Variational Schrödinger Diffusion Models

Wei Deng [*1] **Weijian Luo** [*2] **Yixin Tan** [*3]
Marin Biloš [1] **Yu Chen** [1] **Yuriy Nevmyvaka** [1] **Ricky T. Q. Chen** [4]

Abstract

Schrödinger bridge (SB) has emerged as the go-to method for optimizing transportation plans in diffusion models. However, SB requires estimating the intractable forward score functions, inevitably resulting in the *costly* implicit training loss based on simulated trajectories. To improve the scalability while preserving efficient transportation plans, we leverage variational inference to linearize the forward score functions (variational scores) of SB and restore *simulation-free* properties in training backward scores. We propose the variational Schrödinger diffusion model (VSDM), where the forward process is a multivariate diffusion and the variational scores are adaptively optimized for efficient transport. Theoretically, we use stochastic approximation to prove the convergence of the variational scores and show the convergence of the adaptively generated samples based on the optimal variational scores. Empirically, we test the algorithm in simulated examples and observe that VSDM is efficient in generations of *anisotropic* shapes and yields *straighter* sample trajectories compared to the single-variate diffusion. We also verify the scalability of the algorithm in real-world data and achieve competitive unconditional generation performance in CIFAR10 and conditional generation in time series modeling. Notably, VSDM no longer depends on warm-up initializations and has become tuning-friendly in training large-scale experiments.

1. Introduction

Diffusion models have showcased remarkable proficiency across diverse domains, spanning large-scale generations of image, video, and audio, conditional text-to-image tasks, and adversarial defenses (Dhariwal & Nichol, 2022; Ho et al., 2022; Kong et al., 2021; Ramesh et al., 2022; Zhang et al., 2024). The key to their scalability lies in the closed-form updates of the forward process, highlighting both statistical efficiency (Koehler et al., 2023) and diminished dependence on dimensionality (Vono et al., 2022). Nevertheless, diffusion models lack a distinct guarantee of optimal transport (OT) properties (Lavenant & Santambrogio, 2022) and often necessitate costly evaluations to generate higher-fidelity content (Ho et al., 2020; Salimans & Ho, 2022; Lu et al., 2022; Xue et al., 2023; Luo, 2023).

Alternatively, the Schrödinger bridge (SB) problem (Léonard, 2014; Chen & Georgiou, 2016; Pavon et al., 2021; Caluya & Halder, 2022; De Bortoli et al., 2021), initially rooted in quantum mechanics (Léonard, 2014), proposes optimizing a stochastic control objective through the use of forward-backward stochastic differential equations (FB-SDEs) (Chen et al., 2022b). The alternating solver gives rise to the iterative proportional fitting (IPF) algorithm (Kullback, 1968; Ruschendorf, 1995) in dynamic optimal transport (Villani, 2003; Peyré & Cuturi, 2019). Notably, the intractable forward score function plays a crucial role in providing theoretical guarantees in optimal transport (Chen et al., 2023c; Deng et al., 2024). However, it simultaneously sacrifices the simulation-free property and largely relies on warm-up checkpoints for conducting large-scale experiments (De Bortoli et al., 2021; Chen et al., 2022b). A natural follow-up question arises:

Can we train diffusion models with efficient transport?

To this end, we introduce the variational Schrödinger diffusion model (VSDM). Employing variational inference (Blei et al., 2017), we perform a locally linear approximation of the forward score function, and denote it by the variational score. The resulting linear *forward* stochastic differential equations (SDEs) naturally provide a *closed-form update, significantly enhancing scalability*. Compared to the single-variate score-based generative model (SGM), VSDM is a multivariate diffusion (Singhal et al., 2023). Moreover, hyperparameters are adaptively optimized for *more efficient transportation plans* within the Schrödinger bridge framework (Chen et al., 2022b).

Theoretically, we leverage stochastic approximation (Rob-

[*]Equal contribution [1]Machine Learning Research, Morgan Stanley, NY [2]Peking University [3]Duke University [4]Meta AI (FAIR). Correspondence to: Wei Deng <weideng056@gmail.com>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

bins & Monro, 1951) to demonstrate the convergence of the variational score to the optimal local estimators. Although the global transport optimality is compromised, the notable *simulation-free* speed-ups in training the backward score render the algorithm particularly attractive for training various generation tasks from scratch. Additionally, the efficiency of simulation-based training for the linearized variational score significantly improves owing to computational advancements in convex optimization. We validate the strength of VSDM through simulations, achieving compelling performance on standard image generation tasks. Our contributions unfold in four key aspects:

- We introduce the variational Schrödinger diffusion model (VSDM), a multivariate diffusion with *optimal variational scores* guided by optimal transport. Additionally, the training of backward scores is *simulation-free* and becomes much more scalable.
- We study the convergence of the variational score using stochastic approximation (SA) theory, which can be further generalized to a class of state space diffusion models for future developments.
- VSDM is effective in generating data of *anisotropic* shapes and motivates *straighter* transportation paths via the optimized transport.
- VSDM achieves competitive unconditional generation on CIFAR10 and conditional generation in time series modeling without reliance on warm-up initializations.

2. Related Works

Flow Matching and Beyond Lipman et al. (2023) utilized the McCann displacement interpolation (McCann, 1997) to train simulation-free CNFs to encourage straight trajectories. Consequently, Pooladian et al. (2023); Tong et al. (2023) proposed straightening by using minibatch optimal transport solutions. Similar ideas were achieved by Liu (2022); Liu et al. (2023) to iteratively rectify the interpolation path. Albergo & Vanden-Eijnden (2023); Albergo et al. (2023) developed the stochastic interpolant approach to unify both flow and diffusion models. However, *"straighter" transport* maps may not imply *optimal* transportation plans in general and the couplings are still not effectively optimized.

Dynamic Optimal Transport Finlay et al. (2020); Onken et al. (2021) introduced additional regularization through optimal transport to enforce straighter trajectories in CNFs and reduce the computational cost. De Bortoli et al. (2021); Chen et al. (2022b); Vargas et al. (2021) studied the dynamic Schrödinger bridge with guarantees in entropic optimal transport (EOT) (Chen et al., 2023c); Shi et al. (2023); Peluchetti (2023); Chen et al. (2023b) generalized bridge matching and flow matching based EOT and obtained smoother trajectories, however, *scalability* remains a significant concern for Schrödinger-based diffusions.

3. Preliminaries

3.1. Diffusion Models

The score-based generative models (SGMs) (Ho et al., 2020; Song et al., 2021b) first employ a forward process (1a) to map data to an approximate Gaussian and subsequently reverse the process in Eq.(1b) to recover the data distribution.

$$\mathrm{d}\overrightarrow{\mathbf{x}}_t = \boldsymbol{f}_t(\overrightarrow{\mathbf{x}}_t)\mathrm{d}t + \sqrt{\beta_t}\mathrm{d}\overrightarrow{\mathbf{w}}_t \tag{1a}$$

$$\mathrm{d}\overleftarrow{\mathbf{x}}_t = \left[\boldsymbol{f}_t(\overleftarrow{\mathbf{x}}_t) - \beta_t \nabla \log \rho_t\left(\overleftarrow{\mathbf{x}}_t\right)\right]\mathrm{d}t + \sqrt{\beta_t}\mathrm{d}\overleftarrow{\mathbf{w}}_t, \tag{1b}$$

where $\overleftarrow{\mathbf{x}}_t, \overrightarrow{\mathbf{x}}_t \in \mathbb{R}^d$; $\overrightarrow{\mathbf{x}}_0 \sim \rho_{\text{data}}$ and $\overleftarrow{\mathbf{x}}_T \sim \rho_{\text{prior}}$; $\boldsymbol{f}_t$ denotes the vector field and is often set to $\mathbf{0}$ (a.k.a. VE-SDE) or linear in $\mathbf{x}$ (a.k.a. VP-SDE); $\beta_t > 0$ is the time-varying scalar; $\overrightarrow{\mathbf{w}}_t$ is a forward Brownian motion from $t \in [0, T]$ with $\rho_T \approx \rho_{\text{prior}}$; $\overleftarrow{\mathbf{w}}_t$ is a backward Brownian motion from time T to 0. The marginal density ρ_t of the forward process (1a) is essential for generating the data but remains inaccessible in practice due to intractable normalizing constants.

Explicit Score Matching (ESM) Instead, the conditional score function $\nabla \log \rho_{t|0}(\cdot) \equiv \nabla \log \rho_t\left(\cdot|\overrightarrow{\mathbf{x}}_0\right)$ is estimated by minimizing a user-friendly ESM loss (weighted by λ) between the score estimator $s_t \equiv s_\theta(\cdot, t)$ and exact score (Song et al., 2021b) such that

$$\mathbb{E}_t\left[\lambda_t \mathbb{E}_{\overrightarrow{\mathbf{x}}_0}\mathbb{E}_{\overrightarrow{\mathbf{x}}_t|\overrightarrow{\mathbf{x}}_0}[\|s_t(\overrightarrow{\mathbf{x}}_t) - \nabla \log \rho_{t|0}\left(\overrightarrow{\mathbf{x}}_t\right)\|_2^2]\right]. \tag{2}$$

Notably, both VP- and VE-SDEs yield closed-form expressions for any $\overrightarrow{\mathbf{x}}_t$ given $\overrightarrow{\mathbf{x}}_0$ in the forward process (Song et al., 2021b), which is instrumental for the scalability of diffusion models in real-world large-scale generation tasks.

Implicit Score Matching (ISM) By integration by parts, ESM is equivalent to the ISM loss (Hyvärinen, 2005; Huang et al., 2021; Luo et al., 2024b) and the evidence lower bound (ELBO) follows

$$\log \rho_0(\mathbf{x}_0) \geq \mathbb{E}_{\rho_{T|0}(\cdot)}\left[\log \rho_{T|0}(\mathbf{x}_T)\right] - \frac{1}{2}\int_0^T \mathbb{E}_{\rho_{t|0}(\cdot)}\left[\beta_t \|\mathbf{s}_t\|_2^2 + 2\nabla \cdot (\beta_t \mathbf{s}_t - \boldsymbol{f}_t)\right]\mathrm{d}t.$$

ISM is naturally connected to Song et al. (2020), which supports flexible marginals and nonlinear forward processes but becomes significantly less scalable compared to ESM.

3.2. Schrödinger Bridge

The dynamic Schrödinger bridge aims to solve a full bridge

$$\inf_{\mathbb{P} \in \mathcal{D}(\rho_{\text{data}}, \rho_{\text{prior}})} \mathrm{KL}(\mathbb{P}|\mathbb{Q}), \tag{3}$$

where $\mathcal{D}(\rho_{\text{data}}, \rho_{\text{prior}})$ is the family of path measures with marginals ρ_{data} and ρ_{prior} at $t=0$ and $t=T$, respectively; $\mathbb{Q}$ is the prior process driven by $\mathrm{d}\mathbf{x}_t = \boldsymbol{f}_t(\mathbf{x}_t)\mathrm{d}t + \sqrt{2\beta_t\varepsilon}\mathrm{d}\overrightarrow{\mathbf{w}}_t$. It also yields a stochastic control formulation (Chen et al., 2021; Pavon et al., 2021; Caluya & Halder, 2022).

$$\begin{aligned} &\inf_{\boldsymbol{u}\in\mathcal{U}} \mathbb{E}\left\{\int_0^T \frac{1}{2}\|\boldsymbol{u}_t(\overrightarrow{\mathbf{x}}_t)\|_2^2 \mathrm{d}t\right\} \\ &\text{s.t. } \mathrm{d}\overrightarrow{\mathbf{x}}_t = \left[\boldsymbol{f}_t(\overrightarrow{\mathbf{x}}) + \sqrt{\beta_t}\boldsymbol{u}_t(\overrightarrow{\mathbf{x}})\right]\mathrm{d}t + \sqrt{2\beta_t\varepsilon}\mathrm{d}\overrightarrow{\mathbf{w}}_t \quad (4) \\ &\overrightarrow{\mathbf{x}}_0 \sim \rho_{\text{data}}, \ \ \overrightarrow{\mathbf{x}}_T \sim \rho_{\text{prior}}, \end{aligned}$$

where $\mathcal{U}$ is the family of controls. The expectation is taken w.r.t $\overrightarrow{\rho}_t(\cdot)$, which denotes the PDF of the controlled diffusion (4); ε is the temperature of the diffusion and the regularizer in EOT (Chen et al., 2023c).

Solving the underlying Hamilton–Jacobi–Bellman (HJB) equation and invoking the time reversal (Anderson, 1982) with $\varepsilon = \frac{1}{2}$, *Schrödinger system* yields the desired forward-backward stochastic differential equations (FB-SDEs) (Chen et al., 2022b):

$$\mathrm{d}\overrightarrow{\mathbf{x}}_t = \left[\boldsymbol{f}_t(\overrightarrow{\mathbf{x}}_t) + \beta_t\nabla\log\overrightarrow{\psi}_t(\overrightarrow{\mathbf{x}}_t)\right]\mathrm{d}t + \sqrt{\beta_t}\mathrm{d}\overrightarrow{\mathbf{w}}_t, \quad (5a)$$

$$\mathrm{d}\overleftarrow{\mathbf{x}}_t = \left[\boldsymbol{f}_t(\overleftarrow{\mathbf{x}}_t) - \beta_t\nabla\log\overleftarrow{\varphi}_t(\overleftarrow{\mathbf{x}}_t)\right]\mathrm{d}t + \sqrt{\beta_t}\mathrm{d}\overleftarrow{\mathbf{w}}_t, \quad (5b)$$

$$\text{where } \overrightarrow{\psi}_t(\cdot)\overleftarrow{\varphi}_t(\cdot) = \overrightarrow{\rho}_t(\cdot), \rho_0(\cdot)\sim\rho_{\text{data}}, \ \rho_T(\cdot)\sim\rho_{\text{prior}}.$$

To solve the optimal controls (scores) $(\nabla\log\overrightarrow{\psi}, \nabla\log\overleftarrow{\varphi})$, a standard tool is to leverage the nonlinear Feynman-Kac formula (Ma & Yong, 2007; Karatzas & Shreve, 1998; Chen et al., 2022b) to learn a stochastic representation.

Proposition 1 (Nonlinear Feynman-Kac representation)**.** *Assume Lipschitz smoothness and linear growth condition on the drift $\boldsymbol{f}$ and diffusion g in the FB-SDE (5). Define $\overrightarrow{y}_t = \log\overrightarrow{\psi}_t(\mathbf{x}_t)$ and $\overleftarrow{y}_t = \log\overleftarrow{\varphi}_t(\mathbf{x}_t)$. Then the stochastic representation follows*

$$\begin{aligned} &\overleftarrow{y}_s = \mathbb{E}\left[\overleftarrow{y}_T - \int_s^T \Gamma_\zeta(\overleftarrow{\mathbf{z}}_t; \overrightarrow{\mathbf{z}}_t)\mathrm{d}t \middle| \overrightarrow{\mathbf{x}}_s = \boldsymbol{x}_s\right], \\ &\Gamma_\zeta(\overleftarrow{\mathbf{z}}_t; \overrightarrow{\mathbf{z}}_t) \equiv \frac{1}{2}\|\overleftarrow{\mathbf{z}}_t\|_2^2 + \nabla\cdot\left(\sqrt{\beta_t}\overleftarrow{\mathbf{z}}_t - \boldsymbol{f}_t\right) + \zeta\langle\overleftarrow{\mathbf{z}}_t, \overrightarrow{\mathbf{z}}_t\rangle, \end{aligned} \quad (6)$$

where $\overrightarrow{\mathbf{z}}_t = \sqrt{\beta_t}\nabla\overrightarrow{y}_t$, $\overleftarrow{\mathbf{z}}_t = \sqrt{\beta_t}\nabla\overleftarrow{y}_t$, and $\zeta = 1$.

4. Variational Schrödinger Diffusion Models

SB outperforms SGMs in the theoretical potential of optimal transport and an intractable score function $\nabla\log\overrightarrow{\psi}_t(\mathbf{x}_t)$ is exploited in the forward SDE for more efficient transportation plans. However, there is no free lunch in achieving such efficiency, and it comes with three notable downsides:

- Solving $\nabla\log\overrightarrow{\psi}_t$ in Eq.(5a) for optimal transport is prohibitively costly and may not be necessary (Marzouk et al., 2016; Liu et al., 2023).
- The nonlinear diffusion no longer yields closed-form expression of $\overrightarrow{\mathbf{x}}_t$ given $\overrightarrow{\mathbf{x}}_0$ (Chen et al., 2022b).
- The ISM loss is inevitable and the estimator suffers from a large variance issue (Hutchinson, 1989).

4.1. Variational Inference via Linear Approximation

FB-SDEs naturally connect to the alternating-projection solver based on the IPF (a.k.a. Sinkhorn) algorithm, boiling down the full bridge (3) to a half-bridge solver (Pavon et al., 2021; De Bortoli et al., 2021; Vargas et al., 2021). With $\mathbb{P}_1$ given and $k = 1, 2, ...$, we have:

$$\mathbb{P}_{2k} := \underset{\mathbb{P}\in\mathcal{D}(\rho_{\text{data}},\,\cdot)}{\arg\min} \ \mathrm{KL}(\mathbb{P}\|\mathbb{P}_{2k-1}), \quad (7a)$$

$$\mathbb{P}_{2k+1} := \underset{\mathbb{P}\in\mathcal{D}(\cdot,\,\rho_{\text{prior}})}{\arg\min} \ \mathrm{KL}(\mathbb{P}\|\mathbb{P}_{2k}). \quad (7b)$$

More specifically, Chen et al. (2022b) proposed a neural network parameterization to model $(\overleftarrow{\mathbf{z}}_t, \overrightarrow{\mathbf{z}}_t)$ using $(\overleftarrow{\mathbf{z}}_t^\theta, \overrightarrow{\mathbf{z}}_t^\omega)$, where θ and ω refer to the model parameters, respectively. Each stage of the half-bridge solver proposes to solve the models alternatingly as follows

$$\overleftarrow{\mathcal{L}}(\theta) = -\int_0^T \mathbb{E}_{\overrightarrow{\mathbf{x}}_t\backsim(5a)}\left[\Gamma_1(\overleftarrow{\mathbf{z}}_t^\theta; \overrightarrow{\mathbf{z}}_t^\omega)\mathrm{d}t \middle| \overrightarrow{\mathbf{x}}_0 = \mathbf{x}_0\right] \quad (8a)$$

$$\overrightarrow{\mathcal{L}}(\omega) = -\int_0^T \mathbb{E}_{\overleftarrow{\mathbf{x}}_t\backsim(5b)}\left[\Gamma_1(\overrightarrow{\mathbf{z}}_t^\omega; \overleftarrow{\mathbf{z}}_t^\theta)\mathrm{d}t \middle| \overleftarrow{\mathbf{x}}_T = \mathbf{x}_T\right], \quad (8b)$$

where Γ_1 is defined in Eq.(6) and $\backsim$ denotes the approximate simulation parametrized by neural networks [*]

However, solving the backward score in Eq.(8a) through simulations, akin to the ISM loss, is computationally demanding and affects the scalability in generative models.

