Conditioning Generative Diffusion Models

Training-free and Asymptotically Consistent

V.R. Bockstael

Technische Universiteit Delft

TU Delft

Conditioning Generative Diffusion Models

Training-free and Asymptotically Consistent

V.R. Bockstael

Abstract Generative diffusion is a machine learning technique to generate high-quality samples from complex data distributions. Much of its success can be attributed to the recently developed techniques that flexibly control the data generation process, without additional training effort. These methods control a pre-trained diffusion model towards specific regions of interest, which are determined by external information such as class labels, masks, or text descriptions. However, these approaches are typically based on heuristic guidance techniques and break the consistency on which the theoretical justification of generative diffusion relies. This is problematic when applying these controlled data generation techniques to tasks that are sensitive to distribution characteristics rather than the perceptual quality of individual samples. To this end, we introduce an asymptotically consistent approach for conditioning generative diffusion models without retraining the entire system. We use an importance sampling technique for simulating diffusion bridges, where multiple draws of a guided proposal process are reweighted to resemble paths of the true conditioned denoising process. A theoretical analysis shows that under certain assumptions, our approach has a vanishing error. In an empirical analysis, we find that specific nuances to the performance trade-off appear with a finite amount of computational effort. Specifically, the effectiveness of our approach highly depends on the choice of the proposal process and the allocation of computational effort towards independent runs of our algorithm.

to obtain the degree of Master of Science
at the Delft University of Technology,
to be defended publicly on Monday, May 19, 2025 at 12:00.

Student number:	4590694	
Project duration:	October 1, 2024 – May 19, 2025	
Thesis committee:	dr. ir. J. Bierkens (Supervisor)	TU Delft
	dr. A. Boer	Ortec Finance
	dr. R. Kraaij	TU Delft

This thesis is confidential and cannot be made public until May 19, 2025.

An electronic version of this thesis is available at http://repository.tudelft.nl/.

TU Delft

Contents

1 Introduction

Generative diffusion models have recently appeared as a successful technique to generate high-quality samples from complex high-dimensional data distributions [SE20; HJA20; Cao+23; CKS23]. Notable examples are image generators Dall-E [Ram+21] and Stable Diffusion [Rom+22]. Generative diffusion stands out from deep generative modeling techniques [Ben23] due to its stability and qualitative performance. This is in contrast to variational auto encoders, which are known to have difficulty producing realistic data, or generative adversarial networks, which are typically unstable to train [XKV22].

Successful adoption in domains such as the generation of images [Ram+21; Rom+22; Rad+21], video [Ho+22; Xin+24], sound [Zha+23], and time series [YQ24; SRH24; Yan+24] is largely attributed to the ability to control the generation of the data towards regions of interest. Examples are text-to-image and inpainting tasks. By controlling the generative process, the data distribution can be thoroughly explored through various conditioned distributions and yields a necessary statistical tool for risk-sensitive application areas.

This is particularly important given the uncertainty about the ability of deep generative models to capture the heavy-tailed nature of real-world data effectively [TD25; SSD24; Pan+24]. Therefore, controlled generative diffusion offers a promising alternative by reframing the assessment of heavy-tailed distributions as a rare event simulation task, thereby potentially circumventing the inherent architectural and data limitations of deep generative models. Unfortunately, effectively controlling generative diffusion remains an open challenge. Existing methods tend to rely either on computationally expensive training-based techniques, or, on cheap training-free techniques that can introduce significant inconsistencies with the true conditional distribution.

The expensive conditional generative diffusion methods are consistent, in the sense that they approximate the true conditional distribution, because they are obtained by training the entire system to learn the joint distribution of the data and the additional information [DN21; HS22]. The theoretical guarantees of generative diffusion models [Che+23] directly transfer to this class of methods. However, they require the availability of the additional information at training-time, which is not always the case, and are inflexible to adaptations of the system requirements.

The cheap conditional generative diffusion methods are inconsistent because they are based on heuristic adaptations of a pre-trained unconditional generative diffusion model [He+24; Son+23; Yu+23; Chu+24; Ye+24; Col+23; Lug+22; SE20]. The benefit is that the approaches are highly flexible to adaptations of the system requirements and do not require additional (re)training.

In application areas such as finance [FJ22; Büh+20], renewable [Jia+19; Li+24; Che+18], and weather forecasting [Pri+24; Hua+24], the statistical tasks are often characterized by the modeling of their distributions, rather than individual samples. This is in contrast to image or sound generation, where one is often more interested in the perceptual quality of individual samples instead. Applying generative diffusion to risk-sensitive distribution-sensitive areas introduces new consistency requirements for generative diffusion models.

1.1. Generative Diffusion Models

Generative diffusion is based on a noising procedure that reduces complicated structural information to a trivial state. Essentially, this transfers the unknown distribution of an observed data set, from which it is impossible to sample, to a known distribution, from which it is easy to sample.

The noising process is a stochastic process $(X_t)_{t=0}^T$ on $\mathbb{R}^d$, where $X_0 \sim \mathbb{P}_{X_0}$ represents a sample from the unknown data distribution and $X_T \sim \mathbb{P}_{X_T}$ represents a sample from the known and trivial distribution, that is obtained by adding a sufficiently large amount of noise to the data. The task of generative diffusion is to learn the dynamics of the reverse process, i.e. the denoising process, such that we can sample from $\mathbb{P}_{X_0}$ by running through the reverse process starting at a sample $X_T \sim \mathbb{P}_{X_T}$. This is done by simulating the noising process and simultaneously training a neural network that acts as a denoiser. The noising process can be described by the following:

$$(\textbf{Noising process}) \qquad \mathbb{P}_{X_0} \sim X_0 \overset{\text{noise}}{\to} \dots \overset{\text{noise}}{\to} X_t \overset{\text{noise}}{\to} \dots \overset{\text{noise}}{\to} X_T \sim \mathbb{P}_{X_T}.$$

Generative diffusion models produce high-quality results by leveraging the flexibility and expressiveness of neural networks in a sophisticated way. At a high level, learning the denoising process can be framed as training a neural network, denoted by $\hat{b}_t$ and parameterized by $t \in [0, T)$, to recover the process state X_t from a slightly more perturbed future state X_s, where $s > t$ is typically close to t. This is formalized as minimizing a loss function, commonly the mean squared error,

$$\hat{b}_t = \arg\min_{b_t} \text{Loss}\left(b_t(X_s), X_t\right). \tag{1.1}$$

Then the procedure to generate realistic samples from the data distribution $\mathbb{P}_{X_0}$ amounts to sampling the known distribution $\mathbb{P}_{X_T}$ and performing the following iterative procedure:

$$\text{sample} \quad X_T \sim \mathbb{P}_{X_T} \quad \text{and compute} \quad X_t = \hat{b}_t(X_s) + Z_t, \text{ with } Z_t \sim \mathcal{N}(0, |s-t| I_{d\times d}).$$

Then, we obtain the following chain that describes the reverse process, where the denoise operation is described in the equation above,

$$(\textbf{Denoising process}) \qquad \mathbb{P}_{X_0} \sim X_0 \overset{\text{denoise}}{\leftarrow} \dots \overset{\text{denoise}}{\leftarrow} X_t \overset{\text{denoise}}{\leftarrow} \dots \overset{\text{denoise}}{\leftarrow} X_T \sim \mathbb{P}_{X_T}.$$

In Figure 1.1, an illustration of the noising and denoising process is given. The overall ability to generate high-quality samples from the data distribution depends on the expressive abilities of the neural network and the number of time steps. However, obtaining the denoising process is easier said than done and relies on the neural network's architectural choices, data availability, and the amount of training. These factors make generative diffusion a generally expensive method, both during training and sampling, due to the multiple denoising passes required for each generated sample.

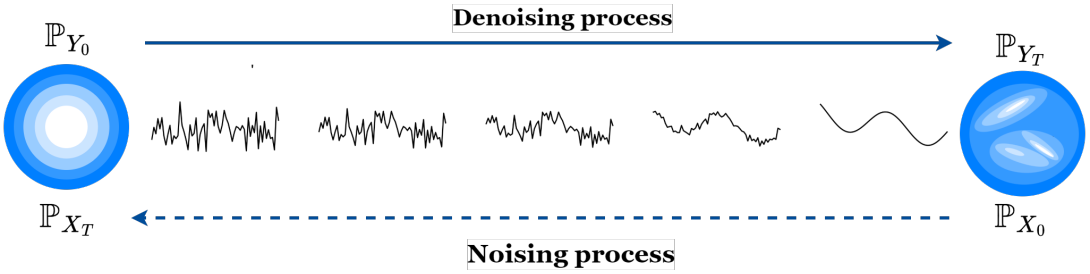


Figure 1.1: Horizontal illustration of generative diffusion. The diagram illustrates the perturbation of a clean signal. This denoising process is depicted by the arrow from left to right. The noising process is depicted by the arrow from right to left. The left side depicts the state of a trivial state drawn from $\mathbb{P}_{X_T}$ (or equivalently $\mathbb{P}_{Y_0}$). Moving towards the right, the sample is denoised iteratively until it resembles a sample drawn from $\mathbb{P}_{X_0}$ (or equivalently $\mathbb{P}_{Y_T}$), depicted on the right-hand side.

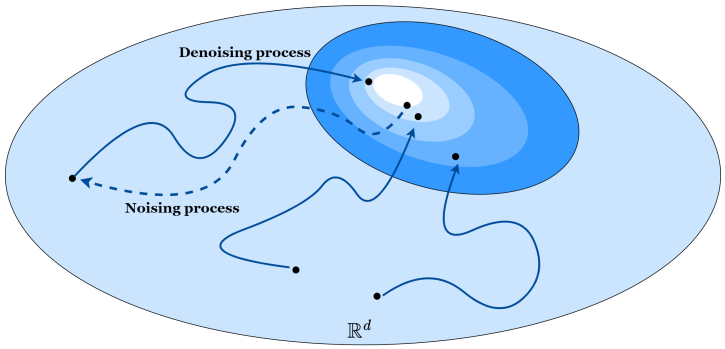


Figure 1.2: Illustration of generative diffusion in $\mathbb{R}^d$. The diagram illustrates the generative diffusion in the state space ($\mathbb{R}^d$), starting from some point sampled with $\mathbb{P}_{Y_0}$ and ending up as being distributed by approximately the data distribution.

Our work is based on the continuous time framework, which is due to Song et al. in 2021, [Son+21] and generalizes the discrete time framework as introduced by Ho et al. in 2020, [HJA20]. The underlying theory of generative diffusion can be established through continuous-time reasoning, which enables a coherent and flexible notational framework that has roots in many existing fields of mathematics and physics. Conceptually, a continuous-time framework is obtained by taking infinitesimally small steps of the above-described discrete-time operation. Then, informally, we can obtain two stochastic differential equations (SDE) that describe the noising process and the denoising process:

$$\begin{aligned} &\text{(Noising SDE)} && \mathrm{d}\,X_t = \mathrm{d}\bar{B}_t, && X_0 \sim \mathbb{P}_{X_0}, \\ &\text{(Denoising SDE)} && \mathrm{d}\,Y_t = \nabla \log p_{X_{T-t}}(Y_t)\mathrm{d}t + \mathrm{d}B_t, && Y_0 \sim \mathbb{P}_{X_T}, \end{aligned}$$

where p_{X_t} is the marginal density of process X at time t. The term $\nabla \log p_{X_{T-t}}$ is often referred to as the score function, specifically of the marginal distribution. The denoising SDE essentially describes a process that satisfies $Y_0 \sim \mathbb{P}_{X_T}$ and $Y_T \sim \mathbb{P}_{X_0}$, and is derived through a classic result on the time-reversal of SDEs by Anderson in 1982 [And82]. In Figure 1.2, an abstract illustration of generative diffusion is given, which we also use later in this chapter to describe conditioning. The marginal density is not known as it ultimately depends on the unknown data distribution $\mathbb{P}_{X_0}$. Therefore, the score function must be learned with a neural network, which is precisely the role of $\hat{b}_t$ in Equation 1.1. A compelling argument for using generative diffusion models is that, given an accurate enough neural network approximation of the score function, theoretical guarantees of the method can be derived with respect to the total variation between the approximated distribution $\mathbb{P}_{Y_T}$ and the true data distribution $\mathbb{P}_{X_0}$ [Che+23].

1.2. Control by Guidance or Conditioning

Now, the goal of controlling the generative diffusions is to approximate a conditional distribution instead of an unconditional one. Say we want to be able to choose some subset $A \subseteq \mathbb{R}^d$ at sample-time and sample from $\mathbb{P}_{X_0}$ conditioned on $X_0 \in A$. For example, we may consider X_0 to satisfy a conditioning on a linear transformation, i.e. $A = \{x \in \mathbb{R}^d : Lx = v\}$, for some $m \times d$ full rank matrix L and vector $v \in \mathbb{R}^d$. Two classes of controlled generation methods can be identified in the literature.

The first class is based on conditioning at training time, which has the convenience that the true conditional distribution is learned, to the extent that the class of neural network approximators limits it. In principle, conditioning at training time is achieved by learning the joint distribution of the data and the additional information. Training-based conditioning is generally not desirable as it requires additional information to be available at training-time and therefore hinders flexibility.

The second class avoids these disadvantages by heuristically guiding the denoising process at sampling-time. The convenience of the training-free method comes at the cost of an accurate approximation of the unconditional distribution, making it problematic in applications where consistency is highly important. For the typical application of generative diffusion, such as image, sound, or video generation, one is often interested in the perceptual quality of individual samples, rather than an accurate representation of the entire data distribution. This emphasis on the realism of the individual samples creates a fundamental mismatch between the state-of-the-art generative diffusion techniques and the domains where the integrity of the distribution is critical.

1.2.1. Conditioning at Training-time

We denote the conditional data distribution by $\mathbb{P}(X_t \in \cdot | X_0 \in A)$ for all $t \in [0, T]$. If we have sufficient available information about event $\{X_0 \in A\}$, we can learn to sample from the conditional distribution through learning the score of the marginal distribution of the conditioned process at training time, by simply appending the information about the set A to the input of the neural network. This way, the denoising process may be simulated for any appropriate choice of A.

This approach is often infeasible, as we often need the condition information to be available or extractable in the dataset, and more importantly, we need the format of the condition to be characterized prior to training the system. Therefore, to add some flexibility to the framework, instead of learning the joint distribution, [DN21] described a procedure based on training a classifier. The approach is due to an application of Bayes' theorem to the marginal density of the process X conditioned on $X_0 \in A$, i.e.,

$$p_{X_t|X_0 \in A}(x) = \frac{\mathbb{P}(X_0 \in A | X_t = x) p_{X_t}(x)}{\mathbb{P}_{X_0}(A)},$$

then it is not hard to see that the conditional score satisfies

$$\underbrace{\nabla \log p_{X_t|X_0 \in A}(x)}_{\text{conditional score}} = \nabla \log \underbrace{\mathbb{P}(X_0 \in A | X_t = x)}_{\text{classifier}} + \underbrace{\nabla \log p_{X_t}(x)}_{\text{unconditional score}} . \tag{1.2}$$

From this, we see that we can use a classifier capable of predicting the probability of event $\{X_0 \in A\}$ given that $X_t = x$ to flexibly turn an unconditional diffusion model into a conditional one. However, this procedure still requires training the classifier alongside the entire diffusion model. Note that an existing classifier trained on the clean data from data distribution $\mathbb{P}_{X_0}$ can not be used effectively, on which we will elaborate in the following subsection.

1.2.2. Guidance at Sampling-time

Alternative methods to control generative diffusion have been introduced to specifically omit the need for (re)training. The basis for these training-free control methods lies in approximating the classifier term in Equation 1.2. Specifically, this is done by leveraging the information that can be obtained from the classifier term at time 0 to guide the denoising process in the direction of satisfying the condition.

Typically, it is easy to determine the event given a true (unperturbed) data sample X_0. Indeed, the score of the marginal distribution of process X at time 0 is based on the term $\mathbb{P}(X_0 \in A | X_0 = x)$. This can often be determined deterministically, by computing the probability directly, or by using an existing classifier module trained on the data. Clearly the same does not hold for the conditional score at time t, which relies on $\mathbb{P}(X_0 \in A | X_t = x)$ as seen in Equation 1.2 and is inherently difficult to determine due to the complicated dynamics of the denoising process.

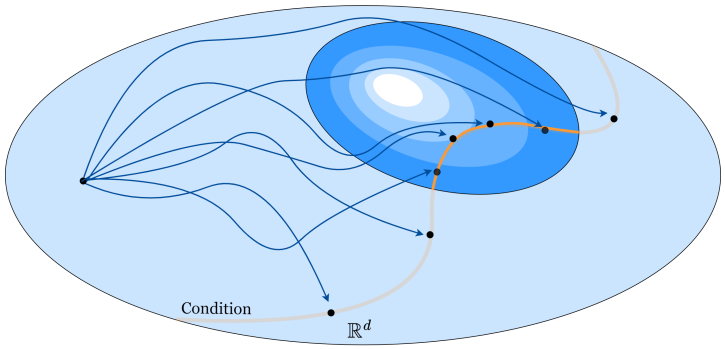


Figure 1.3: Illustration of heuristic guidance of generative diffusion at sampling-time. The orange curve indicates the subspace of $\mathbb{R}^d$ where the condition is satisfied. The blue arrows indicate the guided denoising SDEs that are guaranteed to satisfy the condition at their terminal time, albeit at a mismatch with the true conditioned distribution.

Two foundational approximations of the above equation are replacement guidance [Lug+22], which is based on an approximation of a perturbed state of the condition at time t, and reconstruction guidance [SE20], which is based on an approximation of the clean state X_0 given the perturbed state X_t. The replacement guidance method is typically used to condition on masked data, such as in forecasting and imputation of time series, and image inpainting. In contrast, the reconstruction guidance typically has a more general application area, such as in [Col+23], where the approach is applied to constrained scenario generation. Both these approaches have been extended to more generalized and more effective forms.

For example, Diffusion Posterior Sampling (DPS) [Chu+24] is an approach that is aimed at solving general inverse problems with a generative diffusion model as a prior. It is based on reconstruction guidance and assumes Gaussian or Poisson noise measurements of the reconstruction of the clean state given the noisy state. FreeDoM [Yu+23] improves upon the basic reconstruction guidance approach by applying an efficient denoising strategy that decomposes the entire time-span $[0, T]$ into three separate stages, where only the middle stage is heuristically considered to be important for guiding the diffusion towards satisfying the condition, while minimizing undesirable artifacts of the inconsistencies. Manifold-Preserving Guided Diffusion (MPGD) [He+24] aims to correct inconsistencies of the guidance techniques by making projections of the noisy state at each time step, such that some artifacts are also minimized. In [Ye+24], the authors provide a framework that unifies some of these existing approaches under a collective algorithmic formulation.

While practically impressive, all of the above-specified training-free methods are based on consistency-breaking approximations of the conditional score function. The approaches are often capable of producing high-quality individual samples, but they typically do not reproduce an accurate approximation of the conditional distribution, which remains an open question. The inconsistency of these training-free methods has not gone unnoticed [She+24]. Recent attempts to correct the inconsistency in guided generative diffusion have been introduced [Wu+24; Tri+23]. These methods are based on drawing sample steps from a proposal process, and correcting the error introduced by the approximations through a reweighing with importance weights. These approaches to conditional generative diffusion sampling shed a new light on the problem, as in contrast to the other heuristics, they satisfy the asymptotic precision.

1.3. Research and Contributions

In this work, we study the problem of conditioning generative diffusion models with a training-free and asymptotically consistent approach. This means that for an infinite amount of computational effort our approach should converge to the exact conditional distribution. We assume that we have access to an arbitrarily accurate (unconditional) denoising process Y and study the simulation of the conditioned denoising process. Specifically, we study the conditioning on a linear transformation of the data. Below, we state a formal description of our objective.

Definition 1.1 (Problem Statement)**.** *Let us fix $y_0 \in \mathbb{R}^d$. Consider the denoising process to be driven by the following SDE:*

$$\mathrm{d}Y_t = \underbrace{[-\alpha(t)Y_t + \nabla \log p_{X_t}(Y_t)]}_{b(t,Y_t)}\mathrm{d}t + \sigma(t)\mathrm{d}B_t, \qquad Y_0 = y_0,$$

where the solution Y to this SDE is assumed to exist uniquely. $\alpha : [0,T] \to (-\infty, 0]$ and $\sigma : [0,T] \to (0,\infty)$ are scalar functions that characterize the noise schedules of the generative diffusion model. Let $L \in \mathbb{R}^{m\times d}$ be a full rank matrix with $d > m$ and vector $v \in \mathbb{R}^m$. The objective is to sample Y_T conditioned on

$$LY_T = v. \tag{1.3}$$

The conditioned path measure is denoted by

$$\mathbb{P}^*_Y(\cdot) \overset{\text{def}}{=} \mathbb{P}(Y \in \cdot | LY_T = v).$$

In practice, the generative diffusion model is trained with a collection of samples $\{X_0^{(i)}\}_{i=1}^n \subset \mathbb{R}^d$ from the data distribution $\mathbb{P}_{X_0}$, to obtain an approximate score function. For simplicity, we assume that the score function is obtained exactly. If this is not the case, the results of this work translate nonetheless, despite having its inaccuracies of the score superimposed on that of the exact setting.

Conditioning SDEs in general is a widely studied area. In particular, the canonical Doob's h-transform allows the exact conditioning of Markov processes, e.g., see [PR02]. A well-known example of a conditioned process is a Brownian bridge, which essentially is a Brownian motion that is defined to hit a specific value at time T and can be derived by making use of Doob's h-transform, as we discuss in Chapter 3. Unfortunately, the ability to perform this explicit conditioning is relatively unique.

The SDE that drives the conditioned process is obtained by a change of measure induced by Doob's h-transform. In essence, this enables us to instead consider a different stochastic process $Y^* = (Y_t^*)_{t=0}^T$ that is governed by the following SDE:

$$\mathrm{d}Y_t^* = \left[b(t, Y_t^*) + \sigma^2(t)\nabla \log h(t, Y_t^*)\right]\mathrm{d}t + \sigma(t)\mathrm{d}B_t, \qquad Y_0 \sim \mathbb{P}_{Y_0}, \tag{1.4}$$

where h is the marginal density of the distribution $\mathbb{P}(LY_T \in \cdot | Y_t = x)$ evaluated at v. The change of measure allows us to interchange the conditioned measure $\mathbb{P}^*_Y$, which is conceptually difficult with $\mathbb{P}_{Y^*}$, which is conceptually simple as samples can be obtained by simulating the process driven by the above SDE. The term $\nabla \log h(t,x)$ is typically intractable, and hence it is often the case that we can still not find an explicit form of the conditioned SDE. Therefore, the problem this work aims to address is related to sampling paths that resemble $(Y_t^*)_{t=0}^T$

Our work distinguishes itself from the few related controlled generative diffusion methods [Tri+23; Wu+24], that are also asymptotically consistent, in the order of discretization. Specifically, our approach is based only on a discretization of the path, while their approach is based on a discretization of the weights and a discretization of the path [1]. Furthermore, we study the asymptotically consistent and training-free conditional generation within a continuous-time perspective of generative diffusion, which offers two promising aspects.

[1] Essentially what we describe as an intermezzo in Section 4.2

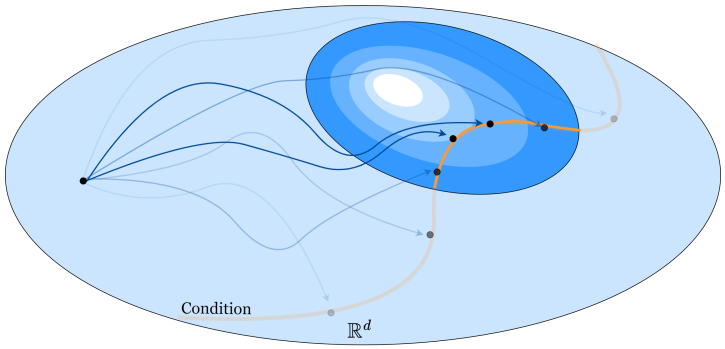


Figure 1.4: Diagram of our conditioned generative diffusion approach. The orange line represents the subset of the state space where the condition is met. The arrows indicate samples of the proposal process that often fail to resemble a sample from the true conditional distribution, indicated by their distance to the high-density area. Our approach is based on reweighing these proposal paths, illustrated by the opacity, to obtain a sample that resembles the true conditioned distribution.

First, the continuous-time framework is theoretically more flexible and allows us to reason about the limiting behaviour of our practical algorithm. This perspective has been fundamental in deriving theoretical guarantees of generative diffusion, such as in [Che+23]. Furthermore, most state-of-the-art accelerated sampling techniques approaches are based on the continuous-time framework, as laid out in [Ma+24].

Second, perhaps more importantly, the continuous-time framework allows us to bridge a gap between the fields of controlled generative diffusion and simulating conditioned stochastic processes. Because the transition density of the unconditioned denoising process Y is not known, conditioning the process to satisfy a specific condition at time T is known to be hard. Contrary to controlled generative diffusion, of which the scientific interest only recently emerged, the more general problem of conditioning continuous time stochastic processes has received widespread attention in the last few decades, going back to work on one-dimensional diffusion bridges [Cla90] to multivariate diffusions bridges [SMZ17] and conditioning with linear observations [BMS20; Mar12; Cor24].

1.3.1. Importance Sampling Technique

An embraced methodology for simulating conditioned stochastic processes bridges uses guided proposal processes, denoted by Y°, that are easy to sample, guaranteed to satisfy the desired conditions, and for which importance weights can be computed with respect to the measure of the true conditioned process Y°. Using these weights, which are values in $(0, \infty)$, importance sampling methods can be used to approximate the intractable measure of the conditioned process.

The weighting promises a theoretically consistent approach that is lacking in many of the controlled generation generative diffusion methods. For $N > 1$, the path-wise importance sampling technique consists of three main steps.

1. Simulate paths from the proposal process $\{(Y^\circ)^{(i)}\}_{i=1}^N$.
2. Compute importance weights $\{(W(Y^\circ))^{(i)}\}_{i=1}^N = \{W^{(i)}\}_{i=1}^N$.
3. Resample the set $\{(Y_T^\circ)^{(i)}\}_{i=1}^N$ with importance weights as sampling probabilities.

Computing the importance weights of paths requires the measures of the proposal process and the target process to be equivalent [2]. In the work by Schauer, van der Meulen and Zanten [SMZ17], which introduced a technique for $L = I_{d\times d}$ and Bierkens et al. [BMS20], which extended it to general L, a specific class of proposal processes is used for which the equivalence w.r.t. the target measure is shown. In this work, we take inspiration from this technique and adapt it to the setting of generative diffusions.

[2] $\mathbb{P}$ and $\mathbb{Q}$ are equivalent if $\mathbb{P}(A) > 0 \iff \mathbb{Q}(A) > 0$

The proposal processes that enable the approach are based on an auxiliary process, denoted by $\tilde{Y}$. This process is chosen to be simple enough such that it is possible to perform an exact conditioning for this stochastic process, i.e. we consider the measure $\tilde{\mathbb{P}}^*(\cdot) = \mathbb{P}(\cdot|L\tilde{Y}_T = v)$. Specifically, if the process $\tilde{Y}$ has Gaussian transition densities, than the marginal density of $L\tilde{Y}_T$ conditioned on $\tilde{Y}_t = x$ function satisfies

$$\tilde{h}(t,x) \overset{\text{def}}{=} \mathbb{P}(L\tilde{Y}_t|\tilde{Y}_t = x) \propto \exp\left(-\frac{1}{2}||L\tilde{\mu}_T(t,x) - v||^2_{(L\tilde{C}_T(t)L^\top)^{-1}}\right),$$

where $\tilde{\mu}_T$ is a function that represents the conditional expectation of $\tilde{Y}_T$ given $\tilde{Y}_t = x$, $\tilde{C}_T$ is a function that represents the covariance. Then, the proposal processes are driven by the following SDE

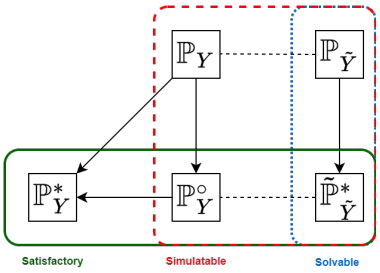
$$\mathrm{d}Y^\circ_t = [b(t,Y^\circ_t) + \sigma^2(t)\nabla \log \tilde{h}(t,Y^\circ_t)]\mathrm{d}t + \sigma(t)\mathrm{d}B_t, \tag{1.5}$$

where $\tilde{h}$ replaces h in Equation 1.4 such that simulating with this SDE is tractable. It follows from the properties of a Gaussian distribution that taking the logarithm and gradient of $\tilde{h}$ gives us

$$\nabla \log \tilde{h}(t,x) = \nabla\tilde{\mu}_T(t,x)^\top L^\top (L\tilde{C}_T(t)L^\top)^{-1}(L\tilde{\mu}_T(t,x) - v).$$

In [BMS20] $\nabla \log \tilde{h}$, it is assumed that the gradient (w.r.t. x) of $\tilde{\mu}_T(t,x)$ is independent of x, which simplifies the form even further and is a sensible choice if the auxiliary process is determined a priori. However, for our setting, due to the high-dimensional state spaces in which generative diffusion is often applied, we may want to adapt the auxiliary process as time moves forward. Especially as the approach intuitively benefits from a choice of $\tilde{Y}$ that is more similar to Y. This leads to the case that the gradient of $\tilde{\mu}_T(t,x)$ w.r.t. x does depend on x.

Under these new circumstances, the validity of the approach by [BMS20] no longer holds directly. Therefore, we must make new assumptions and derive new proofs to verify the technique's validity. In particular, we focus on both the behavior of $||LY^\circ_t - v||_2$ (Theorem 4.1), and the absolute continuity of the measure of Y^* w.r.t the measure of Y° (Theorem 4.2), as t becomes closer to the terminal time T .



Process	Symbol	Measure	Drift
Unconditioned	Y	$\mathbb{P}_Y$	b
Conditioned	Y^*	$\mathbb{P}^*_Y$ (or $\mathbb{P}_{Y^*}$)	$b + \nabla \log h$
Proposal	Y°	$\mathbb{P}^\circ_Y$ (or $\mathbb{P}_{Y^\circ}$)	$b + \nabla \log \tilde{h}$
Auxiliary	$\tilde{Y}$	$\mathbb{P}_{\tilde{Y}}$	$\tilde{b}$
Cond. auxiliary	$\tilde{Y}^*$	$\tilde{\mathbb{P}}^*_{\tilde{Y}}$	$\tilde{b} + \nabla \log \tilde{h}$

Figure 1.5: The solid arrows indicate a change-of-measure operation, that is incited by the associated Radon-Nikodym derivative (Definition 1.1). The dashed lines merely indicate an informal notion of resemblance between the measures. The description of the measures and the associated processes are given in the table. There are three characteristics of the processes: whether they satisfy the condition (satisfactory), whether they can be simulated (simulatable), and whether they have a known exact solution (solvable).

Contribution 1 (Importance sampling algorithm for conditioned generative diffusion)**.** *We derive an importance sampling algorithm for asymptotically consistent conditioning of generative diffusion models, taking inspiration from [BMS20] and [SMZ17]. In particular, we derive a method that uses proposal processes of the form in Equation 1.5 and computes importance weights for the paths with the aim of using importance sampling techniques (Proposition 4.2). We verify that the proposal process satisfies the condition at the time T (Theorem 4.1) and that the proposal and target measures of the paths remain absolutely continuous (Theorem 4.2), under various assumptions on the auxiliary process $\tilde{Y}$.*

1.3.2. Practical Algorithm

As foreshadowed, choosing an auxiliary proposal process is inherently hard. Therefore, we introduce an adaptive auxiliary process that uses the current information of the process at time t to obtain a tractable approximation of the process. This leads to some adjustments to Proposition 4.2 that we do in Proposition 5.5. To show that our approach is asymptotically consistent, we use the approximation of the conditioned probability of some set A with

$$\widehat{P}^*_{N,M}(A) = \frac{\sum_{i=1}^N W^{(i)} \mathbf{1}\{(Y_T^\circ)^{(i)} \in A\}}{\sum_{i=1}^N W^{(i)}}.$$

Here, M denotes the discretization of the sampled processes Y°, and N denotes the number of sampled paths used to approximate the distribution, that we may also refer to as particles. Specifically, we show that the mean squared error vanishes under a few assumptions. These assumptions are primarily related to the auxiliary process, the importance weights, and the set A.

Contribution 2 (Asymptotic consistency of our approach)**.** *We show that our approach is asymptotically consistent (Theorem 5.1), in the sense that the mean squared error converges to zero as we take the limit in $N \to \infty$ and $M \to \infty$. Specifically, we find that under the assumptions laid out in Chapter 4 and Chapter 5, there exists a positive constant C such that for all appropriate[3] sets A, we have*

$$\mathbb{E}\left[(\widehat{P}^*_{N,M}(A) - \mathbb{P}^*_{Y_T}(A))^2\right] \leq C\left(\frac{1}{M} + \frac{1}{N}\right).$$

1.3.3. Numerical Experiments

In the experimental part of this work, we focus on two concrete proposal processes that are obtained from two specific auxiliaries. The first is based on an adaptive drifted Brownian motion, referred to as (G)CDA [4]. In principle, here the auxiliary drift term is assumed to be constant, but it is updated at each sample step, resulting in a consecutively changing choice of auxiliary processes. The second is based on a fixed non-drifted Brownian motion, referred to as ZDA. Here, the auxiliary drift term is assumed to be zero, and is therefore not affected by the adoption of the adaptive framework.

Our experimental results are based on data distributions where we can analytically derive $\nabla \log p_{X_{T-t}}$, to omit the need for training neural networks. We are specifically interested in the practical implications of the asymptotic consistency, primarily focused on the effect of the number of particles N, because the effect of the discretization level M is already quite intuitive in the context of (unconditional) generative diffusion models.

To establish the statistical performance for finite computational effort, we make use of sample-based statistical performance metrics and find a significant impact of our approach even for small values of NI. In addition, it is well known that importance sampling collapses in high-dimensional state spaces [Aga+17; LBB05], which is typically the case for applications of generative diffusion. This will result in only one unique sample, even when we use N particles initially. For this reason, we investigate the effect of the number of particles on the statistical performance in a variety of settings, e.g., low- and high-dimensional, and method configurations, e.g., ZDA and (G)CDA.

Contribution 3 (Empirical Statistical Performance)**.** *We find that using more computational effort typically increases the statistical performance for various metrics, establishing a positive insight into the non-asymptotic behavior of our approach. The significance at which they increase is negatively affected by the dimensionality of the problem. Furthermore, in practically all moderately high-dimensional settings, we need $\mathcal{O}(N)$ computation for a single sample. However, using K independent runs of our approach with each $\mathcal{O}(N/K)$ computational effort can significantly improve the statistical performance, thereby inciting an allocation problem that concerns managing K and N.*

[3] they have to satisfy some mild conditions about the boundary of A and the discretizations of Y°.

[4] (G)CDA is an acronym for (Gradient propagated) Constant Drift Approximation, and ZDA is an acronym for Zero Drift Approximation.

1.4. Outline

In Part I, we lay out the theoretical framework of our work. The primary focus is on generative diffusion and how, at a more fundamental level, (generative) diffusions can be conditioned in exact or approximate ways. In Chapter 2, we describe the continuous time perspective on generative diffusion. We discuss how the noising and denoising processes are precisely related and how, in practice, one can use neural networks to approximate the score function. The chapter also contains illustrative examples of generative diffusion models. In Chapter 3, we first describe how one can formally condition SDEs. Furthermore, we discuss two commonly used heuristics, called replacement guidance and reconstruction guidance, that form the basis of most of the state-of-the-art work in (controlled) generative diffusion.

In Part II, we introduce the importance sampling technique, the associated practical algorithm, and the asymptotic consistency. In Chapter 4, we elaborate on the importance sampling technique that is fundamental to our approach. Here, we describe how importance sampling procedures can be achieved for continuous-time stochastic processes. Specifically, we derive an appropriate version of the approach introduced in [SMZ17] and verify the well-definedness. In Chapter 5, we introduce our practical algorithm based on a discretization and an adaptive proposal process. Apart from deriving the necessary theoretical adjustments that are required for our practical approach to work, we describe an upper bound of the squared error of the particle approximation and the ground truth distribution as a function of the number of particles N and the level of discretization M. The upper bound vanishes in the limit of large N and M, thereby implying an asymptotically consistent method. Finally, we provide details for two practical implementations of the proposal processes based on a zero and constant drift approximation.

In Part III, we provide experimental insights into the empirical performance of our approach. In Chapter 6, we demonstrate the statistical performance in terms of sample-based performance metrics for comparing distributions. We are primarily concerned with how the statistical performance is affected by increasing amounts of computational effort, and how this effect is changed for different dimensionalities of the state space. Apart from comparing the two different proposal strategies (ZDA and (G)CDA), we provide some insight into the particle efficiency, which reduces to a degeneracy within our approach. We study how minor variations of our algorithm may partially resolve some of the issues, and therefore regain some of its efficiency. In Chapter 7, we demonstrate an application of our approach. Specifically, we perform a selection of experiments in the context of conditional scenario generation. First, the generated scenarios are conditioned on masks of certain events happening in the future, such as hitting a certain value at a certain time. Second, we condition the scenarios on information in the frequency domain, such as the average (zeroth frequency) or certain trend and seasonal behaviours (low-pass and band-pass). Finally, we consider conditioning the generated scenarios on inequality conditions, instead of equality conditions, which promises a generalization of our technique to conditions of the form $LY_T \in V$.

In Part IV, we discuss the main results and caveats of this approach with respect to the intersecting research and application areas. In Chapter 8, we focus primarily on recommendations for future work. We describe the assumptions on which our theory relies and how they may be alleviated. Also, we consider improved variants of our proposal processes and how some can be related to existing work. Furthermore, we discuss some possible extensions and provide an outlook for the general technique. In Chapter 9, we provide an executive summary of the main conclusions of this work.

1.5. Preliminaries and Notation

The preliminaries of this work are primarily results from stochastic calculus. The reader is directed to a standard textbook of stochastic differential equations, such as that of Øksendal [Øks03] or Mao [Mao11a]. Furthermore, the works by Schauer et al. [SMZ17] and Bierkens et al. [BMS20] are especially important, as our approach builds on the foundation laid out there. It is also recommended to study the introductory chapters of the work by Corstanje [Cor24], where the technique and preliminaries are described in an approachable manner. After providing some notational remarks, we will define a stochastic process and work towards stating Girsanov's theorem, which enables the measure changes of stochastic processes.

1.5.1. General Notation

The typical state space we consider is $\mathbb{R}^d$ endowed with the Euclidean norm, i.e. $||x|| = \sqrt{x \cdot x}$. Furthermore, we consider the following notation of the energy norm, defined with a $d \times d$ positive definite matrix A, i.e. $||x||_A = \sqrt{x \cdot A \cdot x^\top}$. We denote the matrix square root for positive definite matrices by $A^{1/2}$. Note that the energy norm and the Euclidean norm can be related through $||x||_A = ||A^{1/2}x||$. $||A||$ denotes the operator norm w.r.t. the Euclidean norm, i.e. $||A|| = \sup_{v \in \mathbb{R}^d} \frac{||Av||}{||v||}$. For the operator norm we have the following sub-multiplicative properties $||AB|| \leq ||A||\,||B||$ and $||Av|| \leq ||A||\,||v||$. Furthermore, we often use proportional bounds and therefore, to avoid the notational clutter and ambiguity in writing constants, we may adopt Big-O notation or $\lesssim$ to express the proportional upper bounds instead, i.e.

$$\exists\, C > 0 \text{ independent of } x \text{ such that } \forall x\, f(x) \leq Cg(x) \iff f = \mathcal{O}(g(x)) \iff f(x) \lesssim g(x).$$

1.5.2. Multivariate Calculus

Consider the notation of $x = (x_1 \dots x_d) \in \mathbb{R}^d$. For some function $f : \mathbb{R}^d \to \mathbb{R}^{d'}$ and $h : \mathbb{R}^d \to \mathbb{R}$ we use the following notation

$$\begin{aligned}
&\text{(Gradient)} && \nabla h(x) = \begin{pmatrix} \frac{\partial}{\partial x_1} h(x) & \dots & \frac{\partial}{\partial x_d} h(x) \end{pmatrix} \in \mathbb{R}^d, \\
&\text{(Hessian)} && \nabla^2 h = \begin{pmatrix} \frac{\partial^2}{\partial x_1^2} h(x) & \dots & \frac{\partial^2}{\partial x_1 x_d} h(x) \\ \vdots & \ddots & \vdots \\ \frac{\partial^2}{\partial x_d x_1} h(x) & \dots & \frac{\partial^2}{\partial x_d x_d} h(x) \end{pmatrix} \in \mathbb{R}^{d \times d}, \\
&\text{(Jacobian)} && \nabla f(x) = \begin{pmatrix} \nabla f_1(x) \\ \vdots \\ \nabla f_{d'}(x) \end{pmatrix} \in \mathbb{R}^{d' \times d}, \\
&\text{(Divergence)} && \nabla \cdot f = \frac{\partial}{\partial x_1} f_1(x) + \dots + \frac{\partial}{\partial x_d} f_d(x) \text{ if } d = d', \\
&\text{(Laplacian)} && \Delta h(x) = \operatorname{tr}\left[\nabla^2 h\right] = \frac{\partial^2}{\partial x_1^2} h(x) + \dots + \frac{\partial^2}{\partial x_d^2} h(x),
\end{aligned}$$

and for time dependent functions $f : [0, T] \times \mathbb{R}^d \to \mathbb{R}^{d'}$, we always take the above operations only w.r.t. the x arguments, for example,

$$\nabla g(t, x) = \begin{pmatrix} \frac{\partial}{\partial x_1} g(t, x) & \dots & \frac{\partial}{\partial x_d} g(t, x) \end{pmatrix}.$$

We recall the following differentiation rules,

$$\begin{aligned}
&\text{(Divergence product rule)} && \nabla \cdot (hf) = h \nabla \cdot f + f \cdot \nabla h, \\
&\text{(Gradient chain rule)} && \nabla (h(f(x))) = \nabla f(x)^\top ((\nabla h)(f(x))).
\end{aligned}$$

1.5.3. Itô Processes

A stochastic process is a random variable $X : [0,T] \times \Omega \to \mathbb{R}^d$, where $\mathbb{R}^d$ denotes the state space. Typically, the process is denoted by an index of the time, i.e., such that $X_t(\cdot) \to \mathbb{R}^d$ is a random variable[5]. We denote the path measure of a process $X = (X_t)_{t=0}^T$ with $\mathbb{P}_X$.

Specifically, in our setting, we may consider the triple $(\Omega, \mathcal{F}, \mathbb{P})$ where $\Omega = C([0,T], \mathbb{R}^d)$ which denotes the set of continuous paths in $\mathbb{R}^d$ from time 0 to T. Then $\mathcal{F} = \mathcal{B}(C([0,T], \mathbb{R}^d))$ is the Borel σ-algebra associated to the space of continuous paths in $\mathbb{R}^d$. We refer to a filtration as an ordered collection $(\mathcal{F}_t)_{t=0}^T$ where $\mathcal{F}_t$ is a sub-σ-algebra of $\mathcal{F}$ and $\mathcal{F}_t \subseteq \mathcal{F}_s$ for all $s \leq t \leq T$. We say that a process X is adapted to the filtration $(\mathcal{F}_t)_{t=0}^T$ if the random variable X_t is $\mathcal{F}_t$ measurable.

