Generative Models, Optimal Transport, and PDEs

statistical foundations in the GenAI age

Tengyuan Liang

The University of Chicago

Plan

Prelude: motivation

Sensitivity Analysis: denoising diffusions and Fokker-Planck PDE

Interlude: thoughts on generative models

No-Regret Analysis: Sinkhorn algorithm and Monge-Ampère PDE

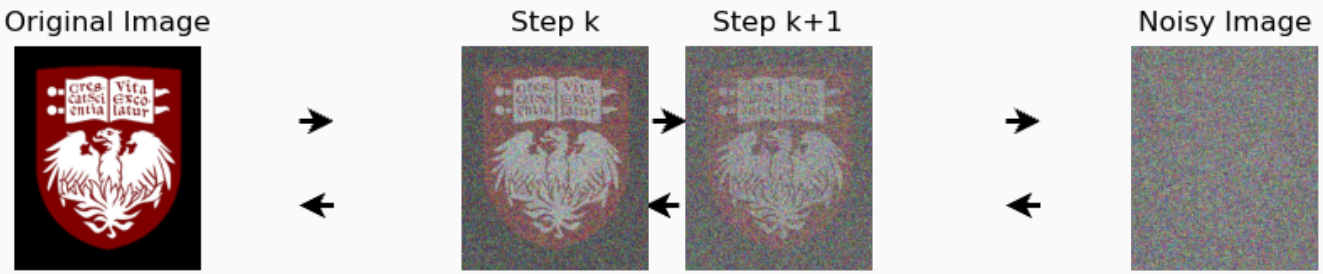
Prelude: motivation

Why do generative models work?

- rethink learning to sample from data distributions
- theories for probabilistic generative models

Optimal Transport Perspectives for Generative Models

diffusion-based generative model



Forward chain: add noise to **data point**

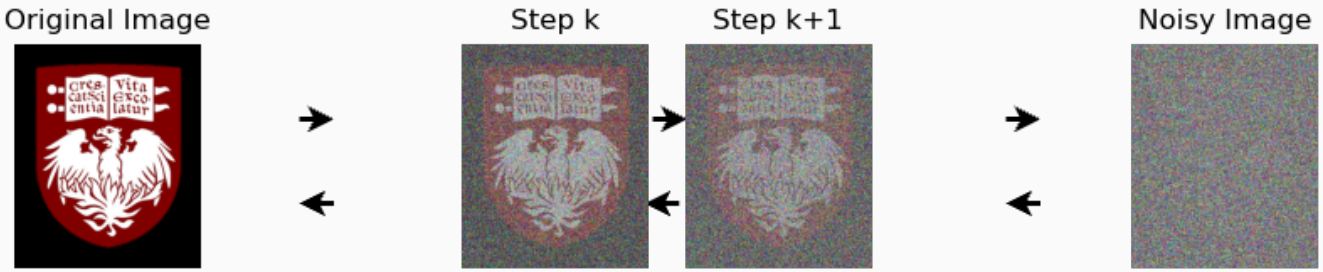
$$X_0 \to \cdots \to X_k \to X_{k+1} \to \cdots \to X_\infty$$

Backward chain: remove noise from **data point**

$$X_0 \leftarrow \cdots \leftarrow X_k \leftarrow X_{k+1} \leftarrow \cdots \leftarrow X_\infty$$

Saremi and Hyvärinen (2019), Sohl-Dickstein, Weiss, Maheswaranathan, and Ganguli (2015), Song and Ermon (2019), and Song, Sohl-Dickstein, et al. (2020)

diffusion-based generative model



Forward chain: Markov chain

diffuse real data to white noise, **learning** process

Backward chain: time-reversal of a Markov chain

denoise white noise to real data, **generative/sampling** process

paradoxical to a theory researcher

Empirically: state-of-the-art results, DALL-E, Stable Diffusion, ...

Song, Durkan, Murray, and Ermon (2021) and Song and Ermon (2020)

(if infinite computational resources, **perfect** generative model)

paradoxical to a theory researcher

Empirically: state-of-the-art results, DALL-E, Stable Diffusion, ...

Song, Durkan, Murray, and Ermon (2021) and Song and Ermon (2020)

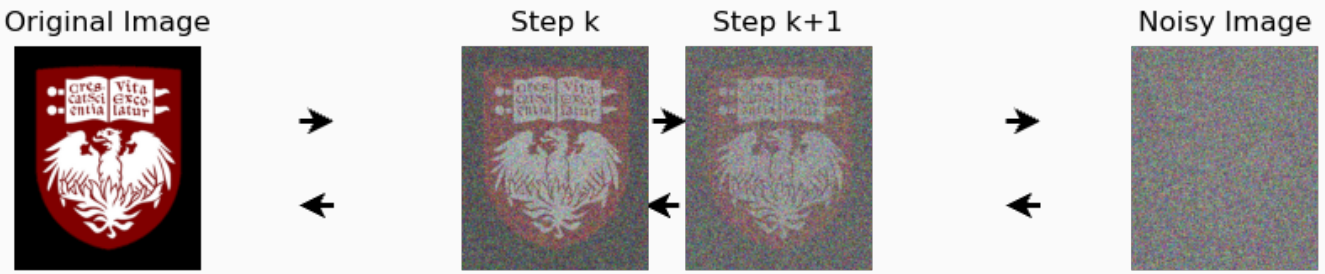
(if infinite computational resources, **perfect** generative model)

However, conceptually problematic in {**probability**, **optimization**, **information**} sense

- **probability**: time-reversal of a Markov chain will get stuck in equilibrium, cannot recover past from future
- **optimization**: if two paths converge to the same point, impossible to recover the initial condition
- **information**: inject noise then denoise, how can one be better off?

(~~if infinite computational resources, **perfect** generative model~~)

resolving the paradox



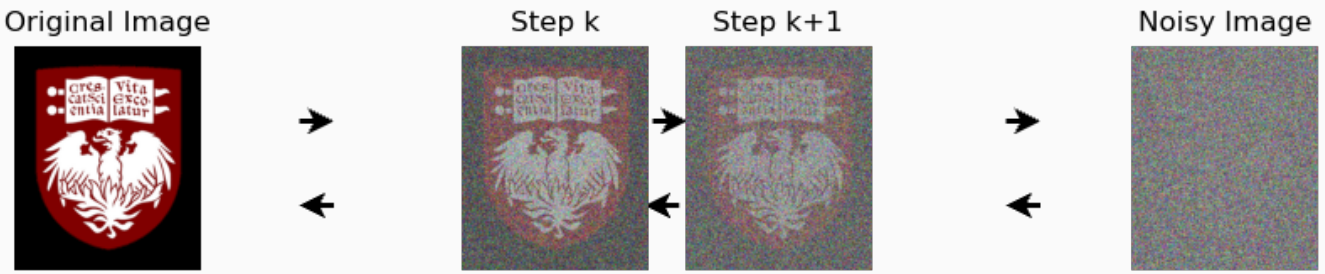
μ: real data distribution, ν: white noise distribution

Forward Diffusion $\mu =: \mu_0 \xrightarrow{\mathbf{f}_1^\mu} \mu_\eta \to \cdots \xrightarrow{\mathbf{f}_k^\mu} \mu_{k\eta} \to \cdots \xrightarrow{\mathbf{f}_T^\mu} \mu_{T\eta} \overset{T\to\infty}{\rightsquigarrow} \nu$

Time Reversal $\mu =: \mu_0 \underset{\mathbf{b}_1^\mu}{\leftarrow} \mu_\eta \leftarrow \cdots \underset{\mathbf{b}_k^\mu}{\leftarrow} \mu_{k\eta} \leftarrow \cdots \underset{\mathbf{b}_K^\mu}{\leftarrow} \mu_{T\eta}$ $\boxed{\underset{?}{\leftarrow} \cdots \underset{\mathbf{b}^\nu}{\leftarrow} \nu}$

infinite computation and data: $\nu \underset{\mathbf{b}^\nu}{\leftarrow} \nu$

resolving the paradox



Time Reversal $\mu =: \mu_0 \underset{\mathbf{b}_1^\mu}{\leftarrow} \mu_\eta \leftarrow \cdots \underset{\mathbf{b}_k^\mu}{\leftarrow} \mu_{k\eta} \leftarrow \cdots \underset{\mathbf{b}_K^\mu}{\leftarrow} \mu_{K\eta}$ $\not\leftarrow \cdots \not\leftarrow \nu$ (? $\mathbf{b}^\nu$)

Backward Denoising $\mu \overset{?}{\leftsquigarrow} \bar{\nu}_0 \underset{\mathbf{b}_1^\mu}{\leftarrow} \bar{\nu}_\eta \leftarrow \cdots \underset{\mathbf{b}_k^\mu}{\leftarrow} \bar{\nu}_{k\eta} \leftarrow \cdots \boxed{\underset{\mathbf{b}_K^\mu}{\leftarrow} \bar{\nu}_{K\eta} := \nu}$

resolving the paradox

Time Reversal $\mu =: \mu_0 \underset{\mathbf{b}_1^\mu}{\leftarrow} \mu_\eta \leftarrow \cdots \underset{\mathbf{b}_k^\mu}{\leftarrow} \mu_{k\eta} \leftarrow \cdots \underset{\mathbf{b}_K^\mu}{\leftarrow} \mu_{K\eta}$ $\not\leftarrow \cdots \not\leftarrow \nu$ (? / $\mathbf{b}^\nu$)

Backward Denoising $\mu \overset{?}{\rightsquigarrow} \bar{\nu}_0 \underset{\mathbf{b}_1^\mu}{\leftarrow} \bar{\nu}_\eta \leftarrow \cdots \underset{\mathbf{b}_k^\mu}{\leftarrow} \bar{\nu}_{k\eta} \leftarrow \cdots \underset{\mathbf{b}_K^\mu}{\leftarrow} \bar{\nu}_{K\eta} := \nu$

infinite computation and data: denoising maps $\mathbf{b}_k^\mu$'s are exact

denoising maps $\mathbf{b}_k^\mu$ carries information of the initial measure μ

(recover the past from the future)

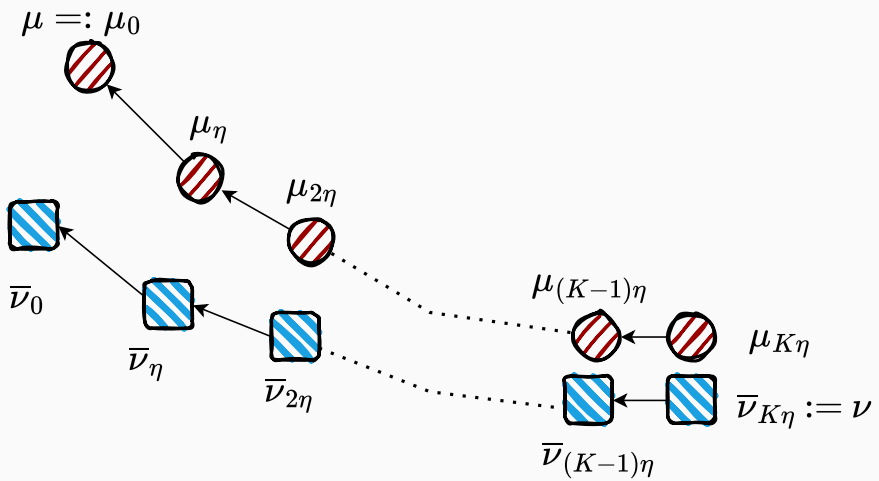
resolving the paradox

Time Reversal $\mu =: \mu_0 \underset{\mathbf{b}_1^\mu}{\leftarrow} \mu_\eta \leftarrow \cdots \underset{\mathbf{b}_k^\mu}{\leftarrow} \mu_{k\eta} \leftarrow \cdots \underset{\mathbf{b}_K^\mu}{\leftarrow} \mu_{K\eta}$ $\not\leftarrow \cdots \not\leftarrow \nu$ (? $\mathbf{b}^\nu$)

