Deep Generative Learning via Schrödinger Bridge
Supplementary Material

A. Proofs

A.1. Proof of Theorem 1

Theorem 1 *(Léonard, 2014) If $\mu, \nu \ll \mathscr{L}$, then SBP admits a unique solution $\mathbf{Q}^* = f^*(X_0)g^*(X_1)\mathbf{P}_\tau$, where f^*, g^* are $\mathscr{L}$-measurable nonnegative functions on $\mathbb{R}^d$ satisfying the Schrödinger system* $\begin{cases} f^*(\mathbf{x})\mathbb{E}_{\mathbf{P}_\tau}\left[g^*\left(X_1\right) \mid X_0 = \mathbf{x}\right] = \frac{\mathrm{d}\mu}{\mathrm{d}\mathscr{L}}(\mathbf{x}), \quad \mathscr{L}-a.e. \\ g^*(\mathbf{y})\mathbb{E}_{\mathbf{P}_\tau}\left[f^*\left(X_0\right) \mid X_1 = \mathbf{y}\right] = \frac{\mathrm{d}\nu}{\mathrm{d}\mathscr{L}}(\mathbf{y}), \quad \mathscr{L}-a.e. \end{cases}$

Proof: Theorem 1 follows from (Léonard, 2014). □

A.2. Proof of Theorem 2

Theorem 2 *(Dai Pra, 1991) Let*

$$\begin{aligned} \mathbf{u}_t^* &= \tau \mathbf{v}_t^* = \tau \nabla_{\mathbf{x}} \log g_t(\mathbf{x}) \\ &= \tau \nabla_{\mathbf{x}} \log \int h_\tau(t, \mathbf{x}, 1, \mathbf{y}) g_1(\mathbf{y}) \mathrm{d}\mathbf{y}. \end{aligned} \tag{1}$$

Then,

$$\mathbf{u}_t^*(\mathbf{x}) \in \arg\min_{\mathbf{u} \in \mathcal{U}} \mathbb{E}\left[\int_0^1 \frac{1}{2}\|\mathbf{u}_t\|^2 \mathrm{d}t\right]$$

s.t.

$$\begin{cases} \mathrm{d}\mathbf{x}_t = \mathbf{u}_t \mathrm{d}t + \sqrt{\tau}\mathrm{d}\mathbf{w}_t, \\ \mathbf{x}_0 \sim q(\mathbf{x}), \quad \mathbf{x}_1 \sim p(\mathbf{x}). \end{cases} \tag{2}$$

Proof: Theorem 2 follows from (Dai Pra, 1991). □

A.3. Proof of Theorem 3

Theorem 3 *Define the density ratio $f(\mathbf{x}) = \frac{q_\sigma(\mathbf{x})}{\Phi_{\sqrt{\tau}}(\mathbf{x})}$. Then for the SDE*

$$\mathrm{d}\mathbf{x}_t = \tau \nabla \log \mathbb{E}_{\mathbf{z} \sim \Phi_{\sqrt{\tau}}}[f(\mathbf{x}_t + \sqrt{1-t}\mathbf{z})]\mathrm{d}t + \sqrt{\tau}\mathrm{d}\mathbf{w}_t \tag{3}$$

with initial condition $\mathbf{x}_0 = \mathbf{0}$, we have $\mathbf{x}_1 \sim q_\sigma(\mathbf{x})$.

And, for the SDE

$$\mathrm{d}\mathbf{x}_t = \sigma^2 \nabla \log q_{\sqrt{1-t}\sigma}(\mathbf{x}_t)\mathrm{d}t + \sigma \mathrm{d}\mathbf{w}_t \tag{4}$$

with initial condition $\mathbf{x}_0 \sim q_\sigma(\mathbf{x})$, we have $\mathbf{x}_1 \sim p_{\mathrm{data}}(\mathbf{x})$.

Proof: Denote

$$f_0(\mathbf{x}) = f^*(\mathbf{x}), \;\; g_1(\mathbf{y}) = g^*(\mathbf{y}),$$

$$\begin{aligned} f_1(\mathbf{y}) &= \mathbb{E}_{\mathbf{P}_\tau}\left[f^*\left(X_0\right) \mid X_1 = \mathbf{y}\right] \\ &= \int h_\tau(0, \mathbf{x}, 1, \mathbf{y}) f_0(\mathbf{x}) \mathrm{d}\mathbf{x}, \end{aligned}$$

$$\begin{aligned} g_0(\mathbf{x}) &= \mathbb{E}_{\mathbf{P}_\tau}\left[g^*\left(X_1\right) \mid X_0 = \mathbf{x}\right] \\ &= \int h_\tau(0, \mathbf{x}, 1, \mathbf{y}) g_1(\mathbf{y}) \mathrm{d}\mathbf{y}. \end{aligned}$$

Then, the Schrödinger system in Theorem 1 can also be characterized by

$$q(\mathbf{x}) = f_0(\mathbf{x})g_0(\mathbf{x}), \;\; p(\mathbf{y}) = f_1(\mathbf{y})g_1(\mathbf{y}) \tag{5}$$

For Eq. (3), let $f_0(\mathbf{x}) = \delta_{\mathbf{0}}(\mathbf{x})$ be the Dirac delta function, $f_1(\mathbf{y}) = \int h_\tau(0, \mathbf{x}, 1, \mathbf{y}) f_0(\mathbf{x}) \mathrm{d}\mathbf{x} = \Phi_{\sqrt{\tau}}(\mathbf{y})$, $g_1(\mathbf{x}) = \frac{q_\sigma(\mathbf{x})}{\Phi_{\sqrt{\tau}}(\mathbf{x})} = f(\mathbf{x})$, $g_0(\mathbf{0}) = \int h_\tau(0, \mathbf{0}, 1, \mathbf{y}) g_1(\mathbf{y}) \mathrm{d}\mathbf{y} = 1$. Then f_i, g_i, $i = 0, 1$ solve Schrödinger system (5) with $q = \delta_{\mathbf{0}}$, $p = q_\sigma$. Define

$$\begin{aligned} g_t(\mathbf{x}) &= \int h_\tau(t, \mathbf{x}, 1, \mathbf{y}) g_1(\mathbf{y}) \mathrm{d}\mathbf{y} = \mathbb{E}_{\mathbf{y} \sim \Phi_{\sqrt{(1-t)\tau}}}[f(\mathbf{y})] \\ &= \sqrt{1-t}\mathbb{E}_{\mathbf{z} \sim \Phi_{\sqrt{\tau}}}[f(\mathbf{x} + \sqrt{1-t}\mathbf{z})]. \end{aligned}$$

By Theorem 2, $\mathbf{u}^*(\mathbf{x}) = \tau \nabla_{\mathbf{x}} \log g_t(\mathbf{x})$ solves the optimal control problem $\min_{\mathbf{u} \in \mathcal{U}} \mathbb{E}\left[\int_0^1 \frac{1}{2}\|\mathbf{u}_t\|^2 \mathrm{d}t\right]$ such that

$$\begin{cases} \mathrm{d}\mathbf{x}_t = \mathbf{u}_t \mathrm{d}t + \sqrt{\tau}\mathrm{d}\mathbf{w}_t \\ \mathbf{x}_0 \sim \delta_{\mathbf{0}}, \quad \mathbf{x}_1 \sim q_\sigma(\mathbf{x}) \end{cases}$$

i.e., the dynamic of Eq. (3) will push $\delta_{\mathbf{0}}$ onto q_σ from $t = 0$ to $t = 1$.