We say that $\mathbb{P}$ is *absolutely continuous* w.r.t. $\mathbb{Q}$ if for all $A \in \mathcal{F}$, we have that $\mathbb{P}(A) > 0$ implies $\mathbb{Q}(A) > 0$. If the absolute continuity between two measures is symmetric, we speak of equivalent measures, denoted by $\mathbb{P} \sim \mathbb{Q}$. The following theorem introduces the Radon-Nikodym derivative that formalizes densities of measures

Theorem 1.1 (Radon-Nikodym theorem)**.** *If $\mathbb{P}$ and $\mathbb{Q}$ are two measures on $(\Omega, \mathcal{F})$ such that $\mathbb{P}$ is absolutely continuous w.r.t. $\mathbb{Q}$ then there exists a random variable Z such that for any $\mathcal{F}$-measurable set A,*

$$\mathbb{P}(A) = \int_A Z \mathrm{d}\mathbb{Q}.$$

The function Z satisfying the above equation is uniquely defined up to a $\mathbb{Q}$-null set, and is typically denoted with $Z \overset{\text{def}}{=} \frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}$.

Now, we can define the marginal density of the stochastic process at time t with a Radon-Nikodym derivative of $\mathbb{P}$ with respect to the Lebesgue measure, i.e.

$$p_{X_t}(x) = \frac{\mathbb{P}(X_t \in \mathrm{d}x)}{\mathrm{d}x}.$$

In a similar viewpoint, we write the transition density of the process with

$$p_{X_s|X_t=x}(y) = \frac{\mathbb{P}(X_s \in \mathrm{d}y | X_t = x)}{\mathrm{d}y}.$$

We use the following notation to describe marginal distributions $\mathbb{P}_{X_t}(\cdot) = \mathbb{P}(X_t \in \cdot)$.
Let $(B_t)_{t>0}$ be a Brownian motion on $\mathbb{R}^d$ and $(X_t)_{t>0}$ be a stochastic process on $\mathbb{R}^d$ governed by the following SDE

$$\mathrm{d}X_t = b(t, X_t)\mathrm{d}t + \sigma(t, X_t)\mathrm{d}B_t, \qquad X_0 \sim \mathbb{P}_{X_0}, \tag{1.6}$$

where the drift coefficient is denoted by $b : [0,T] \times \mathbb{R}^d \to \mathbb{R}^d$ and the diffusion coefficient by $\sigma : [0,T] \times \mathbb{R}^d \to \mathbb{R}^{d \times d}$. In this work, we often consider σ to be a scalar function and will also use it to denote with $\sigma : [0,T] \to \mathbb{R}$. In that case

$$\mathrm{d}X_t = b(t, X_t)\mathrm{d}t + \sigma(t)\mathrm{d}B_t, \qquad X_0 \sim \mathbb{P}_{X_0}. \tag{1.7}$$

The differential form of Equation 1.6 is a shorthand notation for

$$X_t = X_0 + \int_0^t b(s, X_s)\mathrm{d}s + \int_0^t \sigma(s, X_s)\mathrm{d}B_s. \tag{1.8}$$

This class of stochastic processes is referred to as Itô processes. A unique strong solution to the above SDE exists if the coefficients are Lipschitz continuous. Lipschitz continuity is satisfied if

$$||b(t,x) - b(s,y)|| + |\sigma(t,x) - \sigma(s,t)| \lesssim ||x - y|| + |s - t|,$$

for all $x, y \in \mathbb{R}^d$ and $t, s \in [0,T]$. Furthermore, the Lipschitz continuity implies a linear growth condition:

$$||b(t,x)|| + |\sigma(t)| \lesssim (1 + ||x|| + |t|).$$

[5] We use X to denote the noising process and Y to denote the denoising process.

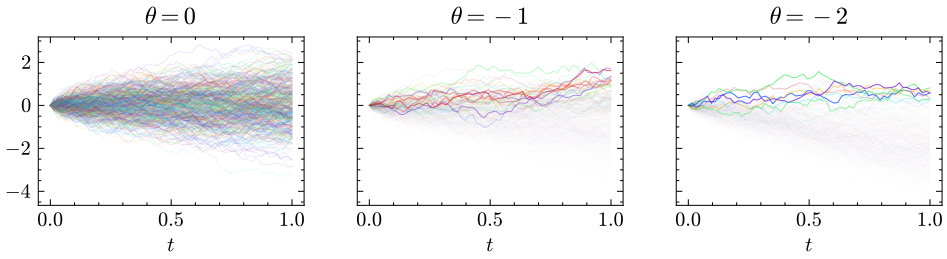


Figure 1.6: Girsanov transformation illustrated for (drifted) Brownian motions. The opacity of the paths represents the Radon-Nikodym derivative of the paths.

1.5.4. Girsanov's theorem

We use Girsanov's theorem to derive Radon-Nikodym derivatives for stochastic processes. An example is given below. In our work, we use the Radon-Nikodym derivatives between the proposal process and the true conditional process to obtain importance weights that can be used to reweigh the sample paths of the proposal process so that they match the true measure of the conditional process.

Theorem 1.2 (Girsanov's Theorem[Øks03])**.** *Suppose that $(X_t)_{t\geq 0}$ is a strong solution to the SDE in Equation 1.6. Let $(B_t)_{t=0}^T$ be a $\mathbb{P}$-Brownian motion and let $\eta : [0,T] \times \mathbb{R}^d : \mathbb{R}^d$ be such that $(\eta(t, X_t))_{t=0}^T$ is an $\mathcal{F}_t$- adapted process that satisfies Novikov's condition, i.e.*

$$\mathbb{E}\left[\exp\left(\frac{1}{2}\int_0^T ||\eta(s, X_s)||^2 \mathrm{d}s\right)\right] < \infty,$$

then the process

$$M_t = \int_0^t \eta(s, X_s)\mathrm{d}B_s$$

is a $\mathbb{P}$-martingale. Moreover, if

$$\mathbb{E}\left[\mathcal{E}(M)_T\right] = 1 \text{ where } \mathcal{E}(M)_T = \exp\left(\int_0^t \eta(s, X_s)\mathrm{d}B_s - \frac{1}{2}\int_0^t ||\eta(s, X_s)||^2 \mathrm{d}s\right),$$

then $\mathcal{E}(M)$, also called the stochasit exponential, is also a $\mathbb{P}$-martingale and specifically the process

$$B'_t = B_t - \int_0^t \eta(s, X_s)\mathrm{d}s,$$

is a Brownian motion under the measure $\mathbb{Q}$, which is defined by the following Radon-Nikodym derivative:

$$\mathcal{E}(M)_T = \left.\frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mathbb{P}}\right|_{\mathcal{F}_T}.$$

Note: *Throughout this work, the context of the Brownian motion and SDEs should clarify under which measure we consider the Brownian motion, to avoid cluttering the notation.*

Example 1.1. *Consider the following stochastic differential equations*

$$\mathrm{d}X_t = \theta \mathrm{d}t + \mathrm{d}B_t,$$

where B is a $\mathbb{P}$-Brownian motion Then Girsanov's theorem tells us that

$$\left.\frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mathbb{P}}\right|_{\mathcal{F}_t}(Y) = \exp\left(-\theta B_t - \frac{1}{2}\theta^2 t\right).$$

where $\mathbb{Q}$ is the measure under which X is a Brownian motion. In Figure 1.6, we have illustrated this example for various values of θ.

1.5.5. Infinitesimal generators

Definition 1.2 (Infinitesimal Generator)**.** *The family of infinitesimal generators* $\{\mathcal{L}_t\}_{t=0}^{T}$*, parametrized by a time parameter* t*, of Markov process* X *is defined by*

$$\mathcal{L}_t f(x) = \lim_{s \to 0} \frac{\mathbb{E}[f(X_{t+s})|X_t = x] - f(x)}{s}.$$

The infinitesimal generator depicts the behavior of the stochastic process through the expectation of a functional of X as it evolves in time, which can be seen from the identity that holds for vanishing s

$$\mathbb{E}[f(X_{t+s})|X_t = x] = s\mathcal{L}_t f(X_t) + \mathcal{O}(s).$$

Definition 1.3 (Domain of infinitesimal generator)**.** *The set of functions for which the limit exists is denoted by* $D_{\mathcal{L}_t}(x)$

$$D_{\mathcal{L}_t}(x) = \left\{ f \in C_0(\mathbb{R}^d) : \mathcal{L}_t f(x) \quad \textit{exists} \right\}.$$

Throughout the thesis, we typically assume that the generator is in the domain whenever we apply it to a function, and it is not stated explicitly.

Proposition 1.1 (Infinitesimal generator of an Itô process)**.** *Let* $(X_t)_{t\geq 0}$ *be driven by the SDE given in Equation 1.6 and let* $f \in D_{\mathcal{L}_t}$*, then*

$$\mathcal{L}_t f(x) = b(t,x) \cdot \nabla f(x) + \frac{1}{2} \operatorname{Tr}\left[\sigma(t,x)\sigma(t,x)^\top \nabla^2 f(x)\right].$$

Corollary 1.1 (Infinitesimal generator of an Itô process with state-independent scalar diffusion coefficient.)**.** *Let* $(X_t)_{t\geq 0}$ *be driven by the SDE given in Equation 1.7 and let* $f \in D_{\mathcal{L}_t}$*, then*

$$\mathcal{L}_t f(x) = b(t,x) \cdot \nabla f(x) + \frac{1}{2}\sigma^2(t)\Delta f(x).$$

If f is a time-dependent function as opposed to just dependent on the state, we use the space-time generator, which is given by $\partial_t f + \mathcal{L}_t f_t$. Here we say that $f_t = f(t,\cdot)$, in the sense that the operator $\mathcal{L}_t$ acts on the state variable.

The infinitesimal generator can be used to specify the Kolmogorov Backward Equation (KBE) and Kolmogorov Forward Equation (KFE) in the following.

Theorem 1.3 (Kolmogorov Backward Equation and Kolmogorov Forward Equation)**.** *Let* $p_{X_s|X_t=x}(y)$ *denote the transition density of process* X*. Then, the Kolmogorov Backward equation is specified as*

$$\frac{\partial\, p_{X_s|X_t=x}(y)}{\partial t} = -b(t,x) \cdot \nabla_x\, p_{X_s|X_t=x}(y) - \frac{1}{2} \operatorname{Tr}\left[\sigma(t,x)\sigma(t,x)^\top \nabla_x^2\, p_{X_s|X_t=x}(y)\right] = -\mathcal{L}_t\, p_{X_s|X_t=x}(y), \tag{1.9}$$

where the infinitesimal generator and the gradient operators act on x*, and not on* y*. Let* p_{X_t} *denote the marginal density of process* X*. Then, the Kolmogorov Forward equation is specified as*

$$\frac{\partial\, p_{X_t}(x)}{\partial t} = -\sum_{i=1}^{d} \frac{\partial}{\partial x_i}(b_i(t,x)p_{X_t}(x)) + \frac{1}{2}\sum_{i=1}^{d}\sum_{j=1}^{d}[(\sigma(t,x)\sigma(t,x)^\top)_{ij} p_{X_t}(x))] = \mathcal{L}_t'\, p_{X_t}(x), \tag{1.10}$$

where $\mathcal{L}_t'$ *is the adjoint operator of* $\mathcal{L}_t$*.*

Part I

Essential Concepts and Theoretical Setting

2

Generative Diffusion

In this chapter, we elaborate on the mathematical description of generative diffusion techniques. In essence, the technique is based on two stochastic processes that can be studied by the stochastic differential equations (SDE) that drive them. In doing so, we elaborate on the relation between the noising SDE and the denoising SDE, which we display here together below for convenient comparison

$$\begin{aligned}
&\textbf{(Noising SDE)} \quad && \mathrm{d}X_t = \bar{\alpha}(t)X_t\mathrm{d}t + \bar{\sigma}(t)\mathrm{d}\bar{B}_t, && \text{with } X_0 \sim \mathbb{P}_{X_0},\\
&\textbf{(Denoising SDE)} \quad && \mathrm{d}Y_t = \left[-\alpha(t)Y_t + \frac{1}{2}\sigma^2(t)\nabla \log p_{X_{T-t}}(Y_t)\right]\mathrm{d}t + \sigma(t)\mathrm{d}B_t, && \text{with } Y_0 \sim \mathbb{P}_{Y_0},
\end{aligned}$$

where $\bar{B}$ and B denote two different independent Brownian motions. Throughout this chapter, we will carefully describe the relations among the scalar functions $\alpha, \sigma, \bar{\alpha}$ and $\bar{\sigma}$, and how they determine the behavior of the noising and denoising process, through what is known as the noise schedule. Under certain ideal settings, it turns out that the marginal distributions of X_t and Y_{T-t} are identical for all $t \in [0, T]$. This justifies the denoising process for generating samples of the data distribution $\mathbb{P}_{X_0}$.

Even in a less-than-ideal setting, theoretical guarantees can be derived for the statistical performance of generative diffusion models. For example, in the theoretical work by Chen et al. [Che+23], it is shown that the total variation between the $\mathbb{P}_{Y_T}$ and $\mathbb{P}_{X_0}$ can be bounded by three sources of errors, i.e.,

$$\mathrm{TV}\left(\mathbb{P}_{X_0}, \mathbb{P}_{Y_T}\right) = \mathcal{O}\left(\epsilon_{\text{forward}} + M^{-1/2} + \epsilon_{\text{score}}\right), \tag{2.1}$$

where $\epsilon_{\text{forward}}$ is an error bound induced by the mismatch between $\mathbb{P}_{Y_0}$ and $\mathbb{P}_{X_T}$, M is the number of discretization steps, and ϵ_{score} is an error bound of the learned score function. The intuition is that the performance of generative diffusion can be improved by choosing appropriate parameters for the noising SDE, choosing a small discretization step, and obtaining an accurate score function. Satisfying these score accuracy aspects is difficult. In particular, the design of the neural networks that approximate the score functions is a deep topic on its own, where their success heavily depends on the structure of the data distribution. Therefore, studying data distributions for which the marginal distribution $\mathbb{P}_{X_t}$ has analytically tractable score functions is worthwhile for developing a theoretical understanding, so we can study our approaches in isolation from the neural network design choices.

The performance guarantee shows that a trade-off between performance and computational effort must be made to determine an appropriate number of discretization steps. The computational drawbacks of choosing a larger number of (Euler-Maruyama) discretization steps are evident, given that every simulated step of the denoising process requires a neural network evaluation.

In Section 2.1, we describe the noising process and its analytical properties. Specifically, we show how the noising process may have explicit transition distributions and how noise schedules can be interpreted. In Section 2.2, we describe the denoising process, which forms the basis of the generative diffusion technique. This denoising SDE is obtained by performing a time-reversal of the noising SDE. Finally, Section 2.3 briefly discusses how the denoising process is learned in practice from a dataset by a procedure called denoising score matching. Furthermore, we provide two example data distributions with a tractable score function.

2.1. Noising Process

The noising process, that we denote by $X = (X_t)_{t=0}^T$, has the purpose of corrupting a clean data sample $X_0 \sim \mathbb{P}_{X_0}$ to a noisy state X_T. Specifically, for some scalar functions $\bar{\alpha}$ and $\bar{\sigma}$, the process is governed by the following SDE:

$$\mathrm{d}X_t = \bar{\alpha}(t)X_t\mathrm{d}t + \bar{\sigma}(t)\mathrm{d}\bar{B}_t, \qquad \text{with } X_0 \sim \mathbb{P}_{X_0}, \tag{2.2}$$

where $(\bar{B}_t)_{t>0}$ is a $\mathbb{R}^d$-valued Brownian motion. For a general choice of $\bar{\alpha}$ and $\bar{\sigma}$, this process resembles an Ornstein-Uhlenbeck process, with time-dependent parameters. Specifically, $\bar{\alpha} : [0,T] \mapsto (-\infty, 0]$ represents the reversion rate of the process while $\bar{\sigma} : [0,T] \mapsto (0,\infty)$ represents the scalar diffusion coefficient of the noising process. We emphasize the use of a barred notation of the coefficients $\bar{\alpha}$ and $\bar{\sigma}$, as we refer to the related coefficients of the denoising process, that appear significantly more frequently in our work, with the unbarred version α and σ. The noising process is chosen to resemble a class of SDEs that at terminal time T have a marginal distribution that roughly matches our choice for $\mathbb{P}_{Y_0}$, for which a typical choice is to use $\mathcal{N}(0, T \cdot I_{d\times d})$ or $\mathcal{N}(0, I_{d\times d})$.

For the remainder of this work, we assume that the scalar functions are chosen such that for all $t \in [0,T]$:

$$-\infty < \int_0^t \bar{\alpha}(s)\mathrm{d}s < \infty \text{ and } 0 < \int_0^t \bar{\sigma}^2(s)\mathrm{d}s < \infty.$$

Under this assumption, the following proposition gives the solution to the noising SDE.

Proposition 2.1 (Noising SDE)**.** *Let us consider the noising SDE in Equation 2.2. Then the following holds:*

1. *The solution to the noising SDE is given by*

$$X_t = \frac{X_0}{\bar{\phi}(t)} + \int_0^t \frac{\bar{\phi}(r)}{\bar{\phi}(t)}\bar{\sigma}(r)\mathrm{d}B_r, \quad \textit{where} \quad \bar{\phi}(t) = \exp\left(-\int_0^t \bar{\alpha}(s)\mathrm{d}s\right). \tag{2.3}$$

2. *the marginal distribution of X_t conditioned on X_0 is given by*

$$X_t|X_0 = x_0 \sim \mathcal{N}\left(\frac{x_0}{\bar{\phi}(t)}, \bar{\gamma}(t)I_{d\times d}\right) \quad \textit{where} \quad \bar{\gamma}(t) = \int_0^t \left(\frac{\bar{\phi}(r)}{\bar{\phi}(t)}\bar{\sigma}(r)\right)^2 \mathrm{d}r. \tag{2.4}$$

Proof. **Item 1.** Consider the function $f(t,x) = \bar{\phi}(t)x$ and observe that $f \in C^{1,2}([0,T]\times\mathbb{R}^d)$. Therefore, we can apply Itô's formula [Øks03]:

$$\mathrm{d}\big(\bar{\phi}(t)X_t\big) = \left[\frac{\partial\bar{\phi}(t)}{\partial t}X_t + \bar{\phi}(t)\bar{\alpha}(t)X_t\right]\mathrm{d}t + \bar{\phi}(t)\bar{\sigma}(t)\mathrm{d}\bar{B}_t.$$

Now, observe that $\bar{\phi}$ solves the following ordinary differential equation

$$\frac{\partial\bar{\phi}}{\partial t}(t) = -\bar{\alpha}(t)\bar{\phi}(t).$$

Then, we obtain

$$\mathrm{d}\big(\bar{\phi}(t)X_t\big) = \bar{\phi}(t)\bar{\sigma}(t)\mathrm{d}\bar{B}_t,$$

for which the solution can be obtained by integrating both sides:

$$\bar{\phi}(t)X_t - \bar{\phi}(0)X_0 = \int_0^t \bar{\phi}(r)\bar{\sigma}(r)\mathrm{d}B_r.$$

Now, dividing both sides with $\bar{\phi}(t)$ and observing that $\bar{\phi}(0) = 1$, gives us the desired result.

Item 2. Now what remains is to show that the marginal distribution of X_t, conditioned on X_0 is a Gaussian. We can use the fundamental properties of the Itô integral (as given in e.g. [Øks03]). Because the integrand in the stochastic integral in Equation 2.3 is deterministic, we know that the stochastic integral is Gaussian distributed. Furthermore, the expectation of the stochastic integral is zero, and therefore

$$\mathbb{E}\left[X_t | X_0\right] = \frac{X_0}{\bar{\phi}(t)}.$$

At last, we know by Itô isometry that the covariance matrix is given by

$$\mathbb{E}\left[\left(\int_0^t \bar{\phi}(r)\bar{\sigma}(r) I_{d\times d} \mathrm{d}B_r\right)\left(\int_0^t \bar{\phi}(r)\bar{\sigma}(r) I_{d\times d} \mathrm{d}B_r\right)^\top\right] = \bar{\gamma}(t) I_{d\times d}$$

where on the left-hand side we have made explicit the fact that $\bar{\phi}$ and $\bar{\sigma}$ are scalar functions and the Brownian motion takes values in $\mathbb{R}^d$. Combining the Gaussianity with the above-derived mean and covariance gives us the desired result. □

A few choices exist for the diffusion coefficient $\bar{\sigma}$ and the reversion rate $\bar{\alpha}$, such that the distribution can be obtained explicitly. First, we consider two simple examples of SDEs that are easy to understand but rarely used in practice.

Definition 2.1 (Simple Noising SDEs)**.** *Two simple noising SDEs are*

$$\begin{aligned} &(\textit{Simple SDE-1}) && \mathrm{d}X_t = \mathrm{d}\bar{B}_t, \\ &(\textit{Simple SDE-2}) && \mathrm{d}X_t = -\frac{1}{2}X_t \mathrm{d}t + \mathrm{d}\bar{B}_t. \end{aligned}$$

Note that the first SDE in the above definition is a Brownian motion, obtained with $\bar{\alpha} = 0$ and $\bar{\sigma} = 1$, and the second is a simple Ornstein-Uhlenbeck process, obtained with $\bar{\alpha} = -\frac{1}{2}$ and $\bar{\sigma} = 1$. An impractical aspect of generative diffusion with these noising SDEs, is that they do not efficiently attribute the computational effort within the time span $[0, T]$. One way to think of this is that when we are noising the data, if we add the noise linearly, i.e., with a constant diffusion coefficient, the clean sample would be quickly submerged under the noise, making it difficult to learn an approximation neural network. Therefore, a common consideration is to use time-varying diffusion and reversion rate parameters. In the context of generative diffusion, this is known as the noise schedule.

2.1.1. Noise Schedules

The adaptation of parameters $\bar{\sigma}$ and $\bar{\alpha}$ that tactically distribute the noise levels along a fixed time span $[0, T]$ can be related to a time-change operation of the simple processes in Definition 2.1. This means that if we sample the time-changed noising process X' at a uniform grid, it is identical in distribution to sampling the Simple SDEs with a non-uniform grid. Precisely, we have for some strictly monotone function $\bar{\beta} : [0, T] \to [0, T]$ that a new process X' can be defined by

$$X'_t \overset{d}{=} X_{\bar{\beta}(t)}.$$

We refer to the time changes as a noise schedule and derive two well-known noising SDEs: the variance-exploding SDE and the variance-preserving SDE.

Definition 2.2 (Noise schedule)**.** *A noise schedule is defined as a continuous, differentiable, and monotone function* $\bar{\beta} : [0, T] \to (0, \infty)$.

A typical example of a noise schedule, that we will employ in the experiments in this work, is $\bar{\beta}(t) = \beta_{\min}\left(1 - \frac{t^2}{T^2}\right) + \beta_{\max}\frac{t^2}{T^2}$ for some $0 < \beta_{\min} < \beta_{\max} < \infty$

Definition 2.3 (Variance Exploding (VE) SDE)**.** *If we set* $\bar{\alpha}(t) = 0$ *and* $\bar{\sigma}(t) = \sqrt{\frac{\mathrm{d}\bar{\beta}(t)}{\mathrm{d}t}}$*, we have the noising SDE that is known as variance exploding (VE) SDE, i.e.,*

$$\mathrm{d}X_t = \sqrt{\frac{\mathrm{d}\bar{\beta}(t)}{\mathrm{d}t}}\mathrm{d}\bar{B}_t.$$

Definition 2.4 (Variance Preserving (VP) SDE)**.** *If we set* $\bar{\alpha}(t) = -\frac{1}{2}\frac{\mathrm{d}\bar{\beta}(t)}{\mathrm{d}t}$ *and* $\bar{\sigma}(t) = \sqrt{\frac{\mathrm{d}\bar{\beta}(t)}{\mathrm{d}t}}$*, we have the noising SDE that is known as variance preserving (VP) SDE, i.e.,*

$$\mathrm{d}X_t = -\frac{1}{2}\frac{\mathrm{d}\bar{\beta}(t)}{\mathrm{d}t}X_t\mathrm{d}t + \sqrt{\frac{\mathrm{d}\bar{\beta}(t)}{\mathrm{d}t}}\mathrm{d}\bar{B}_t.$$

In the following two Lemmas we describe how a time change of the simple SDEs leads to the VE and VP SDE.

Proposition 2.2 (Time-changes for VE-SDE and VP-SDE)**.** *Let* X_t^1 *and* X_t^2 *be the solution to the Simple SDE-1 and Simple SDE-2, respectively. Let us consider all of the below-described processes to have a fixed initial position* $X_0 = x_0$*. Let* $\bar{\beta} : [0, T] \to (0, \infty)$ *be a noise schedule. Then:*

1. *The solution to the variance exploding SDE, denoted by* X^{VE}*, satisfies* $X_t^{\mathrm{VE}} \overset{d}{=} X^1_{\bar{\beta}(t)}$*.*

2. *The solution to the variance preserving SDE, denoted by* X^{VP}*, satisfies* $X_t^{\mathrm{VP}} \overset{d}{=} X^2_{\bar{\beta}(t)}$*.*

Proof. **Item 1.** We have that

$$X^1_{\bar{\beta}(t)} = X_0^1 + \int_0^{\bar{\beta}(t)} \mathrm{d}B_s \text{ and } X^{\mathrm{VE}} = X_0^{\mathrm{VE}} + \int_0^t \sqrt{\frac{\mathrm{d}\bar{\beta}(s)}{\mathrm{d}s}}\mathrm{d}B_s.$$

We can use the same line of reasoning as in the proof of Item 2 of Proposition 2.1. Specifically, we use the properties of the Itô integral to obtain the Gaussianity and the associated mean and covariance. The mean here is X_0 and the covariance is $\bar{\beta}(t)I_{d\times d}$, again making use of Itô isometry. Then, we can find the same mean and covariance for X_t^1. Therefore, the processes have the same marginal distributions conditional on $X_0 = x_0$.

Item 2. We have, using similar techniques as in the proof of 2.1, that

$$X^2_{\bar{\beta}(t)} = \frac{X_0^2}{\exp\left(-\frac{1}{2}\bar{\beta}(t)\right)} + \int_0^{\bar{\beta}(t)} \frac{\exp\left(-\frac{1}{2}s\right)}{\exp\left(-\frac{1}{2}\bar{\beta}(t)\right)}\mathrm{d}B_s \text{ and } X^{\mathrm{VE}} = \frac{X_0^{\mathrm{VE}}}{\bar{\phi}(\bar{\beta}(t))} + \int_0^t \frac{\bar{\phi}(s)}{\bar{\phi}(\bar{\beta}(t))}\sqrt{\frac{\mathrm{d}\bar{\beta}(s)}{\mathrm{d}s}}\mathrm{d}B_s,$$

where

$$\bar{\phi}(t) = \exp\left(-\frac{1}{2}\int_0^t \frac{\mathrm{d}\bar{\beta}(s)}{\mathrm{d}s}\mathrm{d}s\right) = \exp\left(-\frac{1}{2}\bar{\beta}(t)\right).$$

Furthermore, if we now consider $u = \bar{\beta}(s) \Rightarrow \mathrm{d}u = \frac{\mathrm{d}\bar{\beta}(s)}{\mathrm{d}s}\mathrm{d}s$, then

$$\begin{aligned}\mathbb{E}\left[\left|\left|\int_0^t \bar{\phi}(s)\sqrt{\frac{\mathrm{d}\bar{\beta}(s)}{\mathrm{d}s}}\mathrm{d}B_s\right|\right|^2\right] &\overset{\text{Itô Iso.}}{=} \int_0^t \bar{\phi}(s)^2\frac{\mathrm{d}\bar{\beta}(s)}{\mathrm{d}s}\mathrm{d}s \\ &= \int_0^{\bar{\beta}(t)} \exp(-u)\mathrm{d}u \overset{\text{Itô Iso.}}{=} \mathbb{E}\left[\left|\left|\int_0^{\bar{\beta}(t)} \exp\left(-\frac{1}{2}s\right)\mathrm{d}B_s\right|\right|^2\right].\end{aligned}$$

Therefore, the variance of $X_t^{\mathrm{VP}}|X_0^{\mathrm{VP}} = X_0$ matches the variance of $X^2_{\bar{\beta}(t)}|X_0^2 = X_0$, and both processes also share the same mean. Hence, we conclude with similar reasoning as in Item 1. □

Finally, these examples SDEs all lead to an explicit distribution of X_t conditioned on X_0. We show this in the following proposition by deriving the expression for $\bar{\gamma}$, which remains the only unknown in Equation 2.4.

Proposition 2.3. *Let $X = (X_t)_{t\geq 0}$ be a process that is driven by a noising SDE of the form Equation 2.2 and let $\bar{\gamma}$ be defined as in Equation 2.4, then:*

1. *If the driving SDE is a simple SDE-1 (Definition 2.1), we have that*

$$\bar{\gamma}(t) = t.$$

2. *If the driving SDE is a simple SDE-2 (Definition 2.1), we have that*

$$\bar{\gamma}(t) = \frac{1}{2}\left(1 - e^{-t}\right).$$

3. *If the driving SDE is a variance exploding SDE (Definition 2.3), we have*

$$\bar{\gamma}(t) = \bar{\beta}(t).$$

4. *If the driving SDE is a variance-preserving SDE (Definition 2.4), we have*

$$\bar{\gamma}(t) = \frac{1}{2}\left(1 - e^{-\bar{\beta}(t)}\right).$$

Proof. Items 1 and 3 follow directly from the properties of the Itô Integral. Items 2 and 4 can be derived by first noting that the processes are Ornstein-Uhlenbeck processes and are therefore solved by a form that resembles Equation 2.3. Then, the result also follows from the properties of the Itô Integral. □

2.2. Denoising Process

The fundamental theorem underlying generative diffusion models is a time-reversal theorem that Anderson introduced in 1982 [And82]. For generative diffusion, the purpose is to find a denoising process, denoted by Y, that specifies the reverse-time dynamics of the noising process X. Specifically, what we mean by this is that $Y_{T-t} \overset{d}{=} X_t$ for all $0 \leq t \leq T$, which says that the marginal distributions are identical for all t. This is in contrast to the stricter statement that the joint distribution of the processes is equal, i.e. $\{X_t\}_{t=0}^{T} \overset{d}{=} \{Y_{T-t}\}_{t=0}^{T}$. The strength of the latter definition is not necessary for generative diffusion models, as we are ultimately interested in modeling the marginal distribution of process Y at time T.

In writing the denoising process, two possible formulations are possible. First, if we consider the noising SDE from Equation 2.2, then it turns out that the corresponding reverse-time SDE can be written as

$$\mathrm{d}\overleftarrow{X}_t = \left[\bar{\alpha}(t)\overleftarrow{X}_t - \bar{\sigma}^2(t)\nabla \log p_{X_t}(\overleftarrow{X}_t)\right]\mathrm{d}t + \bar{\sigma}(t)\mathrm{d}\overleftarrow{B}_t, \qquad \overleftarrow{X}_T \sim \mathbb{P}_{X_T}. \tag{2.5}$$

This specifies a process where time runs from T to 0 as opposed to from 0 to T. Furthermore, $\overleftarrow{B}$ is a reverse-time Brownian motion. However, a second possibility, which we will adopt, is based on constructing a different process Y that has the same distribution, but has the standard interpretation of time running from 0 to T. Because, for the remainder of the work, we rarely consider the process X, this formulation is favourable as it promotes clarity. The SDE that drives the process Y is written as

$$\mathrm{d}Y_t = -\left[\bar{\alpha}(T-t)Y_t - \bar{\sigma}^2(T-t)\nabla \log p_{X_{T-t}}(Y_t)\right]\mathrm{d}t + \bar{\sigma}(T-t)\mathrm{d}B_t, \quad Y_0 \sim \mathbb{P}_{X_T} \tag{2.6}$$

for which it can be shown to that $X_t \overset{d}{=} \overleftarrow{X}_{T-t} = Y_{T-t}$. This is because the Brownian increments are symmetrically distributed, the $\mathrm{d}t$ term is negated, and occurrences of t are replaced with $T - t$. The following theorem summarizes the above and gives the form of the denoising SDE that we use throughout this work.

Theorem 2.1 (Denoising Process (based on Anderson's time reversal theorem [And82]))**.** *Consider X to be the noising process driven by the SDE of Equation 2.2, and consider the associated denoising SDE that is given by*

$$\mathrm{d}Y_t = \left[-\alpha(t)Y_t + \sigma^2(t)\nabla \log p_{X_{T-t}}(Y_t)\right]\mathrm{d}t + \sigma(t)\mathrm{d}B_t, \qquad Y_0 \sim \mathbb{P}_{X_T} \tag{2.7}$$

where p_{X_t} specifies the marginal density of $(X_t)_{t\geq 0}$ and

$$\alpha(t) = \bar{\alpha}(T-t), \text{ and } \sigma(t) = \bar{\sigma}(T-t). \tag{2.8}$$

(Equation 2.7 is obtained by using these new defined coefficients from Equation 2.8.) Furthermore, assume that a unique strong solution $Y = (Y_t)_{t=0}^T$ to the denoising SDE exists. Then

$$Y_{T-t} \overset{d}{=} X_t \text{ for all } 0 \leq t \leq T,$$

Proof. In the proof below, we derive the form of the reverse SDE. For a more formal derivation, the reader is directed to [And82]. The proof is based on a connection between the SDE for the noising process X and the Kolmogorov backward equation for the transition density $p_{X_s|X_t=x}(y)$ with $s > t$. Using this correspondence, we search for a similar equation for the transition $p_{X_s|X_t=x}(y)$ with $t > s$ such that upon reversing time, this corresponds to the Kolmogorov backward equation (KBE) of the reverse-time denoising process. Let b denote the drift of the process X. The Kolmogorov backward equation for $s > t$ is given by

$$-\partial_t\, p_{X_s|X_t=x}(y) \overset{\text{KBE}}{=} -b(t,x)\nabla_x p_{X_s|X_t=x}(y) - \frac{1}{2}\sigma^2(t)\Delta_x p_{X_s|X_t=x}(y),$$

where we use that the diffusion coefficient is scalar and state-independent, and the gradient operators act on x and not on y. The Kolmogorov forward equation (KFE) is given by

$$-\partial_t p_{X_t}(x) \overset{\text{KFE}}{=} \nabla\cdot(b(t,x)p_{X_t}(x)) - \frac{1}{2}\sigma^2(t)\Delta p_{X_t}(x),$$

where we can also uses that the diffusion coefficient is scalar and state-independent. We are now working towards time derivative of the joint density of X_t and X_s, that we denote with $p_{X_s,X_t}(y,x)$. Note that this can be written as

$$p_{X_s|X_t=x}(y) = \frac{p_{X_s,X_t}(y,x)}{p_{X_t}(x)}. \tag{2.9}$$

Then the partial derivative with respect to t is obtained as follows with a chain rule:

$$\partial_t p_{X_s,X_t}(y,x) \overset{Eq.\ 2.9}{=} \partial_t(p_{X_s|X_t=x}(y)p_{X_t}(x)) \overset{\text{Prod. Rule}}{=} \underbrace{p_{X_t}(x)\partial_t p_{X_s|X_t=x}(y)}_{\text{(I)}} + \underbrace{p_{X_s|X_t=x}(y)\partial_t p_{X_t}(x)}_{\text{(II)}}.$$

By Lemma 2.1, we can write the above in a form that describes the Kolmogorov forward equation in reverse time. The SDE in Equation 2.7 can be shown to be uniquely associated with Equation 2.10. □

Lemma 2.1. *Let X be a noising process that is driven by the following SDE*

$$\mathrm{d}X_t = b(t,X_t) + \sigma(t)\mathrm{d}B_t.$$

Let p_{X_s,X_t} denote joint density of X_t and X_s, then

$$-\partial_t p_{X_s,X_t} = \nabla\cdot((b - \sigma^2\nabla\log p_{X_t})p_{X_t,X_s}) + \frac{1}{2}\sigma^2\Delta p_{X_t,X_s}, \tag{2.10}$$

where we have abbreviated the arguments of the functions.

Proof. The proof is found in Section A.1. □

2.2.1. Examples of Denoising Processes

Again, we consider the four variants of Section 2.1. The following proposition describes the denoising processes associated with the respective noising process.

Proposition 2.4. *Let $X = (X_t)_{t\geq 0}$ be a process that is a driven by a noising SDE of the form in Equation 2.2, then:*

1. *If the driving SDE is a simple SDE-1 (Definition 2.1), we have that*
$$\mathrm{d}Y_t = \frac{1}{2}\nabla \log p_{X_{T-t}}(Y_t)\mathrm{d}t + \mathrm{d}B_t.$$
2. *If the driving SDE is a simple SDE-2 (Definition 2.1), we have that*
$$\mathrm{d}Y_t = \left[\frac{1}{2}Y_t + \nabla \log p_{X_{T-t}}(Y_t)\right]\mathrm{d}t + \mathrm{d}B_t.$$
3. *If the driving SDE is a variance exploding SDE (Definition 2.3), we have*
$$\mathrm{d}Y_t = \left[\frac{\mathrm{d}\beta(t)}{\mathrm{d}t}\nabla \log p_{X_{T-t}}(Y_t)\right]\mathrm{d}t + \sqrt{\frac{\mathrm{d}\beta(t)}{\mathrm{d}t}}\mathrm{d}B_t.$$
4. *If the driving SDE is a variance preserving SDE (Definition 2.4), we have*
$$\mathrm{d}Y_t = \left[\frac{1}{2}\sqrt{\frac{\mathrm{d}\beta(t)}{\mathrm{d}t}}Y_t + \frac{\mathrm{d}\beta(t)}{\mathrm{d}t}\nabla \log p_{X_{T-t}}(Y_t)\right]\mathrm{d}t + \sqrt{\frac{\mathrm{d}\beta(t)}{\mathrm{d}t}}\mathrm{d}B_t.$$

Proof. The proof follows directly by computing the coefficients of the denoising SDE in Equation 2.7 given the coefficients of the noising SDE in Equation 2.2. □

2.3. Score Matching

The score function $\nabla \log p_{X_{T-t}}$ that we use to describe the reverse process in Equation 2.7, is not known in practice. We must construct an approximate function s, e.g., by a neural network, and minimize the discrepancy between s and $\nabla \log p_{X_t}$. Ideally, we would do this by minimizing the L_2 error. This procedure is called Explicit Score Matching (ESM). Specifically, we want to find a suitable function s, such that for all $t \in [0, T]$, the following objective is small:

$$J_{ESM}(s) = \frac{1}{T}\int_0^T \mathbb{E}\left[||s(t, X_t) - \nabla \log p_{X_t}(X_t)||^2\right]\mathrm{d}t. \tag{2.11}$$

Because we do not know $\nabla \log p_{X_t}$, it is impossible to compute the objective directly. Therefore, score matching is done by using a trick introduced by Pascal Vincent in 2011 [Vin11] called Denoising Score Matching (DSM). Utilizing the analytical properties of the noising process, we can separate the score matching objective in terms of an expectation over the distribution of $X_0 \sim \mathbb{P}_{\text{data}}$ and the conditional distribution of $X_t|X_0 \sim \mathbb{P}(X_t \in \cdot|X_0 = x_0)$, which poses a useful form that can be used for practical optimization:

$$J_{DSM}(s) = \frac{1}{T}\int_0^T \mathbb{E}\left[\mathbb{E}\left[||s(t, X_t) - \nabla \log p_{X_t}(X_t|X_0)||^2|X_0\right]\right]\mathrm{d}t. \tag{2.12}$$

The inner expectation is taken over $X_t|X_0$ and the outer expectation is taken over X_0. The minimizer of this objective minimizes the objective in Equation 2.11 under certain mild assumptions [Vin11] and can be practically used by replacing the expectations with Monte Carlo estimates of X_0 by drawing from the data set, and $X_t|X_0$, by simulating the noising processes.

The exact matching of scores can rarely be achieved. This results in what is referred to as the score approximation error, i.e., the term in Equation 2.1. Obtaining an accurate score approximation is often done using neural networks optimized with J_{DSM}. However, to isolate this research on the conditioning of generative diffusion from specific choices in neural network architectures, we rely on analytical score functions. This can be precisely done by limiting our work to the following two examples. Our approach can be applied directly to generative diffusion models and non-exact score approximations in a practical setting. Still, the statistical performance can only be as good as the trained neural networks allow.

2.3.1. Examples

In certain special cases of the data distribution $\mathbb{P}_{X_0}$, the distribution of X_t is known, such as those in Example 2.1 and Example 2.2.

Example 2.1 ($\nabla \log p_{X_{T-t}}$ for Gaussian data)**.** *If $X_0 \sim \mathcal{N}(\mu_0, \Sigma_0)$, and $(X_t)_{t\geq 0}$ is driven by the SDE defined in Equation 2.2, then*

$$X_t \sim \mathcal{N}(\mu_t, \Sigma_t),$$

where

$$\mu_t = \frac{1}{\bar{\phi}(t)}\mu_0 \qquad \text{and} \qquad \Sigma_t = \frac{1}{\bar{\phi}(t)}\Sigma_0 + \bar{\gamma}(t) I_{d\times d}.$$

The associated reverse process is determined by

$$\nabla \log p_{X_{T-t}}(x) = -\Sigma_{T-t}^{-1}(x - \mu_{T-t}). \tag{2.13}$$

To see why this is the case, we know that the solution to the forward process is given by Equation 2.3. From this, it follows that $X_t|X_0 = x_0 \sim \mathcal{N}\left(\frac{x_0}{\bar{\phi}(t)}, \bar{\gamma}(t) I_{d\times d}\right)$. But then using the distribution of X_0, we find that

$$X_t \sim \mathcal{N}\left(\frac{\mu_0}{\bar{\phi}(t)}, \frac{1}{\bar{\phi}(t)}\Sigma_0 + \bar{\gamma}(t) I_{d\times d}\right).$$

The next example we consider is that of a Gaussian mixture model.

Example 2.2 ($\nabla \log p_{X_{T-t}}$ for Gaussian mixture data)**.** *If $X_0 \sim \sum_{k=1}^K w_k \mathcal{N}(\mu_0^{(k)}, \Sigma_0^{(k)})$ with $(w_1, \dots w_k) \subset [0,1]^k$ such that $\sum_{k=1}^K w_k = 1$, and the mean vector is $\mu_0^{(k)}$ and covariance matrix is $\Sigma_0^{(k)}$, then*

$$X_t \sim \sum_{k=1}^K w_k \mathcal{N}(\mu_t^{(k)}, \Sigma_t^{(k)}), \text{ and } p_{X_t}(x) = \sum_{k=1}^K w_k p_{X_t}^{(k)}(x), \text{ where } p_{X_t}^{(k)}(x) \overset{\text{def}}{=} \mathcal{N}\left(x; \mu_t^{(k)}, \Sigma_t^{(k)}\right).$$

where

$$\mu_t^{(k)} = \frac{1}{\bar{\phi}(t)}\mu_0^{(k)} \qquad \text{and} \qquad \Sigma_t^{(k)} = \frac{1}{\bar{\phi}(t)}\Sigma_0^{(k)} + \bar{\gamma}(t) I_{d\times d}.$$

Then the associated reverse process is determined by the following term $\nabla \log p_{X_{T-t}}(x) = \frac{\nabla p_{X_{T-t}}(x)}{p_{X_{T-t}}(x)}$, where we use the log derivative trick. Now, note that by the chain rule

$$\nabla p_{X_{T-t}}^{(k)}(x) = p_{X_{T-t}}^{(k)}(x) \nabla \log p_{X_{T-t}}^{(k)}(x).$$

From this, we find the following expression

$$\nabla \log p_{X_{T-t}}(x) = \sum_{k=1}^K \frac{w_k p_{X_{T-t}}^{(k)}(x)}{p_{X_{T-t}}(x)} \underbrace{\nabla \log p_{X_{T-t}}^{(k)}(x)}_{\text{See Equation 2.13}}. \tag{2.14}$$

This essentially says that the function amounts to a weighted contribution of all the K Gaussian distributions, where the contributions are determined by the factor $\frac{w_k p_{X_{T-t}}^{(k)}(x)}{p_{X_{T-t}}(x)}$. Here w_k represents the prior probability of being drawn from the kth Gaussian, $q_{T_t}^{(k)}(x)$ is the probability of being drawn from the kth Gaussian given x, while $p_{X_{T-t}}(x)$ is the probability of x.

