Consistent Diffusion for Gaussian Blur and General Linear Processes with Additive Gaussian Noise

Kiril Bangachev Victor Butoi Peter Hoffman Gagik Magakyan

May 2025

1 Introduction

Diffusion Overview. Diffusion models have become one of the dominant approaches to generative artificial intelligence in the past decade [HJA20], most notably yielding state-of-the art performance in image and video generation [PEL+24, RBL+22]. Other tasks where diffusion has succeeded include planning problems [JDTL22, CXF+22], inverse problems [DNK+24], and more recently also text generation [LKW+23, WFL+23].

The success of diffusion models can be largely attributed to a good understanding of their theoretical foundations. At a high-level, a diffusion model learns how to sample from a target distribution $\mathcal{D}$ over $\mathbb{R}^D$ by coupling it with a distribution that is easy to sample from, e.g. the standard gaussian $\mathcal{N}(0, I)$, and then sampling from the joint. Any distribution over $\mathbb{R}^d$ can be easily coupled with $\mathcal{N}(0, I)$ via running a simple stochastic differential equation. Different formulations exist, but the classic Variance Preserving one (VP) reads

$$\begin{aligned}&dX_t = -X_t dt + \sqrt{2} dB_t,\\ &\text{where } X_0 \sim \mathcal{D}, \text{ and } dB_t \text{ is standard Brownian motion.}\end{aligned} \tag{Standard VP Forward Process}$$

Here, X_t is distributed like $e^{-t}X_0 + \mathcal{N}(0, (1-e^{-2t})I)$. Hence, as $t \longrightarrow +\infty$, the distribution converges to standard Gaussian.

Producing a sample (X_0, X_T) from the coupling of $\mathcal{D}$ and $\mathcal{N}(0, I)$ (taking T to be large and thus we can think of x_T as Gaussian) is easy if one has sample access to $\mathcal{D}$. One first samples $X_0 \sim \mathcal{D}$ and then numerically runs the (Standard VP Forward Process) with an approximate discretization. However, the practically relevant task is exactly the opposite – sampling from $\mathcal{N}(0, I)$ is easy while sampling from $\mathcal{D}$ is hard. Can we somehow utilize (Standard VP Forward Process) towards this end? It turns out that (Standard VP Forward Process) has a simple reverse process

$$\begin{aligned}&d\overleftarrow{X_t} = (-\overleftarrow{X_t} + \nabla \log p_{T-t}(\overleftarrow{X_t}))dt + \sqrt{2} dB_t,\\ &\text{where } X_\infty \sim \mathcal{N}(0, I), \text{ and } dB_t \text{ is standard Brownian motion.}\end{aligned} \tag{Standard VP Denoising}$$

Again, one can run the same process in reverse, provided that the *score* $\nabla \log p_t(\overleftarrow{X_t})$ is known. Here p_{T-t} is the distribution of X_{T-t} in the forward process.

While the score $\nabla \log p_{T-t}(\overleftarrow{X_t})$ depends on an unknown distribution p_{T-t}, it turns out that an equivalent formulation makes it easy to learn:

$$\nabla \log p_t(x) = \frac{e^{-t}\,\mathbb{E}[X_0 | X_t = x] - x}{1 - e^{-2t}}. \tag{Tweedie's Formula}$$

The conditional distribution $\mathbb{E}[X_0|X_t]$ can be learned from sample data. However, as empirically observed, for good performance of the standard diffusion models,

> A large amount of samples X_0 from the target distribution $\mathcal{D}$ are needed.

This phenomenon has an intuitive explanation. First, the distribution of $\mathbb{E}[X_0|X_t]$ is in high-dimensional. Second, the distribution of $\mathbb{E}[X_0|X_t]$ has many modes (for example, each dog breed is in a different mode of the distribution of natural images; and dog breeds are only a small fraction of all animals, which are only a small fraction of different entities appearing on natural images).

The Problem of Clean Data. While we may have access to a tremendous amount of good dog images, and thus be able to train good models generating dog images, this is not the case in other domains. Important examples in which clean data is scarce or completely absent are X-Ray images of rare-disease patients, expensive-to-produce protein sequences, astronomical images [**?**]. Can we still train a good diffusion model in the absence of clean data? One way to formalize is as follows:

Problem 1. Suppose that instead of having access to samples from $\mathcal{D}$, we only have samples passed through some known noise channel $\mathcal{C}$. That is, we observe samples of the form $\mathcal{C}(X_0)$ where $X_0 \sim \mathcal{D}$. Can we still learn how to sample from $\mathcal{D}$?

Information-theoretically, this problem has a rather straightforward answer. The channel $\mathcal{C}$ should have the following property, as observed in [BPD18]

$$d_{\mathsf{TV}}(X,Y) > 0 \iff d_{\mathsf{TV}}(\mathcal{C}(X),\mathcal{C}(Y)) > 0. \qquad \text{(Distributional Invertibility Condition)}$$

Of course, the main question here is of computational efficiency. When do *practical algorithms* succeed in Problem 1.

The work [DDD24] considers this problem in a setting when the corruption channel is just (Standard VP Forward Process) ran for some known time t_{nature}. That is, instead of having access to p_0, corresponding to the true distribution $\mathcal{D}$, one has access to $p_{t_{\mathsf{nature}}}$ which is the distribution of $e^{-t_{\mathsf{nature}}}X_0 + \mathcal{N}(0,(1-e^{-2t_{\mathsf{nature}}})I)$.

The authors show that in this model, one can still learn how to sample from the uncorrupted distribution $\mathcal{D}$, even though during training the model *never sees clean data*. The key ingredient in the proof is that the true score function $\nabla \log p_t(x) : \mathbb{R}_{\geq 0} \times \mathbb{R}^d \longrightarrow \mathbb{R}^d$ satisfies a certain condition called *consistency* [DDDD23, SDCS23]. The consistency ensures that if $s_\theta(x,t) = \nabla \log p_t(x)$ on $\mathbb{R}_{\geq t_{\mathsf{nature}}} \times \mathbb{R}^d$ and $s_\theta(x,t)$ is consistent on $\mathbb{R}_{\geq 0} \times \mathbb{R}^d$, the functions $s(x,t), \nabla \log p_t(x)$ would evolve in the same way everywhere and, thus, coincide on the entire domain $\mathbb{R}_{\geq 0} \times \mathbb{R}^d$. Hence, training a network $s_\theta(x,t)$ to be consistent everywhere and to approximate $\nabla \log p_t(x)$ for $t \geq t_{\mathsf{nature}}$ gives the desired property. For additional details, have a look on **??**.

Consistency Beyond Additive Gaussian Noise. But what about other noising channels $\mathcal{C}$ in Problem 1? In the real world, we encounter many examples in which the corruptions are not additive Gaussian noise. Natural examples include blurring (mathematically modeled as convolving with a Gaussian kernel), distorting, multiplicative (speckle) noise in astronomical imaging [LAD+14].

In this work, we aim to study Problem 1 under more general linear corruptions models (formally defined in Definition 1) which include as special cases the model of additive Gaussian noise, Gaussian blur and others. We are particularly interested in Gaussian blur due to its practical relevance and interpretability.

1. Consistency: [DDDD23, DDD24]
2. Deblurring: [BBC+22, DDT+22, HS24]

2 Preliminaries

2.1 Model

The general corruption model that we want to study in this work is for linear stochastic processes with additive Gaussian noise [DDT+22].

Definition 1. A linear process on $\mathbb{R}^d$ with additive Gaussian Noise is given by $X_0 \sim \mathcal{D}$ and

$$X_t = D_t X_0 + A_t \eta$$

where $t \longrightarrow D_t, t \longrightarrow A_t$ are sufficiently smooth full-rank deterministic matrix functions for which $D_0 = I, A_0 = 0$ and $\eta \sim \mathcal{N}(0, I)$.

