Revisiting Sampling Strategies for Molecular Generation

Yuyan Ni [1 2 3] **Shikun Feng** [3] **Wei-Ying Ma** [3] **Zhi-Ming Ma** [1] **Yanyan Lan** [3 4]

Abstract

Sampling strategies in diffusion models are critical to molecular generation yet remain relatively underexplored. In this work, we investigate a broad spectrum of sampling methods beyond conventional defaults and reveal that sampling choice substantially affects molecular generation performance. In particular, we identify a maximally stochastic sampling (StoMax), a simple yet underexplored strategy, as consistently outperforming default sampling methods for generative models DDPM and BFN. Our findings highlight the pivotal role of sampling design and suggest promising directions for advancing molecular generation through principled and more expressive sampling approaches.

1. Introduction

Molecular generation has emerged as a crucial task in AI-driven drug and material discovery, enabling the rapid exploration of chemical space for novel and functional compounds. Among the many approaches to generative modeling, diffusion models have recently achieved state-of-the-art performance in 3D molecular generation tasks due to their superior capability in precisely modeling atomic positions.

While extensive research has focused on improving model architectures and training objectives, a key component that remains relatively under-explored is the sampling strategy, the procedure by which new molecules are generated from the learned diffusion models. In most existing approaches, including Equivariant Diffusion Models (EDM) (Hoogeboom et al., 2022) and Geometric Bayesian Flow Networks (GeoBFN) (Song et al., 2023), the sampling process is inherently tied to the model's design. Specifically, these methods adopt the default sampling method of Denoising Diffusion Probabilistic Models (DDPM) (Ho et al., 2020) and Bayesian Flow Networks (BFN) (Graves et al., 2023), which correspond to first-order discretizations of the reverse-time stochastic differential equations (SDEs) of the diffusion process.

However, the default sampling strategies commonly used in diffusion models are not the only theoretically valid choices. In fact, a broader family of sampling methods exists, each defined by distinct assumptions about temporal dependencies in the generative process. These variations can substantially affect the quality of generated samples, and hold untapped potential for improving molecular generation performance.

To further comprehend this design space, we first identify two representative cases that conclude widely adopted strategies: (1) a Markov forward process, which underlies DDPM and BFN samplers, and (2) a Deterministic reverse process, corresponding to DDIM (Song et al., 2021a) and ODE-based sampling (Song et al., 2021b). Beyond these well-studied methods, we introduce a third, intuitive alternative: (3) a conditionally independent reverse process, in which the noisy sample at each reverse timestep is conditionally independent of the previous timestep given the initial data. This method induces maximal stochasticity in the family of sampling methods, we termed maximally stochastic sampling (StoMax).

We systematically evaluate StoMax, a previously underexplored sampling alternative, across multiple molecular generative models, including UniGEM(EDM), UniGEM(BFN) (Feng et al., 2025), EDM, and GeoBFN. Empirically, StoMax consistently outperforms the native samplers of these models on both the QM9 and GEOM-Drugs datasets. Notably, StoMax brings substantial improvements under the DDPM noise schedule, fully leveraging the capacity of pretrained generative models. This enables UniGEM(EDM) to achieve state-of-the-art performance in molecular generation. Interestingly, DDIM and StoMax represent the two extremes of the sampling family in terms of stochasticity, with DDPM and BFN default sampling falling in between. To probe the potential of this design space, we interpolate between the three sampling strategies and find that StoMax yields the best overall sample quality, with a minor trade-off

[1]Academy of Mathematics and Systems Science, Chinese Academy of Sciences [2]University of Chinese Academy of Sciences [3]Institute for AI Industry Research (AIR), Tsinghua University (Work was done during Yuyan's internship at AIR.) [4]Beijing Academy of Artificial Intelligence. Correspondence to: Yanyan Lan <lanyanyan@air.tsinghua.edu.cn>.

Proceedings of the Workshop on Generative AI for Biology at the 42nd International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025.

in diversity.

These empirical observations motivate us to explore more diverse and expressive sampling strategies beyond conventional choices, aiming to further enhance the quality of molecular generation. The results also motivate the development of theoretical frameworks that can explain these empirical gains and guide the design of optimal samplers.

2. Revisiting Sampling Methods in Diffusion Models

2.1. General Reverse SDE

Following the notation introduced in Ni et al. (2025), we describe the noise corruption process in the diffusion model as:

$$\boldsymbol{x}_t = \mu_t \boldsymbol{x}_0 + \sigma_t \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_N), \tag{1}$$

where $\boldsymbol{x}_0, \boldsymbol{x}_t \in \mathbb{R}^N$ denote the clean and corrupted data at time t, respectively. The functions μ_t and σ_t define the noise schedule, with $t \in [0, T]$. This general framework can include DDPM and BFN. For DDPM, it satisfies the variance preserving (VP) assumption $\mu_t^2 + \sigma_t^2 = 1$, while for BFN on continuous data, we have $\sigma_t = \sqrt{\mu_t(1-\mu_t)}$.