3

Controlled Generative Diffusion

In this chapter, we discuss the control of conditioned paths of diffusion models to satisfy certain conditions at time T with the purpose of generating from conditional data distributions. A naive approach is to simulate paths of the denoising process and accept or reject paths based on whether their value at time T actually satisfies our specified condition. However, as the conditions become rarer, which intrinsically happens as the state space is of high dimensionality, the probability of obtaining a satisfactory sample vanishes.

Therefore, a form of guidance is required to steer the paths towards satisfying the condition at time T. Typically, these approaches are inconsistent in the sense that the sampled paths do not resemble paths from the true conditioned denoising process. We underline this discrepancy by studying the derivation of an exact conditioned denoising process, as if we were to have access to the exact score function of the noising process with an initial distribution that does satisfy the condition. Specifically, this can be done by replacing the unconditional score function $\nabla \log p_{X_{T-t}}$ in the standard denoising SDE with a score function that is associated to the conditioned noising process, which we denote by

$$\mathrm{d}Y_t^* = \left[\alpha(t)Y_t^* + \sigma^2(t)\nabla \log p_{X_{T-t}|LX_0=v}(Y_t^*)\right]\mathrm{d}t + \sigma(t)\mathrm{d}B_t, \qquad \text{with } Y_0^* \sim \mathbb{P}_{X_T}. \tag{3.1}$$

Simply simulating this SDE is not possible due to the difficultly in approximating the conditional score function

$$\nabla \log p_{X_{T-t}|LX_0=v}(x),$$

without additional training with data sets that are specifically oriented around the condition. Furthermore, it turns out that we can relate the above SDE to an SDE that is obtained with Doob's h-transform (see e.g. [PR02]), i.e.,

$$\mathrm{d}Y_t^* = \left[\underbrace{\alpha(t)Y_t^* + \overbrace{\sigma^2(t)\nabla \log p_{X_{T-t}}(Y_t^*)}^{\sigma^2(t)\nabla \log p_{X_{T-t}|LX_0=v}(Y_t^*)} + \sigma^2(t)\nabla \log h(t, Y_t^*)}_{b(t,Y_t^*)}\right]\mathrm{d}t + \sigma(t)\mathrm{d}B_t, \text{ with } Y_0^* \sim \mathbb{P}_{X_T}. \tag{3.2}$$

where $h(t,x)$ is the density of $\mathbb{P}(LY_T \in \cdot | Y_t = x)$ evaluated at v, for some suitable $m \times d$ matrix L and $v \in \mathbb{R}^d$. Furthermore, b denotes the unconditional drift, and the decomposition of the conditional score function will be elaborated on in the remainder of this chapter. The function h is based on the generally intractable density of LY_T given Y_t. To circumvent the intractability, one often uses approximation techniques. Typically, the approximate guidance is attached as a heuristic term to the drift coefficient of the unconditional denoising process Y that steers the diffusion towards states that do satisfy the condition at time T.

In Section 3.1, we discuss a primary perspective on controlled data generation that is more acquainted in the context of generative diffusion, which is based on extracting the conditional score given an unconditional one. In Section 3.2, we describe a formal procedure of conditioning diffusions in a more traditional sense using Doob's h-transform. In Section 3.3 we describe two known methods to approximate the conditioned generative diffusion within our mathematical framework: replacement guidance [Lug+22] and reconstruction guidance [SE20].

3.1. Conditional Score Function

The conditional score function describes the underlying relation that is used for practically all training-based and training-free controlled generative diffusion techniques. For the training-based methods, this expression is used during training-time to learn the conditional score. For training-free methods, the expression is used to obtain a heuristic approximation of the conditional score given the pre-trained unconditional score.

We consider a data distribution of X_0 conditioned on $LX_0 = v$ for some set $v \in \mathbb{R}^m$ and full rank matrix $L \in \mathbb{R}^{m\times d}$ with $m < d$. Note that for some $V \subseteq \mathbb{R}^d$, we have that the by adopting the notation of the density

$$\mathbb{P}(LX_0 \in V) = \int_V \left(\int_{Lx=v} p_{X_0}(x)\mathrm{d}x \right) \mathrm{d}v.$$

Because L is full rank, we may consider writing any $x \in \mathbb{R}^d$ as $x = U_1\xi_1 + U_2\xi_2$. where U_1 is a $d \times m$ matrix where the m columns form an orthonormal basis of the column span of L and the $d-m$ columns of U_2 form an orthonormal basis of the null space of L. Let us consider the set $\{x \in \mathbb{R}^d : Lx = v\}$, then specifically it holds that for all x that $LU_1\xi_1 = v$ and $LU_2\xi_2 = 0$. Therefore the density of $\mathbb{P}(LX_0 \in \cdot)$ evaluates to

$$p_{LX_0}(v) = \int_{Lx=v} p_{X_0}(x)\mathrm{d}x = \int_{\mathbb{R}^{d-m}} p_{X_0}(U_1\xi_1 + U_2\xi_2)\mathrm{d}\xi_2.$$

Then, using the above notation, we may also write the function h, evaluated at v as

$$h(t,x) = \int_{\mathbb{R}^{d-m}} p_{Y_T|Y_t=x}\left(U\xi_1 + U\xi_2\right)\mathrm{d}\xi_2, \tag{3.3}$$

which is the density of the measure $\mathbb{P}(LY_T \in \cdot | Y_t = x)$.

We will consider the unconditional score that we denote by $\nabla \log p_{X_{T-t}|LX_0=v}$, and we arrive at a conditioned denoising process by replacing $\nabla \log p_{X_{T-t}}$ in Equation 2.7. It turns out that the process obtained with the conditional score function resembles Doob's h-transform under certain circumstances. The following proposition shows that if we use the interpretation of the conditional score function $\nabla \log p_{X_{T-t}|LX_0=v}$, we obtain an identical conditioned denoising SDE as specified in Equation 3.2. The proposition is based on the assumption that Y_0 is exactly sampled from the distribution of the forward process at time t. The fundamental difference in the derivations is that the starting point of the conditional score derivation is by considering the distribution of X_0, which then relies on the exactness of the noising and denoising processes. In contrast, the derivation with Doob's h-transform is solely based on process Y, and requires no further assumptions about the exactness of the system.

Proposition 3.1. *Assume that Y is a solution to the denoising SDE of Equation 2.7. Assume that the initial distribution of Y is exactly that of X_T, i.e., $Y_0 \sim \mathbb{P}_{X_T}$. Consider that we condition on $LX_0 = v$, or equivalently $LY_T = v$. Then, the conditional score function is*

$$\nabla \log p_{X_{T-t}|LX_0=v}(x) = \nabla \log p_{X_{T-t}}(x) + \nabla \log h(t,x),$$

where h is defined as Equation 3.3

Proof. For some $t \in [0,T]$, we have by Bayes' theorem that the score function of the marginal of the process conditional on $LX_0 = v$ can be written as

$$\nabla \log p_{X_{T-t}|LX_0=v}(x) = \nabla \log \frac{p_{X_{T-t},LX_0}(x,v)}{p_{LX_0}(v)} = \nabla \log p_{X_{T-t}}(x) + \nabla \log \mathbb{P}(LX_0 \in V | X_{T-t} = x),$$

where we have used that $\nabla \log p_{LX_0}(v) = 0$ as the gradient is taken w.r.t x and we have adapted the notation of $p_{X_{T-t},LX_0}(x,v)$ to denote a joint density of X_{T-t} and LX_0. Then if $Y_0 \sim \mathbb{P}_{X_T}$, we have that $Y_T \sim \mathbb{P}_{X_0}$ by Theorem 2.1, so we have our result □

3.2. Doob's h-transform

Now, we connect the result of the previous section to a more general viewpoint. Specifically, Doob's h-transform is a technique that allows continuous-time Markov processes to be conditioned [PR02]. Consider the measure of the path Y given that Y_T satisfies some condition, i.e.

$$\mathbb{P}(Y \in \cdot | LY_T = v).$$

Doob's h transform is a change of measure operation based on the function h. Specifically, we consider the following (random) function:

$$E^h(t) = \frac{h(t, Y_t)}{h(0, Y_0)}.$$

In particular, for our specific choice of h, the process $E^h(t)$ is a martingale and induces a likelihood ratio process with which we can define a new probability measure $\mathbb{P}^*$ by

$$\mathrm{d}\mathbb{P}^*|_{\mathcal{F}_t} = E^h(t)\mathrm{d}\mathbb{P}|_{\mathcal{F}_t}.$$

Furthermore, the process Y has under the law $\mathbb{P}^h$ as infinitesimal generator $\mathcal{L}_t^h$ that satisfies the following relation with $\mathcal{L}_t$:

$$(\partial_t + \mathcal{L}_t^*)f = \frac{1}{h}(\partial_t + \mathcal{L}_t)fh.$$

Because in our work, we only use a single form of h, we call $\mathbb{P}^* \overset{\text{def}}{=} \mathbb{P}^h$ and $\mathcal{L}_t^* \overset{\text{def}}{=} \mathcal{L}_t^h$. In the remainder of the section, we will discuss a few aspects of the technique that we described above. First, we will address under which circumstances the function $E^h(t)$ induces a change of measure and that it gives us the transformed SDE of Equation 3.4. Then, we will address that the changed measure is actually the conditioned measure, that we have defined as $\mathbb{P}^*(\cdot) \overset{\text{def}}{=} \mathbb{P}(\cdot | LY_T = v)$.

Before doing so, we show that E^h is indeed a martingale by studying the stochastic process $(h(t, Y_t))_{t\geq 0}$.

Proposition 3.2. *Let h be as in Equation 3.3 and assume that it lies in the domain of the infinitesimal generator. Let Y be the solution to the denoising SDE given in Equation 2.7, then the stochastic process $(h(t, Y_t))_{t\geq 0}$ is a Martingale w.r.t. $\mathcal{F}_t$.*

Proof. First note that $h(t, Y_t)$ is measurable w.r.t. $\mathcal{F}_t$ because Y_t is measurable w.r.t. $\mathcal{F}_t$ and h is a continuous function, because we assume it lies in the domain of the infinitesimal generator. We show that $h(t, Y_t)$ is a martingale with respect to the natural filtration $\{\mathcal{F}_t\}_{t\geq 0}$. Now, we pick any time s such that $T > s > t$, then the property follows from

$$\begin{aligned}
\mathbb{E}\left[h(s, Y_s)|\mathcal{F}_t\right] &= \int_{\mathbb{R}^d} p_{Y_s|Y_t}(y) \underbrace{\left(\int_{\mathbb{R}^{d-m}} p_{Y_T|Y_s=y}\left(U\xi_1 + U\xi_2\right)\mathrm{d}\xi_2\right)}_{h(s,y)} \mathrm{d}y \\
&= \int_{\mathbb{R}^{d-m}} p_{Y_T|Y_t}\left(U\xi_1 + U\xi_2\right)\mathrm{d}\xi_2 \\
&= h(t, Y_t),
\end{aligned}$$

where the second equality is due to Chapman-Kolmogorov. □

From this result, a particularly useful property of h in relation to the infinitesimal generator follows.

Proposition 3.3 (Space-time harmonic h)**.** *Let $\mathcal{L}_t$ denote the infinitesimal generator of process Y and assume that h in the domain of the infinitesimal generator, which it is if it is twice differentiable w.r.t x and once differentiable with respect to t, then*

$$\partial_t h + \mathcal{L}_t h = 0.$$

We refer to this property as h being space-time harmonic w.r.t. $\mathcal{L}_t$.

Proof. Because $(h(t, Y_t))_{t=0}^T$ is a martingale (Lemma 3.2), we know that the SDE that drives the process $(h(t, Y_t))_{t\geq 0}$ has zero drift. Furthermore, by Itô's formula we have that the zero drift implies that

$$\frac{\partial}{\partial t}h(t,x) + \mathcal{L}_t h(t,x) = 0.$$

Therefore, we recognize the definition of space-time harmonic property of h. □

The above integral definition of h in Equation 3.3 is quite cumbersome and prevents us from computing anything meaningful, such as $\nabla \log h$. Fortunately, for a particular class of processes, i.e., those with Gaussian transition densities, the transform reduces to a simple explicit form, which we show in the following Lemma.

Lemma 3.1 (Gaussian transition density)**.** *Assume that we can write the transition density of process Y as*

$$p_{Y_T|Y_t=x}(y) = \mathcal{N}\left(y; \mu_T(t,x), \Sigma_T(t)\right),$$

for some vector valued function $\mu_T : [0,T] \times \mathbb{R}^d \to \mathbb{R}^d$ and positive definite matrix-valued function $\Sigma_T : [0,T] \to \mathbb{R}^{d\times d}$. Then if L is a $m \times d$ matrix such that $L\Sigma_T(t)L^\top$ is invertible, the h function of Equation 3.3 is given by

$$h(t,x) = \mathcal{N}(v; L\mu_T(t,x), L\Sigma_T(t)L^\top).$$

Proof. To simplify our notation, we use the following notation $\mu := \mu_T(t,x)$ and $\Sigma := \Sigma_T(t)$. If $Y \sim \mathcal{N}(\mu, \Sigma)$, then Y can be written as $Y = \mu + \Sigma^{1/2}Z$ where $Z \sim \mathcal{N}(0, I)$ and $\Sigma^{1/2}$ denotes the matrix square root that is defined for positive definite matrices. Then, we have that $LY = L\mu + L\Sigma^{1/2}Z$. Therefore, if $L\Sigma L^\top$ is invertible, we have that $LY \sim \mathcal{N}(L\mu, L\Sigma L^\top)$. □

Now, we consider the following proposition, which describes the transformed SDE using a Girsanov transformation and establishes the relation to E^h.

Theorem 3.1 (SDE under Doob's h-transform)**.** *Assume that h is a space-time harmonic function with respect to $\mathcal{L}$. The transformed SDE that is induced by the change-of-measure $E^h(t)$, is given by*

$$\mathrm{d}Y_t^* = \underbrace{[b(t, Y_t^*) + \sigma^2(t)\nabla \log h(t, Y_t^*)]}_{b^*(t, Y_t^*)}\,\mathrm{d}t + \sigma(s)\mathrm{d}B_t', \tag{3.4}$$

where B' is a Brownian motion under $\mathbb{P}^*$.

Proof. Our proof is based on an application of Girsanov's theorem. We know that $h(t, Y_t)$ is a martingale by Lemma 3.2. Specifically, we know that it is a solution to the following SDE

$$\mathrm{d}(h(t, Y_t)) = \sigma(t)\nabla h(t, Y_t) \cdot \mathrm{d}B_t,$$

where we use Itô's formula and the fact that h is space-time harmonic w.r.t. $\mathcal{L}_t$, which makes the drift evaluate to zero. Now, consider that

$$\mathrm{d}E^h(t) = \frac{1}{h(0, Y_0)}\sigma(t)\nabla h(t, Y_t) \cdot \mathrm{d}B_t = E^h(t)\sigma(t)\nabla \log h(t, Y_t) \cdot \mathrm{d}B_t,$$

where we use the identity $h(t, Y_t)\nabla \log h(t, Y_t) = \nabla h(t, Y_t)$ in the last equality. By considering that the stochastic exponential (Doléans-Dade exponential) solves the SDE, we obtain that

$$E^h(t) = \exp\left(\int_0^t \sigma(s)\nabla \log h(s, Y_s) \cdot \mathrm{d}B_s - \frac{1}{2}\int_0^t ||\sigma(s)\nabla \log h(s, Y_s)||_2^2 \mathrm{d}s\right).$$

Now, we note that by Girsanov's theorem, we have that under the measure $\mathbb{P}^*$ a new Brownian motion B_t' is defined by

$$\mathrm{d}B_t' = \mathrm{d}B_t - \sigma(t)\nabla \log h(t, Y_t)\mathrm{d}t.$$

Substituting $\mathrm{d}B_t$ with the above identity into the denoising SDE Equation 2.7, gives us

$$\mathrm{d}Y_t = b(t, Y_t)\mathrm{d}t + \sigma(t)\mathrm{d}B'_t + \sigma^2(t)\nabla \log h(t, Y_t)\mathrm{d}t.$$

This SDE specifies an Itô process under measure $\mathbb{P}^*$ where $(B'_t)_{t\geq 0}$ is a Brownian motion and thus gives us the transformed SDE in Equation 3.4 that specifies the transformed Itô process, denoted by Y^*. □

The infinitesimal generators of the process Y under the two different measures are associated in the following way. By the unique characterization of the generator, the following proposition completes the goal of this section.

Proposition 3.4. *Let us define the following infinitesimal generator $\mathcal{L}_t^*$ such that it satisfies*

$$(\partial_t + \mathcal{L}_t^*)f(t,x) = \lim_{s\downarrow t} \frac{\mathbb{E}\left[f(s, Y_s)|Y_t = x, LY_T = v\right] - f(t, Y_t)}{s-t}.$$

Here, the expectation in Definition 1.2 is replaced with a conditional expectation. Then the following holds:

1. $(\partial_t + \mathcal{L}_t^*)f = \frac{1}{h}(\partial_t + \mathcal{L}_t)fh$
2. *$\mathcal{L}_t^*$ is the infinitesimal generator of the conditioned SDE given in Equation 3.4.*

Proof. The proof is found in Section A.2. The outline of the proof is as follows. Item 1 can be shown by rewriting the conditional expectation, such that it resembles the desired form. Then, item 2 follows simply by writing out the infinitesimal generators. □

3.2.1. Examples

The examples we discuss below explicitly define conditioned stochastic processes. These examples are rather specific cases. In particular, the h-transform can rarely be written explicitly because of the unknown transition densities.

Example 3.1 (One-Dimensional Brownian Bridge)**.** *Let B_t be a d-dimensional standard Brownian motion process and assume that we condition $B_t = v$. Then, $h(t,x)$ is written as*

$$h(t,x) = \mathcal{N}(v; x, T-t) \propto \exp\left(-\frac{||v-x||^2}{2(T-t)}\right),$$

where we make use of the Gaussian increments of standard Brownian motion. Therefore, the conditioned process is driven by the following SDE

$$\mathrm{d}B_t^* = \frac{v - B_t^*}{T-t}\mathrm{d}t + \mathrm{d}B_t.$$

This is seen by applying Theorem 3.1 and using the fact that

$$\nabla \log h(t,x) = \frac{(v-x)}{T-t},$$

in combination with $\sigma^2(t) = 1$.

Example 3.2 (Multi-Dimensional Brownian Motion with Linear Condition)**.** *Let B_t be a d-dimensional standard Brownian motion process and assume we condition on $LB_T = v$ for some $m \times d$ matrix L for which $LL^\top$ is invertible. Then, $h(t,x)$ can written as*

$$h(t,x) = \mathcal{N}(v; Lx, LL^\top(T-t)).$$

Then this gives us

$$\mathrm{d}B_t^* = L^\top(LL^\top)^{-1}\frac{v - LB_t^*}{T-t}\mathrm{d}t + \mathrm{d}B_t.$$

This is seen by applying Theorem 3.1 and using the fact that

$$\nabla \log h(t,x) = L^\top(LL^\top)^{-1}\frac{(v-Lx)}{T-t},$$

again in combination with $\sigma^2(t) = 1$.

Remark 3.1 (Sampling Brownian bridges)**.** *The Brownian bridges in Figure 3.1 are sampled with a shrinking discretization. This means that as $t \to T$, the time steps become smaller. Specifically, we use $t_j = (j/M)^{\frac{1}{2.5}}$ for $M = 1000$. This shrinking discretization is required to ensure that the bridges are sufficiently close to satisfying the condition, which is not guaranteed as we cannot sample exactly at time T. This problem does not affect our methods with generative diffusion, specifically the variance preserving with quadratic noise schedule (Definition 2.4), as this is designed to have a vanishing diffusion coefficient for t approaching T. This vanishing discretization is already incorporated in the SDE, which can be seen as a time change operation of a constant diffusion. We will therefore be able to use a uniform discretization of $[0, T]$ without too many issues for the remainder of this work. Moreover, this observation underlines the motivation of the time-changed SDE's such as the variance preserving SDE as opposed to the non-time changed SDE for controlling generative diffusion models.*

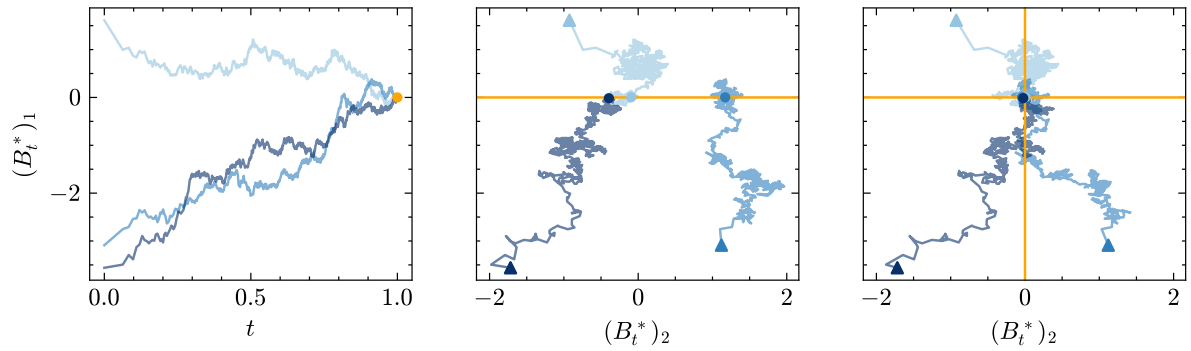


Figure 3.1: Illustration of Brownian Bridges. In the left panel we see one-dimensional Brownian bridges conditioned to hit 0 at $T = 1$. In the right most panel, we see two-dimensional Brownian bridges conditioned to hit $(0, 0)$ at time $T = 1$. In the middle panel, we see two-dimensional Brownian motions conditioned to hit the set $\{(x, 0) : x \in \mathbb{R}\}$. The triangle indicates the starting positions of the particles at time $T = 0$ and the circles indicate the final positions at time $T = 1$

3.3. Heuristic Approximations

Approximating the function h typically causes great difficulties. Therefore, the following forms of approximation are common, which we will refer to with $\hat{g}$. This approximation often induces a severe bias that does not vanish with infinite computational effort. Therefore, the need for a more sophisticated approach is evident, which is the topic for the remainder of this work. Before we turn our attention there, we briefly describe a few approximations for $\hat{g}$ in the formulation of our setting.

The overarching principle is that we use as a guidance term, the gradient of the error of the satisfaction of the condition. Specifically, we use $-\gamma_{\text{scale}}(t)||L\hat{x} - \hat{v}||^2$ for some proxy values, $\hat{x}$ that is a proxy for Y_T, and $\hat{v}$ that is a proxy for v. The scaling is chosen in such a way that, closer to the terminal time T, the error explodes and therefore the paths are drawn toward satisfying the condition at time T. Typically, the guidance scales are heuristically determined with respect to the chosen noising SDE.

3.3.1. Replacement Guidance

One way of approximating h is to consider a proxy for v, that we denote by $\hat{v} : [0, T] \to \mathbb{R}^m$, that is based on the transformed process $(V_t)_{t\geq 0}$, e.g. in our case $V_t = LX_t$. Furthermore, we use the proxy $\hat{x} = x$. Then, our approximation approach is to consider the following expression for the function $\hat{g}$

$$\log \hat{g}(t, x) = -\gamma_{\text{scale}}(t)||Lx - \hat{v}(t)||^2$$

where $\hat{v} = \mathbb{E}[V_{T-t}|V_0 = v]$ and $\gamma_{\text{scale}}(t)$ denotes a scaling factor that is determined heuristically, often inspired by the noise level induced by the forward noising process at time t. Because we know the dynamics of the noising process and the initial position of V, i.e., $V_0 = v$, we can compute the expectation of V_{T-t} relatively easily above exactly.

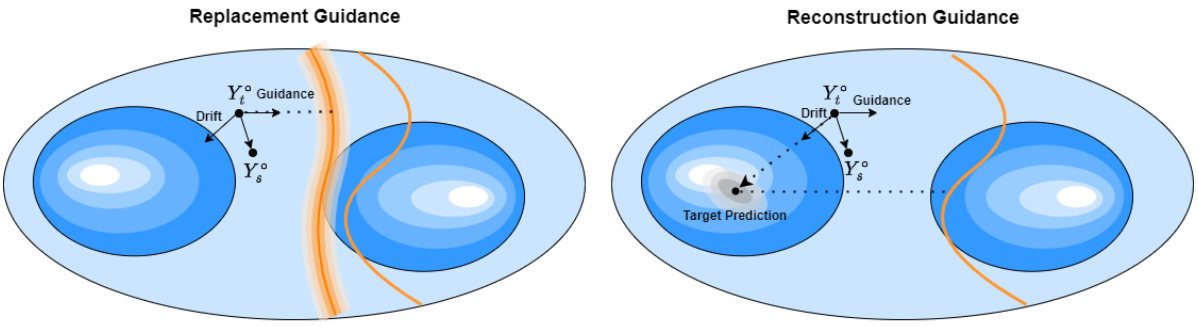


Figure 3.2: Illustration of replacement (left) and reconstruction guidance (right). The blue gradients illustrate a bimodal distribution. The orange curves depict the condition subspace. The arrow denotes the unconditional drift at (t, Y_t°) and the dotted line indicates the guidance direction. The stochastic component of the system is ignored for simplicity. **Left**: In the illustration of replacement guidance, the orange gradient around the orange curve depicts the distribution of the transformed process at time $t < T$, i.e., the distribution of V_{T-t}. The solid orange curve depicts $\mathbb{E}[V_{T-t}|V_0 = v]$. **Right:** In the illustration of the reconstruction guidance, the black dot at the end of the arrow indicates the target prediction of the process at time T. The grey gradient indicates the uncertainty of this target prediction.

3.3.2. Reconstruction Guidance

Another type of approximation is based on a reconstruction of the clean sample given a noisy state. Specifically, reconstruction guidance is based on using Tweedie's formula [Efr11] to obtain a proxy $\hat{x}$, that we denote by $\hat{x} : [0, T] \times \mathbb{R}^d \to \mathbb{R}^d$. Then, the guidance term is determined by the following approximation:

$$\log \hat{g}(t, x) = -\gamma_{\text{scale}}(t)||L\hat{x}(t, x) - v||^2,$$

where the proxy is given by

$$\hat{x}(t, x) = \bar{\phi}(T - t)\left(x + \frac{1}{\sqrt{\bar{\gamma}(T - t)}}\nabla \log p_{X_{T-t}}, (x)\right).$$

where the parameters $\bar{\phi}$ and $\bar{\gamma}$ are described by the noising SDE as in Proposition 2.1. The justification of Tweedie's formula as a proxy for x is obtained from the fact that

$$\nabla \log p_X(t, X_t; 0, X_0) = -\frac{1}{\sqrt{\bar{\gamma}(t)}}\left(\frac{X_0}{\bar{\phi}(t)} - X_t\right),$$

where the proxy is based upon the heuristic exchange of the transition density p_X with a marginal density p_{X_t}.

It is important to note that $\nabla \log \hat{h}$ should also take into account the gradient w.r.t. $\hat{x}(t, x)$. For this reason, the gradient has the following form:

$$\nabla \log \hat{g}(t, x) = -(L\nabla\hat{x}(t, x))^\top \gamma_{\text{scale}}(t)(L\hat{x}(t, x) - v),$$

which is due to an application if the chain rule. Such a propagation of the gradient is not required in replacement guidance where the proxy $\hat{v}$ does not depend on x.

Part II

Theoretical and Methodological Developments

4

Pathwise Importance Sampling

In this chapter, we describe an approach to importance sampling on the space of continuous paths, ultimately intending to correct a bias caused by inconsistent training-free controlled generative diffusion methods and ensure asymptotic consistency. In short, sample paths of a proposal process are sampled, and importance weights are assigned to the paths. The weights are then used to resample with the promise of resembling paths from the true conditioned process.

Importance sampling is typically used to obtain more efficient sampling procedures for computing expectations of rare events. Our application is to generate samples that satisfy a (possibly rare) condition using a different proposal measure. We use a specific class of guided proposals that are guaranteed to satisfy the conditions and have simple-to-derive expressions for the importance weights. These important weights enable the reweighing of proposal paths to make them resemble a set of paths as if they were sampled from the target measure.

The behavior as t approaches T raises two questions. First, we must verify that $||LY_t^\circ - v||$ indeed goes to zero as $t \to T$, such that it is a satisfactory proposal process. If Y° does not satisfy the condition, using it as a proposal process is significantly less attractive. Second, the absolute continuity of the proposal measure $\mathbb{P}^\circ$ w.r.t. the conditioned measure $\mathbb{P}^*$ at time T must be verified to justify using importance sampling. While it is relatively easy to see that the measures are absolutely continuous on $[0, T)$, extending to this $[0, T]$ is not trivial. To study these aspects, we may use a combination of techniques as in [BMS20] and in [SMZ17]. However, we can not solely rely on their results because of our slightly adjusted assumptions on the role of $\nabla\tilde{\mu}_T(t, x)$ as described in Section 1.3. Therefore, we provide additional assumptions and lemmas to support the claims.

In Section 4.1, we describe the importance sampling technique and its favorable characteristics, such as asymptotic consistency of the particle approximation and its convergence rate. Section 4.2 describes how the exact (continuous-time) importance weights can be derived. In addition, as an intermezzo to the buildup of our approach, we describe how an approximation of importance weights can be computed given a discrete-time Euler-Maruyama approximation. In Section 4.3, we study a specific class of guided proposals to derive a simplified form of Girsanov's formula. In addition, we describe how a scalar diffusion coefficient enables a simplified expression of the importance of weights in the typical context of generative diffusion models. In Section 4.4, we discuss the validity of the guided proposal. Specifically, we study the behavior of $||LY_t^\circ - v||$ as $t \uparrow T$ and the absolute continuity of the continuous-time importance weight on the interval $[0, T]$.

4.1. Importance Sampling Technique

The canonical goal of importance sampling is to estimate expectations of functionals that are otherwise difficult to compute. While the use case for conditional sampling with generative diffusion models is slightly different, as we intend to draw samples, the underlying principles are the same. The idea behind using importance sampling for the conditioning of diffusion models is to draw sample paths from the proposal measure $\mathbb{P}_Y^\circ$ and compute importance weights for the sample paths, that can later be used to weight the proposal samples such that expectations can be computed, or resample paths in its entirety to approximately match the measure $\mathbb{P}_Y^*$.

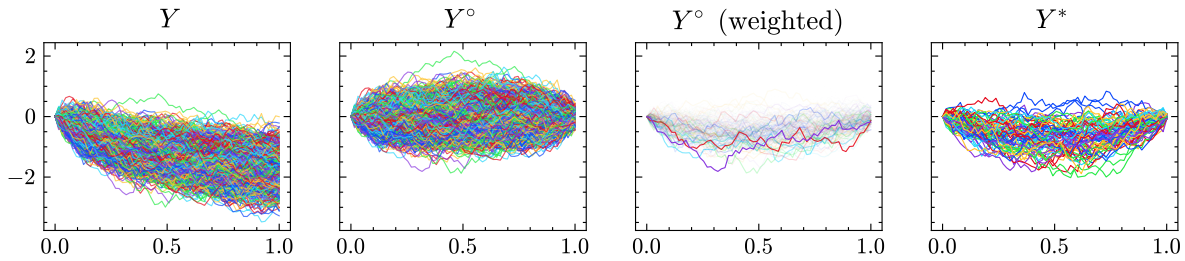


Figure 4.1: Paths of (weighted) proposal process Y°, unconditioned Y process, and conditioned process Y^*. The leftmost panel shows sample paths of an unconditional Ornstein-Uhlenbeck process. The second panel from the left shows paths of the proposal process, which is a Brownian bridge pinned down at 0 at time $T = 1$. The second panel from the right shows the weighted proposal processes. The rightmost figure shows sample paths of a true conditioned process.

Example 4.1. *In this example, we are interested in an Ornstein-Uhlenbeck (OU) process Y that takes values on $\mathbb{R}$ and is conditioned to hit 0 at time $T = 1$. The paths of the unconditioned OU process are described in the leftmost panel of Figure 4.1 and the following SDE drives them*

$$\mathrm{d}Y_t = -2(Y_t + 1)\mathrm{d}t + \mathrm{d}B_t.$$

To obtain an approximation of the measure of Y^*, we sample paths from a Brownian bridge, driven by

$$\mathrm{d}Y_t^\circ = -\frac{Y_t^\circ}{T - t}\mathrm{d}t + \mathrm{d}B_t.$$

Then, we compute the importance weights with Girsanov's formula for sample paths from the Brownian bridge, denoted by Y°. The weighted paths are displayed in the rightmost panel of Figure 4.1. In Figure 4.2 we give an impression of how importance weights can be used to resample. Specifically, we display the histograms of sampled values of $Y^*_{0.5}$ (target), $Y^\circ_{0.5}$ (proposal), and a reweighted set of samples.

Let us first consider a $Y \sim \mathbb{P}^*_Y$, a distribution from which we cannot sample. We wish to compute the following expectation $\mathbb{E}^*[f(Y)]$ for some bounded function f. The key idea is to choose $\mathbb{P}^\circ_Y$ such that: $\mathbb{P}^\circ_Y$ is absolutely continuous w.r.t. $\mathbb{P}^*_Y$, we can sample from $\mathbb{P}^\circ_Y$, and we can compute the desired Radon-Nikodym derivative. Then, one can approximate the expectation

$$\mathbb{E}^*[f(Y)] = \mathbb{E}^\circ\left[f(Y)\frac{\mathrm{d}\mathbb{P}_{Y^*}}{\mathrm{d}\mathbb{P}_{Y^\circ}}(Y)\right] \overset{\text{LLN}}{\approx} \frac{1}{N}\sum_{i=1}^{N} f(Y^{(i)})\frac{\mathrm{d}\mathbb{P}_{Y^*}}{\mathrm{d}\mathbb{P}_{Y^\circ}}(Y^{(i)}), \tag{4.1}$$

where $\{Y^{(i)}\}_{i=1}^N$ are draw from the proposal distribution, i.e. $Y^{(i)} \sim \mathbb{P}^\circ_Y$ i.i.d. for all $i \in \{1, \ldots, N\}$. We denote $\mathbb{E}^\circ[Y]$ by the expectation of Y w.r.t. $\mathbb{P}^\circ_Y$. The approximation becomes almost surely exact in the limit of large N, due to the law of large numbers, denoted by LLN in the approximation of Equation 4.1. This Radon-Nikodym derivative is also referred to as the importance weight of a sample Y, i.e.

$$W^*(Y) = \frac{\mathrm{d}\mathbb{P}^*_Y}{\mathrm{d}\mathbb{P}^\circ_Y}(Y),$$

where $\mathbb{P}^*_Y$ is called the target measure and $\mathbb{P}^\circ_Y$ is called the proposal measure. A set-wise estimator for $\mathbb{P}^*_Y$, that is known as a particle approximation, can be obtained by

$$\frac{1}{N}\sum_{i=1}^{N} W^*(Y^{(i)})\mathbf{1}_A(Y^{(i)}). \tag{4.2}$$

The particle approximation above is identical to the right-hand side of Equation 4.1 for an indicator function. It is important to see that this particle approximation is a random counting measure, which is a measure-valued random element. Therefore, to evaluate convergence results of the particle approximation, we often consider convergence in mean or in mean square error.

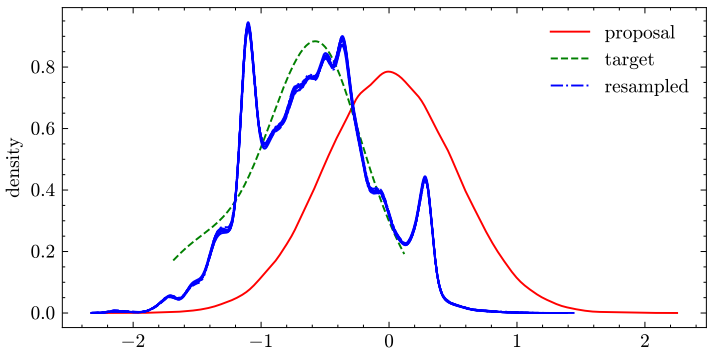


Figure 4.2: Histogram of samples of $Y_{0.5}$ and (reweighted) samples $Y^{\circ}_{0.5}$ from Example 4.1. The histograms are shown as kernel density plots for enhanced visibility. The *proposal* histogram represents samples from $Y^{\circ}_{0.5}$ and the *target* histogram represents samples from $Y_{0.5}$. The *resampled* histogram represents a resampled set of samples of $Y^{\circ}_{0.5}$, where the resample probabilities are obtained with the normalized importance weights. The large spikes in empirical density of the resampled set is due to the few particles with the relatively high importance weights to have a significantly larger presence, giving large spikes in the histogram. This weight imbalance is a common issue with importance sampling in large systems, where numerically sampled naturally diffusing processes belong, even for small state spaces, because of the dependency of the entire paths.

So far, we have shown that the particle approximation is unbiased. We now consider the variance of the particle approximation of the probability of A under $\mathbb{P}^*_Y$. The following lemma shows that for any A, the term the variance can be bounded by a term independent of the choice of A that vanishes for $N \to \infty$.

Lemma 4.1 (Importance Sampling)**.** *Let $\mathbb{E}^{\circ}\left[W^*(Y)^2\right]$ denote the second moment of the importance weight W^*, which has mean 1 and assume $\mathbb{E}^{\circ}\left[W^*(Y)^2\right] \lesssim 1$. Then the following bound holds for any measurable set A, such that the bound is independent of A*

$$\mathbb{E}^{\circ}\left[\left(\frac{1}{N}\sum_{i=1}^{N}\mathbf{1}_A(Y)W^*(Y) - \mathbb{P}^*_Y(A)\right)^2\right] \leq \frac{1}{N}\mathbb{E}^{\circ}\left[(W^*(Y))^2\right] = \mathcal{O}\left(\frac{1}{N}\right),$$

where the expectation is w.r.t. a set of N independent samples drawn from the proposal measure $\mathbb{P}^{\circ}_Y$

Proof. The proof is found in Section A.3 □

It is, however, the case that we cannot compute W^*, due to its dependency on the conditioned probability measure $\mathbb{P}^*_Y$. We can compute some other weight W that satisfies $W \propto W^*$, where the proportionality is independent of the argument of W. This is fundamental to our approach, as will become clear in the next sections. The weight W gives us a slightly different estimator for $\mathbb{P}^*_Y(A)$, which is called the self-normalized particle approximation,

$$P^*_N(A) \overset{\text{def}}{=} \frac{\frac{1}{N}\sum_{i=1}^{N} W(Y^{(i)})\mathbf{1}_A(Y^{(i)})}{\frac{1}{N}\sum_{i=1}^{N} W(Y^{(i)})}. \tag{4.3}$$

While it can be easily derived that this definition also converges to $\mathbb{P}^*_Y$ almost surely, the estimator with finite particles is biased. No closed-form for the bias exists for the general case, however the induced bias can be intuitively understood from the non-linearity of the summation in the denominator of the above equation.

Proposition 4.1 (Self-Normalizing Importance Sampling)**.** *Let $P^*_N(A)$ be defined as in Equation 4.3 and assume $\frac{\mathbb{E}^{\circ}\left[(\mathbb{E}^{\circ}[W]-W)^2\right]}{(\mathbb{E}^{\circ}[W])^2} \lesssim 1$. Then the following holds for all measurable A such that the bound is independent of A:*

$$\mathbb{E}^{\circ}\left[(P^*_N(A) - \mathbb{P}^*_Y(A))^2\right] = \mathcal{O}\left(\frac{1}{N}\right).$$

Proof. The proof is found in Section A.4 □

4.2. Continuous-Time Importance Weights

For computing weights of a process on the entire time span $[0, T]$, we need to find an expression for the following Radon-Nikodym derivative

$$W_T^* := \frac{\mathrm{d}\mathbb{P}_Y^*}{\mathrm{d}\mathbb{P}_Y^\circ}\Big|_{\mathcal{F}_T}.$$

Before we turn our attention to computing these importance weights, we describe a decomposition of the Radon-Nikodym derivative. Consider that the above, if it is defined, can be decomposed into the product of two different Radon-Nikodym derivatives. Specifically, if $Y_0 = y_0 \in \mathbb{R}^d$, then it follows from the previous chapter that

$$W_t^* = \frac{\mathrm{d}\mathbb{P}_Y^*}{\mathrm{d}\mathbb{P}_Y^\circ}\Big|_{\mathcal{F}_t} = \frac{\mathrm{d}\mathbb{P}_Y^*}{\mathrm{d}\mathbb{P}_Y}\Big|_{\mathcal{F}_t}\frac{\mathrm{d}\mathbb{P}_Y}{\mathrm{d}\mathbb{P}_Y^\circ}\Big|_{\mathcal{F}_t} = \frac{h(t, Y_t)}{h(0, y_0)}\frac{\mathrm{d}\mathbb{P}_Y}{\mathrm{d}\mathbb{P}_Y^\circ}\Big|_{\mathcal{F}_t} = \frac{h(t, Y_t)}{h(0, y_0)}W_t.$$

where

$$W_t := \frac{\mathrm{d}\mathbb{P}_Y}{\mathrm{d}\mathbb{P}_Y^\circ}\Big|_{\mathcal{F}_t}.$$

If we choose the proposal processes to satisfy $LY_T^\circ = v$, we have that $h(T, x) = 1$. Therefore, we can write

$$W_T^* = \frac{1}{h(0, y_0)}W_T \propto W_T. \tag{4.4}$$

If we fix y_0, we can omit the term $h(0, y_0)$ in the importance sampler as it is identical for all weights and is therefore canceled out with a simple weight normalization. This justifies the last proportional relation in the above equation.

