SYNTHETIC ASSET PRICE PATHS GENERATION USING DENOISING DIFFUSION PROBABILISTIC MODEL

Shujie Liu
Computational Mathematics
University of Waterloo
Waterloo, Ontario
s477liu@uwaterloo.ca

Justin W.L. Wan
David R. Cheriton School of Computer Science
University of Waterloo
Waterloo, Ontario
justin.wan@uwaterloo.ca

ABSTRACT

Synthetic asset price paths are crucial in various financial applications, such as risk management and strategy testing [2, 10, 17]. In this paper, we present the application of a Denoising Diffusion Probabilistic Model (DDPM) to generate synthetic asset price paths. The objective is not to exactly replicate the original paths, but rather to capture their underlying dynamics, thereby creating plausible yet unseen scenarios. Our methodology innovatively incorporates a discrete cosine transform, which allows the DDPM to learn in the frequency domain, substantially improving its learning efficacy. Unlike traditional approaches which involve predefined assumptions about the original paths' price dynamics, our approach avoids explicit model selection and calibration. Through both qualitative and quantitative assessments, we show that the synthetic paths generated by our DDPM closely align with the dynamics of the original paths, thereby affirming the effectiveness of our approach.

Keywords Asset price generation · deep generative models · denoising diffusion probabilistic model

1 Introduction

Traditional methods for creating synthetic asset prices typically involve selecting a model to describe price dynamics, calibrating its parameters, and then using it to generate new data. This approach, however, comes with considerable challenges. Designing a model that accurately reflect the complex nature of asset price movements is difficult. Additionally, the calibration process is prone to inaccuracies, especially when calibration data is limited.

The recent advancement of deep generative models [7, 11, 13, 14, 19] has opened new possibilities for producing synthetic asset prices. These models, based on neural networks, offer a 'model-free' approach as they do not rely on predefined assumptions about the dynamics of the studied time series. Notable successes have been achieved with generative adversarial networks (GANs) in time-series applications [6, 20]. However, GAN-based methods can be difficult to train and may face stability and model collapse issues.

In this paper, we explore the Denoising Diffusion Probabilistic Model (DDPM) [9, 16, 18], which has been shown to outperform GANs in image synthesis. Here, we focus on applying DDPM to synthesize asset price data. Meanwhile, we introduce a novel preprocessing step involving discrete cosine transform, which allows the DDPM to learn in the frequency domain and significantly enhances its ability to capture the dynamics of asset prices.

2 Denoising Diffusion Probabilistic Model

At its core, DDPM operates by incrementally introducing noise into a set of original data during a forward process and subsequently learning to reverse this noise addition in a backward process. This enables the generation of new data from Gaussian noise, that aligns with the same probability distribution as the original data [4].

In the forward process of DDPM, we assume the observed data points $\mathbf{x}_0$ are sampled from a probability distribution q. This process employs a Markov chain [5] to incrementally add Gaussian noise $\mathcal{N}(\mathbf{0}, \mathbf{I})$ to $\mathbf{x}_0$, following a variance

schedule [15] $\beta_1, ..., \beta_T$, to generate a sequence $\mathbf{x}_1, \ldots, \mathbf{x}_T$. The transition probability is given by:

$$q(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}\left(\sqrt{\bar{\alpha}}\mathbf{x}_0, (1-\bar{\alpha})\mathbf{I}\right),$$

where $\alpha_t = 1 - \beta_t,\ \bar{\alpha} = \prod_{s=1}^{t} \alpha_s$. This closed-form representation enables direct sampling of $\mathbf{x}_t$ from $\mathbf{x}_0$.

The reverse process attempts to counteract the noise addition at each timestep of the Markov chain. It is modeled by a neural network parameterized by θ, which takes noisy inputs $\mathbf{x}_t$ and timestep t, and learns to output approximated Gaussian distributions

$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) \sim \mathcal{N}(\mu_\theta(\mathbf{x}_t, t), \mathbf{\Sigma}_\theta(\mathbf{x}_t, t)),$$

effectively reversing the noise added in the transition from $\mathbf{x}_{t-1}$ to $\mathbf{x}_t$. For a well-trained network, the reconstructed distribution $p_\theta(\mathbf{x}_0)$ from the reverse process should match the original data distribution $q(\mathbf{x}_0)$. By initializing noisy inputs $\hat{\mathbf{x}}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, we can generate synthetic samples $\hat{\mathbf{x}}_0$ from Gaussian noise.