For Eq. (4), let $f_0(\mathbf{x}) = 1$,

$$f_1(\mathbf{y}) = \int h_{\sigma^2}(0, \mathbf{x}, 1, \mathbf{y}) f_0(\mathbf{x}) \mathrm{d}\mathbf{x} = 1,$$

$g_1(\mathbf{x}) = p_{\mathrm{data}}(\mathbf{x})$, $g_0(\mathbf{x}) = \int h_{\sigma^2}(0, \mathbf{x}, 1, \mathbf{y}) g_1(\mathbf{y}) \mathrm{d}\mathbf{y} = q_\sigma(\mathbf{x})$. Then, f_i, g_i, $i = 0, 1$ solve Schrödinger system (5) with $q = q_\sigma$, $p = p_{\mathrm{data}}$ and $\tau = \sigma^2$. Define

$$g_t(\mathbf{x}) = \int h_{\sigma^2}(t, \mathbf{x}, 1, \mathbf{y}) g_1(\mathbf{y}) \mathrm{d}\mathbf{y} = q_{\sqrt{1-t}\sigma}(\mathbf{x}).$$

By Theorem 2, $\mathbf{u}^*(\mathbf{x}) = \sigma^2 \nabla_{\mathbf{x}} \log g_t(\mathbf{x})$ solves the optimal control problem $\min_{\mathbf{u} \in \mathcal{U}} \mathbb{E}\left[\int_0^1 \frac{1}{2}\|\mathbf{u}_t\|^2 \mathrm{d}t\right]$ such that

$$\begin{cases} \mathrm{d}\mathbf{x}_t = \mathbf{u}_t \mathrm{d}t + \sigma \mathrm{d}\mathbf{w}_t \\ \mathbf{x}_0 \sim q_\sigma(\mathbf{x}), \quad \mathbf{x}_1 \sim p_{\mathrm{data}}(\mathbf{x}) \end{cases}$$

i.e., the dynamic of Eq. (4) will push q_σ onto p_{data} from $t = 0$ to $t = 1$. □

A.4. Proof of Theorem 4

Theorem 4 *Assume that the support of $p_{\text{data}}(\mathbf{x})$ is contained in a compact set, and $f(\mathbf{x})$ is Lipschitz continuous and bounded. Set the depth $\mathcal{D}$, width $\mathcal{W}$, and size $\mathcal{S}$ of $\mathcal{NN}_\phi$ as*

$$\mathcal{D} = \mathcal{O}(\log(n)), \mathcal{W} = \mathcal{O}(n^{\frac{d}{2(2+d)}}/\log(n)),$$

$$\mathcal{S} = \mathcal{O}(n^{\frac{d-2}{d+2}}\log(n)^{-3}).$$

Then $\mathbb{E}[\|\hat{f}(\mathbf{x}) - f(\mathbf{x})\|_{L^2(p_{\text{data}})}] \to 0$ as $n \to \infty$.

Proof: Recall that

$$\hat{f}(\mathbf{x}) = \exp(\hat{r}_\phi(\mathbf{x})), \tag{6}$$

where $\hat{r}_\phi \in \mathcal{NN}_\phi$ is the neural network that minimizes the empirical loss:

$$\hat{r}_\phi \in \arg\min_{r_\phi \in \mathcal{NN}_\phi} \hat{L}(r_\phi), \text{ where } \hat{L}(r_\phi) = \frac{1}{n}\sum_{i=1}^{n}[\log(1+\exp(-r_\phi(\widetilde{\mathbf{x}}_i))) + \log(1+\exp(r_\phi(\mathbf{z}_i)))], \tag{7}$$

$\widetilde{\mathbf{x}}_1, ..., \widetilde{\mathbf{x}}_n$ are i.i.d. samples from $q_\sigma(\mathbf{x})$, and $\mathbf{z}_1, ..., \mathbf{z}_n$ are i.i.d. samples from $\Phi_{\sqrt{\tau}}(\mathbf{x})$. Note that $f(\mathbf{x}) = \exp(r^*(\mathbf{x}))$ with

$$r^* \in \arg\min_r \mathcal{L}(r),$$

where $\mathcal{L}(r) = \mathbb{E}_{q_\sigma(\mathbf{x})}\log(1+\exp(-r(\mathbf{x}))) + \mathbb{E}_{\Phi_{\sqrt{\tau}}(\mathbf{x})}\log(1+\exp(r(\mathbf{x})))$.

Theorem 4 follows by showing $\|\hat{r}_\phi - r^*\|_{L^2(p_{\text{data}})} \to 0$ as $n \to \infty$. By the assumption that $r^*(\mathbf{x})$ is Lipschitz continuous on a compact set and bounded, we use L_1 and B_1 to denote its Lipschitz constant and the upper bound. Without loss of generality, we use $E = [-C, C]^d$ to denote its domain. By Lemma 1 (given in A.6) with $L = \log n$, $N = n^{\frac{d}{2(2+d)}}/\log n$, there exists a $\bar{r}_\phi \in \mathcal{NN}_\phi$ with depth $\mathcal{D} = 12\log n + 14 + 2d$, width $\mathcal{W} = 3^{d+3}\max\{d(n^{\frac{d}{2(2+d)}}/\log n)^{\frac{1}{d}}, n^{\frac{d}{2(2+d)}}/\log n + 1\}$, and size $\mathcal{S} = n^{\frac{d-2}{d+2}}/(\log^4 n)$, $\mathcal{B} = 2B_1$, such that

$$\|\bar{r}_\phi - r^*\|_{L^2(p_{\text{data}})} \le 38 L_1 C\sqrt{d} n^{-\frac{1}{d+2}}. \tag{8}$$

Using Taylor expansion and the boundness of $r_\phi \in \mathcal{NN}_\phi$ and r^*, it is easy to show that $\mathcal{L}(r_\phi) - \mathcal{L}(r^*)$ is sandwiched by $\|\bar{r}_\phi - r^*\|^2_{L^2(p_{\text{data}})}$, i.e., $\forall r_\phi \in \mathcal{NN}_\phi$

$$\begin{aligned} C_{1,\mathcal{B}}\|r_\phi - r^*\|^2_{L^2(p_{\text{data}})} &\le \mathcal{L}(r_\phi) - \mathcal{L}(r^*) \\ &\le C_{2,\mathcal{B}}\|r_\phi - r^*\|^2_{L^2(p_{\text{data}})}. \end{aligned} \tag{9}$$

Then,

$$\begin{aligned} &C_{1,\mathcal{B}}\|\hat{r}_\phi - r^*\|^2_{L^2} \le \mathcal{L}(\hat{r}_\phi) - \mathcal{L}(r^*) \\ =&\mathcal{L}(\hat{r}_\phi) - \hat{\mathcal{L}}(\hat{r}_\phi) + \hat{\mathcal{L}}(\hat{r}_\phi) - \hat{\mathcal{L}}(\bar{r}_\phi) \\ &+ \hat{\mathcal{L}}(\bar{r}_\phi) - \mathcal{L}(\bar{r}_\phi) + \mathcal{L}(\bar{r}_\phi) - \mathcal{L}(r^*) \\ \le&2\sup_{r\in\mathcal{NN}_\phi}|\mathcal{L}(r) - \hat{\mathcal{L}}(r)| + C_{2,\mathcal{B}}\|\bar{r}_\phi - r^*\|^2_{L^2(\nu)} \\ \le&2\sup_{r\in\mathcal{NN}_\phi}|\mathcal{L}(r) - \hat{\mathcal{L}}(r)| + 38 C_{2,\mathcal{B}} L_1 C\sqrt{d} n^{-\frac{1}{d+2}}, \end{aligned} \tag{10}$$