We use Girsanov's theorem to derive the importance weight. Specifically, let b denote the drift of the SDE that drives process Y and b° the drift of the SDE that drives our proposal Y°, then note that the diffusion coefficient is a non-zero scalar function on $[0, T]$, and therefore, we may write

$$\eta(t, x) = \frac{b^\circ(t, x) - b(t, x)}{\sigma(t)}$$

Then if η satisfies Novikov's condition, the Radon-Nikodym derivative is given by

$$W_t = \frac{\mathrm{d}\mathbb{P}_Y}{\mathrm{d}\mathbb{P}_Y^\circ}\Big|_{\mathcal{F}_t}(Y^\circ) = \exp\left(\int_0^t \eta(s, Y_s^\circ)^\top \mathrm{d}B_s - \frac{1}{2}\int_0^t ||\eta(s, Y_s^\circ)||^2 \mathrm{d}s\right) \tag{4.5}$$

and induces a change of measure between the proposal process Y° and the unconditional process Y. The formula in Equation 4.5 boils down to a convenient form if we consider the following proposal processes.

$$\mathrm{d}Y_t^\circ = \left[b(t, Y_t^\circ) + \sigma^2(t)\nabla \log \tilde{h}(t, Y_t^\circ)\right]\mathrm{d}t + \sigma(t)\mathrm{d}B_t, \tag{4.6}$$

where $\tilde{h}$ is a tractable approximation of function in Equation 3.3. A beneficial aspect of this form of proposals, is that they incorporate information about the unconditional process through the presence of the unconditional drift term b. In fact, because of the way we have defined b°, we can write

$$\eta(t, x) = \sigma(t)\nabla \log \tilde{h}(t, x) = \sigma(t)\tilde{r}(t, x),$$

where we use the notation $\tilde{r}(t, x) := \nabla \log \tilde{h}(t, x)$.

4.2.1. Intermezzo: The Limit of Discretized Importance Weights

As an intermezzo of the buildup of our method, we elaborate on importance sampling with a discretization of the processes instead. This is previously done in the context of generative diffusion in, for example, [Wu+24]. Understanding such approaches is useful in underlining the conceptual differences and similarities between our approach and some of the existing work.

Typically, the sample paths of diffusions are approximated at finite times, for example due to use of an Euler-Maruyama (EM) approximation, that is defined using the following the following discretization of $[0, T]$, $0 = t_0 < \cdots < t_M = T$, where $t_i = i\frac{T}{M}$. The density of a sampled discrete path $\{\widehat{Y}_{t_i}\}_{i=1}^M$ using the EM approximation is given by

$$\hat{p}(\{\widehat{Y}_{t_i}\}_{i=1}^M) = \prod_{i=1}^M \hat{p}(\widehat{Y}_{t_i}; \widehat{Y}_{t_{i-1}}) \quad \text{where} \quad \hat{p}(\widehat{Y}_s; \widehat{Y}_t) = \mathcal{N}\left(\widehat{Y}_s; \widehat{Y}_t + b(t, \widehat{Y}_t)(s-t), \sigma^2(t)(s-t)\right).$$

Now, the task at hand is to consider the weights that are accumulated by taking the product of M ratios of approximated transition densities

$$\widehat{W}^* = \prod_{i=1}^M \frac{\hat{p}^*(\widehat{Y}_{t_i}; \widehat{Y}_{t_{i-1}})}{\hat{p}^\circ(\widehat{Y}_{t_i}; \widehat{Y}_{t_{i-1}})}. \tag{4.7}$$

Here the transition densities are similarly defined as to $\hat{p}(y_s; y_t)$, but in coherence with the drift coefficients b° and b^* of the proposal and target (conditioned) SDEs that we defined in Chapter 3. Using similar reasoning as for the continuous time weights, the discrete weight can be written in terms of the approximate densities of the proposal process and the unconditional process, i.e.

$$\widehat{W}^* = \frac{1}{h(0, \widehat{Y}_0)} \prod_{i=1}^M \frac{\hat{p}(\widehat{Y}_{t_i}; \widehat{Y}_{t_{i-1}})}{\hat{p}^\circ(\widehat{Y}_{t_i}; \widehat{Y}_{t_{i-1}})} \propto \prod_{i=1}^M \frac{\hat{p}(\widehat{Y}_{t_i}; \widehat{Y}_{t_{i-1}})}{\hat{p}^\circ(\widehat{Y}_{t_i}; \widehat{Y}_{t_{i-1}})} = \widehat{W}. \tag{4.8}$$

Because of the notation of $\widehat{W}^*$ in Equation 4.7, the computation of the approximate (self-normalizing) weights is possible. In particular, we understand the dynamics of the unconditioned and proposal processes. The weights are no longer unbiased due to the bias induced by the Euler-Maruyama, which only vanishes as the discretization becomes infinitely fine-grained, i.e., $M \to \infty$.

Asymptotic Behaviour of Discrete-time Importance Weights

To substantiate the intuition behind the continuous-time importance weight we discuss in the next section, we can look into an illustrative derivation of the Girsanov formula as the convergence of the discrete-time approximation from Equation 4.8.

Lemma 4.2. *Consider $\widehat{W}$ as defined in Equation 4.8, then*

$$\log \widehat{W} = -\sum_{i=1}^M \frac{1}{2} ||\eta_{t_{i-1}}||^2 \frac{T}{M}(t_i - t_{i-1}) + \sum_{i=1}^M \eta_{t_{i-1}} \cdot (B_{t_i} - B_{t_{i-1}}).$$

Proof. See Section A.5 □

Given that $\eta_t = \eta(t, Y_t)$, we have that under regularity conditions, $\log \hat{W}$ from Section 4.2 converges to the log Radon-Nikodym derivative as described in Equation 4.5,

$$\log_{M\to\infty} \log \widehat{W} = -\frac{1}{2}\int_0^T ||\eta_t||^2 \mathrm{d}t + \int_0^T \eta_t \cdot \mathrm{d}B_t,$$

if we use the definition of the Riemann integral and the Itô integral.

4.3. Auxiliary-Guided Proposal Process

The proposal processes in Equation 4.6 are based on a tractable approximation of the function h. Specifically, they are evaluated for some auxiliary process $\tilde{Y} = (\tilde{Y}_t)_{t=0}^T$ that is designed such that we can do the conditioning analytically. We start with the critical characterization of the auxiliary process that enables this analytical tractability, and then in the next chapter, we will give some more concrete examples. At an abstract level, the SDE that drives the auxiliary processes is of the following form:

$$\mathrm{d}\tilde{Y}_t = \tilde{b}(t, \tilde{Y}_t)\mathrm{d}t + \sigma(t)\mathrm{d}B_t, \tag{4.9}$$

for some value $\tilde{b}(t,x) : [0,T] \times \mathbb{R}^d \to \mathbb{R}^d$ and $\sigma(t)$ the diffusion coefficient of the corresponding unconditional denoising SDE as given in Equation 2.7. The defining property of the auxiliary processes is given below.

Definition 4.1 (Defining property of auxiliary processes)**.** *We refer to the processes of the form of Equation 4.9 to belong to the class of auxiliary processes if*

$$\tilde{Y}_T | \tilde{Y}_t \sim \mathcal{N}(\tilde{\mu}_T(t,x), \tilde{C}_T(t)),$$

where $\tilde{\mu}_T : [0,T] \times \mathbb{R}^d \to \mathbb{R}^d$ is a vector function that represents the conditional expectation of $\tilde{Y}_T | \tilde{Y}_t = x$ and $\tilde{C}_T : [0,T] \to \mathbb{R}^{d \times d}$ is a matrix-valued function that specifies the covariance.

From this point, we reserve the term auxiliary process for processes that satisfy the above property. For these auxiliary processes, the gradient log term is explicitly expressed as a direct corollary of Lemma 3.1.

Corollary 4.1 (Auxiliary guidance)**.** *Assume that $\tilde{Y}$ is an auxiliary process and that $\tilde{h}$ is defined by*

$$\tilde{h}(t,x) = \mathbb{P}(L\tilde{Y}_T \in \mathrm{d}v | \tilde{Y}_t = x)/\mathrm{d}v$$

where we use the tilde to denote that we derived for an auxiliary process as opposed to a general process Y with unknown transition densities, then for a full rank matrix $L \in \mathbb{R}^{m \times d}$ with $d > m$, we have that

$$\nabla \log \tilde{h}(t,x) = \nabla \tilde{\mu}(t,x) L^\top (L\tilde{C}_T(t)L^\top)^{-1}(L\tilde{\mu}_T(t,x) - v) \tag{4.10}$$

Note that in Equation 4.10, we have taken the gradient of $\tilde{\mu}(t,x)$ w.r.t. x, due to an application of the chain rule. In many cases of the drift coefficient $\tilde{b}$ the gradient will evaluate to a trivial form, such as a constant 1 if $\tilde{\mu}(t,x) = x$ that follows from a driftless auxiliary process. However, as will become clear in the next chapter (Chapter 5), there is a subtlety in choosing an (adaptive) auxiliary drift, which makes the gradient of $\tilde{\mu}(t,x)$ non-trivial.

Now, it turns out, as shown in [SMZ17], that this specific choice of proposals Y°, in combination with the auxiliary process $\tilde{Y}$, enables a simplified form of the continuous-time importance weights that reduces to an exponential of a Riemann-Stieltjes integral. This integral can consequently be approximated with a Riemann sum on a discretized grid and yields an arguably more convenient form as it does not involve a stochastic integral that may be more difficult to approximate numerically. The fundamental difference between this discretization and the discretization depicted in the intermezzo of the previous section is that here we first derive an exact weight and then discretize, as opposed to discretizing and then deriving the inexact weights.

Proposition 4.2 (Radon-Nikodym derivative (adapted from [SMZ17]))**.** *Let $\tilde{h}$ be space-time harmonic w.r.t. $\tilde{\mathcal{L}} = \{\tilde{\mathcal{L}}_t\}_{t=0}^T$, where $\tilde{\mathcal{L}}$ is the (time-dependent) infinitesimal generator of auxiliary process $\tilde{Y}$. Define the following function*

$$\psi(t) = \exp\left(\int_0^t G(s, Y_s^\circ)\mathrm{d}s\right) \quad \text{where } G(t,x) = (b(t,x) - \tilde{b}(t,x)) \cdot \nabla \log \tilde{h}(t,x) \tag{4.11}$$

Then for $t \in [0,T)$ the laws $\mathbb{P}|_{\mathcal{F}_t}, \mathbb{P}^\circ|_{\mathcal{F}_t}$ and $\mathbb{P}^*|_{\mathcal{F}_t}$ are equivalent and we have

$$\frac{\mathrm{d}\mathbb{P}_Y}{\mathrm{d}\mathbb{P}_Y^\circ}(Y^\circ)\bigg|_{\mathcal{F}_t} = \frac{\tilde{h}(0,y_0)}{\tilde{h}(t, Y_t^\circ)}\psi(t) \tag{4.12}$$

Proof. The proposition and the proof here are, in fact, a special case of (Proposition 1, [SMZ17]). For completeness, we provide the derivation of the simplified formula, which is unique to our setting. Let us define $\tilde{r}_s = \tilde{r}(s, Y_s) = \nabla \log \tilde{h}(s, Y_s)$ and $\tilde{R}_s = \tilde{R}(s, Y_s) = \log \tilde{h}(s, Y_s)$, such that $\nabla \tilde{R}_s = \tilde{r}_s$. Furthermore, we may write $b_t = b(t, Y_t^\circ)$, also for the drift coefficients b° and $\tilde{b}$. The infinitesimal generator of Y° applied to $\tilde{R}$ evaluates to

$$\mathcal{L}_t^\circ \tilde{R}_t = \mathcal{L}_t \tilde{R}_t + \sigma^2(t) \tilde{r}_t \cdot \tilde{r}_t.$$

This follows from the fact that the diffusion coefficients are identical and the drift of Y° satisfies $b^\circ(t, x) = b(t, x) + \sigma^2(t)\tilde{r}(t, x)$. Then

$$\mathcal{L}_t^\circ \tilde{R}_t = (b_t + \sigma^2(t)\tilde{r}_t) \cdot \tilde{r}_t + \frac{1}{2}\sigma^2(t)\Delta \tilde{R}_t = \underbrace{b_t \cdot \tilde{r}_t + \frac{1}{2}\sigma^2(t)\Delta \tilde{R}_t}_{\mathcal{L}_t \tilde{R}_t} + \sigma^2(t)\tilde{r}_t \cdot \tilde{r}_t.$$

Then, we can apply Itô's formula to obtain an expression for $\tilde{R}(t, Y_t)$, i.e.,

$$\tilde{R}_t - \tilde{R}_0 = \int_0^t \left(\frac{\partial}{\partial s} \tilde{R}_s + \mathcal{L} \tilde{R}_s \right) \mathrm{d}s + \int_0^t \sigma^2(s) ||\tilde{r}(s, Y_s^\circ)||_2^2 \mathrm{d}s + \int_0^t \sigma(s) \tilde{r}(s, Y_s^\circ) \cdot \mathrm{d}B_s.$$

The Radon-Nikodym derivative in the left hand side of Equation 4.12 is obtained by the Girsanov formula, with $\eta_t = \sigma(t)^{-1}(b_t - b_t^\circ) = \sigma(t)\tilde{r}_t$. This gives us

$$\log \left. \frac{\mathrm{d}\mathbb{P}_Y}{\mathrm{d}\mathbb{P}_Y^\circ} \right|_{\mathcal{F}_t} = -\int_0^t \sigma(s)\tilde{r}(s, Y_s^\circ) \mathrm{d}B_s - \frac{1}{2}\int_0^t \sigma^2(s) ||\tilde{r}(s, Y_s^\circ)||_2^2 \mathrm{d}s.$$

Therefore, comparing the two above equations, we can write the log Radon-Nikodym derivative without stochastic integrals in the following way

$$\log \left. \frac{\mathrm{d}\mathbb{P}_Y}{\mathrm{d}\mathbb{P}_Y^\circ} \right|_{\mathcal{F}_t} = -(\tilde{R}_t - \tilde{R}_0) + \int_0^t \left(\frac{\partial}{\partial s} \tilde{R}_s + \mathcal{L}_t \tilde{R}_s \right) \mathrm{d}s + \frac{1}{2}\int_0^t \sigma^2(s) ||\tilde{r}(s, Y_s^\circ)||_2^2 \mathrm{d}s.$$

Now, we consider that the function G can be written as

$$G(t, x) = (b(t, x) - \tilde{b}(t, x)) \cdot \tilde{r}(t, x) = (\mathcal{L}_t - \tilde{\mathcal{L}}_t)\tilde{R}_t = \frac{\partial}{\partial t}\tilde{R}_t + \mathcal{L}_t \tilde{R}_t + \frac{1}{2}\sigma^2(t) ||\tilde{r}(s, Y_s^\circ)||_2^2$$

Here, in the second equality we have used the definitions of the infinitesimal generators $\mathcal{L}$ and $\tilde{\mathcal{L}}$. In the third equality, we have used that:

$$\frac{\partial}{\partial t}\tilde{R}_t + \tilde{\mathcal{L}}_t \tilde{R}_t = -\frac{1}{2}\sigma^2(t) ||\tilde{r}(s, Y_s^\circ)||_2^2. \tag{4.13}$$

This identity is due to using the log derivative trick and product rule to write

$$\tilde{\mathcal{L}}_t \tilde{R}_t = \tilde{b}_t \cdot \nabla \tilde{R}_t + \sigma^2(t)\Delta \tilde{R}_t = \frac{1}{\tilde{h}_t}\tilde{b}_t \cdot \tilde{h}_t + \frac{1}{\tilde{h}_t}\frac{1}{2}\sigma^2(t)\Delta \tilde{h}_t - \frac{1}{2}\sigma^2(t)\tilde{r}_t \cdot \tilde{r}_t \quad \text{and} \quad \frac{\partial}{\partial t}\tilde{R}_t = \frac{1}{\tilde{h}_t}\frac{\partial}{\partial t}\tilde{h}_t.$$

Then because $\tilde{h}$ is space-time harmonic, the identity follows. Finally, the claim of the proposition follows by noting that $-(\tilde{R}_t - \tilde{R}_0) = -\log \tilde{h}(t, x)/\tilde{h}(0, x_0)$. □

Remark 4.1 (Simplified form of ψ)**.** *In [SMZ17], the function ψ is given in a different more expansive form. In our case, we consider a scalar and state-independent diffusion coefficient of the SDE that drives Y. This enables us to choose the identical diffusion coefficient among all processes $Y, \tilde{Y}, Y^\circ$, which reduces their result to a simplified expression of Equation 4.11.*

4.4. Behavior at Terminal Time

At this point, we know how to derive the importance weights for any $t < T$. However, studying the behavior of the proposal processes at the terminal time is important for two reasons. First, we study whether the proposals Y° satisfy that LY_t° approaches v as $t \uparrow T$. Otherwise, the proportionality of Equation 4.4 does not hold. Second, the absolute continuity of the laws of the process Y^* and Y° still needs to be established on $[0, T]$ because Proposition 4.2 only holds for $[0, T)$.

Our proofs differ from those in [BMS20], which studies hypoelliptic diffusions, in that we rely on a different available technique due to the uniform ellipticity of the diffusions we study. In particular, we take inspiration from the proof in [SMZ17]. The difference with their proof is that we focus on the general case where $L \neq I$, whereas their work focuses on $L = I$. Furthermore, our assumptions are significantly different from those of both works, which are more appropriate for the adaptive setting we study in the next chapter. Specifically, we explicitly consider the gradient w.r.t. $\tilde{\mu}_T$ as described in Equation 4.10; the results of these existing works do not directly transfer. The idea of the proofs are similar, however certain parts are easier in our setting because we do not have state dependent diffusion coefficients in our unconditional process Y, while other parts are harder due to the dependency on $\nabla\tilde{\mu}_T(t, x)$.

4.4.1. Condition Satisfaction of Y_T°

Before we start, we must make a few assumptions about the auxiliary process $\tilde{Y}$, and in particular the function $\tilde{\mu}_T(t, x)$ and $\tilde{r}$.

Assumption 4.1 (Properties of auxiliary process)**.** *Let L be a full rank $m \times d$ matrix with $d > m$. Then the auxiliary process $\tilde{Y}$, that is driven by the SDE in Equation 4.9, is chosen such that the following properties hold for all $x \in \mathbb{R}^d$ and $t \in [0, T]$:*

1. *$\partial_t \log \tilde{h}$ and $\nabla \log \tilde{h}$ are continuous.*
2. *The Jacobian of the auxiliary drift is bounded, i.e. $||\nabla\tilde{b}(t, x)|| \lesssim 1$.*
3. *The function $\tilde{\mu}_T(t, x)$ approaches x, i.e., $||\tilde{\mu}_T(t, x) - x|| \lesssim (T - t)$.*
4. *$\tilde{\mu}_T$ has a bounded Jacobian $||\nabla\tilde{\mu}_T(t, x)|| \lesssim 1$.*
5. *Let $\tilde{R}(t, x) := \log \tilde{h}(t, x)$. Then, let us call*

$$H(t, x) = \nabla^2 \tilde{R}(t, x),$$

   where ∇^2 specifies the Hessian operator. We assume that $||H(t, x)|| \lesssim (T - t)^{-1}$ and $||H(t, x) - (T - t)^{-1} I|| \lesssim 1$.

6. *Let us define the matrix-valued function $Q : [0, T] \times \mathbb{R}^d \to \mathbb{R}^{d \times m}$ by*

$$Q(t, x) = (\nabla\tilde{\mu}(t, x))^\top L^\top \left(L\tilde{C}_T(t)L^\top\right)^{-1}$$

   such that

$$\nabla \log \tilde{h}(t, x) = (\nabla\tilde{\mu}(t, x))^\top L^\top \left(L\tilde{C}_T(t)L^\top\right)^{-1} (L\tilde{\mu}_T(t, x) - v) = Q(t, x)(L\tilde{\mu}_T(t, x) - v),$$

   Then there exists a matrix-valued function $Q^+ : [0, T] \times \mathbb{R}^d \to \mathbb{R}^{m \times d}$ such that

$$Q^+(t, x)Q(t, x) = I_{m \times m}.$$

   Furthermore, $||Q^+(t, x)|| \lesssim (T - t)$ and $||Q(t, x)|| \leq (T - t)^{-1}$.

A separate assumption that we make is about the combined behavior of the drift coefficients of the auxiliary process and the unconditional process.

Assumption 4.2. *There exists an almost sure finite real random variable V_1 such that*

$$\sup_{t\in[0,T]} ||b(t, Y_t^\circ) - \tilde{b}(t, Y_t^\circ)|| < V_1$$

This assumption requires that the proposal process Y° is well behaved on the interval $[0, T]$. Specifically, consider that $\sup_{t\in[0,T]} ||Y_t^\circ|| < \infty$ implies the assumption by making use of the linear growth of both of these drift coefficients. However, it is not entirely obvious whether this assumption on Y_t° contradicts any of the other assumptions or leads to circular reasoning of the to-be-proven claim, especially when considering Gronwall inequalities.

The theorem below suggests that with the above assumptions, we can control the behavior of $||LY_t^\circ - v||$ such that it vanishes as $t \uparrow T$. The proof, which is given in Subsection 4.4.3, starts by noting that if we want to minimize the quantity $||Lx - v||$, we may study instead $||\tilde{r}(t, x)||$ (Lemma 4.3). For this, we derive an SDE that governs the stochastic process $\tilde{r}(t, Y_t^\circ)$, that we denote by $\tilde{r}_t$ (Lemma 4.4). Then, we rewrite the coefficients of the SDE to find a convenient expression, of which the norm must be bounded. A problematic term is $\nabla^2 \tilde{R}_t$, supposedly an explosive term as $t \to T$, illustrated in the following example. For this reason, dealing with this term is not trivial and therefore takes some significant effort.

Example 4.2. *Consider the setting of Theorem 4.1 and specifically, the case where $\tilde{\mu}_T(t, x) = x$ and $\tilde{C}_T(t) = T - t$. Then the function $\tilde{r}(t, x)$ is*

$$\tilde{r}(t, x) = \nabla \log \tilde{h}(t, x) = L^\top (L\tilde{C}_T(t)L^\top)^{-1}(Lx - v).$$

Then, we have that the Hessian of $\tilde{R}$ is

$$||H(t, x)|| = ||L^\top (L\tilde{C}_T(t)L^\top)^{-1}L|| \propto (T - t)^{-1}.$$

The right-hand side explodes as $t \to T$.

Theorem 4.1 (Condition Satisfaction of Y°)**.** *Let Assumption 4.1 and Assumption 4.2 hold. Then there exists $\epsilon \in (0, \frac{1}{2})$ and an almost sure finite random variable V such that for all $t \in [0, T]$ it holds that*

$$||LY_t^\circ - v|| \leq V(T - t)^{1/2-\epsilon}.$$

4.4.2. Absolute Continuity on $[0, T]$

The Radon-Nikoydm derivative in proposition 4.2 holds only on $[0, T)$. Therefore, to derive a well-behaved method, we must verify that the formula can also be used near time T. This is, however, not trivial, as the behavior of the proposal process near time T may hinder the equivalence of the measures $\mathbb{P}^*$ and $\mathbb{P}^\circ$.

As in [BMS20], we must assume the transition density of the unconditioned process.

Assumption 4.3. *Let $p_{Y_s|Y_t=x}$ and $p_{\tilde{Y}_s|\tilde{Y}_t=x}$ denote the transition densities of Y and $\tilde{Y}$, in that the sense they specify the density of the state at time s conditioned on the state at time t being x. Then there exists a constant $C > 0$ such that for all $t, s \in [0, T]$ and $x, y \in \mathbb{R}^d$.*

$$p_{Y_s|Y_t=x}(y) \leq C p_{\tilde{Y}_s|\tilde{Y}_t=x}(y)$$

Theorem 4.2. *Let Assumption 4.1, Assumption 4.2 and Assumption 4.3 hold and fix $Y_0 = y_0 \in \mathbb{R}^d$. Then the laws of the bridges Y^* and Y° are equivalent on $[0, T]$, the formula in Equation 4.12 holds for all $t \in [0, T]$, and specifically*

$$\frac{\mathrm{d}\mathbb{P}_Y^*}{\mathrm{d}\mathbb{P}_Y^\circ}(Y)\bigg|_{\mathcal{F}_T} = \frac{\tilde{h}(0, y_0)}{h(0, y_0)}\psi(T).$$

4.4.3. proof of Theorem 4.1

The proof that follows is quite tedious and relies on subtle tricks that are available by our assumptions. For a deeper understanding of these tricks, it is advised to inspect the proofs in [BMS20] and in [SMZ17] alongside the proofs here. In particular, we rely on Lemma 4.5 that is a combination of lemmas 14-16 from [SMZ17].

Proof. We know by Lemma 4.3 that $||LY_t^\circ - v|| \lesssim (T-t)(1+||\tilde{r}(t,Y_t^\circ)||)$, so what remains is to find an upper bound for the latter. Specifically, we verify that there exists $\epsilon \in (0,1/2)$ and a finite random variable V such that $||\tilde{r}(t,Y_t^\circ)|| \lesssim V(T-t)^{\epsilon-1}$.

Preparation for Lemma 4.5. By Lemma 4.4, we that $\tilde{r}(t,Y_t^\circ)$ is driven by the following SDE

$$\mathrm{d}\tilde{r}(t,Y_t^\circ) = \left[-\tilde{r}(t,Y_t^\circ)\cdot\nabla\tilde{b}(t,Y_t^\circ) + H(t,Y_t^\circ)(\tilde{b}(t,Y_t^\circ) - b(t,Y_t^\circ))\right]\mathrm{d}t + \sigma(t)H(t,Y_t^\circ)\mathrm{d}B_t.$$

Now, we are preparing to apply Lemma 4.5. To do this, we need to write the term $-\nabla\tilde{b}$ in a specific way, i.e.,

$$\begin{aligned}
K^1(t,x) &= I(T-t)^{-1}\left(1-\frac{\sigma^2(t)}{\sigma^2(T)}\right),\\
K^2(t,x) &= \left(H(t,x) - I(T-t)^{-1}\right)\left(1-\frac{\sigma^2(t)}{\sigma^2(T)}\right),\\
K^3(t,x) &= H(t,x)\left(\frac{\sigma^2(t)}{\sigma^2(T)}-1\right) - \nabla\tilde{b}(t,x).
\end{aligned}$$

Observe that $K := K^1 + K^2 + K^3 = -\nabla\tilde{b}$. Using this notation, we can derive the following:

- **Bounding $(T-t)||\cdot||^2_{K^1}$.** For all $t\in[0,T)$ and $z\in\mathbb{R}^d$, we have that there exists an $\epsilon\in(0,1/2)$ such that the following inequality holds

$$||z||^2_{K^1} = z^\top K^1(t,Y_t^\circ)z = \frac{\left(1-\frac{\sigma^2(t)}{\sigma^2(T)}\right)||z||^2}{T-t} \le \frac{(1-\epsilon)||z||^2}{T-t},$$

  where we use that there exist $\epsilon_0,\epsilon_1,\epsilon_2>0$ such that $\sigma^2(t)\ge\epsilon_0\ge\epsilon_1$, $\sigma^2(T)\le\epsilon_2$, and we pick ϵ_1 small enough such that $\epsilon=\epsilon_1/\epsilon_2\in(0,1/2)$. For example $\epsilon_1=\sigma^2(T)/4$ and $\epsilon_2=\sigma^2(T)$ whenever $\sigma(T)\le\sigma(t)$ for all $t\in[0,T]$.

- **Bounding K^2.** By our assumption on H and the boundedness of $|1-\frac{\sigma^2(t)}{\sigma^2(T)}|$, we know that

$$||K^2(t,x)|| = ||\left(H(t,x)-I(T-t)^{-1}\right)\left(1-\frac{\sigma^2(t)}{\sigma^2(T)}\right)|| \lesssim 1.$$

- **Bounding K^3.** Note that by our assumption $||\nabla\tilde{b}(t,x)||$ is bounded. Therefore, by Lipschitz continuity and boundedness of σ, we can write the following using our assumption on H

$$\begin{aligned}
||K^3(t,x)|| &= ||H(t,x)\left(\frac{\sigma^2(t)}{\sigma^2(T)}-1\right) - \nabla\tilde{b}(t,x)||\\
&\le ||H(t,x)||\frac{1}{\sigma(T)}|\sigma^2(t)-\sigma(T)| + ||\nabla\tilde{b}(t,x)||\\
&\lesssim \frac{1}{T-t}(T-t)+1.
\end{aligned}$$

Now, we use the following notation

$$W(t,x) = H(t,x)(\tilde{b}(t,Y_t^\circ) - b(t,Y_t^\circ)).$$

We can also derive the following:

- **Bounding** $(T-t)||W||$**.** By our assumption on $||\tilde{b}(t,Y_t^\circ) - b(t,Y_t^\circ)||$ and our assumption on H, we have that there exists a finite random variable V_1 such that

$$(T-t)||W(t,x)|| \leq (T-t)||H(t,x)||\,|V_1| \lesssim (T-t)\frac{1}{T-t}V_1 = V_1.$$

Now, we may write the SDE as

$$\mathrm{d}\tilde{r}_t = \left[\underbrace{(K_t^1 + K_t^2 + K_t^3)}_{K_t}\tilde{r}_t + W_t\right]\mathrm{d}t + U_t \mathrm{d}B_t, \tag{4.14}$$

where $U_t := \sigma(t)H(t,Y_t^\circ)$ that satisfies the following.

- **Bounding** $(T-t)||U_t||$**.** By using our assumption on H it directly follows that

$$(T-t)||U_t|| \lesssim 1.$$

Applying Lemma 4.5. We consider the setting of Lemma 4.5. Let us adopt the following notation:

$$K^{23} = K^2 + K^3 \text{ and } K = K^1 + K^{23}.$$

Then, we study the random linear system, described by the following linear ODE:

$$\mathrm{d}\Phi(t) = K(t,Y_t^\circ)\Phi(t)\mathrm{d}t.$$

The matrix function $\Phi(t)$ exists uniquely because $t \to K(t,Y_t^\circ)$ is continuous for each realization Y° (we use the same justification as in [SMZ17]). Furthermore, we have that for any $z \in \mathbb{R}^d$ that K^1 satisfies

$$z^\top K^1(t,Y_t^\circ)z \leq \frac{(1-\epsilon)||z||^2}{T-t}$$

and K^2 is bounded. This allows us to use Lemma 4.5. Specifically, by item 2 of Lemma 4.5, we may write

$$\tilde{r}(t,Y_t^\circ) = \Phi(t)\tilde{r}(0,Y_0^\circ) + \Phi(t)\int_0^t \Phi(s)^{-1}W(s,Y_s^\circ)\mathrm{d}s - \Phi(t)\int_0^t \Phi(s)^{-1}U(s)\mathrm{d}B_s.$$

From this, it follows that we must compute a bound for

$$||\tilde{r}(t,Y_t^\circ)|| \leq \underbrace{||\Phi(t)\tilde{r}(0,Y_0^\circ)||}_{\text{(I)}} + \underbrace{\int_0^t ||\Phi(t)\Phi(s)^{-1}||\,||W(s,Y_s^\circ)||\mathrm{d}s}_{\text{(II)}} + \underbrace{\left|\left|\int_0^t \Phi(t)\Phi(s)^{-1}U_s\mathrm{d}B_s\right|\right|}_{\text{(III)}}.$$

We now proceed with bounding each of the three terms above

1. **(I):** By item 1 of Lemma 4.5 and finiteness of $||Y_0^\circ|| = ||y_0||$, we have that $||\tilde{r}(0,Y_0^\circ)|| \lesssim 1$. Then we find that
$$\text{(I)} \lesssim (T-t)^{\epsilon-1}.$$

2. **(II):** Here we use again use item 1 of Lemma 4.5 to obtain
$$||\Phi(t)\Phi(s)^{-1}||\,||W^2(s,Y_s^\circ)|| \lesssim \left(\frac{T-s}{T-t}\right)^{1-\epsilon}\frac{1}{(T-s)}||W(s,Y_s^\circ)(T-s)|| \lesssim (T-t)^{\epsilon-1}(T-s)^{-\epsilon}V_1.$$
Then, we obtain
$$\text{(II)} \leq \int_0^T V_1\,(T-t)^{\epsilon-1}(T-s)^{-\epsilon}\mathrm{d}s \lesssim V_1(T-t)^{\epsilon-1}\int_0^T (T-s)^{-\epsilon}\mathrm{d}s \lesssim V_1(T-t)^{\epsilon-1}$$

3. **(III):** Here, we use item 3 of Lemma 4.5 with $U(s) = \sigma(s)H(s, Y_s^\circ)$ to obtain an almost surely finite random variable V_2 such that

$$(\text{III}) = \left|\left| \int_0^t \Phi(t)\Phi(s)^{-1}U(s)\mathrm{d}B_s \right|\right| \lesssim V_2(T-t)^{\epsilon-1}.$$

Combining the results, we know that

$$||\tilde{r}(t, Y_t^\circ)|| \lesssim V\,(T-t)^{\epsilon-1},$$

such that $V = V_1 + V_2$ is an almost sure finite random variable. □

Lemmas for Theorem 4.1

Lemma 4.3. *Let Assumption 4.1 hold, then*

$$\frac{||Lx - v||}{T-t} \lesssim 1 + ||\tilde{r}(t,x)||,$$

and

$$||\tilde{r}(t,x)|| \lesssim 1 + \frac{1}{T-t}||Lx - v||.$$

Proof. Let us first consider writing $Lx - v$ as follows:

$$Lx - v = L\tilde{\mu}_T(t,x) - v - (L\tilde{\mu}_T(t,x) - Lx).$$

Then, we can use that there exists a matrix-valued function $Q^+(t,x)$ such that

$$Q^+(t,x)\tilde{r}(t,x) = L\tilde{\mu}_T(t,x) - v.$$

Combining these statements, we obtain

$$Lx - v = Q^+(t,x)\tilde{r}(t,x) - L(\tilde{\mu}_T(t,x) - x).$$

Then, we know that by Assumption 4.1 there exists a constant $C > 0$ such that

$$||\tilde{\mu}_T(t,x) - x|| < C(T-t).$$

Therefore,

$$||Lx - v|| \leq ||Q^+(t,x)||\,||\tilde{r}(t,x)|| + C(T-t).$$

Then using that $||Q^+(t,x)|| \lesssim (T-t)$ we have the desired result. Furthermore, we know that

$$\begin{aligned} ||\tilde{r}(t,x)|| \leq ||Q(t,x)||\,||L\tilde{\mu}_T(t,x) - v|| &\leq ||Q(t,x)||\,(||\tilde{\mu}_T(t,x) - x|| + ||Lx - v||) \\ &\lesssim \frac{1}{T-t}\left((T-t) + ||Lx - v||\right). \end{aligned}$$

□

Lemma 4.4 (SDE of $\tilde{r}_t$)**.** *Let Assumption 4.1 hold. Denote $\tilde{r}_t := \tilde{r}(t, Y_t^\circ)$ where $\tilde{r}(t,x) = \nabla \log \tilde{h}(t,x)$. Furthermore let $\tilde{R}_t := \tilde{R}(t, Y_t^\circ)$ where $\tilde{R}(t,x) = \log \tilde{h}(t,x)$, such that $\nabla\tilde{R} = \tilde{r}$. Then the SDE that governs the stochastic process $(\tilde{r}_t)_{t=0}^T$*

$$\mathrm{d}\tilde{r}(t, Y_t^\circ) = \left[-\tilde{r}(t, Y_t^\circ)\cdot\nabla\tilde{b}(t, Y_t^\circ) + \nabla^2\tilde{R}(t, Y_t^\circ)(\tilde{b}(t, Y_t^\circ) - b(t, Y_t^\circ))\right]\mathrm{d}t + \sigma(t)\nabla^2\tilde{R}(t, Y_t^\circ)\mathrm{d}B_t,$$

with $\tilde{r}_0 = \tilde{r}(0, Y_0^\circ)$ for $Y_0^\circ \sim \mathbb{P}_{Y_0}$. Here ∇^2 denotes the Hessian operator, such that $\nabla^2\tilde{R}(t,x)$ is a $d \times d$ matrix. Furthermore $\nabla\tilde{b}(t,x)$ denotes the Jacobian of the vector-valued function $\tilde{b}$.

Proof. We apply Itô's formula to derive the following

$$\mathrm{d}\tilde{r}(t, Y_t^\circ) = [\partial_t \tilde{r}(t, Y_t^\circ) + \mathcal{L}_t^\circ \tilde{r}(t, Y_t^\circ)]\,\mathrm{d}t + \sigma(t)\nabla\tilde{r}(t, Y_t^\circ)\mathrm{d}B_t,$$

where the generator $\mathcal{L}_t^\circ$ acts on the state argument of $\tilde{r}$. Now, we use that $\partial_t \tilde{r}_t = \partial_t \nabla \tilde{R}_t = \nabla \partial_t \tilde{R}_t$, which is justified because the partial derivatives can be exchanged under the assumption that they are continuous everywhere (Schwarz's theorem). Using Equation 4.13, we have that that

$$\partial_t \tilde{R}_t + \tilde{\mathcal{L}}_t \tilde{R}_t = -\frac{1}{2}\sigma^2(t)||\tilde{r}_t||^2.$$

Therefore, the drift term of the SDE can be written as

$$-\underbrace{\nabla\tilde{\mathcal{L}}_t\tilde{R}_t}_{\text{(I)}} - \underbrace{\frac{1}{2}\sigma^2(t)\nabla||\tilde{r}_t||^2}_{\text{(II)}} + \underbrace{\mathcal{L}_t^\circ \tilde{r}_t}_{\text{(III)}}.$$

The terms can be expressed as

$$\begin{aligned}
\text{(I)} &= \nabla(\tilde{b}\cdot\tilde{r}_t) + \frac{1}{2}\sigma^2(t)\Delta\tilde{r}_t,\\
\text{(II)} &= \sigma^2(t)\tilde{r}_t\nabla\tilde{r}_t,\\
\text{(III)} &= b_t\cdot\nabla\tilde{r}_t + \sigma^2\tilde{r}_t\nabla\tilde{r}_t + \frac{1}{2}\sigma^2(t)\Delta\tilde{r}_t.
\end{aligned}$$

Then, combining the terms gives us $-\text{(I)} - \text{(II)} + \text{(III)} = -\nabla(\tilde{b}\cdot\tilde{r}_t) + b\cdot\nabla\tilde{r}_t$. Using the product rule, we obtain the claimed drift coefficient of the SDE that drives $\tilde{r}_t$. □

Lemma 4.5 (Adapted from Lemma 14, Lemma 15 and Lemma 16 in [SMZ17].)**.** *Let Assumptions 4.1 hold and let a K denote a continuous matrix-valued function that satisfies*

$$\mathrm{d}\tilde{r}_t = [K(t, Y_t^\circ)\tilde{r}_t + W(t, Y_t^\circ)]\,\mathrm{d}t + \sigma(t)\nabla\tilde{r}_t\mathrm{d}B_t,$$

i.e., it establishes the form of Equation 4.14. Furthermore, assume $K(t) = K_1(t) + K_{23}(t)$ with $||K_{23}(t)|| \lesssim 1$ and assume that there exists $\epsilon \in (0, 1/2)$ and $C_0, C_1, C_2 > 0$ such that

$$x^\top K_1(t)x \le \left(C_0 + \frac{1-\epsilon}{T-t}\right)||x||^2 \text{ and } ||K(t)|| \le C_1\frac{1}{T-t} + C_2.$$

If we consider the following random linear system,

$$\mathrm{d}\Phi(t) = K(t, Y_t^\circ)\Phi(t)\mathrm{d}t,$$

such that Φ exists with $\Phi(0) = I$, then we can make the following three claims:

1. *(Lemma 14 in [SMZ17]) There exists a constant C such that for all $0 < s < t < T$*

$$||\Phi(t)\Phi(s)^{-1}|| \le C\left(\frac{T-s}{T-t}\right)^{1-\epsilon}.$$

2. *(Lemma 15 in [SMZ17]) then the solution $\tilde{r}_t$ can be represented as*

$$\Phi(t)\tilde{r}(0,u) + \Phi(t)\int_0^t \Phi(s)^{-1}W(s, Y_s^\circ)\mathrm{d}s - \Phi(t)\int_0^t \Phi(s)^{-1}\sigma(s)\nabla\tilde{r}(s, Y_s^\circ)\mathrm{d}B_s,$$

   in the sense that they are indistinguishable on $[0, T]$.

3. *(Lemma 16 in [SMZ17]) Define $M_t = \Phi(t)\int_0^t \Phi(s)^{-1}U(s)\mathrm{d}W_s$. Assume that $(T-t)||U(t)|| \lesssim 1$ for all $t \in [0, T)$. Then there exists a finite real random variable V_2 such that for all $0 \le t < T$*

$$||M_t|| \le V_2(T-t)^{\epsilon-1}.$$

Proof. We direct the reader to the proofs in [SMZ17].

□

4.4.4. Proof of Theorem 4.2

Proof. The outline of the proof is exactly that of [BMS20] with slight variations in the lemmas. To show absolute continuity on $[0,T]$ we make use of the following stopping times, i.e.

$$\tau_m(Y) = T \wedge \inf_{t\in[0,T]}\{||LY_t - v|| \geq m(T-t)^{1/2-\epsilon}\}. \tag{4.15}$$

This stopping time concerns the first time when the process Y moves too far away from satisfying the condition determined, relative to m. Let us adopt the following notation $\tau_m^\circ = \tau_m(Y^\circ)$, $\tau_m = \tau_m(Y)$ and $\tau_m^* = \tau_m(Y^*)$. Furthermore, note that by Theorem 4.1, we have that $\lim_{m\to\infty}\tau_m^\circ = T$. We will use the stopping times to derive a convergence result with dominated convergence.

In particular, consider the event that $T = \tau_m^\circ$, which is equivalent to saying that for all $t < T$, we have $t < \tau_m^\circ$. Then for $t < \tau_m^\circ$ we have that $|LY_t^\circ - v|| \leq m(T-t)^{1/2-\epsilon}$. We want to show that $\psi(t)$ is bounded on this event, such that we can use a dominated convergence argument. In fact, by Lemma 4.6, we have that there exists a finite random variable K_m such that $\psi(t)\mathbf{1}\{t \leq \tau_m^\circ\} \leq \exp(K_m)$. Furthermore, by Proposition 4.2 we know that for $t < T$

$$\mathbb{E}\left[\mathbf{1}\{t \leq \tau_m^\circ\}\frac{\tilde{h}(0,Y_0^\circ)}{h(0,Y_0^\circ)}\psi(t)\right] = \mathbb{E}\left[\mathbf{1}\{t \leq \tau_m^*\}\frac{\tilde{h}(t,Y_t^*)}{h(t,Y_t^*)}\right]. \tag{4.16}$$

The idea now is to take take $m \to \infty$ and $t \to T$ on both sides.