To motivate simulation-free property, we leverage variational inference (Blei et al., 2017) and study a linear approximation of the forward score $\nabla\log\overrightarrow{\psi}(\mathbf{x}, t) \approx \mathbf{A}_t\mathbf{x}$ with $\boldsymbol{f}_t(\overrightarrow{\mathbf{x}}_t) \equiv -\frac{1}{2}\beta_t\overrightarrow{\mathbf{x}}_t$, which ends up with the variational FB-SDE (VFB-SDE):

$$\mathrm{d}\overrightarrow{\mathbf{x}}_t = \left[-\frac{1}{2}\beta_t\overrightarrow{\mathbf{x}}_t + \beta_t\mathbf{A}_t\overrightarrow{\mathbf{x}}_t\right]\mathrm{d}t + \sqrt{\beta_t}\mathrm{d}\overrightarrow{\mathbf{w}}_t, \quad (9a)$$

$$\mathrm{d}\overleftarrow{\mathbf{x}}_t = \left[-\frac{1}{2}\beta_t\overleftarrow{\mathbf{x}}_t - \beta_t\nabla\log\overrightarrow{\rho}_t(\overleftarrow{\mathbf{x}}_t)\right]\mathrm{d}t + \sqrt{\beta_t}\mathrm{d}\overleftarrow{\mathbf{w}}_t, \quad (9b)$$

where $t\in[0, T]$ and $\nabla\log\overrightarrow{\rho}_t$ is the score function of (9a) and the conditional version is to be derived in Eq.(15).

The half-bridge solver is restricted to a class of OU processes $\mathrm{OU}(\rho_{\text{data}}, \cdot)$ with the initial marginal ρ_{data}.

$$\underset{\mathbb{P}\in\mathcal{D}(\rho_{\text{data}},\cdot)}{\arg\min} \ \mathrm{KL}(\mathbb{P}\|\mathbb{P}_{2k-1}) \Rightarrow \underset{\widehat{\mathbb{P}}\in\mathrm{OU}(\rho_{\text{data}},\cdot)}{\arg\min} \ \mathrm{KL}(\widehat{\mathbb{P}}\|\mathbb{P}_{2k-1}).$$

[*] $\sim$ (resp. $\backsim$) denotes the exact (resp. parametrized) simulation.

By the *mode-seeking* property of the exclusive (reverse) KL divergence (Chan et al., 2022), we can expect the optimizer $\widehat{\mathbb{P}}$ to be a *local estimator* of the nonlinear solution in (7a).

Additionally, the loss function (8b) to learn the variational score $\mathbf{A}_t$, where $t \in [0, T]$, can be simplified to

$$\overrightarrow{\mathcal{L}}(\mathbf{A}) = -\int_0^T \mathbb{E}_{\mathbf{x}_t \backsim (9b)} \left[\Gamma_\zeta(\mathbf{A}_t \mathbf{x}_t; \overleftarrow{\mathbf{z}}_t^\theta) \mathrm{d}t \middle| \overleftarrow{\mathbf{x}}_T = \mathbf{x}_T \right], \tag{10}$$

where Γ_ζ is defined in Eq.(6). Since the structure property $\overrightarrow{\psi}_t \overleftarrow{\varphi}_t = \overrightarrow{\rho}_t$ in Eq.(5) is compromised by the variational inference, we propose to tune ζ in our experiments.

4.2. Closed-form Expression of Backward Score

Assume a prior knowledge of $\mathbf{A}_t$ is given, we can rewrite the forward process (9a) in the VFB-SDE and derive a multivariate forward diffusion (Singhal et al., 2023):

$$\begin{aligned} \mathrm{d}\overrightarrow{\mathbf{x}}_t &= \left[-\frac{1}{2}\beta_t \mathbf{I} + \beta_t \mathbf{A}_t \right] \overrightarrow{\mathbf{x}}_t \mathrm{d}t + \sqrt{\beta_t} \mathrm{d}\overrightarrow{\mathbf{w}}_t \\ &= -\frac{1}{2}\mathbf{D}_t \beta_t \overrightarrow{\mathbf{x}}_t \mathrm{d}t + \sqrt{\beta_t} \mathrm{d}\overrightarrow{\mathbf{w}}_t, \end{aligned} \tag{11}$$

where $\mathbf{D}_t = \mathbf{I} - 2\mathbf{A}_t \in \mathbb{R}^{d \times d}$ is a positive-definite matrix [†]. Consider the multivariate OU process (11). The mean and covariance follow

$$\frac{\mathrm{d}\boldsymbol{\mu}_{t|0}}{\mathrm{d}t} = -\frac{1}{2}\beta_t \mathbf{D}_t \boldsymbol{\mu}_{t|0} \tag{12a}$$

$$\frac{\mathrm{d}\boldsymbol{\Sigma}_{t|0}}{\mathrm{d}t} = -\frac{1}{2}\beta_t \big(\mathbf{D}_t \boldsymbol{\Sigma}_{t|0} + \boldsymbol{\Sigma}_{t|0} \mathbf{D}_t^\intercal \big) + \beta_t \mathbf{I}. \tag{12b}$$

Solving the differential equations with the help of integration factors, the mean process follows

$$\boldsymbol{\mu}_{t|0} = e^{-\frac{1}{2}[\beta \mathbf{D}]_t} \mathbf{x}_0, \tag{13}$$

where $[\beta \mathbf{D}]_t = \int_0^t \beta_s \mathbf{D}_s \mathrm{d}s$. By matrix decomposition $\boldsymbol{\Sigma}_{t|0} = \mathbf{C}_t \mathbf{H}_t^{-1}$ (Särkkä & Solin, 2019), the covariance process follows that:

$$\begin{pmatrix} \mathbf{C}_t \\ \mathbf{H}_t \end{pmatrix} = \exp \left[\begin{pmatrix} -\frac{1}{2}[\beta \mathbf{D}]_t & [\beta \mathbf{I}]_t \\ \mathbf{0} & \frac{1}{2}[\beta \mathbf{D}^\intercal]_t \end{pmatrix} \right] \begin{pmatrix} \boldsymbol{\Sigma}_0 \\ \mathbf{I} \end{pmatrix}, \tag{14}$$

where the above matrix exponential can be easily computed through modern computing libraries. Further, to avoid computing the expensive matrix exponential for high-dimensional problems, we can adopt a diagonal and time-invariant $\mathbf{D}_t$.

Suppose $\boldsymbol{\Sigma}_{t|0}$ has the Cholesky decomposition $\boldsymbol{\Sigma}_{t|0} = \mathbf{L}_t \mathbf{L}_t^\intercal$ for some lower-triangular matrix $\mathbf{L}_t$. We can have a closed-form update that resembles the SGM.

$$\overrightarrow{\mathbf{x}}_t = \boldsymbol{\mu}_{t|0} + \mathbf{L}_t \boldsymbol{\epsilon},$$

where $\boldsymbol{\mu}_{t|0}$ is defined in Eq.(13) and $\boldsymbol{\epsilon}$ is the standard d-dimensional Gaussian vector. The score function follows

$$\begin{aligned} \nabla \log \overrightarrow{\rho}_{t|0}(\overrightarrow{\mathbf{x}}_t) &= -\frac{1}{2}\nabla[(\overrightarrow{\mathbf{x}}_t - \boldsymbol{\mu}_t)^\intercal \boldsymbol{\Sigma}_{t|0}^{-1}(\overrightarrow{\mathbf{x}}_t - \boldsymbol{\mu}_t)] \\ &= -\boldsymbol{\Sigma}_{t|0}^{-1}(\overrightarrow{\mathbf{x}}_t - \boldsymbol{\mu}_t) \\ &= -\mathbf{L}_t^{-\intercal} \mathbf{L}_t^{-1} \mathbf{L}_t \boldsymbol{\epsilon} := -\mathbf{L}_t^{-\intercal} \boldsymbol{\epsilon}. \end{aligned} \tag{15}$$

Invoking the ESM loss function in Eq.(2), we can learn the score function $\nabla \log \overrightarrow{\rho}_{t|0}(\overrightarrow{\mathbf{x}}_t | \overrightarrow{\mathbf{x}}_0)$ using a neural network parametrization $s_t(\cdot)$ and optimize the loss function:

$$\nabla_{\mathbf{A}} \| \mathbf{L}_t^{-\intercal} \boldsymbol{\epsilon} - s_t(\mathbf{x}_t) \|_2^2. \tag{16}$$

One may further consider preconditioning techniques (Karras et al., 2022) or variance reduction (Singhal et al., 2023) to stabilize training and accelerate training speed.

Speed-ups via Time-invariant and Diagonal $\mathbf{D}_t$ If we parametrize $\mathbf{D}_t$ as a time-invariant and diagonal positive-definite matrix, the formula (14) has simpler explicit expressions that do not require calling matrix exponential operators. We present such a result in Corollary 1. For the image generation experiment in Section 7.3, we use such a diagonal parametrization when implementing the VSDM.

Corollary 1. *If $\mathbf{D}_t = \boldsymbol{\Lambda} := \mathrm{diag}(\boldsymbol{\lambda})$, where $\lambda_i \geq 0$, $\forall 1 \leq i \leq d$. If we denote the $\sigma_t^2 := \int_0^t \beta_s \mathrm{d}s$, then matrices $\mathbf{C}_t$ and $\mathbf{H}_t$ has simpler expressions with*

$$\begin{aligned} \mathbf{C}_t &= \boldsymbol{\Lambda}^{-1} \big\{ \exp(\frac{1}{2}\sigma_t^2 \boldsymbol{\Lambda}) - \exp(-\frac{1}{2}\sigma_t^2 \boldsymbol{\Lambda}) \big\} \\ \mathbf{H}_t &= \exp(\frac{1}{2}\sigma_t^2 \boldsymbol{\Lambda}), \end{aligned}$$

which leads to $\mathbf{C}_t \mathbf{H}_t^{-1} = \boldsymbol{\Lambda}^{-1} \big\{ \mathbf{I} - \exp(-\sigma_t^2 \boldsymbol{\Lambda}) \big\}$. As a result, the corresponding forward transition writes

$$\boldsymbol{\mu}_{t|0} = \exp(-\frac{1}{2}\sigma_t^2 \boldsymbol{\Lambda}) \mathbf{x}_0, \ \mathbf{L}_t = \boldsymbol{\Lambda}^{-\frac{1}{2}} \sqrt{\mathbf{I} - \exp(-\sigma_t^2 \boldsymbol{\Lambda})}.$$

In Corrolary 1 detailed in Appendix A, since the matrix $\boldsymbol{\Lambda} = \mathrm{diag}(\boldsymbol{\lambda})$ is diagonal and time-invariant, the matrix exponential and square root can be directly calculated element-wise on each diagonal elements λ_i independently.

4.2.1. Backward SDE

Taking the time reversal (Anderson, 1982) of the forward multivariate OU process (11), the backward SDE satisfies

$$\mathrm{d}\overleftarrow{\mathbf{x}}_t = (-\frac{1}{2}\mathbf{D}_t \beta_t \overleftarrow{\mathbf{x}}_t - \beta_t s_t(\overleftarrow{\mathbf{x}}_t)) \mathrm{d}t + \sqrt{\beta_t} \mathrm{d}\overleftarrow{\mathbf{w}}_t. \tag{17}$$

Notably, with a general PD matrix $\mathbf{D}_t$, the prior distribution follows that $\mathbf{x}_T \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma}_{T|0})$ [‡]. We also note that the prior is now limited to Gaussian distributions, which is not a general bridge anymore.

[†] $\mathbf{D}_t = -2\mathbf{A}_t \in \mathbb{R}^{d \times d}$ when the forward SDE is VE-SDE.

[‡] See the Remark on the selection of ρ_{prior} in section B.1.

4.2.2. Probability Flow ODE

We can follow Song et al. (2021b) and obtain the deterministic process directly:

$$\mathrm{d}\overleftarrow{\mathbf{x}}_t = \left(-\frac{1}{2}\mathbf{D}_t\beta_t\overleftarrow{\mathbf{x}}_t - \frac{1}{2}\beta_t s_t(\overleftarrow{\mathbf{x}}_t)\right)\mathrm{d}t, \qquad (18)$$

where $\mathbf{x}_T \sim \mathrm{N}(\mathbf{0}, \mathbf{\Sigma}_{T|0})$ and the sample trajectories follow the same marginal densities $\overrightarrow{\rho}_t(\mathbf{x}_t)$ as in the SDE.

4.3. Adaptive Diffusion via Stochastic Approximation

Our major goal is to generate high-fidelity data with efficient transportation plans based on the optimal $\mathbf{A}_t^\star$ in the forward process (11). However, the optimal $\mathbf{A}_t^\star$ is not known *a priori*. To tackle this issue, we leverage stochastic approximation (SA) (Robbins & Monro, 1951; Benveniste et al., 1990) to adaptively optimize the variational score $\mathbf{A}_t^{(k)}$ through optimal transport and simulate the backward trajectories.

(1) Simulate backward trajectoriest $\{\overleftarrow{\mathbf{x}}_{nh}^{(k+1)}\}_{n=0}^{N-1}$ via the Euler–Maruyama (EM) scheme of the backward process (17) with a learning rate h.

(2) Optimize variational scores $\{\mathbf{A}_{nh}^{(k)}\}_{n=0}^{N-1}$:

$$\mathbf{A}_{nh}^{(k+1)} = \mathbf{A}_{nh}^{(k)} - \eta_{k+1}\nabla\overrightarrow{\mathcal{L}}_{nh}(\mathbf{A}_{nh}^{(k)}; \overleftarrow{\mathbf{x}}_{nh}^{(k+1)}),$$

where $\nabla\overrightarrow{\mathcal{L}}_{nh}(\mathbf{A}_{nh}^{(k)}; \overleftarrow{\mathbf{x}}_{nh}^{(k+1)})$ is the loss function (10) at time nh and is known as the random field. We expect that the simulation of backward trajectories $\{\overleftarrow{\mathbf{x}}_{nh}^{(k+1)}\}_{n=0}^{N-1}$ given $s_{nh}^{(k+1)}$ helps the optimization of $\mathbf{A}_{nh}^{(k+1)}$ and the optimized $\mathbf{A}_{nh}^{(k+1)}$ in turn contributes to a more efficient transportation plan for estimating $s_{nh}^{(k+2)}$ and simulating the backward trajectories $\{\overleftarrow{\mathbf{x}}_{nh}^{(k+2)}\}_{n=0}^{N-1}$.

Trajectory Averaging The stochastic approximation algorithm is a standard framework to study adaptive sampling algorithms (Liang et al., 2007). Moreover, the formulation suggests to stabilize the trajectories (Polyak & Juditsky, 1992) with averaged parameters $\overline{\mathbf{A}}_{nh}^{(k)}$ as follows

$$\overline{\mathbf{A}}_{nh}^{(k)} = \sum_{i=1}^{k}\mathbf{A}_{nh}^{(i)} = \left(1 - \frac{1}{k}\right)\overline{\mathbf{A}}_{nh}^{(k-1)} + \frac{1}{k}\mathbf{A}_{nh}^{(k)},$$

where $\overline{\mathbf{A}}_{nh}^{(k)}$ is known to be an asymptotically efficient (optimal) estimator (Polyak & Juditsky, 1992) in the local state space $\mathcal{A}$ by assumption A1.

Exponential Moving Average (EMA) Despite guarantees in convex scenarios, the parameter space differs tremendously in different surfaces in non-convex state space $\mathcal{A}$. Empirically, if we want to exploit information from multiple modes, a standard extension is to employ the EMA technique (Trivedi & Kondor, 2017):

$$\overline{\mathbf{A}}_{nh}^{(k)} = (1-\eta)\overline{\mathbf{A}}_{nh}^{(k-1)} + \eta\mathbf{A}_{nh}^{(k)}, \text{where } \eta \in (0,1).$$

The EMA techniques are widely used empirically in diffusion models and Schrödinger bridge (Song & Ermon, 2020; De Bortoli et al., 2021; Chen et al., 2022b) to avoid oscillating trajectories. Now we are ready to present our methodology in Algorithm 1.

Computational Cost Regarding the wall-clock computational time: i) training (linear) variational scores, albeit in a *simulation-based manner*, becomes significantly faster than estimating nonlinear forward scores in Schrödinger bridge; ii) the variational parametrization greatly reduced the number of model parameters, which yields a much-reduced variance in the Hutchinson's estimator (Hutchinson, 1989); iii) since we don't need to update $\mathbf{A}_t$ as often as the backward score model, we can further amortize the training of $\mathbf{A}_t$. In the simulation example in Figure.9(b), VSDM is only 10% slower than the SGM with the same training complexity of backward scores while still maintaining efficient convergence of variational scores.

5. Convergence of Stochastic Approximation

In this section, we study the convergence of $\mathbf{A}_t^{(k)}$ to the optimal $\mathbf{A}_t^\star$, where $t \in [0, T]$ [§]. The primary objective is to show the iterates (19) follow the trajectories of the dynamical system asymptotically:

$$\mathrm{d}\mathbf{A}_t = \nabla\overrightarrow{\mathbf{L}}_t(\mathbf{A}_t)\mathrm{d}s, \qquad (20)$$

where $\frac{\mathrm{d}\mathbf{A}_t}{\mathrm{d}s} = \lim_{\eta\to 0}\frac{\mathbf{A}_t^{(k+1)} - \mathbf{A}_t^{(k)}}{\eta}$ and $\nabla\overrightarrow{\mathbf{L}}_t(\cdot)$ is the mean field at time t:

$$\nabla\overrightarrow{\mathbf{L}}_t(\mathbf{A}_t) = \int_{\mathcal{X}}\nabla\overrightarrow{\mathcal{L}}_t(\mathbf{A}_t; \overleftarrow{\mathbf{x}}_t^{(\cdot)})\overleftarrow{\rho}_t(\mathrm{d}\overleftarrow{\mathbf{x}}_t^{(\cdot)}), \qquad (21)$$

where $\mathcal{X}$ denotes the state space of data $\mathbf{x}$ and $\nabla\overrightarrow{\mathcal{L}}_t$ denotes the gradient w.r.t. $\mathbf{A}_t$; $\overleftarrow{\rho}_t$ is the distribution of the continuous-time interpolation of the discretized backward SDE (22) from $t = T$ to 0. We denote by $\mathbf{A}_t^\star$ one of the solutions of $\nabla\overrightarrow{\mathbf{L}}_t(\mathbf{A}_t^\star) = \mathbf{0}$.

