



**HAL**  
open science

# Time-changed normalizing flows for accurate SDE modeling

Naoufal El Bekri, Lucas Drumetz, Franck Vermet

► **To cite this version:**

Naoufal El Bekri, Lucas Drumetz, Franck Vermet. Time-changed normalizing flows for accurate SDE modeling. 2023. hal-04370186v1

**HAL Id: hal-04370186**

**<https://hal.science/hal-04370186v1>**

Preprint submitted on 2 Jan 2024 (v1), last revised 19 Jan 2024 (v2)

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License

# TIME CHANGED NORMALIZING FLOWS FOR ACCURATE SDE MODELING

Naoufal El Bekri<sup>1,2</sup>, Lucas Drumetz<sup>2</sup>, and Franck Vermet<sup>1</sup>

<sup>1</sup>Univ Brest, UMR CNRS 6205, Laboratoire de Mathématiques de Bretagne Atlantique, France

<sup>2</sup>IMT Atlantique, Lab-STICC, UMR CNRS 6285, Brest, France

## ABSTRACT

The generative paradigm has become increasingly important in machine learning and deep learning models. Among popular generative models are normalizing flows, which enable exact likelihood estimation by transforming a base distribution through diffeomorphic transformations. Extending the normalizing flow framework to handle time-indexed flows gave dynamic normalizing flows, a powerful tool to model time series, stochastic processes, and neural stochastic differential equations (SDEs). In this work, we propose a novel variant of dynamic normalizing flows, a Time Changed Normalizing Flow (TCNF), based on time deformation of a Brownian motion which constitutes a versatile and extensive family of Gaussian processes. This approach enables us to effectively model some SDEs, that cannot be modeled otherwise, including standard ones such as the well-known Ornstein-Uhlenbeck process, and generalizes prior methodologies, leading to improved results and better inference and prediction capability.

**Index Terms**— Stochastic differential equations, deep generative models, normalizing flows, time series, convex neural network.

## 1. INTRODUCTION

Dynamical systems have widespread use in various scientific areas such as finance, geosciences, and physics. The representation of these systems usually involves Ordinary Differential Equations (ODEs), or Stochastic Differential Equations (SDEs) [1] when noise and perturbations, on top of the deterministic component, are considered. Crucial applications include modeling volatility in financial data, or uncertainty quantification and propagation in geosciences. Tackling such systems through time series modeling has witnessed a surge in popularity within the machine learning paradigm, and more recently through the generative approach, for forecasting applications, but also for filtering or interpolation with a notion of uncertainty in the generated sequences.

Popular generative models include Generative Adversarial Networks (GANs) [2] and variational autoencoders (VAEs) [3], but also more recently normalizing flows (NFs) [4, 5] and diffusion/score based models [6]. Though these models can be applied to generate time series, they are not well-suited for the

task because they treat such data as vectors in  $\mathbb{R}^T$ , with  $T$  the number of time steps, without accounting for the causal structure. Adaptations of GANs, VAEs, and NFs to time series data have been carried out in [7, 8],[9, 10],[11, 12], respectively. In this work, we focus on NFs for their capacity to access explicit likelihoods, which is crucial for applications when uncertainty quantification or anomaly detection is required.

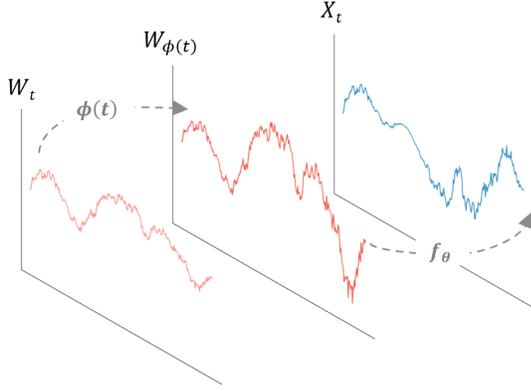
NFs are based on the celebrated change of variables formula that provides an expression of the probability density function of diffeomorphic transformations of a random variable. With carefully chosen transformations (or compositions thereof), if the initial density is tractable (explicit likelihood and easy sampling, in most cases Gaussian), the transformed density can be easily manipulated and sampled as well, provided that the Jacobian of the transformation can be computed efficiently. By considering an infinite number of transformations, representing the limit case of discrete NFs, we can build a Continuous Normalizing Flow (CNF) [13]. This can further improve the computational efficiency of this class of models as we replace the Jacobian by the integration of its trace.

