On efficient, approximate sampling

for high dims scientific computing

Yifan Chen

Courant Institute, New York University

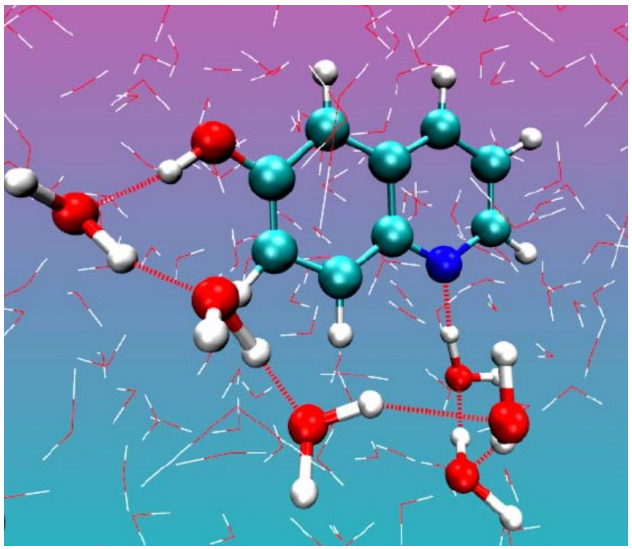
Data Science Seminars, University of Minnesota

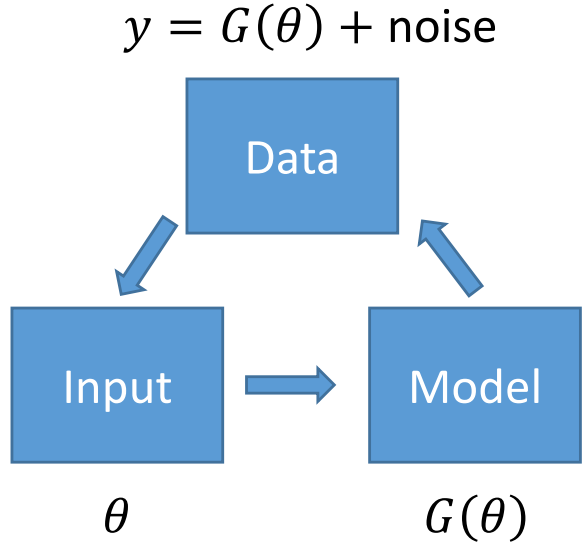
Context

The sampling problem

Goal: draw (approximate) samples from $\pi(x) \propto \exp(-V(x))$

(π or V may be directly modeled or learned from data)





Applications in molecular simulations, Bayes inverse problems, ...

Challenges: probability distributions in high dims

Methodology

Many methods:

- Markov chain Monte Carlo (MCMC), sequential Monte Carlo (SMC), ensemble Kalman methods, variational inference (VI), gradient flows, interacting particle systems, ...
- Leads to dynamics in the space of probability measures

A particular instance: SDEs

$$\mathrm{d}X_t = b_t(X_t)\mathrm{d}t + \sigma_t \mathrm{d}W_t$$

Task: efficient design and implementation of dynamics in high dims

Two Parts

1. **Delocalization of Discretization Bias Phenomenon**
   a nearly dimension-free discretization error phenomenon

2. **Föllmer's Diffusion Process for Probabilistic Forecasting**
   a "statistically optimal" design of diffusion coefficients

Two Parts

1. **Delocalization of Discretization Bias Phenomenon**
   a nearly dimension-free discretization error phenomenon

2. **Föllmer's Diffusion Process for Probabilistic Forecasting**
   a "statistically optimal" design of diffusion coefficients

MCMC Approach through Langevin's Dynamics

(Overdamped) Langevin's dynamics

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

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

Biased scheme: unadjusted Langevin, converging to $\pi_h \neq \pi$

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

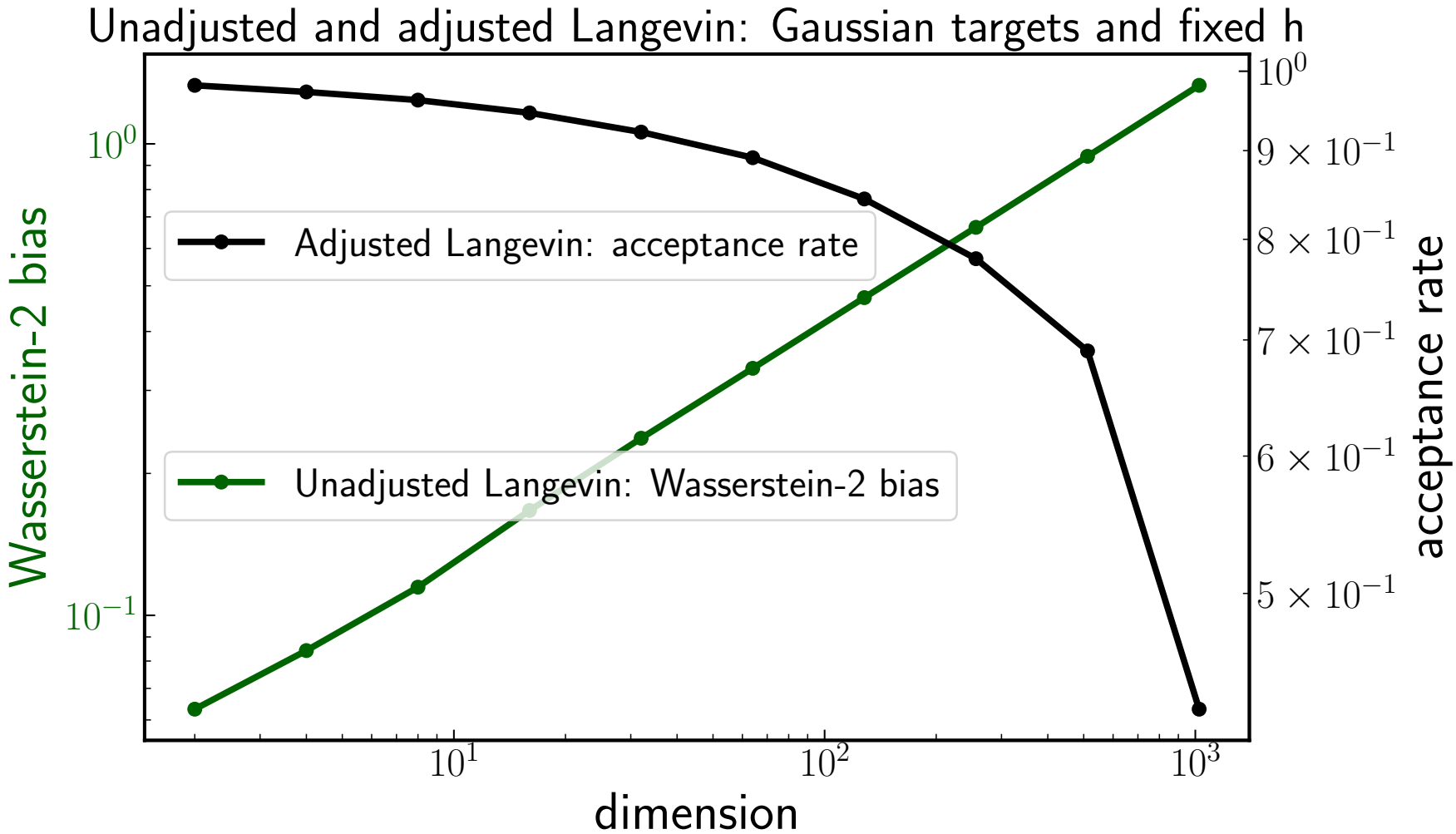
Unbiased scheme: e.g., Metropolis adjusted, converging to π

- MALA: accept w/ some probability, otherwise reject

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

See also the unbiased proximal sampler [Lee, Shen, Tian 2021], [Chen, Chewi, Salim, Wibisono 2022], etc.

Figure: For Fixed h, Both Schemes Degrade as Dimension Increases



- h needs to decrease as d increases, resulting in higher computational costs
- What should be the scaling of h regarding d?

Existing Theoretical Results on Complexity in High Dims

For MALA: h needs to be small for high acceptance rates

- Theories in the literature suggest $h \sim 1/d^{1/3}$ or $1/\sqrt{d}$ or $1/d$

  [Roberts, Rosenthal 1998], [Christensen, Roberts, Rosenthal 2005], [Dwivedi, Chen, Wainwright, Yu 2018], [Chewi, Lu, Ahn, Cheng, Gouic, Rigollet 2021], etc
- This scaling is not avoidable in general

For unadjusted Langevin: h needs to be small for small bias

- Theories in the literature suggest $h \sim 1/\sqrt{d}$ or $h \sim 1/d$

  [Durmus, Moulines, 2019], [Li, Zha, Tao 2022], etc.