Upon the successful training of the reverse process, the sampling procedure is straightforward. We start with a sample drawn from $\mathcal{N}(\mathbf{0}, \mathbf{I})$, representing the noisy data $\hat{\mathbf{x}}_T$ at the final timestep. Then, given $\hat{\mathbf{x}}_t$, we can obtain $\hat{\mathbf{x}}_{t-1}$ by applying the learned reverse process as follows:

$$\hat{\mathbf{x}}_{t-1} = \mu_\theta(\hat{\mathbf{x}}_t, t)\hat{\mathbf{x}}_t + \sigma_t\hat{\mathbf{z}}_t,$$

where $\hat{\mathbf{z}}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. This process sequentially 'denoises' the data by reversing the Markov chain from $t = T$ back to $t = 0$. The final denoised output $\hat{\mathbf{x}}_0$ is the synthetic data generated by the DDPM.

3 Synthetic Asset Price Paths Generation

Consider an asset price path $\{S_t\}_{t=0}^{N-1}$ with length N and initial price S_0. Our proposed method is to generate a synthetic asset price path $\{\hat{S}_t\}_{t=0}^{N-1}$ that replicates the original path's stochastic dynamics using the DDPM.

3.1 Preprocessing of Training Data

We test our model on two commonly used stochastic models for asset price dynamics: 1) Geometric Brownian Motion (GBM) [3], and 2) the Heston model [8, 12]. For each stochastic process and each set of defining parameters, we simulate M paths, $\{S_t^i\}_{t=0}^{N-1}$, $i = 1, \ldots, M$, as our training set.

If we train DDPM on asset price paths directly, the resulting generated paths exhibit unnatural boundary spikes and fail to mimic the original volatility dynamics due to the randomness of asset prices. Inspired by frequency transform in signal processing, we tackle the issue by applying the discrete cosine transform (DCT) [1] to preprocess the raw asset price data. Transforming the raw asset price paths into the frequency domain through DCT allows us to learn the frequency patterns, the "signature" of the stochastic time-domain data. Specifically, the DDPM is able to learn the low-frequency components representing the core patterns of the data, while ignoring the high-frequency components which often correspond to values close to 0.

Note that the time series representing asset price paths in our study are not inherently periodic, which is often assumed when applying DCT. To address this, we apply another idea, mirror reflection, which extends the time series by appending the time-reversed paths to the end of the original. This technique ensures the extended paths do not introduce abrupt transition, while achieving the desired periodicity with a period length of $2N$ and the condition $S_0 = S_{2N-1}$, thereby making them suitable for DCT analysis.

3.2 Training

Training a DDPM entails a forward process where noise is systematically introduced and a reverse process where a neural network learns to revert this noise addition. For this reverse process, we employ a U-Net architecture [9] which processes inputs at various noise levels and their corresponding timesteps, and it learns to predict the noise added at each forward step. Other model architectural configurations and hyperparameter choices for the DDPM are determined by a thorough grid search.

3.3 Sampling & Post-processing

Using the trained DDPM, we can generate synthetic signals which we refer to as the "generated data". Since we performed DCT on the training set, we apply IDCT to the generated data. Also, recall that the asset price paths in the training data are extended with mirror reflection, the time-domain representation of the generated data is actually symmetrical. We select the "second half" of the generated paths as the final generated asset price paths.