We note that the only difference with [DDT+22] is that A_t is a scalar in their set-up, while we allow for matrix values. This is not simply for the sake of generality, but we believe that it may have some practical consequences, see Section 4. We give several examples:

1. (Standard VP Forward Process): Setting $D_t = e^{-t}I, A_t = \sqrt{1-e^{-2t}}I$ (and more generally multiples of the identity), we recover the standard diffusion model.
2. Gaussian Blur [RHS23, BBC+22]: In the Gaussian Blur model, one has that $X_t = G_t * X_0$ where this is the convolution with a kernel corresponding to a Guassian of variance t. This could be equivalently rewritten as the partial differential equation corresponding to the Laplacian: $dX_t = \frac{1}{2}\Delta X_t$. Since the Laplacian is a symmetric linear operator, it can be written as $dX_t = \frac{1}{2}V\Lambda V^T X_t d_t$ for some apropriate diagonal V and orthogonal Λ (to be discussed more in a moment). Hence, this is indeed a noiseless instance of Definition 1 with $X_t = V\exp(\frac{t}{2}\Lambda)V^T X_0$.
3. Gaussian Blue+Noise [DDT+22, HS24]: Similarly, one can mix the two models and obtain the model $dX_t = \frac{1}{2}V\Lambda V^T X_t d_t + \sigma_t \eta$.

Remark 1 (Why is inverting Gaussian blur non-trivial.)**.** Note that in the theoretical model of Gaussian blur, there is an explicit one-to-one deterministic map between clean image X_0 and noisy image X_t given by $X_t = V\exp(\frac{t}{2}\Lambda)V^T X_0$. Why is then this not a trivial task? There are several issues related to discretization. First, since we work over a discrete pixel space, the kernels that we work with are not exactly Gaussian. They are only approximations over the integer lattice which are, furthermore, truncated to a specific window in practice. Thus, if $\hat{G}_t$ corresponds to the practically implemented Kernel, it is not even through that $\hat{G}_1 * \hat{G}_1 = \hat{G}_2$ (even though, in the ideal continuous setting $G_1 * G_1 = G2$). Hence, any naive attempt to invert $G_t * X_0 = V\exp(\frac{t}{2}\Lambda)V^T X_0$ will have certain errors. Due to the exponential growth in $\exp(\frac{t}{2}\Lambda)$, this effect will compound. Hence, a more sophisticated denoising, for example based on diffusion, is needed.

We note that under a somewhat general condition, the process in Definition 1 can be rewritten as an SDE.

Proposition 2.1. *Suppose that we have the process Y_t satisfies $Y_t = D_t X_0 + A_t \eta$ for $\eta \sim \mathcal{N}(0, I)$ where D_t, A_t are both smooth and D_t is invertible for all t. Suppose that for all $0 \leq t \leq T$, they satisfy the following inequality:*

$$\dot{A}_t A_t^T + A_t \dot{A}_t{}^T \succeq \dot{D}_t D_t^{-1} A_t A_t^T + A_t A_t^T D_t^{-T} \dot{D}_t{}^T.$$

Then, Y_t satisfies the SDE

$$dY_t = (\dot{D}_t D_t^{-1}) Y_t dt + \sqrt{\dot{A}_t A_t^T + A_t \dot{A}_t{}^T - \dot{D}_t D_t^{-1} A_t A_t^T - A_t A_t^T D_t^{-T} \dot{D}_t{}^T} dB_t,$$

where $\sqrt{\cdot}$ denotes any PSD square root. In particular, $(Y_t)_{0 \leq t \leq T}$ is a Markov process.

The proof is rather standard and can be found in Appendix B.1. For any Lipschitz test function $h : \mathbb{R}^d \longrightarrow \mathbb{R}$ of compact support, we expresses $\frac{d}{dt} \mathbb{E}[h(Y_t)]$ both using the linear process definition and the corresponding SDE and show that the two expressions coincide.

Perhaps surprisingly, it turns out that the process can be expressed as a Markovian process if and only if it is an SDE. Namely,

Proposition 2.2. *Suppose that $t \longrightarrow D_t, A_t$ are both smooth and D_t is invertible for all t.mit Suppose that for some $T > 0$ and any Gaussian distribution of X_0, there exists some process $(Z_t)_{0 \leq t \leq T}$ which is Markovian and Z_t and $X_t = D_t X_0 + A_t \eta$ have the same marginal distribution for all $0 \leq t \leq T$. Then,*

$$\dot{A}_t A_t^T + A_t \dot{A}_t{}^T \succeq \dot{D}_t D_t^{-1} A_t A_t^T + A_t A_t^T D_t^{-T} \dot{D}_t{}^T.$$

The full proof is in Appendix B.2. The high-level strategy is as follows. Consider some space Θ, a family of distributions $(P_\theta)_{\theta \in \Theta}$ over $\mathbb{R}^d$ and some distribution Q_θ over Θ. Now, consider the following random variables for some $0 \leq t \leq r \leq T$:

$$\theta \sim Q_\theta, X_0 \sim P_\theta, Z_t \sim D_t X_0 + A_t \eta, Z_r \sim D_r X_0 + A_r \eta.$$

Now, if $(Z_t)_{0 \leq t \leq T}$ is indeed Markovian, then clearly

$$\theta \longrightarrow Z_t \longrightarrow Z_r$$

is a Markov chain. In particular, this means that

$$I(\theta; Z_t) \geq I(\theta; Z_r).$$

We will define a collection of possible distributions $(\theta, P_\theta)_{\theta \in \Theta}$ in such a way that $t \longrightarrow I(\theta; Z_t)$ is always differentiable. Hence, the inequality will imply that $\frac{d}{dt} I(\theta; Z_t)$ is non-positive for any $(\theta, P_\theta)_{\theta \in \Theta}$ in the collection. From this, we will derive the conclusion.

2.2 More on Gaussian Blur and Diagonalizable Processes

One instance of interest in Definition 1 for which it is easier to reduce the process to standard diffusion is when the matrices $(D_t, A_t)_{t \geq 0}$ are symmetric and jointly diagonalizable. That is, there exists some orthogonal matrix V such that $D_t = V \Lambda_t V^T$, $A_t = V \Psi_t V^T$ and Λ_t, Ψ_t are both diagonal.

This structure holds for the Laplacian pde associated to Gaussian blur as discussed in [RHS23]. Here, we give a more general presentation.

Then, we can write

$$\begin{aligned}
&X_t = V\Lambda_t V^T X_0 + V\Psi_t V^T \eta \iff \\
&V^T X_t = \Lambda_t V^T X_0 + \Psi_t \eta \iff \\
&Y_t = \Lambda_t Y_0 + \Psi_t \eta,
\end{aligned}$$

where we denoted $Y_t = V^T X_t$ and used the noise η is rotation invariant.

This is the case, for example, for Gaussian Blurring with $\Lambda_t = \exp(\frac{t}{2}\Lambda)$ and any Gaussian noising with covariance of the form $V\Phi V^T$ for a diagonal Φ.

It must be noted that for diagonal matrices, Proposition 2.1 simplifies significantly.