Backward Denoising $\mu \overset{?}{\rightsquigarrow} \bar{\nu}_0 \underset{\mathbf{b}_1^\mu}{\leftarrow} \bar{\nu}_\eta \leftarrow \cdots \underset{\mathbf{b}_k^\mu}{\leftarrow} \bar{\nu}_{k\eta} \leftarrow \cdots \underset{\mathbf{b}_K^\mu}{\leftarrow} \bar{\nu}_{K\eta} := \nu$

Sensitivity analysis

whether the perturbation $\bar{\nu}_{K\eta} \approx \mu_{K\eta}$ will get amplified along the backward denoising chain



what's missing in the current theory

current theory

- sensitivity analysis: on denoising map $\mathbf{b}_k^\mu$, estimated based on finite data, namely $\widehat{\mathbf{b}_k^\mu} \approx \mathbf{b}_k^\mu$
- with infinite data and computation, namely $\mathbf{b}_k^\mu$ is known: non-expansion metric $d(\cdot, \cdot)$

$$d(\mu_{k-1}, \bar{\nu}_{k-1}) / d(\mu_k, \bar{\nu}_k) \leq 1$$

f-divergence, data processing inequality

Not fine-grained enough: the backward denoising chain does not matter

Song, Durkan, Murray, and Ermon (2021), S. Chen et al. (2022), Lee, Lu, and Tan (2022), H. Chen, Lee, and Lu (2023), ...

what's missing in the current theory

current theory

with infinite data and computation, namely $\mathbf{b}_k^\mu$ is known: non-expansion metric $d(\cdot, \cdot)$

$$d(\mu_{k-1}, \bar{\nu}_{k-1})/d(\mu_k, \bar{\nu}_k) \leq 1$$

Not fine-grained enough: the backward denoising chain does not matter

what's missing in the current theory

current theory

with infinite data and computation, namely $\mathbf{b}_k^\mu$ is known: non-expansion metric $d(\cdot,\cdot)$

$$d(\mu_{k-1},\bar{\nu}_{k-1})/d(\mu_k,\bar{\nu}_k) \le 1$$

Not fine-grained enough: the backward denoising chain does not matter

example: current theory is insufficient

1-dim Gaussian $\mu = \mathcal{N}(0,\sigma^2)$, Wasserstein distance $W(\cdot,\cdot)$

$$W(\mu_{k-1},\bar{\nu}_{k-1})/W(\mu_k,\bar{\nu}_k) = 1 + \eta\frac{\sigma^2-1}{e^{k\eta}+\sigma^2-1}$$

$$\begin{cases} > 1, \ \textbf{(expansion)} & \text{if } \sigma^2 > 1 \\ < 1-\eta, \ \textbf{(strict contraction)} & \text{if } \sigma^2 < \frac{1}{2}, \text{ and } k < \frac{1}{2\eta}\log(2(1-\sigma^2)) \end{cases}$$

Ideal theory: the backward denoising chain should matter

technical challenge: analysis under Wasserstein metric is harder than under KL divergence S. Chen et al. (2022)

usefulness of the theory: if the message is that backward denoising chain does not matter

Why do diffusion-based generative models work?

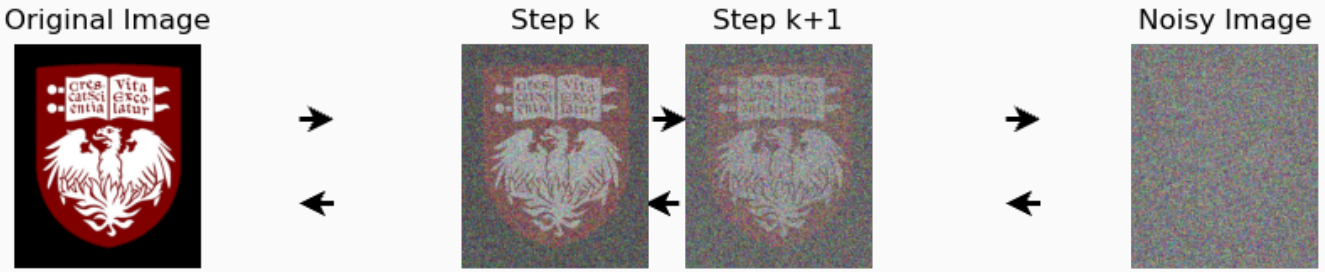
- rethink learning to sample from data distributions
- theories for probabilistic generative models

Optimal Transport Perspectives for Generative Models

a fine-grained theory on the geometric complexity of the diffuse-then-denoise process

Sensitivity Analysis: denoising diffusions and Fokker-Planck PDE

forward chain: diffusion



On data space, x: Langevin diffusion

$$\mathrm{d}\mathbf{X}_t = -\nabla f(\mathbf{X}_t)\mathrm{d}t + \sqrt{2\beta^{-1}}\mathrm{d}\mathbf{B}_t$$

Lift to distribution of data, $\mu(x)$: Fokker-Planck PDE on $\rho = \frac{\mathrm{d}\mu}{\mathrm{d}x}$

$$\partial_t \rho_t = \nabla \cdot \left(\rho_t(\nabla f + \beta^{-1}\nabla \log \rho_t)\right)$$

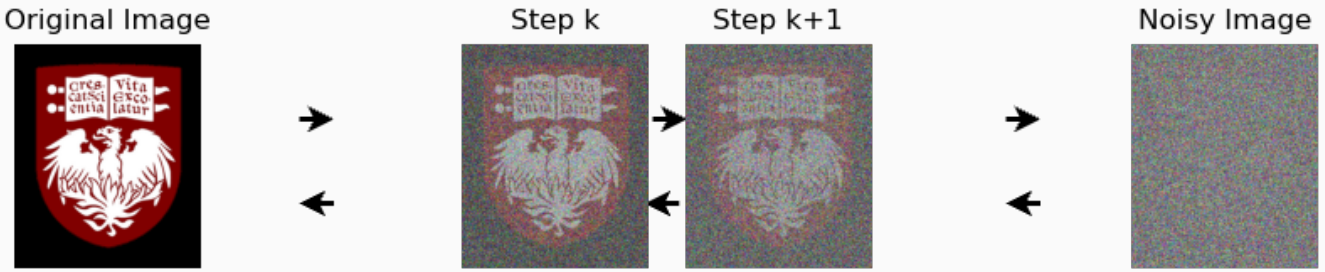
Jordan-Kinderlehrer-Otto:

$$\mu_{t+\eta} := \operatorname*{arg\,min}_{\nu \in \mathcal{P}_2^r(X)} \frac{1}{2\eta} W_2^2(\mu_t, \nu) + \mathcal{G}(\nu), \quad \mathcal{G}(\nu) := \int f \mathrm{d}\nu + \beta^{-1} \int \log\left(\tfrac{\mathrm{d}\nu}{\mathrm{d}x}\right) \mathrm{d}\nu$$

$\nabla f + \beta^{-1}\nabla \log \rho$ is the Wasserstein gradient of $\mathcal{G} : \mathcal{P}_2^r(X) \to \mathbb{R}$

Jordan, Kinderlehrer, and Otto (1998), Ambrosio, Gigli, and Savaré (2008)

forward chain: diffusion



forward step $\mathbf{X}_t \to \mathbf{X}_{t+\eta}$ is easy: add Gaussian noise

$$\mathbf{X}_{t+\eta} = (1-\eta)\mathbf{X}_t + \sqrt{2\beta^{-1}\eta}\mathbf{Z},\ \mathbf{Z} \sim \mathcal{N}(0, I)$$

How about backward step? denoising $\mathbf{X}_{t+\eta} \to \mathbf{X}_t$

backward chain: denoising

$$\mathbf{X}_{t+\eta} = (1-\eta)\mathbf{X}_t + \sqrt{2\beta^{-1}\eta}\mathbf{Z},\ \mathbf{Z} \sim \mathcal{N}(0, I)$$

Goal: denoise $\mathbf{X}_{t+\eta}$ to predict $\mathbf{X}_t$

Score function & Eddington-Tweedie formula

Consider $\mathbf{Y} = \mathbf{X} + \sigma\mathbf{Z},\ \mathbf{Z} \sim \mathcal{N}(0, I_d)$ and $\mathbf{X}, \mathbf{Z}$ are independent. Let $p_{\mathbf{Y}}$ denote the density of $\mathbf{Y}$. Then

$$\underbrace{\nabla \log p_{\mathbf{Y}}(y)}_{\text{score function}} = \frac{1}{\sigma^2}\Big\{\mathbb{E}[\mathbf{X}|\mathbf{Y} = y] - y\Big\}.$$

Score estimation?

$$\min_s\ \mathbb{E}\left[\left\|\frac{(1-\eta)\mathbf{X}_t - \mathbf{X}_{t+\eta}}{2\beta^{-1}\eta} - s(\mathbf{X}_{t+\eta})\right\|^2\right]$$

Minimzer $s(x)$ is the score function $\nabla \log p_{\mu_{t+\eta}}(x)$

Saremi and Hyvärinen (2019) and Song, Sohl-Dickstein, et al. (2020)

Score function $\nabla \log p_{t+\eta}(x)$ induces a DenoiseMap on data level
(based on conditional expectation $\mathbb{E}[\mathbf{X}_t | \mathbf{X}_{t+\eta} = x]$)

However, is the DenoiseMap exact on the distribution level?
Namely, $\text{DenoiseMap}(\mu_{t+\eta}) \overset{?}{=} \mu_t$

backward chain: denoising

Proposition (L., Koriyama and Dharmakeerthi '24)

$$\mu_{t+\eta} := \operatorname*{arg\,min}_{\nu \in \mathcal{P}_2^r(X)} \frac{1}{2\eta} W_2^2(\mu_t, \nu) + \mathcal{G}(\nu), \quad \mathcal{G}(\nu) := \int f \,\mathrm{d}\nu + \beta^{-1} \int \log\left(\frac{\mathrm{d}\nu}{\mathrm{d}x}\right) \mathrm{d}\nu \ .$$

Then any $\eta \in (0, \eta_\star)$, the **optimal transport map** $\mathbf{t}_{\mu_{t+\eta}}^{\mu_t}$ is,

$$\mathbf{t}_{\mu_{t+\eta}}^{\mu_t} = \mathbf{i} + \eta(\nabla f + \beta^{-1} \underbrace{\nabla \log p_{\mu_{t+\eta}}}_{\text{score function}}), \text{ for } \mu_{t+\eta}\text{-a.e. } x \in \mathbb{R}^d \ .$$

$$\mathbf{t}_\nu^\mu := \operatorname*{arg\,min}_{\mathbf{t} \,:\, \mathbf{t}_\# \nu = \mu} \int \|\mathbf{t}(y) - y\|^2 \mathrm{d}\nu(y) \quad \text{Monge}$$

$$W^2(\mu, \nu) := \inf_{\pi \in \Pi(\mu, \nu)} \int \|x - y\|^2 \mathrm{d}\pi(x, y) \quad \text{Kantorovich}$$

$$= \int \|\mathbf{t}_\nu^\mu - \mathbf{i}\|^2 \mathrm{d}\nu$$

Brenier (1987)

backward chain: denoising

Proposition (L., Koriyama and Dharmakeerthi '24)

$$\mu_{t+\eta} := \operatorname*{arg\,min}_{\nu \in \mathcal{P}_2^r(X)} \frac{1}{2\eta} W_2^2(\mu_t, \nu) + \mathcal{G}(\nu), \quad \mathcal{G}(\nu) := \int f \,\mathrm{d}\nu + \beta^{-1} \int \log\left(\frac{\mathrm{d}\nu}{\mathrm{d}x}\right) \mathrm{d}\nu \ .$$

