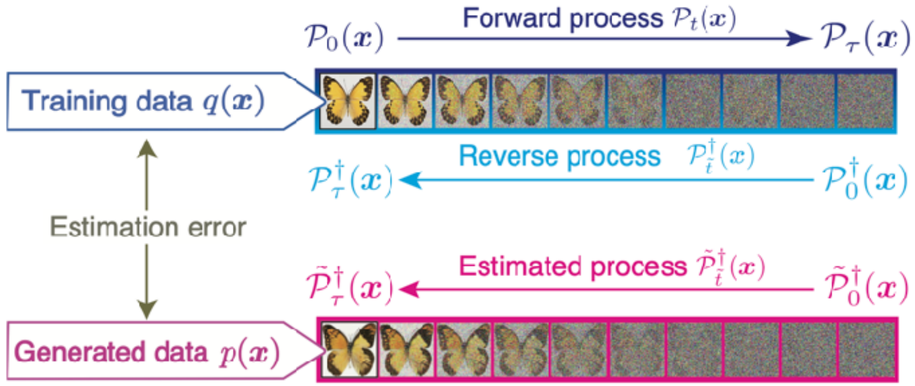
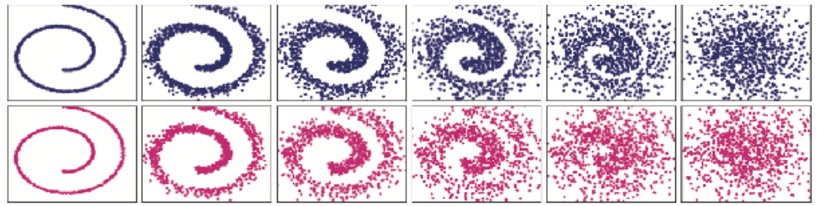
Optimal transport and thermodynamics for the learning: Application to the diffusion model





Sosuke Ito

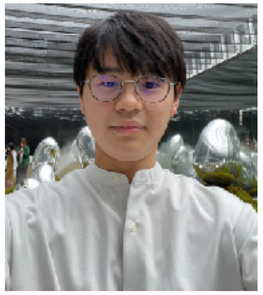
Frontiers in Non-equilibrium Physics 2024, YITP, Jul. 18th, 2024

UTokyo
Universal Biology Institute

Reference and collaborators

Main topic (Diffusion model)

K. Ikeda, T. Uda, D. Okanohara and SI, arXiv:2407.04495.



Kotaro Ikeda (UTokyo)

Tomoya Uda (UTokyo)

Daisuke Okanohara (Preferred Networks Inc.)

Related topic (Thermodynamics and optimal transport)

SI, Information geometry, *Information* Geometry 7.Suppl 1, 441-483 (2024).

M. Nakazato and SI. Phys. Rev. Res. 3, 043093 (2021).
A. Dechant, S-I Sasa and SI. Phys. Rev. Res. 4, L012034 (2022).
A. Dechant, S-I Sasa and SI, Phys. Rev. E. 106, 024125 (2022).
K. Yoshimura, A. Kolchinsky, A. Dechant and SI. Phys. Rev. Res. 5, 013017 (2023).
Y. Fujimoto and SI, Phys. Rev. Res. 6, 013023 (2024).
K. Yoshimura and SI, Phys. Rev. Res. 6, L022057 (2024).
A. Kolchinsky, A. Dechant, K. Yoshimura and SI, arXiv:2206.14599.
R. Nagayama, K. Yoshimura, A. Kolchinsky and SI. arXiv: 2311.16569.
D. Sekizawa, SI, M. Oizumi, arXiv:2312.03489.

Collaborators:
Lab members (+alumni): Muka Nakazato, Kohei Yoshimura, Yuma Fujimoto, Artemy Kolchinsky, Ryan Nagayama
Andreas Dechant (KyotoU), Shin-ichi Sasa (KyotoU), Daiki Sekizawa (UTokyo), Masafumi Oizumi (UTokyo)

Outline

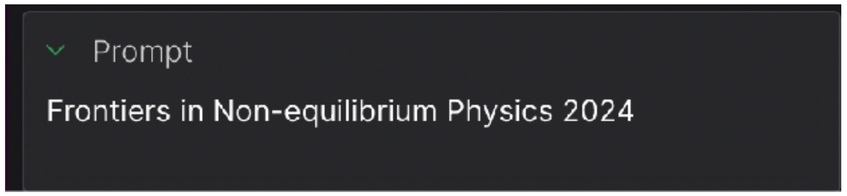
- Introduction: Generative models and diffusion models
- Stochastic thermodynamics based on optimal transport
- Main results: Speed-accuracy trade-off for the diffusion models

K. Ikeda, T. Uda, D. Okanohara and SI, arXiv:2407.04495.

Generative model

Stable diffusion (2022)

- Generative artificial intelligence
- Text-to-image model
- The diffusion models

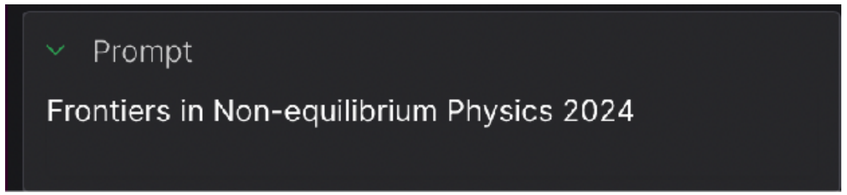


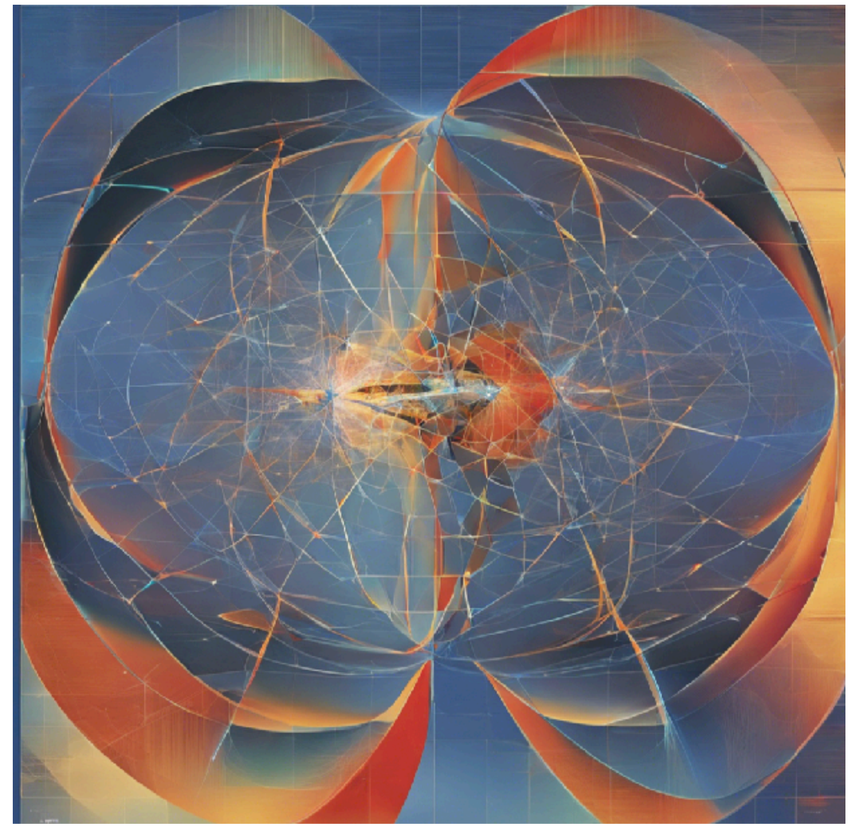
Dream studio by stability.ai **https://beta.dreamstudio.ai/generate**

Generative model

Stable diffusion (2022)

- Generative artificial intelligence
- Text-to-image model
- The diffusion models





Dream studio by stability.ai **https://beta.dreamstudio.ai/generate**

Diffusion model - Original paper (2015)

J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, PMLR, pp. 2256–2265 (2015).

