On the Equivalence of Consistency-Type Models: Consistency Models, Consistent Diffusion Models, and Fokker-Planck Regularization

Chieh-Hsin Lai [1] **Yuhta Takida** [1] **Toshimitsu Uesaka** [1] **Naoki Murata** [1] **Yuki Mitsufuji** [1 2] **Stefano Ermon** [3]

Abstract

The emergence of various notions of "consistency" in diffusion models has garnered considerable attention and helped achieve improved sample quality, likelihood estimation, and accelerated sampling. Although similar concepts have been proposed in the literature, the precise relationships among them remain unclear. In this study, we establish theoretical connections between three recent "consistency" notions designed to enhance diffusion models for distinct objectives. Our insights offer the potential for a more comprehensive and encompassing framework for consistency-type models.

1. Introduction

Score-based generative models (Song & Ermon, 2019; Song et al., 2020b;a; Boffi & Vanden-Eijnden, 2022), commonly referred to as diffusion models (Sohl-Dickstein et al., 2015; Ho et al., 2020), have significantly advanced photorealistic image generation (Saharia et al., 2022; Rombach et al., 2022; Kim et al., 2022) and found applications in diverse domains such as media editing and restoration (Meng et al., 2021; Kawar et al., 2022; Saito et al., 2022; Murata et al., 2023).

Underlying score-based generative model is a (stochastic) differential equation that describes the process of transforming data to noise and vice-versa, which is approximated using a neural (score) network learned from data. Because of the mathematical structure afforded by the underlying differential equation, this neural network needs to satisfy certain consistency properties. Various such notions of consistency have been recently introduced and shown to enhance sample quality (Daras et al., 2023), accelerate sampling speed (Song et al., 2023), and improve likelihood estimation (Lai et al., 2022). We introduce the term *consistency-type model* to encompass and unify these various concepts. It refers to a (diffusion) model that is explicitly designed to align with the underlying trajectory defined by an ordinary differential equation (ODE), stochastic differential equation (SDE), or partial differential equation (PDE). In this study, we aim to provide a theoretical investigation into the relationships between these three consistency-type models. Under certain mild assumptions, we will rigorously establish the equivalence of these independently developed concepts.

[1]Sony AI, Tokyo, Japan [2]Sony Group Corporation, Tokyo, Japan [3]Department of Computer Science, Stanford University, Stanford, CA, USA. Correspondence to: Chieh-Hsin Lai <Chieh-hsin.lai@sony.com>.

Accepted to ICML 2023 Workshop on Structured Probabilistic Inference & Generative Modeling, Honolulu, Hawaii, USA. Copyright 2023 by the author(s).

2. Background

Song et al. (2020b) introduced a stochastic differential equation (SDE) framework that unifies the concepts of denoising score matching (Song & Ermon, 2019) and diffusion models (Sohl-Dickstein et al., 2015; Ho et al., 2020) in continuous time. Especially[1], the process $\{\boldsymbol{x}(t)\}_{t\in[0,T]}$ of adding Gaussian noise

$$\boldsymbol{x}(t) \sim q_t \quad \text{where} \quad q_t(\boldsymbol{x}) = q_{\text{data}}(\boldsymbol{x}) * \mathcal{N}(\boldsymbol{x}; \boldsymbol{0}, \sigma^2(t)\boldsymbol{I})$$

is driven by the following forward SDE

$$d\boldsymbol{x}(t) = g(t)d\boldsymbol{w}_t. \tag{1}$$

Here, we define $q_0 = q_{\text{data}}$, $g(t) := \sqrt{\frac{d\sigma^2(t)}{dt}}$, $*$ as the convolution operator, and $\boldsymbol{w}_t$ as the standard Wiener process. Eq. (1) inherently corresponds to a reverse time SDE from T to 0 under moderate conditions (Anderson, 1982)

$$d\boldsymbol{x}(t) = -g^2(t)\nabla_{\boldsymbol{x}} \log q_t(\boldsymbol{x}(t))dt + g(t)d\bar{\boldsymbol{w}}_t, \tag{2}$$

where $\bar{\boldsymbol{w}}_t$ is a standard Wiener process in reverse time, and $q_t(\boldsymbol{x})$ denotes the ground truth marginal density of $\boldsymbol{x}(t)$ following Eq. (1).

The stochastic process in Eq. (2) automatically associates with a deterministic process, known as the probability flow (PF) ODE. This PF ODE governs the evolution of samples without any diffusion term and guarantees that the trajectories of the samples maintain identical marginal probability

[1]Within this study, our primary emphasis is placed on the variance exploding (VE) SDE (Song et al., 2020b). VE SDE entails a process devoid of drift in the forward SDE formulation. Our discussion can be extended to a broader range of forward SDEs.

densities as the forward SDE (Eq. (1)). The PF ODE is expressed as follows:

$$\frac{d\boldsymbol{x}}{dt}(t) = -\frac{1}{2}g^2(t)\nabla_{\boldsymbol{x}} \log q_t(\boldsymbol{x}(t)). \tag{3}$$

Since $\nabla_{\boldsymbol{x}} \log q_t(\boldsymbol{x}(t))$ is typically unattainable in Eqs. (2) and (3), the denoising score matching (DSM) loss (Vincent, 2011; Song et al., 2020b) is commonly employed to approximate $\nabla_{\boldsymbol{x}} \log q_t(\boldsymbol{x})$ by using a time-conditional neural network $\boldsymbol{s}_{\boldsymbol{\theta}} = \boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}, t)$ over the time interval $[t_0, T]$, where $t_0 \geq 0$ is chosen to be sufficiently small in practice.