This process corresponds to a linear stochastic differential equation (SDE) (Ni et al., 2025):

$$\mathrm{d}\boldsymbol{x}_t = \frac{\dot{\mu}_t}{\mu_t}\boldsymbol{x}_t \mathrm{d}t + \sqrt{2\sigma_t\dot{\sigma}_t - 2\sigma_t^2\frac{\dot{\mu}_t}{\mu_t}}\mathrm{d}w_t, \tag{2}$$

where $\mathrm{d}w_t$ denotes the standard Wiener process.

Interestingly, it is possible to design **a family of reverse processes** (Prop. 4.1 in Xue et al. (2024b)) that share the same marginal probability distributions as equation 2 :

$$\begin{aligned} d\boldsymbol{x}_t = &\left[\frac{\dot{\mu}_t}{\mu_t}\boldsymbol{x}_t - \frac{1+\beta(t)}{2}g^2(t)\nabla_x \log p_t(x)\right]\mathrm{d}t \\ &+ \sqrt{\beta(t)}g_t \mathrm{d}w_t, \end{aligned} \tag{3}$$

where $g_t = \sqrt{2}\sqrt{\sigma_t\dot{\sigma}_t - \sigma_t^2\frac{\dot{\mu}_t}{\mu_t}}$, $\beta(t)$ is any non-negative bounded function. As proved in Appendix E of Song et al. (2021b) and Proposition 4.2 in Xue et al. (2024a) respectively, the sampling processes of both DDPM and BFN can be interpreted as first-order discretizations of the reverse-time SDE in equation 3 with $\beta = 1$.

2.2. Discretized Sampling Strategy with Varying Correlation Hypotheses

To derive concrete sampling strategies, we begin by discretizing the reverse-time SDE in equation 3. Although different discretization schemes may lead to variations in the resulting sampling formula, these differences become negligible when the step size is sufficiently small. As this work focuses on improving sampling quality in the regime of a large number of discretization steps, we adopt a simple first-order discretization:

$$\begin{aligned} \boldsymbol{x}_{t-\Delta t} = &\frac{\mu_{t-\Delta t}}{\mu_t}\boldsymbol{x}_t + \frac{(1+\beta(t))g_t^2\Delta t}{2}\nabla_x \log p_t(\boldsymbol{x}_t) \\ &+ \sqrt{\beta(t)g_t^2\Delta t}\boldsymbol{\epsilon}, \end{aligned} \tag{4}$$

where $g_t^2\Delta t \approx 2\sigma_t^2\frac{\mu_{t-\Delta t}}{\mu_t} - 2\sigma_t\sigma_{t-\Delta t}$, with derivation given in Appendix A.1. In this formulation, the next state is sampled from a Gaussian distribution conditioned on the current state. The optimal mean of this distribution, i.e. the conditional expectation $\mathbb{E}[\boldsymbol{x}_{t-\Delta t}|\boldsymbol{x}_t]$ has an analytic form involving the score function. This expectation depends on the form of the conditional distribution $p(\boldsymbol{x}_{t-\Delta t}|\boldsymbol{x}_t, \boldsymbol{x}_0)$ which captures the correlation between different time steps. According to (Song et al., 2021a), a family of such conditional distributions parameterized by λ_t all satisfy equation 1:

$$p_\lambda(x_{t-\Delta t}|x_t, x_0) = N(\mu_{t-\Delta t}x_0 + \gamma_t\frac{x_t - \mu_t x_0}{\sigma_t}, \lambda_t^2 I), \tag{5}$$

where $\gamma_t^2 + \lambda_t^2 = \sigma_{t-\Delta t}^2$. Based on this, the conditional expectation is given by:

$$\mathbb{E}[\boldsymbol{x}_{t-\Delta t}|\boldsymbol{x}_t] = \frac{\mu_{t-\Delta t}}{\mu_t}\boldsymbol{x}_t + (\frac{\mu_{t-\Delta t}}{\mu_t}\sigma_t^2 - \gamma_t\sigma_t)\nabla_x \log p_t(\boldsymbol{x}_t), \tag{6}$$

with derivation given in Appendix A.2. By aligning the mean in equation 4 with this conditional expectation, we derive the corresponding variance as $2\sigma_t(\sigma_{t-\Delta t} - \gamma_t)$. Substituting this into the original expression, the discretized reverse SDE becomes:

$$\begin{aligned} \boldsymbol{x}_{t-\Delta t} = &\frac{\mu_{t-\Delta t}}{\mu_t}\boldsymbol{x}_t + (\frac{\mu_{t-\Delta t}}{\mu_t}\sigma_t - \gamma_t)\sigma_t\nabla_x \log p_t(\boldsymbol{x}_t) \\ &+ \sqrt{2\sigma_t(\sigma_{t-\Delta t} - \gamma_t)}\boldsymbol{\epsilon}. \end{aligned} \tag{7}$$

2.3. Canonical Sampling Methods and Beyond

In this subsection, we examine the landscape of feasible sampling strategies, each reflecting a different temporal correlation assumption. First of all, we identify two representative cases that together cover the most widely used sampling strategies in existing diffusion models:

- Markov forward process: When the forward diffusion process satisfies the Markov property, i.e., $p(\boldsymbol{x}_t|\boldsymbol{x}_{t-\Delta t}, \boldsymbol{x}_0) = p(\boldsymbol{x}_t|\boldsymbol{x}_{t-\Delta t})$, we prove in the Proposition A.1 and Proposition A.2 that this corresponds to the choice $\gamma_t = \frac{\mu_t\sigma_{t-\Delta t}^2}{\mu_{t-\Delta t}\sigma_t}$. Note that this result does not depend on the VP assumption and both DDPM and BFN sampling belong to this category.
- Deterministic reverse process: When the reverse process becomes deterministic, i.e. $p(\boldsymbol{x}_{t-\Delta t}|\boldsymbol{x}_t, \boldsymbol{x}_0)$ collapses to a point mass, this corresponds to setting $\gamma_t = \sigma_{t-\Delta t}$, as in DDIM sampling.

These two methods span a spectrum of correlation structures from fully stochastic to fully deterministic. However, it remains an open and compelling question whether alternative, potentially superior sampling strategies exist within this family. We now consider a less-explored but intuitively natural alternative that lies at the opposite end of the deterministic method:

- When the reverse process is conditional independent, that is $p(x_{n-1}|x_n, x_0) = p(x_{n-1}|x_0)$. This corresponds to setting $\gamma_t = 0$, which maximizes the variance in both equation 5 and equation 7. We refer to this setting as maximally stochastic sampling (StoMax).

Please note that the categorization here refers to how the mean is chosen during the iterative process, while allowing for multiple theoretically sound options for the variance. By default, we use the variance from the SDE sampling formulation in equation 7. However, alternative choices, such as those used in DDIM or Analytic-DPM (Bao et al., 2022), are also feasible. In the deterministic case, these methods all yield zero variance. In the Markovian case, the performance of different variance choices tends to converge as the number of discrete steps increases (Bao et al., 2022). In the StoMax setting, the variance can be interpreted as a form of temperature control (Ni et al., 2025), offering a trade-off between sample diversity and fidelity.

3. Experimental Results

In this section, we compare the performance of different sampling strategies on molecular generation tasks.

3.1. Settings

Datasets We conduct generation experiments on two commonly used datasets. The QM9 dataset (Ruddigkeit et al., 2012; Ramakrishnan et al., 2014) is a database containing approximately 134,000 organic small molecules, composed of C, H, O, N, and F atoms. Each molecule contains up to 9 heavy atoms. We adopt the same data split strategy as the baseline methods, with 100k molecules in the training set, 18k in the validation set, and 13k in the test set. The other dataset is GEOM-Drugs (Axelrod & Gomez-Bombarelli, 2022), which consists of drug-like molecules. It contains 430,000 molecules, with an average of 44 atoms per molecule and up to 181 atoms in the largest molecule, making it more challenging than QM9. The dataset is split in the same way as in previous works: randomly divided into training, validation, and test sets with a ratio of 8:1:1.

Baselines Our baselines include the classic 3D molecular generation algorithm EDM (Hoogeboom et al., 2022), which performs joint diffusion over discrete atom types and continuous molecular coordinates. GeoBFN (Song et al., 2023) improves the scheduling of discrete atom types and continuous coordinates using Bayesian Flow Networks. UniGEM (Feng et al., 2025) treats molecular generation as a two-stage process: a nucleation phase for generating molecular scaffolds and a growth phase for completing the molecule. In UniGEM, atom type prediction is decoupled and performed only during the growth phase. For coordinate generation, UniGEM can adopt either EDM or GeoBFN, and we refer to these variants as UniGEM(EDM) and UniGEM(GeoBFN), respectively.

Metric We follow the evaluation protocol of prior works (Hoogeboom et al., 2022), generating 10,000 molecules to assess atom stability, molecule stability, validity, and valid & unique (V*U). Consistent with EDM's strategy, covalent bonds are inferred from inter-atomic distances. Atom stability is the portion of atoms with correct valence; molecule stability is the portion of molecules whose atoms all satisfy valence rules. Validity measures the proportion of generated generated 3D structures convertible to valid SMILES via RDKit. The V*U metric calculates the proportion of unique samples among all valid molecules.

Table 1. Unconditional molecular generation results on QM9. For all diffusion-based models, the sampling steps are 1000. Metrics are calculated with 10000 samples generated from each model. Higher values indicate better performance. *: The GeoBFN model is re-trained and evaluated by ourselves, as the original paper did not release the pretrained model.

Models	Sampling	Atom sta(%)	Mol sta(%)	Valid(%)	V*U(%)
Data	-	99.0	95.2	97.7	97.7
EDM	Default	98.7	82.0	91.9	90.7
	StoMax	**98.9**	**87.9**	**94.5**	**92.1**
UniGEM (EDM)	Default	99.0	89.8	95.0	93.2
	StoMax	**99.6**	**96.1**	**98.1**	**93.7**
GeoBFN	Default*	99.3	93.0	96.5	**92.7**
	StoMax	**99.3**	**94.2**	**96.9**	91.9
UniGEM (BFN)	Default	99.3	93.7	97.3	**93.0**
	StoMax	**99.5**	**95.7**	**97.8**	91.3

Table 2. Comparison between the StoMax strategy and the default unconditional sampling for EDM and UniGEM on the GEOM-Drugs dataset. 10,000 molecules were sampled using 1,000 diffusion steps.