Proposition 2.3. *Suppose that we have the process Y_t satisfies $Y_t = \Lambda_t Y_0 + \Psi_t \eta$, where $\Lambda_t > 0, \Psi_t$ are both smooth,diagonal, and we have Λ_t is does not have zero entries for any t. Additionally, for $\Sigma_t = \frac{\Psi_t}{\Lambda_t}$ we have $\frac{d}{dt}\Sigma_t^2 \geq 0$ (element-wise). Define $g_t = \sqrt{\frac{d}{dt}\Sigma_t^2}$. Than, Y_t satisfies the following SDE:*

$$Y_t = -\frac{\dot{\Lambda}_t}{\Lambda_t} Y_t dt + \Lambda_t g_t dB_t,$$

with the reverse SDE:

$$dY_t = \left[\frac{\dot{\Lambda}_t}{\Lambda_t} Y_t - \Lambda_t^2 g_t^2 \nabla_y \log p_t(Y_t)\right] dt + \Lambda_t g_t d\bar{B}_t$$

2.3 Learning the score

Note that by the Lemma A.1 we have that:

$$\nabla_y \log p_t(y_t) = \frac{\Lambda_t \mathbb{E}[Y_0 | Y_t = y_t] - y_t}{\Psi_t^2}, \tag{1}$$

and we can learn $\mathbb{E}[Y_0|Y_t = y_t]$ by optimizing the following loss:

$$L(\theta, t) = \mathop{\mathbb{E}}_{t \sim U[0,T]} \mathop{\mathbb{E}}_{Y_0 \sim p_0} \mathop{\mathbb{E}}_{Y_t \sim \Lambda_t Y_0 + \Psi_t \eta} \|g_\theta(Y_t, t) - Y_0\|_2^2.$$

This method finds the score in the frequency space. However, it is empirically observed that neural networks work well for finding the score in the pixel space. Therefore, we instead learn

$$h_\theta(x, t) = \mathbb{E}[X_0 | X_t = x],$$

by optimizing

$$L^1(\theta, t) = \mathop{\mathbb{E}}_{t \sim U[0,T]} \mathop{\mathbb{E}}_{X_0 \sim p_0} \mathop{\mathbb{E}}_{X_t \sim V\left(\Lambda_t V^\top X_0 + \Psi_t V^\top \eta\right)} \|h_\theta(X_t, t) - X_0\|_2^2.$$

Finnaly, note that we have:

$$\mathbb{E}[Y_0 | Y_t = y_t] = \mathbb{E}[V^\top X_0 | V^\top X_t = y_t] = V^\top \mathbb{E}[X_0 | X_t = V y_t] = V^\top h(V y_t, t).$$

Putting this into 1, gives us:

$$\nabla_y \log p_t(y_t) = \frac{\Lambda_t V^\top h(Vy_t, t) - y_t}{\Psi_t^2},$$

and from Proposition 2.3, our reverse SDE becomes the following:

$$dY_t = \left[\frac{\dot{\Lambda}_t}{\Lambda_t} Y_t - \Lambda_t^2 g_t^2 \frac{\Lambda_t V^\top h(VY_t, t) - Y_t}{\Psi_t^2}\right] dt + \Lambda_t g_t d\bar{B}_t \tag{h-SDE}$$

Using h-SDE, we can define the consistency loss as follows:

$$L^2(\theta, t) = \mathop{\mathbb{E}}_{t \sim U[0,T]} \mathop{\mathbb{E}}_{X_t \sim p_t} \mathop{\mathbb{E}}_{t' \sim U[t-\epsilon, t]} \|h_\theta(X_t, t) - h_\theta(X_{t'}, t')\|_2^2,$$

where $X_{t'}$ is obtained first going to frequency space by calculating $Y_t = V^\top X_t$, than running h-SDE from t to t' to obtain $Y_{t'}$, and than going back to the pixel space by $X_{t'} = VY_{t'}$.

3 Learning the score from noisy data

Consider our process in the frequency space defined as follows:

$$Y_t = \Lambda_t \cdot Y_0 + \Psi_t \cdot \eta,$$

and suppose we have access to samples from a noisy distribution $Y_{t_{\text{nature}}}$. Firstly, we are going to consider a linear transformation of our original process:

$$\tilde{Y}_t = \frac{Y_t}{\Lambda_t},$$

which will result into the following process:

$$\begin{aligned} \tilde{Y}_t &= Y_0 + \frac{\Psi_t}{\Lambda_t} \cdot \eta \\ &= Y_0 + \Sigma_t \cdot \eta, \end{aligned}$$

where we used used $\Lambda_0 = I$. Now following [DDD24], we have:

$$\begin{aligned} \tilde{Y}_t &= Y_0 + \Sigma_t \cdot \eta \\ \tilde{Y}_t &= \tilde{Y}_{t_{\text{nature}}} + \sqrt{\Sigma_t^2 - \Sigma_{t_{\text{nature}}}^2} \cdot \eta. \end{aligned}$$

Calculating $\log \tilde{p}_t(\tilde{y}_t)$ using Lemma A.1 for both of the models and equalizing, gives us:

$$\frac{\mathbb{E}[Y_0 | \tilde{Y}_t = \tilde{y}_t] - \tilde{y}_t}{\Sigma_t^2} = \frac{\mathbb{E}[\tilde{Y}_{t_{\text{nature}}} | \tilde{Y}_t = \tilde{y}_t] - \tilde{y}_t}{\Sigma_t^2 - \Sigma_{t_{\text{nature}}}^2}.$$

This identity is nice but it relates the scores in the transformed frequency space. In order for it to be useful for our framework, we need to get a relation of scores in the pixel space. Firstly, we transform the equation back to the Y_t space:

$$\frac{\mathbb{E}[Y_0 | Y_t = \Lambda_t \tilde{y}_t] - \tilde{y}_t}{\Sigma_t^2} = \frac{\Lambda_{t_{\text{nature}}}^{-1} \mathbb{E}[Y_{t_{\text{nature}}} | Y_t = \Lambda_t \tilde{y}_t] - \tilde{y}_t}{\Sigma_t^2 - \Sigma_{t_{\text{nature}}}^2}.$$

Now, we go to the pixel space as follows:

$$\frac{V^\top \mathbb{E}[X_0|X_t = V\Lambda_t\tilde{y}_t] - \tilde{y}_t}{\Sigma_t^2} = \frac{\Lambda_{t_{\mathsf{nature}}}^{-1} V^\top \mathbb{E}[X_{t_{\mathsf{nature}}}|X_t = V\Lambda_t\tilde{y}_t] - \tilde{y}_t}{\Sigma_t^2 - \Sigma_{t_{\mathsf{nature}}}^2}.$$

Taking $\tilde{y}_t = \Lambda_t^{-1}V^\top x_t$ gives us:

$$\frac{V^\top \mathbb{E}[X_0|X_t = x_t] - \Lambda_t^{-1}V^\top x_t}{\Sigma_t^2} = \frac{\Lambda_{t_{\mathsf{nature}}}^{-1} V^\top \mathbb{E}[X_{t_{\mathsf{nature}}}|X_t = x_t] - \Lambda_t^{-1}V^\top x_t}{\Sigma_t^2 - \Sigma_{t_{\mathsf{nature}}}^2}.$$

Doing algebra, we obtain:

$$\begin{aligned}\mathbb{E}[X_0|X_t = x_t] &= V\left(\frac{\Sigma_t^2}{\Sigma_t^2 - \Sigma_{t_{\mathsf{nature}}}^2}\left(\Lambda_{t_{\mathsf{nature}}}^{-1}V^\top \mathbb{E}[X_{t_{\mathsf{nature}}}|X_t = x_t] - \Lambda_t^{-1}V^\top x_t\right) + \Lambda_t^{-1}V^\top x_t\right)\\ &= V\left(\frac{\Sigma_t^2}{\Sigma_t^2 - \Sigma_{t_{\mathsf{nature}}}^2}\Lambda_{t_{\mathsf{nature}}}^{-1}V^\top \mathbb{E}[X_{t_{\mathsf{nature}}}|X_t = x_t] + \frac{\Sigma_{t_{\mathsf{nature}}}^2}{\Sigma_t^2 - \Sigma_{t_{\mathsf{nature}}}^2}\Lambda_t^{-1}V^\top x_t\right)\end{aligned}$$

This gives us a framework of learning the score for time $t \geq t_{\mathsf{nature}}$, without having a direct access to clean samples X_0. In addition, if we know that our score function satisfies "consistency" ([DDD24]), than it will agree with the original function even for $t \leq t_{\mathsf{nature}}$, allowing us to sample up to Y_0.

We now provide a bound on the accuracy of our learned data distribution with respect to the true distribution under assumptions relating to consistency and smoothness.