By substituting $\nabla_{\boldsymbol{x}} \log q_t(\boldsymbol{x})$ with the learned $\boldsymbol{s}_{\boldsymbol{\theta}}$ in the reverse time SDE described in Eq. (2), and in the PF ODE given by Eq. (3), we obtain parametric counterparts of the reverse time SDE for a stochastic process and the PF PDE for a deterministic process, respectively. Consequently, we have the choice to generate samples by numerically solving either the parametric SDE or the parametric PF PDE in reverse, starting from an initial sample drawn from a predefined prior.

3. Consistency-type models

In this section, we provide an overview of recent literature that incorporates notions of "consistency" in diffusion models. Specifically, we review three notable models: Consistent Diffusion Model (CDM) (Daras et al., 2023), Consistency Model (CM) (Song et al., 2023), and Fokker-Planck (FP) Diffusion (Lai et al., 2022). Table 1 compares the distinguishing characteristics of these consistency-type models.

3.1. CDM (Daras et al., 2023)

Daras et al. (2023) introduced the concept of a "consistent denoiser" for the SDE (2). By leveraging Tweedie's formula (Efron, 2011), a connection can be established between the score function $\nabla_{\boldsymbol{x}} \log q_t(\boldsymbol{x})$ and a denoiser $\boldsymbol{h}\colon \mathbb{R}^D \times [t_0, T] \to \mathbb{R}^D$ conditioned on time

$$\nabla_{\boldsymbol{x}} \log q_t(\boldsymbol{x}) = \frac{\boldsymbol{h}(\boldsymbol{x}, t) - \boldsymbol{x}}{\sigma^2(t)}.$$

Consequently, Eq. (2) can be rearranged as follows:

$$d\boldsymbol{x}(t) = -g^2(t)\Big(\frac{\boldsymbol{h}(\boldsymbol{x}, t) - \boldsymbol{x}}{\sigma^2(t)}\Big)dt + g(t)d\bar{\boldsymbol{w}}_t. \tag{4}$$

This reparameterization gives rise to the concept of a consistent denoiser $\boldsymbol{h}$ (Daras et al., 2023). A denoiser is considered consistent if, on average, it produces estimates of the nearly clean data that align with those obtained by solving Eq. (4) in reverse, regardless of the initial data used. A concise summary of its formal definition is presented below. Furthermore, Fig. 1(a) showcases its corresponding illustration.

Definition 3.1 (Consistent SDE-denoiser (Daras et al., 2023))**.** A function $\boldsymbol{h}\colon \mathbb{R}^D \times [t_0, T] \to \mathbb{R}^D$ is called a *consistent SDE-denoiser* if and only if $\boldsymbol{h}(\boldsymbol{x}, t) = \mathbb{E}_{p^{\text{SDE}}_{[t_0,t],\boldsymbol{h}}}\big[\boldsymbol{x}(t_0)|\boldsymbol{X}_t = \boldsymbol{x}\big]$ for all $\boldsymbol{x} \in \mathbb{R}^D$ and $t \in [t_0, T]$.

Here, $\mathbb{E}_{p^{\text{SDE}}_{[t_0,t],\boldsymbol{h}}}$ denotes the conditional expectation of $\boldsymbol{x}(t_0)$ with respect to the distribution of $p^{\text{SDE}}_{[t_0,t],\boldsymbol{h}}$ along the stochastic trajectory described by the SDE presented in Eq. (4). This trajectory starts with an initial value of $\boldsymbol{x}$ at an arbitrary time t and terminates at a generated sample $\boldsymbol{x}(t_0)$ by running the SDE in Eq. (4) backwards in time.

The aim of Daras et al. (2023) is to train a diffusion model to serve as a consistent SDE-denoiser. However, applying this condition to practical settings is challenging due to the time-consuming process outlined in Definition 3.1, which involves multiple SDE solving by running from t to t_0 in order to accurately evaluate the average. In contrast, Daras et al. (2023) observed that a consistent SDE-denoiser $\boldsymbol{h}$ can be interpreted as a reverse martingale under the same process described in Eq. (2). Fig. 1(b) illustrates this property.

Proposition 3.2 (Daras et al. (2023))**.** *$\boldsymbol{h}$ is a consistent SDE-denoiser if and only if the following properties hold:*

(i) (Reverse martingale) For all $t > t'$ and $\boldsymbol{x}$, we have $\boldsymbol{h}(\boldsymbol{x}, t) = \mathbb{E}_{p^{SDE}_{[t',t],\boldsymbol{h}}}\big[\boldsymbol{h}(\boldsymbol{x}(t'), t')|\boldsymbol{X}_t = \boldsymbol{x}\big]$.

(ii) (Identity at t_0) For all $\boldsymbol{x} \in \mathbb{R}^D$, $\boldsymbol{h}(\boldsymbol{x}, t_0) = \boldsymbol{x}$.