Models	Sampling	Atom sta(%)	Valid(%)
Data	-	86.5	99.9
EDM	Default	81.3	92.6
	StoMax	**86.2**	**99.7**
UniGEM(EDM)	Default	85.1	98.4
	StoMax	**89.5**	**99.9**

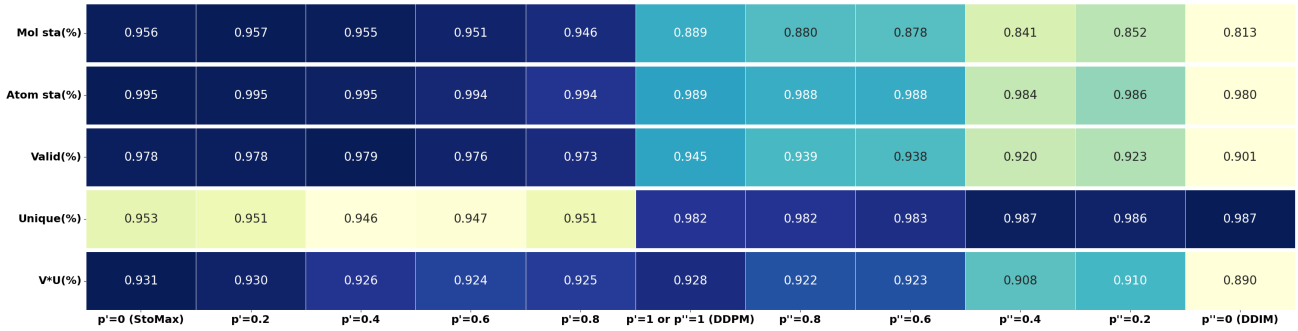

Figure 1. Unconditional molecular generation results on QM9. We evaluate interpolation-based sampling strategies across StoMax, DDPM, and DDIM using the UniGem (EDM) model. Higher values indicate better performance across all metrics.

3.2. Molecular Generation Evaluated on QM9 Dataset

We applied the StoMax strategy to EDM, GeoBFN, and two variants of UniGEM (UniGEM(EDM) and UniGEM(BFN)) and evaluated their unconditional generation performance on the QM9 dataset. The results are summarized in Table 1. Here default sampling for EDM and GeoBFN refers to the sampling method of DDPM and BFN, respectively. StoMax consistently improves quality-related metrics, atom and molecule stability, and validity across all models, demonstrating its general effectiveness in enhancing sample quality. The gains are especially pronounced for models based on EDM. A slight drop in diversity is observed, particularly for GeoBFN and UniGEM(GeoBFN), suggesting a trade-off between stability and uniqueness.

3.3. Molecular Generation Evaluated on GEOM-Drugs

We evaluate StoMax on the larger and more structurally diverse GEOM-Drugs dataset. As shown in Table 2, we apply StoMax to EDM and UniGEM(EDM) and compare the results of unconditional generation with their respective default samplers. StoMax yields notable gains in sample quality. Atom stability improves by around 5 points for EDM and over 4 points for UniGEM. In addition, the validity metric was improved to near-perfect levels (close to 1.0) in both models, clearly highlighting the advantage of StoMax in generating high-quality molecular samples.

3.4. Interpolation

To investigate whether better sampling methods exist within the explored design space, we follow (Song et al., 2021a) and conduct interpolation experiments between different sampling strategies. Notably, the deterministic and StoMax approaches represent two extremes of the sampling family in terms of stochasticity, with the sampling method based on a Markovian forward process lying between them. We thus interpolate between StoMax and DDPM using $\gamma_t = p'\gamma_t^{\text{DDPM}} + (1-p')\gamma_t^{\text{StoMax}} = p'\frac{\mu_t\sigma_{t-\Delta t}^2}{\mu_{t-\Delta t}\sigma_t}$, and between DDPM and DDIM via $\gamma_t = p''\gamma_t^{\text{DDPM}} + (1-p'')\gamma_t^{\text{DDIM}} = p''\frac{\mu_t\sigma_{t-\Delta t}^2}{\mu_{t-\Delta t}\sigma_t} + (1-p'')\sigma_{t-\Delta t}$, where p' and p'' are interpolation coefficients in $[0, 1]$. We evaluate the resulting sampling strategies using the UniGEM(EDM) model.

As shown in Figure 1, increasing the level of stochasticity consistently enhances the stability and validity of generated molecules, though it slightly compromises uniqueness. Notably, the StoMax strategy achieves the highest score on the U×V metric, demonstrating a favorable trade-off between diversity and generation quality. Among the sampling methods considered, StoMax emerges as the most balanced and effective approach.