Existing Theoretical Results on Complexity in High Dims

For MALA: h needs to be small for high acceptance rates

- Theories in the literature suggest $h \sim 1/d^{1/3}$ or $1/\sqrt{d}$ or $1/d$
  [Roberts, Rosenthal 1998], [Christensen, Roberts, Rosenthal 2005], [Dwivedi, Chen, Wainwright, Yu 2018], [Chewi, Lu, Ahn, Cheng, Gouic, Rigollet 2021], etc
- This scaling is not avoidable in general

For unadjusted Langevin: h needs to be small for small bias

- Theories in the literature suggest $h \sim 1/\sqrt{d}$ or $h \sim 1/d$
  [Durmus, Moulines, 2019], [Li, Zha, Tao 2022], etc.
- Some empirical observation: in molecular dynamics simulations using similar unadjusted integrators for billions of atoms, typically employ fixed step sizes of several femtoseconds, regardless of system size d [Leimkuhler, Matthews 2015]

Existing Theoretical Results on Complexity in High Dims

For MALA: h needs to be small for high acceptance rates

- Theories in the literature suggest $h \sim 1/d^{1/3}$ or $1/\sqrt{d}$ or $1/d$
  [Roberts, Rosenthal 1998], [Christensen, Roberts, Rosenthal 2005], [Dwivedi, Chen, Wainwright, Yu 2018], [Chewi, Lu, Ahn, Cheng, Gouic, Rigollet 2021], etc
- This scaling is not avoidable in general

For unadjusted Langevin: h needs to be small for small bias

- Theories in the literature suggest $h \sim 1/\sqrt{d}$ or $h \sim 1/d$
  [Durmus, Moulines, 2019], [Li, Zha, Tao 2022], etc.
- Some empirical observation: in molecular dynamics simulations using similar unadjusted integrators for billions of atoms, typically employ fixed step sizes of several femtoseconds, regardless of system size d [Leimkuhler, Matthews 2015]
- $h = O(1)$ could suffice for accurate averaged observables, e.g. $f(x) = \frac{1}{d}\sum_{i=1}^{d} x^{(i)}$, which satisfies $|\nabla f(x)|_2 \leq |x|_2/\sqrt{d}$
  [Bou-Rabee, Schuh 2023], [Durmus, Eberle 2024]

A Delocalization of Bias Phenomenon for Unadjusted Langevin

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

(informal) for unadjusted Langevin in d dims, $h = O(1/K)$ could suffice for desired accuracy in all K-marginals

- K-marginals: marginal distributions of any K-coordinates
- Rigorous results proved under the assumption $\alpha I \preceq \nabla^2 V \preceq \beta I$ and V is Gaussian/"sparse" (and some generalizations)
- Iteration complexity: $O(K)$ nearly independent of d

($\log d$ terms omitted)

A Delocalization of Bias Phenomenon for Unadjusted Langevin

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

(informal) for unadjusted Langevin in d dims, $h = O(1/K)$ could suffice for desired accuracy in all K-marginals

- K-marginals: marginal distributions of any K-coordinates
- Rigorous results proved under the assumption $\alpha I \preceq \nabla^2 V \preceq \beta I$ and V is Gaussian/"sparse" (and some generalizations)
- Iteration complexity: $O(K)$ nearly independent of d

($\log d$ terms omitted)

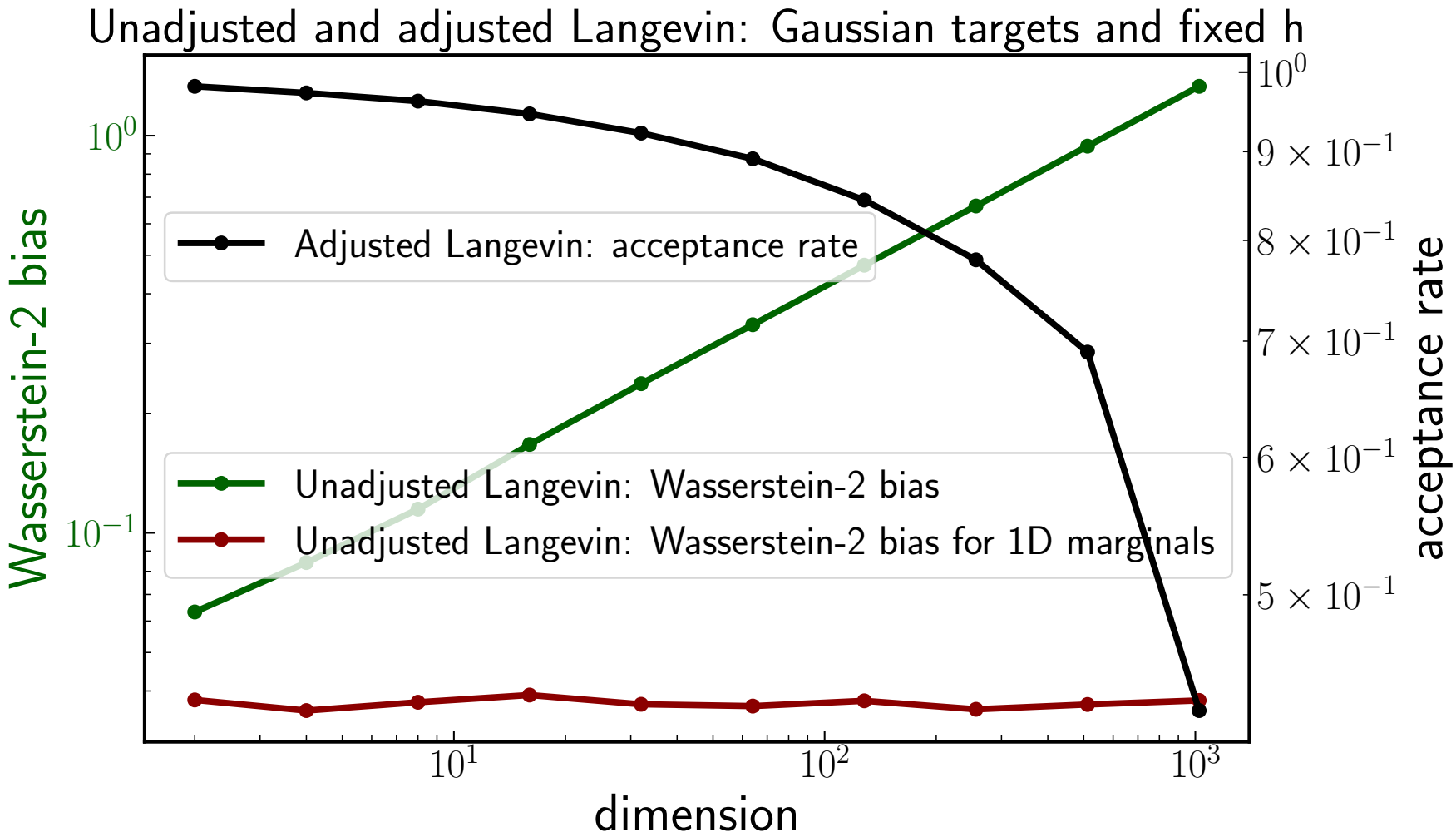
Bias in each individual coordinate behaves nearly dimension-free!

- We refer to this benign dimension dependence phenomenon as

"Delocalization of Bias"

- Phenomenon not shared by unbiased, adjusted approaches!

Updated Figure: If Interested in A Small Number of Coordinates



- Same for K-marginals, if K is independent of dimension (under the assumption of our theorem)
- Approximate, unadjusted approaches can be more scalable

How to Analyze? New Metric for Low Dimensional Marginals

Standard W_p metric: ℓ^2 measures full coordinates

$$W_p(\mu,\nu) = \left(\min_{\gamma \in \Pi(\mu,\nu)} \int |x-y|_2^p \, \gamma(\mathrm{d}x, \mathrm{d}y) \right)^{1/p}$$

New W_{p,ℓ^∞} metric: ℓ^∞ "can" measure a small set of variables

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

The rationale

- $K|x-y|_\infty^p \geq \sum_{t=1}^K |x^{(j_t)} - y^{(j_t)}|^p$ for any $1 \leq j_t \leq d$
- $K^{1/p} \cdot W_{p,\ell^\infty}(\mu,\nu)$ serves as an upper bound for the W_p distance between any K-dimensional marginals of μ and ν

In this work, we consider $p = 2$

Positive Examples: Product Measures and Gaussian Measures

Proposition: 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, for $h \leq 1/\beta$, it holds that

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

Proposition: 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, for $h \leq 1/\beta$, it holds that

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

- W_2 between K-marginals of π_h and π is $O(\sqrt{Kh\log(2d)})$
- Overall bias nearly delocalized accross all 1D marginals

A Negative Example: Strong, Dense Interactions In the Potential

Proposition: W_{2,ℓ^∞} bias for some rotated product measures

Let $\tilde{\pi} = Q\#\pi$, where

- Q is a rotation matrix which satisfies $(Qx)^{(1)} = \frac{1}{\sqrt{d}} \sum_{i=1}^{d} x^{(i)}$
- $\pi = \rho^{\otimes d}$ where ρ is a 1D centered distribution, such that the mean of ρ and the biased ρ_h differs by $\delta > 0$

Then it holds that

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

- We have $\tilde{\pi}_h = Q\#\pi_h$, and

$$\left| \int x^{(1)}(\tilde{\pi} - \tilde{\pi}_h) \right| = \left| \int f(\pi - \pi_h) \right| = \sqrt{d}\delta$$

  where $f(x) = \frac{1}{\sqrt{d}} \sum_{i=1}^{d} x^{(i)}$
- Thus $W_{2,\ell^\infty}(\tilde{\pi}, \tilde{\pi}_h) \geq W_{1,\ell^\infty}(\tilde{\pi}, \tilde{\pi}_h) \geq \left| \int x^{(1)}(\tilde{\pi} - \tilde{\pi}_h) \right| = \sqrt{d}\delta$

A Negative Example: Strong, Dense Interactions In the Potential

Proposition: W_{2,ℓ^∞} bias for some rotated product measures

Let $\tilde{\pi} = Q\#\pi$, where

- Q is a rotation matrix which satisfies $(Qx)^{(1)} = \frac{1}{\sqrt{d}} \sum_{i=1}^{d} x^{(i)}$
- $\pi = \rho^{\otimes d}$ where ρ is a 1D centered distribution, such that the mean of ρ and the biased ρ_h differs by $\delta > 0$

Then it holds that

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

- We have $\tilde{\pi}_h = Q\#\pi_h$, and

$$\left| \int x^{(1)}(\tilde{\pi} - \tilde{\pi}_h) \right| = \left| \int f(\pi - \pi_h) \right| = \sqrt{d}\delta$$

  where $f(x) = \frac{1}{\sqrt{d}} \sum_{i=1}^{d} x^{(i)}$

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

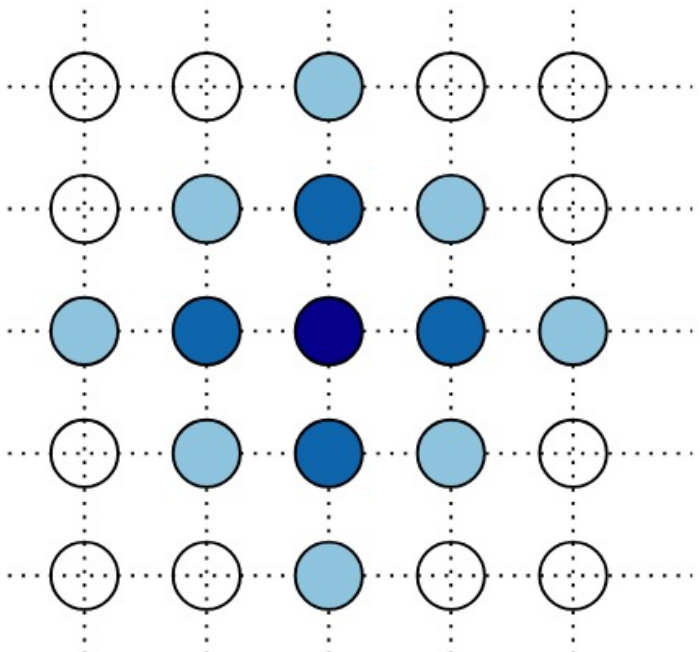
No delocalization, but concentration on one coordinate!