where we use the definition of $\hat{r}_\phi$, r^*, and $\bar{r}_\phi$, as well as (8) and (9). Next, we finish the proof by bounding the empirical process term in (10). Let $\mathbf{O} = (\tilde{\mathbf{x}}, \mathbf{z})$ be the random variable pair, with $\mathbf{x} \sim p_{\text{data}}$, $\mathbf{z} \sim \Phi_{\sqrt{\tau}}$, and $\{\mathbf{O}_i\}_{i=1}^n$ be n i.i.d. copies of $\mathbf{O}$. Let $\mathbf{o} = (\tilde{x}, z) \in \mathbb{R}^d \times \mathbb{R}^d$ be a realization of $\mathbf{O}$, and define

$$b(r, \mathbf{o}) = \log(1+\exp^{-r(\tilde{x})}) + \log(1+\exp^{r(z)}).$$

It is easy to check that $b(r, \mathbf{o})$ is 1-Lipschitz on r, i.e.,

$$|b(r,\mathbf{o}) - b(\tilde{r},\mathbf{o})| \le |r(\tilde{x}) - \tilde{r}(\tilde{x})| + |r(z) - \tilde{r}(z)|. \tag{11}$$

Let $\widetilde{\mathbf{O}}_i$ be a ghost i.i.d. copy of $\mathbf{O}_i$, and $\sigma_i(\epsilon_i)$ be the i.i.d. Rademacher random (standard normal) variables that are independent with $\widetilde{\mathbf{O}}_i$ and $\mathbf{O}_i$, $i = 1, ...n$. We need the following results (12)-(13) to upper bound the expected value of the right hand side term in (10).

$$\mathbb{E}_{\{\mathbf{O}_i\}_{i=1}^n}[\sup_r |\mathcal{L}(r) - \hat{\mathcal{L}}(r)|] \le \mathcal{O}(\mathcal{G}(\mathcal{NN})), \tag{12}$$

where $\mathcal{G}(\mathcal{NN})$ is the Gaussian complexity (Bartlett & Mendelson, 2002) of $\mathcal{NN}_\phi$ defined as

$$\mathcal{G}(\mathcal{NN}) = \mathbb{E}_{\{\mathbf{O}_i,\epsilon_i\}_i^n}[\sup_{r\in\mathcal{NN}_\phi}|\frac{1}{n}\sum_{i=1}^n \epsilon_i b(r, \mathbf{O}_i)|].$$

Proof of (12).
Obviously,

$$\mathcal{L}(r) = \mathbb{E}_{\mathbf{O}}[b(r,\mathbf{O})] = \frac{1}{n}\mathbb{E}_{\widetilde{\mathbf{O}}_i}[b(r,\widetilde{\mathbf{O}}_i)],$$

and

$$\widehat{\mathcal{L}}(r) = \frac{1}{n}\sum_{i=1}^n b(r, \mathbf{O}_i).$$

Let

$$\mathcal{R}(\mathcal{NN}) = \frac{1}{n}\mathbb{E}_{\{\mathbf{O}_i,\sigma_i\}_i^n}[\sup_{r\in\mathcal{NN}_\phi}|\sum_{i=1}^n \sigma_i b(r, \mathbf{O}_i)|]$$

be the Rademacher complexity of $\mathcal{NN}_\phi$ (Bartlett & Mendel-

son, 2002). Then,

$$\begin{aligned}
&\mathbb{E}_{\{\mathbf{O}_i\}_{i=1}^n}[\sup_r|\mathcal{L}(r)-\hat{\mathcal{L}}(r)|]\\
=&\frac{1}{n}\mathbb{E}_{\{\mathbf{O}_i\}_i^n}[\sup_r|\sum_{i=1}^n(\mathbb{E}_{\widetilde{\mathbf{O}}_i}[b(r,\widetilde{\mathbf{O}}_i)]-b(r,\mathbf{O}_i))|]\\
\leq&\frac{1}{n}\mathbb{E}_{\{\mathbf{O}_i,\widetilde{\mathbf{O}}_i\}_i^n}[\sup_r|b(r,\widetilde{\mathbf{O}}_i)-b(r,\mathbf{O}_i)|]\\
=&\frac{1}{n}\mathbb{E}_{\{\mathbf{O}_i,\widetilde{\mathbf{O}}_i,\sigma_i\}_i^n}[\sup_r|\sum_{i=1}^n\sigma_i(b(r,\widetilde{\mathbf{O}}_i)-b(r,\mathbf{O}_i))|]\\
\leq&\frac{1}{n}\mathbb{E}_{\{\mathbf{O}_i,\sigma_i\}_i^n}[\sup_r|\sum_{i=1}^n\sigma_i b(r,\mathbf{O}_i)|]\\
&+\frac{1}{n}\mathbb{E}_{\{\widetilde{\mathbf{O}}_i,\sigma_i\}_i^n}[\sup_r|\sum_{i=1}^n\sigma_i b(r,\widetilde{\mathbf{O}}_i)|]\\
=&2\mathcal{R}(b\circ\mathcal{NN})\\
\leq&4\mathcal{R}(\mathcal{NN})\\
\leq&\mathcal{O}(\mathcal{G}(\mathcal{NN})),
\end{aligned}$$

where the first inequality follows from Jensen's inequality, and the second equality holds since both $\sigma_i(b(r,\widetilde{\mathbf{O}}_i)-b(r,\mathbf{O}_i))$ and $b(r,\widetilde{\mathbf{O}}_i)-b(D,\mathbf{O}_i)$ are governed by the same law, and the last equality holds since the distribution of the two terms are the same. In the third inequality, we use the Lipschitz contraction property of Rademacher complexity, see Theorem 12 in (Bartlett & Mendelson, 2002) and (11). The last inequality holds since the relationship between the Gaussian complexity and the Rademacher complexity, see for Lemma 4 in (Bartlett & Mendelson, 2002).

Next, we bound the Gaussian complexity.

$$\mathcal{G}(\mathcal{NN})\leq \mathcal{O}(\mathcal{B}\sqrt{\frac{n}{\mathcal{DS}\log\mathcal{S}}}\log\frac{n}{\mathcal{DS}\log\mathcal{S}}\exp(-\log^2\frac{n}{\mathcal{DS}\log\mathcal{S}})). \tag{13}$$

Proof of (13).
Since $\mathcal{NN}_\phi$ is closed under negation,

$$\begin{aligned}
\mathcal{G}(\mathcal{NN})&=\mathbb{E}_{\{\mathbf{O}_i,\epsilon_i\}_i^n}[\sup_{r\in\mathcal{NN}_\phi}\frac{1}{n}\sum_{i=1}^n\epsilon_i b(r,\mathbf{O}_i)]\\
&=\mathbb{E}_{\mathbf{O}_i}[\mathbb{E}_{\epsilon_i}[\sup_{r\in\mathcal{NN}_\phi}\frac{1}{n}\sum_{i=1}^n\epsilon_i b(r,\mathbf{O}_i)]|\{\mathbf{O}_i\}_{i=1}^n].
\end{aligned}$$

Conditioning on $\{\mathbf{O}_i\}_{i=1}^n$, $\forall r,\tilde{r}\in\mathcal{NN}_\phi$, it easy to check

$$\mathbb{V}_{\epsilon_i}[\frac{1}{n}\sum_{i=1}^n\epsilon_i(b(r,\mathbf{O}_i)-b(\tilde{r},\mathbf{O}_i))]=\frac{d_{\mathcal{NN}}(r,\tilde{r})}{\sqrt{n}},$$

where

$$d_{\mathcal{NN}}(r,\tilde{r})=\frac{1}{\sqrt{n}}\sqrt{\sum_{i=1}^n(b(r,\mathbf{O}_i)-b(\tilde{r},\mathbf{O}_i))^2}.$$