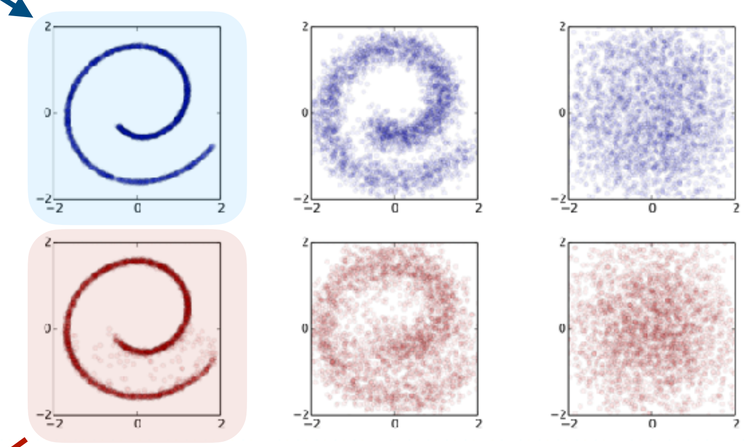
Deep Unsupervised Learning using Nonequilibrium Thermodynamics

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, Surya Ganguli *Proceedings of the 32nd International Conference on Machine Learning*, PMLR 37:2256-2265, 2015.

Abstract

A central problem in machine learning involves modeling complex data-sets using highly flexible families of probability distributions in which learning, sampling, inference, and evaluation are still analytically or computationally tractable. Here, we develop an approach that simultaneously achieves both flexibility and tractability. The essential idea, inspired by non-equilibrium statistical physics, is to systematically and slowly destroy structure in a data distribution through an iterative forward diffusion process. We then learn a reverse diffusion process that restores structure in data, yielding a highly flexible and tractable generative model of the data. This approach allows us to rapidly learn, sample from, and evaluate probabilities in deep generative models with thousands of layers or time steps, as well as to compute conditional and posterior probabilities under the learned model. We additionally release an open source reference implementation of the algorithm.

Training data



Forward diffusion process [learning]

Reverse diffusion process [data generation]

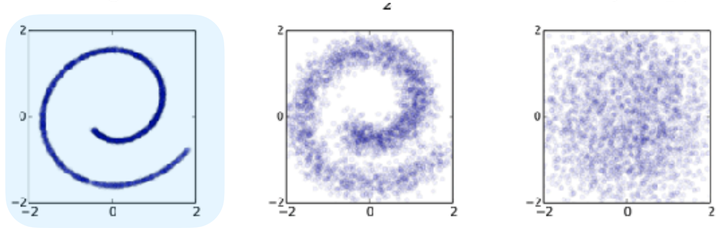
Generated data

Essential idea

J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, PMLR, pp. 2256–2265 (2015).

Training data q

$P_0(\boldsymbol{x}_0) = q(\boldsymbol{x}_0)$ → $P_\tau(\boldsymbol{x}_N)$



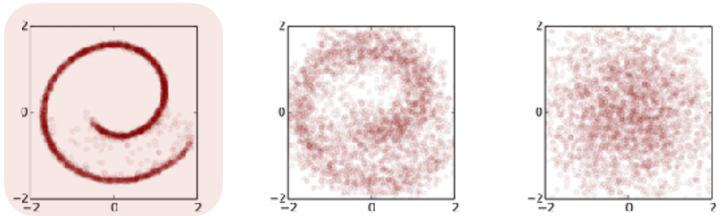
Forward diffusion process [learning]

$$\mathscr{P}^{\mathrm{F}}(\{\boldsymbol{x}\}) = q(\boldsymbol{x}_0)\prod_i T_i(\boldsymbol{x}_{i+1}\,|\,\boldsymbol{x}_i)$$

Estimating the reverse process $\hat{T}_i^\dagger = T_i^\dagger$

$$\mathscr{P}^{\mathrm{E}}(\{\boldsymbol{x}\}) = P_0^\dagger(\boldsymbol{x}_\tau)\prod_i T_i^\dagger(\boldsymbol{x}_i\,|\,\boldsymbol{x}_{i+1})$$

Generated data p



$p(\boldsymbol{x}_0)(\simeq q(\boldsymbol{x}_0))$ ← $P_0^\dagger(\boldsymbol{x}_N)(\simeq P_\tau(\boldsymbol{x}_N))$

Reverse diffusion process [data generation]

Variants of the diffusion models
- Score-based generative model

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., & Poole, B. In International Conference on Learning Representations. (2021).

Score-based generative model

Fokker-Planck equation (Forward diffusion process)

$$\partial_t P_t(\boldsymbol{x}) = -\nabla \cdot (\boldsymbol{\nu}_t(\boldsymbol{x}) P_t(\boldsymbol{x})) \qquad \boldsymbol{\nu}_t(\boldsymbol{x}) = F_t(\boldsymbol{x}) - T_t \nabla \ln P_t(\boldsymbol{x})$$

Estimating $\hat{\boldsymbol{\nu}}_t(\boldsymbol{x}) = F_t(\boldsymbol{x}) - T_t s_t(\boldsymbol{x})$

via the score function $s_t = \nabla \ln P_t$

-Data generation by the reverse stochastic differential equation

$$\dot{\boldsymbol{x}}_{\tilde{t}} = F_{\tau-\tilde{t}}(\boldsymbol{x}_{\tilde{t}}) - 2\hat{\boldsymbol{\nu}}_{\tau-\tilde{t}}(\boldsymbol{x}_{\tilde{t}}) + \sqrt{2T_{\tau-\tilde{t}}}\boldsymbol{\xi}_{\tau-\tilde{t}}$$

($\tilde{t} = \tau - t$:Reversed time)

-Data generation by the ordinary differential equation (probability flow ODE)

$$\dot{\boldsymbol{x}}_{\tilde{t}} = -\hat{\boldsymbol{\nu}}_{\tau-\tilde{t}}(\boldsymbol{x}_{\tilde{t}})$$

Variants of the diffusion models
- Flow-based generative model

Y. Lipman, R. T. Chen, H. Ben-Hamu, M. Nickel, and M. Le, International Conference on Learning Representations (2022)

Flow-based generative model

Continuity equation (Forward process)

$$\partial_t P_t(\boldsymbol{x}) = -\nabla \cdot (\boldsymbol{\nu}_t(\boldsymbol{x}) P_t(\boldsymbol{x}))$$

Estimating the velocity field $\hat{\boldsymbol{\nu}}_t(\boldsymbol{x}) = \boldsymbol{\nu}_t(\boldsymbol{x})$

-Data generation by the ordinary differential equation

$$\dot{\boldsymbol{x}}_{\tilde{t}} = -\hat{\boldsymbol{\nu}}_{\tau-\tilde{t}}(\boldsymbol{x}_{\tilde{t}})$$

($\tilde{t} = \tau - t$:Reversed time)

Examples: Forward diffusion process for accurate data generation

Linear force $F_t(\boldsymbol{x}) = \mathrm{A}_t \boldsymbol{x} + \boldsymbol{b}_t$

Gaussian transition probability

$$P_t^{\mathrm{c}}(\boldsymbol{x}\,|\,\boldsymbol{y}) = \mathcal{N}(\boldsymbol{x}\,|\,\boldsymbol{\mu}_t(\boldsymbol{y}), \Sigma_t) \qquad \boldsymbol{\mu}_0(\boldsymbol{y}) = \boldsymbol{y} \qquad \Sigma_0 = \mathrm{O} \quad \Rightarrow \quad P_t(\boldsymbol{x}) = \int d\boldsymbol{y} P_t^{\mathrm{c}}(\boldsymbol{x}\,|\,\boldsymbol{y}) P_0(\boldsymbol{y})$$

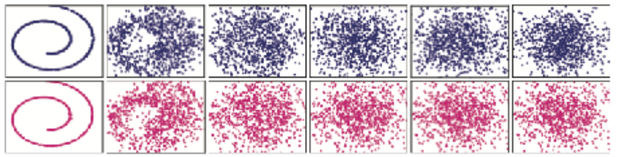
Cosine schedule

A. Q. Nichol, & P. Dhariwal, In *International conference on machine learning* (pp. 8162-8171). PMLR (2021)

$$\boldsymbol{\mu}_t(\boldsymbol{y}) = m_t \boldsymbol{y} \qquad \Sigma_t = \sigma_t^2 \mathrm{I}$$

$$m_t^2 + \sigma_t^2 = 1$$

$$m_t = \cos\left(\frac{\pi}{2}\frac{t}{\tau}\right) \qquad \sigma_t = \sin\left(\frac{\pi}{2}\frac{t}{\tau}\right) \qquad t \in [0,\tau]$$