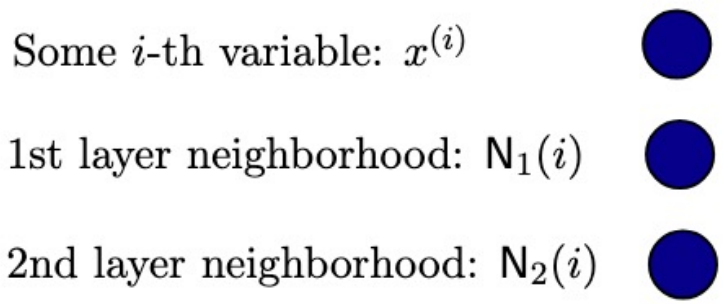
Delocalization for Sparse Potentials

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

(informal) suppose V is log-concave and satisfies the sparsity conditions illustrated in the figure with $s_k \le C(k+1)^n$, then

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





$V(x) = \sum_{i=1}^d V_i(\mathcal{X}^i)$, where $\mathcal{X}^i = \{x^{(j)} : j \in \mathsf{N}_1(i)\}$, and V_i is a function depending on variables in $\mathcal{X}^i$

Sparsity parameter $s_k := \sup_{1 \le i \le d} |\mathsf{N}_k(i)|$

For this example, $s_k = O(k^2)$

Figure: Illustration of a sparse potential covered in our analysis

Summary of the First Part

A "delocalization of bias" phenomenon for unadjusted Langevin

- Nearly d-independent step size and complexity
- Phenomenon not shared by unbiased schemes
- We prove it for log-concave Gaussians and sparse potentials
- Not hold for some potentials with strong, dense interactions
- Asymptotic arguments for general observables and potentials
  (up to first order; see our paper)

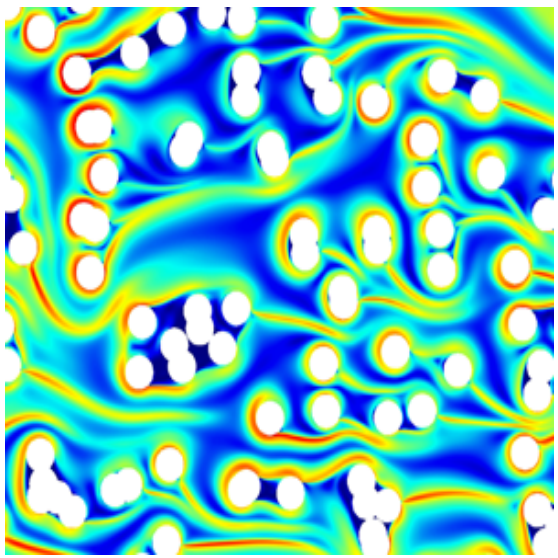
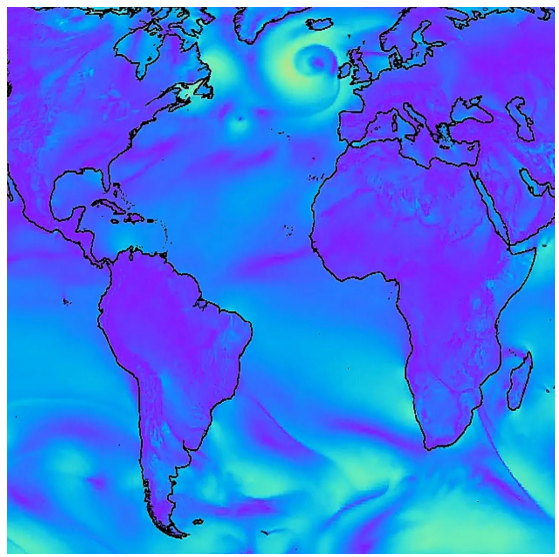
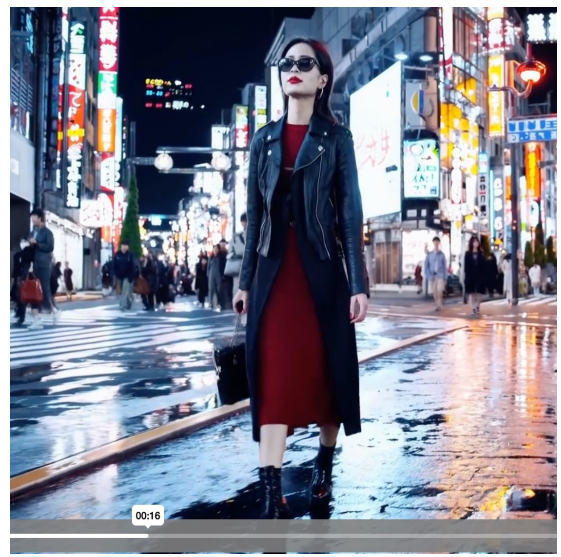
Two Parts

1 Delocalization of Discretization Bias Phenomenon
a nearly dimension-free discretization error phenomenon

2 Föllmer's Diffusion Process for Probabilistic Forecasting
a "statistically optimal" design of diffusion coefficients

Forecasting Problem

Given time series $(y_{k\tau})_{k\in\mathbb{Z}}$, predict $y_{(k+1)\tau}$ from new $y_{k\tau}$

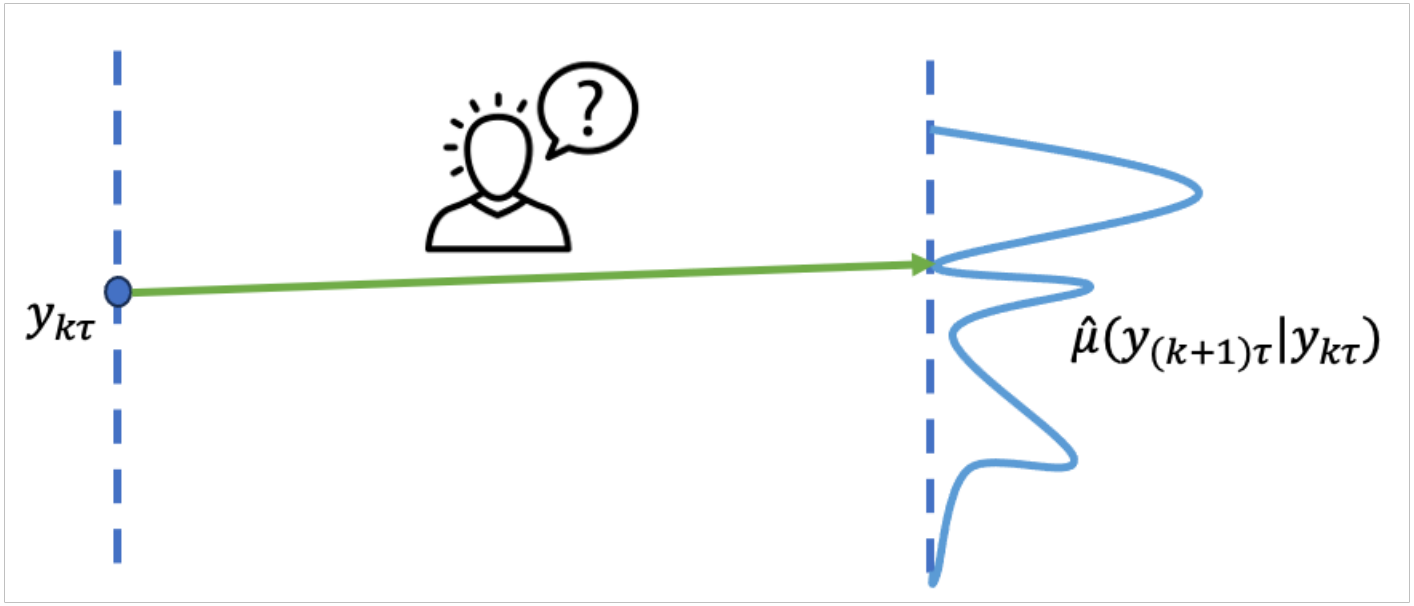




- Examples: fluids, daily weather measurements, video frames
- Assume successive observations $\sim$ joint PDF $\mu(y_{k\tau}, y_{(k+1)\tau})$
- Goal is conditional sampling $y_{(k+1)\tau} \sim \mu(\cdot|y_{k\tau})$

Figure credited to Google online search

Forecasting Problem

Given time series $(y_{k\tau})_{k\in\mathbb{Z}}$, predict $y_{(k+1)\tau}$ from new $y_{k\tau}$



Aim: Learn a generative stochastic dynamics based on data

Let x_0 and x_1 denote the current and forecasting state

Stochastic Interpolants with Diracs Base Distributions

Define the stochastic process $I_s = \alpha_s x_0 + \beta_s x_1 + \sigma_s W_s$

- $\alpha_0 = \beta_1 = 1$ and $\alpha_1 = \beta_0 = \sigma_1 = 0 \rightsquigarrow I_0 = x_0, I_1 = x_1$
- $(x_0, x_1) \sim \mu(x_0, x_1)$ joint distribution
- $W = (W_s)_{s\in[0,1]}$ is a Wiener process with $W \perp (x_0, x_1)$

Then, define the SDE

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

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

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

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

- Fact: $\mathrm{d}I_s = (\dot\alpha_s x_0 + \dot\beta_s x_1 + \dot\sigma_s W_s)\mathrm{d}s + \sigma_s \mathrm{d}W_s$

[Albergo, Vanden-Eijnden, 2022], [Albergo, Boffi, Vanden-Eijnden 2023]

See also [Liu, Gong, Liu 2022], [Lipman et al 2022], ...

