FROM OPTIMAL SCORE MATCHING TO OPTIMAL SAMPLING

Harrison H. Zhou

Department of Statistics and Data Science

Yale University

Zehao Dou
OpenAI

Subhodh Kotekal
Chicago —> MIT

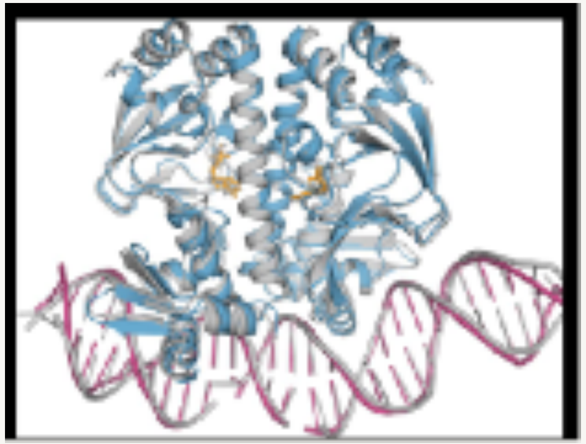
Zhehao Xu
Yale

Table of Contents

Diffusion Models

DALL-E 3
(Betker et al.)

Stable Diffusion 3
(Esser et al.)



AlphaFold 3
(Abramson et al.)

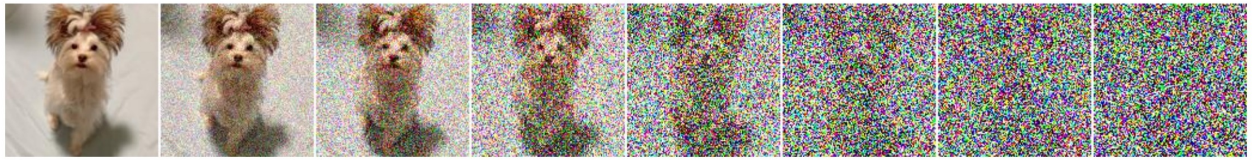
Basics of Diffusion Models

(Ho et al.; Song et al.)

Forward Process:

Forward SDE (data → noise)

$\mathbf{x}(0) \longrightarrow \mathrm{d}\mathbf{x} = \mathbf{f}(\mathbf{x}, t)\mathrm{d}t + g(t)\mathrm{d}\mathbf{w} \longrightarrow \mathbf{x}(T)$



Backward Process:

score function

$\mathbf{x}(0) \longleftarrow \mathrm{d}\mathbf{x} = \left[\mathbf{f}(\mathbf{x}, t) - g^2(t)\boxed{\nabla_{\mathbf{x}} \log p_t(\mathbf{x})}\right] \mathrm{d}t + g(t)\mathrm{d}\bar{\mathbf{w}} \longleftarrow \mathbf{x}(T)$

Reverse SDE (noise → data)

➢ In this talk, the drift $f = 0$ and diffusion coefficient $g(t) = 1$.

➢ The backward process can be replaced by Probability Flow ODE:

$$\mathrm{d}\mathbf{x} = \left[\mathbf{f}(x, t) - \frac{1}{2}g^2(t)\nabla_x \log p_t(\mathbf{x})\right] \mathrm{d}t.$$

with the same marginal distributions at all time steps as reverse SDE.

Score Function

Given $\mu_1,\ \mu_2,\ \cdots,\ \mu_n \sim f$ on [-1,1], how to estimate the score $s(x,t)$ on of distribution at time t? What is the minimax statistical error rate?

$$s(x,t) := \frac{\partial}{\partial x} \log p(x,t)$$

$$p(x,t) = (\varphi_t * f)(x)$$

where $\varphi_t(x) = \frac{1}{\sqrt{2\pi t}} e^{-\frac{x^2}{2t}}$ is the density of Gaussian distribution $\mathcal{N}(0,t)$.

Algorithm: Backward SDE with Estimated Score

Algorithm 1 Diffusion Estimation Algorithm

1: **Input:** Data $\{\mu_i\}_{i=1}^n$ drawn i.i.d. from f.
2: Calculate $\widehat{s}(\cdot, t)$ for all $t \in [0, T]$ with $T := n$.
3: Solve the SDE:

$$\mathrm{d}\widehat{Y}_t = \widehat{s}(\widehat{Y}_t, T - t)\,\mathrm{d}t + \mathrm{d}W_t$$

with initialization $\widehat{Y}_0 \sim \mathcal{N}(0, T)$.
4: **Output:** $\widehat{X}_0 := \widehat{Y}_T \mathbb{1}_{\{|\widehat{Y}_T| \leq 1\}}$.