Here, $\mathbb{E}_{p^{SDE}_{[t',t],\boldsymbol{h}}}$ represents the conditional expectation of $\boldsymbol{x}(t')$ given the distribution of $p^{SDE}_{[t',t],\boldsymbol{h}}$ along the trajectory described by the SDE in Eq. (4), starting with an initial value $\boldsymbol{x}$ at time t and terminates at time t'.

Based on this proposition, Daras et al. (2023) proposed to train a denoiser $\boldsymbol{h}_{\boldsymbol{\theta}}$ by using the denoising score matching (DSM) loss, along with a regularizer which is defined as

$$\frac{1}{2}\Big(\boldsymbol{h}_{\boldsymbol{\theta}}(\boldsymbol{x}, t) - \mathbb{E}_{p^{\text{SDE}}_{[t',t],\boldsymbol{h}_{\boldsymbol{\theta}}}}\big[\boldsymbol{h}_{\boldsymbol{\theta}}(\boldsymbol{x}(t'), t')|\boldsymbol{X}_t = \boldsymbol{x}\big]\Big)^2, \tag{5}$$

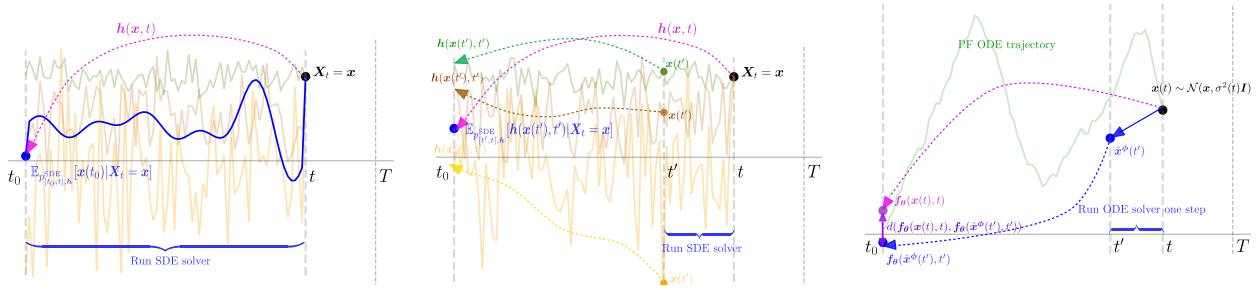
and its purpose is to enforce the reverse martingale property. Their approach demonstrates notable improvement in terms of sample quality.

3.2. CM (Song et al., 2023)

Song et al. (2023) directed their attention to PF ODE in Eq. (3) (a deterministic process). They introduced the notion of a "consistency function" which promotes the model's prediction of nearly clean data that aligns with the trajectory of the PF ODE. In order to establish a better connection with CDM (which will be discussed in Sec. 4.1), we propose a modification to their terminology, replacing "consistency function" with "consistent ODE-denoiser". Below, we provide the formal definition.

Table 1. Comparison of existing consistency-type models.

Models	Purpose	Trajectory	Object of Eq.	Approach
CDM (Daras et al., 2023)	Sample quality	Backward SDE	Samples	DSM + Martingale regularizer
CM (Song et al., 2023)	Sampling speed	PF ODE	Samples	Specific NN structure + New training scheme
FP-Diffusion (Lai et al., 2022)	Likelihood	Score FPE (a PDE)	Scores	DSM + Score FPE-regularizer



(a) Illustration of Def. 3.1. A consistent SDE-denoiser indicates that the SDE-denoiser prediction $\boldsymbol{h}(\boldsymbol{x}, t)$ (endpoint of the magenta arrow) aligns with the average of SDE predictions $\mathbb{E}_{p^{\text{SDE}}_{[t_0,t],\boldsymbol{h}}}\big[\boldsymbol{x}(t_0)|\boldsymbol{X}_t = \boldsymbol{x}\big]$ (blue dot).

(b) Illustration of Prop. 3.2. The SDE-denoiser prediction $\boldsymbol{h}(\boldsymbol{x}, t)$ (endpoint of the magenta arrow) aligns with the average prediction of intermediate points obtained by first applying an SDE solver and subsequently applying the SDE-denoiser $\mathbb{E}_{p^{\text{SDE}}_{[t',t],\boldsymbol{h}}}\big[\boldsymbol{h}(\boldsymbol{x}(t'), t')|\boldsymbol{X}_t = \boldsymbol{x}\big]$ (blue dot).

(c) Illustration of Alg. 2 in (Song et al., 2023). The objective of CM is to align the prediction of the direct denoiser (endpoint of the magenta arrow) with the prediction obtained by first applying a one-step ODE solver and subsequently applying the denoiser.

Figure 1. Illustration of a consistent SDE/ODE-denoiser. The dashed curves depict predictions from denoisers. In (a) and (b), the three erratic curves in light hues represent the SDE trajectories; the blue curve in (a) indicates the deterministic trajectory of the average. In (c), the green curve corresponds to the PF ODE trajectory, and ϕ represents the parameters of a pre-trained score model.