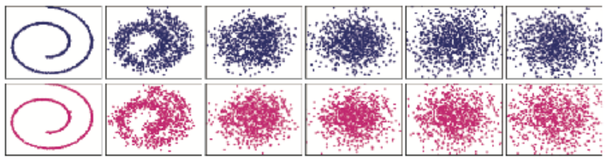


Conditional optimal transport schedule (Approximate optimal transport)

Y. Lipman, R. T. Chen, H. Ben-Hamu, M. Nickel, and M. Le, International Conference on Learning Representations (2022)

$$\boldsymbol{\mu}_t(\boldsymbol{y}) = m_t \boldsymbol{y} \qquad \Sigma_t = \sigma_t^2 \mathrm{I}$$

$$m_t = 1 - \frac{t}{\tau} \qquad \sigma_t = \frac{t}{\tau} \qquad t \in [0,\tau]$$



(Figure from) K. Ikeda, T. Uda, D. Okanohara and SI, arXiv:2407.04495.

Motivation

The diffusion models are inspired by nonequilibrium thermodynamics.

J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, PMLR, pp. 2256–2265 (2015).

Question:

Is stochastic thermodynamics still useful for understanding the current technique (e.g., optimal transport) in the diffusion models?

Our results:

In terms of stochastic thermodynamics based on optimal transport, the accuracy of data generation in the diffusion models can be discussed thermodynamically.

Outline

- Introduction: Generative models and diffusion models
- Stochastic thermodynamics based on optimal transport
- Main results: Speed-accuracy trade-off for the diffusion models

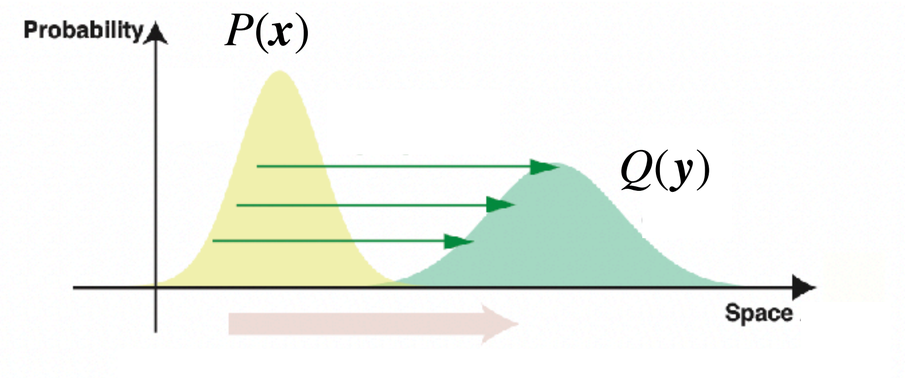
K. Ikeda, T. Uda, D. Okanohara and SI, arXiv:2407.04495.

Optimal transport: p-Wasserstein distance

Textbook: Villani, C. (2009). *Optimal transport: old and new* (Vol. 338, p. 23). Berlin: springer.

p-Wasserstein distance

$$\mathscr{W}_p(P,Q) = \left(\inf_{\pi \in \Pi(P,Q)} \int d\boldsymbol{x} \int d\boldsymbol{y} \pi(\boldsymbol{x},\boldsymbol{y}) \|\boldsymbol{x}-\boldsymbol{y}\|^p \right)^{\frac{1}{p}}$$



$$\Pi(P,Q) = \left\{ \pi(\boldsymbol{x},\boldsymbol{y}) \,\middle|\, \int d\boldsymbol{y}\pi(\boldsymbol{x},\boldsymbol{y}) = P(\boldsymbol{x}), \int d\boldsymbol{x}\pi(\boldsymbol{x},\boldsymbol{y}) = Q(\boldsymbol{y}), \pi(\boldsymbol{x},\boldsymbol{y}) \geq 0 \right\}$$

Metric:

$$① \ \mathscr{W}_p(P,Q) \geq 0 \quad ② \ \mathscr{W}_p(P,Q) = 0 \Leftrightarrow P = Q \quad ③ \ \mathscr{W}_p(P,Q) = \mathscr{W}_p(Q,P)$$

$$④ \ \mathscr{W}_p(P,R) + \mathscr{W}_p(R,Q) \geq \mathscr{W}_p(P,Q)$$

Inequality:

$$p \geq q \geq 1 \Rightarrow \mathscr{W}_p(P,Q) \geq \mathscr{W}_q(P,Q)$$

Optimal transport: Expressions based on dual problems

Textbook: Villani, C. (2009). *Optimal transport: old and new* (Vol. 338, p. 23). Berlin: springer.

1-Wasserstein distance

(Kantorovich-Rubinstein duality)

$$\mathscr{W}_1(P,Q) = \sup_{f \in \mathrm{Lip}^1} [\langle f \rangle_P - \langle f \rangle_Q]$$

$$\langle f \rangle_P = \int d\boldsymbol{x} f(\boldsymbol{x}) P(\boldsymbol{x}) \qquad \mathrm{Lip}^1 = \{ f(\boldsymbol{x}) \,|\, \|\nabla f(\boldsymbol{x})\|^2 \leq 1 \}$$

2-Wasserstein distance

(Benamou-Brenier formula)

$$\mathscr{W}_2(P,Q) = \sqrt{\inf_{\{\boldsymbol{u}_t, Q_t\}_{0 \leq t \leq \tau}} \tau \int_0^\tau dt \int d\boldsymbol{x} \|\boldsymbol{u}_t(\boldsymbol{x})\|^2 Q_t(\boldsymbol{x})}$$

$$\partial_t Q_t(\boldsymbol{x}) = -\nabla \cdot (\boldsymbol{u}_t(\boldsymbol{x}) Q_t(\boldsymbol{x})) \quad Q_0(\boldsymbol{x}) = P(\boldsymbol{x}) \quad Q_\tau(\boldsymbol{x}) = Q(\boldsymbol{x})$$

J-D. Benamou & Y. Brenier. *Numerische Mathematik* 84, 375-393 (2000).

Stochastic thermodynamics for the diffusion systems

Review: U. Seifert, Reports on progress in physics, 75, 126001 (2012).

Fokker-Planck equation

$$\partial_t P_t(\boldsymbol{x}) = -\nabla \cdot (\boldsymbol{\nu}_t(\boldsymbol{x}) P_t(\boldsymbol{x}))$$

$$\boldsymbol{\nu}_t(\boldsymbol{x}) = F_t(\boldsymbol{x}) - T_t \nabla \ln P_t(\boldsymbol{x})$$

The entropy production rate

$$\dot{S}_t^{\mathrm{tot}} = \frac{1}{T_t} \int d\boldsymbol{x} \|\boldsymbol{\nu}_t(\boldsymbol{x})\|^2 P_t(\boldsymbol{x})$$

The entropy production

$$S_\tau^{\mathrm{tot}} = \int_0^\tau dt \dot{S}_t^{\mathrm{tot}}$$

Lower bound on the entropy production rate

Lower bound on the entropy production rate (Excess entropy production rate*)

$$\dot{S}_t^{\mathrm{tot}} \geq \frac{[v_2(t)]^2}{T_t} = \frac{1}{T_t}\int dx \|\boldsymbol{\nu}_t^{\mathrm{ex}}(x)\|^2 P_t(x)$$

Speed in the space of the 2-Wasserstein distance

$$v_2(t) = \lim_{\Delta t \to +0} \frac{\mathscr{W}_2(P_t, P_{t+\Delta t})}{\Delta t} = \sqrt{\int dx \|\boldsymbol{\nu}_t^{\mathrm{ex}}(x)\|^2 P_t(x)}$$