Denote $\mathfrak{C}(\mathcal{NN},d_{\mathcal{NN}},\delta)$ as the covering number of $\mathcal{NN}_\phi$ under the metric $d_{\mathcal{NN}}$ with radius δ, and let $\mathrm{Pdim}_{\mathcal{NN}}$ be the Pseudo-dimension of $\mathcal{NN}_\phi$. Since the diameter of $\mathcal{NN}_\phi$ under $d_{\mathcal{NN}}$ is at most $\mathcal{B}$, we have

$$\begin{aligned}
&\mathcal{G}(\mathcal{NN})\\
\leq&\frac{c}{\sqrt{n}}\mathbb{E}_{\{\mathbf{O}_i\}_{i=1}^n}[\int_0^B\sqrt{\log\mathfrak{C}(\mathcal{NN},d_{\mathcal{NN}},\delta)}\mathrm{d}\delta]\\
\leq&\frac{c}{\sqrt{n}}\mathbb{E}_{\{\mathbf{O}_i\}_{i=1}^n}[\int_0^{\mathcal{B}}\sqrt{\log\mathfrak{C}(\mathcal{NN},d_{\mathcal{NN},\infty},\delta)}\mathrm{d}\delta]\\
\leq&\frac{c}{\sqrt{n}}\int_0^{\mathcal{B}}\sqrt{\mathrm{Pdim}_{\mathcal{NN}}\log\frac{2e\mathcal{B}n}{\delta\mathrm{Pdim}_{\mathcal{NN}}}}\mathrm{d}\delta\\
\leq&c\mathcal{B}\sqrt{\frac{n}{\mathrm{Pdim}_{\mathcal{NN}}}}\log(\frac{n}{\mathrm{Pdim}_{\mathcal{NN}}})\exp(-\log^2(\frac{n}{\mathrm{Pdim}_{\mathcal{NN}}}))\\
\leq&c\mathcal{B}\sqrt{\frac{n}{\mathcal{DS}\log\mathcal{S}}}\log\frac{n}{\mathcal{DS}\log\mathcal{S}}\exp(-\log^2\frac{n}{\mathcal{DS}\log\mathcal{S}}),
\end{aligned}$$

where c is a constant which may vary on different places, the first inequality follows from the chaining Theorem 8.1.3 in (Vershynin, 2018), the second inequality holds due to $\mathfrak{C}(\mathcal{NN},d_{\mathcal{NN}},\delta)\leq\mathfrak{C}(\mathcal{NN},d_{\mathcal{NN},\infty},\delta)$, we use the relationship between the metric entropy and the Pseudo-dimension of the ReLU networks $\mathcal{NN}_\phi$ (Anthony & Bartlett, 2009) in the third inequality, i.e.,

$$\log\mathfrak{C}(\mathcal{NN},d_{\mathcal{NN},\infty},\delta))\leq\mathrm{Pdim}_{\mathcal{NN}}\log\frac{2e\mathcal{B}n}{\delta\mathrm{Pdim}_{\mathcal{NN}}},$$

the fourth inequality follows by some calculation, and the last inequality holds due to the upper bound of Pseudo-dimension for the ReLU network $\mathcal{NN}_\psi$ satisfying

$$\mathrm{Pdim}_{\mathcal{NN}}=\mathcal{O}(\mathcal{DS}\log\mathcal{S}),$$

see (Bartlett et al., 2019).

Finally, by (10)-(13) and the choice of $\mathcal{D}$, $\mathcal{W}$ and $\mathcal{S}$, we get $\mathbb{E}[\|\hat{r}_\phi-r^*\|_{L^2}^2]\leq\mathcal{O}(n^{-\frac{2}{2+d}})\to 0$ as $n\to\infty$. □

A.5. Proof of Theorem 5

Theorem 5 *Assume that $p_{\mathrm{data}}(\mathbf{x})$ is differentiable with bounded support, and $\nabla\log q_{\tilde{\sigma}}(\mathbf{x})$ is Lipschitz continuous and bounded for $(\tilde{\sigma},\mathbf{x})\in[0,\sigma]\times\mathbb{R}^d$. Set the depth $\mathcal{D}$, width $\mathcal{W}$, and size $\mathcal{S}$ of $\mathcal{NN}_\theta$ as*

$$\mathcal{D}=\mathcal{O}(\log(n)),\mathcal{W}=\mathcal{O}(\max\{n^{\frac{d}{2(2+d)}}/\log(n),d\}),$$

$$\mathcal{S}=\mathcal{O}(dn^{\frac{d-2}{d+2}}\log(n)^{-3}).$$

Then $\mathbb{E}[\|\|\widehat{\nabla\log q_{\tilde{\sigma}}}(\mathbf{x})-\nabla\log q_{\tilde{\sigma}}(\mathbf{x})\|_2\|_{L^2(q_{\tilde{\sigma}})}]\to 0$ as $m,n\to\infty$.

Proof: We give the proof for the fixed $\tilde{\sigma}$ case. The case that $\tilde{\sigma}$ vary in a interval can be treated similarly. Recall that

$$\mathbf{s}^* \in \arg\min_{\mathbf{s}} \frac{1}{2}\mathbb{E}_{\mathbf{x}\sim q_{\tilde{\sigma}}(\mathbf{x})}\|\mathbf{s}(\mathbf{x}) - \nabla_{\mathbf{x}} \log q_{\tilde{\sigma}}(\mathbf{x})\|^2$$

is equivalent to $\mathbf{s}^* \in \arg\min_{\mathbf{s}} \mathcal{L}(\mathbf{s})$,

where $\mathcal{L}(\mathbf{s}) = \frac{1}{2}\mathbb{E}_{\mathbf{x}\sim p_{\text{data}}(\mathbf{x}),\mathbf{z}\sim\mathcal{N}(\mathbf{0},\tilde{\sigma}^2\mathbf{I})}\left\|\mathbf{s}(\mathbf{x}+\mathbf{z}) + \frac{\mathbf{z}}{\tilde{\sigma}^2}\right\|^2$.

Since $\widehat{\nabla \log q_{\tilde{\sigma}}}(\mathbf{x}) = \hat{\mathbf{s}}_\theta(\mathbf{x};\tilde{\sigma})$ (we use $\hat{\mathbf{s}}_\theta(\mathbf{x})$ to denote $\hat{\mathbf{s}}_\theta(\mathbf{x};\tilde{\sigma})$ for short), where

$$\hat{\mathbf{s}}_\theta \in \arg\min_{\mathbf{s}_\theta \in \mathcal{NN}_\theta} \hat{\mathcal{L}}(\mathbf{s}_\theta),$$