The aim is to find the optimal solution $\mathbf{A}_t^\star$ to the mean field $\nabla\overrightarrow{\mathbf{L}}_t(\mathbf{A}_t^\star) = \mathbf{0}$. However, we acknowledge that the equilibrium is not unique in general nonlinear dynamical systems. To tackle this issue, we focus our analysis around a neighborhood $\mathbf{\Theta}$ of the equilibrium by assumption A1. After running sufficient many iterations with a small enough

[§] We slightly abuse the notation and generalize $\mathbf{A}_{nh}^{(k)}$ to $\mathbf{A}_t^{(k)}$.

Algorithm 1 Variational Schrödinger Diffusion Models (VSDM). ρ_{prior} is fixed to a Gaussian distribution. η_k is the step size for SA and h is the learning rate for the backward sampling of Eq.(17). $\boldsymbol{\xi}_n$ denotes the standard Gaussian vector at the sampling iteration n. The exponential moving averaging (EMA) technique can be used to further stabilize the algorithm.

repeat

Simulation-free Optimization of Backward Score

Draw $\mathbf{x}_0 \sim \rho_{\text{data}}$, $n \sim \{0, 1, \cdots, N-1\}$, $\boldsymbol{\epsilon} \sim \mathrm{N}(\mathbf{0}, \mathbf{I})$.

Sample $\mathbf{x}_{nh}|\mathbf{x}_0 \sim \mathrm{N}(\boldsymbol{\mu}_{nh|0}, \boldsymbol{\Sigma}_{nh|0})$ by Eq.(13) and (14) given $\mathbf{A}_{nh}^{(k)}$.

Cache $\{\boldsymbol{\mu}_{nh|0}\}_{n=0}^{N-1}$ and $\{\mathbf{L}_{nh}^{-\intercal}\}_{n=0}^{N-1}$ via Cholesky decomposition of $\{\boldsymbol{\Sigma}_{nh}\}_{n=0}^{N-1}$ to avoid repeated computations.

Optimize the score functions $s_{nh}^{(k+1)}$ sufficiently through the loss function $\nabla_\theta \|\mathbf{L}_{nh}^{-\intercal}\boldsymbol{\epsilon} - s_{nh}^{(k+1)}(\mathbf{x}_{nh})\|_2^2$.

Optimization of Variational Score via Stochastic Approximation (SA)

Simulate the backward trajectory $\overleftarrow{\mathbf{x}}_{nh}^{(k+1)}$ given $\mathbf{A}_{nh}^{(k)}$ via Eq.(22), where $\overleftarrow{\mathbf{x}}_{(N-1)}^{(k+1)} \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma}_{(N-1)h|0}^{(k)})$.

Optimize variational score $\mathbf{A}_{nh}^{(k+1)}$ using the loss function (10), where $n \in \{0, 1, \cdots, N-1\}$:

$$\mathbf{A}_{nh}^{(k+1)} = \mathbf{A}_{nh}^{(k)} - \eta_{k+1} \nabla \overrightarrow{\mathcal{L}}_{nh}(\mathbf{A}_{nh}^{(k)}; \overleftarrow{\mathbf{x}}_{nh}^{(k+1)}). \tag{19}$$

until Stage $k = k_{\max}$

Sample $\overleftarrow{\mathbf{x}}_0$ with stochastic (resp. deterministic) trajectories via the discretized Eq.(17) (resp. Eq.(18)).

step size η_k, suppose $\mathbf{A}_t^{(k)} \in \boldsymbol{\Theta}$ is somewhere near one equilibrium $\mathbf{A}_t^\star$ (out of all equilibrium), then by the induction method, the iteration tends to get trapped in the same region as shown in Eq.(32) and yields the convergence to one equilibrium $\mathbf{A}_t^\star$. We also present the variational gap of the (sub)-optimal transport and show our transport is more efficient than diffusion models with Gaussian marginals.

Next, we outline informal assumptions and sketch our main results, reserving formal ones for readers interested in the details in the appendix. We also formulate the optimization of the variational score $\mathbf{A}_t$ using stochastic approximation in Algorithm 2 in the supplementary material.

Assumption A1 (Regularity). *(Positive definiteness) For any $t \geq 0$ and $\mathbf{A}_t \in \mathcal{A}$, $\mathbf{D}_t = \mathbf{I} - 2\mathbf{A}_t$ is positive definite. (Locally strong convexity) For any stable local minimum $\mathbf{A}_t^\star$ with $\nabla \overrightarrow{\mathbf{L}}_t(\mathbf{A}_t^\star) = \mathbf{0}$, there is always a neighborhood $\boldsymbol{\Theta}$ s.t. $\mathbf{A}_t^\star \in \boldsymbol{\Theta} \subset \mathcal{A}$ and $\overrightarrow{\mathbf{L}}_t$ is strongly convex in $\boldsymbol{\Theta}$.*

By the mode-seeking property of the exclusive (reverse) KL divergence (Chan et al., 2022), we only make a mild assumption on a small neighborhood of the solution and expect the convergence given proper regularities.

Assumption A2 (Lipschitz Score). *For any $t \in [0, T]$, the score $\nabla \log \overrightarrow{\rho}_t$ is L-Lipschitz.*

Assumption A3 (Second Moment Bound). *The data distribution has a bounded second moment.*

Assumption A4 (Score Estimation Error). *We have bounded score estimation errors in L^2 quantified by ϵ_{score}.*

We first use the multivariate diffusion to train our score estimators $\{s_t^{(k)}\}_{n=0}^{N-1}$ via the loss function (16) based on the pre-specified $\mathbf{A}_t^{(k)}$ at step k. Similar in spirit to Chen et al. (2023a; 2022a), we can show the generated samples based on $\{s_t^{(k)}\}_{n=0}^{N-1}$ are close in distribution to the ideal samples in Theorem 1. The novelty lies in the extension of single-variate diffusions to multi-variate diffusions.

Theorem 1 (Generation quality, informal). *Assume assumptions A1-A4 hold with a fixed $\mathbf{A}_t^{(k)}$, the generated data distribution is close to the data distributions ρ_{data} such that*

$$\mathrm{TV}(\overleftarrow{\rho}_0^{(k)}, \rho_{data}) \lesssim \exp(-T) + (\sqrt{dh} + \epsilon_{score})\sqrt{T}.$$

To show the convergence of $\mathbf{A}_t^{(k)}$ to $\mathbf{A}_t^\star$, the proof hinges on a stability condition such that the solution asymptotically tracks the equilibrium $\mathbf{A}_t^\star$ of the mean field (20).

Lemma 2 (Local stability, informal). *Assume the assumptions A1 and A2 hold. For $\forall t \in [0, T]$ and $\forall \mathbf{A} \in \boldsymbol{\Theta}$, the solution satisfies a local stability condition such that*

$$\langle \mathbf{A} - \mathbf{A}_t^\star, \nabla \overrightarrow{\mathbf{L}}_t(\mathbf{A}) \rangle \gtrsim \|\mathbf{A} - \mathbf{A}_t^\star\|_2^2.$$

The preceding result illustrates the convergence of the solution toward the equilibrium on average. The next assumption assumes a standard slow update of the SA process, which is standard for theoretical analysis but may not be always needed in empirical evaluations.

Assumption A5 (Step size). *The step size $\{\eta_k\}_{k \in \mathrm{N}}$ is a positive and decreasing sequence*

$$\eta_k \to 0, \ \sum_{k=1}^{\infty} \eta_k = +\infty, \ \sum_{k=1}^{\infty} \eta_k^2 < +\infty.$$

Next, we use the stochastic approximation theory to prove the convergence of $\mathbf{A}_t^{(k)}$ to an equilibrium $\mathbf{A}_t^\star$.

Theorem 2 (Convergence in L^2). *Assume assumptions A1-A5 hold. The variational score $\mathbf{A}_t^{(k)}$ converges to an equilibrium $\mathbf{A}_t^\star$ in L^2 such that*

$$\mathbb{E}[\|\mathbf{A}_t^{(k)} - \mathbf{A}_t^\star\|_2^2] \leq 2\eta_k,$$

where the expectation is taken w.r.t samples from $\overleftarrow{\rho}_t^{(k)}$.

In the end, we adapt Theorem 1 again to show the adaptively generated samples are asymptotically close to the samples based on the optimal $\mathbf{A}_t^\star$ in Theorem 3, which quantifies the quality of data based on more efficient transportation plans.

Theorem 3 (Generation quality of adaptive samples). *Given assumptions A1-A5, the generated sample distribution at stage k is close to the exact sample distribution based on the equilibrium $\mathbf{A}_t^\star$ such that*

$$\mathrm{TV}(\overleftarrow{\rho}_0^\star, \rho_{data}) \lesssim \exp(-T) + (\sqrt{dh} + \epsilon_{score} + \sqrt{\eta_k})\sqrt{T}.$$

6. Variational Gap

Recall that the optimal and variational forward SDEs follow

$$\mathrm{d}\overrightarrow{\mathbf{x}}_t = \left[\boldsymbol{f}_t(\overrightarrow{\mathbf{x}}_t) + \beta_t \nabla \log \overrightarrow{\psi}_t(\overrightarrow{\mathbf{x}}_t)\right]\mathrm{d}t + \sqrt{\beta_t}\mathrm{d}\overrightarrow{\mathbf{w}}_t,$$
$$\mathrm{d}\overrightarrow{\mathbf{x}}_t = \left[\boldsymbol{f}_t(\overrightarrow{\mathbf{x}}_t) + \beta_t \mathbf{A}_t^{(k)}\overrightarrow{\mathbf{x}}_t\right]\mathrm{d}t + \sqrt{\beta_t}\mathrm{d}\overrightarrow{\mathbf{w}}_t,$$
$$\mathrm{d}\overrightarrow{\mathbf{x}}_t = \left[\boldsymbol{f}_t(\overrightarrow{\mathbf{x}}_t) + \beta_t \mathbf{A}_t^\star\overrightarrow{\mathbf{x}}_t\right]\mathrm{d}t + \sqrt{\beta_t}\mathrm{d}\overrightarrow{\mathbf{w}}_t,$$

where we abuse the notion of $\overrightarrow{\mathbf{x}}_t$ for the sake of clarity and they represent three different processes. Despite the improved efficiency based on the ideal $\mathbf{A}_t^\star$ compared to the vanilla $\mathbf{A}_t \equiv \mathbf{0}$, the variational score inevitably yields a sub-optimal transport in general nonlinear transport. We denote the law of the above processes by L, $\mathrm{L}^{(k)}$, and $\mathrm{L}^\star$. To assess the disparity, we leverage the Girsanov theorem to study the variational gap.

Theorem 4 (Variational gap). *Assume the assumption A2 and Novikov's condition hold. Assume $\boldsymbol{f}_t$ and $\nabla \log \overrightarrow{\psi}_t$ are Lipschitz smooth and satisfy the linear growth. The variational gap follows that*

$$\mathrm{KL}(\mathrm{L}\|\mathrm{L}^\star) = \frac{1}{2}\int_0^T \mathbb{E}\left[\beta_t \|\mathbf{A}_t^\star \overrightarrow{\mathbf{x}}_t - \nabla \log \overrightarrow{\psi}_t(\overrightarrow{\mathbf{x}}_t)\|_2^2\right]\mathrm{d}t$$
$$\mathrm{KL}(\mathrm{L}\|\mathrm{L}^{(k)}) \lesssim \eta_k + \mathrm{KL}(\mathrm{L}\|\mathrm{L}^\star).$$

Connections to Gaussian Schrödinger bridge (GSB) When data follows a Gaussian distribution, VSDM approximates the closed-form OT solution of Schrödinger bridge (Janati et al., 2020; Bunne et al., 2023). We refer readers to Theorem 3 (Bunne et al., 2023) for the detailed transportation plans. Compared to the vanilla $\mathbf{A}_t \equiv \mathbf{0}$, we can significantly reduce the variational gap with $\mathrm{KL}(\mathrm{L}\|\mathrm{L}^\star)$ using proper parametrization and sufficient training.

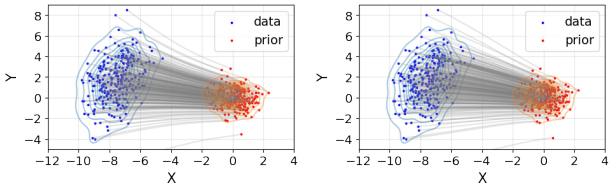
We briefly compare VSDM to SGM and SB in the following:

Properties	SGM	SB	VSDM
Entropic Optimal Transport	×	Optimal	Sub-Optimal
Simulation-Free (backward)	✓	×	✓

7. Empirical Studies

7.1. Comparison to Gaussian Schrodinger Bridge

VSDM is approximating GSB (Bunne et al., 2023) when both marginals are Gaussian distributions. To evaluate the solutions, we run our VSDM with a fixed $\beta_t \equiv 4$ in Eq.(25) in Song et al. (2021b) and use the same marginals to replicate the VPSDE of the Gaussian SB with $\alpha_t \equiv 0$ and $c_t \equiv -2$ in Eq.(7) in Bunne et al. (2023). We train VSDM with 20 stages and randomly pick 256 samples for presentation. We compare the flow trajectories from both models and observe in Figure 1 that the ground truth solution forms an almost linear path, while our VSDM sample trajectories exhibit a consistent alignment with trajectories from Gaussian SB. We attribute the bias predominantly to score estimations and numerical discretization.



(a) GSB (b) VSDM

Figure 1. Gaussian SB (GSB) v.s. VSDM on the flow trajectories.

7.2. Synthetic Data

We test our variational Schrödinger diffusion models (VSDMs) on two synthetic datasets: spiral and checkerboard (detailed in section D.2.1). We include SGMs as the baseline models and aim to show the strength of VSDMs on general shapes with straighter trajectories. As such, we stretch the Y-axis of the spiral data by 8 times and the X-axis of the checkerboard data by 6 times and denote them by spiral-8Y and checkerboard-6X, respectively.

We adopt a monotone increasing $\{\beta_{nh}\}_{n=0}^{N-1}$ similar to Song et al. (2021b) and denote by $\beta_{\min}$ and $\beta_{\max}$ the minimum and maximum of $\{\beta_{nh}\}_{n=0}^{N-1}$. We fix $\zeta = 0.75$ and $\beta_{\min} = 0.1$ and we focus on the study with different $\beta_{\max}$. We find that SGMs work pretty well with $\beta_{\max} = 10$ (SGM-10) on standard isotropic shapes. However, when it comes to

spiral-8Y, the SGM-10 *struggles to recover* the boundary regions on the spiral-8Y data as shown in Figure 2 (top).

Generations of Anisotropic Shapes To illustrate the effectiveness of our approach, Figure 2 (bottom) shows that VSDM-10 accurately reconstructs the edges of the spiral and generates high-quality samples.

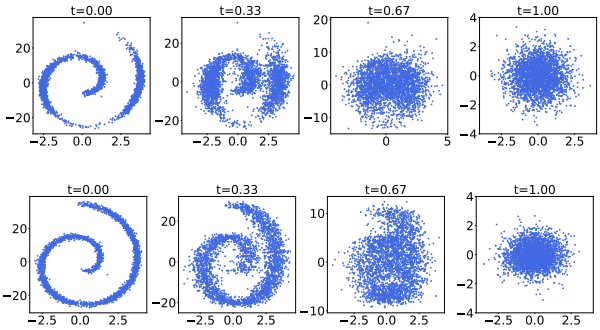
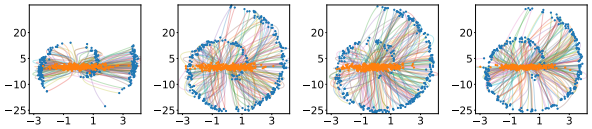


Figure 2. Variational Schrödinger diffusion models (VSDMs, bottom) v.s. SGMs (top) with the same hyperparameters ($\beta_{\max} = 10$).

Straighter Trajectories The SGM-10 fails to fully generate the anisotropic spiral-8Y and increasing $\beta_{\max}$ to 20 or 30 (SGM-20 and SGM-30) significantly alleviates this issue. However, we observe that excessive $\beta_{\max}$ values in SGMs compromise the straightness and leads to inefficient transport, especially in the X-axis of spiral-8Y.



(a) SGM-10 (b) SGM-20 (c) SGM-30 (d) VSDM-10

Figure 3. Probability flow ODE via VSDMs and SGMs. SGM with $\beta_{\max} = 10$ is denoted by SGM-10 for convenience.

Instead of setting excessive $\beta_{\max}$ on both axes, our VSDM-10, by contrast, proposes conservative diffusion scales on the X-axis of spiral-8Y and explores more on the Y-axis of spiral-8Y. As such, we obtain around 40% improvement on the straightness in Figure 3 and Table 4.

Additional insights into a similar analysis of the checkboard dataset, convergence analysis, computational time, assessments of straightness, and evaluations via a smaller number of function evaluations (NFEs) can be found in Appendix D.2.

7.3. Image Data Modeling

Experiment Setup In this experiment, we evaluate the performance of VSDM on image modeling tasks. We choose the CIFAR10 datasetas representative image data to demonstrate the scalability of the proposed VSDM on generative modeling of high-dimensional distributions. We refer to the code base of FB-SDE (Chen et al., 2022b) and use the same forward diffusion process of the EDM model (Karras et al., 2022). Since the training of VSDM is an alternative manner between forward and backward training, we build our implementations based on the open-source diffusion distillation code base (Luo et al., 2024a) [¶], which provides a high-quality empirical implementation of alternative training with EDM model on CIFAR10 data. To make the VSDM algorithm stable, we simplify the matrix $\mathbf{D}_t$ to be diagonal with learnable diagonal elements, which is the case as we introduced in Corollary 1. We train the VSDM model from scratch on two NVIDIA A100-80G GPUs for two days and generate images from the trained VSDM with the Euler–Maruyama numerical solver with 200 discretized steps for generation.

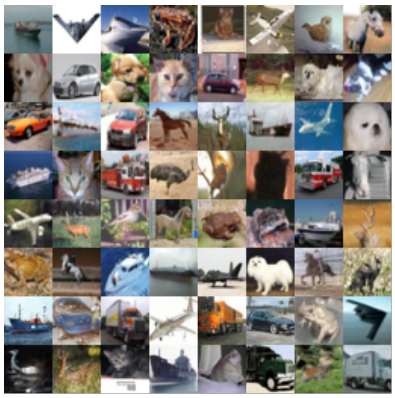


Figure 4. Unconditional generated samples from VSDM on CIFAR10 (32×32 resolution) trained from scratch.