➢ Note that $\mathrm{d}_{\mathrm{KL}}(p(\cdot, T) || \mathcal{N}(0, T)) \lesssim \frac{1}{T}$

Statistical Understanding for Diffusion Models

➢ Girsanov Formula: For the score-based diffusion model with $f(x_t, t) \equiv 0$ and $g(t) = 1$. Denote X_0, $\hat{X}_0$ to be the true distribution and generated distribution. We have:

$$\mathrm{d_{TV}}(\hat{X}_0, X_0)^2 \lesssim \mathrm{d_{TV}}\left(p(\cdot, T), \pi_0\right)^2 + \int_0^T \int_{-\infty}^{\infty} |\hat{s}(x,t) - s(x,t)|^2 \, p(x,t) \, \mathrm{d}x \, \mathrm{d}t.$$

➢ We use the Mean Integrated Squared Error:

$$\mathbb{E}\left(\int_{-\infty}^{\infty} |\hat{s}(x,t) - s(x,t)|^2 \, p(x,t) \, \mathrm{d}x\right)$$

➢ Optimal score matching? Optimal sampling?

Function class for Diffusion Models

➢ Define the Hölder class of probability density function

$$\mathcal{F}_\alpha := \left\{f : [-1,1]^d \to [0,\infty) : f \in \mathcal{H}_\alpha,\ c_d \le f(x) \le C_d\right\}$$

➢ The density estimation in Hölder class $\mathscr{F}_\alpha$ has minimax TV error rate $n^{-\frac{\alpha}{2\alpha+d}}$, and Wasserstein-1 error rate $\max\{n^{-\frac{\alpha+1}{2\alpha+d}}, n^{-1/2}\}$ (up to a *log n* factor for $d = 2$).

Table of Contents

A (Kernel Density) Plug-in Estimator

- ➢ Score function $$s(x,t) = \frac{\psi(x,t)}{p(x,t)}$$

- ➢ Density estimation $$\hat{p}(x,t) = \varphi_t * \hat{f}(x)$$

- ➢ Derivative of density estimation

$$\hat{\psi}(x,t) = \varphi_t(x+1)\hat{f}(-1) - \varphi_t(x-1)\hat{f}(1) + \int_{-1}^{1} \varphi_t(x-\mu)\hat{f}'(\mu)\,\mathrm{d}\mu$$

- ➢ Note that $$\psi(x,t) = \frac{\partial}{\partial x}p(x,t) = (\varphi_t * f)'(x)$$

Risk Upper Bound

Theorem (Upper Bound: Low Noise)

If $\alpha > 0$ and $t < n^{-\frac{2}{2\alpha+1}}$, score estimator $\hat{s}$ achieves the error

$$\sup_{f \in \mathcal{F}_\alpha} \mathbb{E} \left(\int_{\mathbb{R}} |\hat{s}(x,t) - s(x,t)|^2 \, p(x,t) \, \mathrm{d}x \right) \lesssim n^{-\frac{2(\alpha-1)}{2\alpha+1}} \vee t^{\alpha-1}.$$

An Unbiased Estimator

- Score function

$$s(x,t) = \frac{\psi(x,t)}{p(x,t)} = \frac{\int_{-1}^{1} -\frac{x-\mu}{t}\varphi_t(x-\mu)f(\mu)\,\mathrm{d}\mu}{\int_{-1}^{1} \varphi_t(x-\mu)f(\mu)\,\mathrm{d}\mu}$$

- Unbiased estimation for both $\psi(x,t)$ and $p(x,t)$,

$$\hat{\psi}(x,t) = -\frac{1}{n}\sum_{j=1}^{n}\frac{x-\mu_j}{t}\varphi_t(x-\mu_j), \;\; \hat{p}(x,t) = \frac{1}{n}\sum_{j=1}^{n}\varphi_t(x-\mu_j)$$

- Note that $\psi(x,t) = (\varphi_t' * f)(x)$

Risk Upper Bound

Theorem (Upper Bound: High Noise)