4. Discussion and Future Work

Our findings raise compelling questions about the theoretical foundations of optimal sampling in diffusion-based generative models. While Appendix C in (Xue et al., 2024b) suggest that the optimal sampling strategy that minimizes the ELBO corresponds to $\beta(t) = 1$ in equation 3, which aligns with the default samplers used in DDPM and BFN, our empirical results demonstrate that, at least for molecular generation, alternative strategies such as StoMax can yield superior performance.

Despite the strong empirical performance of StoMax on molecular tasks, its rigorous theoretical explanation is still lacking. Interestingly, viewing the continuous formulation in equation 3, the StoMax strategy, characterized by maximal sampling variance, can be seen as approximating Langevin dynamics by taking the limit $\beta(t) \to \infty$, although our discrete implementation still differs from the commonly used Langevin dynamics used in generative models (Song & Ermon, 2019; Saremi & Hyvärinen, 2019). This observation raises the intriguing possibility that Langevin-like samplers may be inherently better suited for molecule generation. However, this remains a hypothesis that requires further systematic investigation.

Moreover, our experiments reveal a clear trade-off between diversity and validity across different sampling strategies. Developing a unified theoretical framework to characterize how different variance schedules affect this trade-off remains an open and promising direction for future work.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

Acknowledgements

This work is supported by Beijing Academy of Artificial Intelligence (BAAI).

References

Axelrod, S. and Gomez-Bombarelli, R. Geom, energy-annotated molecular conformations for property prediction and molecular generation. *Scientific Data*, 9(1):185, 2022.

Bao, F., Li, C., Zhu, J., and Zhang, B. Analytic-dpm: an analytic estimate of the optimal reverse variance in diffusion probabilistic models. *International conference on learning representations*, 2022.

Feng, S., Ni, Y., Lu, Y., Ma, Z.-M., Ma, W.-Y., and Lan, Y. Unigem: A unified approach to generation and property prediction for molecules. *International conference on learning representations*, 2025.

Graves, A., Srivastava, R. K., Atkinson, T., and Gomez, F. Bayesian flow networks. *arXiv preprint arXiv:2308.07037*, 2023.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Hoogeboom, E., Satorras, V. G., Vignac, C., and Welling, M. Equivariant diffusion for molecule generation in 3d. In *International conference on machine learning*, pp. 8867–8887. PMLR, 2022.

Ni, Y., Feng, S., Chi, H., Zheng, B., Gao, H.-a., Ma, W.-Y., Ma, Z.-M., and Lan, Y. Straight-line diffusion model for efficient 3d molecular generation. *arXiv preprint arXiv:2503.02918*, 2025.

Ramakrishnan, R., Dral, P. O., Rupp, M., and Von Lilienfeld, O. A. Quantum chemistry structures and properties of 134 kilo molecules. *Scientific data*, 1(1):1–7, 2014.

Ruddigkeit, L., Van Deursen, R., Blum, L. C., and Reymond, J.-L. Enumeration of 166 billion organic small molecules in the chemical universe database gdb-17. *Journal of chemical information and modeling*, 52(11):2864–2875, 2012.

Saremi, S. and Hyvärinen, A. Neural empirical bayes. *Journal of Machine Learning Research*, 20(181):1–23, 2019.

Song, J., Meng, C., and Ermon, S. Denoising diffusion implicit models. 2021a. URL https://openreview.net/forum?id=St1giarCHLP.

Song, Y. and Ermon, S. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021b. URL https://openreview.net/forum?id=PxTIG12RRHS.

Song, Y., Gong, J., Zhou, H., Zheng, M., Liu, J., and Ma, W.-Y. Unified generative modeling of 3d molecules with bayesian flow networks. In *The Twelfth International Conference on Learning Representations*, 2023.

Xue, K., Zhou, Y., Nie, S., Min, X., Zhang, X., ZHOU, J., and Li, C. Unifying bayesian flow networks and diffusion models through stochastic differential equations. In *Forty-first International Conference on Machine Learning*, 2024a. URL https://openreview.net/forum?id=1jHiq640y1.

Xue, S., Yi, M., Luo, W., Zhang, S., Sun, J., Li, Z., and Ma, Z.-M. Sa-solver: Stochastic adams solver for fast sampling of diffusion models. *Advances in Neural Information Processing Systems*, 36, 2024b.

A. Complementary Proofs

A.1. Discretizing the Reverse-Time SDE

To derive the discrete-time update rule for sampling, we begin by discretizing the reverse-time SDE in equation 3 using the Euler–Maruyama method. The resulting update step takes the following form:

$$\boldsymbol{x}_{t-\Delta t} = \boldsymbol{x}_t - \left[\frac{\dot{\mu}_t}{\mu_t}\boldsymbol{x}_t - \frac{(1+\beta(t))g_t^2}{2}\nabla_x \log p_t(\boldsymbol{x}_t)\right]\Delta t + \sqrt{\beta(t)g_t^2\Delta t}\boldsymbol{\epsilon}, \tag{8}$$

where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_N)$ denotes Gaussian noise.