Performances. We measure the generative performances in terms of the Fretchat Inception Score (FID (Heusel et al., 2017), the lower the better), which is a widely used metric for evaluating generative modeling performances.

Tables 2 summarize the FID values of VSDM along with other optimal-transport-based and score-based generative models on the CIFAR10 datasets (unconditional without labels). The VSDM outperforms other optimal transport-based models with an FID of 2.28. This demonstrates that the VSDM has applicable scalability to model high-dimensional distributions. Figure 7.3 shows some non-cherry-picked unconditional generated samples from VSDM trained on the CIFAR10 dataset.

Convergence Speed. To demonstrate the convergence speed of VSDM along training processes, we record the FID values in Table 1 for a training trail with no warmup on CIFAR10 datasets (unconditional). We use a batch size of 256 and a learning rate of $1e-4$. We use the 2nd-order Heun numerical solver to sample. The result shows that VSDM has a smooth convergence performance.

[¶] See code in https://github.com/pku1wj1994/diff_instruct

Table 1. CONVERGENCE SPEED OF FID VALUES FOR VSDM.

K IMAGES	0	10K	20K	30K	40K	50K	100K	150K	200K	CONVERGE
FID↓ (NFE=35)	406.13	13.13	8.65	6.83	5.66	5.21	3.62	3.29	3.01	2.28

Table 2. CIFAR10 EVALUATION USING SAMPLE QUALITY (FID SCORE). OUR VSDM OUTPERFORMS OTHER OPTIMAL TRANSPORT BASELINES BY A LARGE MARGIN.

CLASS	METHOD	FID ↓
OT	**VSDM (OURS)**	**2.28**
	SB-FBSDE (CHEN ET AL., 2022B)	3.01
	DOT (TANAKA, 2019)	15.78
	DGFLOW (ANSARI ET AL., 2020)	9.63
SGMS	SDE (SONG ET AL. (2021B))	2.92
	SCOREFLOW (SONG ET AL., 2021A)	5.7
	VDM (KINGMA ET AL., 2021)	4.00
	LSGM(VAHDAT ET AL., 2021)	2.10
	EDM(KARRAS ET AL., 2022)	**1.97**

Table 3. FORECASTING RESULTS (LOWER IS BETTER).

CRPS-SUM	ELECTRICITY	EXCHANGE RATE	SOLAR
DDPM	**0.026±0.007**	0.012±0.001	0.506±0.058
SGM	0.045±0.005	0.012±0.002	0.413±0.045
VSDM (OUR)	0.038±0.006	**0.008±0.002**	**0.395±0.011**

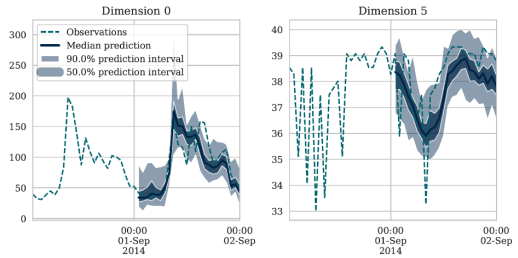


Figure 5. Example for Electricity for 2 (out of 370) dimensions.

7.4. Time Series Forecasting

We use multivariate probabilistic forecasting as a real-world *conditional* modeling task. Let $\{(t_1, \mathbf{x}_1), \ldots, (t_n, \mathbf{x}_n)\}$, $\mathbf{x} \in \mathbb{R}^d$, denote a single multivariate time series. Given a dataset of such time series we want to predict the next P values $\mathbf{x}_{n+1}, \ldots, \mathbf{x}_{n+P}$. In probabilistic modeling, we want to generate forecasts from learned $p(\mathbf{x}_{n+1:n+P}|\mathbf{x}_{1:n})$.

The usual approach is to have an encoder that represents a sequence $\mathbf{x}_{1:i}$ with a fixed-sized vector $\boldsymbol{h}_i \in \mathbb{R}^h$, $\forall i$, and then parameterize the output distribution $p(\mathbf{x}_{i+1}|\boldsymbol{h}_i)$. At inference time we encode the history into $\boldsymbol{h}_n$ and sample the next value from $p(\mathbf{x}_{n+1}|\boldsymbol{h}_n)$, then use $\mathbf{x}_{n+1}$ to get the updated $\boldsymbol{h}_{n+1}$ and repeat until we obtain $\mathbf{x}_{n+P}$.

In the previous works, the output distribution has been specified with a Copulas (Salinas et al., 2019) and denoising diffusion (Rasul et al., 2021). We augment our approach to allow conditional generation which requires only changing the model to include the conditioning vector $\boldsymbol{h}_i$. For that we adopt the U-Net architecture. We use the LSTM neural network as a sequence encoder.

We use three real-world datasets, as described in Appendix D.3. We compare to the SGM and the denoising diffusion approach from Rasul et al. (2021) which we refer to as DDPM. Table 3 shows that our method matches or outperforms the competitors. Figure 5 is a demo for conditional time series generation and more details are presented in Figure 12 to demonstrate the quality of the forecasts.

8. Conclusions and Future Works

The Schrödinger bridge diffusion model offers a principled approach to solving optimal transport, but estimating the intractable forward score relies on implicit training through costly simulated trajectories. To address this scalability issue, we present the variational Schrödinger diffusion model (VSDM), utilizing linear variational forward scores for simulation-free training of backward score functions. Theoretical foundations leverage stochastic approximation theory, demonstrating the convergence of variational scores to local equilibrium and highlighting the variational gap in optimal transport. Empirically, VSDM showcases the strength of generating data with anisotropic shapes and yielding the desired straighter transport paths for reducing the number of functional evaluations. VSDM also exhibits scalability in handling large-scale image datasets without requiring warm-up initializations. In future research, we aim to explore the critically damped (momentum) acceleration (Dockhorn et al., 2022) and Hessian approximations to develop the "ADAM" alternative of diffusion models.

Acknowledgements

We thank Valentin De Bortoli, Tianyang Hu, and the anonymous reviewers for their valuable insights.

Impact Statement

This paper proposed a principled approach to accelerate the training and sampling of generative models using optimal transport. This work will contribute to developing text-to-image generation, artwork creation, and product design. However, it may also raise challenges in the fake-content generation and pose a threat to online privacy and security.

References

Albergo, M. S. and Vanden-Eijnden, E. Building Normalizing Flows with Stochastic Interpolants. In *International Conference on Learning Representation (ICLR)*, 2023.

Albergo, M. S., Bof, N. M., and Vanden-Eijnden, E. Stochastic Interpolants: A Unifying Framework for Flows and Diffusions. *arXiv:2303.08797v1*, pp. 1–48, 2023.

Anderson, B. D. Reverse-time Diffusion Equation Models. *Stochastic Processes and Their Applications*, 12(3):313–326, 1982.

Ansari, A. F., Ang, M. L., and Soh, H. Refining Deep Generative Models via Discriminator Gradient Flow. In *International Conference on Learning Representations*, 2020.

Benveniste, A., Métivier, M., and Priouret, P. *Adaptive Algorithms and Stochastic Approximations*. Berlin: Springer, 1990.

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. Variational Inference: A Review for Statisticians. *Journal of the American Statistical Association*, 112 (518), 2017.

Bunne, C., Hsieh, Y.-P., Cuturi, m., and Krause, A. The Schrödinger Bridge between Gaussian Measures has a Closed Form. In *AISTATS*, 2023.

Caluya, K. F. and Halder, A. Wasserstein Proximal Algorithms for the Schrödinger Bridge Problem: Density Control with Nonlinear Drift. *IEEE Transactions on Automatic Control*, 67(3):1163–1178, 2022.

Chan, A., Silva, H., Lim, S., Kozuno, T., Mahmood, A. R., and White, M. Greedification Operators for Policy Optimization: Investigating Forward and Reverse KL Divergences. *Journal of Machine Learning Research*, 2022.

Chen, H., Lee, H., and Lu, J. Improved Analysis of Score-based Generative Modeling: User-friendly Bounds under Minimal Smoothness Assumptions. In *International Conference on Machine Learning*, pp. 4735–4763, 2023a.

Chen, S., Chewi, S., Li, J., Li, Y., Salim, A., and Zhang, A. R. Sampling is as Easy as Learning the Score: Theory for Diffusion Models with Minimal Data Assumptions. *arXiv preprint arXiv:2209.11215v2*, 2022a.

Chen, T., Liu, G.-H., and Theodorou, E. A. Likelihood Training of Schrödinger Bridge using Forward-Backward SDEs Theory. In *International Conference on Learning Representation (ICLR)*, 2022b.

Chen, T., Gu, J., Dinh, L., Theodorou, E. A., Susskind, J., and Zhai, S. Generative Modeling with Phase Stochastic Bridges. In *arXiv:2310.07805v2*, 2023b.

Chen, Y. and Georgiou, T. Stochastic Bridges of Linear Systems. *IEEE Transactions on Automatic Control*, 61 (2), 2016.

Chen, Y., Georgiou, T. T., and Pavon, M. Stochastic Control Liaisons: Richard Sinkhorn Meets Gaspard Monge on a Schrödinger Bridge. *SIAM Review*, 63(2):249–313, 2021.

Chen, Y., Deng, W., Fang, S., Li, F., Yang, N., Zhang, Y., Rasul, K., Zhe, S., Schneider, A., and Nevmyvaka, Y. Provably Convergent Schrödinger Bridge with Applications to Probabilistic Time Series Imputation. In *ICML*, 2023c.

Chewi, S. *Log-Concave Sampling*. online draft, 2023.

De Bortoli, V., Thornton, J., Heng, J., and Doucet, A. Diffusion Schrödinger Bridge with Applications to Score-Based Generative Modeling. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Deng, W., Chen, Y., Yang, N., Du, H., Feng, Q., and Chen, R. T. Q. Reflected Schrödinger Bridge for Constrained Generative Modeling. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, 2024.

Dhariwal, P. and Nichol, A. Diffusion Models Beat GANs on Image Synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Dockhorn, T., Vahdat, A., and Kreis, K. Score-Based Generative Modeling with Critically-Damped Langevin Diffusion. In *International Conference on Learning Representation (ICLR)*, 2022.

Finlay, C., Jacobsen, J.-H., Nurbekyan, L., and Oberman, A. How to Train Your Neural ODE: the World of Jacobian and Kinetic Regularization. In *ICML*, 2020.

Grathwohl, W., Chen, R. T. Q., Bettencourt, J., Sutskever, I., and Duvenaud, D. FFJORD: Free-form Continuous Dynamics for Scalable Reversible Generative Models. In *International Conference on Learning Representation (ICLR)*, 2019.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In *Neural Information Processing Systems*, 2017.

Ho, J., Jain, A., and Abbeel, P. Denoising Diffusion Probabilistic Models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Ho, J., Chan, W., Saharia, C., Whang, J., Gao, R., Gritsenko, A., Kingma, D. P., Poole, B., Norouzi, M., Fleet, D. J., and Salimans, T. Imagen Video: High Definition Video Generation with Diffusion Models. In *arXiv:2210.02303*, 2022.

Huang, C.-W., Lim, J. H., and Courville, A. A Variational Perspective on Diffusion-Based Generative Models and Score Matching. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Hutchinson, M. F. A Stochastic Estimator of the Trace of the Influence Matrix for Laplacian Smoothing Splines. *Communications in Statistics-Simulation and Computation*, 18(3):1059–1076, 1989.

Hyvärinen, A. Estimation of Non-normalized Statistical Models by Score Matching. *Journal of Machine Learning Research*, 6(24):695–709, 2005.

Janati, H., Muzellec, B., Peyré, G., and Cuturi, M. Entropic Optimal Transport between Unbalanced Gaussian Measures has a Closed Form. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Karatzas, I. and Shreve, S. E. *Brownian Motion and Stochastic Calculus*. Springer, 1998.

Karras, T., Aittala, M., Aila, T., and Laine, S. Elucidating the Design Space of Diffusion-Based Generative Models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Kingma, D. P., Salimans, T., Poole, B., and Ho, J. Variational Diffusion Models. *ArXiv*, abs/2107.00630, 2021.

Koehler, F., Heckett, A., and Risteski, A. Statistical Efficiency of Score Matching: The View from Isoperimetry. In *ICLR*, 2023.

Kong, Z., Ping, W., Huang, J., Zhao, K., and Catanzaro, B. DiffWave: A Versatile Diffusion Model for Audio Synthesis . In *Proc. of the International Conference on Learning Representation (ICLR)*, 2021.

Kullback, S. Probability Densities with Given Marginals. *Ann. Math. Statist.*, 1968.

Lavenant, H. and Santambrogio, F. The Flow Map of the Fokker–Planck Equation Does Not Provide Optimal Transport. *Applied Mathematics Letters*, 133, 2022.

Lee, H., Lu, J., and Tan, Y. Convergence for Score-based Generative Modeling with Polynomial Complexity. *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Léonard, C. A Survey of the Schrödinger Problem and Some of its Connections with Optimal Transport. *Discrete & Continuous Dynamical Systems-A*, 34(4):1533–1574, 2014.

Liang, F., Liu, C., and Carroll, R. J. Stochastic Approximation in Monte Carlo Computation. *Journal of the American Statistical Association*, 102:305–320, 2007.

Lipman, Y., Chen, R. T. Q., Ben-Hamu, H., Nickel, M., and Le, M. Flow Matching for Generative Modeling. In *Proc. of the International Conference on Learning Representation (ICLR)*, 2023.

Liptser, R. S. and Shiryaev, A. N. *Statistics of Random Processes: I. General Theory*. Springer Science & Business Media, 2001.

Liu, Q. Rectified Flow: A Marginal Preserving Approach to Optimal Transport. *arXiv:2209.14577*, 2022.

Liu, X., Gong, C., and Liu, Q. Flow Straight and Fast: Learning to Generate and Transfer Data with Rectified Flow. In *International Conference on Learning Representation (ICLR)*, 2023.

Lu, C., Zhou, Y., Bao, F., Chen, J., Li, C., and Zhu, J. DPM-Solver: A Fast ODE Solver for Diffusion Probabilistic Model Sampling in Around 10 Steps. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Luo, W. A Comprehensive Survey on Knowledge Distillation of Diffusion Models. *arXiv preprint arXiv:2304.04262*, 2023.

Luo, W., Hu, T., Zhang, S., Sun, J., Li, Z., and Zhang, Z. Diff-instruct: A Universal Approach for Transferring Knowledge from Pre-trained Diffusion Models. *Advances in Neural Information Processing Systems*, 36, 2024a.

Luo, W., Zhang, B., and Zhang, Z. Entropy-based Training Methods for Scalable Neural Implicit Samplers. *NeurIPS*, 36, 2024b.

Ma, J. and Yong, J. *Forward-Backward Stochastic Differential Equations and their Applications*. Springer, 2007.

Marzouk, Y., Moselhy, T., Parno, M., and Spantini, A. Sampling via Measure Transport: An Introduction. *Handbook of Uncertainty Quantification*, pp. 1–41, 2016.

McCann, R. J. A Convexity Principle for Interacting Gases. *Advances in mathematics*, 128(1):153–179, 1997.

Øksendal, B. *Stochastic Differential Equations: An Introduction with Applications*. Springer, 2003.

Onken, D., Fung, S. W., Li, X., and Ruthotto, L. OT-Flow: Fast and Accurate Continuous Normalizing Flows via Optimal Transport. In *Proc. of the National Conference on Artificial Intelligence (AAAI)*, 2021.

Pavon, M., Tabak, E. G., and Trigila, G. The Data-driven Schrödinger Bridge. *Communications on Pure and Applied Mathematics*, 74:1545–1573, 2021.

Peluchetti, S. Diffusion Bridge Mixture Transports, Schrödinger Bridge Problems and Generative Modeling. *ArXiv e-prints arXiv:2304.00917v1*, 2023.

Peyré, G. and Cuturi, M. *Computational Optimal Transport: With Applications to Data Science*. Foundations and Trends in Machine Learning, 2019.

Polyak, B. T. and Juditsky, A. Acceleration of Stochastic Approximation by Averaging. *SIAM Journal on Control and Optimization*, 30:838–855, 1992.

Pooladian, A.-A., Ben-Hamu, H., Domingo-Enrich, C., Amos, B., Lipman, Y., and Chen, R. T. Q. Multisample Flow Matching: Straightening Flows with Minibatch Couplings. In *ICML*, 2023.

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., and Chen, M. Hierarchical Text-Conditional Image Generation with CLIP Latents. In *arXiv:2204.06125v1*, 2022.

Rasul, K., Seward, C., Schuster, I., and Vollgraf, R. Autoregressive Denoising Diffusion Models for Multivariate Probabilistic Time Series Forecasting. In *International Conference on Machine Learning*, 2021.

Robbins, H. and Monro, S. A Stochastic Approximation Method. *Annals of Mathematical Statistics*, 22:400–407, 1951.

Ruschendorf, L. Convergence of the Iterative Proportional Fitting Procedure. *Ann. of Statistics*, 1995.

Salimans, T. and Ho, J. Progressive Distillation for Fast Sampling of Diffusion Models. In *ICLR*, 2022.

Salinas, D., Bohlke-Schneider, M., Callot, L., Medico, R., and Gasthaus, J. High-dimensional Multivariate Forecasting with Low-rank Gaussian Copula Processes. *Advances in neural information processing systems*, 2019.

Särkkä, S. and Solin, A. *Applied Stochastic Differential Equations*. Cambridge University Press, 2019.

Shi, Y., De Bortoli, V., Campbell, A., and Doucet, A. Diffusion Schrödinger Bridge Matching. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Singhal, R., Goldstein, M., and Ranganath, R. Where to Diffuse, How to Diffuse, and How to Get Back: Automated Learning for Multivariate Diffusions. In *Proc. of the International Conference on Learning Representation (ICLR)*, 2023.

Song, Y. and Ermon, S. Improved Techniques for Training Score-Based Generative Models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Song, Y., Garg, S., Shi, J., and Ermon, S. Sliced Score Matching: A Scalable Approach to Density and Score Estimation. In *Uncertainty in Artificial Intelligence*, 2020.

Song, Y., Durkan, C., Murray, I., and Ermon, S. Maximum Likelihood Training of Score-Based Diffusion Models . In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021a.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. Score-Based Generative Modeling through Stochastic Differential Equations . In *International Conference on Learning Representation (ICLR)*, 2021b.

Tanaka, A. Discriminator Optimal Transport. In *Neural Information Processing Systems*, 2019.