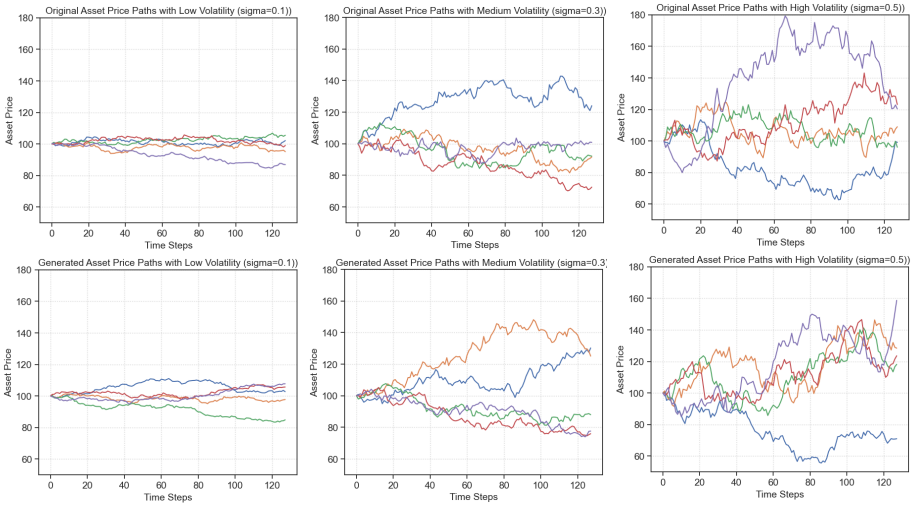

Figure 1: GBM - original paths (top) vs. generated paths (bottom). The different colours represent different price paths and there is no correspondence between paths with the same colour.

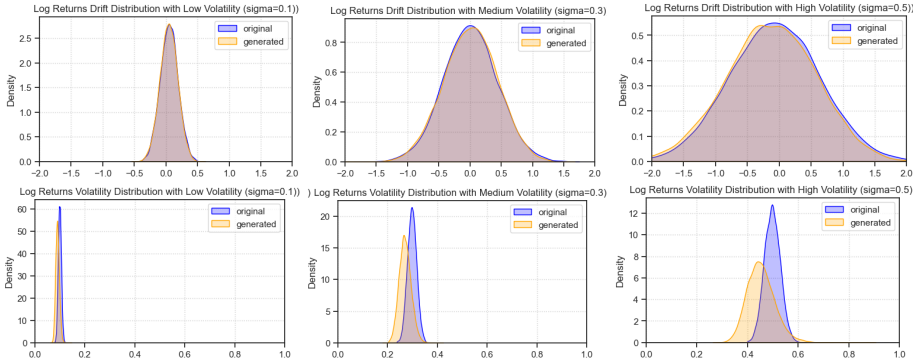

Figure 2: GBM - log returns drift (top) and volatility (bottom) distributions.

4 Results

We first test the DDPM's performance in generating asset price paths when the original paths' dynamics are characterized by GBM. To do this, we consider training sets simulated to follow GBM at three volatility scenarios: low volatility (σ = 10%), medium volatility (σ = 30%), and high volatility (σ = 50%). In all cases, we fix the initial price $S_0 = 100$, time to maturity $T = 0.5$, and the drift rate $\mu = 0.05$.

In Figure 1, for each volatility scenario, we illustrate five randomly selected paths from the training set and five randomly generated paths from the trained DDPM. A visual comparison indicates that across all three scenarios, the generated paths closely mirror the dynamics of the original paths by reflecting similar ranges of volatility, with no apparent issues at the left or right boundaries.

Next, we qualitatively compare the distributions of drift and volatility between the generated and original paths' log returns for each volatility scenario; see Figure 2. The drift distributions exhibit a high degree of congruence. The volatility distributions reveal underestimation in the generated paths compared to the original. Despite these discrepancies, the general resemblance between the shapes of the distributions indicates that the DDPM effectively captures the inherent randomness present in the original paths. Rather than generating a uniform set of paths characterized by identical drift and volatility, the DDPM demonstrates its capability to produce a diverse array of paths. Each path exhibits unique variations, yet collectively, they closely represent the underlying dynamics.