Theorem 3.1. *Let $(X_t)_{t\in[0,T]}$ satisfy the true reverse SDE*

$$\mathrm{d}X_t = f(t, X_t)\,\mathrm{d}t + \Lambda_t\, g_t\,\mathrm{d}B_t, \qquad X_T \sim \mathcal{N}(0, I),$$

with

$$f(t, x) = -\Lambda_t^2\, g_t^2\, \nabla_x s(t, x).$$

Let $(\widehat{X}_t)_{t\in[0,T]}$ satisfy the learned reverse SDE

$$\mathrm{d}\widehat{X}_t = \widehat{f}(t, \widehat{X}_t)\,\mathrm{d}t + \Lambda_t\, g_t\,\mathrm{d}B_t, \qquad \widehat{X}_T \sim \mathcal{N}(0, I),$$

with

$$\widehat{f}(t, x) = -\Lambda_t^2\, g_t^2\, \widehat{s}(t, x).$$

Suppose that for all $t \in [0, T]$ the following hold:

1. *Lipschitz in space:*

$$\|f(t, x) - f(t, y)\| \;\leq\; L\,\|x - y\| \quad \forall\, x, y.$$

2. *Uniform noise bound:*

$$\|\Lambda_t\, g_t\| \;\leq\; \|\Lambda\|_\infty\, \|g\|_\infty =: M^{\frac{1}{2}}.$$

3. *Score-estimation error:*

$$\mathbb{E}_{x_t\sim p_t}\left[\|s(t, x_t) - \widehat{s}(t, x_t)\|^2\right] \;\leq\; \varepsilon^2.$$

   (By consistency the same holds under $\widehat{p}_t$.)

Then the Wasserstein-2 distance between the marginals at time 0 satisfies

$$W_2\big(\widehat{P}_0,\, P_0\big) \;\leq\; \frac{e^{LT} - 1}{L}\, M\,\varepsilon.$$

Please refer to Appendix D for a complete proof.

4 Future Work

In this work, we proposed a generalized consistency framework applicable to a broad set of linear corruptions beyond additive Gaussian noise, emphasizing Gaussian blur. Although we have laid a rigorous theoretical foundation, practical validation through experimental results remains a critical next step. Our initial experimental setup has not yet demonstrated empirical success, underscoring the need for careful tuning and implementation refinements. Moving forward, we plan comprehensive experiments designed to validate our theoretical predictions, specifically targeting Gaussian blur and mixed corruption scenarios. These empirical studies will allow us to assess the robustness and practical efficacy of our generalized consistency framework.

References

[BBC$^+$22] Arpit Bansal, Eitan Borgnia, Hong-Min Chu, Jie S. Li, Hamid Kazemi, Furong Huang, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Cold diffusion: Inverting arbitrary image transforms without noise, 2022.

[BPD18] Ashish Bora, Eric Price, and Alexandros G. Dimakis. AmbientGAN: Generative models from lossy measurements. In *International Conference on Learning Representations*, 2018.

[CXF$^+$22] Cheng Chi, Zhenjia Xu, Siyuan Feng, Eric Cousineau, Yilun Du, Benjamin Burchfiel, Russ Tedrake, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. *The International Journal of Robotics Research*, 0(0):02783649241273668, 2022.

[DDD24] Giannis Daras, Alexandros G. Dimakis, and Constantinos Daskalakis. Consistent diffusion meets tweedie: Training exact ambient diffusion models with noisy data, 2024.

[DDDD23] Giannis Daras, Yuval Dagan, Alexandros G. Dimakis, and Constantinos Daskalakis. Consistent diffusion models: Mitigating sampling drift by learning to be consistent, 2023.

[DDT$^+$22] Giannis Daras, Mauricio Delbracio, Hossein Talebi, Alexandros G. Dimakis, and Peyman Milanfar. Soft diffusion: Score matching for general corruptions, 2022.

[DNK$^+$24] Giannis Daras, Weili Nie, Karsten Kreis, Alexandros G. Dimakis, Morteza Mardani, Nikola B. Kovachki, and Arash Vahdat. Warped diffusion: Solving video inverse problems with image diffusion models. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 101116–101143. Curran Associates, Inc., 2024.

[HJA20] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc.

[HS24] Emiel Hoogeboom and Tim Salimans. Blurring diffusion models, 2024.

[JDTL22] Michael Janner, Yilun Du, Joshua Tenenbaum, and Sergey Levine. Planning with diffusion for flexible behavior synthesis. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 9902–9915. PMLR, 17–23 Jul 2022.

[LAD+14] Antoine Lucas, Oded Aharonson, Charles Deledalle, Alexander G. Hayes, Randolph Kirk, and Elpitha Howington-Kraus. Insights into titan's geology and hydrology based on enhanced image processing of cassini radar data. *Journal of Geophysical Research: Planets*, 119(10):2149–2166, 2014.

[LKW+23] Justin Lovelace, Varsha Kishore, Chao Wan, Eliot Seo Shekhtman, and Kilian Q Weinberger. Latent diffusion for language generation. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[PEL+24] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. In *The Twelfth International Conference on Learning Representations*, 2024.

[RBL+22] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Bjorn Ommer. High-Resolution Image Synthesis with Latent Diffusion Models . In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10674–10685, Los Alamitos, CA, USA, June 2022. IEEE Computer Society.

[RHS23] Severi Rissanen, Markus Heinonen, and Arno Solin. Generative modelling with inverse heat dissipation. In *The Eleventh International Conference on Learning Representations*, 2023.

[SDCS23] Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models, 2023.

[WFL+23] Tong Wu, Zhihao Fan, Xiao Liu, Hai-Tao Zheng, Yeyun Gong, yelong shen, Jian Jiao, Juntao Li, zhongyu wei, Jian Guo, Nan Duan, and Weizhu Chen. AR-diffusion: Auto-regressive diffusion model for text generation. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Appendix

A Appendix A

A.1 Generalized Tweedie formula.

Lemma A.1. *Consider the process defined as follows:*

$$Y_t = \Lambda_t \cdot Y_0 + \Psi_t \cdot \eta.$$

Than, the following equality holds:

$$\log p_t(y_t) = \frac{\Lambda_t\, \mathbb{E}[Y_0|Y_t = y_t] - y_t}{\Psi_t^2}.$$

Proof. We can find the score using calculations from the lecture as follows:

$$\begin{aligned}\log p_t(y_t) &= \int_{y_0} \left(\nabla_{y_t} \log p_t(y_t|Y_0 = y_0)\right) p_0(y_0|Y_t = y_t) dy \\ &= \int_{y_0} \left(\nabla_{y_t} \frac{-\left(y_t - \Lambda_t y_0\right)^2}{2}\right) p_0(y_0|Y_t = y_t) dx \\ &= \frac{\Lambda_t\, \mathbb{E}[Y_0|Y_t = y_t] - y_t}{\Psi_t^2},\end{aligned}$$

where all the operations are element-wise.

□

B On General Linear Processes

B.1 Proof of Proposition 2.1

Proposition B.1. *Suppose that we have the process Y_t satisfies $Y_t = D_t X_0 + A_t \eta$ for $\eta \sim \mathcal{N}(0, I)$ where D_t, A_t are both smooth and D_t is invertible for all t. Suppose that for all $0 \leq t \leq T$, they satisfy the following inequality:*

$$\dot{A}_t A_t^T + A_t \dot{A}_t{}^T \succeq \dot{D}_t D_t^{-1} A_t A_t^T + A_t A_t^T D_t^{-T} \dot{D}_t{}^T.$$

Then, Y_t satisfies the SDE

$$dY_t = (\dot{D}_t D_t^{-1}) Y_t dt + \sqrt{\dot{A}_t A_t^T + A_t \dot{A}_t{}^T - \dot{D}_t D_t^{-1} A_t A_t^T - A_t A_t^T D_t^{-T} \dot{D}_t{}^T} dB_t,$$

where $\sqrt{\cdot}$ denotes any PSD square root. In particular, $(Y_t)_{0 \leq t \leq T}$ is a Markov process.