Tong, A., Malkin, N., Huguet, G., Zhang, Y., Rector-Brooks, J., Fatras, K., Wolf, G., and Bengio, Y. Improving and Generalizing Flow-based Generative Models with Minibatch Optimal Transport. *arXiv:2302.00482v3*, 2023.

Trivedi, S. and Kondor, R. Optimization for Deep Neural Networks. *Slides - University of Chicago*, 2017.

Vahdat, A., Kreis, K., and Kautz, J. Score-based Generative Modeling in Latent Space. *Advances in Neural Information Processing Systems*, 34:11287–11302, 2021.

Vanden-Eijnden, E. Introduction to Regular Perturbation Theory. *Slides*, 2001. URL https://cims.nyu.edu/~eve2/reg_pert.pdf.

Vargas, F., Thodoroff, P., Lamacraft, A., and Lawrence, N. Solving Schrödinger Bridges via Maximum Likelihood. *Entropy*, 23(9):1134, 2021.

Vempala, S. S. and Wibisono, A. Rapid Convergence of the Unadjusted Langevin Algorithm: Isoperimetry Suffices, 2022.

Villani, C. *Topics in Optimal Transportation*, volume 58. American Mathematical Soc., 2003.

Vono, M., Paulin, D., and Doucet, A. Efficient MCMC Sampling with Dimension-Free Convergence Rate using ADMM-type Splitting. *Journal of Machine Learning Research*, 2022.

Xue, S., Yi, M., Luo, W., Zhang, S., Sun, J., Li, Z., and Ma, Z.-M. SA-Solver: Stochastic Adams Solver for Fast Sampling of Diffusion Models. *Advances in Neural Information Processing Systems*, 2023.

Zhang, B., Luo, W., and Zhang, Z. Enhancing Adversarial Robustness via Score-Based Optimization. *Advances in Neural Information Processing Systems*, 36, 2024.

Supplementary Material for "Variational Schrödinger Diffusion Models"

In section A, we study the closed-form expression of matrix exponential for diagonal and time-invariant $\mathbf{D}_t$; In section B, we study the convergence of the adaptive diffusion process; In section C, we study the variational gap of the optimal transport and discuss its connections to Gaussian Schrödinger bridge; In section D, we present more details on the empirical experiments.

Notations: $\mathcal{X}$ is the state space for the data $\mathbf{x}$; $\overleftarrow{\mathbf{x}}_{nh}^{(k)}$ is the n-th backward sampling step with a learning rate h at the k-th stage. η_k is the step size to optimize $\mathbf{A}$. $\mathcal{A}$ is the (latent) state space of $\mathbf{A}$; $\mathbf{A}_t^{(k)}$ is the forward linear score estimator at stage k and time t, $\mathbf{A}_t^\star$ is the equilibrium of Eq.(25) at time t. $\nabla \overrightarrow{\mathcal{L}}_t$ is the random field in the stochastic approximation process and also the loss (10) at time t; $\nabla \overrightarrow{\mathbf{L}}_t$ is the mean field with the equilibrium $\mathbf{A}_t^\star$. Given a fixed $\mathbf{A}_t^{(k)}$ at step k, $\nabla \log \overrightarrow{\rho}_t^{(k)}$ (resp. $\nabla \log \overrightarrow{\rho}_{t|0}^{(k)}$) is the (resp. conditional) forward score function of Eq.(11) at time t and step k; $\mathbf{A}_t^{(k)}$ yields the approximated score function $s_t^{(k)}$ and $\overleftarrow{\rho}_t^{(k)}$ is the distribution of the continuous-time interpolation of the discretized backward SDE (22).

A. Closed-form Expression with Diagonal and Time-Invariant D_t

In this section, we give the proof of Corollary 1.

Proof Denote $\mathbf{D}_t = \mathbf{\Lambda} \coloneqq \mathrm{diag}(\boldsymbol{\lambda})$, where $\lambda_i \geq 0,\ \forall 1 \leq i \leq d$, and $\sigma_t^2 \coloneqq \int_0^t \beta_s \mathrm{d}s$, then by Eq. (14), we have

$$\begin{pmatrix} \mathbf{C}_t \\ \mathbf{H}_t \end{pmatrix} = \exp\left[\begin{pmatrix} -\frac{1}{2}[\beta\mathbf{D}]_t & [\beta\mathbf{I}]_t \\ \mathbf{0} & \frac{1}{2}[\beta\mathbf{D}^\intercal]_t \end{pmatrix}\right]\begin{pmatrix} \mathbf{\Sigma}_0 \\ \mathbf{I} \end{pmatrix} = \exp(\mathbf{M}_t)\begin{pmatrix} \mathbf{\Sigma}_0 \\ \mathbf{I} \end{pmatrix}.$$

Here $[\beta\mathbf{D}]_t = \int_0^t \beta_s \mathbf{D}_t \mathrm{d}s = \sigma_t^2 \mathbf{\Lambda}$. The matrix $\mathbf{M}_t$ is defined as

$$\mathbf{M}_t = \begin{pmatrix} -\frac{1}{2}\sigma_t^2\mathbf{\Lambda} & \sigma_t^2\mathbf{I} \\ \mathbf{0} & \frac{1}{2}\sigma_t^2\mathbf{\Lambda} \end{pmatrix}$$

Therefore, we have

$$\mathbf{M}_t^2 = \begin{pmatrix} (-\frac{1}{2}\sigma_t^2\mathbf{\Lambda})^2 & \mathbf{0} \\ \mathbf{0} & (\frac{1}{2}\sigma_t^2\mathbf{\Lambda})^2 \end{pmatrix},\ \mathbf{M}_t^3 = \begin{pmatrix} (-\frac{1}{2}\sigma_t^2\mathbf{\Lambda})^3 & \sigma_t^2(\frac{1}{2}\sigma_t^2\mathbf{\Lambda})^2 \\ \mathbf{0} & (\frac{1}{2}\sigma_t^2\mathbf{\Lambda})^3 \end{pmatrix},$$
$$\mathbf{M}_t^4 = \begin{pmatrix} (-\frac{1}{2}\sigma_t^2\mathbf{\Lambda})^4 & \mathbf{0} \\ \mathbf{0} & (\frac{1}{2}\sigma_t^2\mathbf{\Lambda})^4, \end{pmatrix},\ \mathbf{M}_t^5 = \begin{pmatrix} (-\frac{1}{2}\sigma_t^2\mathbf{\Lambda})^5 & \sigma_t^2(\frac{1}{2}\sigma_t^2\mathbf{\Lambda})^4 \\ \mathbf{0} & (\frac{1}{2}\sigma_t^2\mathbf{\Lambda})^5 \end{pmatrix},\ \ldots$$

According to the definition of matrix exponential, we have

$$\begin{aligned}
\exp(\mathbf{M}_t) &= [\mathbf{I} + \frac{1}{1!}\mathbf{M}_t + \frac{1}{2!}\mathbf{M}_t^2 + \frac{1}{3!}\mathbf{M}_t^3 + \ldots] \\
&= \begin{pmatrix} \exp(-\frac{1}{2}\sigma_t^2\mathbf{\Lambda}) & \left[\sigma_t^2\mathbf{I} + \frac{1}{3!}\sigma_t^2(\frac{1}{2}\sigma_t^2\mathbf{\Lambda})^2 + \frac{1}{5!}(\frac{1}{2}\sigma_t^2\mathbf{\Lambda})^4 + \ldots\right] \\ \mathbf{0} & \exp(\frac{1}{2}\sigma_t^2\mathbf{\Lambda}) \end{pmatrix} \\
&= \begin{pmatrix} \exp(-\frac{1}{2}\sigma_t^2\mathbf{\Lambda}) & \frac{\sigma_t^2}{\frac{1}{2}\sigma_t^2\mathbf{\Lambda}}\left[(\frac{1}{2}\sigma_t^2\mathbf{\Lambda})^1 + \frac{1}{3!}\sigma_t^2(\frac{1}{2}\sigma_t^2\mathbf{\Lambda})^3 + \frac{1}{5!}(\frac{1}{2}\sigma_t^2\mathbf{\Lambda})^5 + \ldots\right] \\ \mathbf{0} & \exp(\frac{1}{2}\sigma_t^2\mathbf{\Lambda}) \end{pmatrix} \\
&= \begin{pmatrix} \exp(-\frac{1}{2}\sigma_t^2\mathbf{\Lambda}) & \mathbf{\Lambda}^{-1}\left[\exp(\frac{1}{2}\sigma_t^2\mathbf{\Lambda}) - \exp(-\frac{1}{2}\sigma_t^2\mathbf{\Lambda})\right] \\ \mathbf{0} & \exp(\frac{1}{2}\sigma_t^2\mathbf{\Lambda}) \end{pmatrix}.
\end{aligned}$$

Notice that, when we have $\mathbf{\Sigma}_0 = \mathbf{0}$, the expression can be simplified as follows

$$\begin{pmatrix} \mathbf{C}_t \\ \mathbf{H}_t \end{pmatrix} = \exp(\mathbf{M}_t)\begin{pmatrix} \mathbf{0} \\ \mathbf{I} \end{pmatrix} = \begin{pmatrix} \mathbf{\Lambda}^{-1}\left[\exp(\frac{1}{2}\sigma_t^2\mathbf{\Lambda}) - \exp(-\frac{1}{2}\sigma_t^2\mathbf{\Lambda})\right] \\ \exp(\frac{1}{2}\sigma_t^2\mathbf{\Lambda}) \end{pmatrix}.$$

Therefore, $\mathbf{C}_t\mathbf{H}_t^{-1} = \mathbf{\Lambda}^{-1}\{\mathbf{I} - \exp(-\sigma_t^2\mathbf{\Lambda})\}$. As a result, the corresponding forward transition writes

$$\boldsymbol{\mu}_{t|0} = \exp(-\frac{1}{2}\sigma_t^2\mathbf{\Lambda})\mathbf{x}_0$$
$$\mathbf{L}_t = \sigma_t\mathbf{\Lambda}^{-\frac{1}{2}}\sqrt{\mathbf{I} - \exp(-\sigma_t^2\mathbf{\Lambda})}.$$

B. Stochastic Approximation

Stochastic approximation (SA), also known as the Robbins–Monro algorithm (Robbins & Monro, 1951; Benveniste et al., 1990) offers a conventional framework for the study of adaptive algorithms. The stochastic approximation algorithm works by repeating the sampling-optimization iterations in the dynamic setting in terms of simulated trajectories. We present our algorithm in Algorithm 2.

Algorithm 2 The (dynamic) stochastic approximation (SA) algorithm. The (dynamic) SA is a theoretical formulation of Algorithm 1. We assume optimizing the loss function (16) yields proper score estimations $s_t^{(k+1)}$ at each stage k and time t to approximate $\nabla \log \overrightarrow{p}_t^{(k)}(\overrightarrow{\mathbf{x}}_t|\overrightarrow{\mathbf{x}}_0)$ in Eq.(9b).

repeat

Simulation: Sample the backward process from (17) given a fixed $\mathbf{A}_{nh}^{(k)}$

$$\overleftarrow{\mathbf{x}}_{(n-1)h}^{(k+1)} = \left(-\frac{1}{2}(\mathbf{I} - 2\mathbf{A}_{nh}^{(k)})\beta_{nh}\overleftarrow{\mathbf{x}}_{nh}^{(k+1)} - \beta_{nh}s_{nh}^{(k+1)}(\overleftarrow{\mathbf{x}}_{nh}^{(k+1)})\right)h + \sqrt{\beta_{nh}h}\boldsymbol{\xi}_n, \tag{22}$$

where $\overleftarrow{\mathbf{x}}_{(N-1)}^{(k+1)} \sim \mathrm{N}(\mathbf{0}, \mathbf{\Sigma}_{(N-1)h|0}^{(k)})$, $n \in [1, 2, \cdots, N-1]$ and h is the learning rate for the backward sampling (17) via the Euler–Maruyama (EM) discretization. $\boldsymbol{\xi}_n$ denotes the standard Gaussian vector at the sampling iteration n.

Optimization: Minimize the implicit forward loss function (10)

$$\mathbf{A}_{nh}^{(k+1)} = \mathbf{A}_{nh}^{(k)} - \eta_{k+1}\nabla\overrightarrow{\mathcal{L}}_{nh}(\mathbf{A}_{nh}^{(k)}; \overleftarrow{\mathbf{x}}_{nh}^{(k+1)}), \tag{23}$$

where $\nabla\overrightarrow{\mathcal{L}}_{nh}(\mathbf{A}_{nh}^{(k)}; \overleftarrow{\mathbf{x}}_{nh}^{(k+1)})$ is the (dynamic) random field and η_k is the step size. $n \in \{0, 1, \cdots, N-1\}$.

until Stage $k = k_{\max}$

To facilitate the analysis, we assume we only make a one-step sampling in Eq.(23). Note that it is not required in practice and multiple-step extensions can be employed to exploit the cached data more efficiently. The theoretical extension is straightforward and omitted in the proof. We also slightly abuse the notation for convenience and generalize $\mathbf{A}_{nh}$ to $\mathbf{A}_t$.

Theoretically, the primary objective is to show the iterates (19) follow the trajectories of the dynamical system asymptotically:

$$\mathrm{d}\mathbf{A}_t = \nabla\overrightarrow{\mathbf{L}}_t(\mathbf{A}_t)\mathrm{d}s, \tag{24}$$

where $\nabla\overrightarrow{\mathbf{L}}_t(\mathbf{A}_t)$ is the mean field defined as follows:

$$\nabla\overrightarrow{\mathbf{L}}_t(\mathbf{A}_t) = \int_{\mathcal{X}} \nabla\overrightarrow{\mathcal{L}}_t(\mathbf{A}_t; \overleftarrow{\mathbf{x}}_t^{(\cdot)})\overleftarrow{\rho}_t(\mathrm{d}\overleftarrow{\mathbf{x}}_t^{(\cdot)}). \tag{25}$$

We denote by $\mathbf{A}_t^\star$ the solution of $\nabla\overrightarrow{\mathbf{L}}_t(\mathbf{A}_t^\star) = \mathbf{0}$. Since the samples simulated from $\overleftarrow{\rho}_t$ are slightly biased due to the convergence of forward process, discretization error, and score estimation errors as shown in Theorem 1. We expect the mean field is also biased with a perturbed equilibrium. However, by the perturbation theory (Vanden-Eijnden, 2001), the perturbation is mild and controlled by the errors in Theorem 1. Hence although $\mathbf{A}_t^\star$ is not the optimal linear solution in terms of optimal transport, it still yields efficient transportation plans.

Since the exclusive (reverse) KL divergence is known to approximate a single mode (denoted by $\mathbf{A}_t^\star$) in fitting multi-modal distributions, we proceed to assume the following regularity conditions for the solution $\mathbf{A}_t^\star$ and the neighborhood of $\mathbf{A}_t^\star$.

Assumption A1 (Regularity)**.** *(Positive definiteness) For any $t \geq 0$ and $\mathbf{A}_t \in \mathcal{A}$, there exists a constant $\lambda_{\min} > 0$ s.t. $\lambda_{\min}\mathbf{I} \preccurlyeq \mathbf{D}_t = \mathbf{I} - 2\mathbf{A}_t$, where $\mathbf{A} \preccurlyeq \mathbf{B}$ means $\mathbf{B} - \mathbf{A}$ is semi positive definite. (Locally strong convexity) For any stable*

local minimum $\mathbf{A}_t^\star$ *with* $\nabla \overrightarrow{\mathbf{L}}_t(\mathbf{A}_t^\star) = \mathbf{0}$*, there is always a neighborhood* $\mathbf{\Theta}$ *s.t.* $\mathbf{A}_t^\star \in \mathbf{\Theta} \subset \mathcal{A}$ *and* $\overrightarrow{\mathbf{L}}_t$ *is strongly convex in* $\mathbf{\Theta}$*, i.e. there exists fixed constants* $M > m > 0$ *s.t. for* $\forall \mathbf{A} \in \mathbf{\Theta}$*,* $m\mathbf{I} \preccurlyeq \frac{\partial^2 \overrightarrow{\mathbf{L}}_t}{\partial \mathbf{A}^2}(\mathbf{A}) \preccurlyeq M\mathbf{I}$*.*

The first part of the above assumption is standard and can be achieved by an appropriate regularization during the training; the second part only assumes the strong convexity for a small neighborhood $\mathbf{\Theta}$ of the optimum $\mathbf{A}_t^\star$. As such, when conditions for Eq.(31) hold, we can apply the induction method to make sure all the subsequent iterates of $\mathbf{A}_t^{(k)}$ stay in the same region $\mathbf{\Theta}$ and converge to the local minimum $\mathbf{A}_t^\star$. For future works, we aim to explore the connection between m and $\lambda_{\min}$.

Next, we lay out three standard assumptions following Chen et al. (2023a) to conduct our analysis. Similar results are studied by Lee et al. (2022); Chen et al. (2022a) with different score assumptions.

Assumption A2 (Lipschitz Score)**.** *The score function* $\nabla \log \overrightarrow{\rho}_t$ *(*$\nabla \log \overrightarrow{\rho}_{t,\mathbf{A}}$*)*[‖] *is L-Lipschitz in both* $\mathbf{x}$ *and* $\mathbf{A}$ *for any* $t \in [0, T]$*. For any* $\mathbf{A}, \mathbf{B} \in \mathcal{A}$ *and any* $\mathbf{x}, \mathbf{y} \in \mathcal{X}$*, we have*

$$\|\nabla \log \overrightarrow{\rho}_{t,\mathbf{A}}(\mathbf{x}) - \nabla \log \overrightarrow{\rho}_{t,\mathbf{A}}(\mathbf{y})\|_2 \le L\|\mathbf{x} - \mathbf{y}\|_2$$
$$\|\nabla \log \overrightarrow{\rho}_{t,\mathbf{A}}(\mathbf{x}) - \nabla \log \overrightarrow{\rho}_{t,\mathbf{B}}(\mathbf{y})\|_2 \le L\|\mathbf{A} - \mathbf{B}\|$$

where $\|\cdot\|_2$ *is the standard* L^2 *norm and* $\|\cdot\|$ *is matrix norm.*

Assumption A3 (Second Moment Bound)**.** *The data distribution has a bounded second moment* $\mathfrak{m}_2^2 := \mathbb{E}_{\rho_{data}}[\|\cdot\|_2^2] < \infty$*.*

Assumption A4 (Score Estimation Error)**.** *For all* $t \in [0, T]$*, and any* $\mathbf{A}_t$*, we have some estimation error .*

$$\mathbb{E}_{\overrightarrow{\rho}_t}[\|s_t - \nabla \log \overrightarrow{\rho}_t\|_2^2] \le \epsilon_{score}^2.$$

We first use the multivariate diffusion to train our score estimators $\{s_t^{(k)}\}_{n=0}^{N-1}$ via the loss function (16) based on the pre-specified $\mathbf{A}_t^{(k)}$. Following Chen et al. (2023a), we can show the generated samples based on $\{s_t^{(k)}\}_{n=0}^{N-1}$ are close in distribution to the ideal samples in Theorem 1. The novelty lies in the extension of single-variate diffusions to multi-variate diffusions.