We then analyze DDPM's capability in synthesizing asset price paths that capture the Heston model dynamics, which introduces additional complexity by incorporating stochastic volatility. In alignment with our GBM setup, we construct training sets representing three different volatility scenarios by altering the long-term variance θ while holding other

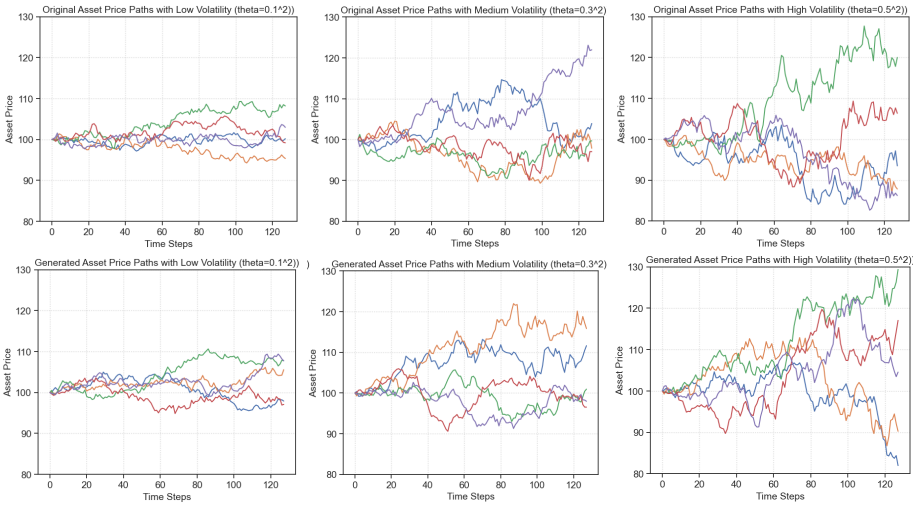

Figure 3: Heston model - original paths (top) vs. generated paths (bottom). The different colours represent different price paths and there is no correspondence between paths with the same colour.

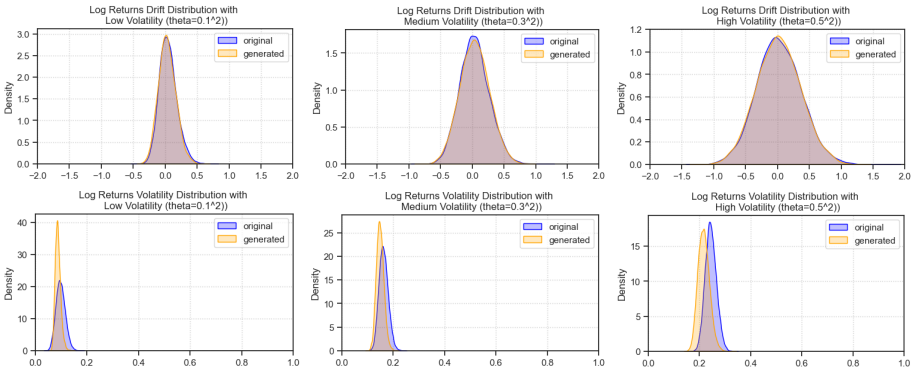

Figure 4: Heston model - log returns drift (top) and volatility (bottom) distributions.

parameters constant. The θ values considered are 0.1^2, 0.3^2, and 0.5^2, corresponding to low, medium, and high volatility scenarios respectively. Note that θ is a variance measure, hence $\sqrt{\theta}$ provides a volatility measure analogous σ in GBM.

We begin with a visual comparison of the original and generated paths for each volatility scenario (see Figure 3). We can see that the original paths in the medium and high volatility scenarios ($\theta = 0.3^2$ and $\theta = 0.5^2$) demonstrate increased volatility level through time. The generated paths mirror this dynamics effectively, with the paths at later timesteps displaying a higher degree of variation.

Figure 4 shows a visual comparison between original and generated paths' log returns drift and volatility distributions. The drift distributions between the original and generated paths show no significant deviations. In line with our findings from the GBM scenarios, the volatility distributions indicate a slight underestimation in the generated paths' volatility. Moreover, when $\theta = 0.1^2$, the generated paths exhibit a volatility distribution with lighter tails compared to those in the original paths. Despite these discrepancies, the DDPM overall successfully generates paths with a range of individual drifts and volatilities, collectively forming distributions that closely mirror those of the original paths.

5 Conclusions

In this paper, we have introduced a novel method for generating synthetic financial asset price paths using the DDPM. Our approach uniquely incorporates a DCT in preprocessing the training data. This technique shifts the learning process from the time domain into the frequency domain, significantly enhancing the DDPM's performance. Our method can generate synthetic paths that resemble the dynamics of the original paths, eliminating the need for explicit assumptions about the model form of the original paths' dynamics and avoiding the traditional calibration process.