NFs have been extended to the dynamic setting by replacing the tractable base distribution with a tractable *stochastic process*, i.e. a Brownian motion [14], making this type of model much more efficient for time series generation. However, it was noted in [15] that these models are theoretically unable to handle some of the most basic and common processes, such as the classical Ornstein-Uhlenbeck process.

Thus, in this paper, we propose a generalization of this class of approaches by using a large family of Gaussian processes as the underlying base process instead of the conventional standard Brownian motion. The Gaussian processes are constructed by transforming through time the standard Brownian motion, giving rise to Time Changed Normalizing Flow (TCNF), a model that possesses mathematical properties that enable it to describe dynamics and SDEs that cannot be captured by previous flow-based models, while maintaining the expressivity of dynamic NFs in other cases. We corroborate these findings by numerical experiments on several well-known processes.

The remainder of this paper is organized as follows: we first provide, in section 2, an overview of the neural SDEs, wherein both the drift and diffusion are neural networks, and the dynamic normalizing flow approach and a discussion on the

inherent limitations of such models. Afterward, in section 3, we introduce our model and describe its characteristics and the training algorithm. Finally, quantitative results are presented in section 4 and compared to other flow-based models, and provide concluding remarks in section 5.



**Fig. 1:** A time change  $\phi$  is applied to the Wiener process (red) to create a new Gaussian process, which is then mapped through a bijection  $f_\theta$  to the observed process (blue).

## 2. BACKGROUND

### 2.1. Neural Stochastic Differential Equations

We consider a filtered probability space  $(\Omega, \mathcal{F}, P)$  and a time horizon  $T$ . A diffusion process  $X = \{X_t\}_{t \in [0, T]}$  is defined by the Itô Stochastic Differential Equation (SDE):

$$dX_t = \mu(X_t, t)dt + \sigma(X_t, t)dW_t, t \in [0, T] \quad (1)$$

where  $W = \{W_t\}_{t \in [0, T]}$  is the  $m$ -dimensional adapted standard Wiener process (or Brownian motion). Functions  $\mu : \mathbb{R}^d \times [0, T] \rightarrow \mathbb{R}^d$  and  $\sigma : \mathbb{R}^d \times [0, T] \rightarrow \mathbb{R}^{d \times m}$  are the drift and diffusion coefficients, respectively. When  $\mu$  and  $\sigma$  are implemented by neural networks, the SDE is designated as a *neural SDE* [16, 17].

Several works have been proposed to learn neural SDEs using different generative modeling frameworks including Variational Autoencoders (VAEs) [9, 10] and Generative Adversarial Networks (GANs) [8]. In this paper, we specifically focus on the normalizing flow paradigm.

### 2.2. Normalizing flows

A normalizing flow [4, 5, 13] is a transformation designed to model a random variable  $X$  and its corresponding complex distribution  $p_X$  through a base distribution  $p_Z$  and a differentiable bijective function  $f : \mathbb{R}^d \rightarrow \mathbb{R}^d$ . Such modeling allows for both exact density estimation and efficient sampling,

by using the change of variable formula for  $X = f(Z)$ :

$$\log p_X(x) = \log p_Z(z) - \log \left| \det \left( \frac{\partial f}{\partial z}(z) \right) \right| \quad (2)$$

Previous works have extended this framework to model time series and stochastic processes by employing a bijection that is continuously indexed by time  $F(\cdot, t)$ , along with a Brownian motion as the base process, giving rise to a Continuous Time Flow Process (CTFP) [14]:

$$X_t = F(W_t, t).$$

Another approach proposed by [15] involves incorporating latent dynamics from an Ornstein-Uhlenbeck process combined with a normalizing flow to effectively model SDEs.