Next, we use the stochastic approximation theory to prove the convergence of $\mathbf{A}_t^{(k)}$ to a local equilibrium $\mathbf{A}_t^\star$ in Theorem 2. In the end, we adapt Theorem 1 again to show the adaptively generated samples are asymptotically close to the samples based on the optimal $\mathbf{A}_t^\star$ in Theorem 3, which further optimizes the transportation plans through a variational formulation. To facilitate the understanding, we summarize the details as follows

$$\begin{array}{ccccccccc} \text{Sample via } \mathbf{A}_t^{(k)} & & \text{Random Field} & & \text{Mean Field} & & \text{Convergence of } \mathbf{A}_t^{(k)} & & \text{Sample via } \mathbf{A}_t^\star \\ s_t^{(k)} & \xRightarrow[\text{Theorem 1}]{\text{Backward Sampling}} & \nabla \overrightarrow{\mathcal{L}}_t(\mathbf{A}_t^{(k)}; \overleftarrow{\mathbf{x}}_t^{(k+1)}) & \xRightarrow{\text{Eq.(25)}} & \nabla \overrightarrow{\mathbf{L}}_t(\mathbf{A}_t^{(k)}) & \xRightarrow[\text{Theorem 2}]{\text{Convergence}} & \mathbf{A}_t^{(k)} \to \mathbf{A}_t^\star & \xRightarrow[\text{Theorem 3}]{\text{Adaptive Sampling}} & \lim_{k\to\infty} \overleftarrow{\mathbf{x}}_t^{(k+1)}. \end{array}$$

Proof of Sketch

- **Part B.1**: The generated samples (backward trajectories) approximate the ideal samples from the fixed $\mathbf{A}_t^{(k)}$.
- **Part B.2**: We employ the SA theory to show the convergence $\mathbf{A}_t^{(k)}$ to the optimal estimator $\mathbf{A}_t^\star$.
- **Part B.3**: The adaptively generated samples approximate the ideal samples from the optimal $\mathbf{A}_t^\star$ asymptotically.

B.1. Convergence of Approximated Samples with a Fixed $\mathbf{A}_t$

The following result is majorly adapted from Theorem 2.1 of Chen et al. (2023a), where the single-variate diffusions are extended to the general multi-variate diffusions.

Recall that the forward samples $\mathbf{x}_t$ are sampled by (11) given a fixed $\mathbf{A}_t$, we denote the density of $\mathbf{x}_t$ by $\overrightarrow{\rho}_t$ with $\overrightarrow{\rho}_0 = \rho_{\text{data}}$. To facilitate the proof, we introduce an auxiliary variable $\mathbf{y}_t$ simulated from (11) with $\mathbf{y}_0 \sim \mathrm{N}(\mathbf{0}, \mathbf{I})$ such that $\mathbf{y}_t$ is always a Gaussian distribution at time t and $\mathrm{KL}(\rho_{\text{data}}\|\mathrm{N}(\mathbf{0}, \mathbf{I}))$ is well defined (not applicable to deterministic initializations for $\mathbf{y}_0$). We denote the auxiliary distribution of $\mathbf{y}_t$ at time t by $\overrightarrow{\rho}_t^\circ$. For a fixed $T > 0$ and score estimations s_t, let $\overleftarrow{\rho}_t$ be the distribution of the continuous-time interpolation of the discretized backward SDE from $t = T$ to 0 with $\overleftarrow{\rho}_T = \overrightarrow{\rho}_T^\circ$. Then generation quality is measured by the distance between $\overleftarrow{\rho}_0$ and ρ_{data}.

[‖]We abstain from using $\nabla \log \overrightarrow{\rho}_{t,\mathbf{A}_t}$ for the sake of clarity. The smoothness w.r.t. $\mathbf{A}_t$ is only used in Eq.(33). When its use may lead to confusion elsewhere, we employ the $\nabla \log \overrightarrow{\rho}_{t,\mathbf{A}}$ notation.

Theorem 1 (Generation quality)**.** *Assume assumptions A2, A3, and A4 hold. Given a fixed $\mathbf{A}_t$ by assumption A1, the generated data distribution via the EM discretization of Eq.(17) is close to the data distributions ρ_{data} such that*

$$\mathrm{TV}(\overleftarrow{\rho}_0, \rho_{data}) \lesssim \underbrace{\sqrt{\mathrm{KL}(\rho_{data}\|\gamma^d)}\exp(-T)}_{\textit{convergence of forward process}} + \underbrace{(L\sqrt{dh} + \mathfrak{m}_2 h)\sqrt{T}}_{\textit{EM discretization}} + \underbrace{\epsilon_{score}\sqrt{T}}_{\textit{score estimation}},$$

where γ^d is the standard Gaussian distribution.

Proof Following Chen et al. (2023a), we employ the chain rule for KL divergence and obtain:

$$\mathrm{KL}(\rho_{\text{data}}\|\overleftarrow{\rho}_0) \le \mathrm{KL}(\overrightarrow{\rho}_T\|\overleftarrow{\rho}_T) + \mathbb{E}_{\overrightarrow{\rho}_T(\mathbf{x})}[\mathrm{KL}(\overrightarrow{\rho}_{0|T}(\cdot\|\mathbf{x})|\overleftarrow{\rho}_{0|T}(\cdot\|\mathbf{x})],$$

where $\overrightarrow{\rho}_{0|T}$ is the conditional distribution of $\mathbf{x}_0$ given $\mathbf{x}_T$ and likewise for $\overleftarrow{\rho}_{0|T}$. Note that the two terms correspond to the convergence of the forward and reverse process respectively. We proceed to prove that

$$\begin{aligned}
&\text{Part I: Forward process} && \mathrm{KL}(\overrightarrow{\rho}_T\|\overleftarrow{\rho}_T) = \mathrm{KL}(\overrightarrow{\rho}_T\|\overrightarrow{\rho}^\circ_T) \lesssim \mathrm{KL}(\rho_{\text{data}}\|\gamma^d)e^{-T},\\
&\text{Part II: Backward process} && \mathbb{E}_{\overrightarrow{\rho}_T(\mathbf{x})}[\mathrm{KL}(\overrightarrow{\rho}_{0|T}(\cdot|\mathbf{x})\|\overleftarrow{\rho}_{0|T}(\cdot|\mathbf{x})] \lesssim (L^2dh + m_2^2h^2)T + \epsilon^2_{\text{score}}T.
\end{aligned}$$

Part I: By the Fokker-Plank equation, we have

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{KL}(\overrightarrow{\rho}_t\|\overrightarrow{\rho}^\circ_t) = -\frac{1}{2}\beta_t J_{\overrightarrow{\rho}^\circ_t}(\overrightarrow{\rho}_t)$$

where

$$J_{\overrightarrow{\rho}^\circ_t}(\overrightarrow{\rho}_t) = \int \overrightarrow{\rho}_t(x)\left\|\nabla \ln \frac{\overrightarrow{\rho}_t(\mathbf{x})}{\overrightarrow{\rho}^\circ_t(\mathbf{x})}\right\|^2 \mathrm{d}\mathbf{x}$$

is the relative Fisher information of $\overrightarrow{\rho}_t$ with respect to $\overrightarrow{\rho}^\circ_t$. Note that for all $t \ge 0$, $\overrightarrow{\rho}^\circ_t$ is a Gaussian distribution and hence satisfies the log-Sobolev inequality (Vempala & Wibisono, 2022). It follows that

$$\mathrm{KL}(\overrightarrow{\rho}_t\|\overrightarrow{\rho}^\circ_t) \le \frac{1}{2\alpha_t}J_{\overrightarrow{\rho}^\circ_t}(\overrightarrow{\rho}_t),$$

where α_t is the log-Sobolev constant of $\overrightarrow{\rho}^\circ_t$. This implies that

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{KL}(\overrightarrow{\rho}_t\|\overrightarrow{\rho}^\circ_t) \le -\alpha_t\beta_t\mathrm{KL}(\overrightarrow{\rho}_t\|\overrightarrow{\rho}^\circ_t).$$

Applying the Grönwall's inequality yields

$$\mathrm{KL}(\overrightarrow{\rho}_t\|\overrightarrow{\rho}^\circ_t) \le e^{-\int_0^t \alpha_s\beta_s \mathrm{d}s}\mathrm{KL}(\overrightarrow{\rho}_0\|\overrightarrow{\rho}^\circ_0) \le e^{-\alpha\int_0^t \beta_s \mathrm{d}s}\mathrm{KL}(\overrightarrow{\rho}_0\|\overrightarrow{\rho}^\circ_0),$$

where the last inequality is followed by Lemma 1 and α is a lower bound estimate of the LSI constant $\inf_{t\in[0,T]}\alpha_t$.

Then by Pinsker's Inequality, we have

$$\mathrm{TV}(\overrightarrow{\rho}_t, \overrightarrow{\rho}^\circ_t) \le \sqrt{2\mathrm{KL}(\overrightarrow{\rho}_t\|\overrightarrow{\rho}^\circ_t)} \le \sqrt{2e^{-\alpha\int_0^t \beta_s \mathrm{d}s}\mathrm{KL}(\overrightarrow{\rho}_0\|\overrightarrow{\rho}^\circ_0)} \lesssim \sqrt{\mathrm{KL}(\rho_{\text{data}}\|\gamma^d)}\exp(-t).$$

Part II: The proof for the convergence of the reverse process is essentially identical to Theorem 2.1 of Chen et al. (2023a), with the only potential replacements being instances of $\|\mathbf{x}_t - \mathbf{x}_{kh}\|_2$ with $\|\mathbf{D}_{T-t}(\mathbf{x}_t - \mathbf{x}_{kh})\|_2$. However, they are equivalent due to Assumption A1. Therefore, we omit the proof here.

In conclusion, the convergence follows that

$$\mathrm{KL}(\rho_{\text{data}}\|\overleftarrow{\rho}_0) \lesssim \mathrm{KL}(\rho_{\text{data}}\|\gamma^d)e^{-T} + (L^2dh + m_2^2h^2)T + \epsilon_{\text{score}}T.$$

And we obtain the final result using the Pinsker's Inequality. □

Lemma 1 (Lower bound of the log-Sobolev constant)**.** *Under the same assumptions and setups in Theorem 1, we have*

$$\inf_{t\in[0,T]} \alpha_t \geq \min\{1, \lambda_{\min}\} =: \alpha \sim O(1).$$

Proof Consider the auxiliary process for $\mathbf{y}_t$:

- Randomness from the initial: By the mean diffusion in Eq.(12a), the conditional mean diffusion of $\mathbf{y}_t$ at time t, denoted by $\mu_{t,\mathbf{y}}$, follows that $\mu_{t,\mathbf{y}} = \mathbb{D}_t \mu_{0,\mathbf{y}}$, where $\mathbb{D}_t = e^{-\frac{1}{2}[\beta \mathbf{D}]_t}$. Since $\mathbf{y}_0 \sim \mathrm{N}(\mathbf{0}, \mathbf{I})$, we know $\mu_{t,\mathbf{y}} \sim \mathrm{N}(\mathbf{0}, \mathbb{D}_t \mathbb{D}_t^\intercal)$.
- Randomness from Brownian motion: the covariance diffusion induced by Brownian motion follows from $\mathbf{\Sigma}_{t|0}$ in Eq.(12b).

Since $\mathbf{y}_0 \sim \mathrm{N}(\mathbf{0}, \mathbf{I})$ and $\mathbf{y}_t$ is an OU process in Eq.(11), we know that $\mathbf{y}_t$ is always a Gaussian distribution at time $t \geq 0$ with mean $\mathbf{0}$. As such, we know that

$$\overrightarrow{\rho}_t^\circ = \mathrm{N}(\mathbf{0}, \mathbb{D}_t \mathbb{D}_t^\intercal + \mathbf{\Sigma}_{t|0}). \tag{26}$$

It follows that

$$\mathrm{TV}(\overrightarrow{\rho}_t, \overrightarrow{\rho}_t^\circ) \leq \sqrt{2e^{-\int_0^t \alpha_s \beta_s \mathrm{d}s} \mathrm{KL}(\overrightarrow{\rho}_0 \| \overrightarrow{\rho}_0^\circ)}.$$

Now we need to bound the log-Sobolev constant α_t of $\overrightarrow{\rho}_t^\circ$. Let $\mathbf{\Sigma}_t = \mathbb{D}_t \mathbb{D}_t^\intercal + \mathbf{\Sigma}_{t|0}$. Recall that if a distribution p is α-strongly log-concave, then it satisfies the log-Sobolev inequality (LSI) with LSI constant α (Vempala & Wibisono, 2022). So for the Gaussian distribution $\overrightarrow{\rho}_t^\circ$, it suffices to bound the (inverse of) smallest eigenvalue of $\mathbf{\Sigma}_t$. Recall from Eq.(12b) that $\mathbf{\Sigma}_t$ satisfies the ODE

$$\frac{\mathrm{d}\mathbf{\Sigma}_t}{\mathrm{d}t} = -\frac{1}{2}\beta_t(\mathbf{D}_t \mathbf{\Sigma}_t + \mathbf{\Sigma}_t \mathbf{D}_t^\intercal) + \beta_t \mathbf{I}, \quad \mathbf{\Sigma}_0 = \mathbf{I}.$$

Fix a normalized vector $\mathbf{x} \in \mathbb{R}^d$ and denote $u_t = \mathbf{x}^\intercal \mathbf{\Sigma}_t \mathbf{x}$ for $t \in [0, T]$. By the cyclical property of the trace, we have

$$\mathbf{x}^\intercal \mathbf{D}_t \mathbf{\Sigma}_t \mathbf{x} = \mathrm{Tr}(\mathbf{x}^\intercal \mathbf{D}_t \mathbf{\Sigma}_t \mathbf{x}) = \mathrm{Tr}(\mathbf{D}_t \mathbf{\Sigma}_t \mathbf{x}\mathbf{x}^\intercal) \geq \lambda_{\min} \mathrm{Tr}(\mathbf{\Sigma}_t \mathbf{x}\mathbf{x}^\intercal) = \lambda_{\min} u_t.$$

It follows that

$$\frac{\mathrm{d}u_t}{\mathrm{d}t} \leq -\lambda_{\min} \beta_t u_t + \beta_t.$$

Applying the Grönwall's inequality tells us that

$$u_t \leq \frac{1}{\lambda_{\min}}(1 - e^{-\lambda_{\min} \int_0^T \beta_s \mathrm{d}s}) + e^{-\lambda_{\min} \int_0^T \beta_s \mathrm{d}s} \leq \max\{1, 1/\lambda_{\min}\}.$$

Since $\mathbf{x}$ can be any normalized vector, we have that the largest eigenvalue of $\mathbf{\Sigma}_t$ is bounded by $\max\{1, 1/\lambda_{\min}\}$ and hence

$$\inf_{t\in[0,T]} \alpha_t \geq \min\{1, \lambda_{\min}\} =: \alpha \sim O(1),$$

where α_t is the log-Sobolev constant of $\overrightarrow{\rho}_t^\circ$. □

Remark: In our theoretical analysis, we introduced an auxiliary variable $\mathbf{y}_0 \sim \gamma^d$ to make sure $\mathrm{KL}(\rho_{\text{data}} \| \gamma^d)$ is well defined. Moreover, the distribution of $\mathbf{y}_T$ is set to $\overrightarrow{\rho}_T^\circ$ in Eq.(26). However, we emphasize that the introduction of $\mathbf{y}_t$ is only for theoretical analysis and we adopt a simpler prior $\mathrm{N}(\mathbf{0}, \mathbf{\Sigma}_{T|0})$ instead of $\mathrm{N}(\mathbf{0}, \mathbb{D}_T \mathbb{D}_T^\intercal + \mathbf{\Sigma}_{T|0})$ in Eq.(26) for convenience.

B.2. Part II: Stochastic Approximation Convergence

$\mathbf{A}_t^{(k)}$ converges to $\mathbf{A}_t^\star$ by tracking a mean-field ODE with some fluctuations along the trajectory. Before we prove the convergence, we need to show the stability property of the mean-field ODE such that small fluctuations of earlier iterates do not affect the convergence to the equilibrium. To that end, we construct a Lyapunov function $\mathbb{V}_t(\mathbf{A}) = \frac{1}{2}m\|\mathbf{A} - \mathbf{A}_t^\star\|_2^2$ to analyze the local stability condition of the solution. This result shows that when the solution is close to the equilibrium $\mathbf{A}_t^\star \in \mathbf{\Theta} \subset \mathbf{A}$, $\mathbf{A}_t$ will asymptotically track the trajectory of the mean field (24) within $\mathbf{\Theta}$ when the step size $\eta_k \to 0$.

Lemma 2 (Local stabiltity)**.** *Assume the assumptions A1 and A2 hold. For any $\mathbf{A} \in \mathbf{\Theta}$, the solution satisfies a local stability condition such that*

$$\langle \mathbf{A} - \mathbf{A}_t^\star, \nabla \mathbb{V}_t(\mathbf{A})\rangle = \langle \mathbf{A} - \mathbf{A}^\star, \nabla \overrightarrow{\mathbf{L}}_t(\mathbf{A})\rangle \geq m\|\mathbf{A} - \mathbf{A}^\star\|_2^2.$$

Proof By the smoothness assumption A2 and Taylor expansion, for any $\mathbf{A} \in \mathbf{\Theta}$, we have

$$\nabla \overrightarrow{\mathbf{L}}_t(\mathbf{A}) = \nabla \overrightarrow{\mathbf{L}}_t(\mathbf{A}^\star) + \mathbf{Hess}\big[\overrightarrow{\mathbf{L}}_t(\widetilde{\mathbf{A}})\big](\mathbf{A} - \mathbf{A}^\star) = \mathbf{Hess}\big[\overrightarrow{\mathbf{L}}_t(\widetilde{\mathbf{A}})\big](\mathbf{A} - \mathbf{A}^\star), \tag{27}$$

where $\mathbf{Hess}\big[\overrightarrow{\mathbf{L}}_t(\mathbf{A})\big]$ denotes the Hessian of $\overrightarrow{\mathbf{L}}_t$ with $\mathbf{A}$ at time t; $\widetilde{\mathbf{A}}$ is some value between $\mathbf{A}$ and $\mathbf{A}_t^\star$ by the mean-value theorem. Next, we can get

$$\langle \mathbf{A} - \mathbf{A}_t^\star, \nabla \overrightarrow{\mathbf{L}}_t(\mathbf{A})\rangle = \mathbf{Hess}\big[\overrightarrow{\mathbf{L}}_t(\widetilde{\mathbf{A}})\big]\|\mathbf{A} - \mathbf{A}^\star\|_2^2 \geq m\|\mathbf{A} - \mathbf{A}^\star\|_2^2,$$

where the last inequality follows by assumption A1. □

Additionally, we show the random field satisfies a linear growth condition to avoid blow up in tails.