First, we consider the left-hand side. Ny the dominated convergence theorem (DCT), we may write

$$\lim_{m\to\infty}\lim_{t\uparrow T}\mathbb{E}\left[\frac{\tilde{h}(0,Y_0^\circ)}{h(0,Y_0^\circ)}\underbrace{\psi(t)\mathbf{1}\{t \leq \tau_m^\circ\}}_{\leq \exp(K_m)}\right] \overset{\text{DCT}}{=} \lim_{m\to\infty}\mathbb{E}\left[\frac{\tilde{h}(0,Y_0^\circ)}{h(0,Y_0^\circ)}\psi(T)\mathbf{1}\{T \leq \tau_m^\circ\}\right].$$

Furthermore, we have that $\{T \leq \tau_m^\circ\} = \{T = \tau_m^\circ\}$, and that $\mathbf{1}\{T = \tau_m^\circ\} \uparrow 1$ as $m \to \infty$. Therefore, we have that by monotone convergence

$$\lim_{m\to\infty}\mathbb{E}\left[\frac{\tilde{h}(0,Y_0^\circ)}{h(0,Y_0^\circ)}\psi(T)\mathbf{1}\{T \leq \tau_m^\circ\}\right] \overset{\text{MCT}}{=} \mathbb{E}\left[\frac{\tilde{h}(0,Y_0^\circ)}{h(0,Y_0^\circ)}\psi(T)\right].$$

Now, we consider the right-hand side of Equation 4.16. Note that

$$\mathbb{E}\left[\mathbf{1}\{t \leq \tau_m^*\}\frac{\tilde{h}(t,Y_t^*)}{h(t,Y_t^*)}\right] = \mathbb{E}\left[\frac{\tilde{h}(t,Y_t^*)}{h(t,Y_t^*)}\right] - \mathbb{E}\left[\mathbf{1}\{t > \tau_m^*\}\frac{\tilde{h}(t,Y_t^*)}{h(t,Y_t^*)}\right],$$

By Lemma 4.7 the first term on the right hand side tends to 1 as $t \uparrow T$, when choosing $g = 1$. So what remains is to show that the second term tends to 0. To show this, we can use the change of measure between the conditioned measure $\mathbb{P}^*$ and the unconditional measure $\mathbb{P}$ to write:

$$h(0,Y_0^*)\mathbb{E}\left[\mathbf{1}\{t > \tau_m^*\}\frac{\tilde{h}(t,Y_t^*)}{h(t,Y_t^*)}\right] = \mathbb{E}\left[\mathbf{1}\{t \geq \tau_m\}\tilde{h}(t,Y_t)\right].$$

By Lemma 4.8, this term vanishes, and considering that $Y_0^* = Y_0^\circ = Y_0$ is chosen to be deterministic, we have the desired result. Specifically, we have that, we have that

$$\lim_{t\uparrow T}\mathbb{E}\left[\frac{\tilde{h}(0,Y_0^\circ)}{h(0,Y_0^\circ)}\psi(t)\right] = 1.$$

By Scheffés lemma we have that $\psi(t) \to \psi(T)$ in L_1. Hence, for any $s < T$ we have that for any bounded $\mathcal{F}_s$ measurable function g

$$\mathbb{E}\left[g(Y^\circ)\frac{\tilde{h}(0,Y_0^\circ)}{h(0,Y_0^\circ)}\psi(T)\right] = \lim_{t\to T}\mathbb{E}\left[g(Y^\circ)\frac{\tilde{h}(t,Y_t^\circ)}{h(t,Y_t^\circ)}\left(\frac{\tilde{h}(0,Y_0^\circ)}{h(0,Y_0^\circ)}\frac{h(t,Y_t^\circ)}{\tilde{h}(t,Y_t^\circ)}\psi(t)\right)\right] \overset{\text{Eq. 4.16}}{=} \lim_{t\to T}\mathbb{E}\left[g(Y^*)\frac{\tilde{h}(t,Y_t^*)}{h(t,Y_t^*)}\right].$$

Then again by Lemma 4.7, we have that this converges to $\mathbb{E}[g(Y^*)]$, which gives us the desired result. □

Lemmas for Theorem 4.2

Lemma 4.6. *Let Assumption 4.1 and Assumption 4.2 hold, then there exists a random variable K_m, such that*

$$\psi(t)\mathbf{1}\{t \le \tau_m^\circ\} \le \exp(K_m).$$

Proof. We use the definition of the function G and Cauchy-Schwarz to write

$$|G(t,x)| \le ||b(t,x) - \tilde{b}(t,x)||\,||\tilde{r}(t,x)||.$$

Then using Lemma 4.3, we find that

$$||\tilde{r}(t,x)|| \lesssim \left(1 + \frac{1}{T-t}||Lx - v||\right)$$

and therefore,

$$|G(t,x)| \lesssim ||b(t,x) - \tilde{b}(t,x)||\left(1 + \frac{1}{T-t}||Lx - v||\right).$$

Then, we can use Assumption 4.2 to bound the drift difference, i.e., there exists an almost sure finite random variable V_1 such that

$$|G(t,Y_t^\circ)| \lesssim V_1\left(1 + \frac{1}{T-t}||LY_t^\circ - v||\right).$$

Then, on the event $\{t \le \tau_m^\circ\}$, we have that for some $\epsilon \in (0,1/2)$, i.e.,

$$||LY_t^\circ - v|| \lesssim m(T-t)^{1/2-\epsilon}.$$

Therefore,

$$|G(t,Y_t^\circ)| \lesssim V_1\left(1 + \frac{m}{T-t}(T-t)^{1/2-\epsilon}\right).$$

Because V_1 is almost sure finite for $\epsilon \in (0,1/2)$, the right-hand side is integrable on $[0,T]$ and the claim follows. □

Lemma 4.7 (Adapted from Lemma 6.4 in [BMS20])**.** *Let Assumption 4.1, Assumption 4.2, and Assumption 4.3 hold, then and $0 < t_1 < t_2 < \cdots < t_n < t < T$ and g be a bounded continuous function on $\mathbb{R}^{nd}$, then*

$$\lim_{t\uparrow T}\mathbb{E}\left[g(Y_{t_1}^*,\dots Y_{t_n}^*)\frac{\tilde{h}(t,Y_t^*)}{h(t,Y_t^*)}\right] = \mathbb{E}\left[g(Y_{t_1}^*,\dots Y_{t_n}^*)\right]$$

Proof. The proof can be found in [BMS20] □

Lemma 4.8 (Adapted from Lemma 6.5 in [BMS20])**.** *Let Assumption 4.1, Assumption 4.2, and Assumption 4.3 hold, then*

$$\mathbb{E}\left[\mathbf{1}\{t \ge \tau_m\}\tilde{h}(t,Y_t)\right] \to 0$$

Proof. The proof can be found in [BMS20]. □

5
Practical Algorithm

In this chapter, we elaborate on a practical interpretation of the pathwise importance sampling technique. We use amenable auxiliary processes based on linearizations of the unconditional drift and consequently have tractable transition densities. The fidelity of the linearized drift is hypothesized to be an important factor for the efficiency of our sampling approach.

However, in practice, it is difficult to choose a reasonable linear approximation for the entire time span $[0, T]$ such that the auxiliary process remains similar to the process $(Y_t)_{t\geq 0}$. Therefore, we update the auxiliary process at multiple times throughout the simulation. This introduces two issues with the approach so far that need to be addressed. First, the adaptation of the auxiliary processes must be considered when computing the Radon-Nikodym derivative of the paths. This requires a minor variation of Proposition 4.2, which we address in this chapter. Second, in the simulation of $(Y_t^\circ)_{t\geq 0}$, we compute gradients of the $\log \tilde{h}$ at the points in the discretization grid. A naive approach of the adaptive constant drift approximation leads to severe instabilities, therefore for a practical method, the gradients must be propagated through the linearization of the drift. This specific aspect required us to perform the additional theoretical validation of the approach in Section 4.4 due to the incorporation of non-trivial expressions for $\nabla\tilde{\mu}_T(t, x)$.

Moreover, an important question is whether our approach is asymptotically consistent. In particular, the canonical results on importance sampling from the previous chapter and the exact formula for the Radon-Nikodym derivative only partially translate to our approach. This is because we consider a discretized version of the proposal process. Therefore, we show that our approach is still asymptotically consistent by inspecting the expected squared error and studying its behavior for large N and large M under various assumptions.

On the practical side, we describe the formal algorithm for conditional sampling with a pre-trained denoising process using our approach. Additionally, we lay out a flexible framework that enables various method configurations. In particular, the algorithm can be implemented with intermediate resampling, a well-known tactic to increase the particle efficiency of importance sampling algorithms, and independent particle pools, a technique to distribute the allocation of computational effort among independent and dependent instances of the algorithm.

In Section 5.1 we study the auxiliary processes with linear approximations of the drift coefficients, and their adaptive mechanism that enables dynamic updates of the auxiliary processes for more accurate proposal processes. In Section 5.2, we discuss the asymptotic consistency of our approach in terms of a vanishing mean squared error. In Section 5.3 we provide an overview of our practical algorithm and the consistent adaptation of intermediate resampling steps. Section 5.4 elaborates on the zero drift approximation and the constant drift approximation.

5.1. Adaptive Auxiliaries

At this point, we recall the construction of the (non-adaptive) proposal process in Equation 4.6, where we superimpose a guidance term $\nabla \log \tilde{h}$ upon the unconditional drift b. This way, some of the dynamics of the unconditional process are effectively carried over to the proposal process.

5.1.1. Local Drift Approximations

Due to the difficulty of choosing a single auxiliary process $\tilde{Y}$ that resembles Y, we simultaneously consider multiple auxiliary processes that we exchange as time passes from 0 to T. Specifically, we mean that for some integer k, we define the process $\tilde{Y}^{(k)}$ to be driven by the following SDE:

$$\mathrm{d}\tilde{Y}^{(k)} = \tilde{b}^{(k)}(t, Y_t^{(k)})\mathrm{d}t + \sigma(t)\mathrm{d}B_t.$$

By the defining property of the auxiliary processes (Definition 4.1), the transition densities of the auxiliary process are Gaussian. Therefore, we have that for the auxiliary process, the guidance term has the form of Equation 4.10. Specifically for some vector valued function $\tilde{\mu}_T^{(k)} : [0,T] \times \mathbb{R}^d \to \mathbb{R}^d$ that represents the conditional expectation of $\tilde{Y}_T^{(k)}$ given $\tilde{Y}_t^{(k)} = x$ and a real valued function $\tilde{C}_T^{(k)} : [0,T] \to \mathbb{R}$ that represents the variance of $\tilde{Y}_T^{(k)}$ given $\tilde{Y}_t^{(k)}$. Note that the covariance matrix is independent of the state, following the state-independent diffusion coefficients. Then, consider any full rank $m \times d$ matrix L such that $L\tilde{C}_T(t)L^\top$ is invertible. It follows that $L\tilde{Y}_T^{(k)}$ given that $\tilde{Y}_t^{(k)} = x$ is Gaussian distributed. Akin to what we show in Lemma 3.1, we may write

$$\tilde{h}^{(k)}(t,x) \propto \exp\left(-\frac{1}{2}||v - L\tilde{\mu}_T^{(k)}(t,x)||^2_{(L\tilde{C}_T^{(k)}(t)L^\top)^{-1}}\right). \tag{5.1}$$

We update the auxiliary process at each step of the discretization of the Euler-Maruyama approximation. This means that the proposal process is adapted through the dynamic auxiliary guidance term. We can formally describe this by the following SDE:

$$\mathrm{d}Y_t^\circ = \left[b(t, Y_t^\circ) + \nabla \log \tilde{h}_{k_t}(t, Y_t^\circ)\right]\mathrm{d}t + \sigma(t)\mathrm{d}B_t, \tag{5.2}$$

with $k_t = \min\{k \in \{1, \dots M\} : t_k \geq t\}$.

A practical sampling scheme is obtained given a current state Y_t°, we sample the new state Y_s° at time $s = t + M^{-1}$ according to the following (Euler-Maruyama) rule, i.e.

$$\widehat{Y}_s^\circ = \widehat{Y}_t^\circ + (b(t, \widehat{Y}_t^\circ) + \nabla \log \tilde{h}_{k_t}(t, \widehat{Y}_t^\circ))M^{-1} + \sigma(t)\sqrt{M}Z, \tag{5.3}$$

where $Z \sim \mathcal{N}(0,1)$.

Remark 5.1 (Condition Satisfaction of $Y_{T,M}$ and $\widehat{Y}_{T,M}$)**.** *Condition satisfaction of $Y_{T,M}^\circ$ follows from the result of Section 4.4 for arbitrary M. Specifically, given that all of the intermediate auxiliary processes independently satisfy the conditions of Theorem 4.1 and Theorem 4.2, the choice of which auxiliary process is used and when should not matter for the result. In the numerical setting, however, the discretized process $\widehat{Y}_{T,M}^\circ$ will not satisfy the condition because we are limited to the last discretization step before T. This is not as problematic because using noise schedules ensures that the diffusion coefficients are small near T, as discussed in Chapter 2. Therefore, in the context of the conditioning of the denoising process, the discretized sample paths of the proposal processes are likely to satisfy the condition for times that are close to T.*

5.1.2. Adaptive Radon-Nikodym derivative

The use of the adaptive auxiliary process requires us to reconsider the use of the importance weights obtained in Proposition 4.2.

Proposition 5.1 (Adaptive Auxiliary Process)**.** *We consider the proposal of the form Equation 5.2 and denote the measure with $\mathbb{P}^{\circ}_{Y}$. Then, the Radon-Nikodym derivative is*

$$\left.\frac{\mathrm{d}\mathbb{P}_Y}{\mathrm{d}\mathbb{P}^{\circ}_Y}\right|_{\mathcal{F}_T}(Y) = W_{T,M} = \prod_{k=1}^{M} \frac{\tilde{h}_k(t_{k+1}, Y^{\circ}_{t_{k+1}})}{\tilde{h}_k(t_k, Y^{\circ}_{t_k})} \exp\left(\int_{t_k}^{t_{k+1}} G_k(s, Y^{\circ}_s)\mathrm{d}s\right). \tag{5.4}$$

Incidentally, we may use the following notation,

$$\rho_k = \frac{\tilde{h}_k(t_{k+1}, Y^{\circ}_{t_{k+1}})}{\tilde{h}_k(t_k, Y^{\circ}_{t_k})}$$

and

$$\psi_k(t) = \exp\left(\int_0^t G_k(s, Y^{\circ}_s)\mathrm{d}s\right).$$

Furthermore, using the notation above, we may refer to the following quantity $\Psi_k = \psi_k(t_{k+1})/\psi_k(t_k)$. This way we can write the weight of Equation 5.4 as

$$W_{T,M} = \prod_{k=1}^{M} \rho_k \Psi_k \tag{5.5}$$

Proof. By Girsanov's formula, we know that the following expression is the specified Radon-Nikodym derivative

$$\exp\left(-\int_0^t \tilde{r}_{s,k_s}\sigma(s)W_s - \frac{1}{2}\int_0^t \sigma(s)\tilde{r}_{s,k_s}\cdot\tilde{r}_{s,k_s}\mathrm{d}s\right),$$

where $\tilde{r}_{s,k_s}$ denotes $\nabla \log \tilde{h}_{k_s}(s, Y^{\circ}_s)$. Now, we break this exponential up into a product of M exponentials, i.e.

$$\prod_{k=1}^{M} \exp\left(-\int_{t_k}^{t_{k+1}} \tilde{r}_{s,k}\sigma(s)W_s - \frac{1}{2}\int_{t_k}^{t_{k+1}} \sigma(s)\tilde{r}_{s,k}\cdot\tilde{r}_{s,k}\mathrm{d}s\right),$$

which allows us to conveniently write the terms $\tilde{r}_k$ instead of $\tilde{r}_{k_s}$. Then, we can follow a similar derivation as in Proposition 4.2, by treating all M auxiliary processes $\tilde{Y}^{(k)}$ independently. This way, we arrive at the desired result. □

To obtain a practical approximation of the importance weight, we must deal with the integral in Equation 5.4. To do this, we make use of a right Riemann sum approximation. This cannot be a left Riemann sum. Otherwise, the term G_k can be degenerately reduced to zero for certain variations of $\tilde{Y}$, e.g., the constant drift approximation that we describe in Section 5.4.

Definition 5.1 (Practical Importance Weight)**.** *Given a set of pairs of samples of $\{(\widehat{Y}^{\circ}_{t_k}, \widehat{Y}^{\circ}_{t_{k+1}})\}_{k=0}^{M-1}$ obtained as described in Equation 5.3, then we define the practical weight as*

$$\widehat{W}_{T,M} = \prod_{k=1}^{M} \frac{\tilde{h}_k(t_{k+1}, \widehat{Y}^{\circ}_{t_{k+1}})}{\tilde{h}_k(t_k, \widehat{Y}^{\circ}_{t_k})} \exp\left(G_k(t_{k+1}, \widehat{Y}^{\circ}_{t_{k+1}})(t_{k+1} - t_k)\right). \tag{5.6}$$

and

$$\widehat{W}_{t,M} = \prod_{k=1, t_k \leq t}^{M} \frac{\tilde{h}_k(t_{k+1}, \widehat{Y}^{\circ}_{t_{k+1}})}{\tilde{h}_k(t_k, \widehat{Y}^{\circ}_{t_k})} \exp\left(G_k(t_{k+1}, \widehat{Y}^{\circ}_{t_{k+1}})(t_{k+1} - t_k)\right).$$

Furthermore, we may use $\widehat{\rho}_k$ to denote the practical equivalent of ρ_k

5.2. Asymptotic Consistency

Now, we work towards understanding the consistency of our approach. In particular, we discuss the squared error of an approximate conditional probability, denoted by $\widehat{P}^*_{N,M}(A)$, compared to the true conditional probability $\mathbb{P}^*_{Y_T}(A)$, for some measurable set A. We define the following

$$\widehat{P}^*_{N,M}(A) = \frac{\sum_{i=1}^N \mathbf{1}_A(\widehat{Y}^\circ_{T,M})\widehat{W}_{T,M}}{\sum_{i=1}^N \widehat{W}_{T,M}}, \tag{5.7}$$

where $\widehat{W}_{T,M}$ is the practical importance weight and $\widehat{Y}^\circ_{T,M}$ is the terminal value of an Euler-Maruyama approximation of Y° as discussed in the previous section. The idea of this section is to show that under certain assumptions on the chosen auxiliary processes, the associated weight function, and the set A, the squared error can be bounded by a sum of two vanishing functions of M and N, i.e.

$$\mathbb{E}\left[|\widehat{P}^*_{N,M}(A) - \mathbb{P}^*_{Y_T}(A)|^2\right] \lesssim \frac{1}{M} + \frac{1}{N}.$$

In particular, we show how the left-hand side of the above term can be decomposed into a term that vanishes for large N as a result of the particle approximation and a different term that vanishes due to the convergence of an Euler-Maruyama discretization. The latter is non-trivial because the discretization also affects the frequency of updating the auxiliary process. Therefore, the main component of this section is showing that for a fixed N, the particle approximation obtained with the discretized version $\widehat{P}^*_{N,M}(A)$ converges to the continuous-time particle approximation $P^*_{N,M}(A)$ as M grows large.

We make the following assumptions that may appear relatively abstract. This is necessary as it is generally hard to make conclusions about the importance weight without explicitly knowing the auxiliary drift $\tilde{b}$ or making assumptions about the unconditional drift b.

Assumption 5.1 (Properties of importance weights)**.** *Let us consider the importance weights $W_{T,M}$ as defined in Equation 5.5 and the practical importance weights as defined in Equation 5.6. Furthermore, we denote the reciprocal sum of the (practical) importance weights as*

$$\mathcal{Z}_{N,M} = \sum_{i=1}^N W^{(i)}_{T,M} \text{ and } \widehat{\mathcal{Z}}_{N,M} = \sum_{i=1}^N \widehat{W}^{(i)}_{T,M}$$

and for $R \subseteq \{1,\dots,N\}$

$$\mathcal{Z}_{N,M,-R} = \sum_{i=1,i\notin R}^N W^{(i)}_{T,M} \text{ and } \widehat{\mathcal{Z}}_{N,M,-R} = \sum_{i=1,i\notin R}^N \widehat{W}^{(i)}_{T,M}.$$

We assume the following properties:

1. *The (practical) importance weights are almost surely positive, i.e.*

$$W_{T,M} > 0 \text{ and } \widehat{W}_{T,M} > 0.$$

2. *The fourth moment of the (practical) importance weight is bounded, i.e.*

$$\mathbb{E}\left[\widehat{W}^4_{T,M}\right] = \mathcal{O}(1) \text{ and } \mathbb{E}\left[W^4_{T,M}\right] = \mathcal{O}(1).$$

3. *The fourth moment of the reciprocal sum of (practical) importance weights satisfies*

$$\mathbb{E}\left[\widehat{\mathcal{Z}}^{-4}_{N,M,-R}\right] = \mathcal{O}\left(\frac{1}{(N-|R|)^4}\right) \text{ and } \mathbb{E}\left[\mathcal{Z}^{-4}_{N,M,-R}\right] = \mathcal{O}\left(\frac{1}{(N-|R|)^4}\right),$$

for any $R \subseteq \{1,\dots,N\}$ with $|R| < 3$.

4. Consider the following notation:

$$I_k = \int_{t_k}^{t_k+M^{-1}} G_k(t, Y_t^\circ)\mathrm{d}t \text{ and } \hat{I}_k = G_k(t_k + M^{-1}, Y^\circ_{t_k+M^{-1}})M^{-1}.$$

Then, the following bounds hold

$$\mathbb{E}\left[\left(\prod_{k=1}^{M} \rho_k\right)^8\right] = \mathcal{O}(1) \text{ and } \mathbb{E}\left[\left(\prod_{k=1}^{M} \widehat{\rho}_k\right)^8\right] = \mathcal{O}(1)$$

and

$$\mathbb{E}\left[\exp\left(\sum_{k=1}^{M} I_k\right)^8\right] = \mathcal{O}(1) \text{ and } \mathbb{E}\left[\exp\left(\sum_{k=1}^{M} \hat{I}_k\right)^8\right] = \mathcal{O}(1)$$

A second assumption we make about the importance weights is slightly less explicit. Specifically, we make an assumption about the function G_k.

Assumption 5.2. *Let Y° denote the proposal process and $\widehat{Y}^\circ$ its Euler-Maruyama approximation. The function G_k as defined in Proposition 5.1 satisfies the following properties for $M > 1$ and $1 \leq k < M$:*

1. We have that

$$\sup_{t\in[t_k,t_k+M^{-1}]} |G_k(t, Y_t^\circ) - G_k(t, Y^\circ_{t_k+M^{-1}})| \lesssim M^{-1} + \sup_{t\in[t_k,t_k+M^{-1}]} ||Y_t^\circ - Y^\circ_{t_k+M^{-1}}||, \tag{5.8}$$

and there exists a finite random variable V such that

$$||Y_t^\circ - Y^\circ_{t_k+M^{-1}}|| \leq VM^{-1}.$$

2. We have that for all valid M and k

$$|G_k(t_k + M^{-1}, Y^\circ_{t_k+M^{-1}}) - G_k(t_k + M^{-1}, \widehat{Y}^\circ_{t_k+M^{-1}})| \lesssim ||Y^\circ_{t_k+M^{-1}} - \widehat{Y}^\circ_{t_k+M^{-1}}||. \tag{5.9}$$

The last assumption we make concerns the set A. Specifically, we want the sets A to be chosen such that the behavior of the (discretized) proposal process around the boundary is sufficiently regular.

Assumption 5.3. *Let $\widehat{Y}^\circ_{T,M}$ denote an Euler-Maruyama approximate of $Y^\circ_{T,M}$, specifically such that*

$$\left(\mathbb{E}\left[||\widehat{Y}^\circ_{T,M} - Y^\circ_{T,M}||^p\right]\right)^{1/p} \lesssim \frac{1}{\sqrt{M}}.$$

This is, for example, shown under certain mild conditions in [AKK18]. Then, then A satisfies for all $M > 1$

$$\mathbb{E}\left[|\mathbf{1}\{\widehat{Y}^{\circ(i)}_{T,M} \in A\} - \mathbf{1}\{Y^{\circ(i)}_{T,M} \in A\}|\right] \lesssim M^{-2}$$

The purpose of these assumptions is to control the behavior of the proposal process and the discretization at the boundaries of sets on which we evaluate the square error. Specifically, we do this by limiting the sets to have a convenient behavior that limits the interplay between the error induced by the discretization and the set's boundary. This is important, as highly irregular set boundaries may limit the guarantee that the discrepancies between the discretized process and the exact process vanish sufficiently fast for bounding the square error.

Now we are ready to state the main theoretical result of this section, of which the main part of the proof is then addressed in the remainder of this section.

Theorem 5.1 (Asymptotic Consistency $\widehat{P}^*_{N,M}$)**.** *Let the denoising process Y satisfy the standard setting, specifically the conditions on the driving SDE for uniqueness and existence of the solution. Let the following assumptions hold for all $k \in \{1, \dots, M\}$ for all $M > 1$:*

- *the auxiliary process $\tilde{Y}^{(k)}$ satisfies Assumption 4.1*
- *the drift of Y and the drift of $\tilde{Y}$ evaluated on Y° satisfies Assumption 4.2.*
- *the unconditional denoising process Y and the auxiliary process $\tilde{Y}$ satisfy Assumption 4.3,*
- *the (practical) importance weights satisfy Assumption 5.1,*
- *the functions G_k (as defined in Proposition 5.1) satisfy Assumption 5.2,*
- *the set A satisfies Assumption 5.3.*

Then, for $N > 2$

$$\mathbb{E}\left[(\widehat{P}^*_{N,M}(A) - \mathbb{P}^*_{Y_T}(A))^2\right] \lesssim \frac{1}{M} + \frac{1}{N}.$$

Proof. We use the fact that $(x+y)^2 \leq 2x^2 + 2y^2$, to decompose the squared error into

$$\frac{1}{2}(\widehat{P}^*_{N,M}(A) - \mathbb{P}^*_{Y_T}(A))^2 \leq (\widehat{P}^*_{N,M}(A) - P^*_{N,M}(A))^2 + (P^*_{N,M}(A) - \mathbb{P}^*_{Y_T}(A))^2,$$

where we denote

$$P^*_{N,M}(A) = \frac{\sum_{i=1}^N \mathbf{1}_A(Y^\circ_{T,M}) W_{T,M}}{\sum_{i=1}^N W_{T,M}},$$

which does depend on M, but not on the discretization. Specifically, $Y^\circ_{T,M}$ is the exact terminal value of the process with the M different auxiliary guidance terms, and $W_{T,M}$ is the exact importance weight. Then, from Assumption 5.1, it follows that the second moments of the weights are bounded, so we can use Proposition 4.1 to bound the second term. Now, we use Lemma 5.2 to bound the first term, which gives us the desired result. □

The proof of this theorem is outlined above and can be primarily decomposed into two separately studied Lemmas that are given throughout the remainder of this section. The error term described in the above equation can be decomposed into an error induced by the discretization and an error induced by the particle approximation. The former requires some explicit effort, because the specific form does not frequently appear in literature. The latter error is more standard, as is portrayed in Section 4.1, specifically Proposition 4.1.

As a final remark, Theorem 5.1 leads to a simple corollary on vanishing bias.

Corollary 5.1 (Vanishing bias)**.** *Under the setting of Theorem 5.1, as $N \to \infty$ and $M \to \infty$, we have that*

$$|\mathbb{E}\left[\widehat{P}^*_{N,M}(A)\right] - \mathbb{P}^*_{Y_T}(A)| \to 0.$$

Proof. Convergence in L_2 implies convergence in L_1 by Hölder's Inequality and because

$$|\mathbb{E}\left[\widehat{P}^*_{N,M}(A)\right] - \mathbb{P}^*_{Y_T}(A)| \leq \mathbb{E}\left[|\widehat{P}^*_{N,M}(A) - \mathbb{P}^*_{Y_T}(A)|\right],$$

the claim follows. □

Remark 5.2 (Relation to a total variation(-like) metric)**.** *To study the relation of Theorem 5.1 with an actual metric for probability distributions, we may consider that a restricted version of the total variation. Specifically, between the expectation of the random measure $\widehat{P}^*_{N,M}$ and the true measure $\mathbb{P}^*_{Y_T}$. It converges to zero due to the vanishing bias, i.e., let $\mathcal{R}$ denote the restriction to those sets that satisfy Assumption 5.3*

$$\sup_{A\in\mathcal{R}}\left|\mathbb{E}\left[\widehat{P}^*_{N,M}(A)\right]-\mathbb{P}^*_{Y_T}(A)\right|\to 0.$$

Clearly, the above expression is only a lower bound on the actual total variation $\mathrm{TV}(\widehat{P}^*_{N,M},\mathbb{P}^*_{Y_T})$, which makes it less meaningfull. This is for two reasons, first being that taking the supremum over the entire σ-algebra $\mathcal{F}$ gives

$$\sup_{A\in\mathcal{R}}|\mathbb{E}\left[\widehat{P}^*_{N,M}(A)\right]-\mathbb{P}^*_{Y_T}(A)|\leq\sup_{A\in\mathcal{F}}|\mathbb{E}\left[\widehat{P}^*_{N,M}(A)\right]-\mathbb{P}^*_{Y_T}(A)|$$

and second, considering the random counting measure itself instead of its expectation gives

$$\sup_{A\in\mathcal{R}}|\mathbb{E}\left[\widehat{P}^*_{N,M}(A)\right]-\mathbb{P}^*_{Y_T}(A)|\leq\sup_{A\in\mathcal{R}}|\widehat{P}^*_{N,M}(A)-\mathbb{P}^*_{Y_T}(A)|.$$

Therefore, because it is difficult to say something meaningful about the actual total variation, we leave these considerations in this direction for future research.

5.2.1. Lemmas of Theorem 5.1

To show the convergence of $(\widehat{P}^*_{N,M}(A)-P^*_{N,M}(A))^2$ (Lemma 5.2), we can make use of the definitions and a useful rewriting, such that most of the terms in our expression the assumptions directly result into a bound on the expectation. First, we inspect the behavior of $(\widehat{W}_{T,M}-W_{T,M})$.

Lemma 5.1 (Square error of $\widehat{W}_{T,M}$)**.** *Let Assumption 5.1 and Assumption 5.2 hold for $\widehat{W}_{T,M}$ and $W_{T,M}$, then*

$$\mathbb{E}\left[(\widehat{W}_{T,M}-W_{T,M})^2\right]\lesssim\frac{1}{M}.$$

Proof. Writing out the expression gives us the following:

$$(\widehat{W}_{T,M}-W_{T,M})^2=\left(\prod_{k=1}^{M}\widehat{\rho}_k\exp\left(\sum_{k=1}^{M}G_k(t_{k+1},\widehat{Y}^\circ_{t_{k+1}})M^{-1}\right)-\prod_{k=1}^{M}\rho_k\exp\left(\sum_{k=1}^{M}\int_{t_k}^{t_k+M^{-1}}G_k(t,Y^\circ_t)\mathrm{d}t\right)\right)^2.$$

We use the notation of Assumption 5.1 and bound it by

$$\max\left\{\prod_{k=1}^{M}\rho_k^2,\prod_{k=1}^{M}\widehat{\rho}_k^2\right\}\left(\exp\left(\sum_{k=1}^{M}\hat{I}_k\right)-\exp\left(\sum_{k=1}^{M}I_k\right)\right)^2.$$

The premise of the proof is based on applying the mean value theorem to the squared difference of the exponential functions. specifically, we consider the that for $a,b\in\mathbb{R}$ there exists a value $c\in(a,b)$ such that

$$(e^a-e^b)=(a-b)e^c,$$

where is a value $c\in(a,b)$. Then

$$(e^a-e^b)^2=e^{2c}(a-b)^2\leq(a-b)^2\cdot\max_{c\in(a,b)}e^{2c},$$

where the latter equality is due to the monotonicity of the exponential function. we can now to derive the following upper bound:

$$(\widehat{W}_{T,M} - W_{T,M})^2 \leq \max\left\{\prod_{k=1}^{M}\rho_k^2, \prod_{k=1}^{M}\widehat{\rho}_k^2\right\}\exp\left(\max\left\{2\sum_{k=1}^{M}I_k, 2\sum_{k=1}^{M}\hat{I}_k\right\}\right)\left(\sum_{k=1}^{M}(I_k - \hat{I}_k)\right)^2.$$

Then, we consider a Taylor approximation of I_k can be made at $(t_k + M^{-1}, Y^{\circ}_{t_k+M^{-1}})$. Specifically, note that for some integral of a real-valued function, we have that by Taylor's theorem that

$$\begin{aligned}
F(t) &= \int_t^a f(s)\mathrm{d}s = F(a) + F'(a)(t-a) + \frac{1}{2}F''(a)(t-a)^2 + \cdots + F^{(n)}\frac{1}{n!}(t-a)^n + \ldots \\
&= \int_a^a f(s)\mathrm{d}s - f(a)(t-a) + R_1(t) \\
&= -f(a)(t-a) + R_1(t),
\end{aligned}$$

such that $|R_1(t)| \propto h_1(t)(t-a)$ with h_1 vanishing as $t \to a$ faster than $t-a$. Applying Taylor's theorem with random functions is not trivial, because the remainder term is generally not guaranteed to be controllable without making significant assumptions. Here, using our assumption on the functions G_k, we can control the behavior of the random-valued remainder term. To show this, we proceed with some additional care. Consider using the above writing the integral I_k as a function of t (and of $\omega \in \Omega$ to avoid a certain level of ambiguity), to obtain

$$\begin{aligned}
I_k(t,\omega) &\overset{\text{def}}{=} \int_t^{t_k+M^{-1}} G_k(t, Y_t^{\circ}(\omega))\mathrm{d}t \\
&= \int_{t_k+M^{-1}}^{t_k+M^{-1}} G_k(t, Y_t^{\circ}(\omega))\mathrm{d}t \\
&\quad - G_k(t_k + M^{-1}, Y^{\circ}_{t_k+M^{-1}}(\omega)) \cdot (t - (t_k + M^{-1})) + R_1(t,\omega).
\end{aligned}$$

Then again, the first integral on the right-hand side evaluates to zero. Therefore, we have that

$$I_k(\omega) = I_k(t_k,\omega)) = G_k(t_k + M^{-1}, Y^{\circ}_{t_k+M^{-1}}(\omega))M^{-1} - R_1(t_k,\omega).$$

Now, we consider that

$$\begin{aligned}
R_1(t,\omega) &= I_k - G_k(t_k + M^{-1}, Y^{\circ}_{t_k+M^{-1}}(\omega))M^{-1} \\
&\leq M^{-1} \sup_{t\in[t_k, t_k+M^{-1}]} G_k(t, Y_t^{\circ}(\omega)) - G_k(t_k + M^{-1}, Y^{\circ}_{t_k+M^{-1}}(\omega))M^{-1} \\
&\leq M^{-1}\left(\sup_{t\in[t_k, t_k+M^{-1}]} G_k(t, Y_t^{\circ}(\omega)) - G_k(t_k + M^{-1}, Y^{\circ}_{t_k+M^{-1}}(\omega))\right).
\end{aligned}$$

We obtain by Assumption 5.2 that $|R_1(t,\omega)| = \mathcal{O}(M^{-2})$. Now that we have arrived at the above bound, which is no longer ambiguous with respect to the remainder term of the Taylor expansion, we drop the dependency on ω in our notation again.

The Taylor expansion conveniently aligns with using the right Riemann sum in the discretization scheme. Therefore,

$$(I_k - \hat{I}_k)^2 \lesssim (G_k(t_k + M^{-1}, Y^{\circ}_{t_k+M^{-1}})M^{-1} + \mathcal{O}(M^{-2}) - G_k(t_k + M^{-1}, \widehat{Y}^{\circ}_{t_k+M^{-1}})M^{-1})^2.$$

Then, again using our assumptions about G_k, we have that

$$(I_k - \hat{I}_k)^2 \lesssim \left(\mathcal{O}(M^{-1})||\widehat{Y}^{\circ}_{t_k+M^{-1}} - Y^{\circ}_{t_k+M^{-1}}|| + \mathcal{O}(M^{-2})\right)^2.$$

We can use this inequality to bound

$$\left(\sum_{k=1}^{M} I_k - \hat{I}_k\right)^2 \overset{\text{CS}}{\leq} M \sum_{k=1}^{M} (I_k - \hat{I}_k)^2 \lesssim ||\widehat{Y}^{\circ}_{t_k+M^{-1}} - Y^{\circ}_{t_k+M^{-1}}||^2 + \mathcal{O}(M^{-2}),$$

where the inequalities are due to Cauchy-Schwarz inequality, the canceling of M^2 with M^{-2} and $(a+b)^2 \leq 2a^2 + 2b^2$.

Summarizing what we have so far gives us the following bound:

$$(\widehat{W}_{T,M} - W_{T,M})^2 \leq \max\left\{\prod_{k=1}^{M} \rho_k^2, \prod_{k=1}^{M} \widehat{\rho}_k^2\right\} \exp\left(\max\left\{2\sum_{k=1}^{M} I_k, 2\sum_{k=1}^{M} \hat{I}_k\right\}\right) \cdot (||\widehat{Y}^{\circ}_{t_k} - Y^{\circ}_{t_k}||^2 + \mathcal{O}(M^{-2})).$$

Now, we consider taking the expectation and use Cauchy-Schwarz to derive

$$\begin{aligned}
&\left(\mathbb{E}\left[\max\left\{\prod_{k=1}^{M} \rho_k^2, \prod_{k=1}^{M} \widehat{\rho}_k^2\right\} \exp\left(\max\left\{2\sum_{k=1}^{M} I_k, 2\sum_{k=1}^{M} \hat{I}_k\right\}\right) \cdot (||\widehat{Y}^{\circ}_{t_k} - Y^{\circ}_{t_k}||^2 + \mathcal{O}(M^{-2}))\right]\right)^2 \\
&\overset{\text{CS}}{\leq} \mathbb{E}\left[\max\left\{\prod_{k=1}^{M} \rho_k^4, \prod_{k=1}^{M} \widehat{\rho}_k^4\right\} \exp\left(\max\left\{4\sum_{k=1}^{M} I_k, 4\sum_{k=1}^{M} \hat{I}_k\right\}\right)\right] \cdot \mathbb{E}\left[(||\widehat{Y}^{\circ}_{t_k} - Y^{\circ}_{t_k}||^2 + \mathcal{O}(M^{-2}))^2\right].
\end{aligned}$$

Now, by the fourth item of Assumption 5.1 we know that the first term is bounded, indeed by Cauchy-Schwarz

$$\begin{aligned}
&\mathbb{E}\left[\max\left\{\prod_{k=1}^{M} \rho_k^4, \prod_{k=1}^{M} \widehat{\rho}_k^4\right\} \exp\left(\max\left\{4\sum_{k=1}^{M} I_k, 4\sum_{k=1}^{M} \hat{I}_k\right\}\right)\right]^2 \\
&\leq \mathbb{E}\left[\max\left\{\prod_{k=1}^{M} \rho_k^8, \prod_{k=1}^{M} \widehat{\rho}_k^8\right\}\right] \mathbb{E}\left[\exp\left(\max\left\{8\sum_{k=1}^{M} I_k, 8\sum_{k=1}^{M} \hat{I}_k\right\}\right)\right] = \mathcal{O}(1),
\end{aligned}$$

which gives us that

$$\mathbb{E}\left[(\widehat{W}_{T,M} - W_{T,M})^2\right] \lesssim \sqrt{\left(\mathbb{E}\left[||\widehat{Y}^{\circ}_{t_k} - Y^{\circ}_{t_k}||^4\right] + 2\mathcal{O}(M^{-2})\mathbb{E}\left[(||\widehat{Y}^{\circ}_{t_k} - Y^{\circ}_{t_k}||^2\right] + \mathcal{O}(M^{-4})\right)}.$$

Now, we use the convergence rate of the Euler-Maruyama approximation as assumed in Assumption 5.3, i.e.,

$$\left(\mathbb{E}\left[||\widehat{Y}^{\circ}_{t_k} - Y^{\circ}_{t_k}||^p\right]\right)^{1/p} \lesssim \frac{1}{\sqrt{M}}.$$

Therefore, we know that

$$\left(\mathbb{E}\left[||\widehat{Y}^{\circ}_{t_k} - Y^{\circ}_{t_k}||^4\right]\right) = \mathcal{O}(M^{-2}) \text{ and } \left(\mathbb{E}\left[||\widehat{Y}^{\circ}_{t_k} - Y^{\circ}_{t_k}||^2\right]\right) = \mathcal{O}(M^{-1}).$$

Using this in the inequality above, we obtain

$$\mathbb{E}\left[(\widehat{W}_{T,M} - W_{T,M})^2\right] \lesssim \sqrt{\mathcal{O}(M^{-2}) + \mathcal{O}(M^{-3}) + \mathcal{O}(M^{-4})} = \mathcal{O}(M^{-1}).$$

which results from concavity of the mapping $\cdot \to \cdot^{1/2}$, Jensen's inequality and the fact that the M^{-2} dominate M^{-3} and M^{-4} as $M \to \infty$. □

Finishing the proof

Now, we use the above-determined Lemma to demonstrate the behavior of $\widehat{P}^*_{N,M}$ as $M \to \infty$. We do this by showing a bound on the mean squared error between $\widehat{P}^*_{N,M}$ and P^*_N, which is the approximation with exact continuous samples of the proposal process.

