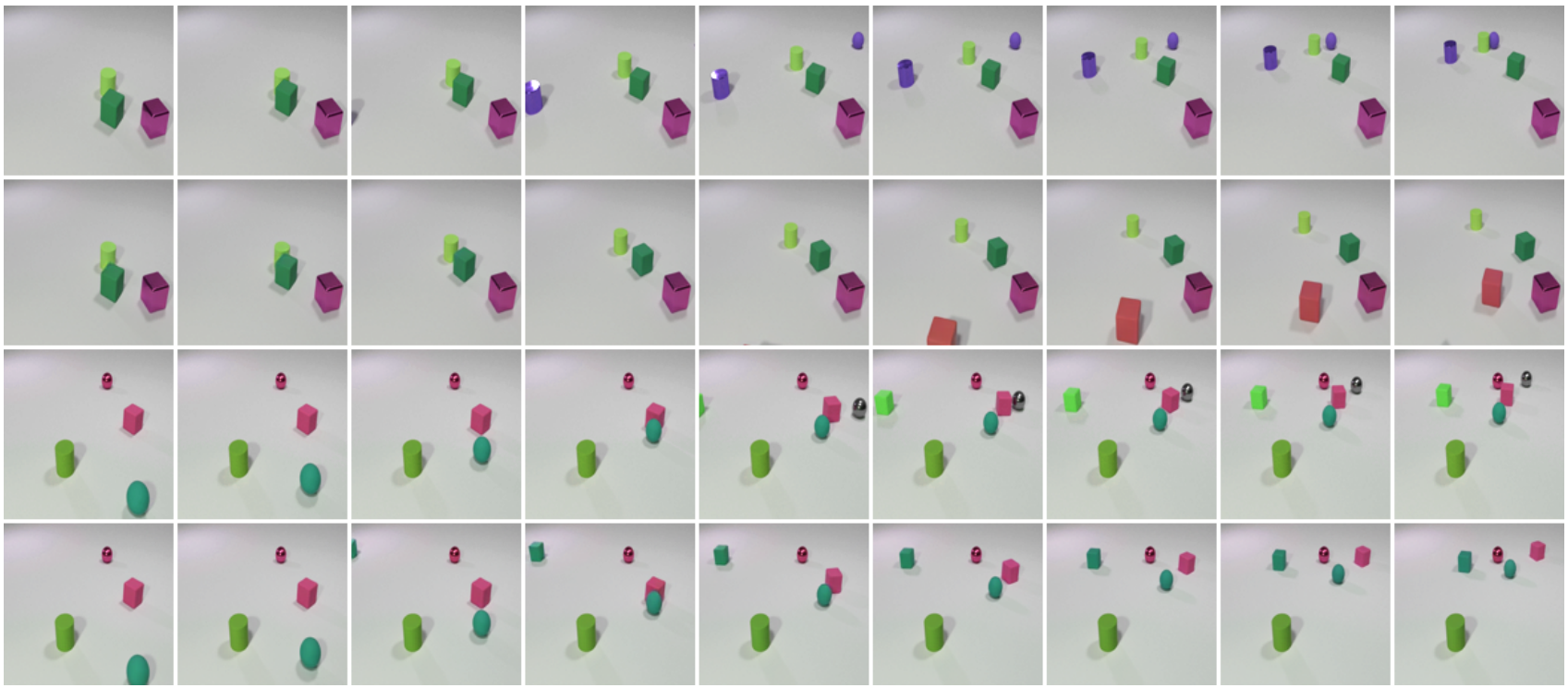


Figure: **Top row:** Real trajectory. **Second row:** Generated trajectory. A new, red cube enters the scene. **Third row:** Real trajectory. **Fourth row:** Generated trajectory. A new green cube enters the scene, and collision physics is respected (green ball hits red cube).

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]

Proof of delocalization of bias: sketch of arguments

Synchronous coupling of Brownian motion

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

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

  coupled with the same B_t

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

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

- Part (b): discretization error = $O(\beta h^{3/2}\sqrt{\log(2d)})$

  (reminiscent of the fact that $\mathbb{E}[|B_t|_\infty^2] \leq t\log(2d)$)

- Part (a):

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

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

- When $\nabla^2 V$ is diagonal, $|H_k|_\infty = |H_k|_2 \le 1 - \alpha h \le \exp(-\alpha h)$ so we get contraction

- In general, H_k is non-diagonal but sparse. We have

$$|H_k|_\infty \le \sqrt{s_1}|H_k|_2 \le \sqrt{s_1}\exp(-\alpha h)$$

  Not a one-step contraction in general

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

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

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

Sketch of arguments: bound discretization errors

How to control $\mathrm{error}(N)$?