Proof. Consider the two processes. The first one is given by

$$Y_t = D_t X_0 + A_t \eta.$$

The second one is given by

$$dZ_t = \dot{D}_t D_t^{-1}) Z_t dt + \sqrt{\dot{A}_t A_t^T + A_t \dot{A}_t{}^T - \dot{D}_t D_t^{-1} A_t A_t^T - A_t A_t^T D_t^{-T} \dot{D}_t{}^T} dB_t, \qquad Z_0 \sim X_0.$$

We will show that Y_t, Z_t have the same distribution for any t. □

Proof. Assume that $\dot{A}_t A_t^T + A_t \dot{A}_t{}^T - \dot{D}_t D_t^{-1} A_t A_t^T - A_t A_t^T D_t^{-T} \dot{D}_t{}^T$ is indeed PSD for any t and let K_t be an arbitrary square root of it. Let $h : \mathbb{R}^d \longrightarrow \mathbb{R}$ be any Lipschitz function with compact support.

Evolution of Linear Process. We compute

$$\begin{aligned}
&\frac{d}{dt}\,\mathbb{E}[h(Y_t)] \\
&= \lim_{\epsilon \longrightarrow 0} \frac{1}{\epsilon}\,\mathbb{E}[h(Y_{t+\epsilon}) - h(Y_t)] \\
&= \lim_{\epsilon \longrightarrow 0} \frac{1}{\epsilon} \mathop{\mathbb{E}}_{X_0,\eta}\left[h(D_{t+\epsilon}X_0 + A_{t+\epsilon}\eta) - h(D_tX_0 + A_t\eta)\right] \\
&= \lim_{\epsilon \longrightarrow 0} \frac{1}{\epsilon}\left(\mathop{\mathbb{E}}_{X_0,\eta}\left[h(D_tX_0 + \epsilon\dot{D}_tX_0 + o(\epsilon) + A_t\eta + \epsilon\dot{A}_t\eta + o(\epsilon))\right] - \mathbb{E}[h(D_tX_0 + A_t\eta)]\right) \\
&= \lim_{\epsilon \longrightarrow 0} \frac{1}{\epsilon}\left(\mathop{\mathbb{E}}_{X_0,\eta}\left[h\Big((D_tX_0 + A_t\eta) + \epsilon\dot{D}_tD_t^{-1}(D_tX_0 + A_t\eta) + \epsilon(\dot{A}_t - \dot{D}_tD_t^{-1}A_t)\eta\Big)\right] - \mathbb{E}[h(D_tX_0 + A_t\eta)]\right)
\end{aligned} \tag{2}$$

Now, observe that $(A_t + \epsilon\dot{A}_t)\eta$ has covariance matrix $A_tA_t^T + \epsilon A_t\dot{A}_t^T + \epsilon\dot{A}_tA_t^T + \epsilon^2\dot{A}_t\dot{A}_t^T$. To relate the above equation to the SDE, we will find some L_t such that for independent Gaussian vectors η, ψ

$$(A_t + \epsilon\dot{D}_tD_t^{-1}A_t)\eta + L_t\phi$$

also has covariance $A_tA_t^T + \epsilon A_t\dot{A}_t^T + \epsilon\dot{A}_tA_t^T + \epsilon^2\dot{A}_t\dot{A}_t^T + o(\epsilon^2)$. That way, the first expectation above can be rewritten as

$$\mathbb{E}[h(Y_t + \epsilon\dot{D}_tD_t^{-1}Y_t + L_t\phi)]$$

and we can relate the expression to the SDE.

Now, such an L_t should satisfy

$$(A_t + \epsilon\dot{D}_tD_t^{-1}A_t)(A_t + \epsilon\dot{D}_tD_t^{-1}A_t)^T + L_tL_t^T = A_tA_t^T + \epsilon A_t\dot{A}_t^T + \epsilon\dot{A}_tA_t^T + \epsilon^2\dot{A}_t\dot{A}_t^T + o(\epsilon^2).$$

Hence, one possible L_t is $L_t = \sqrt{\epsilon}\sqrt{\dot{A}_tA_t^T + A_t\dot{A}_t^T \succeq \dot{D}_tD_t^{-1}A_tA_t^T + A_tA_t^TD_t^{-T}\dot{D}_t^T} + O(\epsilon) = \sqrt{\epsilon}K_t + O(\epsilon)$. In other words, $L_t = O(\sqrt{\epsilon})$ Altogether, going back to (5), we obtain

$$\begin{aligned}
&\frac{d}{dt}\,\mathbb{E}[h(Y_t)] \\
&= \lim_{\epsilon \longrightarrow 0} \frac{1}{\epsilon}\left(\mathop{\mathbb{E}}_{Y_t,\phi}\left[h\Big(Y_t + \epsilon\dot{D}_tD_t^{-1}Y_t + L_t\phi\Big)\right] - \mathbb{E}[h(Y_t)]\right) \\
&= \lim_{\epsilon \longrightarrow 0} \frac{1}{\epsilon}\left(\mathop{\mathbb{E}}_{Y_t,\phi}\left[h(Y_t) + \langle\nabla h(Y_t), \epsilon\dot{D}_tD_t^{-1}Y_t + L_t\psi\rangle\right.\right. \\
&\qquad\left.\left. + \frac{1}{2}\langle\epsilon\dot{D}_tD_t^{-1}Y_t + L_t\psi, \nabla h(Y_t)(\epsilon\dot{D}_tD_t^{-1}Y_t + L_t\psi)\rangle + o(\epsilon) - h(Y_t)\right]\right)
\end{aligned}$$

Now observe that ϕ is a zero-centered Gaussian independent of everything else. Getting rid of the terms that or of the order $o(\epsilon)$, recalling $L_t = \sqrt{\epsilon} K_t + O(\epsilon)$, we are left with

$$\begin{aligned}&\lim_{\epsilon \longrightarrow 0} \frac{1}{\epsilon}\left(\mathop{\mathbb{E}}_{Y_t,\phi}\left[\epsilon\langle \nabla h(Y_t), \dot{D}_t D_t^{-1} Y_t\rangle + \langle \sqrt{\epsilon} K_t \phi, \nabla^2 h(Y_t), \sqrt{\epsilon} K_t \phi\rangle\right]\right)\\&= \mathop{\mathbb{E}}_{Y_t}\left[\langle \nabla h(Y_t), \dot{D}_t D_t^{-1} Y_t\rangle + \mathsf{tr}(\phi^T K_t \nabla^2 h(Y_t) K_t \phi)\right]\\&= \mathop{\mathbb{E}}_{Y_t}[\langle \nabla h(Y_t), \dot{D}_t D_t^{-1} Y_t\rangle + \langle \nabla^2 h(Y_t), K_t K_t^T\rangle]\end{aligned}$$

Hence, if we denote by $p(y,t)$ the density of Y_t, we conclude that the following PDE is satisfied for any h :

$$\frac{d}{dt}\int_{\mathbb{R}^d} h(y)p(y,t)dy = \int_{\mathbb{R}^d}(\langle \nabla h(y), \dot{D}_t D_t^{-1} y\rangle + \langle \nabla^2 h(y), K_t K_t^T\rangle)p(y,t)dy. \tag{3}$$

Using integration by parts twice, we conclude that

$$\begin{aligned}&\int_{\mathbb{R}^d} h(y)\frac{d}{dt}p(y,t)dy\\&= -\int h(y)\langle \dot{D}_t D_t^{-1}, \mathsf{J}_y[yp(y,t)]\rangle dy + \int h(y)\langle K_t K_t^T, \nabla_y^2[p(t,y)]\rangle dy\end{aligned} \tag{4}$$

Hence, the PDE satisfied by the evolution of density of Y_t is

$$\frac{d}{dt}p(y,t) = -\langle \dot{D}_t D_t^{-1}, \mathsf{J}_y[yp(y,t)]\rangle + \langle K_t K_t^T, \nabla_y^2[p(t,y)]\rangle.$$