If $t \leq 1$, the score estimator $\hat{s}$ achieves the error

$$\sup_{f \in \mathcal{F}_\alpha} \mathbb{E}\left(\int_{\mathbb{R}} |\hat{s}(x,t) - s(x,t)|^2 \, p(x,t) \, \mathrm{d}x\right) \lesssim \frac{1}{nt^{3/2}}.$$

Another Unbiased Estimator

➢ Score function

$$s(x,t) = \frac{\psi(x,t)}{p(x,t)} = \frac{\int_{-1}^{1} -\frac{x-\mu}{t}\varphi_t(x-\mu)f(\mu)\,\mathrm{d}\mu}{\int_{-1}^{1} \varphi_t(x-\mu)f(\mu)\,\mathrm{d}\mu}$$

where

$$\psi(x,t) = (\varphi_t' * f)(x) = \int_{-1}^{1} -\frac{x-\mu}{t}\varphi_t(x-\mu)f(\mu)\,\mathrm{d}\mu = -\frac{x}{t}p(x,t) + \frac{1}{t}\int_{-1}^{1} \mu\varphi_t(x-\mu)f(\mu)\,\mathrm{d}\mu.$$

➢ Unbiased estimation

$$\hat{\psi}(x,t) = -\frac{x}{t}\cdot\hat{p}(x,t) + \frac{1}{nt}\sum_{j=1}^{n}\mu_j\varphi_t(x-\mu_j)$$

➢ Note that $\psi(x,t) = (\varphi_t' * f)(x)$

Risk Upper Bound

Theorem (Upper Bound: Very High Noise)

If $t > 1$, the score estimator $\hat{s}$ achieves the error

$$\sup_{f \in \mathcal{F}_\alpha} \mathbb{E} \left(\int_{\mathbb{R}} |\hat{s}(x,t) - s(x,t)|^2 \, p(x,t) \, \mathrm{d}x \right) \lesssim \frac{1}{nt^2}.$$

Goal: Rate-Optimal Score Matching

Theorem (Main Result)

Let $\alpha > 0$. The minimax estimation error for score matching at time step t is:

$$\inf_{\hat{s}} \sup_{f \in \mathcal{F}_\alpha} \mathbb{E}\left(\int_{-\infty}^{\infty} \left|\hat{s}(x,t) - s(x,t)\right|^2 p(x,t)\,\mathrm{d}x\right)$$
$$\asymp \frac{1}{nt^2} \wedge \frac{1}{nt^{3/2}} \wedge \left(n^{-\frac{2(\alpha-1)}{2\alpha+1}} + t^{\alpha-1}\right)$$

for all $t \geq 0$.

The transition points are: $t = 1$ and $t = n^{-\frac{2}{2\alpha+1}}$.

Table of Contents

Lower Bound

➢ For illustration, we shift focus to proving a lower bound for the target

$$\frac{\partial}{\partial x}p(x,t) = (\varphi_t * f)'(x)$$

with respect to L^2-norm.

➢ Introduce free parameter $\rho > 0$, for unknown $b_i \in \{0,1\}$, set unknown density:

$$f_b(\mu) = \frac{1}{2}1_{\{|\mu|\leq 1\}} + \epsilon^\alpha \sum_{i=1}^{m} b_i w\left(\frac{\mu - x_i}{\rho}\right)$$

such that $\rho \geq \epsilon$ (to satisfy Hölder) and $m \asymp \frac{1}{\rho}$.

➢ Classical construction: $\rho = \epsilon = n^{-\frac{1}{2\alpha+1}}$

Lower Bound

➢ A key inequality:

$$\| ((f_b - f_{b'}) * \varphi_t)' \|^2 \geq \frac{\epsilon^{2\alpha}}{\rho^2} \cdot d_{\mathrm{Ham}}(b, b') \cdot \rho \cdot \left(1 - \frac{Ct}{\rho^2}\right)$$

➢ If a function $h \in C^\infty(\mathbb{R})$ is compactly supported, we have

$$\|h * \varphi_t\|^2 \geq \|h\|^2 - t\|h'\|^2.$$

Rate-Optimal Score Matching

After combining up everything, we have the minimax rate for score matching at any $t > 0$.