$\hat{\mathcal{L}}(\mathbf{s}_\theta) = \sum_{i=1}^n \left\|\mathbf{s}_\theta(\mathbf{x}_i + \mathbf{z}_i) + \frac{\mathbf{z}_i}{\tilde{\sigma}_j^2}\right\|^2 /(2n)$, $\mathbf{x}_i$ are i.i.d. samples from p_{data}, and $\mathbf{z}_i$ are i.i.d. samples from $\Phi_{\tilde{\sigma}}$, $i = 1, ..., n$. What we need to prove is

$$\begin{aligned}&\mathbb{E}_{\mathbf{x}_i,\mathbf{z}_i}[\|\|\mathbf{s}^* - \hat{\mathbf{s}}_\theta\|_2\|^2_{L^2(q_{\tilde{\sigma}})}]\\ =&\mathbb{E}_{\mathbf{x}_i,\mathbf{z}_i}[\mathbb{E}_{\mathbf{x}\sim q_{\tilde{\sigma}}}[\|\mathbf{s}^*(\mathbf{x}) - \hat{\mathbf{s}}_\theta(\mathbf{x})\|^2]] \to 0\end{aligned}$$

as $n \to \infty$. Since the functional $\mathcal{L}$ and $\hat{\mathcal{L}}$ are both quadratic, it is easy to conclude that

$$\begin{aligned}&\mathbb{E}_{\mathbf{x}\sim q_{\tilde{\sigma}}}[\|\mathbf{s}^*(\mathbf{x}) - \hat{\mathbf{s}}_\theta(\mathbf{x})\|^2]\\ =&\mathbb{E}_{\mathbf{x}\sim p_{\text{data}},\mathbf{z}\sim\Phi_{\tilde{\sigma}}}[\|\mathbf{s}^*(\mathbf{x}+\mathbf{z}) - \hat{\mathbf{s}}_\theta(\mathbf{x}+\mathbf{z})\|^2]\\ =&\mathcal{L}(\hat{\mathbf{s}}_\theta) - \mathcal{L}(\mathbf{s}^*)\\ =&\mathcal{L}(\hat{\mathbf{s}}_\theta) - \hat{\mathcal{L}}(\hat{\mathbf{s}}_\theta) + \hat{\mathcal{L}}(\hat{\mathbf{s}}_\theta) - \hat{\mathcal{L}}(\bar{\mathbf{s}}_\theta)\\ &+ \hat{\mathcal{L}}(\bar{\mathbf{s}}_\theta) - \mathcal{L}(\bar{\mathbf{s}}_\theta) + \mathcal{L}(\bar{\mathbf{s}}_\theta) - \mathcal{L}(\mathbf{s}^*)\\ \le&2\sup_{\mathbf{s}\in\mathcal{NN}_\theta}|\mathcal{L}(\mathbf{s}) - \hat{\mathcal{L}}(\mathbf{s})| + \mathbb{E}_{\mathbf{x}\sim q_\sigma}[\|\bar{\mathbf{s}}_\theta(\mathbf{x}) - \mathbf{s}^*\|_2^2]\\ \le&2\sup_{\mathbf{s}\in\mathcal{NN}_\theta}|\mathcal{L}(\mathbf{s}) - \hat{\mathcal{L}}(\mathbf{s})| + \inf_{\bar{\mathbf{s}}\in\mathcal{NN}_\theta}\mathbb{E}_{\mathbf{x}\sim q_\sigma}[\|\bar{\mathbf{s}}_\theta(\mathbf{x}) - \mathbf{s}^*\|_2^2],\end{aligned} \tag{14}$$

where we use $\hat{\mathbf{s}}$ as a minimizer and $\bar{\mathbf{s}}$ as an arbitrary element of $\mathcal{NN}_\theta$ in the first inequality, and we take infimum over $\bar{\mathbf{s}} \in \mathcal{NN}_\theta$ in the second inequality. We need to bound the two terms on the right hand side of (14). The terms $\inf_{\bar{\mathbf{s}}\in\mathcal{NN}_\theta}\mathbb{E}_{\mathbf{x}\sim q_\sigma}[\|\bar{\mathbf{s}}_\theta(\mathbf{x}) - \mathbf{s}^*\|_2^2]$ and $\sup_{\mathbf{s}\in\mathcal{NN}_\theta}|\mathcal{L}(\mathbf{s}) - \hat{\mathcal{L}}(\mathbf{s})|$ are the so called approximation error and statistical error. They can be bounded by using the similar technique when we prove (8) and (10), respectively. Here we directly give the bounds and omit the details. By setting,

$$\mathcal{D} = \mathcal{O}(\log(n)), \mathcal{W} = \mathcal{O}(n^{\frac{d}{2(2+d)}}/\log(n)),$$

$$\mathcal{S} = \mathcal{O}(n^{\frac{d-2}{d+2}}\log(n)^{-3}).$$

Then

$$\inf_{\bar{\mathbf{s}}\in\mathcal{NN}_\theta}\mathbb{E}_{\mathbf{x}\sim q_\sigma}[\|\bar{\mathbf{s}}_\theta(\mathbf{x}) - \mathbf{s}^*\|_2^2] \le \mathcal{O}(dn^{-\frac{2}{d+2}}),$$

$$\sup_{\mathbf{s}\in\mathcal{NN}_\theta}|\mathcal{L}(\mathbf{s}) - \hat{\mathcal{L}}(\mathbf{s})| \le \mathcal{O}(n^{-\frac{2}{d+2}}).$$

Thus, Theorem 5 follows by plugging these above two displays into (14) and setting $n \to \infty$. □

Theorem 6 *Under Assumptions 1-4,*

$$\mathbb{E}[\mathcal{W}_2(\text{Law}(\mathbf{x}_{N_2}), p_{\text{data}})] \to 0, \text{ as } n, N_1, N_2, N_3 \to \infty,$$

where $\mathcal{W}_2$ is the 2-Wasserstein distance between two distributions.

Proof: Recall that

$$D_1(t,\mathbf{x}) = \nabla \log \mathbb{E}_{\mathbf{z}\sim\Phi_{\sqrt{\tau}}}[f(\mathbf{x} + \sqrt{1-t}\mathbf{z})],$$

$$D_2(t,\mathbf{x}) = \nabla \log q_{\sqrt{1-t}\sigma}(\mathbf{x}),$$

and

$$h_{\sigma,\tau}(\mathbf{x}_1,\mathbf{x}_2) = \exp\left(\frac{\|\mathbf{x}_1\|^2}{2\tau}\right)p_{\text{data}}(\mathbf{x}_1 + \sigma\mathbf{x}_2).$$

We recall the assumptions

Assumption 1 $\text{supp}(p_{\text{data}})$ *is contained in a ball with radius R, and $p_{\text{data}} > c > 0$ on its support.*

Assumption 2 $\|D_i(t,\mathbf{x})\|^2 \le C_1(1 + \|\mathbf{x}\|^2)$, $\forall \mathbf{x} \in \text{supp}(p_{\text{data}})$, $t \in [0,1]$, *where $C_1 \in \mathbb{R}$ is a constant.*

Assumption 3 $\|D_i(t_1,\mathbf{x}_1) - D_i(t_2,\mathbf{x}_2)\| \le C_2(\|\mathbf{x}_1 - \mathbf{x}_2\| + |t_1 - t_2|^{1/2})$, $\forall \mathbf{x}_1, \mathbf{x}_2 \in \text{supp}(p_{\text{data}}), t_1, t_2 \in [0,1]$. *$C_2 \in \mathbb{R}$ is another constant.*

Assumption 4 *$h_{\sigma,\tau}(\mathbf{x}_1,\mathbf{x}_2)$, $\nabla_{\mathbf{x}_1}h_{\sigma,\tau}(\mathbf{x}_1,\mathbf{x}_2)$, p_{data} and ∇p_{data} are L-Lipschitz functions.*