To simplify the expression, we expand $\mu_{t-\Delta t}$ and $\sigma_{t-\Delta t}$ using a first-order Taylor approximation:

$$\mu_{t-\Delta t} = \mu_t - \dot{\mu}_t\Delta t + O((\Delta t)^2), \sigma_{t-\Delta t} = \sigma_t - \dot{\sigma}_t\Delta t + O((\Delta t)^2) \tag{9}$$

Recall that $g_t^2 = 2\left(\sigma_t\dot{\sigma}_t - \sigma_t^2\frac{\dot{\mu}_t}{\mu_t}\right)$, applying the above Taylor expansions, we approximate:

$$g_t^2\Delta t = 2\sigma_t\dot{\sigma}_t\Delta t - 2\sigma_t^2\frac{\dot{\mu}_t}{\mu_t}\Delta t \approx 2\sigma_t^2\frac{\mu_{t-\Delta t}}{\mu_t} - 2\sigma_t\sigma_{t-\Delta t}. \tag{10}$$

Substituting this into equation 8, and collecting terms, we obtain the final discrete-time update:

$$\boldsymbol{x}_{t-\Delta t} = \frac{\mu_{t-\Delta t}}{\mu_t}\boldsymbol{x}_t + \frac{(1+\beta(t))g_t^2\Delta t}{2}\nabla_x \log p_t(\boldsymbol{x}_t) + \sqrt{\beta(t)g_t^2\Delta t}\boldsymbol{\epsilon}, \tag{11}$$

where $g_t^2\Delta t \approx 2\sigma_t^2\frac{\mu_{t-\Delta t}}{\mu_t} - 2\sigma_t\sigma_{t-\Delta t}$. This formulation offers a general discrete-time sampling framework encompassing a broad family of strategies, each determined by specific choices of $\beta(t)$

A.2. Analytic Expectation of the Reverse Probability

We begin by connecting the score function to the conditional expectation of the original data:

$$\nabla_{\boldsymbol{x}_t}\log p_t(\boldsymbol{x}_t) = \frac{\int f(\boldsymbol{x}_0)\nabla_{\boldsymbol{x}_t}p(\boldsymbol{x}_t|\boldsymbol{x}_0)d\boldsymbol{x}_0}{p(\boldsymbol{x}_t)} = \frac{\int f(\boldsymbol{x}_0)p(\boldsymbol{x}_t|\boldsymbol{x}_0)(-\frac{\boldsymbol{x}_t-\mu_t\boldsymbol{x}_0}{\sigma_t^2})d\boldsymbol{x}_0}{p(\boldsymbol{x}_t)} = -\frac{\boldsymbol{x}_t}{\sigma_t^2} + \frac{\mu_t}{\sigma_t^2}\mathbb{E}[\boldsymbol{x}_0|\boldsymbol{x}_t], \tag{12}$$

which directly yields the Tweedie formula:

$$\mathbb{E}[\boldsymbol{x}_0|\boldsymbol{x}_t] = \frac{1}{\mu_t}\left(\boldsymbol{x}_t + \sigma_t^2\nabla_{\boldsymbol{x}_t}\log p_t(\boldsymbol{x}_t)\right) \tag{13}$$

To compute the expectation of the reverse sample $\boldsymbol{x}_{t-\Delta t}$ given $\boldsymbol{x}_t$, we write:

$$\begin{aligned}\mathbb{E}[\boldsymbol{x}_{t-\Delta t}|\boldsymbol{x}_t] &= \int \boldsymbol{x}_{t-\Delta t}p(\boldsymbol{x}_{t-\Delta t}|\boldsymbol{x}_t)\mathrm{d}\boldsymbol{x}_{t-\Delta t} = \int \boldsymbol{x}_{t-\Delta t}\int p(\boldsymbol{x}_{t-\Delta t}|\boldsymbol{x}_t,\boldsymbol{x}_0)p(\boldsymbol{x}_0|\boldsymbol{x}_t)\mathrm{d}\boldsymbol{x}_0\mathrm{d}\boldsymbol{x}_{t-\Delta t} \\ &= \int\left(\int \boldsymbol{x}_{t-\Delta t}p(\boldsymbol{x}_{t-\Delta t}|\boldsymbol{x}_t,\boldsymbol{x}_0)\mathrm{d}\boldsymbol{x}_{t-\Delta t}\right)p(\boldsymbol{x}_0|\boldsymbol{x}_t)\mathrm{d}\boldsymbol{x}_0\end{aligned} \tag{14}$$

Using equation 5, we know that: $\mathbb{E}[\boldsymbol{x}_{t-\Delta t}|\boldsymbol{x}_t,\boldsymbol{x}_0] = \mu_{t-\Delta t}x_0 + \gamma_t\frac{x_t-\mu_t x_0}{\sigma_t}$, where $\gamma_t = \sqrt{\sigma_{t-\Delta t}^2 - \lambda_t^2}$.