Definition 3.3 (Consistent ODE-denoiser (Song et al., 2023))**.** Given a solution trajectory $\{\boldsymbol{x}(t)\}_{t\in[t_0,T]}$ of the PF ODE in Eq. (3). Given a time-dependent vector field $\boldsymbol{f}$. $\boldsymbol{f}$ is called a *consistent ODE-denoiser* if it satisfies

$$\boldsymbol{f}(\boldsymbol{x}(t), t) = \boldsymbol{x}(t_0) \quad \text{for all} \quad t \in [t_0, T]. \tag{6}$$

A straightforward corollary of the definition is that a consistent ODE-denoiser $\boldsymbol{f}$ must fulfill the condition:

$$\boldsymbol{f}(\boldsymbol{x}(t), t) = \boldsymbol{f}(\boldsymbol{x}(t'), t') \quad \text{for all} \quad t, t' \in [t_0, T]. \tag{7}$$

The goal of (Song et al., 2023) is to train a network $\boldsymbol{f}_{\boldsymbol{\theta}}$ to satisfy both Eqs. (6) and (7). Eq. (6) is guaranteed by employing a specific network design, while Eq. (7) is learned through the minimization of a "distance" measured by $d(\cdot,\cdot)$:

$$d(\boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{x}(t), t), \boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{x}(t'), t')). \tag{8}$$

Thanks to the specific design of the network, the learned consistent ODE-denoiser is capable of performing one-step sampling. We visually depict the distillation training of Song et al. (2023) (Algorithm 2) in Fig. 1(c).

3.3. FP-Diffusion (Lai et al., 2022)

Lai et al. (2022) established an equivalent system of PDEs known as the "*score Fokker-Planck equation*" (*score FPE*) for the ground truth score (see Eq. (9)), which is built upon the classic FPE introduced by Fokker (1914); Planck (1917). The score FPE describes the temporal evolution of the ground truth score once the forward SDE is given. We present their findings in the following proposition.

Proposition 3.4 (Lai et al. (2022))**.** *The ground truth score* $\boldsymbol{s} = \boldsymbol{s}(\boldsymbol{x}, t) := \nabla_{\boldsymbol{x}} \log q_t(\boldsymbol{x})$ *satisfies the score FPE*

$$\partial_t \boldsymbol{s} = \frac{1}{2} g^2(t) \nabla_{\boldsymbol{x}} \big(\text{div}_{\boldsymbol{x}}(\boldsymbol{s}) + \|\boldsymbol{s}\|_2^2\big). \tag{9}$$

Lai et al. (2022) noted that DSM pre-trained score-based diffusion models often do not conform to the underlying score FPE (which is satisfied by the ground truth score). To address this issue, they proposed a regularization approach by minimizing the residuals of the score FPE, referred to as the *score FPE-regularizer*, into the DSM loss (as described in Eq. (19) of Lai et al. (2022)). The authors provided both theoretical and numerical evidence that their proposed model, known as *FP-Diffusion*, can enhance likelihood estimation.

4. Relation of consistency-type models

4.1. CDM and CM

In this section, we establish a connection between CDM and CM. Specifically, we demonstrate that the notion of a con-

sistent SDE-denoiser can be transformed into a consistent ODE-denoiser when the underlying trajectory is governed by a PF ODE.

To rigorously formulate the theorem, we incorporate a parameter $\lambda \geq 0$ that establishes a connection between Eqs. (2) and (3) as Karras et al. (2022)

$$d\boldsymbol{x}(t) = -(\frac{1+\lambda}{2})g^2(t)\boldsymbol{s}(\boldsymbol{x},t)dt + \lambda g(t)\bar{\boldsymbol{w}}_t. \tag{10}$$

Furthermore, we introduce $\mathbb{E}_{p^{\lambda-\text{SDE}}_{[t',t],\boldsymbol{h}}}$ analogous to Proposition 3.1, to denote the conditional expectation of $\boldsymbol{x}(t')$ based on the distribution of $p^{\lambda-\text{SDE}}_{[t',t],\boldsymbol{h}}$ along the stochastic trajectory described by the λ-parametrized SDE presented in Eq. (10) starting from $\boldsymbol{x}(t)$ and terminating at time t'. It is observed that when $\lambda = 1$, Eq. (10) coincides with Eq. (2) and $p^{1-\text{SDE}}_{[t',t],\boldsymbol{h}}$ is equivalent to $p^{\text{SDE}}_{[t_0,t],\boldsymbol{h}}$. On the other hand, when $\lambda = 0$, Eq. (10) becomes the PF ODE in Eq. (3). We now present the relationship of CDM and CM as the following theorem.

Theorem 4.1 (Relationship of CDM and CM)**.** *If $\lambda = 0$, then a consistent SDE-denoiser is equivalent to a consistent ODE-denoiser.*