These models have demonstrated the effectiveness of dynamic NFs in capturing the complex behavior present in various types of stochastic processes and SDEs. However, it is important to highlight that these models have inherent limitations. One constraint arises when applying Itô's lemma to CTFP in order to derive the Ornstein-Uhlenbeck (OU) process described by the equation:

$$dY_t = -a(Y_t - b)dt + \sigma dW_t \quad (3)$$

Indeed, by applying Itô's lemma to  $F(W_t, t)$  we obtain:

$$\begin{aligned} dF(W_t, t) &= \frac{\partial F}{\partial t}(W_t, t)dt + \frac{\partial F}{\partial x}(W_t, t)dW_t \\ &+ \frac{1}{2} \frac{\partial^2 F}{\partial x^2}(W_t, t)dt \end{aligned} \quad (4)$$

By comparing both Eq. (3) and Eq. (4), we deduce that to model the OU process, we need to have  $\frac{\partial F}{\partial x}(W_t, t) = \sigma$ , implying that  $F(W_t, t) = \sigma W_t + g(t)$ , where  $g$  is a given differentiable function. However, differentiating this relation wrt  $t$  and plugging in (3) leads to the following condition:

$$\frac{dg}{dt}(t) + ag(t) - ab = -a\sigma W_t \quad (5)$$

Eq. (5) is not feasible as the left-hand side is a deterministic function of  $t$  while the right-hand side is a stochastic one depending on  $W_t$ . Thus, CTFP shows limitations and falls short in its ability to effectively model various stochastic processes.

In the following section, we propose a model that can address this limitation and achieve improved results.

## 3. DYNAMIC NORMALIZING FLOW WITH TIME CHANGE

### 3.1. Time changed Normalizing Flow

We propose to model an observed stochastic process, denoted as  $X = \{X_t\}_{t \in [0, T]}$ , by combining a normalizing flow and a time-changed Wiener process to accurately capture

the dynamic behavior of  $X_t$  based on a realized time series  $\{(x_{t_i}, t_i)\}_{i=1}^n$ . We introduce the concept of a time-changed normalizing flow (TCNF), defined as follows:

$$X_t = f_\theta(W_{\phi(t)}, \phi(t)), \quad \forall t \in [0, T],$$

where  $f_\theta(\cdot, t) : \mathbb{R}^d \rightarrow \mathbb{R}^d$  is a differentiable bijection parameterized by  $\theta$ , while  $W_{\phi(t)}$  denotes a Brownian motion with a time change [18, Sect. 1 Chap. 0]. The time change is given by  $\phi : \mathbb{R}^+ \rightarrow \mathbb{R}^+$ , which is a measurable, positive and increasing function. The measurable and positive properties ensure the correct definition of  $W_{\phi(t)}$ , while the increasing property ensures the existence of its moments. Consequently, the neural network modeling the time change function must have intrinsically positive and increasing attributes. The property of time change has important applications as it produces a family of Gaussian processes that are more general than the Brownian motion. The Dubins-Schwarz theorem [18, Theorem 5.1.6] further emphasizes this property as it states that every local martingale is simply a time-changed Brownian motion. Therefore, by making the base process of our model a time-changed Brownian motion, we can accurately capture all instances of local martingales and semimartingales, thus generalizing the CTFP setting. In fact, the solution to Eq. (3) can be expressed as follows:

$$Y_t = Y_0 e^{-at} + b(1 - e^{-at}) + \frac{\sigma e^{-at}}{\sqrt{2a}} W_{e^{2at} - 1} \quad (6)$$

which can be properly modeled by TCNF. More general cases like processes with time-dependent volatility can also be expressed via a time-change and therefore modeled by TCNF. Finally, for  $\phi(t) = t$  we recover the CTFP setting, which is suitable for modeling SDEs that do not require a time-change as the Geometric Brownian Motion (GBM) [1].