Evolutin of SDE. We can perform an analogous analysis with the SDE. Namely, we have that

$$\begin{aligned}&\frac{d}{dt}\mathbb{E}[h(Z_t)]\\&= \mathbb{E}[\langle \nabla h(Z_t), dZ_t\rangle + \frac{1}{2}\langle dZ_t, \nabla^2 h(Z_t), dZ_t\rangle]\\&= \mathbb{E}[\langle h(Z_t), (\dot{D}_t D_t^{-1})Z_t\rangle]dt + \mathbb{E}[\frac{1}{2}\langle K_t \nabla^2 h(Z_t), K_t\rangle]dt\end{aligned} \tag{5}$$

where we used that $(dt)^2 = 0, dtdB_t = dB_t dt = 0, (dB_t)^2 = dt$. Hence, we reach the same condition as before for the density q of Z_t:

$$\int h(z)\frac{d}{dt}q(z,t)dz = \int \Big(\langle \nabla h(z), (\dot{D}_t D_t^{-1})z\rangle + \frac{1}{2}\langle \nabla^2 h(z), K_t K_t^T\Big)q(z,t)dt.$$

As p, q coincide at $t = 0$ and satisfy the same linear PDE, $p = q$ for any $z \in \mathbb{R}^d, t \in \mathbb{R}$. □

B.2 Proof of Proposition 2.2

Proposition B.2. *Suppose that for some $T > 0$ and any Gaussian distribution of X_0, there exists some process $(Z_t)_{0\le t\le T}$ which is Markovian and Z_t and $X_t = D_t X_0 + A_t \eta$ have the same marginal distribution for all $0 \le t \le T$. Then,*

$$\dot{A}_t A_t^T + A_t \dot{A}_t{}^T \succeq \dot{D}_t D_t^{-1} A_t A_t^T + A_t A_t^T D_t^{-T} \dot{D}_t{}^T.$$

Proof. Recall the high-level approach after Proposition 2.2.

Choice of distributions. More concretely, let $\theta \sim \mathcal{N}(0, K^2)$ for some symmetric matrix K and let $P_\theta = \mathcal{N}(\theta, L^2)$ for some other symmetric L. Then, we have

$$\theta = K\phi, \quad X_0 \sim \mathcal{N}(K\phi, L^2) = K\phi + L\xi, \quad Z_t \sim D_tX_0 + A_t\eta = D_tK\phi + D_tL\xi + A_t\eta,$$

where ϕ, ξ, η are iid samples from $\mathcal{N}(0, I)$. Hence, we have that

$$\begin{aligned}
&I(\theta; Z_t) \\
&= I(K\phi; D_tK\phi + D_tL\xi + A_t\eta) \\
&= h(D_tK\phi + D_tL\xi + A_t\eta) - h(D_tK\phi + D_tL\xi + A_t\eta | K\phi) \\
&= h(D_tK\phi + D_tL\xi + A_t\eta) - h(D_tL\xi + A_t\eta).
\end{aligned}$$

Using that the differential entropy of $\mathcal{N}(\mu, \Sigma)$ in d dimensions is $\frac{d}{2}\ln(2\pi) + \frac{1}{2}\ln\det\Sigma + \frac{d}{2}$, we obtain

$$I(\theta; Z_t) = \frac{1}{2}\ln\det(D_tK^2D_t^T + D_tL^2D_t^T + A_tA_t^T) - \frac{1}{2}\ln\det(D_tL^2D_t^T + A_tA_t^T).$$

These functions are clearly differentiable. Hence, for any K, L it is the case that

$$\begin{aligned}
&\frac{d}{dt}I(\theta; Z_t) \leq 0 \iff \\
&\frac{d}{dt}\ln\det(D_tK^2D_t^T + D_tL^2D_t^T + A_tA_t^T) \leq \frac{d}{dt}\ln\det(D_tL^2D_t^T + A_tA_t^T) \iff \\
&-\frac{d}{dt}\ln\det(D_t) + \frac{d}{dt}\ln\det(D_tK^2D_t^T + D_tL^2D_t^T + A_tA_t^T) - \frac{d}{dt}\ln\det(D_t^T) \leq \\
&\qquad -\frac{d}{dt}\ln\det(D_t) + \frac{d}{dt}\ln\det(D_tK^2D_t^T + D_tL^2D_t^T + A_tA_t^T) - \frac{d}{dt}\ln\det(D_t^T) \iff \\
&\frac{d}{dt}\ln\det(K^2 + L^2 + X_tX_t^T) \leq \frac{d}{dt}\ln\det(L^2 + X_tX_t^T)
\end{aligned} \tag{6}$$

where $X_t = D_t^{-1}A_t$. Now, we use the fact that $\frac{d}{dt}\ln\det M(t) = \langle M^{-1}(t), \frac{d}{dt}M(t)\rangle$ for any smooth symmetric full-rank matrix function $t \longrightarrow M(t)$. Hence, the above inequality implies that for any symmetric K, L such that $L^2 + X_tX_t^T$ is full rank,

$$\langle (K^2 + L^2 + X_tX_t^T)^{-1}, \dot{X}_tX_t^T + X_t\dot{X}_t{}^T\rangle \leq \langle (L^2 + X_tX_t^T)^{-1}, \dot{X}_tX_t^T + X_t\dot{X}_t{}^T\rangle$$

Choosing the free parameters. Now, choose L such that $L^2 + X_tX_t^T = \lambda I$ for any large positive λ, say $\lambda = 2\lambda_{\max}(X_tX_t^T)$. Let $K = \sqrt{\epsilon}v$ for some unit vector v and some small positive ϵ. Then, we have

$$\begin{aligned}
&\langle (\lambda I + \epsilon vv^T)^{-1}, \dot{X}_tX_t^T + X_t\dot{X}_t{}^T\rangle \leq \langle (\lambda I)^{-1}, \dot{X}_tX_t^T + X_t\dot{X}_t{}^T\rangle \iff \\
&\langle \frac{1}{\lambda}I - \frac{\epsilon}{\lambda(\lambda+\epsilon)}vv^T, \dot{X}_tX_t^T + X_t\dot{X}_t{}^T\rangle \leq \langle \frac{1}{\lambda}I, \dot{X}_tX_t^T + X_t\dot{X}_t{}^T\rangle \iff \quad 0 \leq \langle \epsilon vv^T, \dot{X}_tX_t^T + X_t\dot{X}_t{}^T\rangle.
\end{aligned}$$

Endgame. As this holds for any v, this is equivalent to $\dot{X}_tX_t^T + X_t\dot{X}_t{}^T \succeq 0$. Expanding this, we obtain

$$\begin{aligned}
&\dot{X}_tX_t^T + X_t\dot{X}_t{}^T \succeq 0 \iff \\
&(D_t^{-1}A_t)'(D_t^{-1}A_t)^T + (D_t^{-1}A_t)((D_t^{-1}A_t)^T)' \succeq 0 \iff \\
&(-D_t^{-1}\dot{D}_tD_t^{-1}A_t + D_t^{-1}\dot{A}_t)A_t^TD_t^{-T} + D_t^{-1}A_t(\dot{A}_t{}^TD_t^{-T} - A_t^TD_t^{-T}\dot{D}_t{}^TD_t^{-T}) \succeq 0 \iff \\
&D_t^{-1}\dot{A}_tA_t^TD_t^{-T} + D_t^{-1}A_t\dot{A}_t{}^TD_t^{-T} \succeq D_t^{-1}\dot{D}_tD_t^{-1}A_tA_t^TD_t^{-T} + D_t^{-1}A_tA_t^TD_t^{-T}\dot{D}_t{}^TD_t^{-T}.
\end{aligned}$$

Now, we can multiply by D_t on the left and D_t^T on the right and obtain

$$\dot{A}_t A_t^T + A_t \dot{A}_t^T \succeq \dot{D}_t D_t^{-1} A_t A_t^T + A_t A_t^T D_t^{-T} \dot{D}_t^T,$$

which finishes the proof. □

C Consistency proof for diagonal non-uniform processes.