Proof. Consider a denoiser $\boldsymbol{h}$ and $\lambda = 0$. This is equivalent to the formulation presented in Definition 3.1, but replacing the trajectory governed by the SDE in Eq. (4) with a PF ODE described by Eq.(3). Then for all $t \in [t_0, T]$, we have $\boldsymbol{h}(\boldsymbol{x}(t), t) = \mathbb{E}_{p^{0-\text{SDE}}_{[t_0,t],\boldsymbol{h}}}\left[\boldsymbol{x}(t_0)|\boldsymbol{X}_t = \boldsymbol{x}(t)\right] = \boldsymbol{x}(t_0)$, owing to the deterministic nature of the ODE. Alternatively, we can arrive at the same conclusion by leveraging Proposition 3.2. The reverse martingale property implies that $\boldsymbol{h}(\boldsymbol{x}(t), t) = \mathbb{E}_{p^{0-\text{SDE}}_{[t',t],\boldsymbol{h}}}\left[\boldsymbol{h}(\boldsymbol{x}(t'), t')|\boldsymbol{X}_t = \boldsymbol{x}(t)\right] = \boldsymbol{h}(\boldsymbol{x}(t'), t')$ holds for all $t' < t$, which is a consequence of the ODE trajectory being deterministic. In particular, if we set $t' = t_0$ and consider condition (ii) of Proposition 3.2, we also obtain $\boldsymbol{h}(\boldsymbol{x}(t), t) = \boldsymbol{h}(\boldsymbol{x}(t_0), t_0) = \boldsymbol{x}(t_0)$ for all $t \in [t_0, T]$. □

Both perspectives presented in the proof, namely Definition 3.1 and Proposition 3.2, ultimately yield the same conclusion: the theoretical correlation between Definition 3.1 and Definition 3.3. Indeed, Figs. 1(b) and (c) illustrate an analogy between a consistent SDE-denoiser and ODE-denoiser. Furthermore, it is worth noting that when we incorporate the CDM's regularizer in Eq. (5) along the PF ODE, it simplifies to $\frac{1}{2}\left(\boldsymbol{h}_{\boldsymbol{\theta}}(\boldsymbol{x},t) - \boldsymbol{h}_{\boldsymbol{\theta}}(\boldsymbol{x}(t'),t')\right)^2$, which coincides with Eq. (8) when the measurement d is taken as a $\frac{1}{2}$-weighted MSE.

4.2. FP-Diffusion and CDM

According to Daras et al. (2023), a consistent SDE-denoiser $\boldsymbol{h}$ is sufficient to guarantee the fulfillment of the score FPE by its corresponding induced score $\boldsymbol{s}(\boldsymbol{x},t) := \frac{\boldsymbol{h}(\boldsymbol{x},t)-\boldsymbol{x}}{\sigma^2(t)}$. In this section, we establish the necessity of this condition, thereby demonstrating the equivalence between these two concepts.

Theorem 4.2 (Equivalence of consistent SDE-denoiser and satisfaction of score FPE)**.** *Let $\boldsymbol{h}\colon \mathbb{R}^D \times [t_0, T] \to\colon \mathbb{R}^D$ be a smooth denoiser so that $\boldsymbol{h}(\boldsymbol{x}, t_0) = \boldsymbol{x}$, for all $\boldsymbol{x}$. Then $\boldsymbol{h}$ is a consistent SDE-denoiser if and only if $\boldsymbol{s}$ satisfies Eq. (9).*

Proof. The proof is built upon the approach presented in (Daras et al., 2023). It is important to note that once the forward SDE (Eq. (1)) is defined, it naturally corresponds to a reverse SDE (Eq. (2)). By employing the multidimensional Itô's Lemma (Daras et al., 2023, Lemma A.2) to Eq. (2), we obtain

$$d\boldsymbol{h} = \left[\frac{\partial \boldsymbol{h}}{\partial t} - \frac{1}{2}g^2(t)\Delta_{\boldsymbol{x}}\boldsymbol{h} - g^2(t)\mathcal{J}_{\boldsymbol{h}} \cdot \boldsymbol{s}\right]dt + g(t)\mathcal{J}_{\boldsymbol{h}}d\bar{\boldsymbol{w}}_t. \tag{11}$$

Here $\left(\Delta_{\boldsymbol{x}}\boldsymbol{F}(\boldsymbol{x},t)\right)_i := \sum_{j=1}^D \partial^2_{x_j} F_i(\boldsymbol{x},t)$ denotes the $\boldsymbol{x}$-Laplacian of a vector field $\boldsymbol{F} := (F_i)_{i=1}^D \colon \mathbb{R}^D \times [t_0, T] \to \mathbb{R}^D$ with a fixed time t. Moreover, $\mathcal{J}_{\boldsymbol{F}}$ denotes the Jacobian of $\boldsymbol{F}$ with respect to $\boldsymbol{x}$ (with a fixed time t). We know from Proposition 3.2 of Daras et al. (2023) that $\boldsymbol{h}$ is a consistent SDE-denoiser if and only if $\boldsymbol{h}$ is a reverse martingale. Indeed, it is equivalent to driftlessness of Eq. (11) by applying Proposition A.3. Namely, the following equation is valid

$$\partial_t \boldsymbol{h} = \frac{1}{2}g^2(t)\Delta_{\boldsymbol{x}}\boldsymbol{h} + g^2(t)\mathcal{J}_{\boldsymbol{h}} \cdot \boldsymbol{s}.$$

Finally, Lemma A.4, establishing a connection between the score FPE and a PDE that a denoiser $\boldsymbol{h}$ must satisfy, implies the equivalence of fulfilling the score FPE by $\boldsymbol{s}$. □

Despite the theoretical equivalence of a consistent SDE-denoiser and the fulfillment of its score FPE, the empirical experiments conducted by Daras et al. (2023) and Lai et al. (2022) used different approaches and achieved different outcomes. Daras et al. (2023) enforced consistency through regularization to promote the martingale property, whereas Lai et al. (2022) applied a regularizer to ensure satisfaction of the score FPE. These different methods result in distinct loss landscapes and optimization dynamics. We defer the empirical investigation of consistency-type models to future research.