- For $N = 1$:

$$
\begin{aligned}
&\mathbb{E}[|\overline{Y}_{(k+1)h} - Y_{(k+1)h}|_\infty^2] \\
=&\mathbb{E}[|\int_{kh}^{(k+1)h} \nabla V(Y_t) - \nabla V(Y_{kh}) \mathrm{d}t|_\infty^2] \\
\leq& h \int_{kh}^{(k+1)h} \mathbb{E}[|\nabla V(Y_t) - \nabla V(Y_{kh})|_\infty^2] \mathrm{d}t \\
\leq& h \int_{kh}^{(k+1)h} \int_0^1 \mathbb{E}[|\nabla^2 V(uY_t + (1-u)Y_{kh})(Y_t - Y_{kh})|_\infty^2] \mathrm{d}u \mathrm{d}t \\
\leq& h s_1 \beta^2 \int_{kh}^{(k+1)h} \mathbb{E}[|Y_t - Y_{kh}|_\infty^2] \mathrm{d}t = h s_1 \beta^2 \cdot O(h^2 \log(2d))
\end{aligned}
$$

Sketch of arguments: bound discretization errors

How to control $\mathrm{error}(N)$?

- For $N = 2$:

$$\begin{aligned}
&\mathbb{E}[|H_k(\overline{Y}_{kh} - Y_{kh})|_\infty^2] \\
\leq & h \int_{(k-1)h}^{kh} \mathbb{E}[|H_k(\nabla V(Y_t) - \nabla V(Y_{(k-1)h}))|_\infty^2] \mathrm{d}t \\
\leq & h \int_{(k-1)h}^{kh} \int_0^1 \mathbb{E}[|H_k(\nabla^2 V(uY_t + (1-u)Y_{(k-1)h}))(Y_t - Y_{(k-1)h})|_\infty^2] \mathrm{d}u \mathrm{d}t
\end{aligned}$$

- Now, how to bound $|H_k(\nabla^2 V(uY_t + (1-u)Y_{(k-1)h}))|_\infty$?
- A simple bound

$$|H_k(\nabla^2 V(uY_t + (1-u)Y_{(k-1)h}))|_\infty \leq \sqrt{s_2}\beta \exp(-\alpha h)$$

- Issue: The bound does take into account sparsity, but the sparsity growth s_2 does not depend on h

Sketch of arguments: sparsity growth bound

Consider the general N-case

- Let $J_N = |H_{k+N-1}H_{k+N-2}\cdots H_k(\nabla^2 V(uY_t + (1-u)Y_{(k-1)h})|_\infty$, then simple bound $|J_N|_\infty \leq \beta\sqrt{s_N}\exp(-\alpha N h)$

  The issue again is that s_N does not depend on h

- Improved bound by using sparsity bound for terms involving small powers of h and using maximum bound for terms involving large powers of h

$$|J_N|_\infty \leq \beta(\sqrt{s_r}\exp(-\alpha N h) + \sqrt{d}\exp(-r))$$

  for any $r \geq e^2 N h \beta$

- In particular, taking $r_N = \lceil e^2 N h \beta + \log\sqrt{d}\rceil$ leads to

$$|J_N|_\infty \leq 2\beta\sqrt{s_{r_N}}\exp(-\alpha N h)$$

  Here r_N scales with physical time Nh

Sketch of arguments: back to discretization errors

Back to the estimate of $\mathrm{error}(N)$

- For $N = 2$:

$$
\begin{aligned}
&\mathbb{E}[|H_k(\overline{Y}_{kh} - Y_{kh})|_\infty^2] \\
\leq & h \int_{(k-1)h}^{kh} \mathbb{E}[|H_k(\nabla V(Y_t) - \nabla V(Y_{(k-1)h}))|_\infty^2]\mathrm{d}t \\
\leq & h \int_{(k-1)h}^{kh} \int_0^1 \mathbb{E}[|H_k(\nabla^2 V(uY_t + (1-u)Y_{(k-1)h}))(Y_t - Y_{(k-1)h})|_\infty^2]\mathrm{d}u\mathrm{d}t \\
\leq & 4h s_{r_2} \beta^2 \exp(-2\alpha h) \int_{(k-1)h}^{kh} \mathbb{E}[|Y_t - Y_{(k-1)h}|_\infty^2]\mathrm{d}t \\
= & 4h s_{r_2} \beta^2 \exp(-2\alpha h) \cdot O(h^2 \log(2d))
\end{aligned}
$$

Sketch of arguments: back to discretization errors

Putting everything together

- For general N:

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

- Therefore, we get

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

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