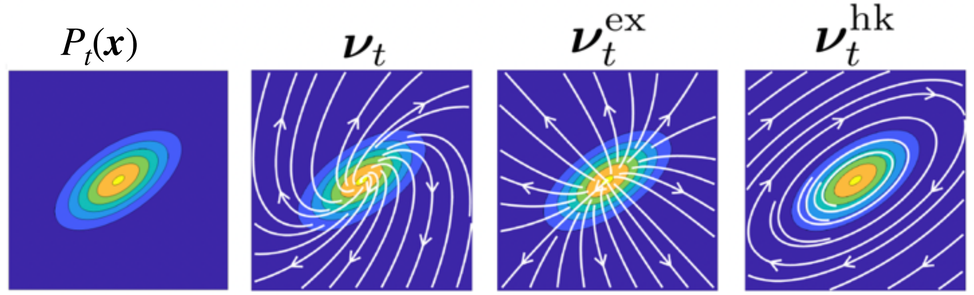
cf.) Benamou-Brenier formula

$$\partial_t P_t(x) = -\nabla\cdot(\boldsymbol{\nu}_t(x)P_t(x)) = -\nabla\cdot(\boldsymbol{\nu}_t^{\mathrm{ex}}(x)P_t(x))$$

$\boldsymbol{\nu}_t^{\mathrm{ex}}(x) = \nabla\phi_t(x)$:conservative (gradient flow)

$\nabla\cdot(\boldsymbol{\nu}_t^{\mathrm{hk}}(x)P_t(x)) = 0$:non-conservative

$\boldsymbol{\nu}_t^{\mathrm{hk}}(x) = \boldsymbol{\nu}_t(x) - \boldsymbol{\nu}_t^{\mathrm{ex}}(x)$ (cyclic)



(Figure from) D. Sekizawa, SI and M. Oizumi, arXiv:2312.03489.

M. Nakazato and SI. Phys. Rev. Res. 3, 043093 (2021).
A. Dechant, S-I Sasa and SI. Phys. Rev. Res. 4, L012034 (2022).

* Maes, C., & Netočný, K. *Journal of Statistical Physics*, *154*, 188-203 (2014).

Minimum entropy production and geodesic (optimal transport)

Time-independent temperature $T_t = T = \mathrm{const}.$

Thermodynamic speed limit

$$S_\tau^{\mathrm{tot}} \geq \frac{[\int_0^\tau dt v_2(t)]^2}{\tau T} \geq \frac{[\mathscr{W}_2(P_0, P_\tau)]^2}{\tau T}$$

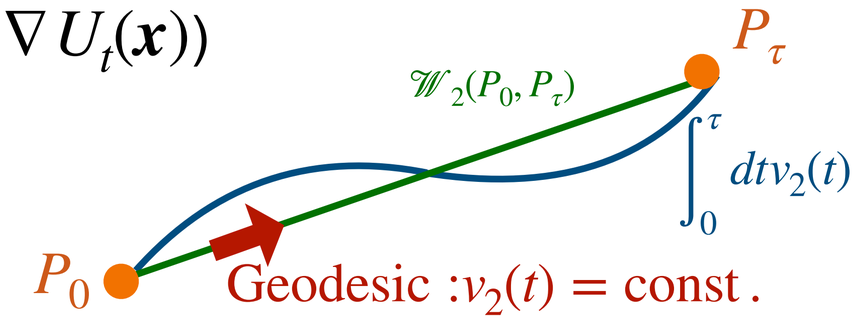
E. Aurell, K. Gawędzki, C. Mejía-Monasterio, R. Mohayaee, & P. Muratore-Ginanneschi, *Journal of statistical physics*, *147*, 487-505 (2012).
M. Nakazato and SI. Phys. Rev. Res. 3, 043093 (2021).

Minimum entropy production: Geodesic + Conservative

$$S_\tau^{\mathrm{tot}} = \frac{[\mathscr{W}_2(P_0, P_\tau)]^2}{\tau T}$$

$\dot{S}_t^{\mathrm{tot}} = \dfrac{[v_2(t)]^2}{T}$:Conservative ($\boldsymbol{\nu}_t(\boldsymbol{x}) = \nabla \phi_t(\boldsymbol{x})$ or $\boldsymbol{F}_t(\boldsymbol{x}) = -\nabla U_t(\boldsymbol{x})$)

$v_2(t) = \dfrac{\mathscr{W}_2(P_0, P_\tau)}{\tau} = \mathrm{const}.$:Geodesic (optimal transport)



Thermodynamic uncertainty relation for the excess entropy production rate

Thermodynamic uncertainty relation

$$\dot{S}_t^{\mathrm{tot}} \geq \frac{[v_2(t)]^2}{T_t} \geq \frac{|\partial_t \langle r \rangle_{P_t}|^2}{T_t \langle \|\nabla r\|^2 \rangle_{P_t}}$$

A. Dechant, S-I Sasa and SI. Phys. Rev. Res. 4, L012034 (2022).
A. Dechant, S-I Sasa and SI, Phys. Rev. E. 106, 024125 (2022).

cf.) Cramér–Rao bound: SI and A. Dechant, *Physical Review X*, *10*, 021056 (2020).

$r(\boldsymbol{x})$: time-independent observable

$$v_2(t) \geq v_r(t)$$

(Normalized) speed of observable $r(\boldsymbol{x})$

$$v_r(t) = \frac{|\partial_t \langle r \rangle_{P_t}|}{\sqrt{\langle \|\nabla r\|^2 \rangle_{P_t}}}$$

Speed in the space of the 2-Wasserstein distance is the upper bound on the speed of any observable.

cf.) $\mathscr{W}_2(P,Q) \geq \mathscr{W}_1(P,Q), r(\boldsymbol{x}) \in \mathrm{Lip}^1$

R. Nagayama, K. Yoshimura, A. Kolchinsky and SI. arXiv: 2311.16569.

Analogous to thermodynamic speed limit and thermodynamic uncertainty relation

Question:

Is stochastic thermodynamics still useful for understanding the current technique (e.g., optimal transport) in the diffusion models?

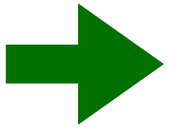
Stochastic thermodynamics

Optimal transport
= Minimum entropy production

Trade-offs

$$\dot{S}_t^{\mathrm{tot}} \geq \frac{[v_2(t)]^2}{T_t} \qquad S_\tau^{\mathrm{tot}} \geq \frac{[\int_0^\tau dt v_2(t)]^2}{\tau T} \qquad v_2(t) \geq v_r(t)$$

Analogy



Diffusion models

(Approximate) optimal transport
= Accurate data generation
(empirical finding)