Some calculation shows

$$\begin{aligned}&D_1(t,\mathbf{x}) = \nabla \log \mathbb{E}_{\mathbf{z}\sim\Phi_{\sqrt{\tau}}}[f(\mathbf{x} + \sqrt{1-t}\mathbf{z})]\\ =&\frac{\mathbb{E}_{\mathbf{z}\sim\Phi_{\sqrt{\tau}}}\left[f(\mathbf{x} + \sqrt{1-t}\mathbf{z})\nabla \log f(\mathbf{x} + \sqrt{1-t}\mathbf{z})\right]}{\mathbb{E}_{\mathbf{z}\sim\Phi_{\sqrt{\tau}}}[f(\mathbf{x} + \sqrt{1-t}\mathbf{z})]},\end{aligned} \tag{15}$$

and

$$\nabla \log f(\mathbf{x}) = \nabla \log q_\sigma(\mathbf{x}) + \mathbf{x}/\tau. \tag{16}$$

Let $\hat{D}_1(t,\mathbf{x})$ be an estimated version of $D_1(t,\mathbf{x})$ by replacing $f(\mathbf{x})$ and $\nabla \log f(\mathbf{x}) = \nabla \log q_\sigma(\mathbf{x}) + \mathbf{x}/\tau$ with $\hat{f}(\mathbf{x})$ and $\hat{\mathbf{s}}_\theta(\mathbf{x};\sigma) + \mathbf{x}/\tau$, respectively. By Theorem 4 and 5, we know that

$$\hat{D}_1(t,\mathbf{x}) \to D_1(t,\mathbf{x}) \text{ as } n \to \infty.$$

Similarly, we know that

$$\hat{D}_2(t,\mathbf{x}) = \hat{\mathbf{s}}_\theta(\mathbf{x};\sqrt{1-t}\sigma) \to D_2(t,\mathbf{x}), \text{ as } n \to \infty.$$

Recall that the iteration of state 1 in our Schrödinger Bridge algorithm reads

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \frac{\tau}{N_1}\mathbf{b}(t_k, \mathbf{x}_k) + \sqrt{\frac{\tau}{N_1}}\boldsymbol{\epsilon}_k, \quad \mathbf{x}_0 = \mathbf{0}, \ \ k = 0, ...N_1 - 1, \tag{17}$$

where

$$\mathbf{b}(t_k, \mathbf{x}_k) = \frac{\sum_{i=1}^{N_3} \hat{f}(\tilde{\mathbf{x}}_i)[\hat{\mathbf{s}}_\theta(\tilde{\mathbf{x}}_i, \sigma) + \sqrt{(1-t_k)/\tau}\mathbf{z}_i]}{\sum_{i=N_3+1}^{2N_3} \hat{f}(\tilde{\mathbf{x}}_i)} + \frac{\mathbf{x}_k}{\tau},$$

$\tilde{\mathbf{x}}_i = \mathbf{x}_k + \sqrt{\tau(1-t_k)}\mathbf{z}_i$, $i = 1, ..., 2N_3$, $t_k = \frac{k}{N_1}$, $\{\mathbf{z}_i\}_{i=1}^{2N_3}$, and $\boldsymbol{\epsilon}_k \sim \mathscr{N}(\mathbf{0}, \mathbf{I})$. Note that $\mathbf{b}(t, \mathbf{x})$ is a Monte Carlo version of $\hat{D}_1(t, \mathbf{x})$ and converges to it as the number of samples $N_3 \to \infty$. Then, $\forall (t, \mathbf{x})$

$$\mathbf{b}(t, \mathbf{x}) \to D_1(t, \mathbf{x}) \ \text{ as } \ n, N_3 \to \infty. \tag{18}$$

By Assumption 1 and Assumption 4, we can show that the above consistency results hold uniformly for $(t, \mathbf{x}) \in [0, 1] \times \text{supp}(p_{\text{data}})$. The Euler-Maruyama method for solving for SDE (3) with step size $s = 1/N_1$, $t_k = k/N_1$ reads

$$X_{k+1} = X_k + \frac{\tau}{N_1}D_1(t_k, X_k) + \sqrt{\frac{\tau}{N_1}}\boldsymbol{\epsilon}_k, \quad X_0 = \mathbf{0}, \ \ k = 0, ..., N_1 - 1. \tag{19}$$

Under our Assumptions 2 and 3, SDE (3) admits a strong solution and (22)-(23) in Lemma 2 hold (see A.6). By the classical theory of Euler-Maruyama methods for solving SDEs (Higham, 2001),

$$\mathcal{W}_2(\text{Law}(X_{N_1}), q_\sigma) = \mathcal{O}(1/\sqrt{N_1}) \to 0 \ \text{ as } \ N_1 \to \infty.$$

Using the triangle inequality, we prove

$$\mathcal{W}_2(\text{Law}(\mathbf{x}_{N_1}), q_\sigma) \to 0 \ \text{ as } \ n, N_3, N_1 \to \infty, \tag{20}$$

by showing

$$\mathcal{W}_2(\text{Law}(\mathbf{x}_{N_1}), \text{Law}(X_{N_1})) \to 0 \ \text{ as } \ n, N_3 \to \infty.$$

Recall the definition of $\mathbf{x}_k$ in (17) and X_k in (19). We have

$$\begin{aligned}
&\|\mathbf{x}_k - X_k\|_2^2 \\
\le&\|\mathbf{x}_{k-1} - X_{k-1}\|_2^2 \\
&+ \left(\frac{\tau}{N_1}\|D_1(t_{k-1}, X_{k-1}) - b(t_{k-1}, \mathbf{x}_{k-1})\|_2 \mathrm{d}\ell\right)^2 \\
&+ 2\frac{\tau}{N_1}\|\mathbf{x}_{k-1} - X_{k-1}\|_2 \\
&\cdot \|D_1(t_{k-1}, X_{k-1}) - b(t_{k-1}, \mathbf{x}_{k-1})\|_2 \\
\le&(1 + \tau/N_1)\|X_{k-1} - \mathbf{x}_{k-1}\|_2^2 \\
&+ (\tau/N_1 + \tau^2/N_1^2) \\
&\cdot \|D_1(t_{k-1}, X_{k-1}) - b(t_{k-1}, \mathbf{x}_{k-1})\|_2^2 \\
\le&(1 + \tau/N_1)\|X_{k-1} - \mathbf{x}_{k-1}\|_2^2 \\
&+ 2(\tau/N_1 + \tau^2/N_1^2) \\
&\cdot \|D_1(t_{k-1}, X_{k-1}) - D_1(t_{k-1}, \mathbf{x}_{k-1})\|_2^2 \\
&+ 2(\tau/N_1 + \tau^2/N_1^2) \\
&\cdot \|D_1(t_{k-1}, \mathbf{x}_{k-1}) - b(t_{k-1}, \mathbf{x}_{k-1})\|_2^2 \\
\le&(1 + \tau/N_1)\|X_{k-1} - \mathbf{x}_{k-1}\|_2^2 \\
&+ 2C_2(\tau/N_1 + \tau^2/N_1^2)\|X_{k-1} - \mathbf{x}_{k-1}\|_2^2 \\
&+ 2(\tau/N_1 + \tau^2/N_1^2)o(1) \\
=&(1 + \tau/N_1 + 2C_2(\tau/N_1 + \tau^2/N_1^2))\|X_{k-1} - \mathbf{x}_{k-1}\|_2^2 \\
&+ 2(\tau/N_1 + \tau^2/N_1^2)o(1).
\end{aligned}$$

where the third inequality holds by Assumption 3 and (18). Taking expectation on the above display, we get

$$\begin{aligned}
&\mathbb{E}[\|\mathbf{x}_k - X_k\|_2^2] \\
\le&(1 + \tau/N_1 + 2C_2(\tau/N_1 + \tau^2/N_1^2))\mathbb{E}[\|X_{k-1} - \mathbf{x}_{k-1}\|_2^2] \\
&+ 2(\tau/N_1 + \tau^2/N_1^2)o(1).
\end{aligned}$$