C.1 Notations

Throughout this section, for $u, v \in R^d$, by $u \cdot v$ we mean element-wise multiplication and for a real $\alpha \in R$, by u^α we mean element-wise power operation. For a function $F : R^d \to R^d$, denote by $J_F \in R^{d \times d}$ the Jakobian $(J_F)_{i,j} = \frac{\partial F_i}{\partial x_j}$ and by $\Delta^2 F \in R^{d \times d}$, the matrix where $(\Delta^2 F)_{i,j} = \frac{\partial F_i}{\partial x_j^2}$. Additionally, letters with an overline are used to denote vector-valued functions, e.g., $\bar{g}(t) : R \to R^d$. Finally, consider our main SDE, given as follows:

$$dX_t = \bar{g}(t) dB_t, x_0 \sim p_0. \tag{7}$$

C.2 Preliminaries

Consider the following backward SDE:

$$dX_t = -\bar{g}(t)^2 \cdot \bar{\sigma}(t)^{-2} \cdot (h(X_t, t) - X_t)\, dt + \bar{g}(t) \cdot d\bar{B}_t \tag{h-SDE}$$

Property 1 (**Consistency**). A function $h : R^d \times [0, 1] \to R^d$, is said to be *consistent* iff for all $t \in [0, 1)$ and $x \in R^d$, we have:

$$h(x, t) = \mathbb{E}_h[X_0 | X_t = x].$$

Property 2 (**Conservative vector field**). Let $h : R^d \times [0, 1] \to R^d$. We say h induces a conservative vector field if for any $t \in [0, 1]$, there exists some probability density $p(\cdot, t)$ such that:

$$s(x, t) = \bar{g}(t)^{-2} (h(x, t) - x) = \nabla \log p(x, t).$$

Now, we will show that if our score satisfies both Property 1 and Property 2, then it agrees with the score induced from p. Furthermore, if it agrees with the true score in an open set, than it agrees everywhere. The proof will be very similar to [DDDD23], with the difference that instead of a scalar noising model, we have a vector values noising model.

Remark 2. Note that our theory doesn't exactly correspond to our implementation. The reason is that for implementation, the SDE is run in the frequency space, and the score is estimated in the pixel space. However, currently our theory entirely focuses on operating in the transformed frequency space

$$\tilde{Y}_t = \frac{Y_t}{\Lambda_t}.$$

C.3 Proof development

Lemma C.1. *Suppose the function h satisfies the consistency property. Than s defined as:*

$$s(x,t) = \bar{\sigma}(t)^{-2} \cdot (h(x,t) - x)\,,$$

satisfies the following PDE:

$$\frac{\partial s}{\partial t} - J_s\left(\bar{g}(t)^2 \cdot s\right) - \frac{1}{2}\left(\Delta^2 s\right)\bar{g}(t)^2.$$

Proof. By definition, we have:

$$dX_t = -\bar{g}(t)^2 \cdot s(X_t, t)dt + \bar{g}(t) \cdot d\bar{B}_t.$$

Now, applying multi-dimensional Ito's formula for the function h, gives us:

$$dh(X_t,t) = \left(\frac{\partial h}{\partial t} - J_h\left(\bar{g}(t)^2 \cdot s\right) - \frac{1}{2}(\Delta^2 h)\bar{g}(t)^2\right)dt + J_h\left(\bar{g}(t) \cdot d\bar{B}_t\right).$$

Since from Property 1 it follows that h is a reverse martingale and hence the dt part needs to be 0, from which we obtain:

$$\frac{\partial h}{\partial t} - J_h\left(\bar{g}(t)^2 \cdot s\right) - \frac{1}{2}(\Delta^2 h)\bar{g}(t)^2 \ = 0. \tag{8}$$

Now, remember that by definition we have:

$$h = x + \bar{\sigma}(t)^2 \cdot s.$$

Therefore, putting this into 8, gives us:

$$\begin{aligned}
0 &= \frac{\partial}{\partial t}(x + \bar{\sigma}(t)^2 \cdot s) - J_{x+\bar{\sigma}(t)^2 \cdot s}\left(\bar{g}(t)^2 \cdot s\right) - \frac{1}{2}\left(\Delta^2(x + \bar{\sigma}(t)^2 \cdot s)\right)\bar{g}(t)^2 \\
&= \frac{d}{dt}\bar{\sigma}(t)^2 \cdot s + \bar{\sigma}(t)^2 \cdot \frac{\partial}{\partial t}s - \left(I + \mathrm{diag}(\bar{\sigma}(t)^2)J_s\right)\left(\bar{g}(t)^2 \cdot s\right) - \frac{1}{2}\left(\mathrm{diag}(\bar{\sigma}(t)^2)(\Delta^2 s)\bar{g}(t)^2\right) \\
&= \bar{g}(t)^2 \cdot s + \bar{\sigma}(t)^2 \cdot \frac{\partial}{\partial t}s - \bar{g}(t)^2 \cdot s - \mathrm{diag}(\bar{\sigma}(t)^2)J_s(\bar{g}(t)^2 \cdot s) - \frac{1}{2}\left(\mathrm{diag}(\bar{\sigma}(t)^2)(\Delta^2 s)\bar{g}(t)^2\right) \\
&= \bar{\sigma}(t)^2 \cdot \frac{\partial}{\partial t}s - \bar{\sigma}(t)^2 \cdot J_s(\bar{g}(t)^2 \cdot s) - \frac{1}{2}\bar{\sigma}(t)^2 \cdot (\Delta^2 s)\bar{g}(t)^2.
\end{aligned}$$

Dividing by $\bar{\sigma}(t)$ element-wise finishes the proof.

□

Now, using Property 2, there is a probability field $p(\cdot, t)$, such that we have:

$$s(x,t) = \bar{g}(t)^{-2}\left(h(x,t) - x\right) = \log p(x,t),$$

for any $t \in [0,1]$. Now denote by $p'(x,t)$ the probability density function defined by 7 with $p_0 = p(\cdot, 0)$. Define $s'(x,t) = \log p'(x,t)$. Than, since the real score is self-consistent, it follows that s' also satisfies:

$$\frac{\partial s}{\partial t} - J_s\left(\bar{g}(t)^2 \cdot s\right) - \frac{1}{2}\left(\Delta^2 s\right)\bar{g}(t)^2 = 0.$$

Since we have that $s'(\cdot, 0) = s(\cdot, 0)$, than by the uniqueness of the PDE solution (`[should be true]`), we mush that $s(\cdot, t) = s'(\cdot, t)$ for any $t \in [0, 1]$. Hence, we conclude that $h(\cdot, t) = h'(\cdot, t)$ for any $t \in [0, 1]$.

The second part of the analogous theorem from [DDDD23], follows similarly by showing that for a fixed t_0, $p(\cdot, t_0)$ is an analystical function.

D Proof of approximation bound, Theorem 3.1

Theorem D.1. *Let $(X_t)_{t\in[0,T]}$ satisfy the true reverse SDE*

$$\mathrm{d}X_t = f(t, X_t)\,\mathrm{d}t + \Lambda_t\, g_t\, \mathrm{d}B_t, \qquad X_T \sim \mathcal{N}(0, I),$$

with

$$f(t, x) = -\Lambda_t^2\, g_t^2\, \nabla_x s(t, x).$$

Let $(\widehat{X}_t)_{t\in[0,T]}$ satisfy the learned reverse SDE

$$\mathrm{d}\widehat{X}_t = \widehat{f}(t, \widehat{X}_t)\,\mathrm{d}t + \Lambda_t\, g_t\, \mathrm{d}B_t, \qquad \widehat{X}_T \sim \mathcal{N}(0, I),$$

with

$$\widehat{f}(t, x) = -\Lambda_t^2\, g_t^2\, \widehat{s}(t, x).$$

Suppose that for all $t \in [0, T]$ the following hold:

1. *Lipschitz in space:*
$$\|f(t, x) - f(t, y)\| \;\le\; L\,\|x - y\| \quad \forall\, x, y.$$
2. *Uniform noise bound:*
$$\|\Lambda_t\, g_t\| \;\le\; \|\Lambda\|_\infty\, \|g\|_\infty =: M^{\frac{1}{2}}.$$
3. *Score-estimation error:*
$$\mathbb{E}_{x_t\sim p_t}\big[\|s(t, x_t) - \widehat{s}(t, x_t)\|^2\big] \;\le\; \varepsilon^2.$$
(By consistency the same holds under $\widehat{p}_t$.)