Lemma 5.2 (Square error of $\widehat{P}^*_{N,M}(A)$)**.** *Let Assumption 5.1 and Assumption 5.3 hold, then for $N > 2$*

$$\mathbb{E}\left[(\widehat{P}^*_{N,M}(A) - P^*_N(A))^2\right] \lesssim M^{-1}.$$

Proof. The idea of this proof is to split the error with the following inequality using that $(a+b)^2 \leq 2a^2 + 2b^2$

$$\frac{1}{2}(P^*_N(A) - \widehat{P}^*_{N,M}(A))^2 \leq \underbrace{\left(\widehat{P}^*_{N,M}(A) - \frac{1}{\widehat{\mathcal{Z}}_{N,M}}\sum_{i=1}^N \mathbf{1}\{Y_T^{\circ(i)} \in A\} W_{T,M}^{(i)}\right)^2}_{\text{(I)}} + \underbrace{\left(\frac{1}{\widehat{\mathcal{Z}}_{N,M}}\sum_{i=1}^N \mathbf{1}\{Y_T^{\circ(i)} \in A\} W_{T,M}^{(i)} - P^*_N(A)\right)^2}_{\text{(II)}}.$$

Then note that the first term is written as

$$\text{(I)} = \left(\widehat{P}^*_{N,M}(A) - \frac{1}{\widehat{\mathcal{Z}}_{N,M}}\sum_{i=1}^N \mathbf{1}\{Y_T^{\circ(i)} \in A\} W_{T,M}^{(i)}\right)^2 = \frac{1}{\widehat{\mathcal{Z}}^2_{N,M}}\left(\sum_{i=1}^N (\mathbf{1}\{Y_{T,M}^{\circ(i)} \in A\} - \mathbf{1}\{Y_T^{\circ(i)} \in A\}) W_{T,M}^{(i)}\right)^2,$$

and the second term can be written as

$$\text{(II)} = \left(\frac{1}{\widehat{\mathcal{Z}}_{N,M}}\sum_{i=1}^N \mathbf{1}\{Y_T^{\circ(i)} \in A\} W_{T,M}^{(i)} - P^*_N(A)\right)^2 = \left(\sum_{i=1}^N \mathbf{1}\{Y_T^{\circ(i)} \in A\}\left(\frac{1}{\widehat{\mathcal{Z}}_{N,M}}\widehat{W}_{T,M}^{(i)} - \frac{1}{\mathcal{Z}_{N,M}} W_{T,M}^{(i)}\right)\right)^2.$$

First term. Using Cauchy-Schwarz inequality, we have that

$$\begin{aligned}
\text{(I)} &\leq \frac{1}{\widehat{\mathcal{Z}}^2_{N,M}}\left(\sum_{i=1}^N (\mathbf{1}\{Y_{T,M}^{\circ(i)} \in A\} - \mathbf{1}\{Y_T^{\circ(i)} \in A\})^2\right)\left(\sum_{i=1}^N \left(W_{T,M}^{(i)}\right)^2\right) \\
&= \left(\sum_{i=1}^N \frac{1}{\widehat{\mathcal{Z}}_{N,M}}(\mathbf{1}\{Y_{T,M}^{\circ(i)} \in A\} - \mathbf{1}\{Y_T^{\circ(i)} \in A\})^2\right)\left(\sum_{i=1}^N \frac{1}{\widehat{\mathcal{Z}}_{N,M}}\left(W_{T,M}^{(i)}\right)^2\right) \\
&\leq \left(\sum_{i=1}^N \frac{1}{\widehat{\mathcal{Z}}_{N,M,-\{i\}}}(\mathbf{1}\{Y_{T,M}^{\circ(i)} \in A\} - \mathbf{1}\{Y_T^{\circ(i)} \in A\})^2\right)\left(\sum_{i=1}^N \frac{1}{\widehat{\mathcal{Z}}_{N,M,-\{i\}}}\left(W_{T,M}^{(i)}\right)^2\right).
\end{aligned}$$

The latter inequality is due to the fact that $\widehat{\mathcal{Z}}_{N,M,-\{i\}} < \widehat{\mathcal{Z}}_{N,M}$ and $\mathcal{Z}_{N,M,-\{i\}} < \mathcal{Z}_{N,M}$, because of the positivity of the importance weights. Then again, applying Cauchy-Schwarz gives us:

$$
\begin{aligned}
\mathbb{E}\left[(I)\right]^2 &\overset{\text{CS}}{\leq} \mathbb{E}\left[\left(\sum_{i=1}^{N}\frac{1}{\widehat{\mathcal{Z}}_{N,M,-\{i\}}}(\mathbf{1}\{Y_{T,M}^{\circ(i)}\in A\}-\mathbf{1}\{Y_{T}^{\circ(i)}\in A\})^2\right)^2\right]\mathbb{E}\left[\left(\sum_{i=1}^{N}\frac{1}{\widehat{\mathcal{Z}}_{N,M,-\{i\}}}\left(W_{T,M}^{(i)}\right)^2\right)^2\right]\\
&\overset{\text{CS}}{\leq} \mathbb{E}\left[N\sum_{i=1}^{N}\frac{1}{\widehat{\mathcal{Z}}^2_{N,M-\{i\}}}(\mathbf{1}\{Y_{T,M}^{\circ(i)}\in A\}-\mathbf{1}\{Y_{T}^{\circ(i)}\in A\})^4\right]\mathbb{E}\left[N\sum_{i=1}^{N}\frac{1}{\widehat{\mathcal{Z}}^2_{N,-\{i\}}}\left(W_{T,M}^{(i)}\right)^4\right]\\
&\overset{\text{Lin. \& Indep.}}{=} N\sum_{i=1}^{N}\mathbb{E}\left[\frac{1}{\widehat{\mathcal{Z}}^2_{N,M-\{i\}}}\right]\mathbb{E}\left[(\mathbf{1}\{Y_{T,M}^{\circ(i)}\in A\}-\mathbf{1}\{Y_{T}^{\circ(i)}\in A\})^4\right]N\sum_{i=1}^{N}\mathbb{E}\left[\frac{1}{\widehat{\mathcal{Z}}^2_{N,M,-\{i\}}}\right]\mathbb{E}\left[\left(W_{T,M}^{(i)}\right)^4\right]\\
&\overset{\text{Ident.}}{=} N^4\mathbb{E}\left[\frac{1}{\widehat{\mathcal{Z}}^2_{N-1,M}}\right]^2\mathbb{E}\left[(\mathbf{1}\{Y_{T,M}^{\circ}\in A\}-\mathbf{1}\{Y_{T}^{\circ(1)}\in A\})^4\right]\mathbb{E}\left[(W_{T,M})^4\right].
\end{aligned}
$$

A bound can be obtained from using Assumption 5.3, combined with moment bounds on the weights and normalization that are given in Assumption 5.1, respectively. The bound is as follows:

$$
\mathbb{E}\left[(\mathrm{I})'\right]\lesssim N^2\frac{1}{M}\sqrt{\mathbb{E}\left[(W_{T,M})^4\right]}\sqrt{\mathbb{E}\left[\frac{1}{\widehat{\mathcal{Z}}^4_{N-1,M}}\right]}=N^2\frac{1}{M}\mathcal{O}(1)\sqrt{\frac{1}{(N-1)^4}}\lesssim\frac{1}{M}.
$$

Second term. Now for the term (II), we use again a Cauchy-Schwarz inequality

$$
(\mathrm{II})\overset{\text{CS}}{\leq}\left(\sum_{i=1}^{N}\mathbf{1}\{Y_{T}^{\circ(i)}\in A\}\right)\left(\sum_{i=1}^{N}\left(\frac{1}{\widehat{\mathcal{Z}}_{N,M}}\widehat{W}_{T,M}^{(i)}-\frac{1}{\mathcal{Z}_{N,M}}W_{T,M}^{(i)}\right)^2\right)=(\mathrm{II})'.
$$

This can be bounded by

$$
(\mathrm{II})'\overset{\text{CS}}{\leq} N\sum_{i=1}^{N}\left(\frac{1}{\widehat{\mathcal{Z}}_{N,M,-\{i\}}}\widehat{W}_{T,M}^{(i)}-\frac{1}{\mathcal{Z}_{N,M,-\{i\}}}W_{T,M}^{(i)}\right)^2. \tag{5.10}
$$

We now consider the following decomposition for some positive constant C, using $(a+b+c)^2\leq 3a^2+3b^2+3c^2$, i.e.,

$$
\begin{aligned}
&\frac{1}{3}\left(\frac{1}{\widehat{\mathcal{Z}}_{N,M}}\widehat{W}_{T,M}^{(i)}-\frac{1}{\mathcal{Z}_{N,M}}W_{T,M}^{(i)}\right)^2\\
&=\frac{1}{3}\left(\left(\frac{1}{\widehat{\mathcal{Z}}_{N,M}}\widehat{W}_{T,M}^{(i)}-C\cdot\widehat{W}_{T,M}^{(i)}\right)+\left(C\cdot\widehat{W}_{T,M}^{(i)}-C\cdot W_{T,M}^{(i)}\right)+\left(C\cdot W_{T,M}^{(i)}-\frac{1}{\mathcal{Z}_{N,M}}W_{T,M}\right)\right)^2\\
&\leq\left(\frac{1}{\widehat{\mathcal{Z}}_{N,M}}\widehat{W}_{T,M}^{(i)}-C\cdot\widehat{W}_{T,M}^{(i)}\right)^2+C^2\cdot(\widehat{W}_{T,M}^{(i)}-W_{T,M}^{(i)})^2+\left(C\cdot W_{T,M}^{(i)}-\frac{1}{\mathcal{Z}_{N,M}}W_{T,M}^{(i)}\right)^2\\
&\leq(\widehat{W}_{T,M}^{(i)})^2\left(\frac{1}{\widehat{\mathcal{Z}}_{N,M}}-C\right)^2+(W_{T,M}^{(i)})^2\left(\frac{1}{\mathcal{Z}_{N,M}}-C\right)^2+C^2\cdot(\widehat{W}_{T,M}^{(i)}-W_{T,M}^{(i)})^2.
\end{aligned}
$$

Specifically, we choose $C=\frac{1}{\widehat{\mathcal{Z}}_{N,M}}$, such that we have

$$
\left(\frac{1}{\widehat{\mathcal{Z}}_{N,M}}\widehat{W}_{T,M}^{(i)}-\frac{1}{\mathcal{Z}_{N,M}}W_{T,M}^{(i)}\right)^2\leq(W_{T,M}^{(i)})^2\left(\frac{1}{\mathcal{Z}_{N,M}}-\frac{1}{\widehat{\mathcal{Z}}_{N,M}}\right)^2+\frac{1}{\widehat{\mathcal{Z}}^2_{N,M}}\cdot(\widehat{W}_{T,M}^{(i)}-W_{T,M}^{(i)})^2.
$$

This enables us to write

$$\frac{1}{3}\left(\frac{1}{\widehat{\mathcal{Z}}_{N,M}}\widehat{W}_{T,M}^{(i)} - \frac{1}{\mathcal{Z}_{N,M}}W_T^{(i)}\right)^2 \lesssim (W_{T,M}^{(i)})^2\left(\frac{1}{\mathcal{Z}_{N,M,-\{i\}}} - \frac{1}{\widehat{\mathcal{Z}}_{N,M,-\{i\}}}\right)^2 + \frac{1}{\widehat{\mathcal{Z}}_{N,M,-\{i\}}^2}(\widehat{W}_{T,M}^{(i)} - W_{T,M}^{(i)})^2.$$

Then, because of the independence, identical distribution and linearity, we can write the expectation of Equation 5.10 as

$$N^2\mathbb{E}\left[\left(\frac{1}{\widehat{\mathcal{Z}}_{N-1,M}}\widehat{W}_{T,M} - \frac{1}{\mathcal{Z}_{N-1,M}}W_T\right)^2\right] \lesssim N^2\mathbb{E}\left[W_{T,M}^2\right]\mathbb{E}\left[\left(\frac{1}{\mathcal{Z}_{N-1,M}} - \frac{1}{\widehat{\mathcal{Z}}_{N-1,M}}\right)^2\right] \tag{5.11}$$

$$+ N^2\mathbb{E}\left[\frac{1}{\widehat{\mathcal{Z}}_{N-1,M}^2}\right]\mathbb{E}\left[(\widehat{W}_{T,M} - W_{T,M})^2\right]. \tag{5.12}$$

By Lemma 5.1 the last term is $\mathcal{O}(M^{-1]})$ and we know that by our assumptions

$$\mathbb{E}\left[W_{T,M}^2\right] \lesssim 1 \quad \text{and} \quad \mathbb{E}\left[\frac{1}{\widehat{\mathcal{Z}}_{N-1,M}^2}\right] \leq \sqrt{\mathbb{E}\left[\frac{1}{\widehat{\mathcal{Z}}_{N-1,M}^4}\right]} \lesssim ???$$

What remains is to use $a^{-1} - b^{-1} \leq (a-b)\max\{a^{-2}, b^{-2}\}$ to derive

$$\begin{aligned}\left(\frac{1}{\mathcal{Z}_{N-1,M,}} - \frac{1}{\widehat{\mathcal{Z}}_{N-1,M}}\right)^2 &\leq \max\left\{\frac{1}{\mathcal{Z}_{N-1,M}}, \frac{1}{\widehat{\mathcal{Z}}_{N-1,M}}\right\}^4 \left(\mathcal{Z}_{N-1,M} - \widehat{\mathcal{Z}}_{N-1,M}\right)^2 \\ &\overset{\text{CS}}{\leq} \max\left\{\frac{1}{\mathcal{Z}_{N-1,M}}, \frac{1}{\widehat{\mathcal{Z}}_{N-1,M}}\right\}^4 N\sum_{i=1}^{N-1}\left(W_{T,M}^{(i)} - \widehat{W}_{T,M}^{(i)}\right)^2 \\ &\leq N\sum_{i=1}^{N-1}\left(\max\left\{\frac{1}{\mathcal{Z}_{N-1,M,-\{i\}}}, \frac{1}{\widehat{\mathcal{Z}}_{N-1,M,-\{i\}}}\right\}^4 (\widehat{W}_{T,M}^{(i)} - W_{T,M}^{(i)})^2\right).\end{aligned}$$

Then the expectation can be bounded in the following way

$$\begin{aligned}\mathbb{E}\left[\left(\frac{1}{\mathcal{Z}_{N-1}} - \frac{1}{\widehat{\mathcal{Z}}_{N-1}}\right)^2\right] &\leq \mathbb{E}\left[N\sum_{i=1}^{N-1}\left(\max\left\{\frac{1}{\mathcal{Z}_{N-1,M,-\{i\}}}, \frac{1}{\widehat{\mathcal{Z}}_{N-1,M,-\{i\}}}\right\}^4 (\widehat{W}_{T,M}^{(i)} - W_{T,M}^{(i)})^2\right).\right] \\ &\overset{\text{Lin. \& Indep.}}{=} N(N-1)\mathbb{E}\left[\max\left\{\frac{1}{\mathcal{Z}_{N-2,M}}, \frac{1}{\widehat{\mathcal{Z}}_{N-2,M}}\right\}^4\right]\mathbb{E}\left[(\widehat{W}_{T,M} - W_{T,M})^2\right] \\ &\lesssim N(N-1)\frac{1}{(N-2)^4}\frac{1}{M},\end{aligned}$$

where the last inequality is due to our assumptions and Lemma 5.1. Using this bound in Equation 5.11 shows that term (II) is also $\mathcal{O}(M^{-1})$/ Then combining all our derived bounds, we find that both terms (I) and (II) are $\mathcal{O}(M^{-1})$ and independent of N for $N > 2$, so we have our desired result. □

5.3. Algorithmic Details

In Algorithm 1, we give a formal description of the practical algorithm that is contributed in this work. Before we describe it, we introduce two practical aspects that enable a more efficient usage of the importance sampling technique.

5.3.1. Intermediate Resampling

A well-known phenomenon of importance sampling is the degeneracy of importance weights. In the context of stochastic processes, this degeneracy is driven by both the number of time steps and the dimensionality of the state space. The collapse is caused by almost all particles having low weights, while only a few particles do not. A commonly made adjustment is the intermediate resampling technique, which, combined with the importance sampling procedures, is also known as sequential Monte Carlo. The idea of intermediate resampling is that we refresh the particles at intermediate times by resampling proportional to the computed importance weights. Therefore, the hope is that at time T, more high quality particles are available. The intermediate resampling at time t is done by drawing resampled indices, denoted by $I^{(i)}$, from the following multinomial distribution, i.e.,

$$I^{(i)} \sim \mathrm{Cat}\left(\left\{\frac{(\widehat{W}_{t_k,M})^{(i)}}{\sum_{m=1}^{N}(\widehat{W}_{t_k,M})^{(m)}}\right\}_{i=1}^{N}\right) \quad \text{for } i \in \{1,\dots,N\}. \tag{5.13}$$

Using these sampled indices, it can be seen that the multinomial resampling operation is unbiased. Here, $I^{(i)}$ is the index of the resampled particle replacing particle i, drawn from a categorical distribution defined by the normalized weights.

To see this, let $\{I^{(i)}\}_{i=1}^{N}$ be a set of i.i.d. random variables distributed according to Equation 5.13, then

$$\mathbb{E}\left[(Y_t^{\circ})^{(I^{(i)})}|\{Y_t^{(i)}\}_{i=1}\right] = \sum_{i=1}^{N} Y_t^{\circ(i)} \frac{(\widehat{W}_{t,M})^{(i)}}{\sum_{m=1}^{N}(\widehat{W}_{t,M})^{(m)}},$$

where the expectation is taken w.r.t. the sample index set. The identity can be derived by evaluating the expectation of the discrete categorical probability distribution.

5.3.2. Independent Particle Pools

Furthermore, we propose to consider parameterizing the method configuration with a number K, that divides N, and represents the number of independent pools of particles. One independent pool is a set of N/K particles that start at the same initial state, among which the weights are normalized and resampled if necessary. This simple addition provides an intuitive way to flexibly allocate resources towards diversity and quality, even when the dimensionality of the state space is so high that the weights collapse onto a single particle with and without intermediate resampling.

We may now choose different initial positions for the different pools. However, because we only normalize weights within each pool, the consistency of the importance sampling procedure is not broken. Formally, we write this resampling operation as follows. For particle index $i \in \{1,\dots,N\}$ consider $\mathcal{P}_l \subseteq \{1,\dots,N\}$ that denotes the indices corresponding to the l-th pool. Then, for each pool $\mathcal{P}_l$, the resampling operation is performed independently, i.e., for all l

$$I^{(i)} \sim \mathrm{Cat}\left(\left\{\frac{(\widehat{W}_{t,M})^{(j)}}{\sum_{m\in\mathcal{P}_l}(\widehat{W}_{t,M})^{(m)}}\right\}_{j\in\mathcal{P}_l}\right), \quad \text{for } i \in \mathcal{P}_l. \tag{5.14}$$

Here, again $I^{(i)}$ is the index of the resampled particle replacing particle i, drawn from a categorical distribution defined by the normalized weights within its pool.

5.3.3. Algorithm Overview

The outline of the algorithm is as follows. First, in the initial step, an initial state is sampled for each K independent particle pool. Second, in the sample step, R consecutive steps are sampled for all particles. Third, the weights are computed according to Equation 5.6. Fourth, the weights are used to resample the particles. Then, if the number of sampled steps is not at the discretization level M, we go back to step 2. Otherwise, the particles return. If the index is omitted, the procedure is identical for all indices.

In the previous section, we assume that in the algorithm configuration, no intermediate resampling is used, i.e., $R = M$, and there is a single pool of particles, i.e., $K = 1$. However, it should not be difficult to extend the results for the different configurations. First, $K > 1$ amounts to changing N to N/K in the results below. Second, choosing $R < M$ amounts to partitioning the complete time span $[0, T]$, and considering the results locally within the time spans.

Algorithm 1.
***Input**: Discretization level M, Number of particles N, Number of pools K, Number of steps between resample times R.*
***Output**: $\{Y^{(i)}\}_{i=1}^{N} \subset \mathbb{R}^d$*

1. *(**Initial step**) Sample K initial states $\{y_0^{(l)}\}_{l=1}^{K}$ from the initial distribution $\mathbb{P}_{Y_0}$ and set*

$$(\widehat{Y}_0^{\circ})^{(i)} = y_0^{(l)} \text{ for } i \in \mathcal{P}_l,$$

   where $\{\mathcal{P}_l\}_{l=1}^{K}$ partitions $\{1, \dots, N\}$ into K evenly sized sets.

2. *(**Sample step**) Sample $Y_{t_j}^{\circ}, \dots, Y_{t_{j+R}}^{\circ}$ as is described in Equation 5.3*

3. *(**Weight step**) Compute practical incremental weights from time j to $j + R$ according to Equation 5.6, specifically*

$$\widehat{W}_{t_j,M} = \prod_{k=j}^{j+R} \frac{\tilde{h}_k(t_{k+1}, \widehat{Y}_{t_{k+1}}^{\circ})}{\tilde{h}_k(t_k, \widehat{Y}_{t_k}^{\circ})} \exp\left(G_k(t_{k+1}, \widehat{Y}_{t_{k+1}}^{\circ})(t_{k+1} - t_k)\right).$$

4. *(**Resample step**) sample new indices according to Equation 5.14, update $j = j + R$ and set*

$$(\widehat{Y}_{t_j}^{\circ})^{(i)} = (\widehat{Y}_{t_j}^{\circ})^{(I)^{(i)}}.$$

   if $j < M$ go to step 2 otherwise stop and return $\{(\widehat{Y}_{t_M}^{\circ})^{(i)}\}_{i=1}^{N}$.

Remark 5.3 (Typical configurations)**.** *The following configurations of Algorithm 1 are associated for choices of K and R*

- ***Proposal** $K = N$, R free.*
- ***IS** $K = 1$, $R = M$*
- ***SMC** $K = 1$, $R = 1$*

If we choose $K = N$, we use no importance sampling, and in this case, the output is not changed by resampling, so R is free. If we choose $R = M$ and $K = 1$, we have the standard path-wise importance resampling method. If we choose $K = 1$ and $R = 1$, we have a canonical sequential Monte Carlo method that resamples at every step.

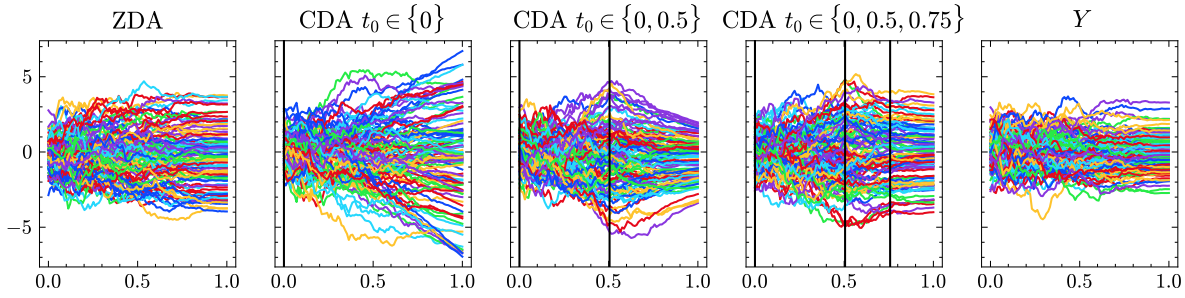


Figure 5.1: Illustration of unconditioned process and auxiliary processes. The leftmost panels display the ZDA auxiliary process. The middle three figures display three instances of the constant drift approximation auxiliaries with different update times. The rightmost figure displays the unconditional process. The vertical lines represent the update times of the constant drift, i.e., the different values for t_0.

5.4. Concrete Auxiliaries

The simplest auxiliary process we choose is a time-inhomogeneous Brownian motion, the zero-drift approximation. Because the adaptive framework does not affect the zero-drift approximation, the algorithm and theory is significantly simplified in this case.

Definition 5.2 (Zero-Drift Approximation (ZDA))**.** *The auxiliary process with zero drift is driven by the following SDE*

$$\mathrm{d}\tilde{Y}_t = \sigma(t)\mathrm{d}B_t,$$

where σ is the same scalar diffusion coefficient as in the unconditional denoising process Y (Equation 2.7).

To compute $\tilde{h}(t,x)$, we observe that the transition distribution of $\tilde{Y}$ is given by

$$\tilde{Y}_T|\tilde{Y}_t \sim \mathcal{N}\left(\tilde{Y}_t, \int_t^T \sigma^2(r)\mathrm{d}r I_{d\times d}\right).$$

Here, we recognize the functions

$$\tilde{C}_T(t) = I_{d\times d}\int_t^T \sigma^2(r)\mathrm{d}r, \text{ and } \tilde{\mu}_T(t,x) = x,$$

as in the context of Lemma 3.1. Making use of the fact that σ is a scalar function, we can easily write out the gradient log term in Equation 4.6 as follows

$$\nabla \log \tilde{h}(t,x) = \left(\int_t^T \sigma^2(r)\mathrm{d}r\right)^{-1} L^\top (LL^\top)^{-1}(v - Lx),$$

with $L \in \mathbb{R}^{m\times d}$, such that $(LL^\top)^{-1}$ exists. The practical incremental importance weights, defined in Equation 5.6, for the ZDA proposal amount to

$$G_k(t,x) = b(t,x)\cdot \nabla \log \tilde{h}(t,x).$$

It is important to note that $\nabla\tilde{\mu}_T(t,x) = I_{d\times d}$.

The constant drift auxiliary process is based on a constant approximation of the drift $b(t,x)$ at (t_0,x_0). Using this constant approximation, we can obtain an auxiliary process $(\tilde{Y}_t)_{t=t_0}^T$.

Definition 5.3 ((G)CDA Auxiliary process)**.** *Let $t_0 \in [0,T)$. Then the auxiliary process based on a constant drift approximation is driven by the following SDE for $t \geq t_0$,*

$$\mathrm{d}\tilde{Y}_t = b(t_0,x_0)\mathrm{d}t + \sigma(t)\mathrm{d}B_t, \qquad \tilde{Y}_{t_0} = x_0,$$

where σ is the same scalar diffusion coefficient as in the unconditional denoising process Y (Equation 2.7).

To compute the $\tilde{h}(t,x)$, as in Equation 5.1 for some $t > t_0$, we observe that the transition distribution of $\tilde{Y}$ is given by

$$\tilde{Y}_T|\tilde{Y}_t = x \sim \mathcal{N}\left(x + b(t_0,x_0)(T-t), \int_t^T \sigma^2(s)\mathrm{d}s\right)$$

Note that this approximation resembles an Euler-Maruyama approximation, without approximating the diffusion coefficient.

$$\tilde{C}_T(t) = I\int_t^T \sigma^2(r)\mathrm{d}r \text{ and } \tilde{\mu}_T(t,x) = x + b(t_0,x_0)(T-t),$$

where I is the $d \times d$ identity matrix.

An important realization when employing the conditioning based on auxiliary SDEs is that the chosen constant drift depends on the values of the auxiliary process at t_0. In the zero drift processes we described earlier, this aspect does not play a role, because $\tilde{\mu}_T(t,x) = x$ and therefore $\nabla\tilde{\mu}(t,x) = 1$. However, severe instabilities may occur if the gradient is not propagated in the case that $\nabla\tilde{\mu}(t,x) \neq 1$.

At this point, it is essential to recall that the linear drift of the auxiliary processes $(\tilde{Y}_t)_{t\geq 0}$, ensures that

$$\log\tilde{h}(t,x) \propto ||L\tilde{\mu}_T(t,x) - v||^2_{(L\tilde{C}_T(t)L^\top)^{-1}}$$

where the function $\tilde{\mu}_T(t,x)$ represents the conditional expectation of $\tilde{Y}_T$ given $\tilde{Y}_t = x$. If we compute a gradient of $\log\tilde{h}$ at (t_0,x_0), it is important to consider that the $\log\tilde{h}$ depends on x_0 (through $\tilde{\mu}_T$). Therefore, the instabilities may be resolved by propagating the gradients at the expansion points (t_0,x_0) through $\tilde{\mu}_T$. To be precise, this means that the gradient of the log of $\tilde{h}$ of CDA
and therefore

$$\nabla\tilde{\mu}_T(t,x) = \begin{cases} 1 + (T-t)\nabla b(t,x) & \text{if } (t,x) = (t_0,x_0), \\ 1 & \text{otherwise.} \end{cases}$$

We refer to this gradient propagation strategy is as GCDA, which is an acronym for Gradient propagated Constant Drift Approximation.

Part III

Numerical Experiments and Application

6 Empirical Analysis

In this chapter, we evaluate the performance of our methods in settings with finite computational effort through an empirical analysis. Our analysis primarily concerns the statistical performance of the different variants of our method, and how it improves for increased computational effort. Following this, we analyze the effect of slightly varied method configurations, i.e., with intermediate resampling and/or pooled particles..

The theory of importance sampling promises a strictly increasing performance as the number of particles grows. However, the steepness of the performance increase may heavily depend on the specific data distributions and method configurations. Therefore, we choose a variety of experiments to investigate the performance. The choice of auxiliary process primarily distinguishes the method configurations we evaluate. In particular, we consider the zero drift approximation (ZDA, Definition 5.2) and the gradient propagated constant drift approximation (GCDA, Definition 5.3). Furthermore, we use a variance-preserving diffusion model with a quadratic noise schedule (Definition 2.4).

The primary goal of this chapter is to measure the statistical performance, i.e., the ability to approximate the conditional distribution, in a variety of ways. The overarching principle is to quantify a discrepancy between the approximate conditional (empirical) distribution that we obtain from samples of our approach and an (pseudo-)exact conditional (empirical) distribution that we obtain from samples of an exact conditioned model. Here, we make use of the exact conditioned model that is available in our simplistic experimental setting. Partially, our findings align with the intuition behind importance sampling. However, they also convey some surprising behavior within GCDA that does not have an increasing performance in low-dimensional state spaces.

A secondary goal is to evaluate the effective sample size, a typical quantification of the efficiency of the proposal distributions in the context of importance sampling, of our approaches for varying the number of dimensions. To be precise, we study the nature of the statistical computational tradeoff that importance sampling induces. The canonical effect of dimensionality on importance sampling results in the well-known weight degeneracy phenomenon, which refers to the collapse of a diverse set of particles into a single unique one. Here, we find that this phenomenon is practically insurmountable, due to the exponential decay of the weights being driven by both the dimensionality of the problem as well as the number of time steps used to sample the denoising process. To this end, we experiment with the use of intermediate resampling and the allocation of computational resources among different independent pools to demonstrate how the efficiency of our approach can be partially preserved.

Name	symbol
Approximate KL Divergence	$\widehat{D}_{\mathrm{KL}}$
Mean Square Error (of the conditional expectation)	MSE
Kolmogorov Smirnov Statistic	D_{KS}
Wasserstein 2 Distance	W_2
Sliced Wasserstein Distance	SWD

Table 6.1: Table containing statistical performance metrics. The Kolmogorov-Smirnov Statistic and the Wasserstein 2 distance are computed with the Python package `Scikit-learn`. The sliced Wasserstein distance is computed with the Python package `pythonot`. The approximate KL divergence and the mean square error can be computed directly with sample means and variances.

In Section 6.1, we study the performance of the core method configurations. Specifically, we measure the statistical performance of the method for different configurations under a highly simplified, low-dimensional setting. Here, we primarily focus on varying the number of particles used in the approach, the dimensionality of the problems, and the number of time discretization steps. In Section 6.2, we empirically study the interplay between the dimensionality and the statistical performance. In Section 6.3, we inspect the behaviour of the different proposal processes. Specifically, we study the effective sample sizes and associated weight degeneracy. Furthermore, we observe how intermediate resampling, which is commonly used as a remedy for the weight degeneracy problem, only offers a partial mitigation in our setting. Finally, in Section 6.4, we study the particle allocation problem to independent particle pools.

For notational clarity, we now drop the dependency on time from the random variables as we are only interested in the value of Y at time T, i.e., if we write $Y := Y_T$. We compute the performance based on two sample sets, one set of samples of a ground-truth conditional distribution $\mathcal{Y} = \{Y^{(i)}\}_{i=1}^N$, that is approximately distributed by $\mathbb{P}^*_{X_0}$, and one set obtained with our approach $\widehat{\mathcal{Y}} = \{\widehat{Y}^{(i)}\}_{i=1}^N$ as obtained by Alg. 1, where the distribution above implicitly depends also on the number of particle pools K, the intermediate resampling time step R, and the choice of the proposal process (ZDA or GCDA).

6.1. Empirical Convergence Rates

We first study the empirical convergence rates in small problems, such that the low-dimensional intuition of importance sampling is demonstrated. Specifically, we compute the statistical performance for various configurations according to some of the performance metrics described in Table 6.1.

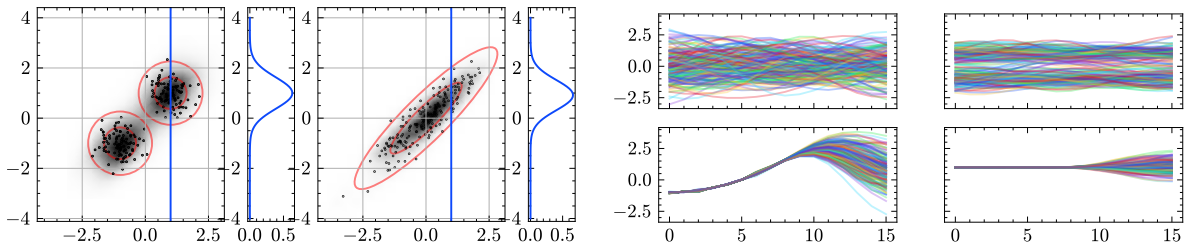
Specifically, we use a Kolmogorov-Smirnov test statistic to determine whether the empirical cumulative distribution function (CDF) of a sample obtained with the generative diffusion significantly resembles the true CDF. The quantity D_{KS} represents the largest absolute difference between two CDFs and thus quantifies the worst-case deviation. Therefore, the rate of improvement is expected to be relatively slow as we increase the number of particles.

Furthermore, given the simplicity of our examples, we can also use an approximate KL divergence. Assuming that the underlying distributions are approximately Gaussian, the approximate KL divergence provides a useful scalar performance metric for quantifying the discrepancy. Let m and a denote the sample average and variance of $\{Y^{(i)}\}_{i=1}^N$ and the sample variance and let $\hat{m}$ and $\hat{a}$ denote the sample average and variance of $\{\widehat{Y}^{(i)}\}_{i=1}^N$ then the approximate KL-divergence is denoted by

$$\widehat{D}_{\mathrm{KL}}(\mathcal{Y}, \widehat{\mathcal{Y}}) \overset{\mathrm{def}}{=} D_{\mathrm{KL}}(\mathcal{N}(m, a) \parallel \mathcal{N}(\hat{m}, \hat{a})) \tag{6.1}$$

This is a very simplified estimator and is strictly limited to the one-dimensional case, where such an approximation is plausible.

A different way to measure the statistical performance is the Wasserstein-2 distance W_2. This is an actual metric on the space of distributions that is sensitive to the overall shape of the distribution, rather than just the centers or most extreme absolute difference between CDFs, and provides valuable information about the discrepancy between two distributions.



(a) Example 6.2 and Example 6.1 **(b)** Example 6.3 and Example 6.4

Figure 6.1: Illustration of experimental settings. (**Left**) The scatter plots indicate samples of the two-dimensional distribution described in Example 6.2 and Example 6.1. The grey shaded area indicates the density of the distributions, and the red ellipses display the covariance of the individual Gaussians. The vertical blue line indicates the conditioning of the horizontal dimension. The vertical marginal of the conditional distributions are displayed in the plots besides the respective scatter plots. (**Right**) The panels show samples of 16 dimensional multivariate samples from Example 6.3 in the left columns and Example 6.4 in the right column. The top row shows samples of the unconditional distributions, and the bottom row shows samples from the ground truth conditional distributions.

6.1.1. Effect of Number of Particles

The following two examples describe the data models and conditions we use in the two-dimensional examples. In Figure 6.1a, we see samples of the example data models.

Example 6.1 (Bivariate)**.** *Consider the following bivariate Gaussian model*

$$\mathbb{P}_{X_0} = \mathcal{N}\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1 & 0.9\\0.9 & 1\end{pmatrix}\right). \tag{6.2}$$

The condition we consider is based on a matrix $L = \begin{pmatrix}1 & 0\end{pmatrix}$ *and a vector* $v = 1$*. Note that a ground truth model can be (informally) described by having the following mean and covariance matrix*

$$\begin{pmatrix}1\\0.9\end{pmatrix} \text{ and } \begin{pmatrix}0 & 0\\0 & 0.19\end{pmatrix}, \tag{6.3}$$

which are computed as described in Section 2.3 of [Bis06]. The zero variance indicates that the value is deterministic. This is therefore not a true covariance matrix that defines a Gaussian distribution. Therefore, in practice, we may add $I\epsilon$ *to the covariance matrix for a negligible* $\epsilon > 0$ *(*$\epsilon = 10^{-6}$*), such that we can easily sample from the Gaussian distribution. We adopt this practical aspect for the remainder of the conditioned distributions.*

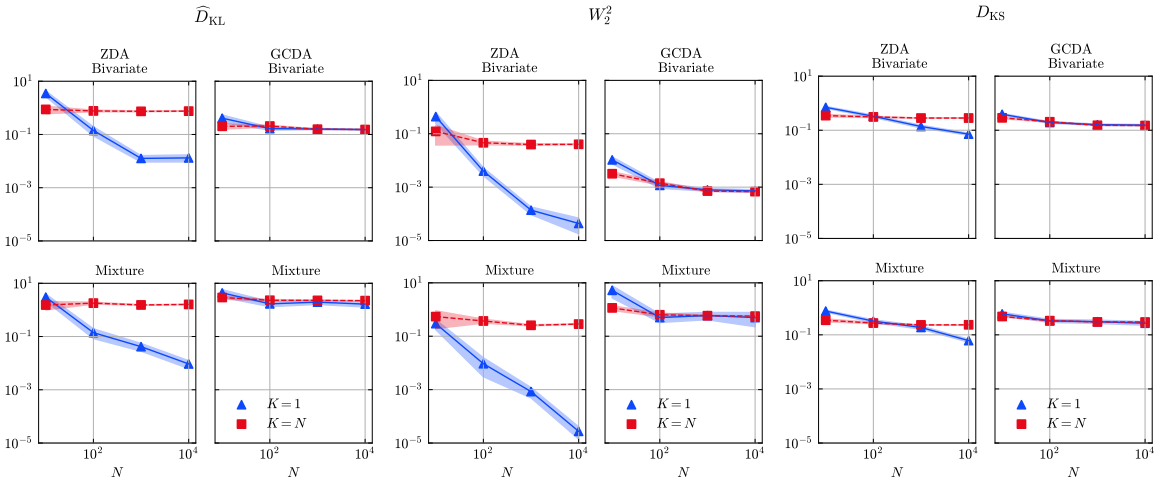
Example 6.2 (Bivariate Bimodal Mixture)**.** *Consider the following bivariate bimodal model*

$$\mathbb{P}_{X_0} = \frac{1}{2}\mathcal{N}(-1, \Sigma) + \frac{1}{2}\mathcal{N}(1, \Sigma), \text{ where } \Sigma = \begin{pmatrix}0.2 & 0\\0 & 0.2\end{pmatrix} \tag{6.4}$$

The condition in this example is identical to that of Example 6.1. The ground-truth conditional distribution is approximated by $\mathcal{N}(1, \Sigma)$*, which is justifiable as the contribution of the Gaussian centered at* -1 *is negligible to the conditional data distribution due to the variances being chosen small enough.*

In Figure 6.2, we depict statistical performance measures of different method configurations for the contexts of Example 6.1 and Example 6.2. The red lines indicate the configuration with no importance sampling because the number of particle pools K equals the number of particles, i.e., every particle is independent. The blue lines indicate importance sampling with a single particle pool, i.e., $K = 1$. Comparing the method's behavior with the proposals ZDA and GCDA gives us two observations.

For the ZDA approach in both experimental settings, all performance metrics appear to decrease as N grows, albeit at different rates. This aligns with the theory of importance sampling, as discussed in Section 4.1. In particular, W_2^2 appears to follow roughly $\mathcal{O}(1/N)$, for both the mixture and bivariate. The quantities $\widehat{D}_{\mathrm{KL}}$ and D_{KS} also decrease at a relatively constant rate on a logarithmic scale. In these experiments, all performance metrics possibly converge to a value that is dominated by the discretization error induced by the fixed choice for M.



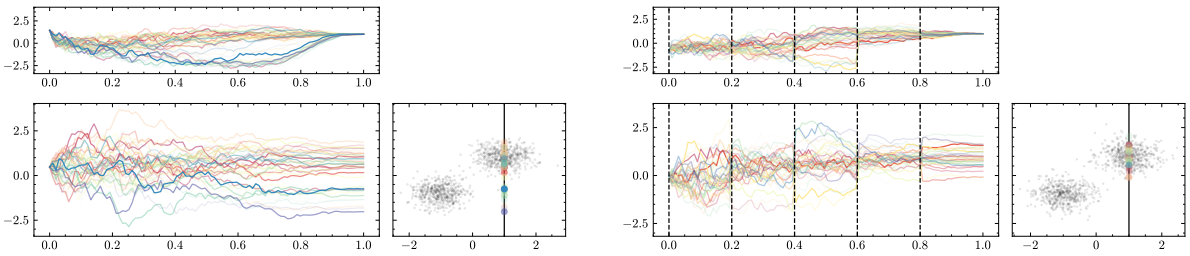
(a) Approximate KL divergence **(b)** Squared Wasserstein-2 distance **(c)** KS statistic D_{KS}

Figure 6.2: Statistical performance with bivariate data models. The horizontal axis shows the number of particles, and the vertical axis shows the statistical discrepancy (lower is better). The red line indicates the performance of N independent particles ($K = N$) and the blue line indicates the performance of a single particle pool ($K = 1$) where we use importance sampling. The shaded areas indicate the standard error of the mean, which is computed from 10 independent runs. Both the horizontal and vertical axes are on a logarithmic scale. The three figures show four panels for different combinations of the two bivariate models (Example 6.1 and Example 6.2) and the two proposal processes (ZDA; Definition 5.2, GCDA; Definition 5.3). The leftmost figure shows the approximate KL Divergence, the middle figure shows the squared Wasserstein-2 distance, and the rightmost figure shows the Kolmogorov-Smirnov statistic.

Surprisingly, GCDA does not show such a significant increase in statistical performance. In fact, for all performance metrics, the statistical performance remains constant for increased computational effort. A possible explanation is that the proposals obtained with GCDA are practically always far from a good candidate sample for the true conditional distribution. Therefore, even for larger numbers of particles (e.g. $N = 10^4$), the importance sampling procedure provides no improvement.

In Figure 6.3a we find the weighted paths of the GCDA proposal processes in a 2-dimensional bimodal example (Example 6.2). In the top panel, we see the paths along the first dimension, i.e., the dimension that we condition to end up at 1, and in the lower panel, we see the paths along the second dimension. In the right panel, we see a scatter plot of the unconditional distribution in black and the resulting (weighted) samples of the GCDA proposal. Because of the bimodal nature of the data distribution, the unconditional paths have a bifurcation. In the top panel, we see that the paths that would otherwise lead to samples close to the mode of $(-1, -1)$ are drawn back to the other mode, i.e. see the time span from 0.4, where the bifurcation starts, and 0.9 where the branches are merged again. These are unrealistic paths for the unconditional process, and one would expect that these paths are weighted relatively low. However, it turns out that this does not happen. A possible explanation is that along the entire path, which is used to compute the importance weight, the undesirable shape at the end of the path can be negligible.

Now that we have a better understanding of the behavior of GCDA, we look into the effect of minor variations of our approach to see whether these can improve the method. As a first attempt to enhance the approach for combating the failure, we use intermediate resampling as described in Section 5.3. Intermediate resampling would resample particles at multiple points in the time span $[0, T)$ instead of just at time T. This appears promising, as it supposedly could remove low-weight particles before they become too problematic, and replace them with high-weight particles. This way, phenomena like the bifurcation may not happen in the first place. In Figure 6.3b, we see this variation in action. The vertical dashed lines indicate the intermediate resampling times. Here, we find that the problematic paths are eliminated and the conditional sample at time T matches the true conditional distribution better than the naive GCDA proposals.



(a) GCDA **(b)** GCDA with intermediate resampling ($R = 20$)

Figure 6.3: In-depth display of failure and fixes of GCDA in bivariate mixture data. The two figures display sample paths of our GCDA method, without intermediate resampling (left) and with intermediate resampling (right). The dashed vertical lines in the right figure indicate the intermediate resampling times. The setting is Example 6.2.

6.2. Dimensional Scalability

Now, we address the effect of the dimensionality of the problem. To achieve this, we must consider a different set of statistical performance metrics. First, the Kolmogorov-Smirnov statistic is difficult to compute for multivariate distributions. Second, the approximate KL-divergence becomes unreliable as estimating the sample covariance is increasingly more difficult in higher dimensions. Finally, the Wasserstein-distance, despite being well-defined in multivariate settings, becomes challenging due to the increased computational complexity.

For this reason, we consider different performance metrics, which are also described in Table 6.1. Specifically, in the multivariate case, a useful simplification of the approximated KL divergence can be made by only comparing the mean squared error (MSE) of the empirical conditional expectations, that is, $\mathrm{MSE}(\mathcal{Y}, \hat{\mathcal{Y}}) = \frac{1}{d}||m - \hat{m}||_2^2$, where d is the dimensionality. This approach is effective because the average statistic (i.e., the mean) can still be estimated with relatively high accuracy even if the full covariance structure is difficult to capture. As a replacement for the Wasserstein distance, we rely on the sliced Wasserstein distance, which is computed by projecting the high-dimensional samples onto random one-dimensional subspaces, computing the one-dimensional Wasserstein distance for each projection, and averaging the results [Bon+15].

6.2.1. Interplay between number of particles and dimensionality

To study the effect of dimensionality, we use the following two examples. Samples from these (conditional) distributions are displayed in Figure 6.1b.