Thus equation 14 can be further reduced to

$$\mathbb{E}[\boldsymbol{x}_{t-\Delta t}|\boldsymbol{x}_t] = \mu_{t-\Delta t}\mathbb{E}[\boldsymbol{x}_0|\boldsymbol{x}_t] + \gamma_t\frac{x_t - \mu_t\mathbb{E}[\boldsymbol{x}_0|\boldsymbol{x}_t]}{\sigma_t} \tag{15}$$

Substituting $\mathbb{E}[\boldsymbol{x}_0|\boldsymbol{x}_t]$ from equation 12, we have:

$$\mathbb{E}_p[\boldsymbol{x}_{t-\Delta t}|\boldsymbol{x}_t] = \boldsymbol{x}_{t-\Delta t} = \frac{\mu_{t-\Delta t}}{\mu_t}\boldsymbol{x}_t + (\frac{\mu_{t-\Delta t}}{\mu_t}\sigma_t - \gamma_t)\sigma_t\nabla_x\log p_t(\boldsymbol{x}_t) \tag{16}$$

thus we proved equation 7.

A.3. Reverse-Time SDE with Analytic Expectation

By aligning the mean in equation 4 with equation 6, we obtain:

$$\begin{aligned}\frac{(1+\beta(t))g_t^2\Delta t}{2} &= (\frac{\mu_{t-\Delta t}}{\mu_t}\sigma_t - \gamma_t)\sigma_t \\ \beta(t)g_t^2\Delta t &= 2(\frac{\mu_{t-\Delta t}}{\mu_t}\sigma_t - \gamma_t)\sigma_t - g_t^2\Delta t \\ &\approx 2(\frac{\mu_{t-\Delta t}}{\mu_t}\sigma_t - \gamma_t)\sigma_t - (2\sigma_t^2\frac{\mu_{t-\Delta t}}{\mu_t} - 2\sigma_t\sigma_{t-\Delta t}) \\ &= 2\sigma_t(\sigma_{t-\Delta t} - \gamma_t)\end{aligned} \tag{17}$$

The approximation in the third line follows from equation 10. This result provides the sampling variance and confirms the expression in equation 7.

A.4. Sampling with Markov Forward Process Assumption

Proposition A.1. *Assume the forward diffusion process satisfies the Markov property:*

$$p(\boldsymbol{x}_t|\boldsymbol{x}_{t-\Delta t}, \boldsymbol{x}_0) = p(\boldsymbol{x}_t|\boldsymbol{x}_{t-\Delta t}). \tag{18}$$

We show that this leads to the variance parameter in equation 5 becomes:

$$\lambda_t = \frac{\sigma_{t-\Delta t}}{\sigma_t}\sqrt{\sigma_t^2 - \frac{\mu_t^2}{\mu_{t-\Delta t}^2}\sigma_{t-\Delta t}^2}. \tag{19}$$

Then $\gamma_t = \sqrt{\sigma_{t-\Delta t}^2 - \lambda_t^2} = \frac{\sigma_{t-\Delta t}^2\mu_t}{\sigma_t\mu_{t-\Delta t}}$.

Proof. Following the derivation approach in equation 5, the conditional distribution $p(\boldsymbol{x}_t|\boldsymbol{x}_{t-\Delta t}, \boldsymbol{x}_0)$ under general variance $\tilde{\lambda}_t$ takes the form:

$$p(\boldsymbol{x}_t|\boldsymbol{x}_{t-\Delta t}, \boldsymbol{x}_0) = \mathcal{N}\left(\mu_t\boldsymbol{x}_0 + \tilde{\gamma}_{t-\Delta t}\cdot\frac{\boldsymbol{x}_{t-\Delta t} - \mu_{t-\Delta t}\boldsymbol{x}_0}{\sigma_{t-\Delta t}}, \tilde{\lambda}_{t-\Delta t}^2\mathbf{I}\right), \tag{20}$$

where $\tilde{\gamma}_{t-\Delta t} = \sqrt{\sigma_t^2 - \tilde{\lambda}_{t-\Delta t}^2}$.

Now, consider the identity:

$$p(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-\Delta t}, \boldsymbol{x}_0)\, p(\boldsymbol{x}_{t-\Delta t} \mid \boldsymbol{x}_0) = p(\boldsymbol{x}_{t-\Delta t} \mid \boldsymbol{x}_t, \boldsymbol{x}_0)\, p(\boldsymbol{x}_t \mid \boldsymbol{x}_0), \tag{21}$$

and substitute the Gaussian expressions into both sides. By matching the exponents, we obtain the following equality:

$$\frac{\left\|\boldsymbol{x}_{t-\Delta t} - \mu_{t-\Delta t}\boldsymbol{x}_0 - \gamma_t\cdot\frac{\boldsymbol{x}_t - \mu_t\boldsymbol{x}_0}{\sigma_t}\right\|^2}{2\lambda_t^2} + \frac{\|\boldsymbol{x}_t - \mu_t\boldsymbol{x}_0\|^2}{2\sigma_t^2} \tag{22}$$

$$= \frac{\left\|\boldsymbol{x}_t - \mu_t\boldsymbol{x}_0 - \tilde{\gamma}_{t-\Delta t}\cdot\frac{\boldsymbol{x}_{t-\Delta t} - \mu_{t-\Delta t}\boldsymbol{x}_0}{\sigma_{t-\Delta t}}\right\|^2}{2\tilde{\lambda}_{t-\Delta t}^2} + \frac{\|\boldsymbol{x}_{t-\Delta t} - \mu_{t-\Delta t}\boldsymbol{x}_0\|^2}{2\sigma_{t-\Delta t}^2}. \tag{23}$$