Acknowledgments

This work is partially supported by Natural Sciences and Engineering Research Council of Canada (NSERC).

References

[1] N. Ahmed, T. Natarajan, and K.R. Rao. Discrete cosine transform. *IEEE Transactions on Computers*, C-23(1):90–93, 1974.

[2] Samuel A. Assefa, Danial Dervovic, Mahmoud Mahfouz, Robert E. Tillman, Prashant Reddy, and Manuela Veloso. Generating synthetic data in finance: Opportunities, challenges and pitfalls. In *Proceedings of the First ACM International Conference on AI in Finance*, ICAIF '20, New York, NY, USA, 2021. Association for Computing Machinery.

[3] Fischer Black and Myron Scholes. The pricing of options and corporate liabilities. *Journal of Political Economy*, 81(3):637–654, 1973.

[4] Sam Bond-Taylor, Adam Leach, Yang Long, and Chris G. Willcocks. Deep generative modelling: A comparative review of VAEs, GANs, normalizing flows, energy-based and autoregressive models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11):7327–7347, November 2022.

[5] Ka Chan, C. Lenard, and Terence Mills. An introduction to Markov chains. In *Proceedings of the MAV 49th Annual Conference*, La Trobe University, Bundoora, VIC, Australia, 12 2012.

[6] Federico Gatta, Fabio Giampaolo, Edoardo Prezioso, Gang Mei, Salvatore Cuomo, and Francesco Piccialli. Neural networks generative models for time series. *Journal of King Saud University - Computer and Information Sciences*, 34(10, Part A):7920–7939, 2022.

[7] Jie Gui, Zhenan Sun, Yonggang Wen, Dacheng Tao, and Jieping Ye. A review on generative adversarial networks: Algorithms, theory, and applications. *IEEE Transactions on Knowledge and Data Engineering*, 35(4):3313–3332, 2023.

[8] Steven L. Heston. A closed-form solution for options with stochastic volatility with applications to bond and currency options. *The Review of Financial Studies*, 6(2):327–343, 1993.

[9] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS 2020)*, Vancouver, Canada, 2020.

[10] James Hurst, Kirill Mayorov, and Joseph Francois Tagne Tatsinkou. The generation of synthetic data for risk modelling. *Journal of Risk Management in Financial Institutions*, 15(3):260–269, 2022.

[11] Abdul Jabbar, Xi Li, and Bourahla Omar. A survey on generative adversarial networks: Variants, applications, and training. *ACM Comput. Surv.*, 54(8), oct 2021.

[12] Herb Johnson and David Shanno. Option pricing when the variance is changing. *The Journal of Financial and Quantitative Analysis*, 22(2):143–151, 1987.

[13] Diederik P. Kingma and Max Welling. An introduction to variational autoencoders. *Foundations and Trends® in Machine Learning*, 12(4):307–392, 2019.

[14] Pengzhi Li, Yan Pei, and Jianqiang Li. A comprehensive survey on design and application of autoencoder in deep learning. *Applied Soft Computing*, 138:110176, 2023.

[15] Alex Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models, 2021. arXiv preprint arXiv:2102.09672.

[16] Jonas Oppenlaender. The creativity of text-to-image generation. In *Proceedings of the 25th International Academic Mindtrek Conference*, Academic Mindtrek '22, page 192–202, New York, NY, USA, 2022. Association for Computing Machinery.

[17] B.K. Pagnoncelli, D. Ramírez, H. Rahimian, et al. A synthetic data-plus-features driven approach for portfolio optimization. *Computational Economics*, 62:187–204, 2023.

[18] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[19] Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Wentao Zhang, Bin Cui, and Ming-Hsuan Yang. Diffusion models: A comprehensive survey of methods and applications. *ACM Comput. Surv.*, 56(4), nov 2023.

[20] Jinsung Yoon, William Zame, and Mihaela Schaar. Estimating missing data in temporal data streams using multi-directional recurrent neural networks. *IEEE Transactions on Biomedical Engineering*, PP, 11 2017.