Theorem (Main Result)

Let $\alpha > 0$. The minimax estimation error for score matching at time step t is:

$$\inf_{\hat{s}} \sup_{f \in \mathcal{F}_\alpha} \mathbb{E}\left(\int_{-\infty}^{\infty} |\hat{s}(x,t) - s(x,t)|^2 \, p(x,t) \, \mathrm{d}x \right) \asymp \frac{1}{nt^2} \wedge \frac{1}{nt^{3/2}} \wedge \left(n^{-\frac{2(\alpha-1)}{2\alpha+1}} + t^{\alpha-1} \right)$$

for all $t \geq 0$.

The transition points are: $t = 1$ and $t = n^{-\frac{2}{2\alpha+1}}$.

Some Comments

- ➢ Early stopping is not required.
- ➢ At different t, we have different estimation accuracy. Sharp minimax rates without extraneous logarithmic terms.
- ➢ Consider all smoothness α

Table of Contents

Rate-Optimal Estimation under TV

Finally, we apply Girsanov's theorem and take integral of the score matching minimax rate:

$$\mathbb{E}\left(\mathrm{d}_{\mathrm{TV}}(\hat{f}, f)^2\right) \lesssim \int_0^{n^{-\frac{2}{2\alpha+1}}} \left(n^{-\frac{2(\alpha-1)}{2\alpha+1}} + t^{\alpha-1}\right) \mathrm{d}t + \int_{n^{-\frac{2}{2\alpha+1}}}^{T\wedge 1} \frac{1}{nt^{3/2}}\,\mathrm{d}t + \int_{T\wedge 1}^{T\vee 1} \frac{1}{nt^2}\,\mathrm{d}t \asymp n^{-\frac{2\alpha}{2\alpha+1}}.$$

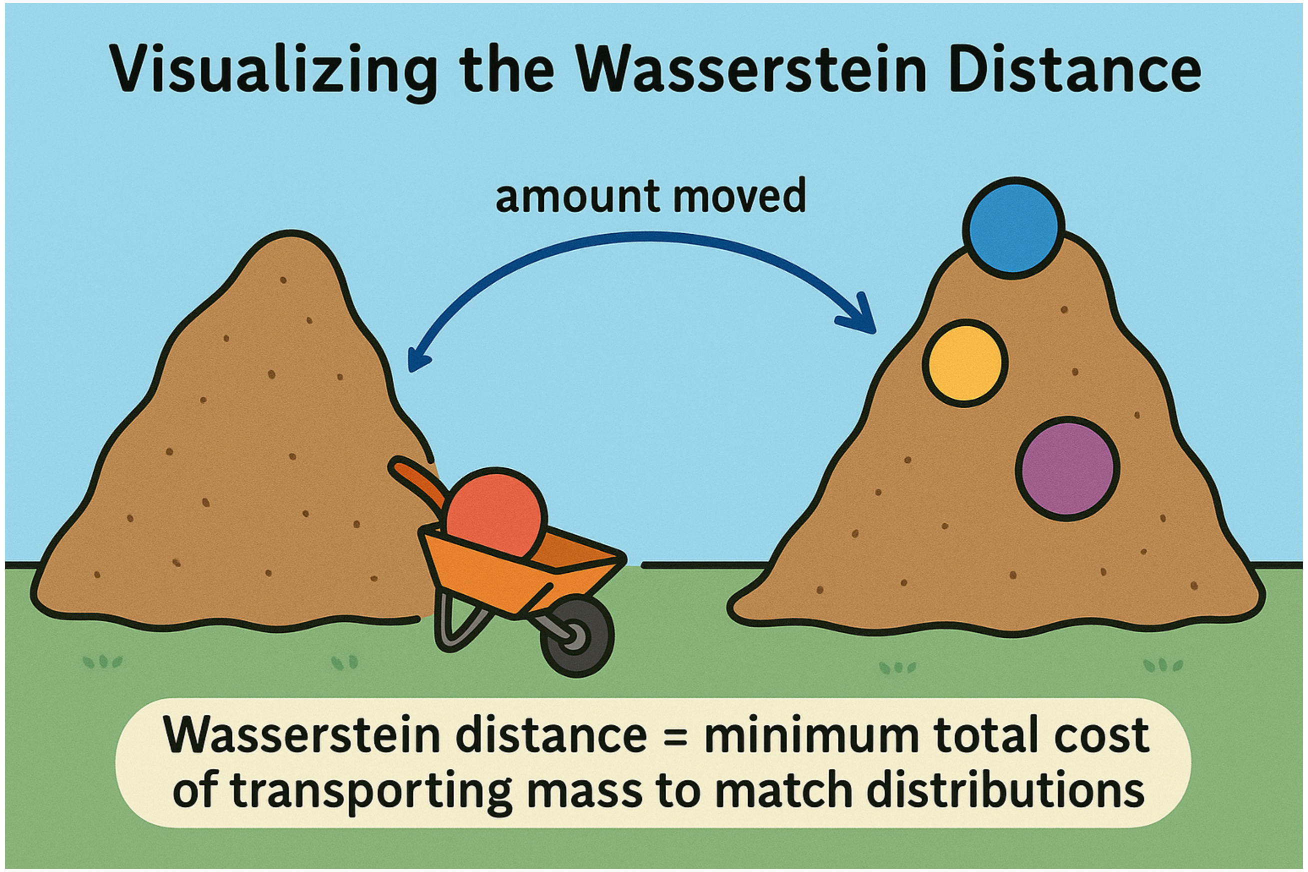
Theorem (Main Result)

