Using Denoising Diffusion Models for Portfolio Optimization and Synthetic Data Generation

Ritvik Varada
AlgoGators
University of Florida
ritvik.varada@ufl.edu

ABSTRACT

The advent of advanced machine learning models has revolutionized data-driven approaches in finance. This paper explores the application of Denoising Diffusion Probabilistic Models (DDPMs) in financial markets, specifically for generating synthetic data and optimizing portfolio weights. By leveraging the temporal adaptability of AutoRegressive Diffusion Models (ARDPMs), this research demonstrates how diffusion models can generate synthetic time series data that mirrors real market dynamics and predicts future returns distributions for robust portfolio construction. The results show promise in enhancing risk-adjusted returns while maintaining a market-neutral stance, offering a compelling case for integrating diffusion models into modern investment strategies.

1. INTRODUCTION

In financial modeling, predictive accuracy and robust portfolio optimization are critical for consistent performance in volatile markets. Traditional machine learning methods such as Long Short-Term Memory (LSTM) networks and transformers have been employed for financial time series forecasting, but these models often fall short in capturing the probabilistic nature of return distributions.

Denoising Diffusion Probabilistic Models (DDPMs) have emerged as a powerful alternative for data

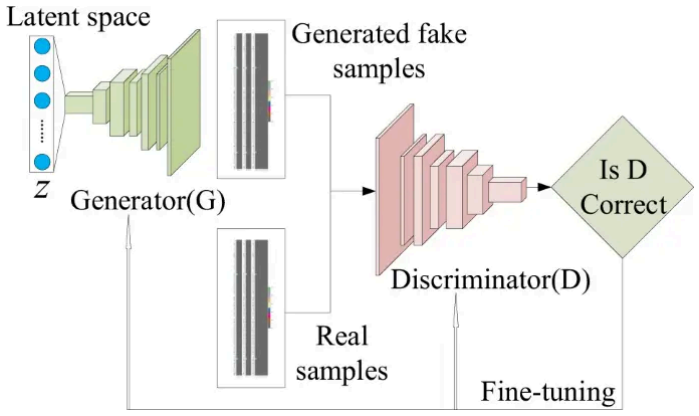
generation and probabilistic forecasting. Originally developed for high-quality image synthesis, DDPMs operate through a two-step Markov chain process: a forward process that gradually adds noise to data and a reverse process that denoises it to recover the original distribution. By extending these models with autoregressive conditioning, it becomes possible to model temporal dependencies in financial data and predict future returns. This research explores how these innovations can be applied to generate synthetic time series data and construct optimized portfolios using forecasted return distributions.

2. LITERATURE REVIEW

Generative models such as Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs) have been widely explored in finance for synthetic data generation and probabilistic modeling.

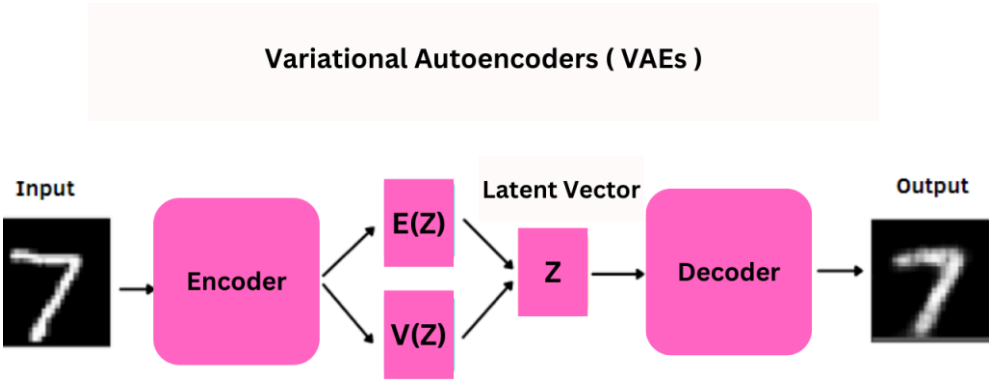
2.1 GANs

Generative Adversarial Networks (GANs), introduced by Goodfellow et al. (2014), have been widely used for generating synthetic data across various domains, including finance. GANs consist of two competing neural networks: a generator that creates synthetic data and a discriminator that evaluates the authenticity of the generated data. Through iterative training, the generator learns to produce increasingly realistic outputs.