Example 6.3 (Multivariate)**.** *The multivariate distributions we consider are defined by*

$$\mathbb{P}_{X_0} = \mathcal{N}(0, \Sigma(l)) \text{ where } \Sigma_{ij}(l) = \exp\left(-\frac{||i-j||^2}{d^2 l}\right). \tag{6.5}$$

Specifically, we choose $l = 0.1$. Furthermore the condition we consider is obtained with $L \in \{0,1\}^{\lfloor d/2 \rfloor, d}$, such that $L_{ij} = 1$ if $i = j$ and 0 otherwise, for $i \in \{1, \dots, \lfloor d/2 \rfloor\}$ and $j \in \{1, \dots, d\}$. The vector v is defined by $v_i = 2 \cdot (i \cdot \lfloor d/2 \rfloor)^2 - 1$, for $i \in \{1, \dots, \lfloor d/2 \rfloor\}$. This corresponds to a quadratic slope from -1 to 1 for the first half of the dimensions. The ground-truth conditional distribution is obtained via [Bis06]

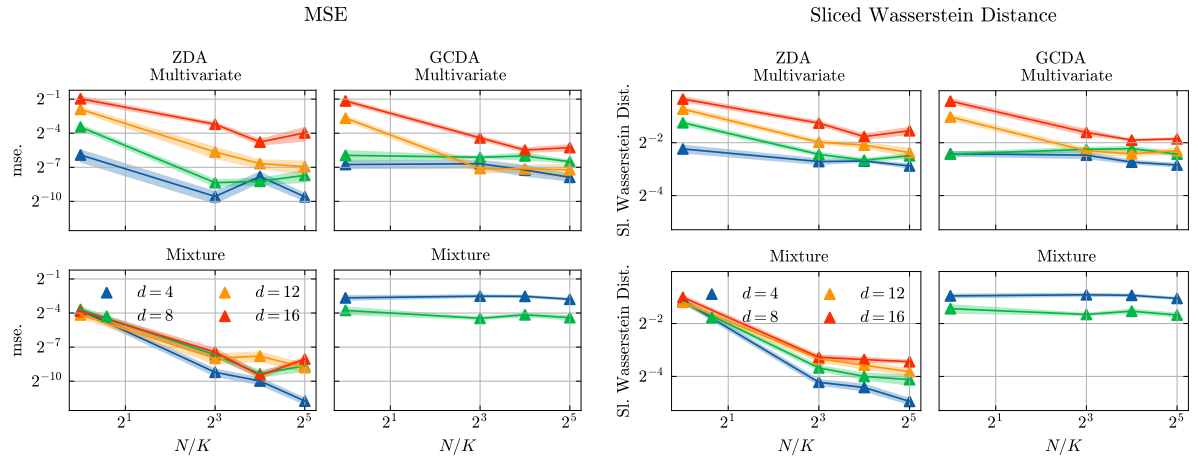
$$\mathcal{N}(\mu + \Sigma L^\top (L\Sigma L^\top)^1 (v - L\Sigma), \Sigma - \Sigma L^\top (L\Sigma L^\top)^{-1} L\Sigma + \epsilon I),$$

where we write $\Sigma := \Sigma(l)$ for brevity and $\epsilon > 0$ is negligible ($\epsilon = 10^{-6}$).

Example 6.4 (Bimodal Mixture)**.** *The bimodal distribution we consider is defined as follows*

$$\mathbb{P}_{X_0} = \frac{1}{2}\mathcal{N}(1, \Sigma(l)) + \frac{1}{2}\mathcal{N}(-1, \Sigma(l)),$$

where $\Sigma(l)$ is defined as in Equation 6.5. Again, we pick $l = 0.1$. The matrix L in the condition we study in this setting is identical to that of Example 6.3, and the vector v is a vector of ones. The true conditional distribution here is again assumed to be $\mathcal{N}(1, \Sigma(l))$.



(a) Mean squared error of conditional expectation

(b) Sliced Wasserstein Distance.

Figure 6.4: Statistical performance in higher-dimensional data models. The horizontal axis shows the number of particles, and the vertical axis shows the statistical discrepancy (lower is better). The number of independent particle pools is fixed to $K = 64$ for all experiments. The colored lines indicate the settings with a varying number of dimensions. The shaded areas indicate the standard error of the mean, which is computed from 5 independent runs. Both the horizontal and vertical axes are on a logarithmic scale. The two figures show four panels for different combinations of the two multivariate models (Multivariate; Example 6.3, Multivariate Mixture; Example 6.4) and the two proposal processes (ZDA; Definition 5.2, GCDA; Definition 5.3). The left figure shows the mean squared error of the conditional expectation and the right figure shows the sliced Wasserstein distance. Our approach to computing the analytical score in the context of Gaussian mixture data suffers from numerical issues that appear in high dimensions. Therefore, we have omitted these results for $d > 8$.

In Figure 6.4 the statistical performance of our approach is displayed for Example 6.3 and Example 6.4. We adopt a slightly different interpretation of the horizontal axis. In particular, we choose to use a value $K = 64$ and vary N, such that the ratio N/K spans values from 1 to 32. This means that $N/K = 1$ is identical to the setting where no importance sampling is used, and $N/K = 32$ suggests that we use 32 particles in each of the $K = 64$ particle pools to obtain our final set of samples. The different colors indicate different dimensionalities of the state space, spanning values from $d = 4$ to $d = 16$. It is expected that the dimensionality makes the approach more difficult, as is recognized by a vertical translation of the performance slope, associated with the dimensionality of the problem. Again, we consider the cases of ZDA and GCDA separately.

First, when considering ZDA as a proposal, we find that the dimensionality greatly affects the unimodal setting, but not so much on the multimodal one. Most of the slopes tend to overlap for both performance quantities. This can be partially explained by the somewhat easier conditioning problem of Example 6.4, as opposed to Example 6.3. In the latter, the curvature of the data, as displayed in the bottom left panel of Figure 6.1b, is an important characteristic of the conditional distribution. On the contrary, in the bottom right panel of the figure, the curvature plays less of a role. In this sense, it can be hypothesized that finding conditional distributions that are primarily characterized by their mean (as in Example 6.2) is easy compared to ones that are primarily characterized by their covariance structure (as in Example 6.3), in which the dimensionality plays a multiplicative role with respect to the difficulty.

Second, when considering the GCDA proposals, we find that for higher values of d, in contrast to small dimensional problems, there is a significantly increased performance. Specifically, we see that for $d = 12$ and $d = 16$, the statistical performance in the setting of Example 6.3 has a similar slope as the ones of the ZDA proposal. This effect can be explained by the benefits of having a theoretically more accurate proposal, starting to outweigh the unresolved failure modes, even to the extent that using GCDA appears to outperform ZDA, which is a reassuring observation that aligns better with the initial hypothesis that GCDA is a better proposal than ZDA as it inherently establishes more fidelity to the unconditioned denoising process by updating the drift.

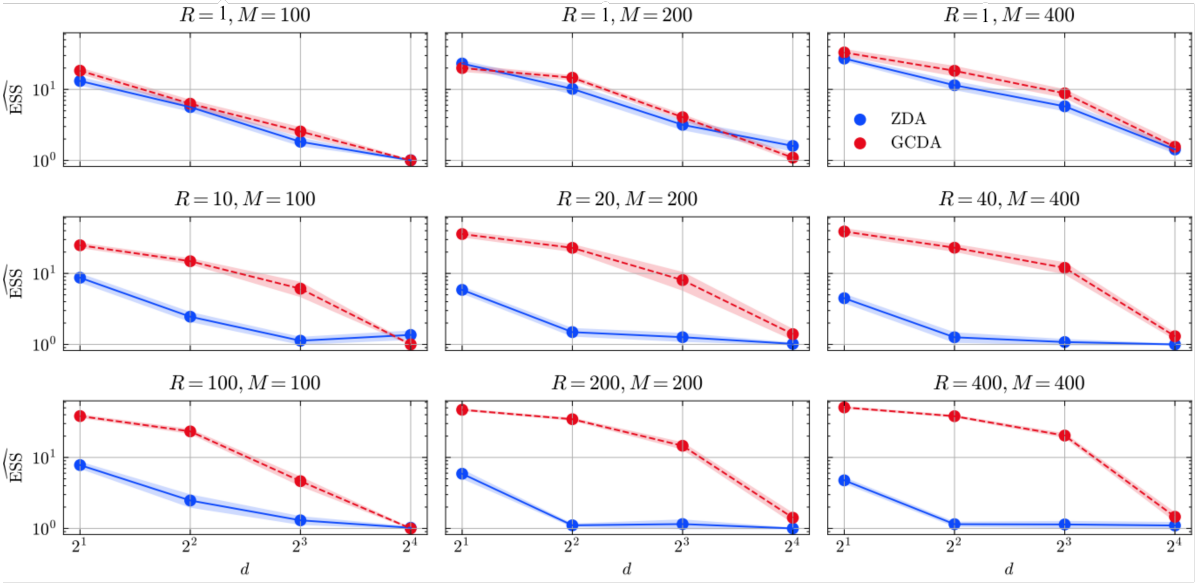


Figure 6.5: Effective sample size for varying dimensionality and method configurations. The horizontal axis depicts the dimensionality of the data, and the vertical axes show the effective sample size, denoted by $\widehat{\text{ESS}}$. The experiments are performed in the setting of the multivariate data model (Example 6.3). Each of the nine panels show the results of experiments that are ran with different values for the number of discretization time steps M and the steps between intermediate resampling R as described in Section 5.3, where $R = 1$ corresponds to resampling every step and $R = M$ corresponds to resampling only at the end. The blue line indicates the effective sample size of ZDA (Definition 5.2) and the red line that of GCDA (Definition 5.3). The vertical axis is in logarithmic scale with base 10, and the horizontal axis is in logarithmic scale with base 2.

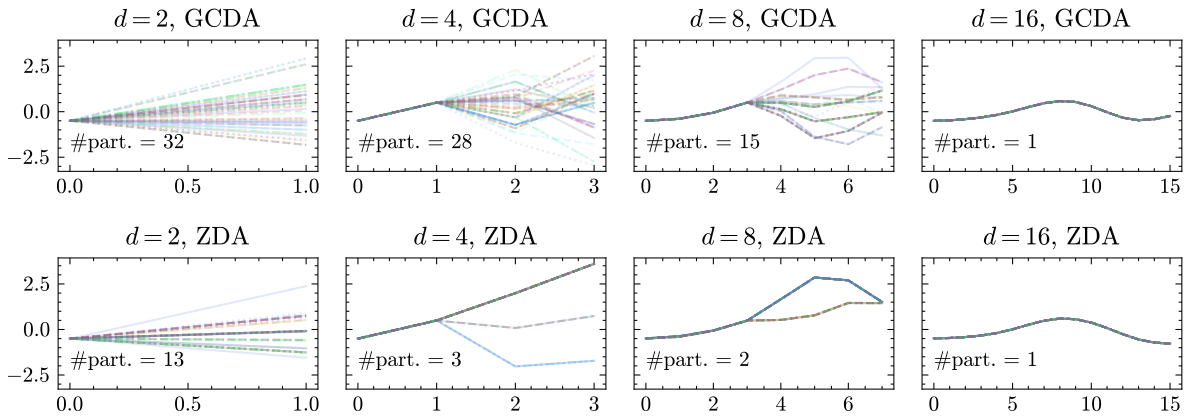
6.3. Particle Efficiency

Now, we study the effectiveness of the two proposal processes in terms of their effective sample size (ESS). The ESS is related to the second moment of the (practical) importance weight, here abstractly referred to with W. In particular, ESS and its approximation are defined as

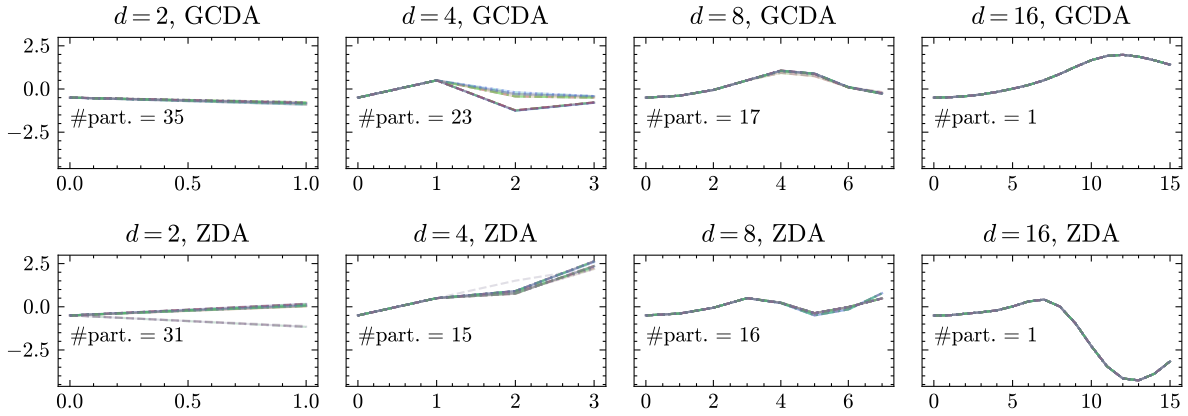
$$\text{ESS} = N\frac{\mathbb{E}\left[W\right]^2}{\mathbb{E}\left[W^2\right]} \quad \text{and} \quad \widehat{\text{ESS}} = N\frac{\left(N^{-1}\sum_{i=1}^{N}\widehat{W}^{(i)}\right)^2}{N^{-1}\sum_{i=1}^{N}(\widehat{W}^{(i)})^2} = \frac{\left(\sum_{i=1}^{N}\widehat{W}^{(i)}\right)^2}{\sum_{i=1}^{N}(\widehat{W}^{(i)})^2} \tag{6.6}$$

It takes values from 1 to N, where an $\widehat{\text{ESS}}$ close to 1 suggests that a single weight is close to 1 while the remaining weights are close to zero, and an $\widehat{\text{ESS}}$ close to N means that all importance weights are roughly the same. Generally, for larger dimensional problems, the importance sampling procedures are likely to exhibit effective sample sizes that are close to 1, a concept known as weight degeneracy. In our case, the measurement of the diffusion processes at M intermediate steps brings us into this regime almost immediately, even when the dimensionality of the state space is only moderately high.

In Figure 6.5, we display $\widehat{\text{ESS}}$ for different method configurations and a varying dimensionality from 2 to 16 in the context of Example 6.3. We first focus on the bottom row, which displays the $\widehat{\text{ESS}}$ of samples obtained with the standard method configuration we have considered so far, without intermediate resampling ($R = M$). Interestingly, we see a large discrepancy between the GCDA approach and the ZDA approach, where ZDA has a rapidly vanishing $\widehat{\text{ESS}}$, especially for a larger number of discretization steps ($M = 400$). In the context of importance sampling, such a display of the $\widehat{\text{ESS}}$ typically suggests that the proposal with a larger $\widehat{\text{ESS}}$ is more efficient. In our case, this is only partially true. Although GCDA seems significantly more effective here in this multivariate example (Example 6.3), it still has a vanishing $\widehat{\text{ESS}}$ when dimensionality of the data is moderately high ($d = 16$).



(a) Without intermediate resampling ($M = 100, R = 100$)



(b) With intermediate resampling ($M = 400, R = 1$)

Figure 6.6: Comparison of collapsing particle pools with and without resampling. Both figures display samples of Example 6.3 in different dimensionalities for different method configurations. The top figure displays the samples obtained with $M = R = 100$ and the bottom figure with $M = 400$ and $R = 1$. These configurations correspond to the bottom left and the top right panel of Figure 6.5, respectively. The number labeled by #part. depicts the number of unique samples out of the 64 generated samples.

A small effective sample size typically results in the collapse of an entire pool of particles into only a few unique particles. More unique particles are favorable in our context as they provide more samples for the same computational power. Unfortunately, we see in Figure 6.5 that the $\widehat{\mathrm{ESS}}$ of both the ZDA and GCDA approaches vanish for high dimensions, albeit at different rates.

Now, we study this effect by displaying unique samples obtained from our method with a single pool of particles ($N = 64, K = 1$) for increasing dimensions. In Figure 6.6a we display individual samples obtained with both GDDA and ZDA approaches for different dimensionalities in the context of Example 6.3. We see the vanishing effective particle size reflected in the vanishing number of unique particles. The ZDA approach appears less efficient, as it only produces 3 unique particles for $d = 4$, while GCDA still produces 28. Notably, both approaches have only a single unique particle for $d = 16$, which suggests that for large-dimensional problems, both approaches are equally (in)efficient.

A candidate solution to the problem of weight degeneracy is to use an intermediate resampling procedure as outlined in Algorithm 1. In Figure 6.5, the top and the middle row show the effective sample sizes obtained with intermediate resampling, i.e., method configurations deviate from the standard version where ($R = M$). Specifically, in the top row, we see the results for choosing $R = 1$, which essentially refers to resampling at every sample step. In the middle row, the results are shown when every $R = M/10$ steps, a resampling operation is performed. The $\widehat{\text{ESS}}$ obtained with GCDA is not affected too much by intermediate resampling. On the other hand, the $\widehat{\text{ESS}}$ obtained with ZDA is affected when $R = 1$, making the methods appear equally efficient as GCDA.

In Figure 6.6b, we see that the number of unique particles is higher, which corresponds with the expectation given the higher $\widehat{\text{ESS}}$. However, the particles are not independent and are nearly identical. This observation underlines that in moderately high dimensions, a pool of N particles reduces to a single sample, even when using intermediate resampling. This is in contrast to from the low-dimensional intuition of importance sampling, where obtaining a diverse set of samples from a pool of N particles may be possible, as seen in the leftmost column of Figure 6.6a.

6.4. Particle Allocation

The purpose of using independent particle pools is to enhance the effectiveness of our approach without attempting to improve the $\widehat{\text{ESS}}$ by efficiently allocating the N particles to K independent runs of our algorithm. In Figure 6.7, we show the results of an experiment that computes the statistical performance of our method in the context of Example 6.3 with $d = 16$. We fix $N = 1028$ and vary K from 1 to N, corresponding to a configuration with a single particle pool and a configuration without any importance sampling, respecitvely. We find a minimum for both variants (ZDA and GCDA) located at around $K = 64$. This observation aligns with our hypothesis that efficiently allocating resources among independent pools can significantly improve the statistical performance.

We briefly focus our attention on understanding this phenomenon by considering the following simplistic proxy of the problem. Assume that we fix N and that the statistical performance of an oracle approach, e.g. by sampling from the exact conditional distribution, improves the performance with $\mathcal{O}(1/K)$. We make this assumption for two reasons. First, it aligns with the rate at which many estimators converge to the true value, e.g., a sample average. Second, we pick $\mathcal{O}(1/K)$ and not $\mathcal{O}(1/N)$, because we know that with K independent pools we get K independent samples instead of N, due to collapsing particle pools. Now, we know that our approach converges to the exact oracle approach with rate $\mathcal{O}((N/K)^{-1})$, i.e., $\mathcal{O}(K)$, as a result of the convergence of importance sampling. Therefore, under these assumptions, the statistical performance of our approach may be related to $\mathcal{O}(1/K) + \mathcal{O}(K)$, which corresponds to the slopes we find in the figures.

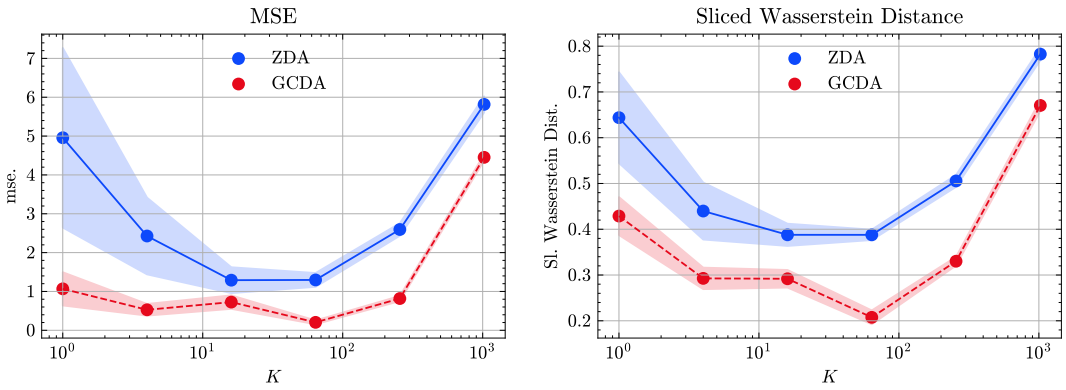


Figure 6.7: Effect of varying number of particle pools. The horizontal axis displays the number of independent particle pools ranging from $K = 1$, referring to a single particle pool to $K = N = 1024$, referring to N independent particles. The vertical axis shows the statistical performance metric. The data model used in the experiment is Example 6.3 with $d = 16$.

7

Application to Scenario Generation

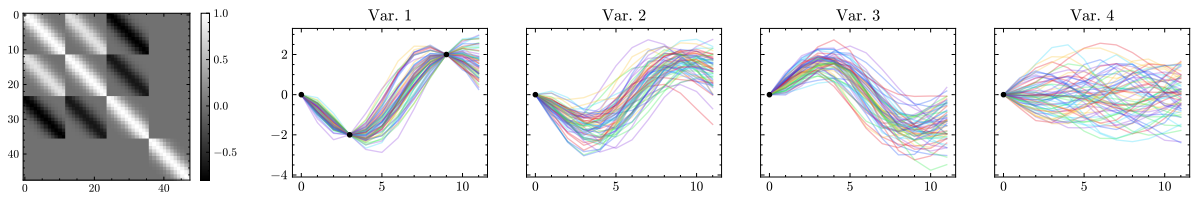
In this chapter, we turn our attention to an application of conditioned generative diffusion in the context of scenario generation. We assume the existence of a pre-trained generative diffusion model that is able to generate unconditional scenarios accurately. The task is to generate scenarios within a specific region of interest, that is, sample from a conditional distribution. This can enhance decision-making by allowing flexible ways to explore risks in various contexts such as finance, economics, meteorology, and climate research. As the focus lies on purely demonstrating the application to scenario generation, we again rely on simplistic Gaussian distributions for the unconditional data. For the same reason, we limit the study to the standard ZDA proposal.

For the performance evaluation, we use a sliced Wasserstein distance and an MSE between the conditional means. In Table 7.1, we depict an overview of the evaluation of the approach in a few examples described throughout this chapter. In most cases, using our approach improves the performance compared to using the pure proposals, even with a relatively small number of particles. However, we also find a counterexample of this, which sheds light on a nuance exhibited by our experimental framework.

In Section 7.1, we study the application of a mask condition in multivariate scenarios. In Section 7.2, we describe conditions on the frequency domain of the scenarios that can be obtained with a discrete cosine transform. In Section 7.3, we describe how our approach can be (heuristically) applied to inequalities.

	(N, K)	Example 7.1	Example 7.2	Example 7.3	Example 7.3	Example 7.5
SWD	(64, 64)	0.8261 (0.0686)	0.2972 (0.0458)	0.0375 (0.0085)	0.5066 (0.0451)	0.6947 (0.0715)
	(64, 2048)	***0.3370 (0.0366)***	0.3676 (0.0624)	0.0246 (0.0058)	***0.3836 (0.0546)***	***0.1692 (0.0348)***
	(2048, 2048)	0.8152 (0.0326)	***0.2873 (0.0143)***	***0.0143 (0.0034)***	0.4972 (0.0333)	0.6275 (0.0359)
MSE	(64, 64)	22.6485 (3.7419)	0.2177 (0.1941)	0.0043 (0.0038)	1.3319 (0.5223)	5.4471 (1.5212)
	(64, 2048)	***1.1378 (0.4946)***	0.8151 (0.5462)	0.0018 (0.0018)	***0.4581 (0.2437)***	***0.0904 (0.0473)***
	(2048, 2048)	21.5198 (0.6581)	***0.1430 (0.0548)***	***0.0004 (0.0007)***	1.2477 (0.1762)	4.6359 (0.2124)

Table 7.1: Result of an application to scenario generation. The quantities are averaged over 10 independent runs. The value in the parentheses gives the standard deviation of the value. Bold-italic entries specify the best performance in terms of that metric.



(a) Equation 7.1. **(b)** Samples from the ground-truth conditioned data.

Figure 7.1: Illustration of Example 7.1. (**Left**): the covariance matrix described in Equation 7.1. The image shows 4 by 4 submatrices that specify the covariances of the distribution of the different variables. Bright colors represent (positive) covariance, and dark colors represent negative covariance. (**Right**): an exact sample of the multivariate scenario distribution. The black dots indicate the mask on which the scenarios are conditioned.

7.1. Multivariate Mask Conditions

To model a k-variate scenario measured at p observation times, we consider a pk-dimensional distribution. Specifically, the tractable Gaussian model relies on two matrices: one $p \times p$ matrix that produces the temporal covariance structures and one $k \times k$ matrix that produces the cross-variate covariance.

Example 7.1 (Multivariate Mask)**.** *The example we consider has $k = 4$ variates and $p = 12$ observations. The mask condition we consider is based on conditioning on a few observations of one of the variables. Let $\Sigma(l)$ represent a 12×12 covariance matrix specified by Equation 6.5 from Example 6.3, where we again pick $l = 0.1$. Then in this example, we define a 48×48 covariance matrix by a Kronecker product between two matrices, i.e.*

$$\Sigma = \begin{pmatrix} 1 & 0.8 & -0.8 & 0 \\ 0.8 & 1 & -0.64 & 0 \\ -0.8 & 0.64 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \otimes \Sigma(l), \tag{7.1}$$

where the left matrix indicates the cross-variable correlations. In Figure 7.1a, the covariance matrix Σ is displayed. The condition we consider is obtained with a matrix $L \in \{0,1\}^{6\times 48}$ that has all elements set to zero except for $\{(1,1),(2,4),(3,10),(4,13),(5,25),(6,37)\}$. Furthermore, we choose $v = \big(0,-2,2,0,0,0\big)^\top$. This condition specifies that all 4 variables start at zero and variable 1 is conditioned to hit -2 at time 4 and to hit 2 at time 10. We use the same technique as in Example 6.3 to obtain a ground-truth conditioned distribution.

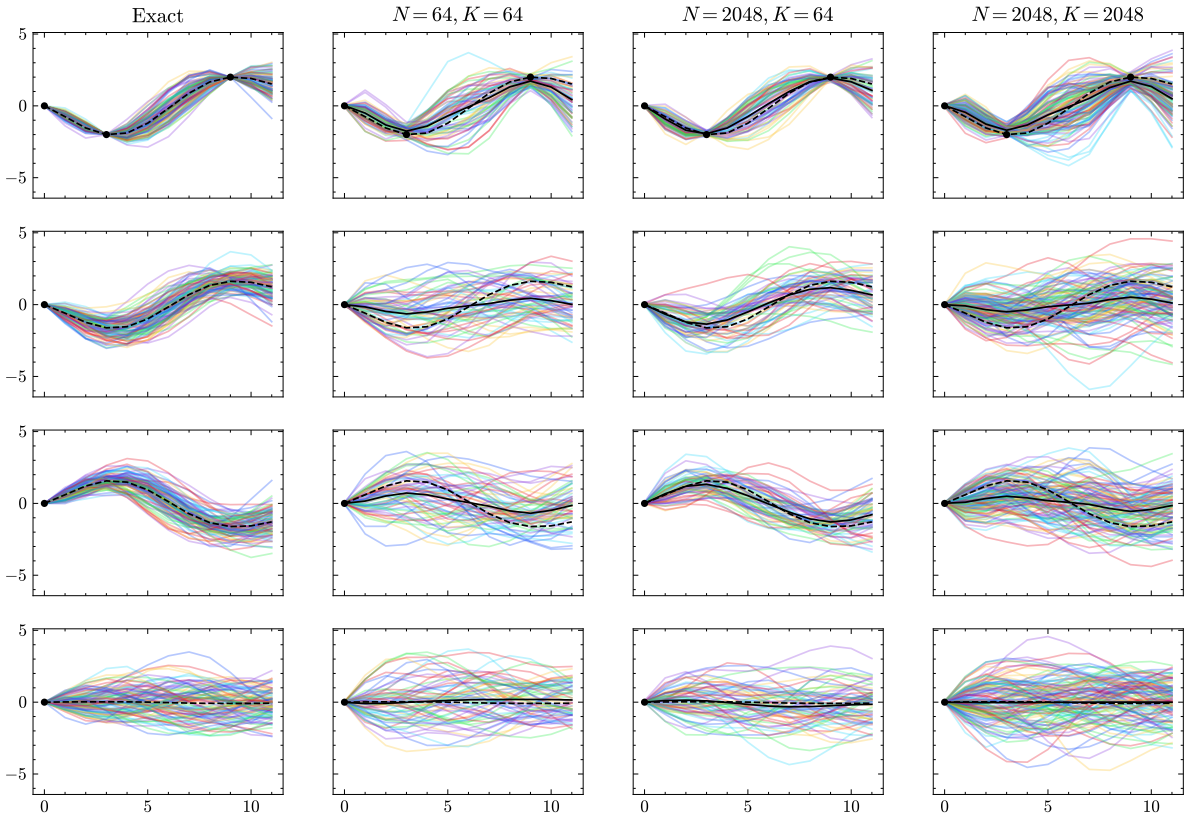


Figure 7.2: Samples of Example 7.1 obtained with different method configurations. The rows of the figure indicate the 4 different variables and the columns represent the 4 different sampling methods. The left most column displays a sample obtained from the exact conditioned distribution akin to Figure 7.1b. The remaining three columns indicate three method configurations for the number of particles N and the number of particle pools K. In the second and fourth column from the left N is equal to K, which means that no importance sampling is used. In the third column, we have $N > K$, which indicates that importance sampling is used within the independent particle pools that have 32 particles each. All experiments are run with the ZDA (Definition 5.2) proposal and $M = 200$ discretization time steps.

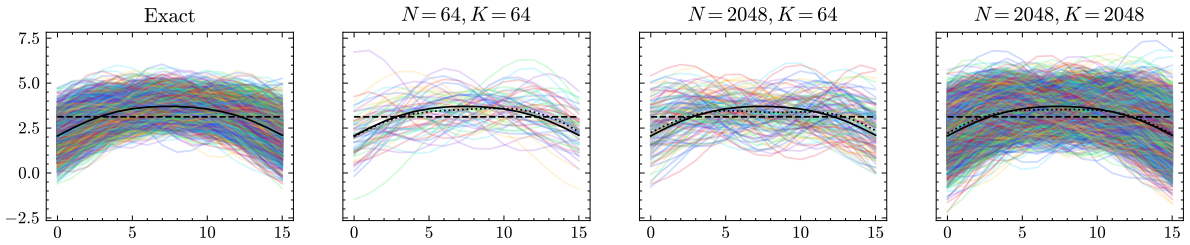


Figure 7.3: Conditioning on zeroth frequency. The panels display scenarios conditioned on the zeroth frequency as described in Example 7.2. The leftmost figure shows a sample of scenarios drawn from the exact conditional distribution. The remaining panels illustrate samples of our approach. The solid black line depicts the ground truth mean, obtained from the exact sample. The dotted line depicts the sample mean obtained from the methods. The dashed line indicates $\bar{v}$ as described in the example.

In Figure 7.2, we find samples of the conditional scenarios specified in Example 7.1, with three different combinations of number of particles N and number of independent particle pools K. The rows represent the 4 different variables, and the columns represent four different approaches that are used to sample the scenarios. In Table 7.1, we find the associated statistical performance metrics.

In the leftmost column, we find a ground truth sample, akin to Figure 7.1b. In the second-to-right column, a pure proposal sample is given with $N = 64$ and $K = 64$. Here, we find that the sample mean obtained with this approach, depicted by the dotted black line, differs from the sample mean of the ground truth, indicated by the solid black line. If we switch to using our importance sampling approach, by increasing the number of particles N to 2048, and keeping the number of independent particle pools at $K = 64$, we find that the sample means are much closer to the ground truth means. On the other hand, if we do increase the particles $N = 2048$ and increase the number of particle pools to $K = 2048$, as in the right-most column, such that we obtain an approach that has roughly the same computational requirements as the $N = 2048, K = 64$ setting, but the procedure draws purely proposal samples, we see that the sample means are not closer to the ground truths. In Table 7.1, we see a significant performance increase when using importance sampling for the masked scenario generation example.

7.2. Frequency Domain Conditions

A condition in the frequency domain can be obtained through using a discrete cosine transform (DCT), which can be described by the following

$$\mathrm{DCT}\,(x)_i = 2\sum_{j}^{d-1} x_j \cos\left(\frac{\pi i(2j+1)}{2d}\right).$$

Applying this operation to the rows of the identity matrix $I_d \in \mathbb{R}^{d\times d}$ yields the DCT matrix $U_{\mathrm{DCT}} \in \mathbb{R}^{d\times d}$, whose i-th row corresponds to the DCT basis vector of frequency i:

$$(U_{\mathrm{DCT}})_{ij} = 2\cos\left(\frac{\pi i(2j+1)}{2d}\right), \quad \text{for } i,j = 0,\ldots,d-1.$$

Thus, the DCT transform can be expressed compactly in matrix form as

$$\mathrm{DCT}(x) = U_{\mathrm{DCT}}x.$$

Conditioning in the frequency domain can be done using this matrix. Specifically, we can select a number $m < d$ and retain only the first m frequencies, i.e., the first m rows of U_{DCT}. For simplicity, we refer to the i-th frequency component of x as:

$$f_i = \mathrm{DCT}(x)_i = (L)_i x.$$

This perspective allows for various interpretations of conditioning. For all examples below, we consider the unconditional Gaussian data distribution of Example 6.3.

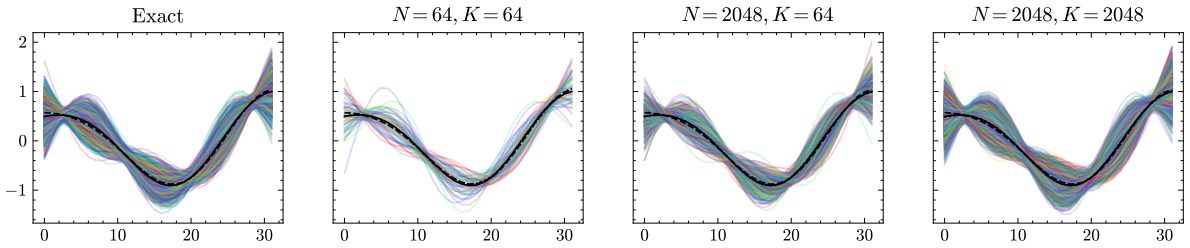


Figure 7.4: Conditioning on low-frequency components. The leftmost panel shows a sample of 200 scenarios drawn from the exact conditional distribution. The remaining figures specify samples of different method configurations. The solid black line depicts the ground truth mean, obtained from the exact sample. The dotted line depicts the sample mean obtained from the methods. The dashed line indicates the low-pass scenarios. The solid line and the dashed and dotted lines overlap in each method configuration, suggesting that our approach does not have a clear benefit. The

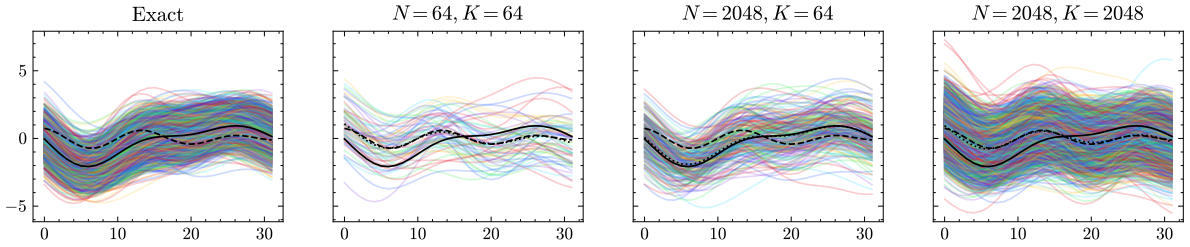


Figure 7.5: Conditioning on mid-frequency components The leftmost panels show 200 scenarios drawn from the exact conditional distribution. The remaining figures specify samples of different method configurations. The solid black line depicts the ground truth mean, obtained from the exact sample. The dotted line depicts the sample mean obtained from the methods. The dashed line indicates the band-passed scenarios.

Example 7.2 (Conditioning on the mean (zeroth frequency))**.** *If we choose only the first row of the DCT matrix (i.e., $i = 0$), then we are effectively conditioning on the mean of the signal. Since the zeroth DCT basis vector is constant, we have: $L = \begin{pmatrix} 2 & 2 & \dots & 2 \end{pmatrix}$. If we then choose $v = \bar{v}/(2d)$, the condition format $Lx = v$ enforces the sample mean of x to be $\bar{v}$. The results of this example are given in Figure 7.3*

Example 7.3 (Conditioning on low-frequency components)**.** *We condition the low-frequency components, i.e. on f_0, f_1, f_2, f_3. This corresponds to conditioning on a low-pass filtering of the scenarios, which preserves trend structures and ignores high-frequency information.*

Example 7.4 (Conditioning on mid-frequency components)**.** *We condition on mid-frequency components, i.e. on f_4, f_5, f_6, f_7. This corresponds to conditioning on a band-pass filtering of the scenarios.*

In Table 7.1, we see a performance increase when using importance sampling for only some of the frequency domain conditioning examples. In particular, those that are associated with the conditioning on low-frequency components (Example 7.2 and Example 7.3) are not affected by importance sampling. This suggests that the conditioning on trends is easy enough in this case, and the proposals are already close enough. On the other hand, using our approach for conditioning on mid-frequency components (Example 7.4) does show a significant performance increase.

7.3. Inequality Conditions

We have postponed the introduction of conditioning on inequalities, as it is fundamentally different to the approach so far. This is because, in principle, we cannot describe an exact auxiliary guidance term $\tilde{h}$ that is similar to the ones we discuss in Chapter 4. Therefore, we make a different approximation of the guidance term, which fundamentally breaks the line of reasoning we use to achieve asymptotic consistency up to this point. Nevertheless, because of our interest towards generalizing beyond singleton conditions, we do provide some illustrative experimental results that explore how the importance sampling behaves for proposals that stretch beyond our theoretically well-understood setting.

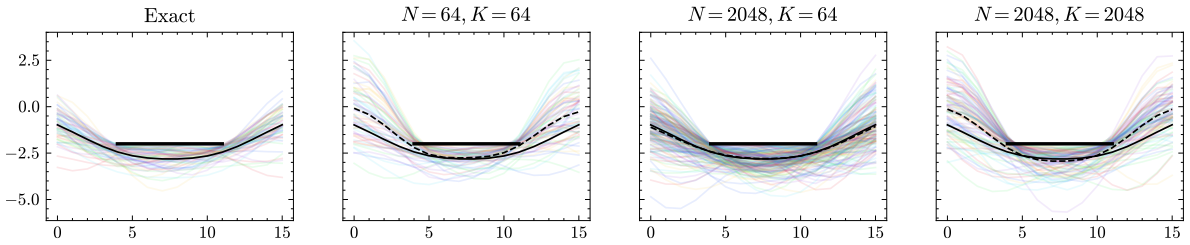


Figure 7.6: Conditioning on inequalities. The leftmost panel shows a sample scenario drawn from the unconditional distribution that satisfies the condition. The remaining figures specify samples of different method configurations. The solid lines indicate the mean of the exact conditioned distribution, and the dashed lines indicate the mean of the approximated conditioned distribution.

For conditions of the form $LY_T \in V \subseteq \mathbb{R}^m$, a guidance term does exist that exactly establishes a conditioned SDE. Specifically, we can make use of

$$\tilde{h}(t, x; V) = \int_V \tilde{h}(t, x; v) \mathrm{d}v,$$

where $\tilde{h}(t, x; v)$ is the auxiliary guidance term for the condition $L\tilde{Y}_t = v$. Unfortunately, this quantity is difficult to compute. Therefore, we use a heuristic method that is based on our approach and can be extended to conditions of the above type by considering the following guidance term,

$$\tilde{h}_{\text{approx}}(t, x) = \tilde{h}(t, x; \hat{v}_x) \text{ where } \hat{v}_x = \arg\min_{v \in V} ||v - L\tilde{\mu}_T(t, x)||. \tag{7.2}$$

The intuition here is that we determine a singleton value $\hat{v}_x$ to derive an equality condition dependent on the prediction of $\tilde{Y}_T$ given $\tilde{Y}_t = x$. It is an open question how the theory relates to choosing such an approximate method.

Example 7.5. *Let us consider the Gaussian data distribution of Example 6.3. Let* $L \in \{0, 1\}^{\lfloor d/4 \rfloor \times d}$ *defined by*

$$L_{ij} = \begin{cases} 1 & \text{if } i = j + \lfloor d/4 \rfloor, \\ 0 & \text{otherwise.} \end{cases}$$

This corresponds to a mask on the middle part of the scenario. Now, we consider $V = \{v \in \mathbb{R}^m : v_i \leq -1 \text{ for all } i \in\}$. *This essentially corresponds to an inequality condition of the form* $LY_t \leq -1$, *where* 1 *is a* m*-dimensional vector of ones. In this case, we have that Equation 7.2 is solved by*

$$(\hat{v}_x)_i = \begin{cases} -1 & \text{if } L\tilde{\mu}_T(t, x) \geq -1, \\ L\tilde{\mu}_T(t, x) & \text{otherwise.} \end{cases}$$

The results of this example are found in Figure 7.6. To obtain a ground truth, in this example, we draw 10^6 *samples from the unconditional distribution and only keep the samples that satisfy the condition. Note that this is only possible for non-rare conditions, such as the one we use here. If the condition is too rare, we can no longer easily obtain a sample from the ground truth conditioned distribution. In Table 7.1, we find a significant performance improvement when using our approach.*

Part IV

Discussion and Conclusion

8 Discussion

In this chapter, we summarize our main results and how they relate to recommendations for future work. The goal of this study is to sample the denoising process of a generative diffusion model conditioned on additional information without additional retraining of neural networks. To this end, we have taken inspiration from an existing simulation approach for diffusion bridges by [SMZ17] and [BMS20], which uses tractable guided proposal processes in combination with importance sampling techniques to guarantee asymptotic consistency.

Our theoretical findings distinguish themselves from those surrounding this fundamental approach due to the incorporation of an adaptive auxiliary framework. In particular, the propagation of gradients makes our work different and requires additional work to validate the condition satisfaction of the proposals and the absolute continuity in Chapter 4. Our core theoretical result is the asymptotic consistency of Chapter 5. It is important to note that this only expresses the asymptotic behavior, as the constant factors in the proportional upper bound have not been determined in this work, and are only known to be independent of the number of particles N and the number of discretization time steps M. It is well-known that the constant that determines the convergence upper bound of importance sampling exponentially depends on the dimensionality problem [Aga+17]. This fact will likely play a significant role in our asymptotic consistency result.

For evaluating the behavior of our approach with finite computational effort, we make use of an empirical study of the statistical performance in Chapter 6. In particular, our approach generally behaves intuitively: using more particles often results in an increased performance. Furthermore, as is also expected, a larger number of dimensions requires more and more computational effort to satisfy the same level of statistical performance. Nevertheless, our approach can significantly improve the statistical performance even for a small number of particles, as demonstrated in the context of conditional scenario generation in Chapter 7. For certain specific conditions, the effect of using importance sampling is negligible. This means that the unweighted proposals are already approximately distributed as the true conditioned distribution, which may be due to the simplistic nature of our experimental settings.

Finally, in Section 7.3, we have provided an appetizer of an extension of our approach. In particular, we experimented with inequality conditions and found positive results. However, it is not entirely clear how severe the consistency is broken when we choose our guidance term in such a heuristic way. This direction is left for future research, along with conditions based on non-linear transformations of the data. A direct application of the theoretical guarantees derived in this work is unlikely to withstand these generalized contexts. From a practical perspective, these extensions would make our approach more in line with the work in [Wu+24; Tri+23], where the conditioning of generative diffusion is not limited to linear conditions and therefore utilizes the entire flexibility of the generative diffusion guidance methods.