Let $\alpha > 0$. The minimax estimation error for score matching at time step t is:

$$\inf_{\hat{s}} \sup_{f\in\mathcal{F}_\alpha} \mathbb{E}\left(\int_{-\infty}^{\infty} \left|\hat{s}(x,t) - s(x,t)\right|^2 p(x,t)\,\mathrm{d}x\right) \asymp \frac{1}{nt^2} \wedge \frac{1}{nt^{3/2}} \wedge \left(n^{-\frac{2(\alpha-1)}{2\alpha+1}} + t^{\alpha-1}\right)$$

for all $t \geq 0$.

Rate-Optimal Estimation Under Wasserstein Distance



- For $d = 1$, if X_0 has a CDF F and X_1 has a CDF G, the optimal transport is $G^{-1}(F(X_0))$

Rate-Optimal Estimation Under Wasserstein Distance

- Using dyadic grid $t_i = 2^i/n$ for $i = 0, \ldots, N = \lfloor \log_2(n) \rfloor$, we decompose:

$$\mathbb{E}(\mathrm{W}_1(X_0, \widehat{X}_0)) = \mathbb{E}(\mathrm{W}_1(Y_T^0, \widehat{Y}_T^T)) \leq \sum_{i=0}^{N} \mathbb{E}(\mathrm{W}_1(Y_T^{t_i}, Y_T^{t_{i+1}}))$$

- "Transport mass": $\sqrt{\int_{t_i}^{t_{i+1}} \int_{\mathbb{R}} \mathbb{E}(|s(x,t) - \widehat{s}(x,t)|^2)\, p(x,t)\, \mathrm{d}x\, \mathrm{d}t}$

- $$\mathbb{E}(\mathrm{W}_1(Y_T^{t_i}, Y_T^{t_{i+1}})) \lesssim \sqrt{t_{i+1}} \cdot \sqrt{\int_{t_i}^{t_{i+1}} \int_{\mathbb{R}} \mathbb{E}(|s(x,t) - \widehat{s}(x,t)|^2)\, p(x,t)\, \mathrm{d}x\, \mathrm{d}t}$$

- Upper bound $1/\sqrt{n}$

Extension to Multivariate Case

To generalize our result to the multivariate case with dimension $d > 1$, we define the class:

$$\mathcal{H}_\alpha(L) = \Big\{ f : [-1,1]^d \to [0,\infty) : \int_{[-1,1]^d} f(x)\,\mathrm{d}x = 1, f \text{ is continuous},$$

$$f \text{ admits all } \lfloor \alpha \rfloor \text{ partial derivatives on } (-1,1)^d,$$

$$\text{and } \max_{S \in [d]^{\lfloor \alpha \rfloor}} |\partial_S f(x) - \partial_S f(y)| \leq L||x-y||^{\alpha - \lfloor \alpha \rfloor} \text{ for all } x, y \in (-1,1)^d \Big\}.$$

$$\mathcal{F}_\alpha := \{ f : [-1,1]^d \to [0,\infty) : f \in \mathcal{H}_\alpha,\ c_d \leq f(x) \leq C_d \}$$

The proof techniques are the same as the $d = 1$ case, with only trivial differences.