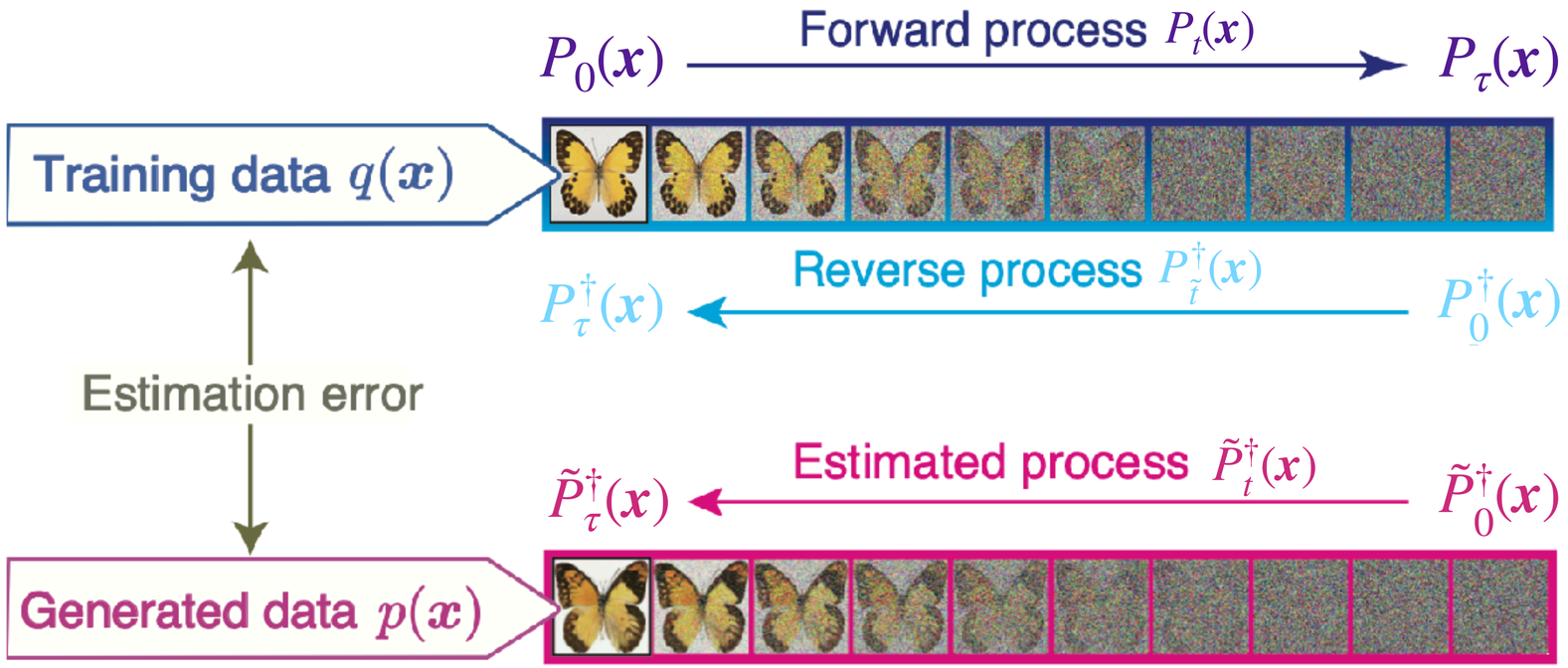
Trade-offs: Our results

Outline

- Introduction: Generative models and diffusion models
- Stochastic thermodynamics based on optimal transport
- Main results: Speed-accuracy trade-off for the diffusion models

K. Ikeda, T. Uda, D. Okanohara and SI, arXiv:2407.04495.

Estimation error in the diffusion models



$$P_t(x) = P^\dagger_{\tau-t}(x)$$

$$P^\dagger_0(x) \neq \tilde{P}^\dagger_0(x)$$

Estimation error (measured by the 1-Wasserstein distance)

$$\mathscr{W}_1(p, q)$$

e.g.,) K. Oko, S. Akiyama & T. Suzuki, In *International Conference on Machine Learning* (pp. 26517-26582). PMLR (2023).

Perturbation and response

Forward process/Reverse process

$$\partial_t P_t(\boldsymbol{x}) = -\nabla \cdot (\boldsymbol{\nu}_t(\boldsymbol{x}) P_t(\boldsymbol{x})) \quad P_t(\boldsymbol{x}) = P^\dagger_{\tau-t}(\boldsymbol{x})$$

$$\partial_{\tilde{t}} P^\dagger_{\tilde{t}}(\boldsymbol{x}) = \nabla \cdot (\boldsymbol{\nu}_{\tau-\tilde{t}}(\boldsymbol{x}) P^\dagger_{\tilde{t}}(\boldsymbol{x})) \quad \tilde{t} = \tau - t$$

Estimated process

[Probability flow ODE/Flow-based generative modeling]

$$\partial_{\tilde{t}} \tilde{P}^\dagger_{\tilde{t}}(\boldsymbol{x}) = \nabla \cdot (\boldsymbol{\nu}_{\tau-\tilde{t}}(\boldsymbol{x}) \tilde{P}^\dagger_{\tilde{t}}(\boldsymbol{x}))$$

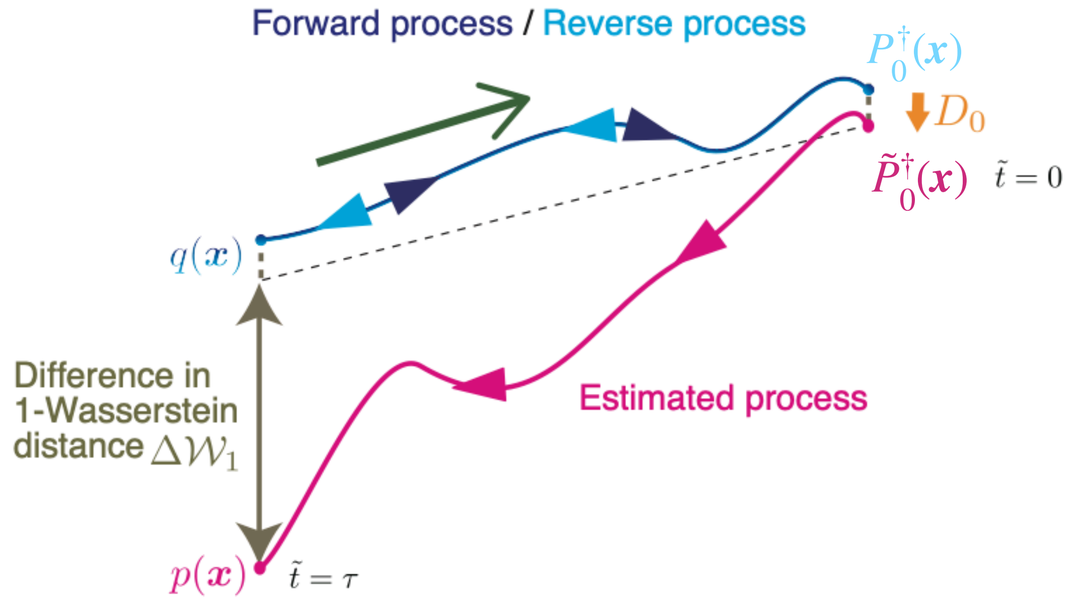
Initial perturbation

$$D_0 = \int d\boldsymbol{x} \frac{(P^\dagger_0(\boldsymbol{x}) - \tilde{P}^\dagger_0(\boldsymbol{x}))^2}{P^\dagger_0(\boldsymbol{x})} : \chi^2\text{-divergence}$$

Response function

$$\frac{\Delta \mathscr{W}_1^2}{D_0} = \frac{[\mathscr{W}_1(p,q) - \mathscr{W}_1(P^\dagger_0, \tilde{P}^\dagger_0)]^2}{D_0}$$

Estimation error / Perturbation



$\frac{\Delta \mathscr{W}_1^2}{D_0}$ is **small**. ⇒ Data generation is **robust** to the initial perturbation.

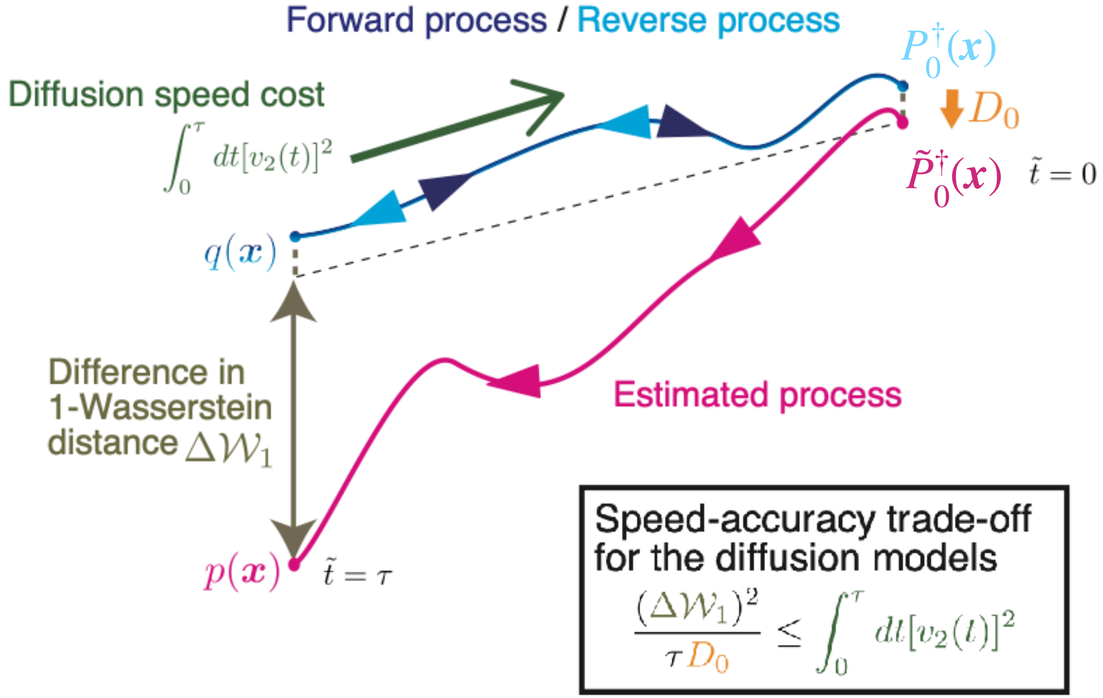
Main results:
Speed-accuracy trade-off for the diffusion models

Speed-accuracy trade-off

$$\frac{\Delta \mathscr{W}_1^2}{\tau D_0} \leq \int_0^\tau dt T_t \dot{S}_t^{\mathrm{tot}}$$

Conservative case
($\boldsymbol{\nu}_t(\boldsymbol{x}) = \nabla \phi_t(\boldsymbol{x})$ or $\boldsymbol{F}_t(\boldsymbol{x}) = -\nabla U_t(\boldsymbol{x})$)

$$\frac{\Delta \mathscr{W}_1^2}{\tau D_0} \leq \int_0^\tau dt [v_2(t)]^2$$



The **robustness** of data generation is generally limited by **the diffusion speed** $v_2(t)$

(or the entropy production rate $\dot{S}_t^{\mathrm{tot}}$) in the forward process.