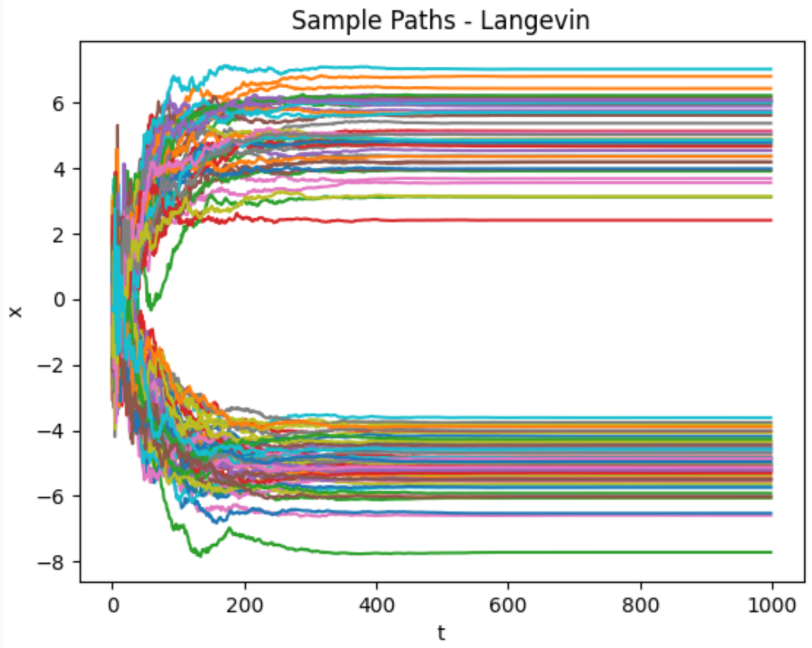
Then any $\eta \in (0, \eta_\star)$, the **optimal transport map** $\mathbf{t}_{\mu_{t+\eta}}^{\mu_t}$ is,

$$\mathbf{t}_{\mu_{t+\eta}}^{\mu_t} = \mathbf{i} + \eta(\nabla f + \beta^{-1} \underbrace{\nabla \log p_{\mu_{t+\eta}}}_{\text{score function}}), \text{ for } \mu_{t+\eta}\text{-a.e. } x \in \mathbb{R}^d \ .$$

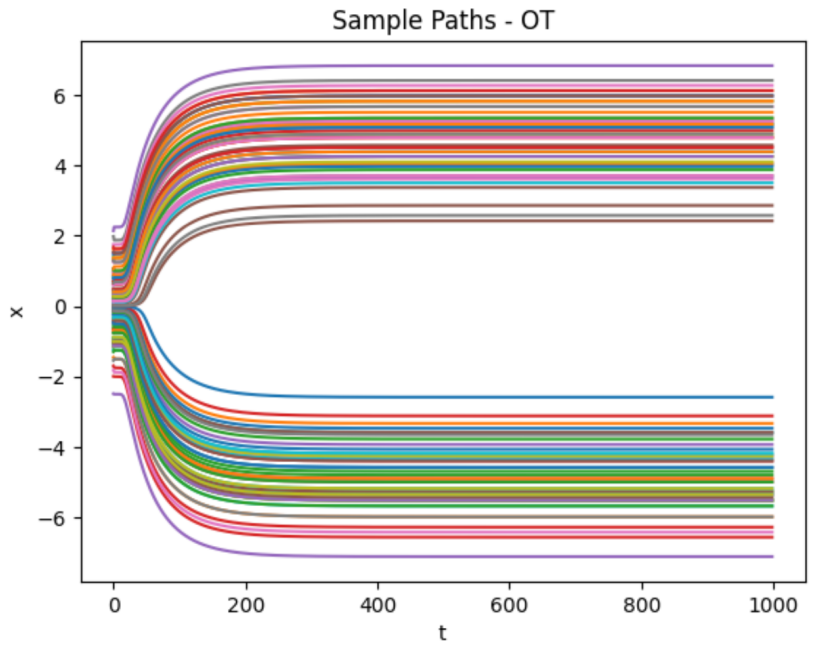
Denoising with score is exact on the distribution level, and is optimal in Wasserstein sense.

$$X_{(k-1)\eta} = X_{k\eta} + \eta \cdot \left(X_{k\eta} + \beta^{-1} \cdot \widehat{\mathbf{s}}_{k\eta}(X_{k\eta}) \right), \qquad \text{ODE}$$

$$X_{(k-1)\eta} = X_{k\eta} + \eta \cdot \left(X_{k\eta} + 2\beta^{-1} \cdot \widehat{\mathbf{s}}_{k\eta}(X_{k\eta}) \right) + \sqrt{2\beta^{-1}\eta} \cdot \mathcal{N}(0,1), \qquad \text{SDE}$$



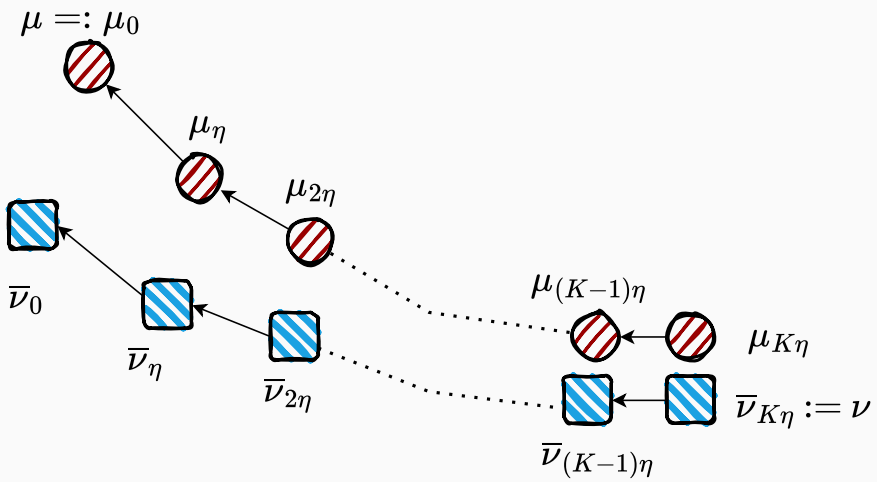
(a) Backward Transport via SDE



(b) Backward Transport via ODE

	statistical	computational	role
score function	conditional mean	optimization step	denoising

sensitivity analysis?



uncertainty in denoising?

Proposition (L., Koriyama and Dharmakeerthi '24)

Consider $\mathbf{Y} = \mathbf{X} + \sigma\mathbf{Z},\ \mathbf{Z} \sim \mathcal{N}(0, I_d)$ and $\mathbf{X}, \mathbf{Z}$ are independent. Let $p_{\mathbf{Y}}$ denote the density of $\mathbf{Y}$. Then

$$\underbrace{\nabla^2 \log p_{\mathbf{Y}}(y)}_{\text{curvature function}} = \frac{1}{\sigma^2}\Big\{ \underbrace{Cov[\tfrac{\mathbf{X}}{\sigma}|\mathbf{Y} = y]}_{\text{denoising uncertainty}} - I_d \Big\}.$$

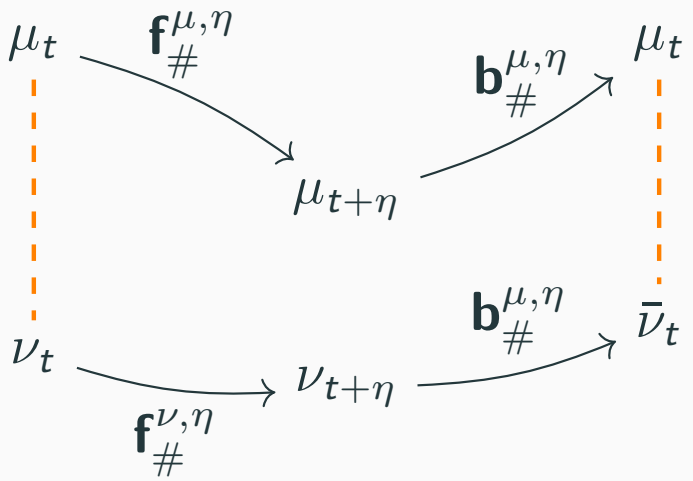
Connects optimization geometry to statistical uncertainty

Turns out crucial in sensitivity analysis

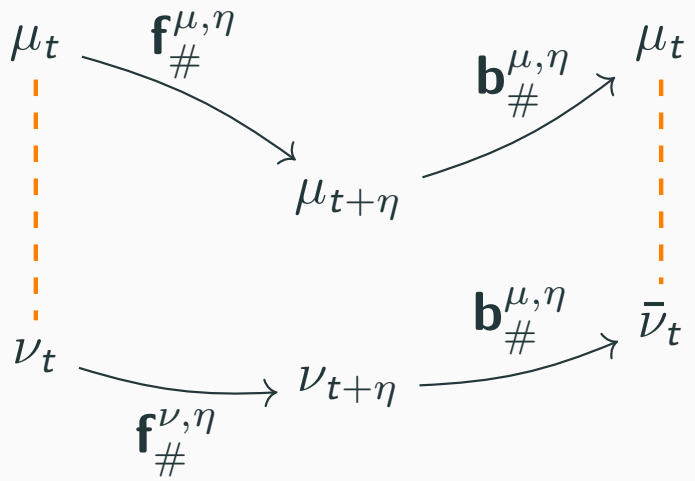
	statistical	computational	role
score function	conditional mean	optimization step	denoising
curvature function	conditional var	optimization guarantee	localization

we discover the crucial role of second-order information in the diffuse-then-denoise process

warm-up: one-step diffuse-then-denoise



warm-up: one-step diffuse-then-denoise



any gain for forward-then-backward step?

Consider $\beta^{-1} = 0$, $\mu_0 = \delta_x$ and $\nu_0 = \delta_y$, two Dirac measures

Forward $W(\mu_0, \nu_0) = \|x - y\|$, Backward $W(\mu_\eta, \bar{\nu}_\eta) = (1-\eta)\|x - y\|$,

Forward-then-Backward $W(\mu_0, \bar{\nu}_0) = (1-\eta)^{-1} W(\mu_\eta, \bar{\nu}_\eta) = \|x - y\| = W(\mu_0, \nu_0)$.