Learning the Drift via Square Loss Regression

Definitions

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

- Conditional expectation $\rightsquigarrow$ square loss regression
- The drift $b_s(x, x_0)$ is the unique minimizer of

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

- Loss function is simulation-free: $W_s \overset{d}{=} \sqrt{s} z$ with $z \sim \mathsf{N}(0, \mathsf{I})$
- Parametrize $\hat{b}_s$ by neural nets and optimize L_b

Performance: Forecasting 2D Stochastically Forced Navier Stokes

2d NSE with Stochastic Forcing

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

- $v = \nabla^{\perp}\psi = (-\partial_y \psi, \partial_x \psi)$ is the velocity
- ψ is the stream function, solution to $-\Delta\psi = \omega$
- $\mathrm{d}\eta$ is white-in-time random forcing on a few Fourier modes
- $\nu = 10^{-3}, \alpha = 0.1, \epsilon = 1$
- Ergodicity shown in [Hairer, Mattingly, 2006]

Goal: Forecast $\omega_{t+\tau}$ from ω_t under stationarity

2D Stochastic Navier Stokes Experiments

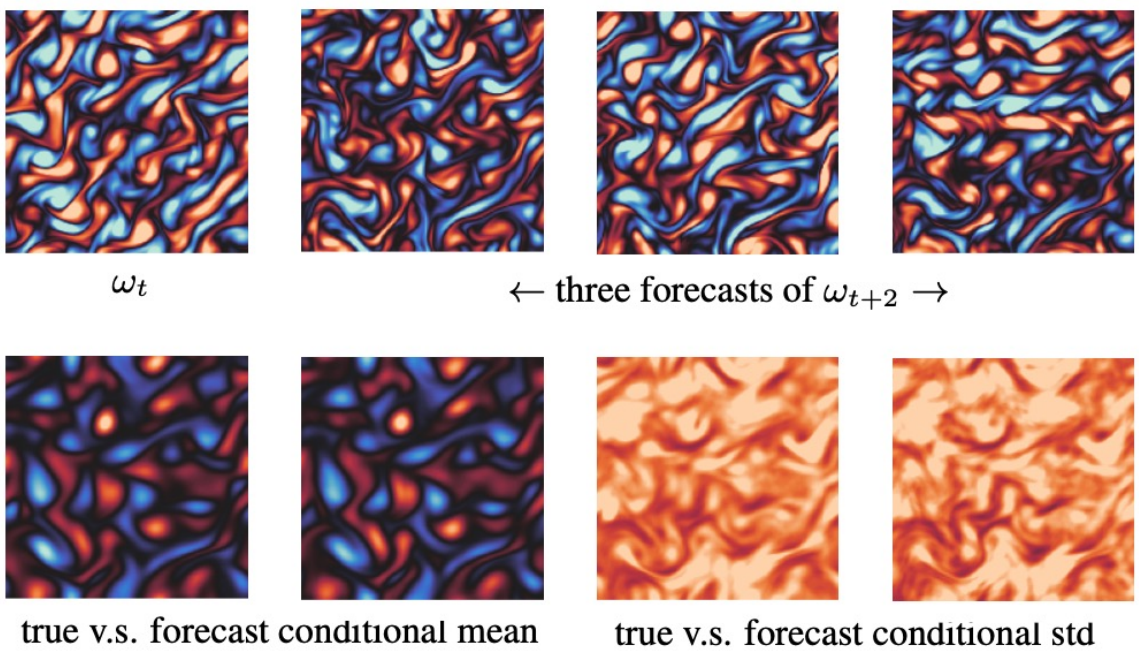


Figure: Lag $\tau = 2$ (autocorrelation 10%). Resolution 128×128, using $200K$ data pairs for training 2M-parameter-Unet

- Necessity of probabilistic over deterministic forecasting
- **Forecasting efficiency**: for this example 100 times faster than running the stochastic PDE simulation

Long time relative error	Föllmer	Deterministic
Total enstrophy (Iterate 100 times)	5.6e-3	3.0e-1

Forecasting with Incomplete Observation

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

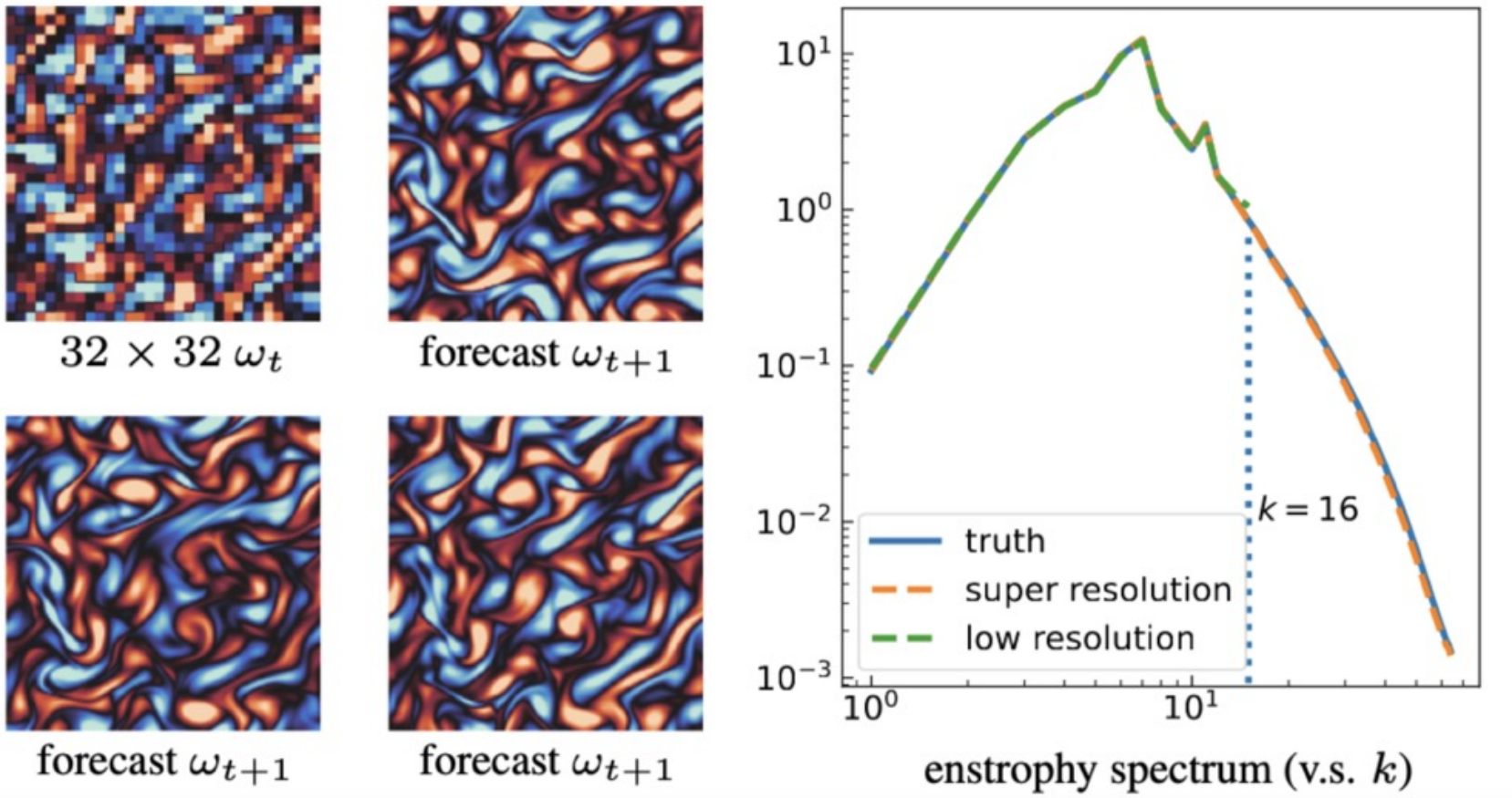


Figure: Probabilistic forecasting with low resolution input, using $200K$ data pairs for training 2M-parameter-Unet

Tuning drift and diffusion terms

It holds that $\text{Law}(X_s) = \text{Law}(X_s^g)$ for

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

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

- Based on the fact $\nabla \cdot (\rho \nabla \log \rho) = \Delta \rho$
- $\rho_s(x|x_0)$ is the PDF of $X_s \overset{d}{=} I_s|x_0$, which satisfies[1]

$$\nabla \log \rho_s(x|x_0) = A_s \left(\beta_s b_s(x, x_0) - c_s(x, x_0)\right)$$

  - $A_s = [s\sigma_s(\dot{\beta}_s \sigma_s - \beta_s \dot{\sigma}_s)]^{-1}$
  - $c_s(x, x_0) = \dot{\beta}_s x + (\beta_s \dot{\alpha}_s - \dot{\beta}_s \alpha_s) x_0$

  $\rightsquigarrow$ with an estimator $\hat{b}_s$, one can also easily get $\hat{b}_s^g$