5. Conclusion

In this research, we provide a theoretical bridge between different consistency-type models. The results of this study have the potential to inspire the development of a comprehensive framework that ensures consistency and facilitates the simultaneous achievement of several desired benefits, such as accurate likelihood estimation, efficient sampling speed, and high sample quality.

References

Anderson, B. D. Reverse-time diffusion equation models. *Stochastic Processes and their Applications*, 12(3):313–326, 1982.

Boffi, N. M. and Vanden-Eijnden, E. Probability flow solution of the fokker-planck equation. *arXiv preprint arXiv:2206.04642*, 2022.

Daras, G., Dagan, Y., Dimakis, A. G., and Daskalakis, C. Consistent diffusion models: Mitigating sampling drift by learning to be consistent. *arXiv preprint arXiv:2302.09057*, 2023.

Efron, B. Tweedie's formula and selection bias. *Journal of the American Statistical Association*, 106(496):1602–1614, 2011.

Fokker, A. D. Die mittlere energie rotierender elektrischer dipole im strahlungsfeld. *Annalen der Physik*, 348(5): 810–820, 1914.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

Karatzas, I., Karatzas, I., Shreve, S., and Shreve, S. E. *Brownian motion and stochastic calculus*, volume 113. Springer Science & Business Media, 1991.

Karras, T., Aittala, M., Aila, T., and Laine, S. Elucidating the design space of diffusion-based generative models. *arXiv preprint arXiv:2206.00364*, 2022.

Kawar, B., Elad, M., Ermon, S., and Song, J. Denoising diffusion restoration models. *arXiv preprint arXiv:2201.11793*, 2022.

Kim, D., Kim, Y., Kang, W., and Moon, I.-C. Refining generative process with discriminator guidance in score-based diffusion models. *arXiv preprint arXiv:2211.17091*, 2022.

Klenke, A. *Probability theory: a comprehensive course*. Springer Science & Business Media, 2013.

Lai, C.-H., Takida, Y., Murata, N., Uesaka, T., Mitsufuji, Y., and Ermon, S. Improving score-based diffusion models by enforcing the underlying score fokker-planck equation. 2022.

Meng, C., Song, Y., Song, J., Wu, J., Zhu, J.-Y., and Ermon, S. Sdedit: Image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021.

Murata, N., Saito, K., Lai, C.-H., Takida, Y., Uesaka, T., Mitsufuji, Y., and Ermon, S. Gibbsddrm: A partially collapsed gibbs sampler for solving blind inverse problems with denoising diffusion restoration, 2023.

Øksendal, B. Stochastic differential equations. In *Stochastic differential equations*, pp. 65–84. Springer, 2003.

Planck, V. Über einen satz der statistischen dynamik und seine erweiterung in der quantentheorie. *Sitzungberichte der*, 1917.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022.

Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E., Ghasemipour, S. K. S., Ayan, B. K., Mahdavi, S. S., Lopes, R. G., et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022.

Saito, K., Murata, N., Uesaka, T., Lai, C.-H., Takida, Y., Fukui, T., and Mitsufuji, Y. Unsupervised vocal dereverberation with diffusion-based generative models. *arXiv preprint arXiv:2211.04124*, 2022.

Schilling, R. L. *Brownian Motion: A Guide to Random Processes and Stochastic Calculus*. Walter de Gruyter GmbH & Co KG, 2021.

Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., and Ganguli, S. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pp. 2256–2265. PMLR, 2015.

Song, Y. and Ermon, S. Generative modeling by estimating gradients of the data distribution. *Advances in Neural Information Processing Systems*, 32, 2019.

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*, pp. 574–584. PMLR, 2020a.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020b.

Song, Y., Dhariwal, P., Chen, M., and Sutskever, I. Consistency models. *arXiv preprint arXiv:2303.01469*, 2023.

Vincent, P. A connection between score matching and denoising autoencoders. *Neural computation*, 23(7):1661–1674, 2011.

A. Auxiliary lemmas and propositions

In this section, we present essential lemmas for proving Theorem 4.2.

Let us begin by revisiting the definition of a *local martingale* (Karatzas et al., 1991), which presents a localized version of the martingale property. A stochastic process $\{\boldsymbol{X}(t)\}_{t\in[t_0,T]}$ is considered a local martingale if there exists a sequence of stopping times $\{\tau_k\}_{k=1}^{\infty}$ (which are random variables) satisfying the following conditions: (i) τ_k is almost surely increasing, (ii) τ_k diverges to ∞ almost surely, and (iii) the stopped process $\boldsymbol{X}(\min\{t,\tau_k\})$ is a martingale.

It is well-known that a diffusion process without drift (e.g., Eq. (12)) is a local martingale in general. However, it is important to note that without additional conditions, it does not necessarily qualify as a true martingale. Here, we demonstrate a sufficient condition to imply the true martingale property of a local martingale. For the proof of this proposition, we refer to the provided source (Karatzas et al., 1991).