No net gain when $\beta^{-1} = 0$.

forward step

$$\mu_{t+\eta} := \mathbf{f}^{\mu,\eta}_{\#}\mu_t, \text{ where } \mathbf{f}^{\mu,\eta} = \mathbf{i} - \eta(\nabla f + \beta^{-1}\nabla \log p_{\mu_t})$$
$$\nu_{t+\eta} := \mathbf{f}^{\nu,\eta}_{\#}\nu_t, \text{ where } \mathbf{f}^{\nu,\eta} = \mathbf{i} - \eta(\nabla f + \beta^{-1}\nabla \log p_{\nu_t})$$

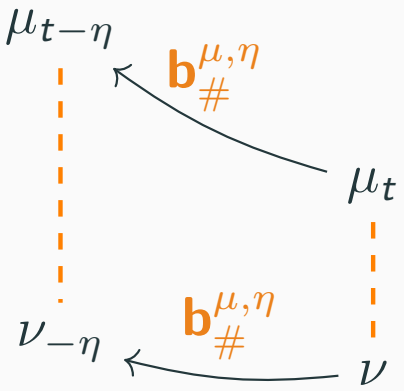
forward contraction

Assume for some $\lambda \in \mathbb{R}_+$, $\nabla^2 f(x) \succeq \lambda \cdot I_d, \ \forall x \in X$.

One-step forward process satisfies

$$\limsup_{\eta \to 0} \frac{1}{\eta} \frac{W_2^2(\mu_{t+\eta}, \nu_{t+\eta}) - W_2^2(\mu_t, \nu_t)}{W_2^2(\mu_t, \nu_t)} \leq -2\lambda \ .$$

backward step: sensitivity



$$\mu_{t-\eta} := \mathbf{b}^{\mu,\eta}_{\#}\mu_t,$$
$$\nu_{-\eta} := \mathbf{b}^{\mu,\eta}_{\#}\nu,$$
$$\text{where } \mathbf{b}^{\mu,\eta} = \mathbf{i} + \eta(\nabla f + \beta^{-1}\nabla \log p_{\mu_t})$$

Theorem (L., Koriyama and Dharmakeerthi '24)

Assume for some $\lambda \in \mathbb{R}_+$, $\nabla^2 f(x) \preceq \lambda \cdot I_d$, and for some $\zeta \in \mathbb{R}$

$$\nabla^2 \log p_{\mu_t}(x) \preceq -\zeta \cdot I_d, \ \forall x \in X \ .$$

One-step backward map satisfies

$$\sup_{\nu \in \mathcal{P}_2^r(X)} \frac{1}{\eta} \frac{W_2^2(\mathbf{b}^{\mu,\eta}_{\#}\mu_t, \mathbf{b}^{\mu,\eta}_{\#}\nu) - W_2^2(\mu_t, \nu)}{W_2^2(\mu_t, \nu)} \leq 2(\lambda - \beta^{-1}\zeta) \ .$$

The inequality is sharp when μ_t is Gaussian.

diffuse-then-denoise process: one-step

Consider Ornstein–Uhlenbeck process where $f(x) = \|x\|^2/2$.

Forward contraction:

$$\frac{W_2^2(\mu_{t+\eta}, \nu_{t+\eta})}{W_2^2(\mu_t, \nu_t)} \le 1 - \underline{2\eta} + O(\eta^2) \ .$$

Backward possible expansion:

$$\frac{W_2^2(\mu_t, \bar{\nu}_t)}{W_2^2(\mu_{t+\eta}, \nu_{t+\eta})} \le 1 + \underline{2\eta} - \boxed{2\eta\beta^{-1}\zeta} + O(\eta^2) \ .$$

diffuse-then-denoise process: one-step

Consider Ornstein–Uhlenbeck process where $f(x) = \|x\|^2/2$.

Forward contraction:

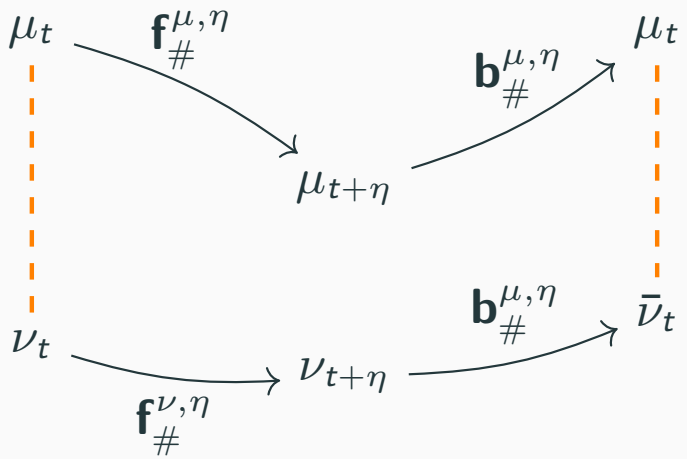
$$\frac{W_2^2(\mu_{t+\eta}, \nu_{t+\eta})}{W_2^2(\mu_t, \nu_t)} \le 1 - \underline{2\eta} + O(\eta^2) \,.$$

Backward possible expansion:

$$\frac{W_2^2(\mu_t, \bar{\nu}_t)}{W_2^2(\mu_{t+\eta}, \nu_{t+\eta})} \le 1 + \underline{2\eta} - \boxed{2\eta\beta^{-1}\zeta} + O(\eta^2) \,.$$

Net gain of diffuse-then-denoise:

$$\frac{W_2^2(\mu_t, \mathbf{b}_\#^{\mu,\eta}\mathbf{f}_\#^{\nu,\eta}\nu_t)}{W_2^2(\mu_t, \nu_t)} \le 1 - 2\eta\beta^{-1}\zeta + O(\eta^2) \,.$$



curvature $\nabla^2 \log p_{\mu_t}$: **net effect** of diffuse-then-denoise process

curvature at all scales: chaining

Theorem (L., Koriyama and Dharmakeerthi '24), chaining

Define the diffuse-then-denoise map

$$\mathbf{f}^{\nu}_{[K]} := \mathbf{f}^{\nu,\eta}_{K} \circ \cdots \circ \mathbf{f}^{\nu,\eta}_{2} \circ \mathbf{f}^{\nu,\eta}_{1} \ ,$$
$$\mathbf{b}^{\mu}_{[K]} := \mathbf{b}^{\mu,\eta}_{1} \circ \mathbf{b}^{\mu,\eta}_{2} \circ \cdots \circ \mathbf{b}^{\mu,\eta}_{K} \ .$$

Assume there exists a sequence of $\zeta_{k\eta} \in \mathbb{R}$ such that

$$\nabla^2 \log p_{\mu_{k\eta}}(x) \preceq -\zeta_{k\eta} \cdot I_d, \ \forall x \in X, \ \forall k = 1, 2, \ldots, K \ .$$

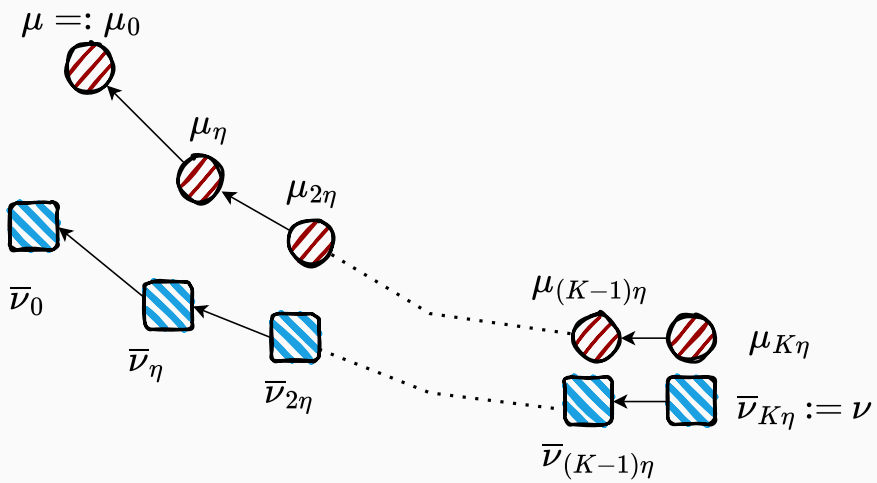
Then the diffuse-then-denoise process on ν with fixed K steps satisfies

$$\frac{W_2^2\left(\mu, \left(\mathbf{b}^{\mu}_{[K]} \circ \mathbf{f}^{\nu}_{[K]}\right)_{\#}\nu\right)}{W_2^2(\mu,\nu)} \leq \exp\left(-2\eta\beta^{-1}\sum_{k=1}^{K}\zeta_{k\eta} + O(\eta^2)\right) .$$

our theory: new insights

This paper: effectiveness of diffuse-then-denoise process

- Net gain of diffuse-then-denoise process: $\beta^{-1} \neq 0$
- Curvature at all scales matters
- If log-concave, $\zeta_{k\eta} > 0$ across all time scales k



our theory: new insights

This paper: effectiveness of diffuse-then-denoise process

- Net gain of diffuse-then-denoise process: $\beta^{-1} \neq 0$
- Curvature at all scales matters
- If log-concave, $\zeta_{k\eta} > 0$ across all time scales k
- Beyond log-concavity: worst-case curvature $\rightarrow$ average notion of curvature

multi-scale complexity: beyond log-concavity

Interlude: thoughts on generative models

discussion and future work

	hard to sample	easy to sample
hard to learn	avoid	generative adversarial networks
easy to learn	diffusion-based models	?

Static analysis: Empirical Processes and Optimal Transport

- Generative adversarial networks: one-step push-forward, complexity tradeoff between generator and discriminator

  Liang (2021) JMLR

- Optimal transport and generative models: (reversible)-Gromov-Wasserstein, forward and backward maps, isomorphism, cycle-consistency

  Hur, Guo, and Liang (2024) SIMODS

Dynamic analysis: Partial Differential Equations and Optimal Transport

- Fokker-Planck PDE: probabilistic diffusion models, forward and backward chains, curvature matters

  Liang, Dharmakeerthi, and Koriyama (2024)

- Monge-Ampère PDE: leverage curvature to design a new chain/flow, regret analysis via new evolution variational inequality

  Deb and Liang (2025)

  Guo, Hur, Liang, and Ryan (2022) COLT

No-Regret Analysis: Sinkhorn algorithm and Monge-Ampère PDE

OT and Monge-Ampère PDE

- Given target $\frac{\mathrm{d}\mu}{\mathrm{d}x} = e^{-f}$, and reference $\frac{\mathrm{d}\nu}{\mathrm{d}x} = e^{-g}$
  Find a convex function $\psi : \mathbb{R}^d \to \mathbb{R}$

$$(\nabla\psi)_{\#}\nu = \mu$$

- (Static) Monge-Ampère equation:

$$-f(\nabla\psi(y)) + g(y) + \log\det\big(\nabla^2\psi(y)\big) = 0\,.$$

- Parabolic Monge-Ampère equation:

$$\frac{\partial\psi_t}{\partial t}(y) = -f(\nabla\psi_t(y)) + g(y) + \log\det(\nabla^2\psi_t(y))\,.$$

  continuous time limit of Sinkhorn algorithm

new EVI

Define the density at step k as $\rho_k := (\nabla\psi_k)\#e^{-g}$ and consider

$$\frac{\nabla\psi_{k+1} - \nabla\psi_k}{\eta_k} = -\boldsymbol{\xi}_k, \quad \text{where} \quad \boldsymbol{\xi}_k := \nabla\big(\log\big(\rho_k/e^{-f}\big) \circ \nabla\psi_k\big)\,.$$

Theorem (Deb and L. '25), evolution variational inequality

Assume that $m_k := \inf_y \nabla^2\psi_k(y) > 0$. Given any probability density π, let π_k denote pdf of $(\nabla\psi_k^*)\#\pi$. Then