- Obtained a family of SDEs can be used for generation purposes

[1] derived using Stein's identity or Tweedies formula

Optimal Tuning of g

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

Consider the KL between the path measures of

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

Then, KL is minimized if we set $g_s = g_s^{\mathsf{F}}$ with

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

- Recall $I_s = \alpha_s x_0 + \beta_s x_1 + \sigma_s W_s$, so the interpretation:

$$\frac{\beta_s}{\sqrt{s}\sigma_s} = \text{signal-to-noise ratio}$$

Föllmer's Processes

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

If $\beta_s/(\sqrt{s}\sigma_s)$ is non-decreasing, then $X^{g^{\mathsf{F}}}$ is an Föllmer process

- Föllmer processes solve the Schrödinger bridge problem when one endpoint is a point mass, offering an entropy-regularized solution to optimal transport
- Stochastic control formulation: (a_s, g_s depends on I_s)

$$\begin{aligned} &\min_u \int_0^1 \frac{1}{2|g_s|^2}\mathbb{E}|u_s(X_s)|^2 \mathrm{d}s, \\ \text{s.t. } &\mathrm{d}X_s = a_s(X_s)\mathrm{d}s + u_s(X_s)\mathrm{d}s + g_s \mathrm{d}B_s, \\ &X_0 = x_0,\ X_1 \sim \mu(\cdot|x_0) \end{aligned}$$

- Theorem provides a new interpretation of Föllmer as the minimizer of the KL of the exact forecasting process from the estimated one, which is more tailored to statistical inference

Föllmer process [Föllmer, 1986] wide applications

Effects of Tuning the Diffusion Coefficient in the Generative SDEs

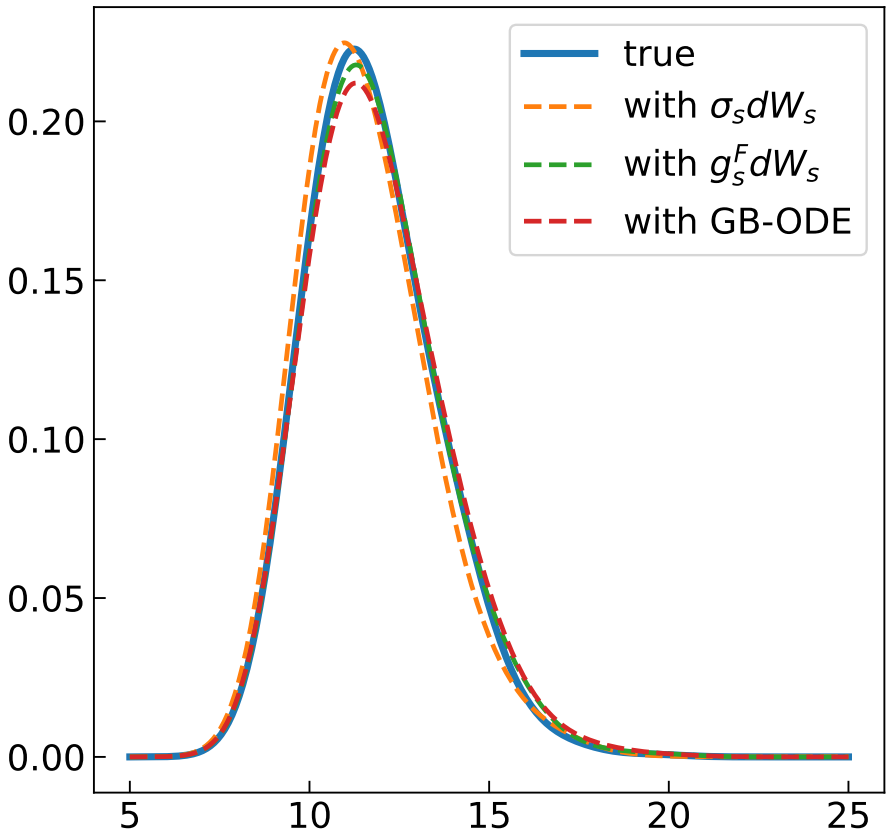


Figure: 1D density of total enstrophy

Comparison between

- Truth
- SDE samples with $\sigma_s dW_s$
- SDE with $g_s^{\mathrm{F}} dW_s$ (Föllmer process)
- ODE with Gaussian bases (conditional flow matching)

	KL: truth versus generation
SDE with $\sigma_s dW_s$	8.49e-3±1.57e-3
SDE with $g_s^{\mathrm{F}} dW_s$	2.79e-3±9.19e-4

Summary

Efficient, approximate dynamical sampling in high dims

Goal: understand and design efficient dynamics and their numerics

- A "delocalization of bias" phenomenon for unadjusted Langevin
  - Nearly d-independent step size and complexity: scalable
  - Phenomenon not shared by unbiased schemes
  - Generalization and algorithmic use of the phenomenon

- Föllmer's processes for probabilistic forecasting
  - "Statistically optimal" tuning of the diffusion coefficient
  - Schrödinger bridge between Diracs and target distribution
  - General design questions for the interpolation process I_s

Thank You!

Back-Up Slides

Forecasting Videos: CLEVER Datasets

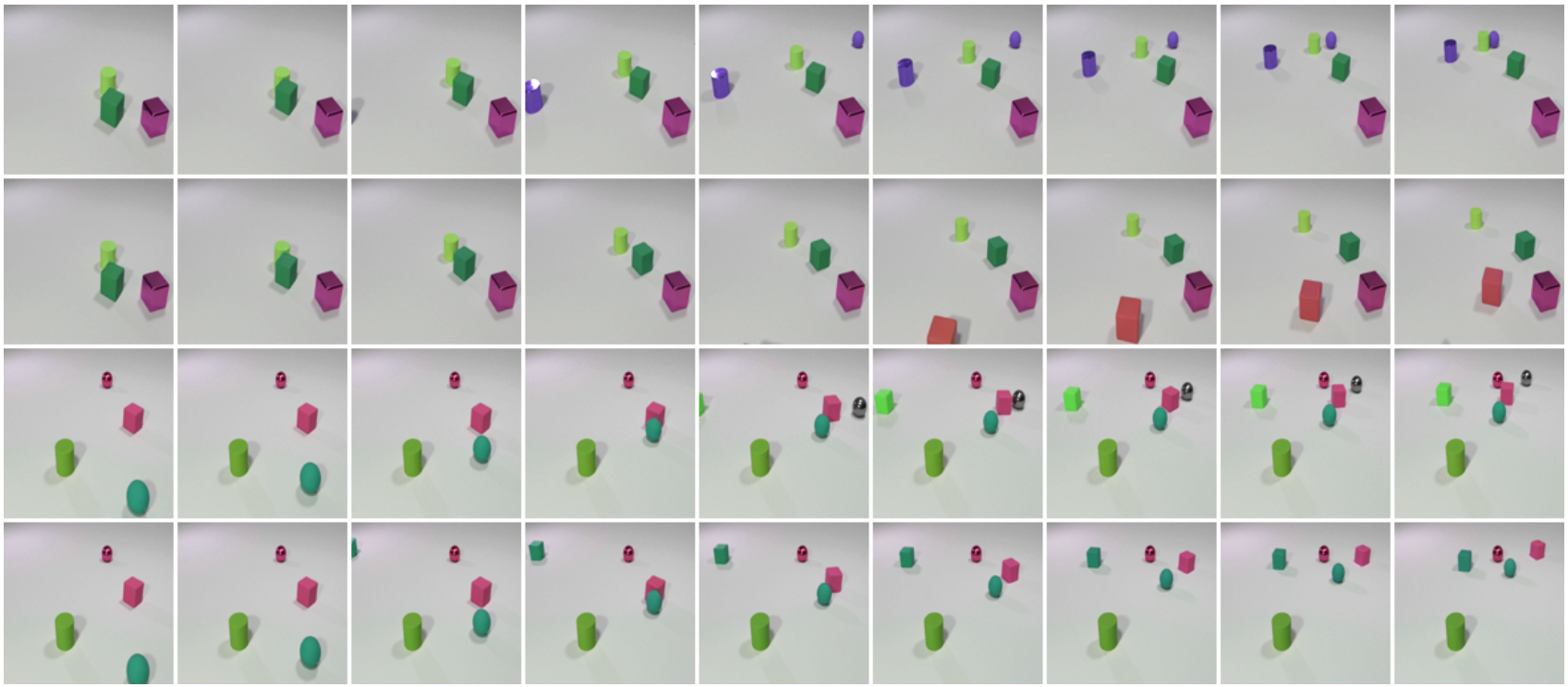


Figure: **Top row:** Real trajectory. **Second row:** Generated trajectory. A new, red cube enters the scene. **Third row:** Real trajectory. **Fourth row:** Generated trajectory. A new green cube enters the scene, and collision physics is respected (green ball hits red cube).