Main results:
Speed-accuracy trade-off for the diffusion models (Instantaneous)

Instantaneous speed-accuracy trade-off

$$\frac{|\,\partial_t \mathscr{W}_1(\tilde{P}^\dagger_{\tau-t}, P^\dagger_{\tau-t})\,|^2}{D_0} \leq T_t \dot{S}^{\mathrm{tot}}_t$$

Conservative case ($\boldsymbol{\nu}_t(\boldsymbol{x}) = \nabla \phi_t(\boldsymbol{x})$ or $\boldsymbol{F}_t(\boldsymbol{x}) = -\nabla U_t(\boldsymbol{x})$)

$$v_{\mathrm{loss}}(t) \leq v_2(t)$$

$$v_{\mathrm{loss}}(t) = \frac{|\,\partial_t \mathscr{W}_1(\tilde{P}^\dagger_{\tau-t}, P^\dagger_{\tau-t})\,|}{\sqrt{D_0}}$$

cf.) Thermodynamic uncertainty relation $v_r(t) \leq v_2(t)$

Sketch of proof: Instantaneous trade-off

$$\partial_{\tilde{t}} P^\dagger_{\tilde{t}}(\boldsymbol{x}) = \nabla \cdot (\boldsymbol{\nu}_{\tau-\tilde{t}}(\boldsymbol{x}) P^\dagger_{\tilde{t}}(\boldsymbol{x}))$$
$$\partial_{\tilde{t}} \tilde{P}^\dagger_{\tilde{t}}(\boldsymbol{x}) = \nabla \cdot (\boldsymbol{\nu}_{\tau-\tilde{t}}(\boldsymbol{x}) \tilde{P}^\dagger_{\tilde{t}}(\boldsymbol{x}))$$
$$\tilde{t} = \tau - t$$

$$\partial_t [P^\dagger_{\tau-t}(\boldsymbol{x}) - \tilde{P}^\dagger_{\tau-t}(\boldsymbol{x})] = -\nabla \cdot (\boldsymbol{\nu}_t(\boldsymbol{x}) [P^\dagger_{\tau-t}(\boldsymbol{x}) - \tilde{P}^\dagger_{\tau-t}(\boldsymbol{x})])$$

Continuity equation

$f \in \mathrm{Lip}^1$

$$|\partial_t(\langle f \rangle_{P^\dagger_{\tau-t}} - \langle f \rangle_{\tilde{P}^\dagger_{\tau-t}})|^2 = \left(\int d\boldsymbol{x} f(\boldsymbol{x}) \partial_t [P^\dagger_{\tilde{t}}(\boldsymbol{x}) - \tilde{P}^\dagger_{\tilde{t}}(\boldsymbol{x})] \right)^2$$
$$= \left(\int d\boldsymbol{x}\, \nabla f(\boldsymbol{x}) \cdot \boldsymbol{\nu}_t(\boldsymbol{x}) [P^\dagger_{\tau-t}(\boldsymbol{x}) - \tilde{P}^\dagger_{\tau-t}(\boldsymbol{x})] \right)^2$$

Cauchy-Schwartz inequality
+ 1-Lipshitz ($\|\nabla f(\boldsymbol{x})\| \leq 1$)

$$\leq \underbrace{\left(\int d\boldsymbol{x} \|\boldsymbol{\nu}_t(\boldsymbol{x})\|^2 P_t(\boldsymbol{x}) \right)}_{T_t \dot{S}^{\mathrm{tot}}_t} \underbrace{\left(\int d\boldsymbol{x} \frac{[P^\dagger_{\tau-t}(\boldsymbol{x}) - \tilde{P}^\dagger_{\tau-t}(\boldsymbol{x})]^2}{P^\dagger_{\tau-t}(\boldsymbol{x})} \right)}_{D_0}$$

(Time-independent)

+ Kantrovich-Rubinstein duality

$\exists f \in \mathrm{Lip}^1$ $\quad |\partial_t \mathscr{W}_1(P^\dagger_{\tau-t}, \tilde{P}^\dagger_{\tau-t})|^2 \leq |\partial_t(\langle f \rangle_{P^\dagger_{\tau-t}} - \langle f \rangle_{\tilde{P}^\dagger_{\tau-t}})|^2$

Instantaneous speed-accuracy trade-off

$$\frac{|\partial_t \mathscr{W}_1(\tilde{P}^\dagger_{\tau-t}, P^\dagger_{\tau-t})|^2}{D_0} \leq T_t \dot{S}^{\mathrm{tot}}_t$$

Sketch of proof: speed-accuracy trade-off

Instantaneous speed-accuracy trade-off

$$\frac{|\partial_t \mathscr{W}_1(\tilde{P}^\dagger_{\tau-t}, P^\dagger_{\tau-t})|^2}{D_0} \leq T_t \dot{S}^{\mathrm{tot}}_t$$

$$\int_0^\tau dt T_t \dot{S}^{\mathrm{tot}}_t \geq \int_0^\tau dt \frac{|\partial_t \mathscr{W}_1(\tilde{P}^\dagger_{\tau-t}, P^\dagger_{\tau-t})|^2}{D_0}$$

Cauchy-Schwartz inequality

$$\geq \frac{(\Delta \mathscr{W}_1)^2}{\tau D_0}$$

Speed-accuracy trade-off

$$\frac{\Delta \mathscr{W}_1^2}{\tau D_0} \leq \int_0^\tau dt T_t \dot{S}^{\mathrm{tot}}_t$$

“Optimal" forward process for accurate data generation

$$\frac{\Delta \mathscr{W}_1^2}{\tau D_0} \leq \int_0^\tau dt[v_2(t)]^2$$

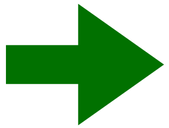
Minimizing the upper bound

$$\int_0^\tau dt[v_2(t)]^2 \geq \frac{\mathscr{W}_2(P_0, P_\tau)^2}{\tau}$$

cf.) Minimum entropy production

$$v_2(t) = \frac{\mathscr{W}_2(P_0, P_\tau)}{\tau} = \text{const.}$$

:Geodesic (optimal transport)



$$\int_0^\tau dt[v_2(t)]^2 = \frac{\mathscr{W}_2(P_0, P_\tau)^2}{\tau}$$

Minimum value

The "optimal" forward process is a dynamics driven by optimal transport (i.e., geodesic in the space of the 2-Wasserstein distance).