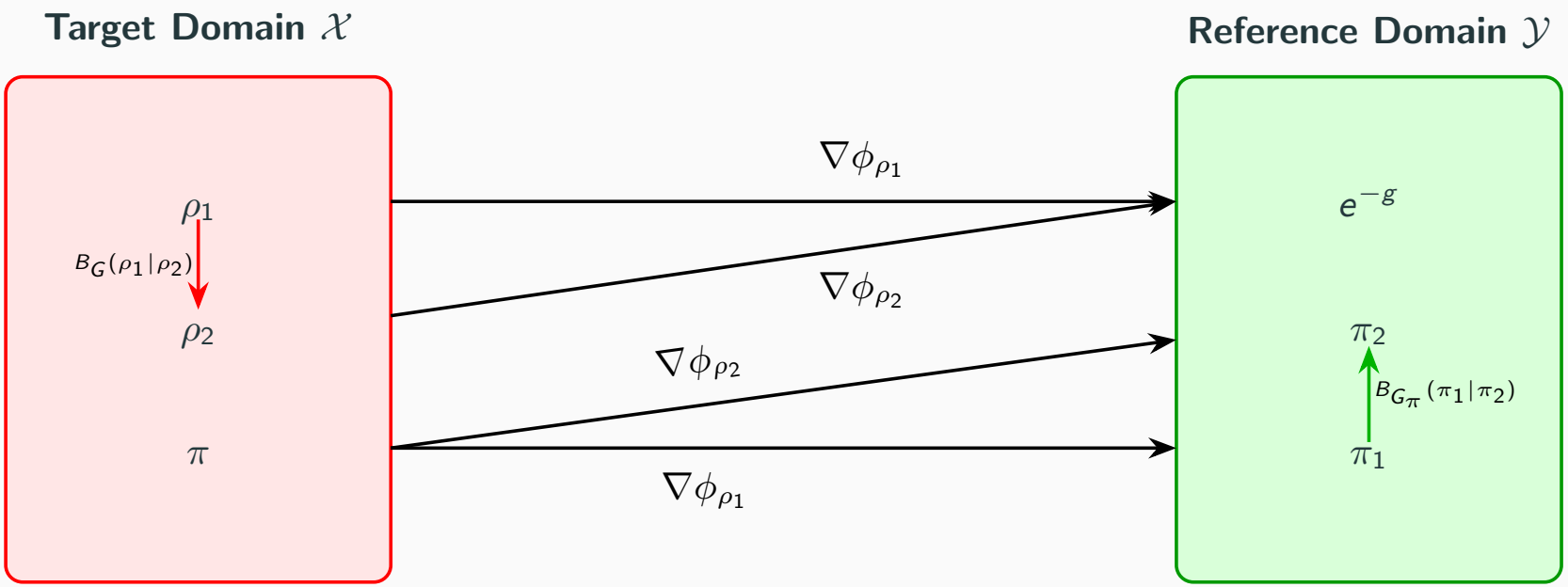
$$\begin{aligned}
&KL(\rho_k|e^{-f}) - KL(\pi|e^{-f}) \\
&= \boxed{\frac{B_G(\pi|\rho_k) - B_G(\pi|\rho_{k+1})}{\eta_k} + \frac{B_{G_\pi}(\pi_k|\pi_{k+1})}{\eta_k}} - KL(\pi|\rho_k) \\
&\le \frac{B_G(\pi|\rho_k) - B_G(\pi|\rho_{k+1})}{\eta_k} + \frac{1}{2}\frac{\eta_k}{m_{k+1}}\int \|\boldsymbol{\xi}_k\|^2 \mathrm{d}\pi_k - KL(\pi|\rho_k)\,.
\end{aligned}$$

Guo, Hur, Liang, and Ryan (2022)

new EVI

Theorem (Deb and L. '25), evolution variational inequality

$$
\begin{aligned}
& KL(\rho_k|e^{-f}) - KL(\pi|e^{-f}) \\
= & \boxed{\frac{B_G(\pi|\rho_k) - B_G(\pi|\rho_{k+1})}{\eta_k} + \frac{B_{G_\pi}(\pi_k|\pi_{k+1})}{\eta_k}} - KL(\pi|\rho_k) \\
\leq & \frac{B_G(\pi|\rho_k) - B_G(\pi|\rho_{k+1})}{\eta_k} + \frac{1}{2}\frac{\eta_k}{m_{k+1}} \int \|\xi_k\|^2 \mathrm{d}\pi_k - KL(\pi|\rho_k) \, .
\end{aligned}
$$



regret to convergence

Theorem (Deb and L. '25), convergence

Average-iterate convergence:

- Constant time-invariant stepsize $\eta_k \equiv \eta$,

$$KL(\bar{\rho}_T | e^{-f}) = O(\tfrac{1}{T})$$

regret to convergence

Theorem (Deb and L. '25), convergence

Last-iterate convergence:

- Time-invariant stepsize $\eta_k \equiv \frac{1}{\sqrt{T}}$, assume $\inf_y \nabla^2 \psi_k(y) \geq m \cdot I_d$,

$$KL(\rho_T | e^{-f}) = O(\tfrac{1}{\sqrt{T}})$$

- Adaptive stepsize $\eta_k = \frac{\lambda}{k+1}$, assume in addition $\sup_y \nabla^2 \psi_k(y) \leq M \cdot I_d$, and reference $g(\cdot)$ λ-strongly convex,

$$KL(\rho_T | e^{-f}) = O(\tfrac{\log(T)}{T})$$

- Time-invariant stepsize $\eta_k \equiv \frac{\log(T)}{T}$,

$$B_G(e^{-f} | \rho_T) = O(\tfrac{\log(T)}{T})$$

Monge-Ampère neural-PDE sampler

Algorithm 1: Monge-Ampère Neural-PDE via Logistic Regression

Input : T, total number of steps; $\{\eta_k\}_{k\leq T}$, the step-sizes; initialize a neural network function $\psi_0 : \mathbb{R}^d \to \mathbb{R}$, and set $k = 0$.

Output: A neural network function $\psi_T : \mathbb{R}^d \to \mathbb{R}$, and samples from $(\nabla\psi_T)\#e^{-g}$.

while $k < T$ **do**

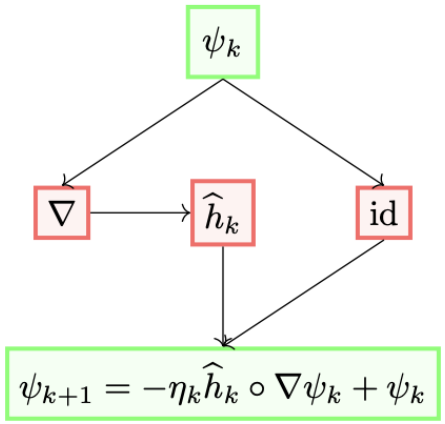
Sampling step: Given i.i.d. data $X_1, \cdots, X_n \sim e^{-f}$, sample $\tilde{X}_1, \cdots, \tilde{X}_n \sim \rho_k = (\nabla\psi_k)\#e^{-g}$, augment to $\{(X_i, L_i = 0)\}_{i=1}^n \cup \{(\tilde{X}_i, \tilde{L}_i = 1)\}_{i=1}^n$, and form its empirical distribution as $(X, L) \sim \widehat{\gamma}_k$;

Learning step: Estimate a neural network discriminator function $\widehat{h}_k$ as in (5.1) ;

Neural Network update: Define a new neural network with the residual architecture shown on the right $\psi_{k+1} = -\eta_k \widehat{h}_k \circ \nabla\psi_k + \psi_k$;

Set $k \leftarrow k + 1$;

end



Monge-Ampère neural-PDE sampler

Algorithm 2: Monge-Ampère Neural-PDE via Score Matching

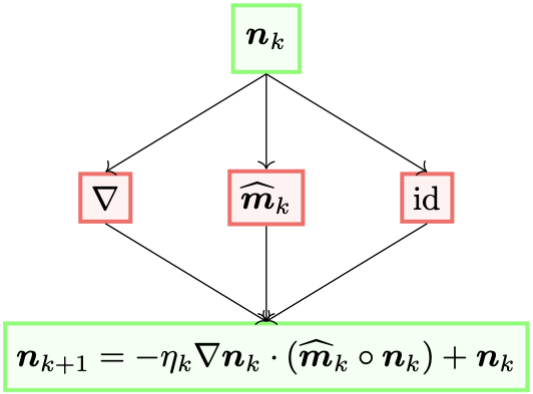
Input : T, total number of steps; $\{\eta_k\}_{k\leq T}$, the step-sizes; initialize a vector-valued neural network $\boldsymbol{n}_0 : \mathbb{R}^d \to \mathbb{R}^d$, and set $k = 0$.

Output: A vector-valued neural network function $\boldsymbol{n}_T : \mathbb{R}^d \to \mathbb{R}^d$, and samples from $(\boldsymbol{n}_T)\#e^{-g}$.

while $k < T$ **do**

- **Sampling step**: Obtain i.i.d. samples from $\rho_k = (\boldsymbol{n}_k)\#e^{-g}$, form the empirical measure $\widehat{\rho}_k$;
- **Learning step**: Estimate a neural network score $\widehat{\boldsymbol{\sigma}}_k$ as in (5.2), and define the corresponding $\widehat{\boldsymbol{m}}_k$ as in (5.3) ;
- **Neural Network update**: Define a new neural network with the residual architecture shown on the right $\boldsymbol{n}_{k+1} = -\eta_k \nabla \boldsymbol{n}_k \cdot (\widehat{\boldsymbol{m}}_k \circ \boldsymbol{n}_k) + \boldsymbol{n}_k$, here $\nabla \boldsymbol{n}_k : \mathbb{R}^d \to \mathbb{S}^{d\times d}_+$, and $\widehat{\boldsymbol{m}}_k \circ \boldsymbol{n}_k : \mathbb{R}^d \to \mathbb{R}^d$ and '$\cdot$' denotes the matrix-vector product ;
- Set $k \leftarrow k + 1$;

end



Thank you!