Extension to Multivariate Case

Theorem (Extension to d-dimensional Case)

Let $\alpha > 0$. The minimax estimation error for score matching at time step t is:

$$\inf_{\hat{s}} \sup_{f \in \mathcal{F}_\alpha} \mathbb{E}\left(\int_{-\infty}^{\infty} |\hat{s}(x,t) - s(x,t)|^2 \, p(x,t) \, \mathrm{d}x\right) \asymp \frac{1}{nt^{d/2+1}} \wedge \left(n^{-\frac{2(\alpha-1)}{2\alpha+d}} + t^{\alpha-1}\right)$$

for all $0 \leq t \lesssim 1$.

Rate-Optimal Estimation Under Total Variation Distance

Theorem

For $\alpha > 0$, there exists a constant $C = C(\alpha, L, d)$ depending only on α, L, and d such that:

$$\sup_{f \in \mathcal{F}_\alpha} \mathbb{E}\left(\mathrm{d}_{\mathrm{TV}}(X_0, \widehat{X}_0)\right) \leq Cn^{-\frac{\alpha}{2\alpha+d}},$$

where $\widehat{X}_0$ is given by Algorithm 1.

Rate-Optimal Estimation Under Wasserstein Distance

Theorem

For $\alpha \geq 1$, there exists a constant $C = C(\alpha, L, d)$ depending only on α, L and d such that

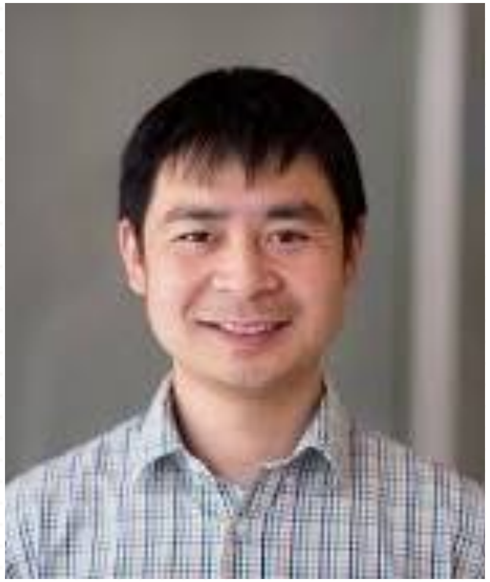
$$\sup_{f \in \mathcal{F}_\alpha} \mathbb{E}\left(\mathrm{W}_1(X_0, \widehat{X}_0)\right) \leq \begin{cases} Cn^{-\frac{1}{2}} & \text{if } d = 1, \\ Cn^{-\frac{1}{2}} \log(n) & \text{if } d = 2, \\ Cn^{-\frac{\alpha+1}{2\alpha+d}} & \text{if } d \geq 3, \end{cases}$$

where $\widehat{X}_0$ is given by Algorithm 1.

This result matches the upper bound in (Niles-Weed 2022) and achieves the minimax rate for Wasserstein distance (up to a logarithmic factor when $d = 2$).

Table of Contents

Estimation of Optimal Transport



Qiang Liu
UT Austin

Leda Wang
Yale

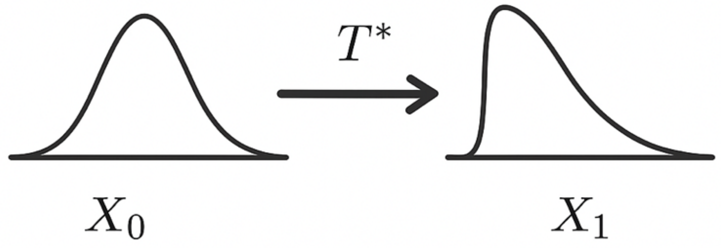
Zhehao Xu
Yale

Rectified Flow (Liu, 2022)

- ➢ Find a transport T from X_0 to X_1, $(Z_0, Z_1) = \text{Rectify}(X_0, X_1)$, by a flow $\mathrm{d}Z_t = v_t(Z_t)\mathrm{d}t$, where $v_t(x) = \mathbb{E}[X_1 - X_0 \mid tX_1 + (1-t)X_0 = x]$.

- ➢ Reduce the transport cost: $\mathbb{E}[\|Z_1 - Z_0\|^2] \leq \mathbb{E}[\|X_1 - X_0\|^2]$.