Treating $\boldsymbol{x}_t - \mu_t\boldsymbol{x}_0$ and $\boldsymbol{x}_{t-\Delta t} - \mu_{t-\Delta t}\boldsymbol{x}_0$ as independent variables and matching the coefficients, we obtain the constraint:

$$\frac{\sigma_{t-\Delta t}^2}{\sigma_t^2} = \frac{\lambda_t^2}{\tilde{\lambda}_{t-\Delta t}^2}. \tag{24}$$

Under the Markov assumption, the conditional distribution $p(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-\Delta t}, \boldsymbol{x}_0)$ must be independent of $\boldsymbol{x}_0$. Therefore, from equation 20, the mean term must satisfy:

$$\mu_t - \tilde{\gamma}_{t-\Delta t}\cdot\frac{\mu_{t-\Delta t}}{\sigma_{t-\Delta t}} = 0, \tag{25}$$

which implies: $\tilde{\gamma}_{t-\Delta t} = \frac{\mu_t \sigma_{t-\Delta t}}{\mu_{t-\Delta t}}$.

Substituting into the definition of $\tilde{\lambda}_{t-\Delta t}$:

$$\tilde{\lambda}^2_{t-\Delta t} = \sigma_t^2 - \tilde{\gamma}^2_{t-\Delta t} = \sigma_t^2 - \left(\frac{\mu_t \sigma_{t-\Delta t}}{\mu_{t-\Delta t}}\right)^2 = \sigma_t^2 - \frac{\mu_t^2}{\mu^2_{t-\Delta t}} \sigma^2_{t-\Delta t}. \tag{26}$$

Finally, using the earlier ratio between λ_t and $\tilde{\lambda}_{t-\Delta t}$, we obtain:

$$\lambda_t^2 = \tilde{\lambda}^2_{t-\Delta t} \cdot \frac{\sigma^2_{t-\Delta t}}{\sigma_t^2} = \frac{\sigma^2_{t-\Delta t}}{\sigma_t^2} \left(\sigma_t^2 - \frac{\mu_t^2}{\mu^2_{t-\Delta t}} \sigma^2_{t-\Delta t}\right). \tag{27}$$

□

Note that our derivation holds for any noise schedule $\{\mu_t, \sigma_t\}$, without assuming VP or any particular form of μ_t and σ_t.

Proposition A.2. *We verify that the sampling formulation of BFN for continuous data satisfies the expression for λ_t derived under the assumption of a Markov forward process in Proposition A.1.*

Proof. The original BFN sampling procedure for continuous data, as proposed in (Graves et al., 2023), involves multiple iterative parameters. By deriving a closed-form expression, the sampling update can be simplified into the following compact form (see (Xue et al., 2024a) for detailed derivation):

$$\boldsymbol{x}_{t-\Delta t} = \frac{\mu_{t-\Delta t}}{\mu_t} \boldsymbol{x}_t + \frac{\mu_t - \mu_{t-\Delta t}}{\sqrt{\mu_t(1-\mu_t)}} \boldsymbol{\epsilon}_\theta + \sqrt{\frac{1-\mu_{t-\Delta t}}{1-\mu_t}(\mu_{t-\Delta t} - \mu_t)}\boldsymbol{\epsilon}. \tag{28}$$

where $\boldsymbol{\epsilon}_\theta$ is the predicted noise and $\boldsymbol{\epsilon}$ is standard Gaussian noise.

To establish its correspondence with the reverse SDE derived under the Markov assumption, it suffices to verify that the mean term in equation 28 matches that in equation 7, with the parameter $\gamma_t = \frac{\sigma^2_{t-\Delta t}\mu_t}{\sigma_t \mu_{t-\Delta t}}$ in the Markov forward process case:

$$(\frac{\mu_{t-\Delta t}}{\mu_t}\sigma_t - \gamma_t)\sigma_t \boldsymbol{s}_\theta = -(\frac{\mu_{t-\Delta t}}{\mu_t}\sigma_t - \frac{\sigma^2_{t-\Delta t}\mu_t}{\sigma_t \mu_{t-\Delta t}})\boldsymbol{\epsilon}_\theta = \frac{\mu_t - \mu_{t-\Delta t}}{\sqrt{\mu_t(1-\mu_t)}}\boldsymbol{\epsilon}_\theta, \tag{29}$$

where $\sigma_t^2 = \mu_t(1-\mu_t)$ for any t in BFN schedule and $\boldsymbol{s}_\theta$ denotes the score function predictor, which is related to the noise predictor via $\boldsymbol{s}_\theta(\boldsymbol{x}_t, t) = -\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t)/\sigma_t$. □