Tengyuan Liang, Kulunu Dharmakeerthi, and Takuya Koriyama (2024). Denoising Diffusions with Optimal Transport: Localization, Curvature, and Multi-Scale Complexity. arXiv:2411.01629

Nabarun Deb, and Tengyuan Liang (2025). No-Regret Generative Modeling via Parabolic Monge-Ampère PDE. arXiv:2504.09279

multi-scale complexity: beyond log-concavity

SNR, localization, and tail

Definition (L., Koriyama and Dharmakeerthi '24), integrated tail

Given a measure $\mu \in \mathcal{P}_2(\mathbb{R}^d)$, define for any SNR $\mathsf{r} \in \mathbb{R}_{\geq 0}$

$$\mathbf{Y}_{\mathsf{r}} := \mathsf{r}\mathbf{X} + \mathbf{Z}, \ (\mathbf{X}, \mathbf{Z}) \sim \mu \otimes \mathcal{N}(0, I_d)$$

SNR, localization, and tail

Definition (L., Koriyama and Dharmakeerthi '24), integrated tail

Given a measure $\mu \in \mathcal{P}_2(\mathbb{R}^d)$, define for any SNR $\mathsf{r} \in \mathbb{R}_{\geq 0}$

$$\mathbf{Y}_\mathsf{r} := \mathsf{r}\mathbf{X} + \mathbf{Z},\ (\mathbf{X}, \mathbf{Z}) \sim \mu \otimes \mathcal{N}(0, I_d)$$

Define the localization function, and the associated random variable

$$L_\mathsf{r}(y) = \| \mathit{Cov}[\mathsf{r}\mathbf{X}|\mathbf{Y}_\mathsf{r} = y]\|_{op}\ ,\ \text{and}\ \mathbf{L}_\mathsf{r} = \| \mathit{Cov}[\mathsf{r}\mathbf{X}|\mathbf{Y}_\mathsf{r}]\|_{op}\ .$$

and denote the survival function of $\mathbf{L}_\mathsf{r}$ as $s_\mathsf{r}(\cdot) : \mathbb{R}_+ \to [0, 1]$

$$s_\mathsf{r}(u) := \mathbb{P}(\mathbf{L}_\mathsf{r} > u)\ .$$

SNR, localization, and tail

Definition (L., Koriyama and Dharmakeerthi '24), integrated tail

Given a measure $\mu \in \mathcal{P}_2(\mathbb{R}^d)$, define for any SNR $\mathsf{r} \in \mathbb{R}_{\geq 0}$

$$\mathbf{Y}_\mathsf{r} := \mathsf{r}\mathbf{X} + \mathbf{Z},\ (\mathbf{X}, \mathbf{Z}) \sim \mu \otimes \mathcal{N}(0, I_d)$$

Define the localization function, and the associated random variable

$$L_\mathsf{r}(y) = \| \mathit{Cov}[\mathsf{r}\mathbf{X}|\mathbf{Y}_\mathsf{r} = y]\|_{op}\ ,\ \text{and}\ \mathbf{L}_\mathsf{r} = \| \mathit{Cov}[\mathsf{r}\mathbf{X}|\mathbf{Y}_\mathsf{r}]\|_{op}\ .$$

and denote the survival function of $\mathbf{L}_\mathsf{r}$ as $s_\mathsf{r}(\cdot) : \mathbb{R}_+ \to [0, 1]$

$$s_\mathsf{r}(u) := \mathbb{P}(\mathbf{L}_\mathsf{r} > u)\ .$$

integrated tail: $h_\mu(\delta, \mathsf{r}) := \int_{1-\delta}^{\infty} s_\mathsf{r}(u)\mathrm{d}u\ .$

multi-scale SNR

Ornstein-Uhlenbeck Process with initialization $X_0 \sim \mu$, and potential $f(x) = \|x\|^2/2$

$$\mathrm{d}\mathbf{X}_t = -\mathbf{X}_t \mathrm{d}t + \sqrt{2\beta^{-1}}\mathrm{d}\mathbf{B}_t \ .$$

Then the distribution of $\mathbf{X}_t, \forall t \in \mathbb{R}_+$

$$\mathbf{X}_t \overset{\mathcal{L}}{\sim} e^{-t}\mathbf{X}_0 + \sqrt{\beta^{-1}(1-e^{-2t})}\mathbf{Z}, \ \mathbf{Z} \sim \mathcal{N}(0, I_d) \ .$$

multi-scale SNR

Ornstein-Uhlenbeck Process with initialization $X_0 \sim \mu$, and potential $f(x) = \|x\|^2/2$

$$\mathrm{d}\mathbf{X}_t = -\mathbf{X}_t \mathrm{d}t + \sqrt{2\beta^{-1}}\mathrm{d}\mathbf{B}_t \ .$$

Then the distribution of $\mathbf{X}_t, \forall t \in \mathbb{R}_+$

$$\mathbf{X}_t \overset{\mathcal{L}}{\sim} e^{-t}\mathbf{X}_0 + \sqrt{\beta^{-1}(1-e^{-2t})}\mathbf{Z}, \ \mathbf{Z} \sim \mathcal{N}(0, I_d) \ .$$

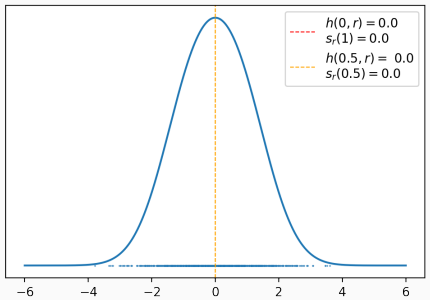
multi-scale SNR

$$\underbrace{\nabla^2 \log p_{\mu_t}(x)}_{\text{curvature function}} = -\frac{1}{\mathrm{s}^2(t)} \left\{ I_d - Cov[\mathrm{r}(t)\mathbf{X}_0 | \mathbf{X}_t = x] \right\}$$

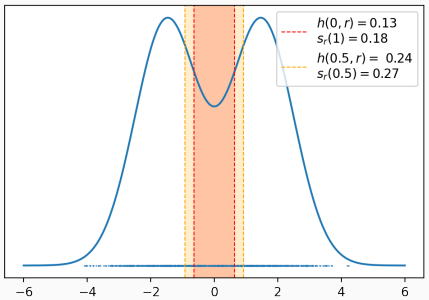
$$\mathrm{r}(t) := \frac{e^{-t}}{\sqrt{\beta^{-1}(1-e^{-2t})}}, \ \mathrm{s}(t) := \frac{1}{\sqrt{\beta^{-1}(1-e^{-2t})}}$$

simplest non-log-concave example: intuition

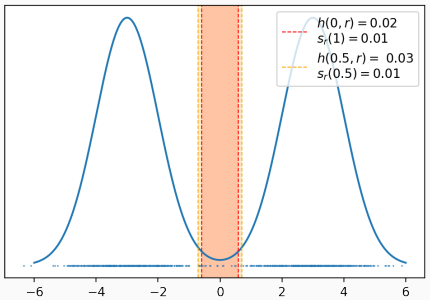
$$\mu = \tfrac{1}{2}\delta_{-1} + \tfrac{1}{2}\delta_{-1}$$



(a) r = 0.71



(b) r = 1.50

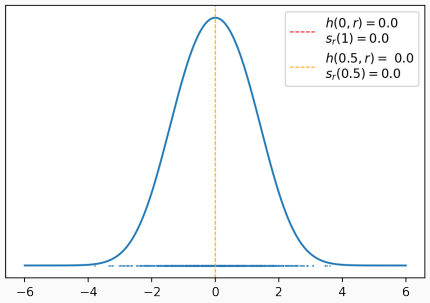


(c) r = 3.00

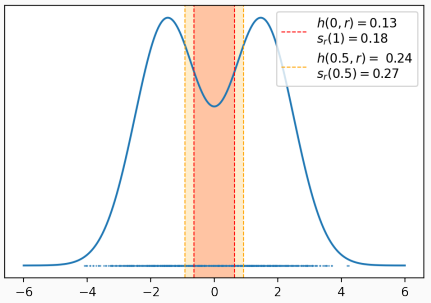
We plot the density $p_{\mathbf{Y}_r}(\cdot)$, for three SNR r's. Red shaded area corresponds to non-log-concave region with $\nabla^2 \log p_{\mathbf{Y}_r}(\cdot) > -\delta$ with $\delta = 0$, and Orange shaded area corresponds to $\delta = 0.5$.

simplest non-log-concave example: intuition

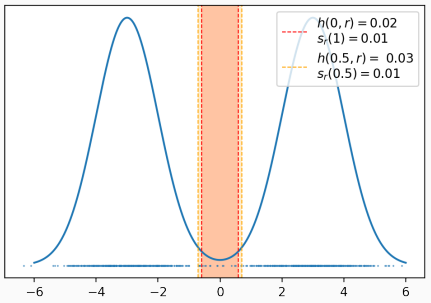
$$\mu = \tfrac{1}{2}\delta_{-1} + \tfrac{1}{2}\delta_{-1}$$



(a) r = 0.71



(b) r = 1.50



(c) r = 3.00

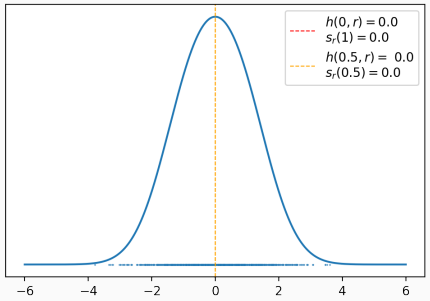
Localization and Curvature: $L_r(y) \leq 1 - \delta$ iff $\nabla^2 \log p_{\mathbf{Y}_r}(y) \preceq -\delta \cdot I_d$, namely, $p_{\mathbf{Y}_r}(y)$ strongly log-concave at y.

These y's are good locations with strong negative curvature: accurate denoising/localization from y.

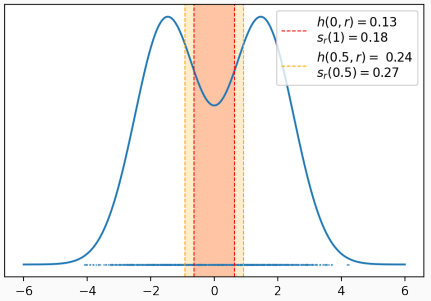
Locations where diffuse-then-denoise is beneficial.

simplest non-log-concave example: intuition

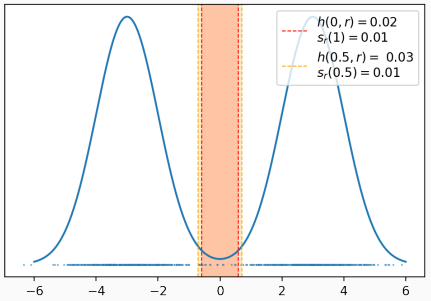
$$\mu = \tfrac{1}{2}\delta_{-1} + \tfrac{1}{2}\delta_{-1}$$



(a) r = 0.71



(b) r = 1.50



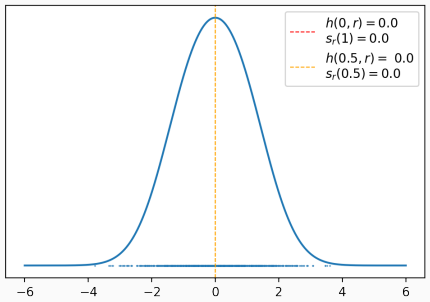
(c) r = 3.00

Survival Function: $s_r(1-\delta) := \mathbb{P}\big(L_r(\mathbf{Y}_r) > 1-\delta\big)$ tells us the probability of bad locations with possibly non-log-concavity where the backward denoising is hard.