2.2 VAEs

Variational Autoencoders (VAEs), introduced by Kingma and Welling (2013), are another popular generative modeling approach. VAEs use a probabilistic framework to map input data into a latent space, where a generative decoder reconstructs the original data. This process allows VAEs to model complex distributions while providing a probabilistic interpretation.



2.3 Diffusion Models vs. GANs and VAEs

Diffusion models, such as Denoising Diffusion Probabilistic Models (DDPMs), have emerged as a robust alternative to GANs and VAEs for synthetic data generation and probabilistic forecasting. Compared to GANs, diffusion models offer more stable training and avoid issues like mode collapse. They also provide a principled probabilistic framework, addressing the deterministic nature of GAN outputs. Similarly, diffusion models surpass VAEs in generating realistic and diverse outputs, overcoming the smoothing effect often seen in VAE-generated data.

In the context of financial applications, the probabilistic nature of diffusion models makes them

particularly suited for modeling return distributions and uncertainty. Furthermore, their ability to generate high-fidelity synthetic data provides a valuable resource for backtesting and stress testing investment strategies.

While GANs and VAEs have laid the groundwork for generative modeling in finance, diffusion models represent a significant advancement, offering enhanced flexibility and predictive accuracy. By leveraging the strengths of these generative paradigms, future research can further improve the integration of machine learning into quantitative finance.

3. Algorithmic Explanation

3.1 Forward Process (Noise Addition)

The forward process incrementally adds Gaussian noise to the data x_0 over T discrete time steps, resulting in a noisy version of the data x_T. This process can be thought of as a stochastic markov chain.

3.2 Reverse Process (Denoising)

The reverse process aims to recover x0x_0 from xTx_T by learning the conditional distributions and using a neural network to approximate the noise. The training objective minimizes the difference between the added noise and the predicted noise, typically using a re-weighted mean squared error (MSE). The following graphic demonstrates the reverse process.

3.3 Generation Process

To synthesize data, the model samples x_T from a Gaussian distribution and iteratively denoises it back to x_0 using the reverse process. The final output resembles the underlying data distribution. The following table demonstrates the algorithm fully:

Algorithm 1 Training

repeat
  $\mathbf{x}_0 \sim q(\mathbf{x}_0)$
  $t \sim \text{Uniform}(\{1, \dots, T\})$
  $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
  Take gradient descent step on
    $\nabla_\theta \left\| \boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, t) \right\|^2$
until converged

Algorithm 2 Sampling

$\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
for $t = T, \dots, 1$ **do**
  $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ if $t > 1$, else $\mathbf{z} = \mathbf{0}$
  $\mathbf{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}} \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t) \right) + \sigma_t \mathbf{z}$
end for
return $\mathbf{x}_0$

4. Novel Expansion ~ AutoRegressive DDPMs

ARDPMs excel in modeling sequential data by incorporating temporal context into the denoising process. Historical data, such as recent asset returns or prior time steps, is used as input to inform the generation of future predictions. This conditioning allows ARDPMs to learn and replicate the temporal dependencies in financial time series, capturing complex patterns like momentum, mean reversion, and volatility clustering.

4.1 Temporal Encoding

ARDPMs excel in modeling sequential data by incorporating temporal context into the denoising process. Historical data, such as recent asset returns or prior time steps, is used as input to inform the generation of future predictions. This conditioning allows ARDPMs to learn and replicate the temporal dependencies in financial time series, capturing complex patterns like momentum, mean reversion, and volatility clustering.

4.2 Probabilistic Forecasting

Unlike traditional time-series models that produce point estimates, ARDPMs generate a probability density function (PDF) for future data points. This probabilistic approach provides a comprehensive view of potential outcomes, including confidence intervals, enabling more informed decision-making in uncertain environments.

4.3 Iterative Denoising

ARDPMs follow the two-stage process of DDPMs, where data is first corrupted with noise in the forward process and then denoised step-by-step in the reverse process. The temporal conditioning integrates historical patterns at each step, allowing the model to recover structured data that reflects both the underlying distribution and sequential dependencies.