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]

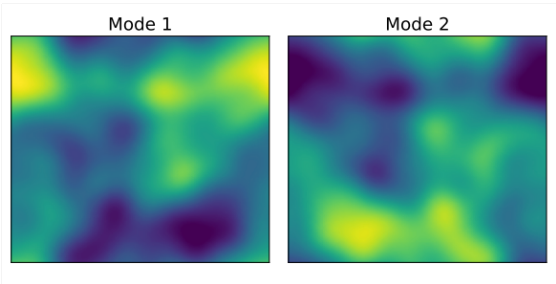
More: Approximate Sampling for High Dims Scientific Computing

model based

e.g., given priors and likelihood functions

Fisher-Rao gradient flows for Bayes inverse problems **[Chen, Huang, Huang, Reich, Stuart 2023]**

Efficient, multimodal, and derivative-free samplers

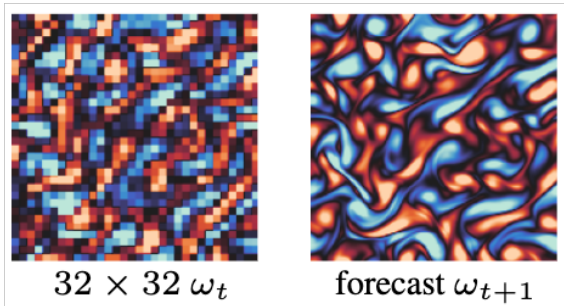


data driven

given only data samples

Föllmer's processes for probabilistic forecasting **[Chen, Goldstein, Hua, Albergo, Boffi, Vanden-Eijnden 2024]**

Statistical KL-optimality of Föllmer's processes

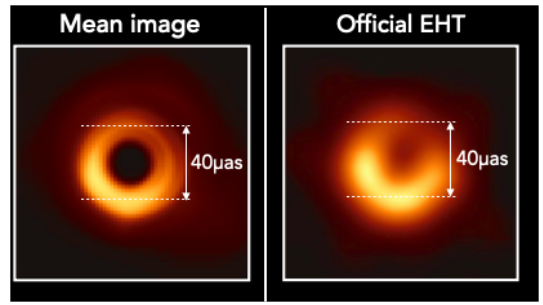


model and data

e.g., prior given as data, plus likelihood

Plug-and-Play Monte Carlo for probabilistic imaging **[Sun, Wu, Chen, Feng, Bouman 2023]**

Provable consistency for integrating generative priors



Sketch of Arguments

- 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 $\mathbb{E}[|B_t|_\infty^2] \leq t\log(2d)$)

Sketch of Arguments

- 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 $\mathbb{E}[|B_t|_\infty^2] \leq t\log(2d)$)

Sketch of Arguments

- 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

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

Sketch of Arguments: Multiple-step Coupling

- One-step iteration

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

- Moving back and two-step iterations

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

Sketch of Arguments: Multiple-step Coupling

- One-step iteration

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

- Moving back and two-step iterations

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

- N-step iterations

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

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

Sketch of Arguments: Bound Discretization Errors

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

- For $N = 1$:

$$
\begin{aligned}
&\mathbb{E}[|\overline{Y}_{(k+1)h} - Y_{(k+1)h}|_\infty^2] \\
=&\mathbb{E}[|\int_{kh}^{(k+1)h} \nabla V(Y_t) - \nabla V(Y_{kh})\mathrm{d}t|_\infty^2] \\
\leq& h\int_{kh}^{(k+1)h} \mathbb{E}[|\nabla V(Y_t) - \nabla V(Y_{kh})|_\infty^2]\mathrm{d}t \\
\leq& h\int_{kh}^{(k+1)h}\int_0^1 \mathbb{E}[|\nabla^2 V(uY_t + (1-u)Y_{kh})(Y_t - Y_{kh})|_\infty^2]\mathrm{d}u\mathrm{d}t \\
\leq& hs_1\beta^2 \int_{kh}^{(k+1)h} \mathbb{E}[|Y_t - Y_{kh}|_\infty^2]\mathrm{d}t = hs_1\beta^2 \cdot O(h^2\log(2d))
\end{aligned}
$$

Sketch of Arguments: Bound Discretization Errors

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

- For $N = 2$:

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

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 $|H_k(\nabla^2 V(uY_t + (1-u)Y_{(k-1)h}))|_\infty$?

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

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

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

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

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:

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

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

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

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 $N = \lceil \frac{\log(2\sqrt{d})}{h\alpha} \rceil$

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

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

Asymptotic Arguments for the Bias of Observables

Bias of Observables [Chen, Cheng, Niles-Weed, Weare 2024]

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

$$\int f\pi - \int f\pi_h = \frac{1}{4}h\left(\int (-2\Delta f + |\nabla \log \pi|_2^2 f)\pi\right) + o(h)$$

Moreover, we also have the following formula:

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

Poisson argument: Let $\mathcal{L}$ and $\mathcal{L}_h$ be the generators of Langevin dynamics and unadjusted Langevin [Mattingly, Stuart, Tretyakov 2010]

- $\mathcal{L}u = \nabla \log \pi \cdot \nabla u + \Delta u$,
  $\mathcal{L}_h u(x) = \frac{1}{h}(\mathbb{E}[u(x + h\nabla \log \pi(x) + \sqrt{2h}\xi)] - u(x))$
- Let $\mathcal{L}u = f$. Then, we get

$$\int f\pi - \int f\pi_h = -\int \mathcal{L}u\pi_h = \int (\mathcal{L}_h u - \mathcal{L}u)\pi_h, \quad \ldots$$

Delocalization of Bias for Observables

Bias of Observables [Chen, Cheng, Niles-Weed, Weare 2024]

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

$$\int f\pi - \int f\pi_h = \frac{1}{4}h\left(\int (-2\Delta f + |\nabla \log \pi|_2^2 f)\pi\right) + o(h)$$

Moreover, we also have the following formula:

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

- If $\pi(x) = \mathcal{N}(x; m, \Sigma)$, then $\int f(\Delta \log \pi)\pi = 0$. The first order term $\int \pi \Delta f$ only depends on the coordinates that f takes
- This delocalization of observable bias can be generalized to

$$\pi(x) \propto \exp(-V(x)) \propto \mathcal{N}(x; m, \Sigma)\exp(-U(x))$$

i.e., perturbation of Gaussians

Plug-and-Play Coupled Dynamics for Probabilistic Imaging

- Bayes priors, learned as a stochastic dynamics via generative modeling
- plug-and-play: couple dynamics with likelihood for posterior sampling
- we achieve this coupling using annealed Langevin and split-Gibbs samplers with rigorous consistency guarantee [Sun, Wu, **Chen**, Feng, Bouman 2023] [Wu, Sun, **Chen**, Zhang, Yue, Bouman 2024]