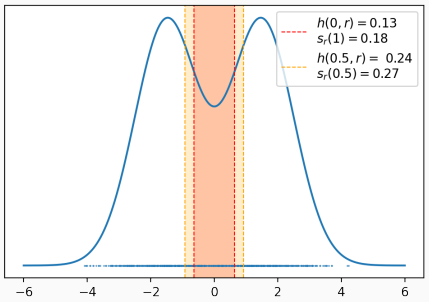
Mass that may induce a large expansion in the diffuse-then-denoise process.

simplest non-log-concave example: intuition

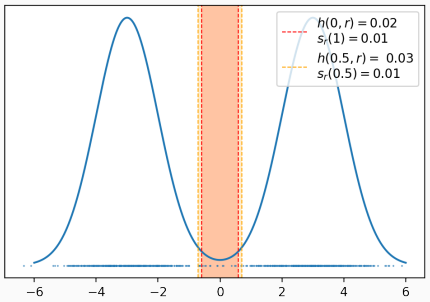
$$\mu = \tfrac{1}{2}\delta_{-1} + \tfrac{1}{2}\delta_{-1}$$



(a) r = 0.71



(b) r = 1.50



(c) r = 3.00

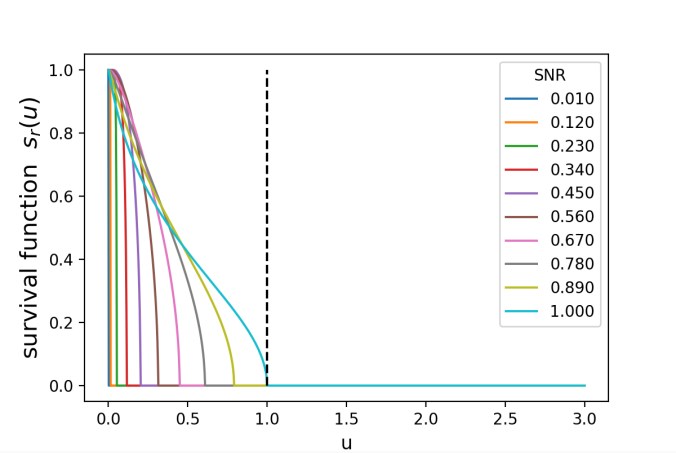
Integrated Tail: slow growth in the integrated tail function $\delta \to h_\mu(\delta, \mathrm{r})$ implies that one can take an effectively large δ

- bad locations with positive curvatures $\Rightarrow$ a negligible expansion effect
- good locations with negative curvature $\Rightarrow$ induce a contraction effect to offset

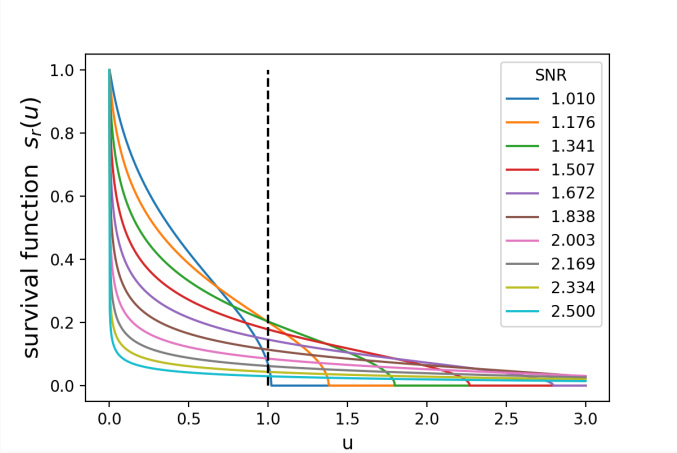
This complexity quantifies an overall notion of curvature.

simplest non-log-concave example: intuition

$$\mu = \tfrac{1}{2}\delta_{-1} + \tfrac{1}{2}\delta_{-1}$$



(a) $s_r(u)$ Low SNR



(b) $s_r(u)$ High SNR

Question: what's the hardest SNR?

multi-scale complexity: theory for non-log-concave

Theorem (L., Koriyama and Dharmakeerthi '24), non-log-concave

For any $\delta \in [0,1]$, as $\eta \to 0$,

$$\limsup_{\nu \in \mathcal{M}(\mu_t, M): \nu \overset{W_2}{\to} \mu_t} \frac{1}{\eta} \frac{W_2^2(\mathbf{b}_\#^{\mu,\eta}\mu_t, \mathbf{b}_\#^{\mu,\eta}\nu) - W_2^2(\mu_t, \nu)}{W_2^2(\mu_t, \nu)}$$
$$\leq 2 - \frac{2}{1-e^{-2t}} \Big[\underbrace{\delta}_{\text{contraction}} - M \cdot \underbrace{h_\mu(\delta, \mathsf{r}(t))}_{\text{integrated tail}} \Big] .$$

multi-scale complexity: theory for non-log-concave

Theorem (L., Koriyama and Dharmakeerthi '24), non-log-concave

For any $\delta \in [0,1]$, as $\eta \to 0$,

$$\limsup_{\nu \in \mathcal{M}(\mu_t, M): \nu \overset{W_2}{\to} \mu_t} \frac{1}{\eta} \frac{W_2^2(\mathbf{b}_{\#}^{\mu,\eta}\mu_t, \mathbf{b}_{\#}^{\mu,\eta}\nu) - W_2^2(\mu_t, \nu)}{W_2^2(\mu_t, \nu)}$$
$$\leq 2 - \frac{2}{1-e^{-2t}} \Big[\underbrace{\delta}_{\text{contraction}} - M \cdot \underbrace{h_\mu(\delta, \mathsf{r}(t))}_{\text{integrated tail}} \Big] .$$

$$\mathcal{M}(\mu, M) := \{\nu \in \mathcal{P}_2^r(X) \ : \ \frac{\sup_{x \in \mathrm{Dom}(\mu)} \|(\mathbf{t}_\mu^\nu - \mathbf{i})(x)\|^2}{\int \|(\mathbf{t}_\mu^\nu - \mathbf{i})(x)\|^2 \mathrm{d}\mu} \leq M\} .$$

Here M controls the richness of perturbations around μ.

multi-scale complexity: theory for non-log-concave

Theorem (L., Koriyama and Dharmakeerthi '24), non-log-concave

For any $\delta \in [0,1]$, as $\eta \to 0$,

$$\limsup_{\nu \in \mathcal{M}(\mu_t, M): \nu \overset{W_2}{\to} \mu_t} \frac{1}{\eta} \frac{W_2^2(\mathbf{b}_{\#}^{\mu,\eta}\mu_t, \mathbf{b}_{\#}^{\mu,\eta}\nu) - W_2^2(\mu_t, \nu)}{W_2^2(\mu_t, \nu)}$$
$$\leq 2 - \frac{2}{1-e^{-2t}} \Big[\underbrace{\delta}_{\text{contraction}} - M \cdot \underbrace{h_\mu(\delta, \mathsf{r}(t))}_{\text{integrated tail}} \Big] .$$

probe the multi-scale complexity: empirics

empirics

probing the multi-scale complexity

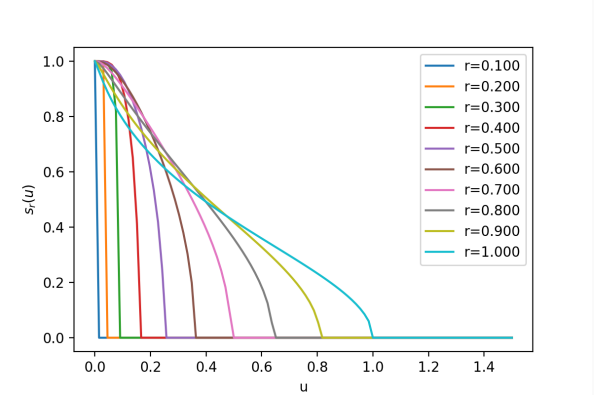
$$
\begin{aligned}
s_{\mathsf{r}}(u) &= \mathbb{P}(\mathbf{L}_{\mathsf{r}} > u), \\
m^*(\mathsf{r}) &= \min_{\delta \in [0,1]} \frac{h_\mu(\delta, \mathsf{r})}{\delta}, \\
\zeta_M^*(t) &= \sup_{\delta \in [0,1]} \tfrac{1}{1-e^{-2t}} \left[\delta - M \cdot h_\mu(\delta, \mathsf{r}(t))\right].
\end{aligned}
$$

Actionable theory: given a target distribution, empirical or with explicit form, we can visualize the functions above.

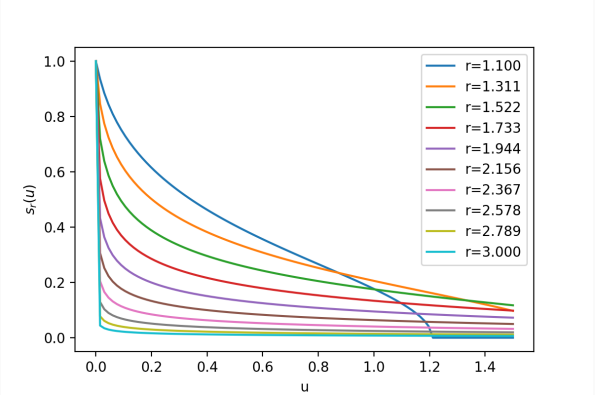
- For any empirical measure μ_0, a analytic expression for the empirical version of $\mathbf{L}_r(y)$.
- Simulate $\mathbf{Y}_{\mathsf{r}}$ by sampling $\mathsf{r}\mathbf{X} + \mathcal{N}(0,1)$, for $\mathbf{X} \sim \mu_0$.
- Obtain the empirical survival function, $s_{\mathsf{r}}(u)$. $m^*(\mathsf{r})$, and $\zeta_M^*(t)$.

non-log-concave example 1: two-point mass

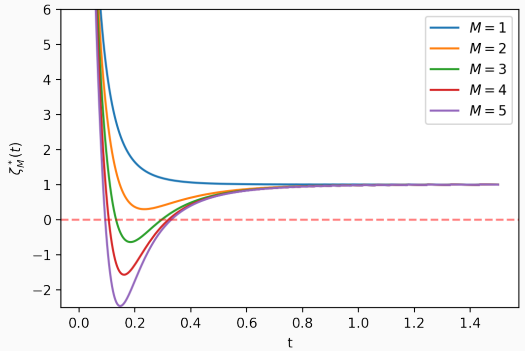
$$X_0 \sim \tfrac{1}{2}\delta_{-\mu} + \tfrac{1}{2}\delta_{\mu} \text{ for some } \mu > 0$$



(a)



(b)

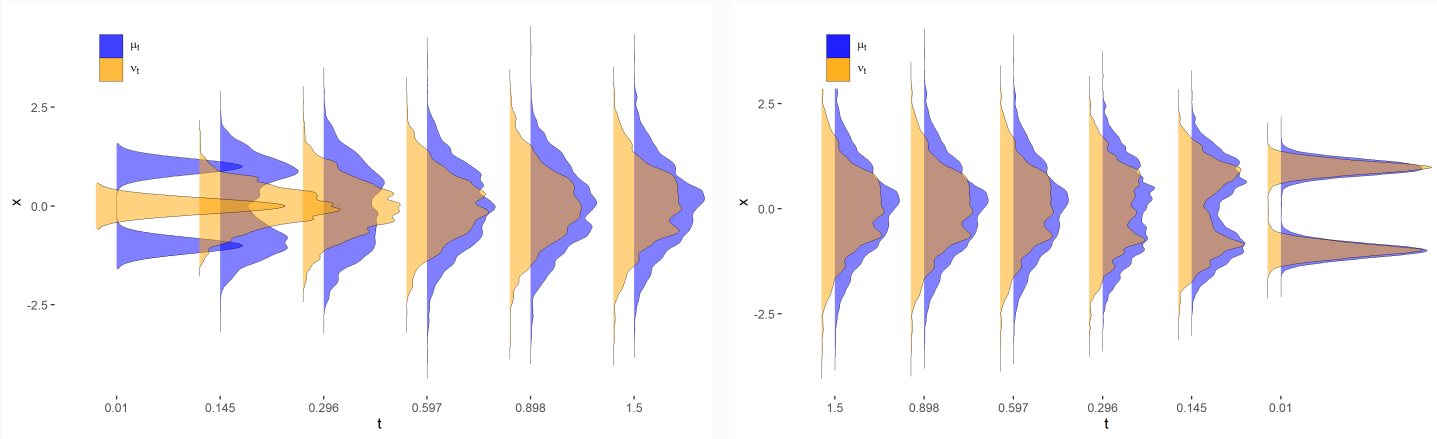


(c)

(a) $s_r(u)$ Low SNR. (b) $s_r(u)$ High SNR. (c) $\zeta^*_M(t)$.