5. Applications in Finance

ARDPMs excel in modeling sequential data by incorporating temporal context into the denoising process. Historical data, such as recent asset returns or prior time steps, is used as input to inform the generation of future predictions. This conditioning allows ARDPMs to learn and replicate the temporal dependencies in financial time series, capturing complex patterns like momentum, mean reversion, and volatility clustering.

5.1 Time Series Forecasting

ARDPMs are particularly well-suited for forecasting asset returns or price movements in financial markets, providing predictions over both one-step (next time point) and multi-step horizons. By leveraging temporal dependencies through their autoregressive nature, these models capture intricate patterns such as momentum effects, volatility clustering, and mean reversion. Additionally, ARDPMs produce probabilistic forecasts in the form of probability density functions (PDFs), which quantify the uncertainty surrounding predicted values. This probabilistic approach equips financial professionals with a more comprehensive understanding of potential outcomes, enabling them to make robust decisions under uncertain conditions. For example, traders can use ARDPM forecasts to evaluate risk-reward trade-offs for specific strategies, while portfolio managers can plan for different market scenarios based on the range of predicted outcomes.

5.2 Portfolio Optimization

The probabilistic forecasts provided by ARDPMs play a crucial role in portfolio construction and optimization. By generating PDFs for expected returns, ARDPMs allow for precise weight allocation based on confidence levels in return predictions. Assets with higher confidence in positive returns can be given greater weights, while assets with uncertain or negative outlooks can have reduced exposure or be used for hedging purposes. This enables the construction of balanced portfolios that achieve targeted

risk-adjusted returns while controlling for downside risk. Moreover, ARDPMs facilitate the design of market-neutral strategies by offsetting long and short positions, minimizing systemic risk. By incorporating uncertainty directly into portfolio optimization, ARDPMs provide a significant advantage over traditional models that rely on point estimates.

5.3 Synthetic Data Generation

ARDPMs excel at generating realistic synthetic financial time series that retain the statistical and temporal properties of real market data. This synthetic data can serve multiple purposes in quantitative finance. For instance, it can be used to augment historical datasets, providing a richer foundation for training machine learning models or testing trading strategies. In scenarios with limited or incomplete historical data, synthetic data generated by ARDPMs can enable backtesting and stress testing of investment strategies under a wide range of simulated market conditions. Furthermore, synthetic data is invaluable for ensuring privacy, as it can replicate key features of real data without exposing sensitive information. By bridging data gaps and enabling extensive testing, ARDPMs enhance the robustness and reliability of financial models.

Conclusions

AutoRegressive Denoising Diffusion Probabilistic Models (ARDPMs) represent a groundbreaking advancement in applying machine learning to finance, combining the strengths of probabilistic modeling and temporal forecasting. Their ability to capture complex sequential dependencies and generate realistic probabilistic predictions makes them a versatile tool for a wide range of applications, including time-series forecasting, portfolio optimization, and synthetic data generation. By quantifying uncertainty, ARDPMs enable more robust decision-making, particularly in volatile and dynamic financial markets. Despite challenges such as computational complexity and sensitivity to regime shifts, ARDPMs offer a unique combination of adaptability, precision, and resilience that sets them apart from

traditional models. Future work focused on enhancing scalability, integrating adaptive mechanisms, and exploring hybrid architectures could unlock their full potential, positioning ARDPMs as a cornerstone of modern quantitative finance.

References

1. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial networks. *Communications of the ACM*, 63(11), 139-144.
2. Kingma, D. P., & Welling, M. (2013). Auto-encoding variational Bayes. *arXiv preprint arXiv:1312.6114*.
3. Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. *arXiv preprint arXiv:2006.11239*.
4. Tashiro, Y., Song, Y., Ermon, S., & Zahavy, T. (2021). CSDI: Conditional Score-Based Diffusion Models for Probabilistic Time Series Imputation. *arXiv preprint arXiv:2111.07903*.
5. Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, 31(3), 307-327.
6. Hull, J. C., & White, A. (1998). Incorporating volatility updating into the historical simulation method for value-at-risk. *Journal of Risk*, 1(1), 5-19.
7. Tsay, R. S. (2005). Analysis of financial time series. *Wiley-Interscience*.
8. Rasmussen, C. E., & Williams, C. K. I. (2006). Gaussian Processes for Machine Learning. *The MIT Press*.
9. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 5998-6008.