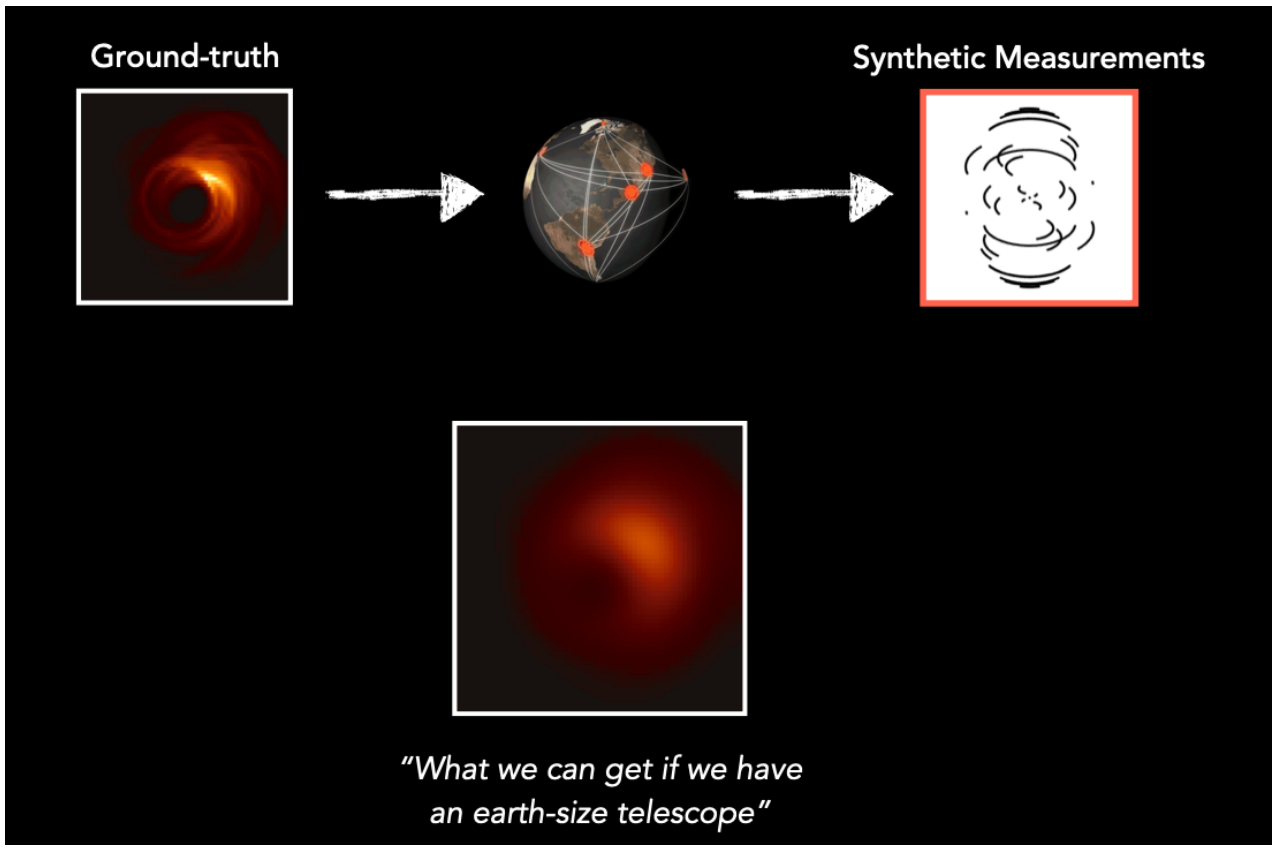
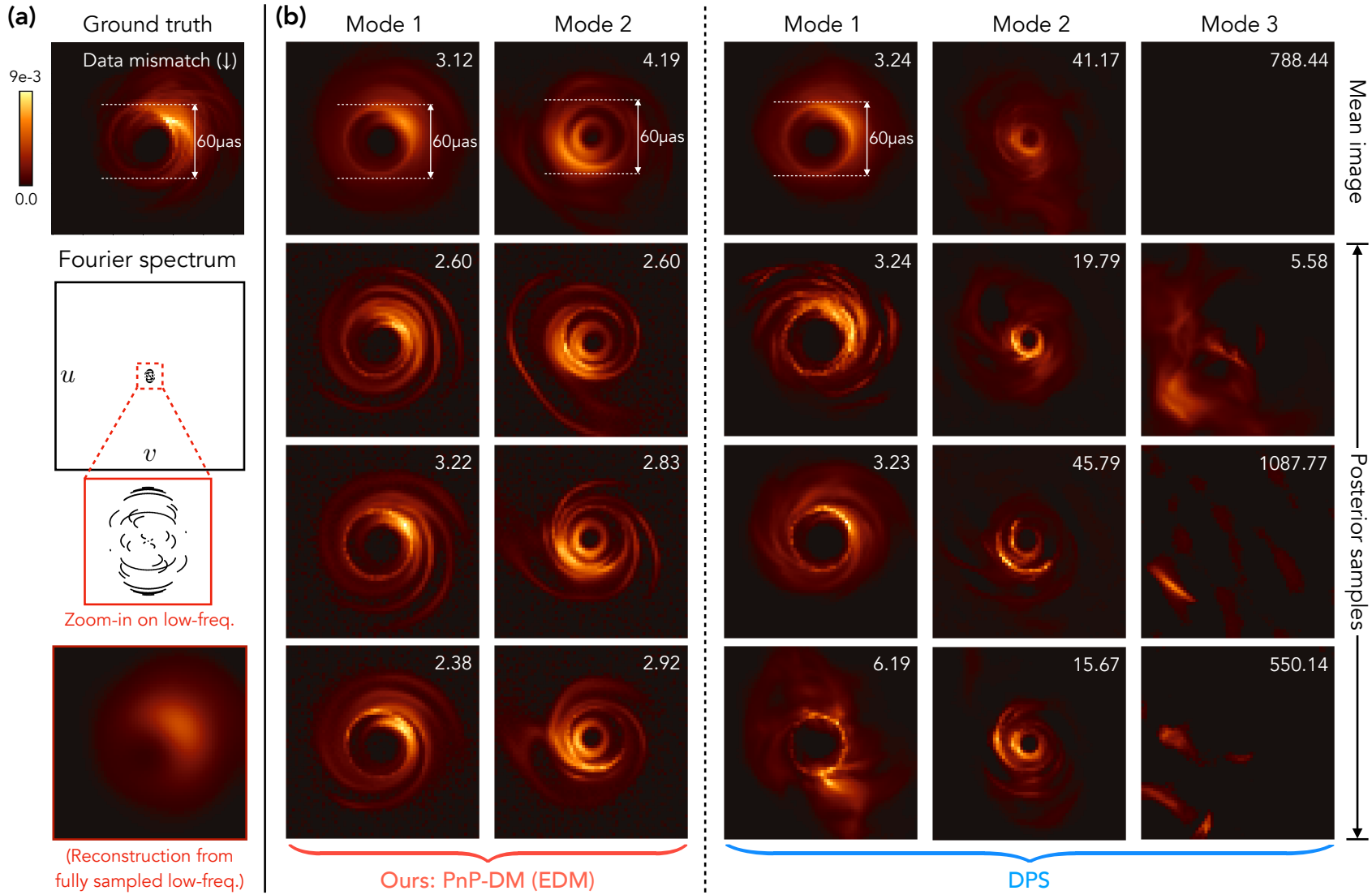
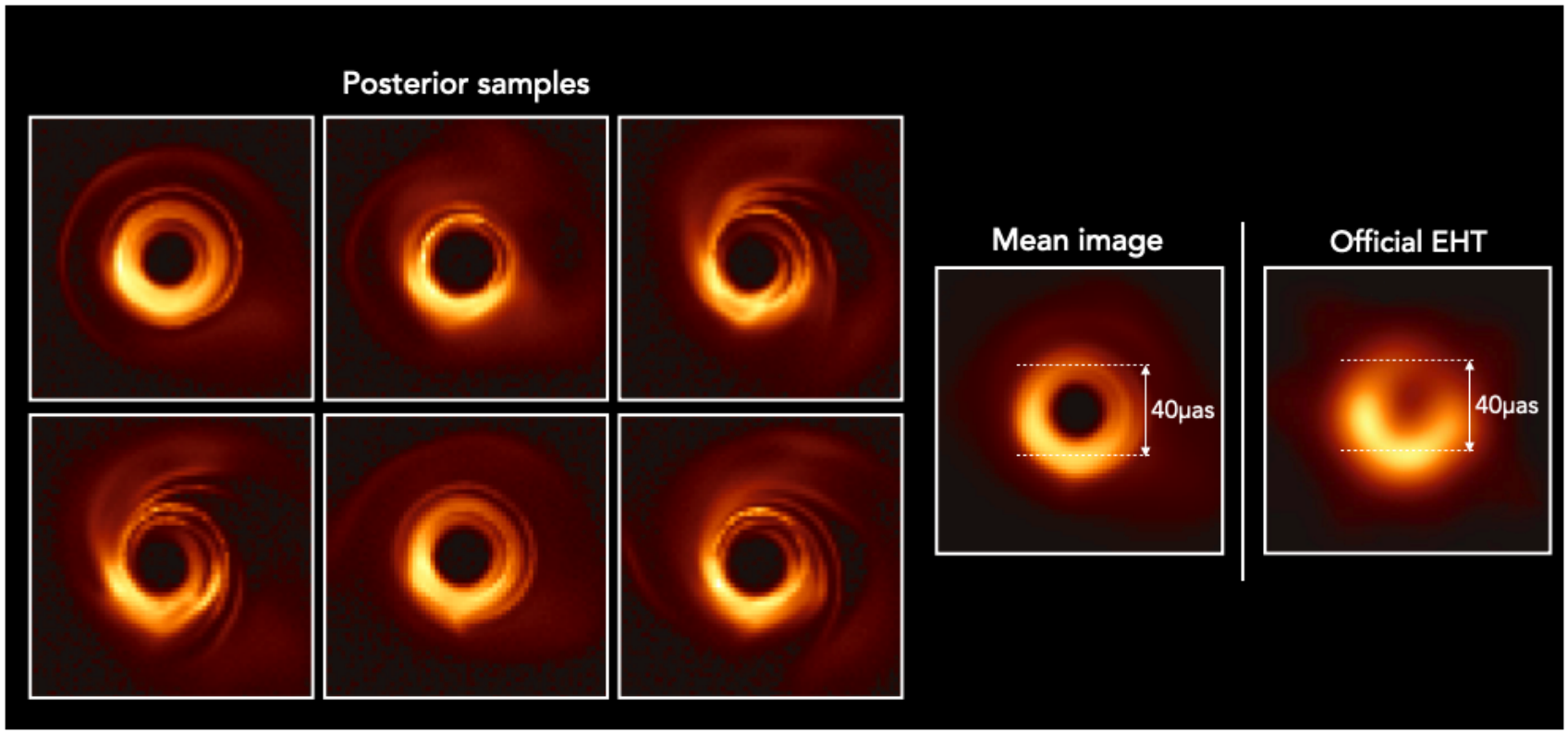


Illustration of the problem

Synthetic Data



Real Data



Challenges

The Bayes inverse problem

$$\rho_{\text{post}}(\theta) \propto \rho(y|\theta)\rho_{\text{prior}}(\theta) \propto \exp(-\Phi_R(\theta, y))$$

$$\text{where } \Phi_R(\theta, y) = \frac{1}{2}\|\Sigma_\eta^{-\frac{1}{2}}(y - G(\theta))\|^2 + \frac{1}{2}\|\Sigma_0^{-\frac{1}{2}}(\theta - r_0)\|^2$$

1. Evaluation of G is expensive: require large scale PDE solvers
2. The posterior distribution $\rho_{\text{post}}(\theta)$ can have multiple modes
3. The gradient of Φ_R may not available or even feasible

Ask for **fast, multimodal, and derivative-free** Bayes sampler

Fisher-Rao Gradient Flow

Fisher-Rao gradient flow of KL divergence

$$\frac{\partial \rho_t}{\partial t} = \rho_t \big(\log \rho_{\text{post}} - \log \rho_t\big) - \rho_t \mathbb{E}_{\rho_t}[\log \rho_{\text{post}} - \log \rho_t]$$

- KL divergence

$$\mathcal{E}(\rho) = \mathrm{KL}[\rho \| \rho_{\text{post}}] = \int \rho \log\Big(\frac{\rho}{\rho_{\text{post}}}\Big)\, \mathrm{d}\theta$$

- Fisher-Rao metric tensor

$$M(\rho)^{-1}\psi = \rho(\psi - \mathbb{E}_\rho[\psi])$$

- The gradient flow equation

$$\frac{\partial \rho_t}{\partial t} = -M(\rho_t)^{-1} \frac{\delta \mathcal{E}}{\delta \rho}|_{\rho=\rho_t} = -M(\rho_t)^{-1}(\log \rho_t - \log \rho_{\text{post}})$$

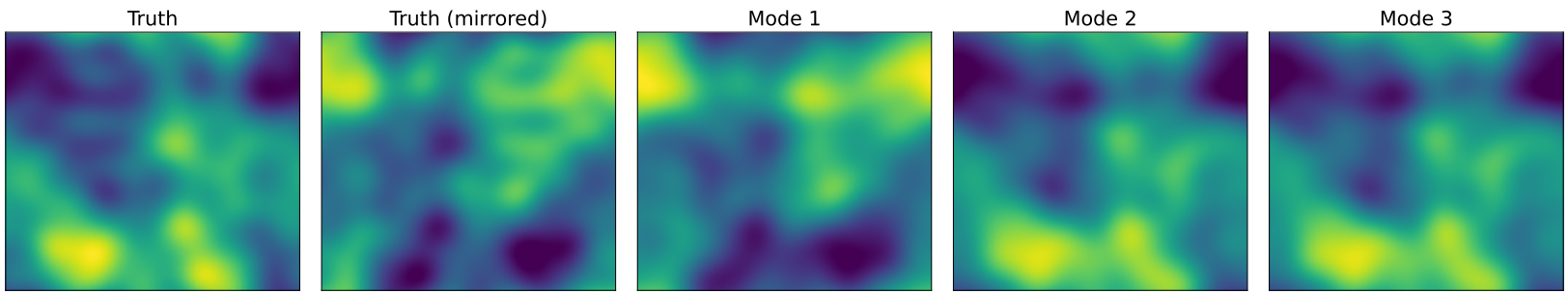
Information geometry [Amari 2016], [Ay, Jost, Lê, Schwachhöfer, 2017]

See also: Wasserstein gradient flow, Stein variational gradient flow, ...