non-log-concave example 1: two-point mass

$$X_0 \sim \tfrac{1}{2}\delta_{-\mu} + \tfrac{1}{2}\delta_{\mu} \text{ for some } \mu > 0$$

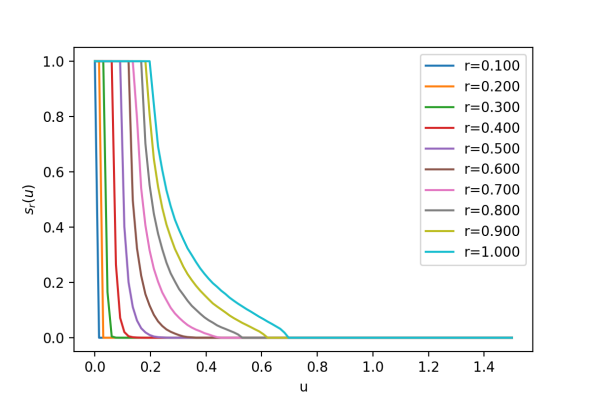


(a) Forward Diffusion via OU **(b)** Backward Transport via OT

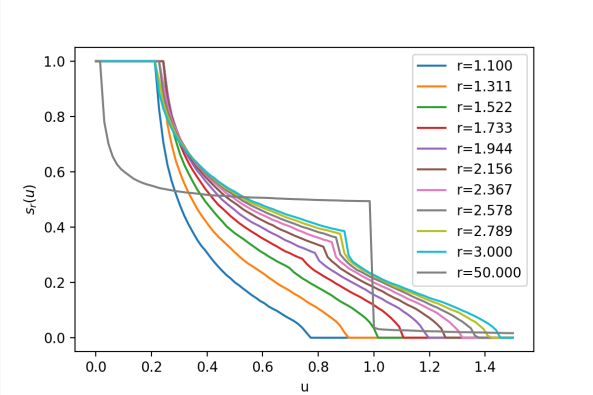
(a) Forward diffusion initialized at $\mu_0 = 0.5\delta_1 + 0.5\delta_{-1}$, $\nu_0 = \delta_0$. $T = 100$, $\eta = 0.01$. (b) Backward transport initialized at μ_T, ν_T, and applying the backward OT map $\mathbf{b}^{\mu,\eta}$.

non-log-concave example 2: mixture of point mass and Gaussian

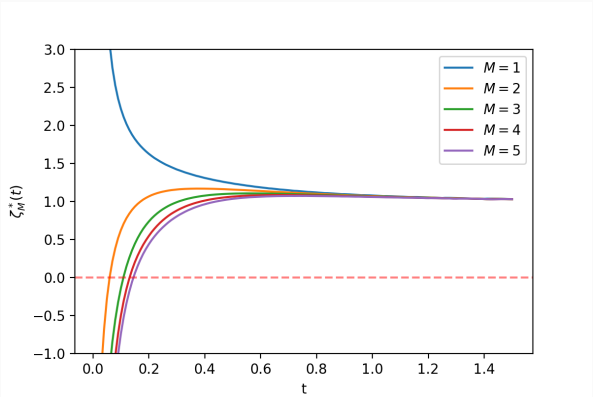
$$X_0 \sim \tfrac{1}{2}\delta_0 + \tfrac{1}{2}\mathcal{N}(0,1)$$



(a)



(b)

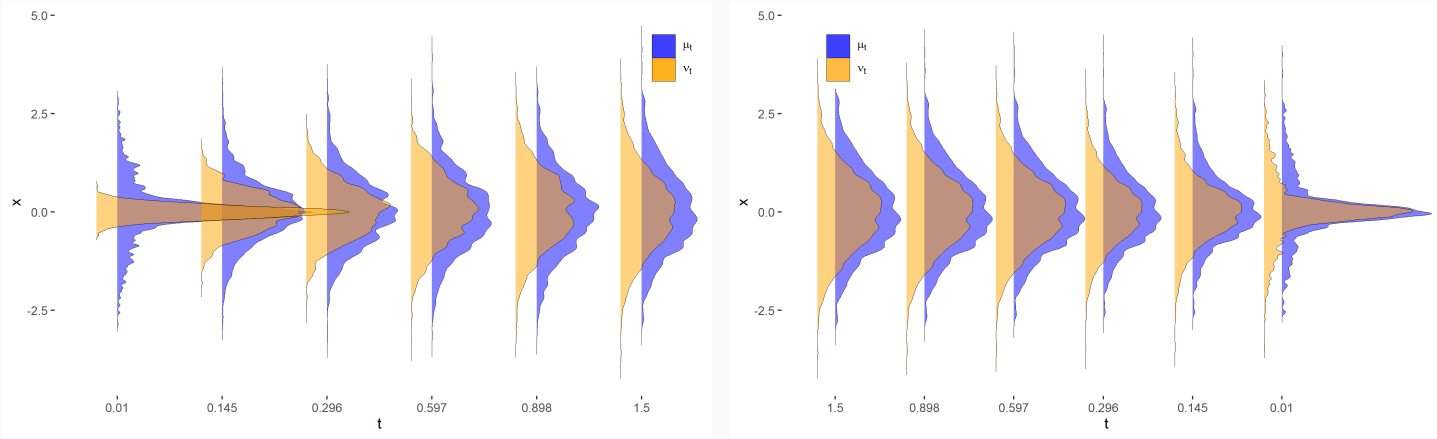


(c)

(a) $s_r(u)$ Low SNR. (b) $s_r(u)$ High SNR. (c) $\zeta_M^*(t)$.

non-log-concave example 2: mixture of point mass and Gaussian

$$X_0 \sim \tfrac{1}{2}\delta_0 + \tfrac{1}{2}\mathcal{N}(0,1)$$

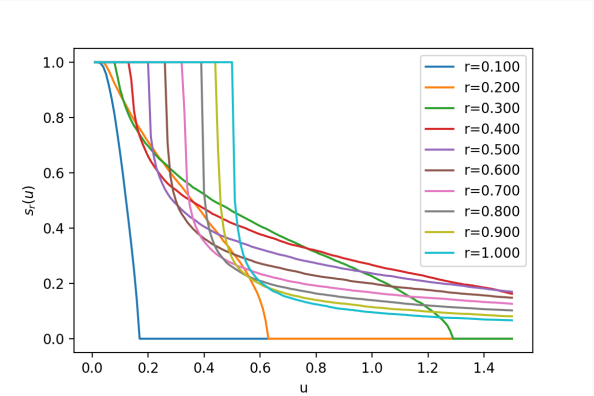


(a) Forward Diffusion via OU **(b)** Backward Transport via OT

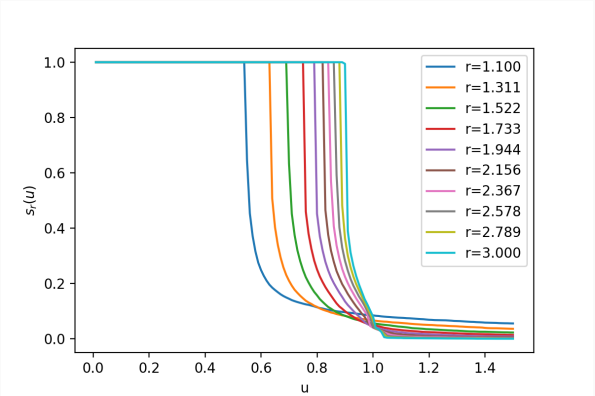
(a) Forward diffusion initialized at $\mu_0 = 0.5\delta_0 + 0.5\mathcal{N}(0,1)$, $\nu_0 = \delta_0$. $T = 100$, $\eta = 0.01$. (b) Backward transport initialized at μ_T, ν_T, and applying the backward OT map $\mathbf{b}^{\mu,\eta}$.

non-log-concave example 3: heterogeneous Gaussian mixture

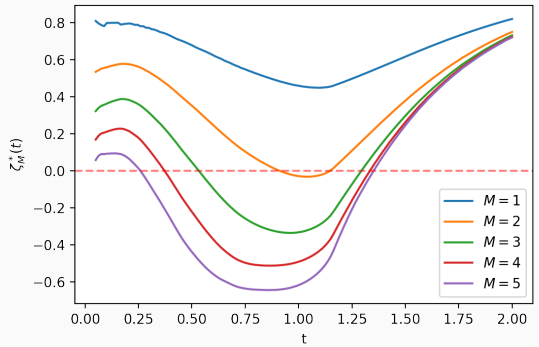
$$X_0 \sim \sum_{i=1}^{m} p_i \cdot \mathcal{N}(\mu_i, \sigma_i^2)$$



(a)



(b)

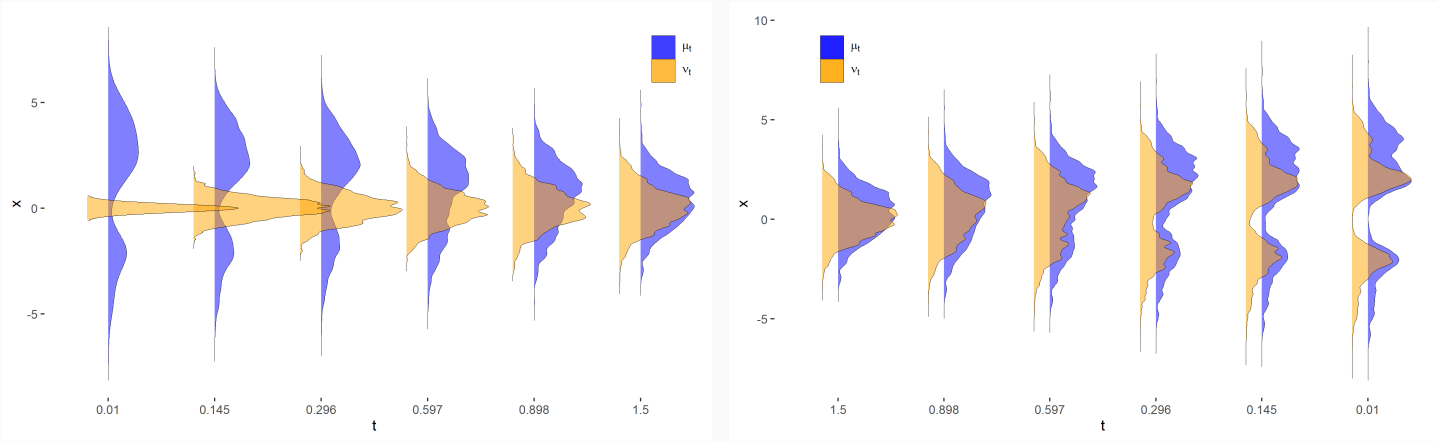


(c)

(a) $s_r(u)$ Low SNR. (b) $s_r(u)$ High SNR. (c) $\zeta_M^*(t)$.

non-log-concave example 3: heterogeneous Gaussian mixture

$$X_0 \sim \sum_{i=1}^{m} p_i \cdot \mathcal{N}(\mu_i, \sigma_i^2)$$



(a) Forward Diffusion via OU **(b)** Backward Transport via OT

(a) Forward diffusion initialized at
$\mu_0 = 0.1\mathcal{N}(-4, 1) + 0.2\mathcal{N}(-2, 0.5) + 0.4\mathcal{N}(2, 0.5) + 0.3\mathcal{N}(4, 1),\ \nu_0 = \delta_0.\ T = 100,\ \eta = 0.01.$
(b) Backward transport initialized at μ_T, ν_T, and applying the backward OT map $\mathbf{b}^{\mu,\eta}$.