“Suboptimal” forward process for accurate data generation

Theorem N. Shaul, R. T. Chen, M. Nickel, M. Le, and Y. Lipman, in International Conference on Machine Learning, PMLR, pp. 30883–30907 (2023)

If the number of data N_{D} is small enough compared to the dimension of the data n_{d} ($N_{\mathrm{D}}/\sqrt{n_{\mathrm{d}}} \to 0$),

$$\int_0^\tau dt[v_2(t)]^2 \simeq n_{\mathrm{d}} \int_0^\tau dt[(\partial_t \sigma_t)^2 + (\partial_t m_t)^2]$$

$$P_t(\boldsymbol{x}) = \int d\boldsymbol{y} P_t^{\mathrm{c}}(\boldsymbol{x}\,|\,\boldsymbol{y}) P_0(\boldsymbol{y})$$

$$P_t^{\mathrm{c}}(\boldsymbol{x}\,|\,\boldsymbol{y}) = \mathcal{N}(\boldsymbol{x}\,|\,m_t \boldsymbol{y}, \sigma_t^2 \mathrm{I})$$

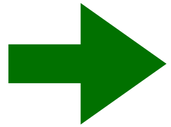
$$\frac{\Delta \mathscr{W}_1^2}{\tau D_0} \leq \int_0^\tau dt[v_2(t)]^2 \simeq n_{\mathrm{d}} \int_0^\tau dt[(\partial_t \sigma_t)^2 + (\partial_t m_t)^2]$$

Minimizing the approximate upper bound (suboptimal)

“Suboptimal” forward process: Conditional optimal transport schedule

$$n_{\mathrm{d}} \int_0^{\tau} dt[(\partial_t \sigma_t)^2 + (\partial_t m_t)^2] \geq n_{\mathrm{d}} \frac{(\sigma_0 - \sigma_\tau)^2 + (m_0 - m_\tau)^2}{\tau}$$

$$m_t = 1 - \frac{t}{\tau} \qquad \sigma_t = \frac{t}{\tau}$$



$$n_{\mathrm{d}} \int_0^{\tau} dt[(\partial_t \sigma_t)^2 + (\partial_t m_t)^2] = n_{\mathrm{d}} \frac{(\sigma_0 - \sigma_\tau)^2 + (m_0 - m_\tau)^2}{\tau}$$

:Conditional optimal transport schedule

Minimum value

The “suboptimal" forward process is a dynamics driven by the conditional optimal transport schedule.

“Suboptimal” forward process: Cosine schedule

Constraint: $m_t^2 + \sigma_t^2 = 1 \Rightarrow (m_t, \sigma_t) = (\cos\theta_t, \sin\theta_t)$

$$n_{\mathrm{d}} \int_0^\tau dt[(\partial_t \sigma_t)^2 + (\partial_t m_t)^2] \geq n_{\mathrm{d}} \frac{(\theta_0 - \theta_\tau)^2}{\tau}$$

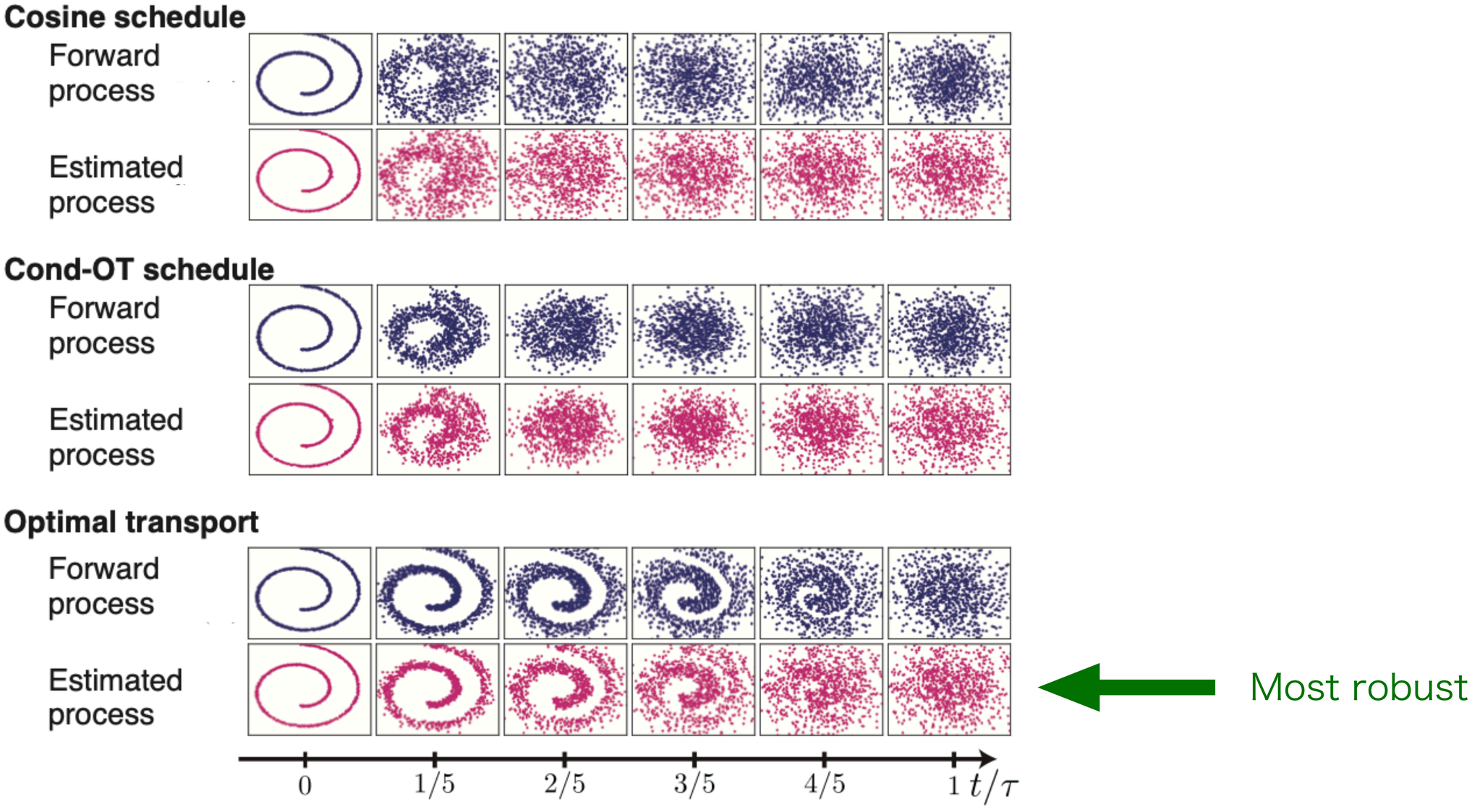
$$m_t = \cos\left(\frac{\pi}{2}\frac{t}{\tau}\right) \quad \sigma_t = \sin\left(\frac{\pi}{2}\frac{t}{\tau}\right) \Rightarrow n_{\mathrm{d}} \int_0^\tau dt[(\partial_t \sigma_t)^2 + (\partial_t m_t)^2] = n_{\mathrm{d}} \frac{(\theta_0 - \theta_\tau)^2}{\tau}$$

:Cosine schedule

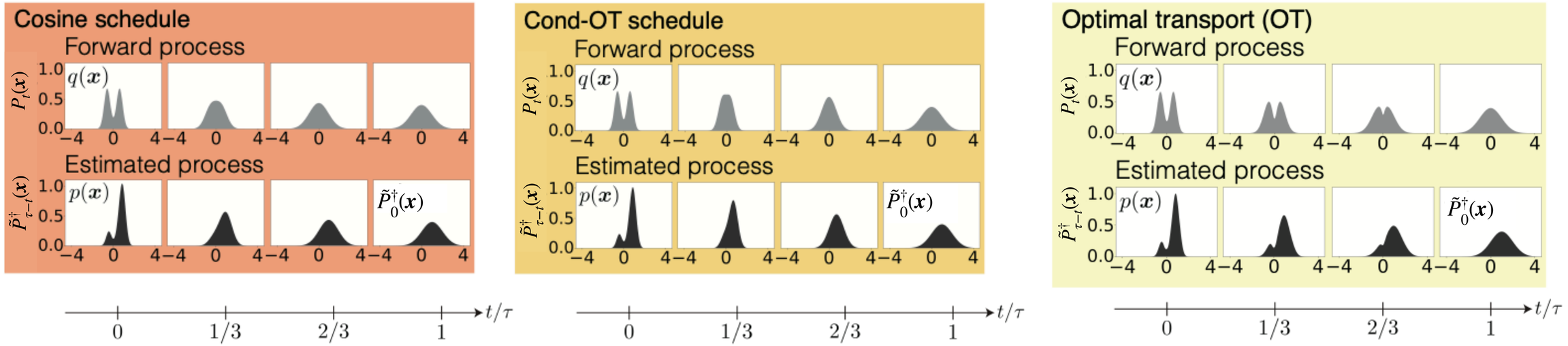
Minimum value

The “suboptimal" forward process under the constraint is a dynamics driven by the cosine schedule.