Lemma 3 (Linear growth)**.** *Assume the assumptions A2 and A3 hold. There exists a constant $C > 0$ such that $\forall \mathbf{A}_t^{(k)} \in \mathbf{\Theta}$ at the SA step k and time t, the random field is upper bounded in L^2 such that*

$$\mathbb{E}[\|\nabla \overrightarrow{\mathcal{L}}_t(\mathbf{A}_t^{(k)}, \overleftarrow{\mathbf{x}}_t^{(k+1)})\|_2^2 | \mathcal{F}_k] \leq C(1 + \|\mathbf{A}_t^{(k)} - \mathbf{A}_t^\star\|_2^2) := C(1 + \|\mathbf{G}_t^{(k)}\|_2^2),$$

where the trajectory $\overleftarrow{\mathbf{x}}_t^{(k+1)}$ is simulated by (22); $\mathcal{F}_k$ is a σ-filtration formed by $(\overleftarrow{\mathbf{x}}_t^{(1)}, \mathbf{A}_t^{(1)}, \overleftarrow{\mathbf{x}}_t^{(2)}, \mathbf{A}_t^{(2)}, \cdots, \overleftarrow{\mathbf{x}}_t^{(k)}, \mathbf{A}_t^{(k)})$.

Proof By the unbiasedness of the random field, we have

$$\mathbb{E}[\nabla \overrightarrow{\mathcal{L}}_t(\mathbf{A}_t^{(k)}; \overleftarrow{\mathbf{x}}_t^{(k+1)}) - \nabla \overrightarrow{\mathbf{L}}_t(\mathbf{A}_t^{(k)}) | \mathcal{F}_k] = \mathbf{0}. \tag{28}$$

It follows that

$$\begin{aligned}
\mathbb{E}[\|\nabla \overrightarrow{\mathcal{L}}_t(\mathbf{A}_t^{(k)}; \overleftarrow{\mathbf{x}}_t^{(k+1)})\|_2^2 | \mathcal{F}_k] &= \mathbb{E}[\|\nabla \overrightarrow{\mathcal{L}}_t(\mathbf{A}_t^{(k)}; \overleftarrow{\mathbf{x}}_t^{(k+1)}) - \nabla \overrightarrow{\mathbf{L}}_t(\mathbf{A}_t^{(k)}) + \nabla \overrightarrow{\mathbf{L}}_t(\mathbf{A}_t^{(k)}))\|_2^2 | \mathcal{F}_k] \\
&= \mathbb{E}[\|\nabla \overrightarrow{\mathcal{L}}_t(\mathbf{A}_t^{(k)}; \overleftarrow{\mathbf{x}}_t^{(k+1)}) - \nabla \overrightarrow{\mathbf{L}}_t(\mathbf{A}_t^{(k)})\|_2^2 | \mathcal{F}_k] + \|\nabla \overrightarrow{\mathbf{L}}_t(\mathbf{A}_t^{(k)})\|_2^2 \\
&\leq \sup \mathbb{E}[\|\nabla \overrightarrow{\mathcal{L}}_t(\mathbf{A}_t^{(k)}; \overleftarrow{\mathbf{x}}_t^{(k+1)}) - \nabla \overrightarrow{\mathbf{L}}_t(\mathbf{A}_t^{(k)})\|_2^2 | \mathcal{F}_k] + M^2\|\mathbf{A}_t^{(k)} - \mathbf{A}_t^\star\|_2^2,
\end{aligned} \tag{29}$$

where the last inequality follows by assumption A1 and Eq.(27).

By assumption A2 and A3 and the process (17), we know $\sup \mathbb{E}[\|\nabla \overrightarrow{\mathcal{L}}_t(\mathbf{A}_t^{(k)}; \overleftarrow{\mathbf{x}}_t^{(k+1)}) - \nabla \overrightarrow{\mathbf{L}}_t(\mathbf{A}_t^{(k)})\|_2^2 | \mathcal{F}_k] < \infty$. Denote by $C := \max\{\sup \mathbb{E}[\|\nabla \overrightarrow{\mathcal{L}}_t(\mathbf{A}_t^{(k)}; \overleftarrow{\mathbf{x}}_t^{(k+1)}) - \nabla \overrightarrow{\mathbf{L}}_t(\mathbf{A}_t^{(k)})\|_2^2 | \mathcal{F}_k], M^2\}$, we can conclude that

$$\mathbb{E}[\|\nabla \overrightarrow{\mathcal{L}}_t(\mathbf{A}_t^{(k)}; \overleftarrow{\mathbf{x}}_t^{(k+1)})\|_2^2 | \mathcal{F}_k] \leq C(1 + \|\mathbf{A}_t^{(k)} - \mathbf{A}_t^\star\|_2^2).$$

□

Next, we make standard assumptions on the step size following Benveniste et al. (1990) (page 245).

Assumption A5 (Step size)**.** *The step size $\{\eta_k\}_{k\in\mathrm{N}}$ is a positive and decreasing sequence*

$$\eta_k \to 0, \quad \sum_{k=1}^{\infty} \eta_k = +\infty, \quad \lim_{k\to\infty} \inf \left(2m\frac{\eta_k}{\eta_{k+1}} + \frac{\eta_{k+1} - \eta_k}{\eta_{k+1}^2}\right) := \kappa > 0.$$

A standard choice is to set $\eta_k := \frac{A}{k^\alpha + B}$ for some $\alpha \in (\frac{1}{2}, 1]$ and some suitable constants $A > 0$ and $B > 0$.

Theorem 2 (Convergence in L^2)**.** *Assume assumptions A1, A2, A3, A4, and A5 hold. The variational score $\mathbf{A}_t^{(k)}$ in algorithm 2 converges to a local minimizer $\mathbf{A}_t^\star$. In other words, given a large enough $k \geq k_0$, where $\eta_{k_0} \leq \frac{1}{2}$, we have*

$$\mathbb{E}[\|\mathbf{A}_t^{(k)} - \mathbf{A}_t^\star\|_2^2] \leq 2\eta_k,$$

where the expectation is taken w.r.t samples from $\overleftarrow{\rho}_t^{(k)}$.

Proof To show $\mathbf{A}_t^{(k)}$ converges to $\mathbf{A}_t^\star$, we first denote $\mathbf{G}_t^{(k)} = \mathbf{A}_t^{(k)} - \mathbf{A}_t^\star$. Subtracting $\mathbf{A}^\star$ on both sides of Eq.(23):

$$\mathbf{G}_t^{(k+1)} = \mathbf{G}_t^{(k)} - \eta_{k+1}\nabla\overrightarrow{\mathcal{L}}_t(\mathbf{A}_t^{(k)}; \overleftarrow{\mathbf{x}}_t^{(k+1)}).$$

By the unbiasedness of the random field, we have

$$\mathbb{E}[\nabla\overrightarrow{\mathcal{L}}_t(\mathbf{A}_t^{(k)}; \overleftarrow{\mathbf{x}}_t^{(k+1)}) - \nabla\overrightarrow{\mathbf{L}}_t(\mathbf{A}_t^{(k)})|\mathcal{F}_k] = \mathbf{0}. \tag{30}$$

Taking the expectation in L^2, we have

$$\begin{aligned}\mathbb{E}[\|\mathbf{G}_t^{(k+1)}\|_2^2|\mathcal{F}_k] &= \|\mathbf{G}_t^{(k+1)}\|_2^2 - 2\eta_{k+1}\mathbb{E}\big[\langle\mathbf{G}_t^{(k)}, \nabla\overrightarrow{\mathcal{L}}_t(\mathbf{A}^{(k)}; \overleftarrow{\mathbf{x}}_t^{(k+1)})\rangle\big] + \eta_{k+1}^2\mathbb{E}\big[\|\nabla\overrightarrow{\mathcal{L}}_t(\mathbf{A}_t^{(k)}; \overleftarrow{\mathbf{x}}_t^{(k+1)})\|_2^2|\mathcal{F}_k\big] \\ &= \|\mathbf{G}_t^{(k+1)}\|_2^2 - 2\eta_{k+1}\langle\mathbf{G}_t^{(k)}, \nabla\overrightarrow{\mathbf{L}}_t(\mathbf{A}_t^{(k)})\rangle + \eta_{k+1}^2\mathbb{E}\big[\|\nabla\overrightarrow{\mathcal{L}}_t(\mathbf{A}_t^{(k)}; \overleftarrow{\mathbf{x}}_t^{(k+1)})\|_2^2|\mathcal{F}_k\big],\end{aligned}$$

where the second equality is followed by the unbiasedness property in Eq.(30).

Applying the stepsize assumption A5, we have

$$\eta_{k+1} - \eta_k + 2m\eta_k\eta_{k+1} \geq C\eta_{k+1}^2.$$

Then for $\eta_k \leq \frac{1}{2}$, we have

$$2(\eta_{k+1} - \eta_k + \eta_k\eta_{k+1}(2m - \eta_{k+1}C)) \geq C\eta_{k+1}^2.$$

Rewrite the above equation as follows

$$2\eta_{k+1} \geq (1 - 2\eta_{k+1}m + C\eta_{k+1}^2)(2\eta_k) + C\eta_{k+1}^2.$$

By the induction method, we have

- Given some large enough $k \geq k_0$, where $\eta_{k_0} \leq \frac{1}{2}$, $\mathbf{A}_t^{(k)}$ is in some subset $\mathbf{\Theta}$ [**] of $\mathcal{A}$ that follows

$$\mathbb{E}[\|\mathbf{G}_t^k\|_2^2] \leq 2\eta_k. \tag{31}$$

- Applying Eq.(B.2) and Eq.(B.2), respectively, we have

$$\begin{aligned}\mathbb{E}[\|\mathbf{G}_t^{(k+1)}\|_2^2|\mathcal{F}_k] &\leq (1 - 2\eta_{k+1}m)\mathbb{E}[\|\mathbf{G}_t^{(k)}\|_2^2] + \eta_{k+1}^2\mathbb{E}\big[\|\nabla\overrightarrow{\mathcal{L}}_t(\mathbf{A}_t^{(k)}; \overleftarrow{\mathbf{x}}_t^{(k+1)})\|_2^2|\mathcal{F}_k\big] \\ &\leq (1 - 2\eta_{k+1}m + C\eta_{k+1}^2)\mathbb{E}[\|\mathbf{G}_t^{(k)}\|_2^2] + C\eta_{k+1}^2, \\ &\leq (1 - 2\eta_{k+1}m + C\eta_{k+1}^2)(2\eta_k) + C\eta_{k+1}^2 \\ &\leq 2\eta_{k+1},\end{aligned} \tag{32}$$

  where the first inequality is held by the stability property in Lemma 2 and the last inequality is followed by the growth property in Lemma 3.

Since $\mathbf{A}_t^\star, \mathbf{A}_t^{(k)} \in \mathbf{\Theta}$, Eq.(32) implies that $\mathbf{A}_t^{(k+1)} \in \mathbf{\Theta}$, which concludes the proof. □

[**] By assumption A1, such $\mathbf{\Theta} \subset \mathcal{A}$ exists, otherwise it implies that the mean field function is a constant and conclusion holds as well.

B.3. Part III: Convergence of Adaptive Samples based on The Optimal $\mathbf{A}^\star$

We have evaluated the sample quality in Theorem 1 based on a fixed $\mathbf{A}_t$, which, however, may not be efficient in terms of transportation plans. To evaluate the sample quality in terms of the limiting optimal $\mathbf{A}^\star$, we provide the result as follows:

Theorem 3. *Given assumptions A1-A5, the generated sample distribution at stage k is ϵ-close to the exact sample distribution $\overrightarrow{\rho}^\star_T$ based on the equilibrium $\mathbf{A}^\star_t$ such that*

$$\mathrm{TV}(\overleftarrow{\rho}^\star_0, \rho_{data}) \lesssim \sqrt{\mathrm{KL}(\rho_{data}\|\gamma^d)}\exp(-T) + (L\sqrt{dh} + L\mathfrak{m}_2 h)\sqrt{T} + (\epsilon_{score} + \sqrt{\eta_k})\sqrt{T}.$$

Proof By assumption A4, for any $\mathbf{A}_t^{(k)} \in \mathcal{A}$, we have

$$\mathbb{E}_{\overrightarrow{\rho}_t^{(k)}}[\|s_t^{(k)} - \nabla \log \overrightarrow{\rho}_t^{(k)}\|_2^2] \leq \epsilon_{\text{score}}^2.$$

Combining Theorem 2 and the smoothness assumption A2 of the score function $\nabla \log \overrightarrow{\rho}_t^{(k)}$ w.r.t $\mathbf{A}_t^{(k)}$, we have

$$\mathbb{E}_{\overrightarrow{\rho}_t^{(k)}}[\|\nabla \log \overrightarrow{\rho}_t^{(k)} - \nabla \log \overrightarrow{\rho}_t^\star\|_2^2] \lesssim \eta_k. \tag{33}$$

It follows that the score function $s_t^{(k)}$ is also close to the optimal $\nabla \log \overrightarrow{\rho}_t^\star$ in the sense that

$$\begin{aligned}
&\mathbb{E}_{\overrightarrow{\rho}_t^{(k)}}[\|s_t^{(k)} - \nabla \log \overrightarrow{\rho}_t^\star\|_2^2] \\
\lesssim\ &\mathbb{E}_{\overrightarrow{\rho}_t^{(k)}}[\underbrace{\|s_t^{(k)} - \nabla \log \overrightarrow{\rho}_t\|_2^2}_{\text{by Assumption A4}}] + \mathbb{E}_{\overrightarrow{\rho}_t^{(k)}}[\underbrace{\|\nabla \log \overrightarrow{\rho}_t^{(k)} - \nabla \log \overrightarrow{\rho}_t^\star\|_2^2}_{\text{by Eq.(33)}}] \\
\lesssim\ &\epsilon_{\text{score}}^2 + \eta_k.
\end{aligned} \tag{34}$$

Applying Theorem 1 with the adaptive score error in Eq.(34) to replace $\epsilon_{\text{score}}^2$ concludes the proof. □

Remark: The convergence of samples based on the adaptive algorithms is slightly weaker than the standard one due to the adaptive update, but this is necessary because $\mathbf{A}_t^\star$ is more transport efficient than a vanilla $\mathbf{A}_t$.

C. Variational Gap

Recall that the optimal forward SDE in the forward-backward SDEs (5) follows that

$$\mathrm{d}\overrightarrow{\mathbf{x}}_t = \left[\boldsymbol{f}_t(\overrightarrow{\mathbf{x}}_t) + \beta_t \nabla \log \overrightarrow{\psi}_t(\overrightarrow{\mathbf{x}}_t)\right]\mathrm{d}t + \sqrt{\beta_t}\mathrm{d}\overrightarrow{\mathbf{w}}_t. \tag{35}$$

The optimal variational forward SDE follows that

$$\mathrm{d}\overrightarrow{\mathbf{x}}_t = \left[\boldsymbol{f}_t(\overrightarrow{\mathbf{x}}_t) + \beta_t \mathbf{A}_t^\star \overrightarrow{\mathbf{x}}_t\right]\mathrm{d}t + \sqrt{\beta_t}\mathrm{d}\overrightarrow{\mathbf{w}}_t. \tag{36}$$

The variational forward SDE at the k-th step follows that

$$\mathrm{d}\overrightarrow{\mathbf{x}}_t = \left[\boldsymbol{f}_t(\overrightarrow{\mathbf{x}}_t) + \beta_t \mathbf{A}_t^{(k)} \overrightarrow{\mathbf{x}}_t\right]\mathrm{d}t + \sqrt{\beta_t}\mathrm{d}\overrightarrow{\mathbf{w}}_t. \tag{37}$$

Since we only employ a linear approximation of the forward score function, our transport is only sub-optimal. To assess the extent of this discrepancy, we leverage the Girsanov theorem to study the variational gap.

We denote the law of the processes by $\mathrm{L}(\cdot)$ in Eq.(35), $\mathrm{L}^\star(\cdot)$ in Eq.(36) and $\mathrm{L}^{(k)}(\cdot)$ in Eq.(37), respectively.

Theorem 4. *Assume assumptions A2 and A3 hold. Assume $\boldsymbol{f}_t$ and $\nabla \log \overrightarrow{\psi}_t$ are Lipschitz smooth and satisfy the linear growth condition. Assume the Novikov's condition holds for $\forall \mathbf{A}_t \in \mathcal{A}$, where $t \in [0, T]$:*

$$\mathbb{E}\left[\exp\left(\frac{1}{2}\int_0^T \|\beta_t \mathbf{A}_t \overrightarrow{\mathbf{x}}_t - \beta_t \nabla \log \overrightarrow{\psi}_t(\overrightarrow{\mathbf{x}}_t)\|_2^2 \mathrm{d}t\right)\right] < \infty.$$

The variational gap (VG) via the linear parametrization is upper bounded by

$$\mathrm{KL}(\mathrm{L}\|\mathrm{L}^\star) = \frac{1}{2}\int_0^T \mathbb{E}_{\overrightarrow{\rho}_t}\left[\beta_t\|\mathbf{A}_t^\star\overrightarrow{\mathbf{x}}_t - \nabla\log\overrightarrow{\psi}_t(\overrightarrow{\mathbf{x}}_t)\|_2^2\mathrm{d}t\right]$$
$$\mathrm{KL}(\mathrm{L}\|\mathrm{L}^{(k)}) \lesssim \eta_k + \mathrm{KL}(\mathrm{L}\|\mathrm{L}^\star).$$

Proof

By Girsanov's formula (Liptser & Shiryaev, 2001), the Radon–Nikodym derivative of $\mathrm{L}(\cdot)$ w.r.t. $\mathrm{L}^\star(\cdot)$ follows that

$$\frac{\mathrm{dL}}{\mathrm{dL}^\star}(\overrightarrow{\mathbf{x}}) = \exp\left(\int_0^T \sqrt{\beta_t}\Big(\mathbf{A}_t^\star\overrightarrow{\mathbf{x}}_t - \nabla\log\overrightarrow{\psi}_t(\overrightarrow{\mathbf{x}}_t)\Big)\mathrm{d}\mathbf{w}_t - \frac{1}{2}\int_0^T \beta_t\|\mathbf{A}_t^\star\overrightarrow{\mathbf{x}}_t - \nabla\log\overrightarrow{\psi}_t(\overrightarrow{\mathbf{x}}_t)\|_2^2\mathrm{d}t\right),$$

where $\mathbf{w}_t$ is the Brownian motion under the Wiener measure. Consider a change of measure (Øksendal, 2003; Chewi, 2023)

$$\mathbf{w}_t = \widetilde{\mathbf{w}}_t - \mathrm{d}\big[\mathbf{w},\mathbf{M}\big]_t, \quad \mathrm{d}\mathbf{M}_t = \big\langle\sqrt{\beta_t}\big(\mathbf{A}_t^\star\overrightarrow{\mathbf{x}}_t - \nabla\log\overrightarrow{\psi}_t(\overrightarrow{\mathbf{x}}_t)\big), \mathrm{d}\mathbf{w}_t\big\rangle,$$

where $\widetilde{\mathbf{w}}_t$ is a L-standard Brownian motion and satisfies martingale property under the L measure.