From the above display and the fact that $\mathbf{x}_0 = X_0 = \mathbf{0}$, we can conclude that

$$\begin{aligned}
&\mathbb{E}[\|\mathbf{x}_k - X_k\|_2^2] \\
\le&2(k-1)(\tau/N_1 + \tau^2/N_1^2)o(1) \le 2(\tau + \tau^2/N_1)o(1), \\
&\forall \ 1 \le k \le N_1.
\end{aligned}$$

Thus, we have

$$\mathcal{W}_2(\text{Law}(X_{N_1}), \text{Law}(\mathbf{x}_{N_1})) \to 0, \ \text{ as } \ n, N_3 \to \infty. \tag{21}$$

The consistency results (20) for the first stage in Schrödinger Bridge algorithm has been established. For the second stage, the iteration reads

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \frac{\sigma^2}{N_2}\mathbf{b}(\mathbf{x}_k) + \frac{\sigma}{\sqrt{N_2}}\boldsymbol{\epsilon}_k, \quad k = 0, ..., N_2 - 1, \mathbf{x}_0 = \mathbf{x}_{N_1},$$

where $\mathbf{b}(\mathbf{x}_k) = \hat{\mathbf{s}}_\theta(\mathbf{x}_k, \sqrt{1-\frac{k}{N_2}}\sigma)$ and $\boldsymbol{\epsilon}_k \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. The Euler-Maruyama method for solving for SDE (4) with step size $s = 1/N_2$, $t_k = k/N_2$ reads

$$X_{k+1} = X_k + \frac{\sigma^2}{N_2} D_2(t_k, X_k) + \sqrt{\frac{\sigma}{N_2}}\boldsymbol{\epsilon}_k,$$
$$X_0 \sim q_\sigma, \quad k = 0, ..., N_2 - 1.$$

Then, the consistency results of the second stage can be proved similarly by repeating the part between Equation (19) and Equation (21) and using the consistency results of the first stage, we omit the details here. □

A.6. Additional Lemmas

Lemma 1 *Let f be a uniformly continuous function defined on $E \subseteq [-R, R]^d$. For arbitrary $L \in \mathbb{N}^+$ and $N \in \mathbb{N}^+$, there exists a function ReLU network f_ϕ with width $3^{d+3}\max\left\{d\left\lfloor N^{1/d}\right\rfloor, N+1\right\}$ and depth $12L+14+2d$ such that*

$$\|f - f_\phi\|_{L^\infty(E)} \leq 19\sqrt{d}\omega_f^E\left(2RN^{-2/d}L^{-2/d}\right),$$

where, $\omega_f^E(t)$ is the modulus of continuity of f satisfying $\omega_f^E(t) \to 0$ as $t \to 0^+$.

Proof: This is Theorem 4.3 in (Shen et al., 2019). □

Lemma 2 *Let $\mathbf{x}_t$ be the solution of SDE (3). Under Assumption 2, we have*

$$\mathbb{E}[\|\mathbf{x}_t\|_2^2] \leq C_{1,\tau,d}\exp(\tau^2 t), \quad \forall t \in [0,1], \tag{22}$$

$$\mathbb{E}\left[\|\mathbf{x}_{t_2} - \mathbf{x}_{t_1}\|_2^2\right] \leq C_{2,\tau,d}((t_2-t_1)^2 + (t_2-t_1)), \quad \forall t_1, t_2 \in [0,1]. \tag{23}$$

Proof: By the definition of $\mathbf{x}_t$ in (3), we have $\|\mathbf{x}_t\|_2 \leq \int_0^t \tau\|D_1(\ell, \mathbf{x}_\ell)\|_2 \mathrm{d}\ell + \sqrt{\tau}\|\mathbf{w}_t\|_2$. Then,

$$\begin{aligned}\|\mathbf{x}_t\|_2^2 &\leq 2\tau^2\left(\int_0^t \|D_1(\ell, \mathbf{x}_\ell)\|_2 \mathrm{d}\ell\right)^2 + 2\tau\|\mathbf{w}_t\|_2^2 \\ &\leq 2\tau^2 t\int_0^t \|D_1(\ell, \mathbf{x}_\ell)\|_2^2 \mathrm{d}\ell + 2\tau\|\mathbf{w}_t\|_2^2 \\ &\leq 2\tau^2 t\int_0^t C_1[\|\mathbf{x}_\ell\|_2^2 + 1]\mathrm{d}\ell + 2\tau\|\mathbf{w}_t\|_2^2,\end{aligned}$$

where the first inequality holds due to the inequality $(a+b)^2 \leq 2a^2 + 2b^2$, the last inequality holds by Assumption 2. Thus,

$$\begin{aligned}\mathbb{E}[\|\mathbf{x}_t\|_2^2] &\leq 2\tau^2 t\int_0^t C_1(\mathbb{E}[\|\mathbf{x}_\ell\|_2^2] + 1)\mathrm{d}\ell + 2\tau\mathbb{E}[\|\mathbf{w}_t\|_2^2] \\ &\leq 2\tau^2 C_1\int_0^t \mathbb{E}[\|\mathbf{x}_\ell\|_2^2]\mathrm{d}\ell + (2\tau^2 C_1 + 2\tau d).\end{aligned}$$

Then, (22) follows from the above display and the Bellman-Gronwall inequality.

Again, by the definition of $\mathbf{x}_t$ in (3), we have

$$\|\mathbf{x}_{t_2} - \mathbf{x}_{t_1}\|_2 \leq \int_{t_1}^{t_2} \tau\|D_1(\mathbf{x}_\ell, \ell)\|_2 \mathrm{d}\ell + \sqrt{\tau}\|\mathbf{w}_{t_2} - \mathbf{w}_{t_1}\|_2,$$

Then,

$$\begin{aligned}&\|\mathbf{x}_{t_2} - \mathbf{x}_{t_1}\|_2^2 \\ \leq& 2\tau^2\left(\int_{t_1}^{t_2} \|D_1(\mathbf{x}_\ell, \ell)\|_2 \mathrm{d}\ell\right)^2 + 2\tau\|\mathbf{w}_{t_2} - \mathbf{w}_{t_1}\|_2^2 \\ \leq& 2\tau^2(t_2 - t_1)\int_{t_1}^{t_2} \|D_1(\mathbf{x}_\ell, \ell)\|_2^2 \mathrm{d}\ell + 2\tau\|\mathbf{w}_{t_2} - \mathbf{w}_{t_1}\|_2^2 \\ \leq& 2\tau^2(t_2 - t_1)\int_{t_1}^{t_2} C_1[\|\mathbf{x}_\ell\|_2^2 + 1]\mathrm{d}\ell + 2\tau\|\mathbf{w}_{t_2} - \mathbf{w}_{t_1}\|_2^2,\end{aligned}$$

where the last inequality holds by by Assumption 2. Taking expectations on both sides and using (22), we get (23). □

B. Hyperparameter Settings

For the two-dimensional toy example, we set batch size to be 1000, and use the Adam optimizer (Kingma & Ba, 2015) for both the score estimator and the density ratio estimator. We use learning rate $lr = 0.0001$ and exponential decay rates $betas = (0.5, 0.999)$ for the moment estimates when training the score estimator, and use $lr = 0.001$, $betas = (0.5, 0.999)$ and L2 penalty $weight\_decay = 0.1$ for the density ratio estimator. For the image datasets, the batch size is 128 for both networks. We use $lr = 0.0001$, $betas = (0.9, 0.999)$ and $eps = 10^{-8}$ for the score estimator, and $lr = 10^{-5}$, $betas = (0.5, 0.999)$ and $weight\_decay = 1.0$ for the density ratio estimator.