### 3.2. Time change function

To handle the time change function, we employ a convex neural network that ensures a positive gradient, thereby guaranteeing a monotone output. Specifically, we utilize the M-MGN architecture [19] based on  $K$  network modules defined as follows:

$$\begin{aligned} \tilde{t}_k &= W_k \times t + b_k, \\ \text{M-MGN}(t) &= a + V^\top V t + \sum_{k=1}^K s_k(\tilde{t}_k) \times W_k^\top \sigma_k(\tilde{t}_k) \end{aligned} \quad (7)$$

where  $W_k, b_k \in \mathbb{R}^{l \times 1}$  are respectively weight and bias vectors of the  $k^{\text{th}}$  layer,  $\sigma_k : \mathbb{R}^l \rightarrow \mathbb{R}^l$  is an activation function and  $s_k : \mathbb{R}^l \rightarrow \mathbb{R}$  its antiderivative.  $a \in \mathbb{R}$  and  $V \in \mathbb{R}^{l \times 1}$  are additional network parameters. As the result is not necessarily positive, we apply a translation of the M-MGN output to ensure that the time change is positive.

### 3.3. Training algorithm

The purpose is to train the TCNF in order to maximize the log-likelihood of the observed dataset  $\{(x_{t_i}, t_i)\}_{i=1}^n$ :

$$L = \log p_{X_{t_1}, \dots, X_{t_n}}(x_{t_1}, \dots, x_{t_n}) \quad (8)$$

To compute Eq.(8), we use the change of variable formula and leverage the independence of increments  $W_{\phi(t_i)} - W_{\phi(t_{i-1})}$ . Thus, the log-likelihood is expressed as follows:

$$\begin{aligned} L &= \sum_{i=1}^n \log p_{W_{\phi(t_i)} | W_{\phi(t_{i-1})}}(w_{\phi(t_i)}) \\ &\quad - \log \left| \det \frac{\partial f_\theta(w_{\phi(t_i)}, \phi(t_i)}{\partial W_{\phi(t_i)}} \right|, \end{aligned} \quad (9)$$

where  $w_{\phi(t_i)} = f_\theta^{-1}(x_{t_i}; \phi(t_i))$  and  $p_{W_{\phi(t_i)} | W_{\phi(t_{i-1})}}$  denotes the conditional Gaussian distribution with mean  $W_{\phi(t_{i-1})}$  and variance  $\phi(t_i) - \phi(t_{i-1})$ . This constitutes one notable difference from the log-likelihood of CTFP which uses a Gaussian distribution with the same mean but variance  $t_i - t_{i-1}$ .

## 4. EXPERIMENTS

### 4.1. Toy datasets

To evaluate the performance of our proposed model, we conducted experiments on three toy datasets comprising univariate time series. These datasets were generated by sampling from three different stochastic processes. Also, in our experiments we used an identical architecture as CTFP, leveraging CNFs.

The first dataset (**Toy-SDE1**) was generated by discretizing the OU process, given by the equation:  $dX_t = -\theta(X_t - \mu)dt + \sigma dW_t$ , where  $\mu$  and  $\sigma$  represent constant parameters for the drift and volatility terms, respectively. Parameter  $\theta$  captures the speed at which a given sample path of the process converges towards the drift term. This dataset aims to assess the model's ability to capture the dynamics of time changes.

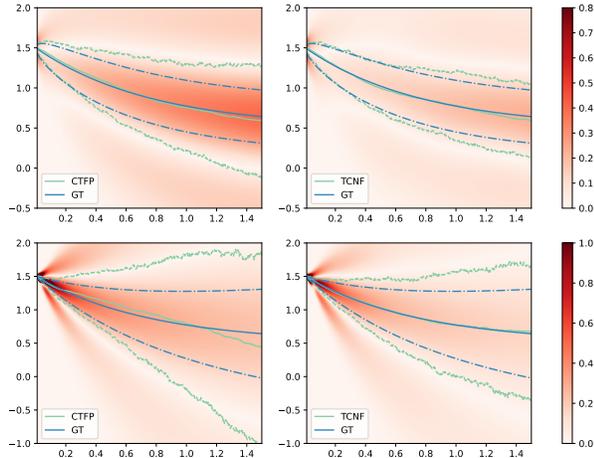
The second dataset (**Toy-SDE2**) was generated based on the equation:  $dX_t = -\theta(X_t - \mu)dt + \sigma\sqrt{t}dW_t$ , describing an OU with a time-dependent diffusion coefficient, and is used to test the model's capability to capture time transformations with increasing complexity. Notably, this SDE is of interest as it is commonly used in score-based models [20], where noise is gradually introduced during the training process.