Now the variational gap is upper bounded by

$$\begin{aligned}
\mathrm{KL}(\mathrm{L}(\cdot)\|\mathrm{L}^\star(\cdot)) &= -\mathbb{E}_{\mathrm{L}(\cdot)}\left[\log\frac{\mathrm{dL}(\cdot)}{\mathrm{dL}^\star(\cdot)}\right] \\
&= \mathbb{E}_{\mathrm{L}(\cdot)}\left[\int_0^T \sqrt{\beta_t}\Big(\mathbf{A}_t^\star\overrightarrow{\mathbf{x}}_t - \nabla\log\overrightarrow{\psi}_t(\overrightarrow{\mathbf{x}}_t)\Big)\mathrm{d}\widetilde{\mathbf{w}}_t + \frac{1}{2}\int_0^T \beta_t\|\mathbf{A}_t^\star\overrightarrow{\mathbf{x}}_t - \nabla\log\overrightarrow{\psi}_t(\overrightarrow{\mathbf{x}}_t)\|_2^2\mathrm{d}t\right] \\
&= \frac{1}{2}\mathbb{E}_{\mathrm{L}(\cdot)}\left[\int_0^T \beta_t\|\mathbf{A}_t^\star\overrightarrow{\mathbf{x}}_t - \nabla\log\overrightarrow{\psi}_t(\overrightarrow{\mathbf{x}}_t)\|_2^2\mathrm{d}t\right] \\
&= \frac{1}{2}\int_0^T \mathbb{E}\left[\beta_t\|\mathbf{A}_t^\star\overrightarrow{\mathbf{x}}_t - \nabla\log\overrightarrow{\psi}_t(\overrightarrow{\mathbf{x}}_t)\|_2^2\right]\mathrm{d}t.
\end{aligned}$$

Similarly, applying $(a+b)^2 \le 2a^2 + 2b^2$, we have

$$\begin{aligned}
\mathrm{KL}(\mathrm{L}(\cdot)\|\mathrm{L}^{(k)}(\cdot)) &\le \frac{3}{2}\int_0^T \mathbb{E}\left[\beta_t\big(\underbrace{\|\mathbf{A}_t^{(k)}\overrightarrow{\mathbf{x}}_t - \mathbf{A}_t^\star\overrightarrow{\mathbf{x}}_t\|_2^2}_{\text{convergence of SA}} + \underbrace{\|\mathbf{A}_t^\star\overrightarrow{\mathbf{x}}_t - \nabla\log\overrightarrow{\psi}_t(\overrightarrow{\mathbf{x}}_t)\|_2^2}_{\text{variational gap based on } \mathbf{A}_t^\star}\big)\right]\mathrm{d}t \\
&\lesssim \eta_k + \int_0^T \mathbb{E}\left[\beta_t\|\mathbf{A}_t^\star\overrightarrow{\mathbf{x}}_t - \nabla\log\overrightarrow{\psi}_t(\overrightarrow{\mathbf{x}}_t)\|_2^2\right]\mathrm{d}t.
\end{aligned}$$

□

D. Experimental Details

D.1. Parametrization of the Variational Score

For the general transport, there is no closed-form update and we adopt an SVD decomposition with time embeddings to learn the linear dynamics in Figure 6. The number of parameters is reduced by thousands of times, which have greatly reduced the training variance (Grathwohl et al., 2019).

D.2. Synthetic Data

D.2.1. Checkerboard Data

The generation of the checkerboard data is presented in Figure. 7. The probability path is presented in Figure. 8. The conclusion is similar to the spiral data.

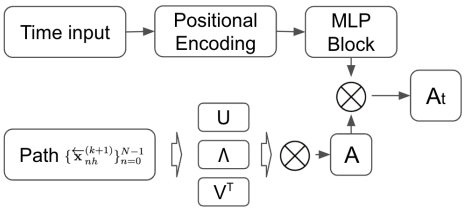


Figure 6. Architecture of the linear module. Both U and V are orthogonal matrices and Λ denotes the singular values.

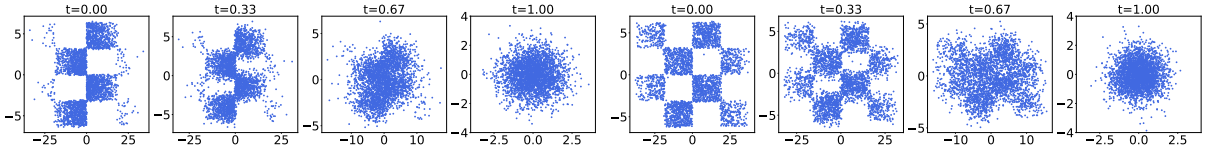
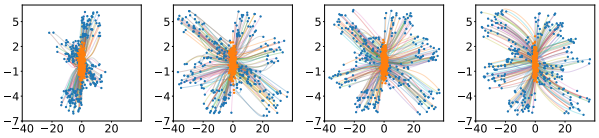


Figure 7. Variational Schrödinger diffusion models (VSDMs, right) v.s. SGMs (left) with the same hyperparameters ($\beta_{\max} = 10$).

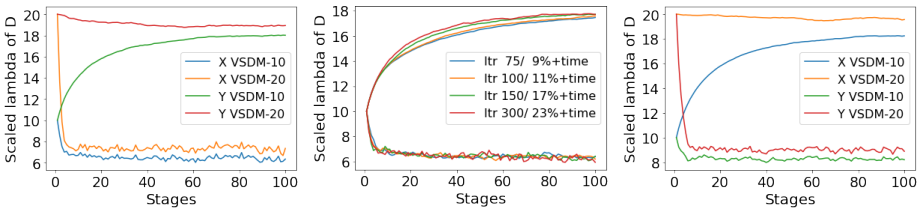


(a) SGM-10 (b) SGM-20 (c) SGM-30 (d) VSDM-10

Figure 8. Probability flow ODE via VSDMs and SGMs. SGM with $\beta_{\max} = 10$ is denoted by SGM-10 for convenience.

D.2.2. Convergence and Computational Time

Convergence Study Under the same setup, VSDM-10 adaptively learns $\mathbf{A}_t$ (and $\mathbf{D}_t$) on the fly and adapts through the pathological geometry via optimal transport. For the spiral-8Y data, the Y-axis of the singular values of $\mathbf{D}_t$ (scaled by $\beta_{\max}$) converges from 10 to around 19 as shown in Figure 9. The singular value of the X-axis quickly converges from 10 to a conservative scale of 7. We also tried VSDM-20 and found that both the Y-axis and X-axis converge to similar scales, which justifies the stability.



(a) Spiral-8Y (b) Convergence v.s. time (c) Checkboard-6X

Figure 9. Optimization of $\mathbf{\Lambda}$ of $\mathbf{D}$ scaled by $\beta_{\max}$ (scaled lambda of $\mathbf{D}$) of VSDM-10 and VSDM-20.

Computational Time We tried different budgets to train the variational scores and observed in Figure 9(b) that 300 iterations yield the fastest convergence among the 4 choices but also lead to 23% extra time compared to SGM. Reducing the number of iterations impacts convergence minimally due to the linearity of the variational scores and significantly reduces the training time.

D.2.3. Evaluation of The Straightness

Straighter trajectories lead to a smaller number of functional evaluations (NFEs). In section D.2.3, we compare VSDM-20 with SGM-20 with NFE=6 and NFE=8 using the same computational budget and observe in Figure 10 and 11 the superiority of the VSDM model in generating more details.

To evaluate the straightness of the probability flow ODE, similar in spirit to Pooladian et al. (2023), we define our straightness

metric by approximating the second derivative of the probability flow (18) as follows

$$\mathrm{S}(i) = \int_0^T \mathbb{E}_{\overleftarrow{\mathbf{x}}_t \sim \overleftarrow{\rho}_t} \left[\left| \frac{\mathrm{d}^2 \overleftarrow{\mathbf{x}}_t(i)}{\mathrm{d}t^2} \right| \right] \mathrm{d}t, \tag{38}$$

where $i \in \{1, 2\}$, $\overleftarrow{\mathbf{x}}_t(1)$ and $\overleftarrow{\mathbf{x}}_t(2)$ denote the X-axis and Y-axis, respectively. $S \geq 0$ and $S = 0$ only when the transport is a straight path.

We report the straightness in Table 4 and find the improvement of VSDM-10 over SGM-20 and SGM-30 is around 40%. We also tried VSDM-20 on both datasets and found a significant improvement over the baseline SGM methods. However, despite the consistent convergence in Figure 9, we found VSDM-20 still performs slightly worse than VSDM-10, which implies the potential to tune $\beta_{\max}$ to further enhance the performance.

Table 4. STRAIGHTNESS METRIC DEFINED IN EQ.(38) VIA SGMS AND VSDM WITH DIFFERENT $\beta_{\max}$'S. SGM WITH $\beta_{\max} = 10$ (SGM-10) FAILS TO GENERATE DATA OF ANISOTROPIC SHAPES AND IS NOT REPORTED.

STRAIGHTNESS (X / Y)	SPIRAL-8Y	CHECKERBOARD-6X
SGM-20	8.3 / 49.3	53.5 / 11.0
SGM-30	9.4 / 57.3	64.6 / 13.1
VSDM-20	**6.3** / 45.6	49.4 / **7.4**
VSDM-10	**5.5 / 38.7**	**43.9 / 6.5**

D.2.4. A SMALLER NUMBER OF FUNCTION EVALUATIONS

We also compare our VSDM-20 with SGM-20 based on a small number of function evaluations (NFE). We use probability flow to conduct the experiments and choose a uniform time grid for convenience. We find that both models cannot fully generate the desired data with NFE=6 in Figure 10 and VSDM appears to recover more details, especially on the top and bottom of the spiral. For the checkboard data, both models work nicely under the same setting and we cannot see a visual difference. With NFE=8 in Figure 11, we observe that our VSDM-20 works remarkably well on both datasets and is slightly superior to SGM-20 in generating the corner details.

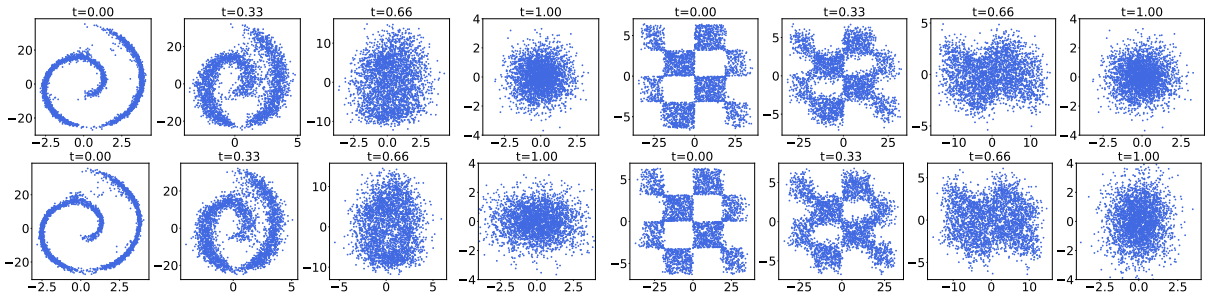


Figure 10. Variational Schrödinger diffusion models (bottom) v.s. SGMs (top) with the same hyperparameters ($\beta_{\max} = 20$) and six function evaluations (NFE=6). Both models are generated by probability flow ODE.

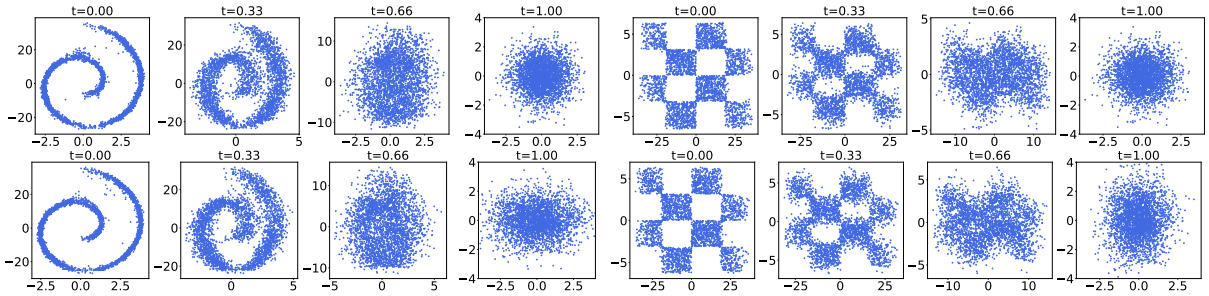


Figure 11. Variational Schrödinger diffusion models (bottom) v.s. SGMs (top) with the same hyperparameters ($\beta_{\max} = 20$) and eight function evaluations (NFE=8). Both models are generated by probability flow ODE.

D.3. Multivariate Probabilistic Forecasting

Data. We use publicly available datasets. Exchange rate has 6071 8-dimensional measurements and a daily frequency. The goal is to predict the value over the next 30 days. Solar is an hourly 137-dimensional dataset with 7009 values. Electricity is also hourly, with 370 dimensions and 5833 measurements. For both, we predict the values over the next day.

Training. We adopt the encoder-decoder architecture as described in the main text, and change the decoder to either our generative model or one of the competitors. The encoder is an LSTM with 2 layers and a hidden dimension size 64. We train the model for 200 epochs, where each epoch takes 50 model updates. In case of our model we also alternate between two training directions at a predefined rate. The neural network parameterizing the backward direction has the same hyperparameters as in (Rasul et al., 2021), that is, it has 8 layers, 8 channels, and a hidden dimension of 64. The DDPM baseline uses a standard setting for the linear beta-scheduler: $\beta_{\text{min}} = 0.0001$, $\beta_{\text{max}} = 0.1$ and 150 steps.

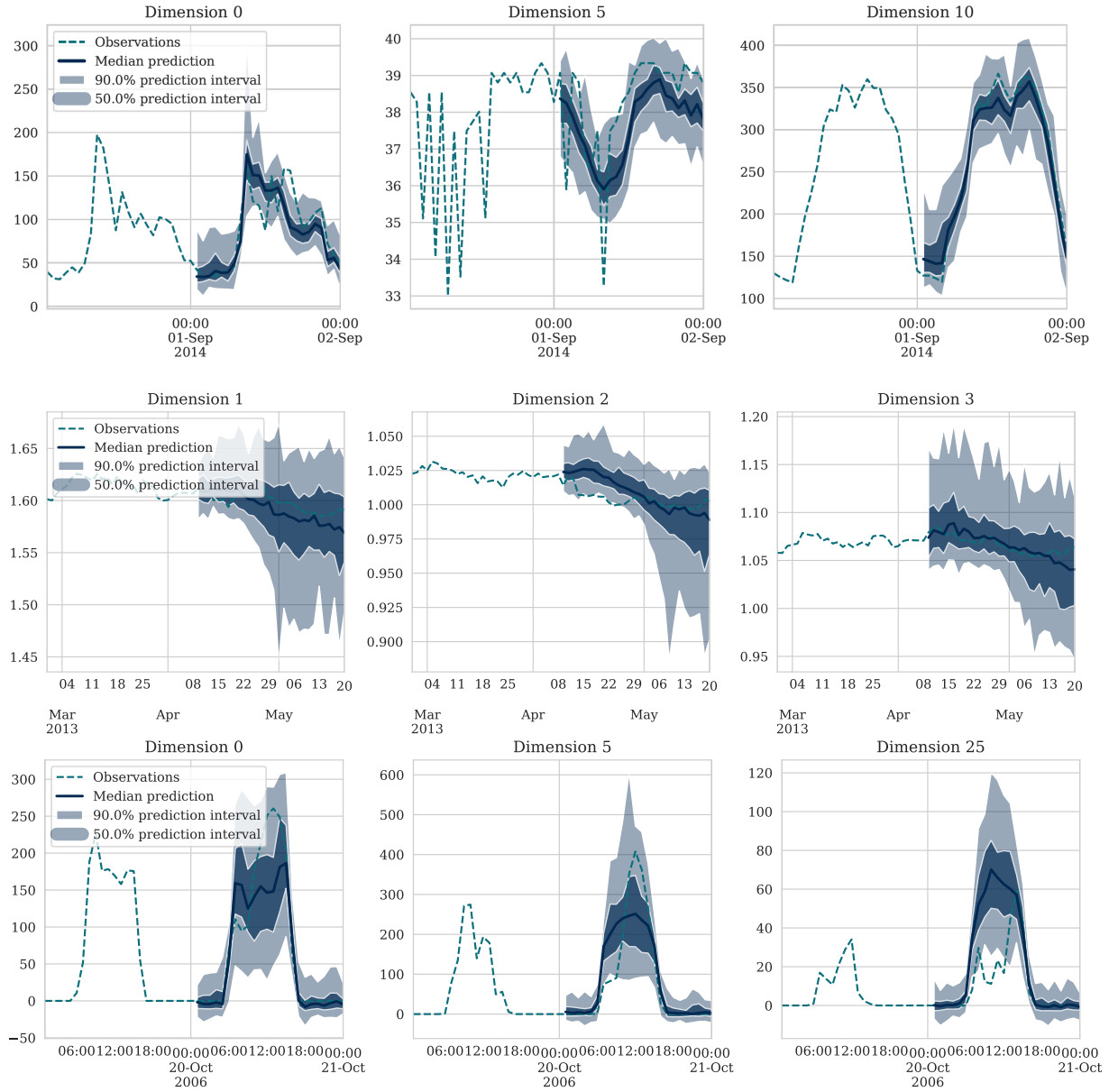


Figure 12. Example forecasts for Electricity (top), Exchange (middle), and Solar (bottom) datasets using our VSDM model. We show 3 out of 370, 8, and 137 dimensions, respectively.