Fisher-Rao Gradient Flow and Kalman Approximation

$$\text{(Dynamics)} \quad \frac{\partial \rho_t}{\partial t} = \rho_t\big(\log \pi - \log \rho_t\big) - \rho_t \mathbb{E}_{\rho_t}[\log \pi - \log \rho_t]$$

- the unique gradient flow that is diffeomorphism-invariant and independent of normalization constants, and converges uniformly
  [Cencov 2000], [**Chen**, Huang, Huang, Reich, Stuart 2023]
  Theory of functional inequalities: [Carrillo, **Chen**, Huang, Huang, Wei 2024]
- we use Kalman methodology for derivative-free, Gaussian mixture approximations; this leads to Gaussian mixture Kalman inversion (GMKI): a Kalman-filter variant of Gaussian mixture VI
  [**Chen**, Huang, Huang, Reich, Stuart 2024]
- experiments: 2D Navier Stokes in $[0, 2\pi]^2$, viscosity $\nu = 0.01$. Goal: recover initial vorticity from 56 point observations at $T = 0.25, 0.5$



True vorticity (128 basis func.); modes obtained by 3-modal GMKI with 50 iterations

High-dimensional Bimodal Problem

Consider 2d NSE on a periodic domain $D = [0, 2\pi] \times [0, 2\pi]$

$$\frac{\partial \omega}{\partial t} + (v \cdot \nabla)\omega - \nu \Delta \omega = \nabla \times f$$

- Viscosity $\nu = 0.01$
- Non-zero mean background velocity $v_b = [0, 2\pi]$
- $f(x_1, x_2) = [0, \cos(4x_1)]$
- **Goal:** learn initial vorticity based on observed vorticity at some observation points at later times $T = 0.25, 0.5$
- Gaussian process prior on initial vorcitity (we keep the first 128 Karhunen-Loève expansion coefficients and use data to learn these coefficients $\theta \in \mathbb{R}^{128}$)

Multimodal Setting: Symmetry in Observations

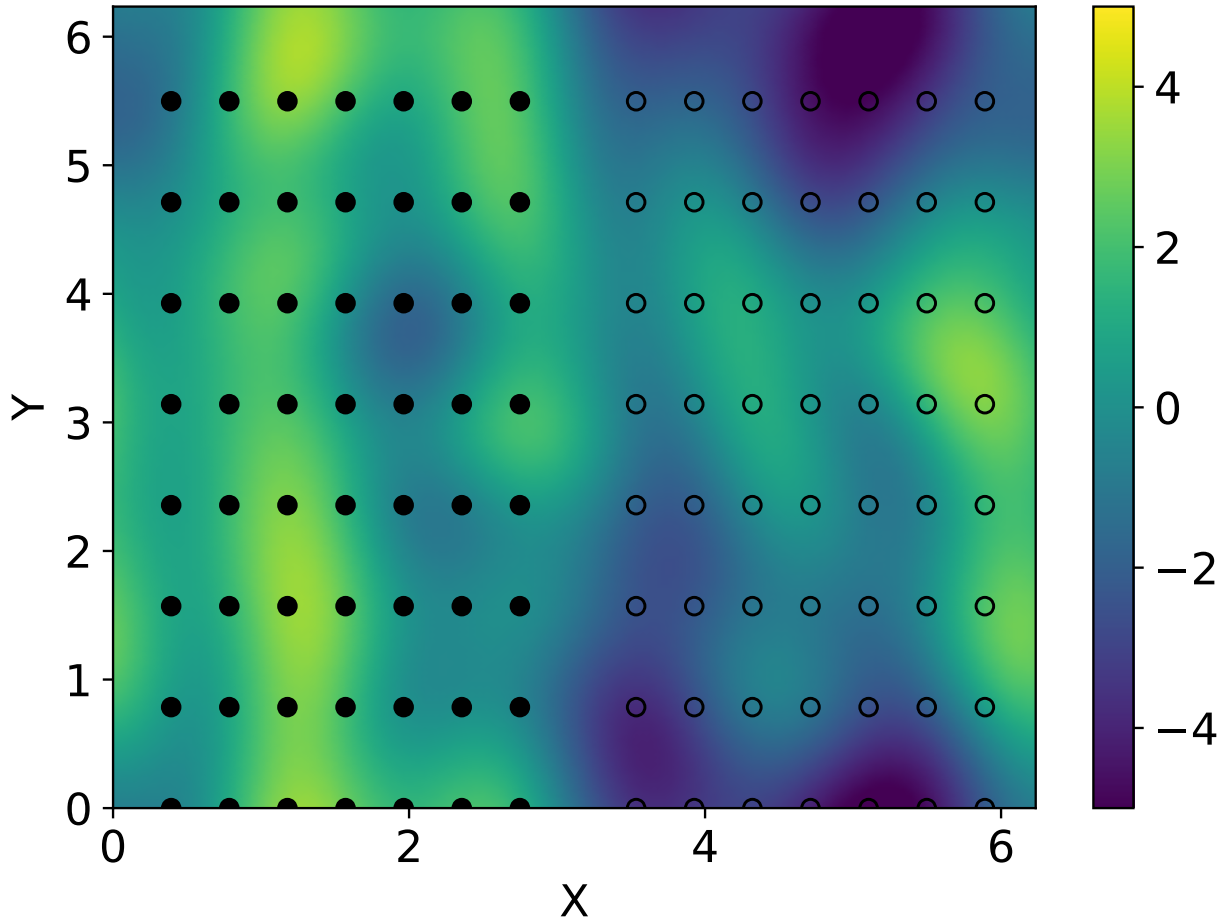


Figure: Vorticity observations $\omega([x_1, x_2]) - \omega([2\pi - x_1, x_2])$ at 56 equidistant points (solid black dots)

Results for Learning Initial Vorticity in 2D NSE: $K = 3$

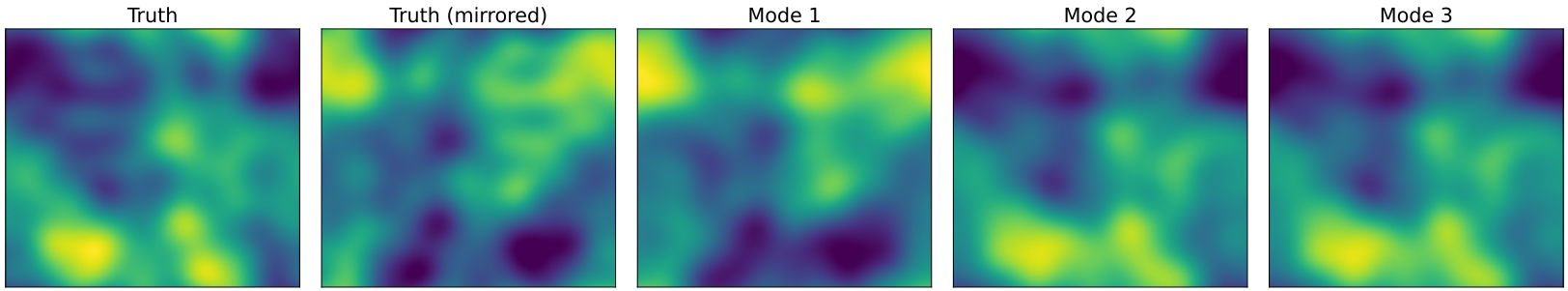


Figure: The true vorticity field, and these modes obtained by GMKI

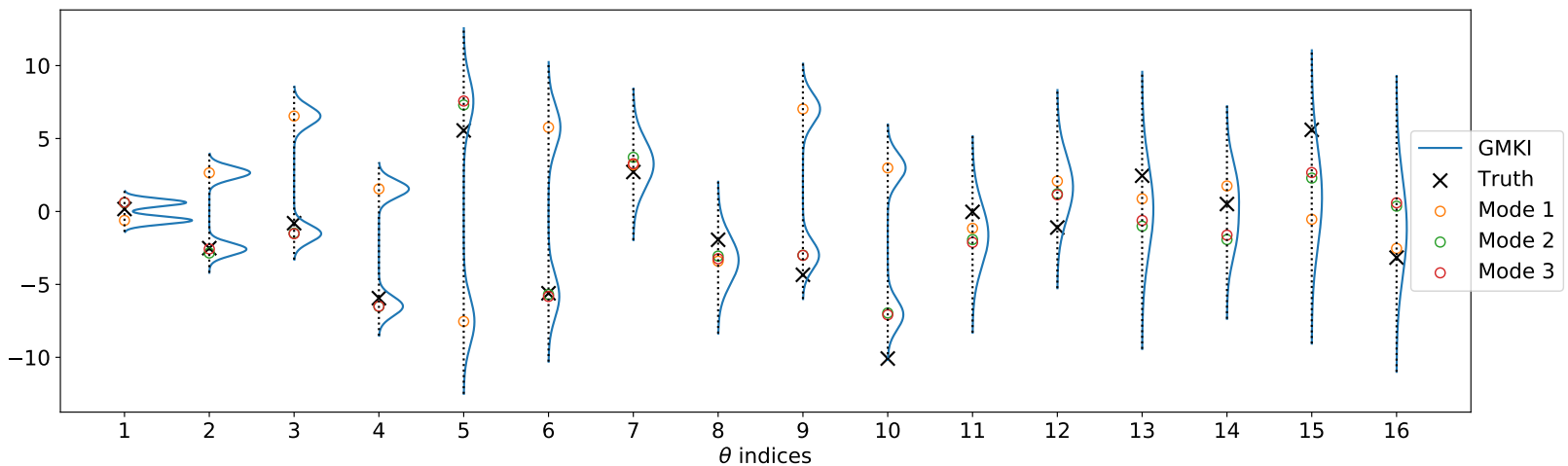


Figure: The truth KL expansion coefficients θ_i (black crosses), and mean estimations of θ_i for each modes (circles) and the associated marginal distributions obtained GMKI at the 50th iteration