Lemma A.1. *Suppose that* $\mathbb{E}\big[\int_{t_0}^{T}\|\boldsymbol{G}(\boldsymbol{X}(\tau),\tau)\|_2^2\,d\tau\big]<\infty$. *Then the Itô integral defined as*

$$\boldsymbol{X}(t)=\boldsymbol{X}(t_0)+\int_{t_0}^{t}\boldsymbol{G}(\boldsymbol{X}(\tau),\tau)d\boldsymbol{w}_\tau \tag{12}$$

is a martingale.

Next, we present a classic result that establishes the constancy of any continuous local martingale with bounded variation. The proof of this lemma can be found in the reference (Karatzas et al., 1991; Schilling, 2021).

Lemma A.2. *Any continuous local martingale of bounded variation is constant.*

Now, we can establish a characterization of the martingale property for solutions of SDEs, which asserts the non-trivial equivalence between a martingale and a "driftless" Itô process.

Proposition A.3. *Assume that a (Itô's) stochastic process* $\{\boldsymbol{X}(t)\}_{t\in[t_0,T]}$ *is the strong solution of the following reverse time SDE on* $[t_0,T]$

$$d\boldsymbol{X}(t)=\boldsymbol{F}(\boldsymbol{X}(t),t)dt+\boldsymbol{G}(\boldsymbol{X}(t),t)d\bar{\boldsymbol{w}}_t. \tag{13}$$

Here $\boldsymbol{F},\boldsymbol{G}\colon\mathbb{R}^D\times[t_0,T]\to\mathbb{R}^D$ *are vector fields satisfying some smoothness conditions* [2]*, and assumed to be Lipschitz and sub-linear. Namely, there is a constant* $C>0$ *such that*

(i) (Lipschitzness) For all $\boldsymbol{x},\boldsymbol{y}\in\mathbb{R}^D$ *and* $t\in[t_0,T]$

$$\|\boldsymbol{F}(\boldsymbol{x},t)-\boldsymbol{F}(\boldsymbol{y},t)\|_2+\|\boldsymbol{G}(\boldsymbol{x},t)-\boldsymbol{G}(\boldsymbol{y},t)\|_2\le C\,\|\boldsymbol{x}-\boldsymbol{y}\|_2\,.$$

(ii) (Sub-linearity) For all $\boldsymbol{x}\in\mathbb{R}^D$ *and* $t\in[t_0,T]$

$$\|\boldsymbol{F}(\boldsymbol{x},t)\|+\|\boldsymbol{G}(\boldsymbol{x},t)\|\le C(1+\|\boldsymbol{x}\|_2).$$

Then $\boldsymbol{X}(t)$ *is a reverse martingale if and only if Eq. (13) is driftless, i.e.,* $\boldsymbol{F}\equiv 0$.

The proposition can be understood intuitively by taking expectation of Eq. (13) conditioned on its history $\{\boldsymbol{X}(s)\}_{s\le t}$

$$\frac{d\mathbb{E}[\boldsymbol{X}(t)]}{dt}=\mathbb{E}[\boldsymbol{F}(\boldsymbol{X}(t),t)].$$

$\mathbb{E}[\boldsymbol{G}(\boldsymbol{X}(t),t)d\boldsymbol{w}_t]$ is zero because of the zero mean of the standard Wiener process $\boldsymbol{w}_t$. Therefore, the expectation remains constant (and consequently, independent of the past) if and only if $\boldsymbol{F}$ is identically equal to zero. Nevertheless, proving this proposition requires more intricate analysis, and we provide a detailed argument in Appx. B.1.

In the next Lemma, we bridge the score FPE and a PDE that a denoiser $\boldsymbol{h}$ should satisfy. To be consistent with notations in (Daras et al., 2023), we can derive the score FPE (Eq. (9)) of the ground truth score $\boldsymbol{s}$ as

$$\partial_t\boldsymbol{s}=\frac{1}{2}g^2(t)\Delta_{\boldsymbol{x}}\boldsymbol{s}+g^2(t)\mathcal{J}_{\boldsymbol{s}}\cdot\boldsymbol{s}. \tag{14}$$

Here $\big(\Delta_{\boldsymbol{x}}\boldsymbol{F}(\boldsymbol{x},t)\big)_i:=\sum_{j=1}^{D}\partial_{x_j}^2F_i(\boldsymbol{x},t)$ denotes the $\boldsymbol{x}$-Laplacian of a vector field $\boldsymbol{F}:=(F_i)_{i=1}^{D}\colon\mathbb{R}^D\times[t_0,T]\to\mathbb{R}^D$ with a fixed time t. Moreover, $\mathcal{J}_{\boldsymbol{F}}$ denotes the Jacobian in $\boldsymbol{x}$ of $\boldsymbol{F}$ (with a fixed time t).

[2] Please refer to (Øksendal, 2003, Theorem 5.2.1).