Examples of optimal and suboptimal dynamics for the diffusion models: Swiss roll



Examples of optimal and suboptimal dynamics for the diffusion models: Gaussian mixture



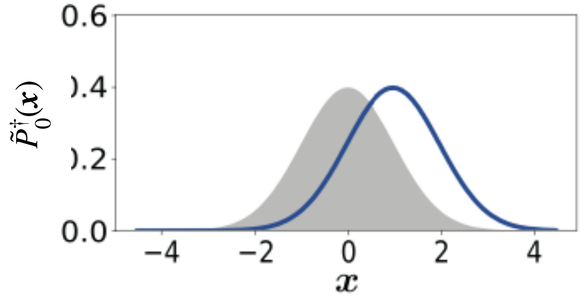
$P_t(\boldsymbol{x})$:Forward process

$\tilde{P}^\dagger_{\tau-t}(\boldsymbol{x})$:Estimated process

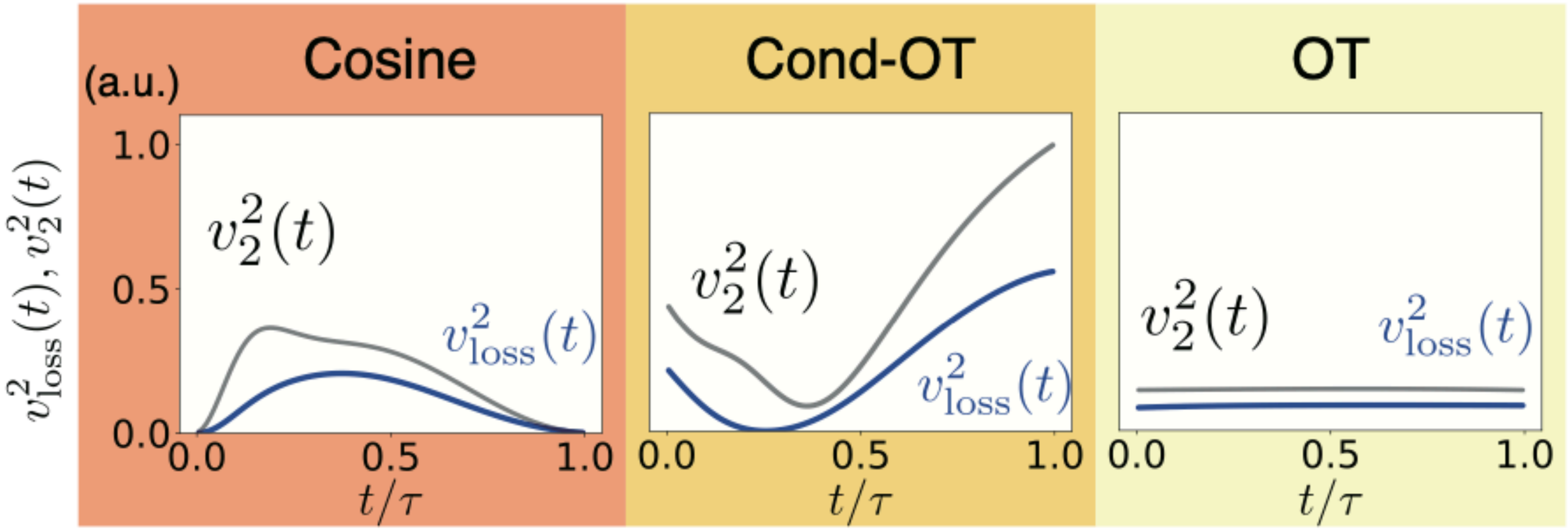
The data structure (the two peaks) is well recovered even during the dynamics of the estimated process in the case of optimal transport.

Initial perturbation

Gaussian distribution with different mean



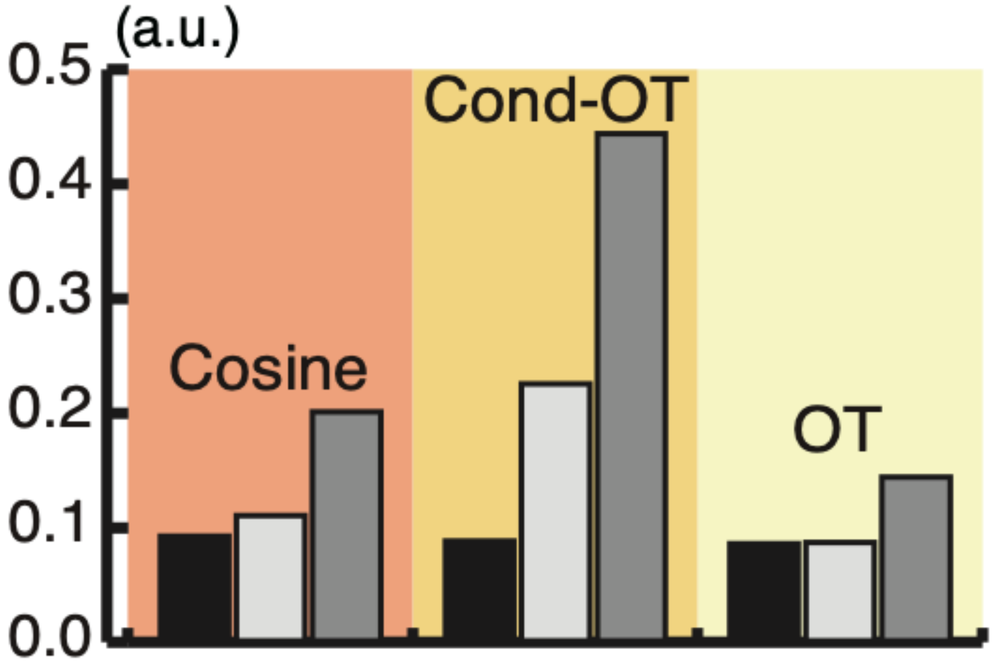
Speed-accuracy trade-off for the diffusion models (Instantaneous)



$$\frac{|\,\partial_t \mathscr{W}_1(\tilde{P}^\dagger_{\tau-t}, P^\dagger_{\tau-t})\,|^2}{D_0} = [v_{\mathrm{loss}}(t)]^2 \leq [v_2(t)]^2$$

In the case of optimal transport, the data structure is not well rapidly changed during the dynamics of the estimated process.

Speed-accuracy trade-off for the diffusion models



$$\frac{(\Delta \mathscr{W}_1)^2}{\tau D_0} \leq \int_0^\tau dt[v_{\mathrm{loss}}(t)]^2 \leq \int_0^\tau dt[v_2(t)]^2$$

The bounds are tighter in the case of the optimal transport compared to the cosine and conditional optimal transport schedules.

The value of $(\Delta \mathscr{W}_1)^2/(\tau D_0)$ for the optimal transport is the smallest for any schedules.

$\frac{(\Delta \mathscr{W}_1)^2}{\tau D_0}$ $\int_0^\tau dt[v_{\mathrm{loss}}(t)]^2$ $\int_0^\tau dt[v_2(t)]^2$

Noise schedules	Values of $(\Delta \mathscr{W}_1)^2/(\tau D_0)$
Cosine	9.1884×10^{-2}
Cond-OT	8.7810×10^{-2}
OT	8.5375×10^{-2}

Interestingly, the cosine and conditional optimal transport schedules work well in the data generation for this simple case because the response function $(\Delta \mathscr{W}_1)^2/(\tau D_0)$ is small enough.

Summary

- We used the technique of stochastic thermodynamics and optimal transport to discuss the accurate data generation in the diffusion models.

- We derived the trade-off relationship between the robust data generation to the initial perturbation and the diffusion speed cost given by the 2-Wasserstein distance or the entropy production rate.

- We discuss the optimality and suboptimality of the forward diffusion process in terms of the trade-off, and we found the theoretical validity of the well-used methods (i.e., the cosine and the conditional optimal transport schedule.)

For more information and examples, see K. Ikeda, T. Uda, D. Okanohara and SI, arXiv:2407.04495.

Take-home message:

Stochastic thermodynamics (based on optimal transport) is useful for generative AI.