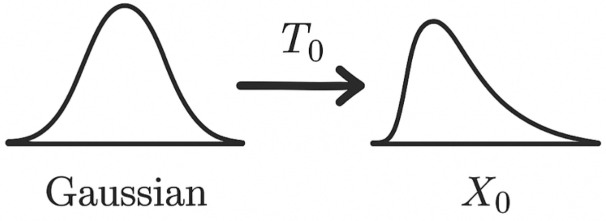
- ➢ It can be shown that applying rectified flow iteratively to obtain the optimal transport T^*: $(Z_0, Z_1) = \text{Rectify}(X_0, X_1)$.

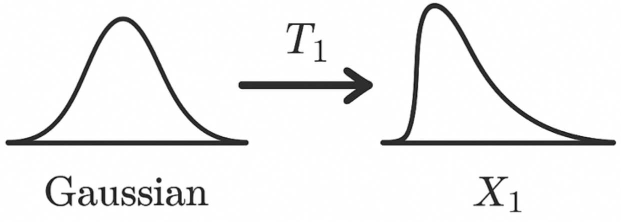


Rectified Flow and Score Matching

- If $Y_0 \sim \mathrm{N}(0,\mathrm{I})$, it can be show that $v_t(y) = \dfrac{y}{t} + \dfrac{t-1}{t}\nabla \log(p_t(y))$, where p_t is the PDF of $Y_t = tY_1 + (1-t)Y_0$.

- Apply rectified flow with a source distribution Gaussian to estimate $\mathscr{L}(X_0)$ to $\mathscr{L}(X_1)$. Denote the estimators by $\mathscr{L}(\widehat{X}_0)$ and $\mathscr{L}(\widehat{X}_1)$ respectively.

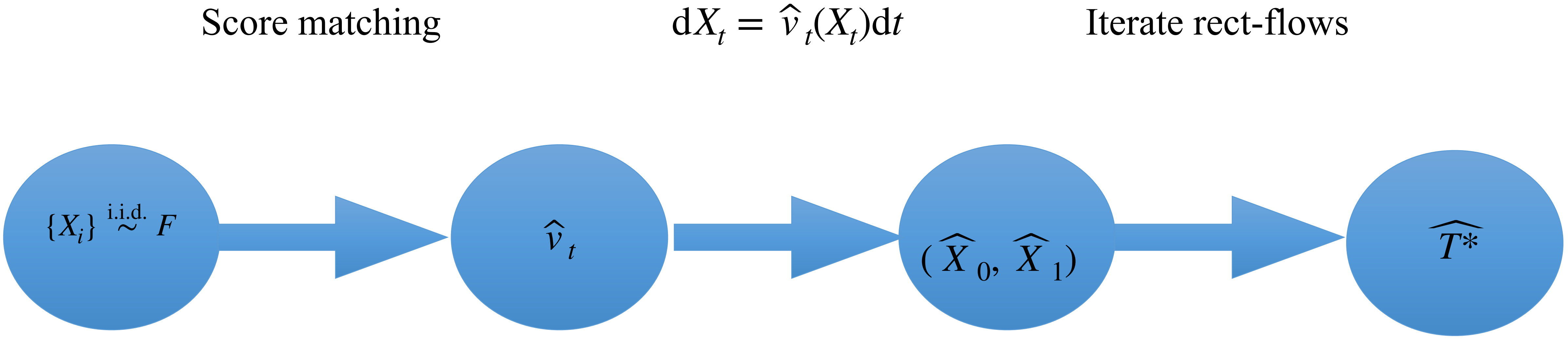




Optimal Estimation of Optimal Transport

- Apply rectified flow iteratively to obtain the optimal transport $\widehat{T^*}$ from $\mathscr{L}(\widehat{X}_0)$ to $\mathscr{L}(\widehat{X}_1)$.

- $\mathbb{E}_{x\sim X_0}[\|T^*(X) - \widehat{T^*}(X)\|^2] \lesssim max\{\mathrm{W}_2^2(X_0, \widehat{X}_0), \mathrm{W}_2^2(X_1, \widehat{X}_1)\}$(Manole et al., 2024). Obtain rate-optimal estimation under the Hölder class assumption of this talk.



Summary

- Rate-Optimal Score Matching
- Distribution Estimation under Total-variation and Wasserstein Distances
- Optimal Transport Estimation