In Section 8.1, we discuss our assumptions for our theoretical results. In Section 8.2, we briefly discuss our findings related to the choice of the proposals, and more importantly, how future work can be directed towards improving our approach. In Section 8.3, we discuss an outlook and motivation for the future development of conditioned generative diffusion models.

8.1. Assumptions

A caveat of our theoretical results is that they rely on relatively abstract assumptions, some of which are obtained as substitutes for hard-to-prove, but natural, claims. For future work, it is advised to lift some of these assumptions or make them more concrete, such that their verification is easier. Here, we briefly discuss our assumptions from Chapter 4 and Chapter 5, why they are chosen, and how they might be lifted.

Assumption 4.1 is put in place to be a replacement for the core assumptions that are made about the proposal and auxiliary processes in [BMS20] and [SMZ17]. Assumption 4.1 at first glance, purely relates to the choice of the auxiliary process $\tilde{Y}$. However, as described in the context of (G)CDA (Proposition **??**), the function $\tilde{\mu}_T(t,x)$ may also depend on the unconditional drift, and therefore the learned denoising neural network. In practice, it may hence not be easy to verify claims on $\tilde{\mu}_T(t,x)$ and $\tilde{\mu}_T(t,x)$. To further develop the theoretical understanding, deepening the relationship between the properties of neural networks and the properties required for the assumption is recommended.

Assumption 4.2, is needed to fill in a gap that arises when extending the outline of the proof of [SMZ17] to conditions on linear transformations. It essentially imposes some relatively mild assumptions about the combined behavior of the unconditional drift, the auxiliary drift, and the proposal process Y° on a finite time interval. It, for example, will directly follow if both drifts have bounded norms, or if Y° has a bounded norm and the drifts have linear growth. Applying proof techniques of [BMS20] may help lift the assumption, but this is left for future research.

Assumption 4.3 is taken from [BMS20] and specifies some relation between the unconditional and auxiliary transition densities. It is required to use Lemma 4.6 in the proof of Theorem 4.2. It is generally challenging to verify this assumption, given that it depends on the unknown unconditional transition density.

Assumption 5.1 describes the bounds that are required on the moments of the importance weights and the practical importance weights. These are required for proof of Theorem 5.1. Making assumptions on the moments of the importance weights is relatively standard in importance sampling convergence results and in the context of conditioning generative diffusion, e.g. in [Wu+24]. Note that from Proposition 4.1, we know that importance sampling convergence only requires a bound on the second moment of the importance weights. This may invite the idea that a relaxation of the assumption above is possible. However, because we need to deal with the discretization additionally, the current derivation of the converging upper bound of the squared error relies on a bounded fourth moment. The assumption may be lifted by specifying alternative assumptions on the unconditional process Y and the proposal process Y° instead. However, this may lead to even stricter requirements and is therefore beyond the scope of this research.

Assumption 5.2 is implicitly about the importance weights by assuming a certain behavior of the function G_k in Equation 5.6. The assumption is satisfied if G_k is a Lipschitz continuous function, which it is when $(\tilde{b} - b) \cdot \nabla \log \tilde{h}^k$ is Lipschitz continuous. This follows, for example, if $\tilde{b}$ and b are bounded and Lipschitz continuous and $\nabla \log \tilde{h}$ is Lipschitz continuous. These specific properties seem relatively strict, so future research is required to lift the assumption.

Assumption 5.3 specifies the behavior of the Euler-Maruyama approximation of Y_T° at the boundary of the set A. In the proof of Theorem 5.1, it is difficult to remove this assumption. However, conceptually it aligns with the convergence of the Euler-Maruyama approximation and therefore is not a strict assumption if A is not chosen in an adversarial way.

8.2. Improved Auxiliaries

Our choices regarding the simple auxiliaries call for concrete recommendations on research directions that focus on improving the proposal processes we have introduced in this thesis. The proposals we introduced are based on an auxiliary process for which we can tractably compute the conditioned process. As we have hinted at before in Section 5.3, in many cases it is possible to exactly condition processes with linear (in the state) drift coefficients, due to the presence of an explicit transition normal distribution. Up to this point, we have mainly focused on auxiliary processes that have at most a constant drift, despite being adapted at every discretization. Therefore, it is promising to study auxiliaries with a linear approximation of the drift instead of a constant. However, this work does not include such a (first-order) approximation because the computational and mathematical tractability can suffer greatly from the additional layer of complexity. Here, we lay out a few initial suggestions that may be discussed and whether they are worth considering for future work.

8.2.1. First order expansion of drift

A natural improvement is to use a (gradient propagated) linear drift approximation, abbreviated to (G)LDA, of the form:

$$\tilde{b}(t,x) = b(t_0,x_0) + J_b(t_0,x_0)\cdot(x-x_0), \tag{8.1}$$

where J_b is the Jacobian matrix of the unconditional drift b. This is, in principle, a first-order Taylor expansion of b w.r.t. the state argument at x_0. The associated SDE drives a multivariate time-dependent Ornstein-Uhlenbeck process, and it utilizes the full capacity of the technique of [SMZ17]. Therefore, conditioning the auxiliary process relies on the function $\Phi(t)$, which is the solution to the following matrix ODE:

$$\frac{\mathrm{d}\Phi(t)}{\mathrm{d}t} = J_b(t_0,x_0)\Phi(t), \tag{8.2}$$

for which details on solving linear SDEs can be found in e.g. [Mao11b]. Solving Φ can be (computationally) difficult. While on its own, the computational effort seems dominated by the many matrix multiplications in the passes of neural networks, the repeated use of such an operation poses a significant bottleneck.

8.2.2. Expansions with respect to time

An additional improvement of the proposals is to take the time parameter of the drift function into account. Unfortunately, simply replacing t_0 with t in (G)CDA as given by Definition 5.3, or (G)LDA as given by Equation 8.1, is not tractable, because of the time-dependency of the neural network that is used to approximate the score function. A workaround is to additionally consider the drift approximated as taking the Taylor expansion with respect to the time argument instead of just the state argument. This leads to the following two auxiliary drift coefficients:

$$\tilde{b}(t,x) = b(t_0,x_0) + \partial_t b(t_0,x_0)\,(t-t_0), \tag{8.3}$$

in correspondence to the (G)CDA, and

$$\tilde{b}(t,x) = b(t_0,x_0) + J_b(t_0,x_0)(x-x_0) + \partial_t b(t_0,x_0)\,(t-t_0), \tag{8.4}$$

in correspondence to (G)LDA. It is unclear how much of an improvement these proposals may be over their respective counterparts. The hypothesized subtlety of this improvement, combined with the added complexity to the method, has reserved this improvement for future work.

8.2.3. Reconstruction SDE as an Auxiliary Process

As it turns out, another approximation with a linear drift promises to be beneficial and does not require an eigenvalue decomposition. In particular, the approximation considers the known part of the drift $\alpha(t)x$, which depends on x, but not the score function. This is then achieved by choosing the auxiliary drift to be

$$\tilde{b}(t,x) = \alpha(t)x + \sigma^2(t)\nabla \log q_{T-t_0}(x_0). \tag{8.5}$$

This drift term considers part of the dynamics of the unconditional process in a linear approximation and part of the dynamics in a constant way. There is an interesting link between this linear approximation and reconstruction guidance as introduced in [SE20]. For simplicity, we assume a generative diffusion model with $\alpha(t) = 0$, e.g., a variance exploding model. In that case, the solution of the auxiliary process at time T is

$$\tilde{Y}_T = x_0 + \nabla \log q_{T-t_0}(x_0) \int_{t_0}^{T} \sigma^2(s)\mathrm{d}s + \int_{t_0}^{T} \sigma(s)\mathrm{d}B_s. \tag{8.6}$$

Taking the expectation shares great resemblance with Tweedie's formula ([Efr11]), and hence the reconstruction principle as described in Section 3.3. This line of research would probably give a more accurate auxiliary process than (G)CDA, with less computational effort than the (G)LDA from Equation 8.1.

8.3. Outlook

The debated limitations of deep generative models [TD25], particularly in capturing rare events or tail behavior, present a bleak outlook for their use in domains requiring reliable modeling of risk and uncertainty. While there are attempts to generalize generative diffusion models from Itô processes to Levy processes to capture heavy-tailed distributions [SSD24], it is far from a solved problem.

However, these limitations also motivate the further development of the conditioned generative diffusion technique, particularly the consistent ones. Specifically, here the deep generative model is leveraged not as a complete solution to a statistical task, but rather as a flexible prior from which we sample under constraints as a form of rare-event simulation. This perspective shifts the focus from learning an accurate generative model to designing more effective biased sampling mechanisms, i.e., guided proposals, and correction mechanisms, i.e., importance sampling methods. This practical approach is valuable in domains where rare events are of central interest.

To make this more concrete, we consider a setting inspired by risk management. For example, we consider some (non-linear) function $\varphi : \mathbb{R}^d \to \mathbb{R}$ that maps high-dimensional realized scenarios to a scalar outcome that specifies an associated reward, such as risk metrics, risk-adjusted portfolio returns, or operational profits. If the reward function evaluated on the random data (or scenario) exhibits significant heavy tail behavior, i.e., there, it is unlikely that generative diffusion will be able to learn the correct distribution. Instead, due to the simple range of the reward function, we can choose a value v and attempt to simulate data with our approach that satisfies $\varphi(Y_T) \leq v$.

More broadly, this viewpoint offers a principled way to explore risk-sensitive regions of the data space. However, to fully capitalize on the technique, significant extensions from the work at hand are needed. In particular, the extension to non-linear conditions and non-singleton observations is a recommended research direction. Within these extended approaches, the impact on the theoretical aspects will likely pose significant difficulties.

9 Conclusion

The greater scope of this work is to understand how recent impressive empirically performing generative diffusion models can be used for distribution-sensitive application areas, underlining the need for an appreciation for the safety of the rapidly developing research field. Our main objective within this scope has been to use the flexible controlled generative diffusion techniques fundamental in many statistical tasks, such as scenario-based risk assessment.

Existing approaches to controlling generative diffusion either rely on computationally inefficient retraining of the systems or on heuristics that provide little to no guarantees about the quality of the generated data. We have utilized importance sampling methods from a broader area of continuous-time stochastic processes to ensure that we obtain theoretical guarantees about the consistency of the conditioned generative diffusion models. By adjusting these existing theoretical frameworks to the context of generative diffusion models, we were able to bridge the gap between the theoretically consistent approaches and the practical field of generative diffusion. In doing so, we obtained theoretical validation of the method, a derivation of its asymptotic consistency, and numerical demonstrations of the statistical performance.

Aside from theory, our numerical results cover data generation problems in a theoretically convenient setting, where experiments with low-dimensional and high-dimensional tasks give us insight into the empirical performance of our approach. The approach incites a tradeoff between computational effort and statistical accuracy. In practice, the performance may depend on the specific allocation of resources between the number of particles and the number of desired unique samples. Yet, it is not entirely clear how the interplay between the amount of computational effort and the dimensionality of the problem affects the result, and to what extent our results apply to real-world applications of generative diffusion.

This work promotes generative diffusion models as a practical statistical tool for distribution-sensitive applications. While our method offers promising results, questions remain concerning dimensionality scalability, computational efficiency, and robustness in real-world environments. Continued research and development are needed to unlock the potential of generative modeling in safety-critical domains.

Acknowledgements

This thesis would not have been possible without the guidance and support of my supervisor, Joris Bierkens. I am also grateful to the members of my reading committee, Alex Boer and Richard Kraaij, for their valuable feedback. I want to thank Ortec Finance for the research opportunity and my colleagues there for their support, especially Afrasiab Kadhum and Noud Riemens, whose help in proofreading this thesis was greatly appreciated. I also want to thank Alexandru Băbeanu, whose mentorship during my previous master's thesis significantly shaped the work presented here. As I now close this chapter at TU Delft with two MSc degrees, I extend my deep appreciation to the staff at EEMCS for their continuous support, encouragement, and flexibility throughout this challenging, sometimes unconventional, and ultimately rewarding academic journey.

Appendix

A

Additional Lemmas and Proofs

A.1. Proof of Lemma 2.1

Proof. Consider that

$$\partial_t p_{X_s,X_t}(y,x) = \underbrace{p_{X_t}(x)\partial_t p_{X_s|X_t=x}(y)}_{\text{(I)}} + \underbrace{p_{X_s|X_t=x}(y)\partial_t p_{X_t}(x)}_{\text{(II)}}. \tag{A.1}$$

We may use the Kolmogorov Backward Equation to write

$$\partial_t p_{X_s|X_t=x}(y) \overset{\text{KBE}}{=} -\mathcal{L}_t p_{X_s|X_t=x}(y) = -b(t,x)\cdot\nabla p_{X_s|X_t=x}(y) - \frac{\sigma^2(y)}{2}\Delta p_{X_s|X_t=x}(y). \tag{A.2}$$

Here the operator acts on x and not on the y. Then, we use the following computations that

$$\begin{aligned}
\nabla p_{X_s|X_t=x} &= \frac{1}{p_{X_t}}\nabla p_{X_t,X_s} + p_{X_t,X_s}\nabla\frac{1}{p_{X_t}}\\
\Delta p_{X_s|X_t=x} &= \frac{1}{p_{X_t}}\Delta p_{X_t,X_s} + 2\nabla\frac{1}{p_{X_t}}\cdot\nabla p_{X_t,X_s} + p_{X_t,X_s}\Delta\frac{1}{p_{X_t}}\\
\nabla\frac{1}{p_{X_t}} &= -\frac{1}{p_{X_t}^2}\nabla p_{X\,t}\\
\Delta\frac{1}{p_{X_t}} &= \frac{1}{p_{X_t}^3}\nabla p_{X_t}\cdot\nabla p_{X_t} - \frac{1}{p_{X_t}^2}\Delta p_{X_t}
\end{aligned}$$

we may write

$$\begin{aligned}
\text{(I)} &= -b\cdot\left(\nabla p_{X_t,X_s} - p_{X_s,X_t}\frac{\nabla p_{X_t}}{p_{X_t}}\right)\\
&\quad - \frac{\sigma^2}{2}\left(\Delta p_{X_t,X_s} - 2\frac{\nabla p_{X_t}}{p_{X_t}}\cdot\nabla p_{X_t,X_s} + \left(p_{X_t,X_s}\frac{\nabla p_{X_t}}{p_{X_t}}\cdot\frac{\nabla p_{X_t}}{p_{X_t}} - \frac{p_{X_t,X_s}}{p_{X_t}}\Delta p_{X_t}\right)\right)\\
&= -b\cdot\nabla p_{X_t,X_s} + p_{X_s,X_t}b\cdot\frac{\nabla p_{X_t}}{p_{X_t}}\\
&\quad - \frac{\sigma^2}{2}\Delta p_{X_t,X_s} + \sigma^2\frac{\nabla p_{X_t}}{p_{X_t}}\cdot\nabla p_{X_t,X_s} - \frac{\sigma^2}{2}p_{X_t,X_s}\frac{\nabla p_{X_t}}{p_{X_t}}\cdot\frac{\nabla p_{X_t}}{p_{X_t}} + \frac{\sigma^2}{2}\frac{p_{X_t,X_s}}{p_{X_t}}\Delta p_{X_t}
\end{aligned}$$

Now, we use the KFE to write

$$\text{(II)} = p_{X_s|X_t=x}\partial_t p_{X_t} = -\frac{p_{X_t,X_s}}{p_{X_t}}p_{X_t}\nabla\cdot b - \frac{p_{X_t,X_s}}{p_{X_t}}b\cdot\nabla p_{X_t} + \frac{p_{X_t,X_s}}{p_{X_t}}\frac{\sigma^2}{2}\Delta p_{X_t}$$

Combining all terms gives us

$$
\begin{aligned}
(\mathrm{I})+(\mathrm{II}) &= -b\cdot\nabla p_{X_t,X_s} + p_{X_s,X_t} b\cdot\frac{\nabla p_{X_t}}{p_{X_t}} \\
&\quad - \frac{\sigma^2}{2}\Delta p_{X_t,X_s} + \sigma^2\frac{\nabla p_{X_t}}{p_{X_t}}\cdot\nabla p_{X_t,X_s} - \frac{\sigma^2}{2}p_{X_t,X_s}\frac{\nabla p_{X_t}}{p_{X_t}}\cdot\frac{\nabla p_{X_t}}{p_{X_t}} + \frac{\sigma^2}{2}\frac{p_{X_t,X_s}}{p_{X_t}}\Delta p_{X_t} \\
&\quad - \frac{p_{X_t,X_s}}{p_{X_t}}p_{X_t}\nabla\cdot b - \frac{p_{X_t,X_s}}{p_{X_t}}b\cdot\nabla p_{X_t} + \frac{p_{X_t,X_s}}{p_{X_t}}\frac{\sigma^2}{2}\Delta p_{X_t} \\
&= -b\cdot\nabla p_{X_t,X_s} - \frac{p_{X_t,X_s}}{p_{X_t}}p_{X_t}\nabla\cdot b + \sigma^2\frac{\nabla p_{X_t}}{p_{X_t}}\cdot\nabla p_{X_t,X_s} \\
&\quad + \sigma^2\frac{p_{X_t,X_s}}{p_{X_t}}\Delta p_{X_t} - \frac{\sigma^2}{2}p_{X_t,X_s}\frac{\nabla p_{X_t}}{p_{X_t}}\cdot\frac{\nabla p_{X_t}}{p_{X_t}} - \frac{\sigma^2}{2}\Delta p_{X_t,X_s}
\end{aligned}
$$

Then, consider

$$
\nabla\cdot((b-\sigma^2\nabla\log p_{X_t})p_{X_t,X_s}) = b\cdot\nabla p_{X_t,X_s} + p_{X_t,X_s}\nabla\cdot b - \sigma^2\nabla\log p_{X_t}\cdot\nabla p_{X_t,X_s} - p_{X_t,X_s}\sigma^2\Delta\log p_{X_t}
$$

where we can recognize all terms but the last. For this, we write

$$
\begin{aligned}
\Delta\log p_{X_t} &= \nabla\cdot\frac{\nabla p_{X_t}}{p_{X_t}} \\
&= \frac{1}{p_{X_t}}\Delta p_{X_t} - \frac{1}{p_{X_t}^2}\nabla p_{X_t}\cdot\nabla p_{X_t}
\end{aligned}
$$

Now, combining the terms, we obtain the desired expression.

□

A.2. Proof of Proposition 3.4

Let $f\in C^{1,2}([0,T]\times\mathbb{R}^d)$. **Item 1.** The first part of the proof is adapted from Appendix D. in [BMS20]. By Chapman-Kolmogorov, we have that

$$
\mathbb{E}\left[f(s,Y_s)|Y_t=x, LY_T=v\right] = \int f(s,y)p_{Y_s|Y_t=x}(y)\frac{h(s,y)}{h(t,x)}\mathrm{d}y
$$

We use that for some infinitesimal generator $\mathcal{L}_t$ and function $f\in C^{1,2}([0,T]\times\mathbb{R}^d)$, we have that

$$
(\partial_t+\mathcal{L}_t)f(t,x) = \lim_{s\downarrow t}\frac{\mathbb{E}\left[f(s,Y_s)|Y_t=x\right]-f(t,Y_t)}{s-t} \tag{A.3}
$$

Therefore,

$$
\begin{aligned}
(\partial_t+\mathcal{L}_t^*)f(t,x) &= \lim_{s\downarrow t}\frac{1}{h(t,x)}\frac{\mathbb{E}\left[f(s,Y_s)h(t,Y_t)|Y_t=x, LY_T=v\right]-f(t,Y_t)h(t,Y_t)}{s-t} \\
&= \lim_{s\downarrow t}\frac{1}{h(t,x)}\frac{\int f(s,y)p_{Y_s|Y_t=x}(y)h(s,y)\mathrm{d}y - f(t,Y_t)h(t,Y_t)}{s-t} \\
&= \lim_{s\downarrow t}\frac{1}{h(t,x)}\frac{\mathbb{E}\left[f(s,Y_s)h(s,Y_s)|Y_t=x\right]-f(t,Y_t)h(t,Y_t)}{s-t} \\
&= \frac{1}{h(t,x)}((\partial_t+\mathcal{L}_t)fh)(t,x)
\end{aligned}
$$

Item 2. We know that by the product rule

$$\begin{aligned}\mathcal{L}_t fh &= hb\cdot\nabla f + fb\cdot\nabla h + \frac{\sigma^2}{2}\Delta fh \\ &= hb\cdot\nabla f + fb\cdot\nabla h + \frac{\sigma^2}{2}f\Delta h + \sigma^2\nabla f\cdot\nabla h + h\frac{\sigma^2}{2}\Delta f\end{aligned}$$

Then

$$\begin{aligned}\frac{1}{h}(\partial_t+\mathcal{L}_t)fh &= \partial_t f + \frac{f}{h}\partial_t h + b\cdot\nabla f + \frac{f}{h}b\cdot\nabla h + \frac{f}{h}\frac{\sigma^2}{2}\Delta h + \frac{\sigma^2}{h}\nabla f\cdot\nabla h + \frac{\sigma^2}{2}\Delta f \\ &= \partial_t f + \frac{f}{h}\left(\partial_t h + \mathcal{L}_t h\right) + b\cdot\nabla f + \frac{\sigma^2}{h}\nabla f\cdot\nabla h + \frac{\sigma^2}{2}\Delta f.\end{aligned}$$

Then using that h is space-time harmonic, we may write

$$\begin{aligned}\frac{1}{h}(\partial_t+\mathcal{L}_t)fh &= \partial_t f + b\cdot\nabla f + \frac{\sigma^2}{h}\nabla f\cdot\nabla h + \frac{\sigma^2}{2}\Delta f \\ &= \partial_t f + (b+\sigma^2\nabla\log h)\cdot\nabla f + \frac{\sigma^2}{2}\Delta f.\end{aligned}$$

From this, we recognize the coefficients of the conditioned SDE in Equation 3.4.

A.3. Proof of Lemma 4.1

Proof. Let us temporarily denote

$$P_N^*(A) = \frac{1}{N}\sum_{i=1}^N W^*(Y^{(i)})\mathbf{1}_A(Y^{(i)}).$$

Writing out the variance gives us

$$\mathbb{E}^\circ\left[(P_N^*(A))^2\right] - 2\mathbb{E}^\circ\left[P_N^*(A)\right]\mathbb{P}_Y^*(A) + (\mathbb{P}_Y^*(A))^2 = \mathbb{E}^\circ\left[(P_N^*(A))^2\right] - (\mathbb{P}_Y^*(A))^2.$$

Then $(P_N^*(A))^2$ can be evaluated to be

$$\frac{1}{N^2}\sum_{i=1}^N\left(W^*(Y^{(i)})\right)^2\mathbf{1}_A(Y^{(i)}) + \frac{1}{N^2}\sum_{i=1}^N\sum_{j=1,i\neq j}^N\left(W^*(Y^{(i)})\right)\mathbf{1}_A(Y^{(i)})\left(W^*(Y^{(j)})\right)\mathbf{1}_A(Y^{(j)}).$$

Because $Y^{(i)}$ and $Y^{(j)}$ are independent draws of $\mathbb{P}_Y^\circ$, we have that the expectation of the above amounts to

$$\frac{1}{N}\mathbb{E}^\circ\left[\left(W^*(Y^{(1)})\right)^2\mathbf{1}_A(Y^{(1)})\right] + \frac{N-1}{N}\mathbb{E}^\circ\left[\left(W^*(Y^{(1)})\right)\mathbf{1}_A(Y^{(1)})\right]^2.$$

Using the fact that W^* is the Radon-Nikodym derivative, we can write the above as the first term of the left hand side of the inequality below. Then, the inequality follows from the non-negativity of $\mathbb{P}_Y^*(A)$ and the fact that $\mathbf{1}_A \leq 1$.

$$\left(\frac{1}{N}\mathbb{E}^\circ\left[\left(W^*(Y^{(1)})\right)^2\mathbf{1}_A(Y^{(1)})\right] + \frac{N-1}{N}(\mathbb{P}_Y^*(A))^2\right) - (\mathbb{P}_Y^*(A))^2 \leq \frac{1}{N}\mathbb{E}^\circ\left[\left(W^*(Y^{(1)})\right)^2\right]$$

□

A.4. Proof of Proposition 4.1

Proof. The proof roughly follows that of [Aga+17]. Let us adopt the notation of $\mathbf{1}_A = \mathbf{1}_A(Y)$ and $W = W(Y)$ and consider that

$$\mathbb{P}^*_Y(A) = \frac{\mathbb{E}\left[\mathbf{1}_A W\right]}{\mathbb{E}\left[W\right]}.$$

Now, we consider the following rewriting:

$$\begin{aligned}
P^*_N(A) - \mathbb{P}^*_Y(A) &= \left(\frac{\frac{1}{N}\sum_{i=1}^N W^{(i)} \mathbf{1}_A^{(i)}}{\frac{1}{N}\sum_{i=1}^N W^{(i)}} - \frac{\mathbb{E}^\circ\left[\mathbf{1}_A W\right]}{\mathbb{E}^\circ\left[W\right]} \right) \\
&= \left(\frac{1}{\frac{1}{N}\sum_{i=1}^N W^{(i)}} - \frac{1}{\mathbb{E}^\circ\left[W\right]} \right) \frac{1}{N}\sum_{i=1}^N W^{(i)} \mathbf{1}_A^{(i)} - \frac{1}{\mathbb{E}^\circ\left[W\right]} \left(\frac{1}{N}\sum_{i=1}^N W^{(i)} \mathbf{1}_A^{(i)} - \mathbb{E}^\circ\left[\mathbf{1}_A W\right] \right) \\
&= \frac{1}{\mathbb{E}^\circ[W]} \left(\left(\mathbb{E}^\circ\left[W\right] - \frac{1}{N}\sum_{i=1}^N W^{(i)} \right) P^*_N(A) - \left(\frac{1}{N}\sum_{i=1}^N W^{(i)} \mathbf{1}_A^{(i)} - \mathbb{E}^\circ\left[\mathbf{1}_A W\right] \right) \right)
\end{aligned}$$

Then we use $(x-y)^2 \leq 2(x^2+y^2)$, $P^*_N(A) \leq 1$, and $\mathbf{1}_A \leq 1$ to bound the expectation

$$\mathbb{E}^\circ\left[(P^*_N(A) - \mathbb{P}^*_Y(A))^2\right] \lesssim \frac{1}{(\mathbb{E}^\circ[W])^2} \mathbb{E}^\circ\left[\left(\mathbb{E}^\circ\left[W\right] - \frac{1}{N}\sum_{i=1}^N W^{(i)} \right)^2 \right].$$

From this it is not hard to see that

$$\mathbb{E}^\circ\left[(P^*_N(A) - \mathbb{P}^*_Y(A))^2\right] \lesssim \frac{1}{N} \frac{\mathbb{E}^\circ\left[\left(\mathbb{E}^\circ\left[W\right] - W\right)^2\right]}{(\mathbb{E}^\circ[W])^2} \lesssim \frac{1}{N}.$$

□

A.5. Proof of Lemma 4.2

Let us consider the ratio between the two known Gaussian transition densities, i.e.

$$\frac{\hat{p}(y_s; y_t)}{\hat{p}^\circ(y_s; y_t)} = \frac{\mathcal{N}\left(y_s; y_t + b(t, y_t)M^{-1}, \sigma^2(t)M^{-1}\right)}{\mathcal{N}\left(y_s; y_t + b^\circ(t, y_t)M^{-1}, \sigma^2(t)M^{-1}\right)}.$$

where we use that

$$y_s = y_t + b^\circ_t M^{-1} + (B_s - B_t)\sigma_t \text{ for } (B_s - B_t) \sim \mathcal{N}(0, I(s-t))$$

This gives us that

$$||y_s - y_t - b^\circ_t M^{-1}||^2 = ||(y_t + b^\circ_t M^{-1} + (B_s - B_t)\sigma_t) - y_t - b^\circ_t M^{-1}||^2 = ||(B_s - B_t)\sigma_t||^2$$

and

$$||y_s - y_t - b_t M^{-1}||^2 = ||(y_t + b^\circ_t M^{-1} + (B_s - B_t)\sigma_t) - y_t - b_t M^{-1}||^2 = ||b^\circ_t M^{-1} - b_t M^{-1} + (B_s - B_t)\sigma_t||^2$$

Decomposing the product, we obtain

$$M^{-12}||b^\circ_t - b_t||^2 + 2M^{-1}(b^\circ_t - b_t) \cdot ((B_s - B_t)\sigma_t) + ||(B_s - B_t)\sigma_t||^2$$

Let us now consider writing $\eta = b_t - b_t^\circ$

$$M^{-1}{}^2\sigma_t^2||\eta_t||^2 - 2M^{-1}\sigma_t^2\eta_t \cdot (B_s - B_t) + ||(B_s - B_t)\sigma_t||^2$$

Then, the log of the weight defined in equation (4.8), can be written as

$$\begin{aligned}
\log \hat{W} &= \sum_{(t,s)\in G} ||(B_s - B_t)\sigma_t^2||^2(2M^{-1}\sigma_t^2)^{-1} - ||b_t^\circ - b_t + (B_s - B_t)\sigma_t^2||^2(2M^{-1}\sigma_t^2)^{-1} \\
&= \sum_{(t,s)\in G} -M^{-2}\sigma_t^2||\eta_t||^2(2M^{-1}\sigma_t^2)^{-1} + 2M^{-1}\sigma_t^2\eta_t \cdot (B_s - B_t)(2M^{-1}\sigma_t^2)^{-1} \\
&= -\sum_{(t,s)\in G} \frac{1}{2}||\eta_t||^2M^{-1} + \sum_{(t,s)\in G} \eta_t \cdot (B_s - B_t),
\end{aligned}$$

which gives us the desired result.

Bibliography

[Aga+17] S. Agapiou, O. Papaspiliopoulos, D. Sanz-Alonso, and A. M. Stuart. *Importance Sampling: Intrinsic Dimension and Computational Cost*. 2017. DOI: 10.48550/arXiv.1511.06196.

[AKK18] S. Aida, T. Kikuchi, and S. Kusuoka. "The rates of the L^p-convergence of the Euler-Maruyama and Wong-Zakai approximations of path-dependent stochastic differential equations under the Lipschitz condition". In: *Tohoku Mathematical Journal* 70.1 (2018), pp. 65–95. ISSN: 0040-8735, 2186-585X. DOI: 10.2748/tmj/1520564419.

[And82] B. D. Anderson. "Reverse-time diffusion equation models". en. In: *Stochastic Processes and their Applications* 12.3 (1982), pp. 313–326. ISSN: 03044149. DOI: 10.1016/0304-4149(82)90051-5.

[Ben23] J. Benton. "Generative models: theory and applications". English. http://purl.org/dc/dcmitype/Text. University of Oxford, 2023.

[Bis06] C. M. Bishop. *Pattern recognition and machine learning*. en. Information science and statistics. New York: Springer, 2006. ISBN: 978-0-387-31073-2.

[BMS20] J. Bierkens, F. v. d. Meulen, and M. Schauer. "Simulation of elliptic and hypo-elliptic conditional diffusions". In: *Advances in Applied Probability* 52.1 (2020), pp. 173–212. ISSN: 0001-8678, 1475-6064. DOI: 10.1017/apr.2019.54.

[Bon+15] N. Bonneel, J. Rabin, G. Peyré, and H. Pfister. "Sliced and Radon Wasserstein Barycenters of Measures". en. In: *Journal of Mathematical Imaging and Vision* 51.1 (2015), pp. 22–45. ISSN: 1573-7683. DOI: 10.1007/s10851-014-0506-3.

[Büh+20] H. Bühler, B. Horvath, T. Lyons, I. P. Arribas, and B. Wood. *A Data-driven Market Simulator for Small Data Environments*. 2020. DOI: 10.48550/arXiv.2006.14498.

[Cao+23] H. Cao, C. Tan, Z. Gao, Y. Xu, G. Chen, P.-A. Heng, and S. Z. Li. *A Survey on Generative Diffusion Model*. 2023. DOI: 10.48550/arXiv.2209.02646.

[Che+18] Y. Chen, Y. Wang, D. Kirschen, and B. Zhang. *Model-Free Renewable Scenario Generation Using Generative Adversarial Networks*. 2018.

[Che+23] S. Chen, S. Chewi, J. Li, Y. Li, A. Salim, and A. R. Zhang. *Sampling is as easy as learning the score: theory for diffusion models with minimal data assumptions*. en. 2023. DOI: 10.48550/arXiv.2209.11215.

[Chu+24] H. Chung, J. Kim, M. T. Mccann, M. L. Klasky, and J. C. Ye. *Diffusion Posterior Sampling for General Noisy Inverse Problems*. 2024. DOI: 10.48550/arXiv.2209.14687.

[CKS23] Z. Chang, G. A. Koulieris, and H. P. H. Shum. *On the Design Fundamentals of Diffusion Models: A Survey*. 2023. DOI: 10.48550/arXiv.2306.04542.

[Cla90] J. Clark. "The simulation of pinned diffusions". In: *29th IEEE Conference on Decision and Control*. 1990, 1418–1420 vol.3. DOI: 10.1109/CDC.1990.203845.

[Col+23] A. Coletta, S. Gopalakrishan, D. Borrajo, and S. Vyetrenko. "On the constrained time-series generation problem". In: *Proceedings of the 37th International Conference on Neural Information Processing Systems*. NIPS '23. Red Hook, NY, USA: Curran Associates Inc., 2023, pp. 61048–61059.

[Cor24] M. A. Corstanje. "Guiding techniques for conditioning Markov processes". en. dr. Vrije Universiteit Amsterdam, 2024. DOI: 10.5463/thesis.792.

[DN21] P. Dhariwal and A. Nichol. *Diffusion Models Beat GANs on Image Synthesis*. 2021. DOI: 10.48550/arXiv.2105.05233.

[Efr11] B. Efron. "Tweedie's Formula and Selection Bias". In: *Journal of the American Statistical Association* 106.496 (2011), pp. 1602–1614. ISSN: 0162-1459. DOI: 10.1198/jasa.2011.tm11181.

[FJ22] S. Flaig and G. Junike. “Scenario Generation for Market Risk Models Using Generative Neural Networks”. In: *Risks* 10.11 (2022), pp. 1–28.

[He+24] Y. He, N. Murata, C.-H. Lai, Y. Takida, T. Uesaka, D. Kim, W.-H. Liao, Y. Mitsufuji, J. Z. Kolter, R. Salakhutdinov, and S. Ermon. “MANIFOLD PRESERVING GUIDED DIFFUSION”. en. In: (2024).

[HJA20] J. Ho, A. Jain, and P. Abbeel. *Denoising Diffusion Probabilistic Models*. 2020. DOI: 10.48550/arXiv.2006.11239.

[Ho+22] J. Ho, T. Salimans, A. Gritsenko, W. Chan, M. Norouzi, and D. J. Fleet. *Video Diffusion Models*. 2022. DOI: 10.48550/arXiv.2204.03458.

[HS22] J. Ho and T. Salimans. *Classifier-Free Diffusion Guidance*. 2022. DOI: 10.48550/arXiv.2207.12598.

[Hua+24] L. Huang, L. Gianinazzi, Y. Yu, P. D. Dueben, and T. Hoefler. *DiffDA: a Diffusion Model for Weather-scale Data Assimilation*. 2024. DOI: 10.48550/arXiv.2401.05932.

[Jia+19] C. Jiang, Y. Chen, Y. Mao, Y. Chai, and M. Yu. *Forecasting Spatio-Temporal Renewable Scenarios: a Deep Generative Approach*. 2019. DOI: 10.48550/arXiv.1903.05274.

[LBB05] B. Li, T. Bengtsson, and P. Bickel. *Curse-of-dimensionality revisited: Collapse of importance sampling in very high-dimensional systems*. eng. 2005.

[Li+24] H. Li, H. Yu, Z. Liu, F. Li, X. Wu, B. Cao, C. Zhang, and D. Liu. “Long-term scenario generation of renewable energy generation using attention-based conditional generative adversarial networks”. en. In: *Energy Conversion and Economics* 5.1 (2024), pp. 15–27. ISSN: 2634-1581. DOI: 10.1049/enc2.12106.

[Lug+22] A. Lugmayr, M. Danelljan, A. Romero, F. Yu, R. Timofte, and L. V. Gool. *RePaint: Inpainting using Denoising Diffusion Probabilistic Models*. 2022. DOI: 10.48550/arXiv.2201.09865.

[Ma+24] Z. Ma, Y. Zhang, G. Jia, L. Zhao, Y. Ma, M. Ma, G. Liu, K. Zhang, J. Li, and B. Zhou. *Efficient Diffusion Models: A Comprehensive Survey from Principles to Practices*. 2024. DOI: 10.48550/arXiv.2410.11795.

[Mao11a] X. Mao. “1 - Brownian Motions and Stochastic Integrals”. In: *Stochastic Differential Equations and Applications (Second Edition)*. Ed. by X. Mao. Woodhead Publishing, 2011, pp. 1–46. ISBN: 978-1-904275-34-3. DOI: 10.1533/9780857099402.1.

[Mao11b] X. Mao. “3 - Linear Stochastic Differential Equations”. In: *Stochastic Differential Equations and Applications (Second Edition)*. Ed. by X. Mao. Woodhead Publishing, 2011, pp. 91–106. ISBN: 978-1-904275-34-3. DOI: 10.1533/9780857099402.91.

[Mar12] J.-L. J. L. Marchand. “Conditionnement de processus markoviens”. fr. PhD thesis. Université Rennes 1, 2012.

[Øks03] B. Øksendal. “Stochastic Differential Equations”. In: *Stochastic Differential Equations*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2003, pp. 65–84. ISBN: 978-3-540-04758-2 978-3-642-14394-6. DOI: 10.1007/978-3-642-14394-6_5.

[Pan+24] K. Pandey, J. Pathak, Y. Xu, S. Mandt, M. Pritchard, A. Vahdat, and M. Mardani. *Heavy-Tailed Diffusion Models*. 2024.

[PR02] Z. Palmowski and T. Rolski. “A technique for exponential change of measure for Markov processes”. In: *Bernoulli* 8.6 (2002), pp. 767–785. ISSN: 1350-7265.

[Pri+24] I. Price, A. Sanchez-Gonzalez, F. Alet, T. R. Andersson, A. El-Kadi, D. Masters, T. Ewalds, J. Stott, S. Mohamed, P. Battaglia, R. Lam, and M. Willson. *GenCast: Diffusion-based ensemble forecasting for medium-range weather*. 2024. DOI: 10.48550/arXiv.2312.15796.

[Rad+21] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. *Learning Transferable Visual Models From Natural Language Supervision*. 2021. DOI: 10.48550/arXiv.2103.00020.

[Ram+21] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever. *Zero-Shot Text-to-Image Generation*. 2021. DOI: 10.48550/arXiv.2102.12092.

[Rom+22] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. *High-Resolution Image Synthesis with Latent Diffusion Models*. 2022. DOI: 10.48550/arXiv.2112.10752.

[SE20] Y. Song and S. Ermon. *Generative Modeling by Estimating Gradients of the Data Distribution*. 2020. DOI: 10.48550/arXiv.1907.05600.

[She+24] Y. Shen, X. Jiang, Y. Wang, Y. Yang, D. Han, and D. Li. *Understanding and Improving Training-free Loss-based Diffusion Guidance*. 2024. DOI: 10.48550/arXiv.2403.12404.

[SMZ17] M. Schauer, F. v. d. Meulen, and H. v. Zanten. "Guided proposals for simulating multi-dimensional diffusion bridges". In: *Bernoulli* 23.4A (2017). ISSN: 1350-7265. DOI: 10.3150/16-BEJ833.

[Son+21] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole. *Score-Based Generative Modeling through Stochastic Differential Equations*. 2021. DOI: 10.48550/arXiv.2011.13456.

[Son+23] J. Song, Q. Zhang, H. Yin, M. Mardani, M.-Y. Liu, J. Kautz, Y. Chen, and A. Vahdat. "Loss-Guided Diffusion Models for Plug-and-Play Controllable Generation". en. In: *Proceedings of the 40th International Conference on Machine Learning*. PMLR, 2023, pp. 32483–32498. DOI: 10.5555/3618408.3619753.

[SRH24] M. F. Sikder, R. Ramachandranpillai, and F. Heintz. *TransFusion: Generating Long, High Fidelity Time Series using Diffusion Models with Transformers*. en. 2024. DOI: 10.48550/arXiv.2307.12667.

[SSD24] D. Shariatian, U. Simsekli, and A. Durmus. *Denoising Lévy Probabilistic Models*. 2024. DOI: 10.48550/arXiv.2407.18609.

[TD25] E. Tam and D. B. Dunson. *On the Statistical Capacity of Deep Generative Models*. 2025. DOI: 10.48550/arXiv.2501.07763.

[Tri+23] B. L. Trippe, J. Yim, D. Tischer, D. Baker, T. Broderick, R. Barzilay, and T. Jaakkola. *Diffusion probabilistic modeling of protein backbones in 3D for the motif-scaffolding problem*. 2023. DOI: 10.48550/arXiv.2206.04119.

[Vin11] P. Vincent. "A Connection Between Score Matching and Denoising Autoencoders". en. In: *Neural Computation* 23.7 (2011), pp. 1661–1674. ISSN: 0899-7667, 1530-888X. DOI: 10.1162/NECO_a_00142.

[Wu+24] L. Wu, B. L. Trippe, C. A. Naesseth, D. M. Blei, and J. P. Cunningham. *Practical and Asymptotically Exact Conditional Sampling in Diffusion Models*. 2024. DOI: 10.48550/arXiv.2306.17775.

[Xin+24] Z. Xing, Q. Feng, H. Chen, Q. Dai, H. Hu, H. Xu, Z. Wu, and Y.-G. Jiang. *A Survey on Video Diffusion Models*. 2024. DOI: 10.48550/arXiv.2310.10647.

[XKV22] Z. Xiao, K. Kreis, and A. Vahdat. *Tackling the Generative Learning Trilemma with Denoising Diffusion GANs*. 2022. DOI: 10.48550/arXiv.2112.07804.

[Yan+24] Y. Yang, M. Jin, H. Wen, C. Zhang, Y. Liang, L. Ma, Y. Wang, C. Liu, B. Yang, Z. Xu, J. Bian, S. Pan, and Q. Wen. *A Survey on Diffusion Models for Time Series and Spatio-Temporal Data*. en. 2024. DOI: 10.48550/arXiv.2404.18886.

[Ye+24] H. Ye, H. Lin, J. Han, M. Xu, S. Liu, Y. Liang, J. Ma, J. Zou, and S. Ermon. *TFG: Unified Training-Free Guidance for Diffusion Models*. 2024. DOI: 10.48550/arXiv.2409.15761.

[YQ24] X. Yuan and Y. Qiao. *Diffusion-TS: Interpretable Diffusion for General Time Series Generation*. 2024.

[Yu+23] J. Yu, Y. Wang, C. Zhao, B. Ghanem, and J. Zhang. *FreeDoM: Training-Free Energy-Guided Conditional Diffusion Model*. 2023. DOI: 10.48550/arXiv.2303.09833.

[Zha+23] C. Zhang, C. Zhang, S. Zheng, M. Zhang, M. Qamar, S.-H. Bae, and I. S. Kweon. *A Survey on Audio Diffusion Models: Text To Speech Synthesis and Enhancement in Generative AI*. 2023. DOI: 10.48550/arXiv.2303.13336.