Then the Wasserstein-2 distance between the marginals at time 0 satisfies

$$W_2\big(\widehat{P}_0,\, P_0\big) \;\le\; \frac{e^{LT} - 1}{L}\, M\, \varepsilon.$$

Proof. We begin by coupling the true and learned SDEs using the same Brownian motion. Namely, first draw

$$X_T \sim \mathcal{N}(0, I)\,,$$

then set $\widehat{X}_T := X_T$, and fix a Brownian motion $(B_t)_{t\in[0,T]}$. Then evolve in reverse time from T to 0:

$$\mathrm{d}X_t = f(t, X_t)\,\mathrm{d}t + \Lambda_t\, g_t\, \mathrm{d}B_t, \tag{9}$$

$$\mathrm{d}\widehat{X}_t = \widehat{f}(t, \widehat{X}_t)\,\mathrm{d}t + \Lambda_t\, g_t\, \mathrm{d}B_t. \tag{10}$$

Therefore

$$\mathrm{d}X_t - \mathrm{d}\widehat{X}_t = \big(f(t, X_t) - \widehat{f}(t, \widehat{X}_t)\big)\,\mathrm{d}t,$$

and we set

$$Y_t := X_t - \widehat{X}_t.$$

$$\mathrm{d}Y_t \;=\; \big(f(t,X_t) - \widehat{f}(t,\widehat{X}_t)\big)\,\mathrm{d}t, \qquad Y_T = 0.$$

Now let $F(y) = \|y\|^2$, so that

$$\nabla F(y) = 2y, \quad \nabla^2 F(y) = 2I.$$

By Itô's formula,

$$\mathrm{d}\,\|Y_t\|^2 = 2\big\langle Y_t,\; f(t,X_t) - \widehat{f}(t,\widehat{X}_t)\big\rangle\,\mathrm{d}t. \tag{*}$$

Next define the "score-error" field

$$\Delta_f(t,x) \;:=\; f(t,x) - f(t,\widehat{x}) \;=\; \Lambda_t^2\, g_t^2\big[\widehat{s}(t,x) - s(t,x)\big].$$

Then we decompose

$$\begin{aligned} f(t,X_t) - \widehat{f}(t,\widehat{X}_t) &= \big(f(t,X_t) - f(t,\widehat{X}_t)\big) + \big(f(t,\widehat{X}_t) - \widehat{f}(t,\widehat{X}_t)\big) \\ &= f(t,X_t) - f(t,\widehat{X}_t) \;+\; \Delta_f(t,\widehat{X}_t). \end{aligned}$$

In this setting, we can think of the difference between the first two terms being the drift due to $X_t - \widehat{X}_t$, then $\Delta_f(t,\widehat{X}_t)$ measures the model error at $\widehat{X}_t$.

We can now use the L-smooth assumption on $\|Y_t\|$, mainly

$$\|f(t,X_t) - f(t,\widehat{X}_t)\| \;\le\; L\|X_t - \widehat{X}_t\| \;=\; L\|Y_t\|.$$

Now by $(*)$

$$\frac{\mathrm{d}}{\mathrm{d}t}\,\|Y_t\| = \frac{\langle Y_t,\, \mathrm{d}Y_t\rangle}{\|Y_t\|} \;\le\; \|\mathrm{d}Y_t\| \quad \text{by Cauchy–Schwarz.}$$

Thus

$$\frac{\mathrm{d}}{\mathrm{d}t}\,\|Y_t\| \;\le\; L\|Y_t\| \;+\; \|\Delta_f(t,\widehat{X}_t)\|.$$

for convenience, define $u(t) := \|Y_t\|$ & $g(t) := \|\Delta_f(t,\widehat{X}_t)\|$

so from before we have

$$u'(t) \le L\,u(t) + g(t)$$

with the condition $u(T) = 0$. We now seek a bound on $u(0)$.

For this, recall Gronwall's Inequality, which states that if

$$u'(t) \le L\,u(t) + g(t), \quad u(0) = u_0,$$

then $\forall t \ge 0$

$$u(t) \le e^{Lt}u_0 + \int_0^t e^{L(t-s)} g(s)\,\mathrm{d}s. \tag{3}$$

But note that our initial condition $u(T) := 0$, so we simply run backwards in time. Mainly, set $\tau := T - t$ then as $t \to T$ down to 0, $\tau \to 0$ up to T.

So $\forall t \geq 0$

$$u(t) \leq e^{L(t-T)} u(T) + \int_T^t e^{L(t-s)} g(s)\,\mathrm{d}s.$$

using $u(T) = 0$, we obtain

$$u(0) = \|Y_0\| \leq \int_T^0 e^{L(0-s)}\, g(s)\,\mathrm{d}s = \int_0^T e^{L(T-s)}\, g(s)\,\mathrm{d}s = \int_0^T e^{L(T-s)} \|\Delta_f(s, \widehat{X}_s)\|\,\mathrm{d}s.$$

Squaring and taking expectations, then applying Cauchy–Schwarz in time,

$$\mathbb{E}\Big[\Big(\int_0^T e^{L(T-s)} g(s)\,\mathrm{d}s\Big)^2\Big] \;\leq\; \Big(\int_0^T e^{L(T-s)}\,\mathrm{d}s\Big) \int_0^T e^{L(T-s)}\; \mathbb{E}\big[g(s)^2\big]\,\mathrm{d}s.$$

Hence

$$\mathbb{E}\,\|Y_0\|^2 \;\leq\; \Big(\int_0^T e^{L(T-s)}\,\mathrm{d}s\Big) \int_0^T e^{L(T-s)}\; \mathbb{E}\,\|\Delta_f(s, \widehat{X}_s)\|^2\,\mathrm{d}s.$$

By A2–A3, $\|\Delta_f(s, \widehat{X}_s)\| \leq \|\Lambda_s g_s\|^2\, \|\widehat{s} - s\| \leq M\, \|\widehat{s} - s\|$ and $\mathbb{E}\,\|\widehat{s} - s\|^2 \leq \varepsilon^2$, so $\mathbb{E}\,\|\Delta_f\|^2 \leq M^2 \varepsilon^2$. Noting $\int_0^T e^{L(T-s)}\,\mathrm{d}s = (e^{LT} - 1)/L$, we get

$$\mathbb{E}\,\|Y_0\|^2 \;\leq\; M^2\, \varepsilon^2 \Big(\tfrac{e^{LT}-1}{L}\Big)^2.$$

so

$$\mathbb{E}\|X_0 - \widehat{X}_0\|^2 \leq M^2\, \varepsilon^2 \Big(\frac{e^{LT} - 1}{L}\Big)^2$$

We connect this to $W_2(\widehat{P}_0, P_0)$. Recall that

$$W_2(\widehat{P}_0, P_0) = \inf_{\pi \in \Pi(\widehat{P}_0, P_0)} \Big(\int \|x - y\|^2\, \mathrm{d}\pi(x, y)\Big)^{1/2}.$$

Thus for any coupling π of $\widehat{P}_0$ and P_0 we get

$$W_2(\widehat{P}_0, P_0) \;\leq\; \Big(\int \|x - y\|^2\, \mathrm{d}\pi(x, y)\Big)^{1/2} = \mathbb{E}\big[\|\widehat{X}_0 - X_0\|^2\big]^{1/2} = \mathbb{E}\big[\|Y_0\|^2\big]^{1/2}.$$

So

$$W_2(\widehat{P}_0, P_0) \;\leq\; \big(\mathbb{E}\|Y_0\|^2\big)^{1/2} = M\, \Big(\frac{e^{LT} - 1}{L}\Big)\, \varepsilon$$

□

E Question to discuss

Technically we can put a consistency on our desired form of SDE. Which one should we consider.

- For CIFAR is the frequency matrix 32x32 or 3x32x32.
- need to check if model timesteps is int or float