C. Network Architectures

The score estimator $\hat{\mathbf{s}}_\theta(\cdot, \cdot)$ and the density ratio estimator $\hat{f}(\cdot) = \exp(\hat{r}_\phi(\cdot))$ are parameterized with fully connected networks for the 2D example. The details are listed in Tables 1 and 2.

Table 1. $\hat{\mathbf{s}}_\theta$ for 2D example. $\mathbf{T}$ represents the sinusoidal embeddings (Vaswani et al., 2017) of time t.

LAYER	DETAIL	OUTPUT SIZE
FULLY CONNECTED	LINEAR	256
	ADD LINEAR$_1(\mathbf{T})$	256
	RELU	256
FULLY CONNECTED	LINEAR	512
	ADD LINEAR$_2(\mathbf{T})$	512
	RELU	512
FULLY CONNECTED	LINEAR	2

Table 2. $\hat{r}_\phi$ for 2D example.

LAYER	DETAIL	OUTPUT SIZE
FULLY CONNECTED	LINEAR	256
	RELU	256
FULLY CONNECTED	LINEAR	512
	RELU	512
FULLY CONNECTED	LINEAR	1

For image datasets, we parameterize the density ratio estimator with a residual network. The structure of $\hat{r}_\phi$ is list in Table 3. Our choice of network architecture for $\hat{\mathbf{s}}_\theta$ follows the implementation of the noise predictor $\boldsymbol{\epsilon}_\theta$ in (Song et al., 2021) which is a U-Net (Ronneberger et al., 2015) based on a Wide ResNet (Zagoruyko & Komodakis, 2016).

Table 3. $\hat{r}_\phi$ with $32 \times 32 \times 3$ resolution.

LAYER	DETAIL	OUTPUT SIZE
CONV BLOCK	CONV 5×5	$32 \times 32 \times 128$
	RELU	$32 \times 32 \times 128$
RESIDUAL BLOCK	CONV 5×5	$32 \times 32 \times 128$
	RELU	$32 \times 32 \times 128$
RESIDUAL BLOCK	CONV 3×3	$32 \times 32 \times 128$
	RELU	$32 \times 32 \times 128$
RESIDUAL BLOCK	CONV 3×3	$32 \times 32 \times 128$
	RELU	$32 \times 32 \times 128$
CONV BLOCK	CONV 3×3	$32 \times 32 \times 128$
	RELU	$32 \times 32 \times 128$
FULLY CONNECTED	LINEAR	1

D. More Implementation Details

When training $\hat{f}(\mathbf{x})$, we substract an estimated image mean $\bar{\mathbf{x}}$ from samples in p_{data} to center the data distributions at the origin. The data pre-processing is slightly different when training $\hat{\mathbf{s}}_\theta(\mathbf{x})$, where the samples $\mathbf{x}$ from p_{data} are only rescaled to $[-0.5, 0.5]$. We match the output $\hat{\mathbf{s}}_\theta(\mathbf{x}+\mathbf{z}, \sigma)$ with $\frac{\mathbf{z}}{\bar{\sigma}^2}$ instead of $-\frac{\mathbf{z}}{\bar{\sigma}^2}$ in the denoising score matching objective. To make our algorithm be correctly implemented, we shift the input by adding $\bar{\mathbf{x}} - 0.5$ when using $\hat{\mathbf{s}}_\theta(\mathbf{x})$, and adjust the sign of the output accordingly.

For image generation, there exist very small noises in the generated samples. To eliminate the negative effects induced by noises, we run one additional denoising step after stage 2, by repeating the last step without injecting any noise:

$$\mathbf{x}_{N_2} = \mathbf{x}_{N_2} + \frac{\sigma_0^2}{N_2}\mathbf{b}(\mathbf{x}_{N_2}), \quad \mathbf{b}(\cdot) = \hat{\mathbf{s}}_\theta(\cdot, \sqrt{\frac{1}{N_2}}\sigma_0).$$

We use one Tesla V100 GPU to run the experiments on CIFAR-10, and one RTX 6000 GPU to run the experiments on CelebA.

E. Additional Experiment Results

Here we first list the quantitive results with $\sigma^2 \in \{0.5, 2.0, 5.0\}$, where results with $\sigma^2 = 1.0$ are already presented in the paper. We compare the results with different τ values starting $\tau_{\min} = \sigma^2$. The results are presented in Tables 4, 5 and 6.

Table 4. FID and Inception Score on CIFAR-10 with $\sigma^2 = 0.5$.

τ	0.5	1.0	1.5	2.0
FID	46.59	19.57	**18.73**	20.86
IS	5.92	7.83	**8.13**	8.09
τ	2.5	3.0	3.5	
FID	21.28	21.03	20.40	
IS	8.05	7.98	8.00	

Table 5. FID and Inception Score on CIFAR-10 with $\sigma^2 = 2.0$.

τ	2.0	2.5	3.0	3.5
FID	28.92	22.37	14.52	12.45
IS	7.06	7.50	7.97	**7.98**
τ	4.0	4.5	5.0	
FID	**12.27**	12.58	12.87	
IS	7.91	7.86	7.81	

Table 6. FID and Inception Score on CIFAR-10 with $\sigma^2 = 5.0$.

τ	5.0	5.5	6.0	6.5
FID	17.80	17.52	18.24	16.46
IS	7.67	**7.68**	7.66	**7.68**
τ	7.0	7.5	8.0	
FID	15.71	15.45	**15.41**	
IS	7.64	7.62	7.59	

Next we demonstrate that our algorithm is still effective with less discretization steps. The additional results show that our algorithm can be largely accelerated, maintaining almost the same performance, as shown in Table 7.

Table 7. Quantitative evaluation with different discretization steps. We let the numbers of discretization steps of two stages be the same as $N_1 = N_2 = N$.

N	100	200	500	1000
FID	14.87	13.59	12.85	12.32
IS	7.94	8.02	8.02	8.14

References

Anthony, M. and Bartlett, P. L. *Neural network learning: Theoretical foundations*. cambridge university press, 2009.

Bartlett, P. L. and Mendelson, S. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002.

Bartlett, P. L., Harvey, N., Liaw, C., and Mehrabian, A. Nearly-tight vc-dimension and pseudodimension bounds for piecewise linear neural networks. *Journal of Machine Learning Research*, 20:1–17, 2019.

Dai Pra, P. A stochastic control approach to reciprocal diffusion processes. *Applied mathematics and Optimization*, 23(1):313–329, 1991.

Higham, D. J. An algorithmic introduction to numerical simulation of stochastic differential equations. *SIAM review*, 43(3):525–546, 2001.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

Léonard, C. A survey of the schrodinger problem and some of its connections with optimal transport. *DYNAMICAL SYSTEMS*, 34(4):1533–1574, 2014.

Ronneberger, O., Fischer, P., and Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 234–241, 2015.

Shen, Z., Yang, H., and Zhang, S. Deep network approximation characterized by number of neurons. *arXiv preprint arXiv:1906.05497*, 2019.

Song, J., Meng, C., and Ermon, S. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. In *Advances in Neural Information Processing Systems*, pp. 6000–6010, 2017.

Vershynin, R. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

Zagoruyko, S. and Komodakis, N. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.