Lemma A.4. *$\boldsymbol{s}$ satisfies the score FPE (or equivalently, Eq. (14)) if and only if $\boldsymbol{h}$ satisfies*

$$\partial_t \boldsymbol{h} = \frac{1}{2} g^2(t) \Delta_{\boldsymbol{x}} \boldsymbol{h} + g^2(t) \mathcal{J}_{\boldsymbol{h}} \cdot \boldsymbol{s}. \tag{15}$$

B. Proofs

B.1. Proof of Proposition. A.3

Proof. We establish the case for forward time SDEs and martingales, as the reverse time case can be derived through a similar line of reasoning (Klenke, 2013). We first prove the sufficient implication. Suppose that $\boldsymbol{X}(t)$ is a martingale. Then the Itô process $\boldsymbol{U}(t)$ defined as

$$\begin{aligned} \boldsymbol{U}(t) &:= \int_{t_0}^{t} \boldsymbol{F}(\boldsymbol{X}(\tau), \tau) d\tau \\ &= \boldsymbol{X}(t) - \boldsymbol{X}(t_0) - \int_{t_0}^{t} \boldsymbol{G}(\boldsymbol{X}(\tau), \tau) d\boldsymbol{w}_\tau \end{aligned}$$

is a local martingale by the preservation of the local martingale property. It can be observed that $\boldsymbol{U}(t)$ possesses a continuous path and bounded variation on the interval $[t_0, T]$, as a result of property (i) and the assumed smoothness. By utilizing Lemma A.2, we conclude that $U(t)$ is constant, specifically zero. Consequently, Eq. (13) is driftless.

Next, we prove the necessity. Suppose that $\boldsymbol{F} \equiv 0$. Thanks to conditions (i) and (ii) (and some additional technical conditions), the existence and uniqueness of the solution to Eq. (13) is guaranteed (see in (Øksendal, 2003, Theorem 5.2.1)) and the solution is finite in the L_2-sense

$$\mathbb{E}\Big[\int_{t_0}^{T} \|X(\tau)\|_2^2 \, d\tau \Big] < \infty.$$

We will show that it implies $\mathbb{E}\big[\int_{t_0}^{T} \|\boldsymbol{G}(\boldsymbol{X}(\tau), \tau)\|_2^2 \, d\tau \big] < \infty$. Consequently, according to Lemma A.1, it guarantees that $\boldsymbol{X}$ is a martingale. The sub-linearity of $\boldsymbol{G}$ indicates

$$\|\boldsymbol{G}(\boldsymbol{x}, t)\|_2^2 \leq C^2 (1 + \|x\|_2)^2 \leq 2C^2 (1 + \|x\|_2^2).$$

So

$$\mathbb{E}\Big[\int_{t_0}^{T} \|\boldsymbol{G}(\boldsymbol{X}(\tau), \tau)\|_2^2 \, d\tau \Big] \lesssim (T - t_0) + \mathbb{E}\Big[\int_{t_0}^{T} \|\boldsymbol{X}(\tau)\|_2^2 \, d\tau \Big] < \infty.$$

Here we use $\lesssim$ to absorb multiplicative constants. Consequently, utilizing Lemma A.1 and relying on the uniqueness of the strong solution, it follows that $\boldsymbol{X}$ is a martingale. □

B.2. Proof of Lemma. A.4

Proof. Suppose that $\boldsymbol{s}$ satisfies Eq. (14). We know that $\boldsymbol{h}(\boldsymbol{x}, t) = \boldsymbol{x} + \sigma^2(t) \boldsymbol{s}(\boldsymbol{x}, t)$. By taking ∂_t and using the chain rule, we obtain

$$\begin{aligned} \partial_t \boldsymbol{h} &= \frac{d\sigma^2(t)}{dt} \boldsymbol{s} + \sigma^2(t) \partial_t \boldsymbol{s} \\ &= g^2(t) \boldsymbol{s} + \sigma^2(t) \Big(\frac{1}{2} g^2(t) \Delta_{\boldsymbol{x}} \boldsymbol{s} + g^2(t) \mathcal{J}_{\boldsymbol{s}} \cdot \boldsymbol{s} \Big). \end{aligned} \tag{16}$$

Notice that $\Delta_{\boldsymbol{x}} \boldsymbol{s}(\boldsymbol{x}, t) = \Delta_{\boldsymbol{x}} \big(\frac{\boldsymbol{h}(\boldsymbol{x}, t) - \boldsymbol{x}}{\sigma^2(t)} \big) = \frac{1}{\sigma^2(t)} \Delta_{\boldsymbol{x}} \boldsymbol{h}(\boldsymbol{x}, t)$ and that $\mathcal{J}_{\boldsymbol{s}} = \frac{1}{\sigma^2(t)} (\mathcal{J}_{\boldsymbol{h}} - \boldsymbol{I})$, where $\boldsymbol{I}$ denote the identity matrix on $\mathbb{R}^D$. So Eq. (16) becomes

$$\begin{aligned} \partial_t \boldsymbol{h} &= g^2(t) \boldsymbol{s} + \frac{1}{2} g^2(t) \Delta_{\boldsymbol{x}} \boldsymbol{h} + g^2(t) (\mathcal{J}_{\boldsymbol{h}} - \boldsymbol{I}) \cdot \boldsymbol{s} \\ &= \frac{1}{2} g^2(t) \Delta_{\boldsymbol{x}} \boldsymbol{h} + g^2(t) \mathcal{J}_{\boldsymbol{h}} \cdot \boldsymbol{s}, \end{aligned}$$

which indicates Eq. (15) is fulfilled by $\boldsymbol{h}$. The reverse implication follows with a similar computation. □