Finally, The third dataset (**Toy-SDE3**) involved the geometric brownian motion, described by the equation:  $dX_t = \mu X_t dt + \sigma X_t dW_t$ , where  $\mu$  and  $\sigma$  represent constant parameters for the drift and volatility terms, respectively. This dataset was designed to showcase the capacity of TCNF to handle SDEs where no time change is required, effectively learning the simple function  $\phi(t) = t$ . This shows that our approach can encompass the CTFP framework.

The quantitative comparison is carried by comparing estimations of the mean  $m_{X_t}$ , the standard-deviation  $\sigma_{X_t}$ , the

Dataset	Model	$m_{X_t}$	$\sigma_{X_t}$	IQR	$p_{X_t}$
Toy-SDE1	CTFP	$0.022 \pm 0.009$	$0.241 \pm 0.012$	$0.886 \pm 0.021$	0.062
	TCNF	<b><math>0.019 \pm 0.007</math></b>	<b><math>0.109 \pm 0.009</math></b>	<b><math>0.700 \pm 0.014</math></b>	<b>0.033</b>
Toy-SDE2	CTFP	$0.059 \pm 0.011$	$0.523 \pm 0.018$	$1.517 \pm 0.037$	0.087
	TCNF	<b><math>0.024 \pm 0.012</math></b>	<b><math>0.322 \pm 0.016</math></b>	<b><math>1.183 \pm 0.024</math></b>	<b>0.061</b>
Toy-SDE3	CTFP	<b><math>0.035 \pm 0.020</math></b>	<b><math>0.086 \pm 0.056</math></b>	$1.103 \pm 0.037$	<b>0.005</b>
	TCNF	$0.036 \pm 0.018$	$0.099 \pm 0.072$	<b><math>1.083 \pm 0.038</math></b>	0.007

**Table 1:** Quantitative analysis: We display mean absolute errors (MAE) for estimating respectively the mean, standard deviation, interquartile range  $IQR = Q_3 - Q_1$ , and density. The estimations are based on 1000 sample paths over 1000 iterations.



**Fig. 2:** Comparison between TCNF and CTFP on toy-SDE1 (Top) and toy-SDE2 (Bottom). In each figure we depict density MAE between the flow-based models and the ground truth, along the mean (continuous line) and the IQR (dashed line).

interquartile range  $IQR = Q_3 - Q_1$ , and the density  $p_{X_t}$  respectively. For each model, we compute the mean absolute errors (MAE) against the ground-truth values. The mean, standard deviation, and quartiles are estimated based on 1000 sample paths over 1000 iterations, while the density is estimated by the change of variable formula on a grid consisting of 1000 spatial points and 500 temporal points within the time interval  $[0, T = 1.5]$ . The quantitative results reported in Table 1 and Figure 2 show first that TCNF shows no loss in generality as it handles cases where no time change is required, and second that our model exhibits superior estimation capability, as it can capture the behavior of time changed ground truth solutions.

#### 4.2. Real-world datasets

To further assess our model’s capacity to capture complex dynamics, we train it on two real-world datasets: Crypto-forecasting (**Crypto**) [21] and Electric Consumption Load (**ECL**) [22]. The Crypto dataset contains historical prices of

Dataset	Model	$m_{X_t}$	$\sigma_{X_t}$
Crypto	CTFP	$0.083 \pm 0.007$	$0.232 \pm 0.009$
	TCNF	<b><math>0.025 \pm 0.001</math></b>	<b><math>0.131 \pm 0.005</math></b>
ECL	CTFP	$0.771 \pm 0.108$	$26.972 \pm 8.634$
	TCNF	<b><math>0.299 \pm 0.007</math></b>	<b><math>2.182 \pm 2.081</math></b>

**Table 2:** Quantitative analysis of real world dataset: We display mean and standard deviation estimation errors.

various cryptocurrencies such as Bitcoin or Ethereum. We focused our analysis on modeling Ethereum returns over the 2020 period. The ECL dataset comprises electricity consumption data from multiple clients within a 15-minute interval. We choose to model the consumption of client ‘200’ as its data offers one of the most extended time series.

The results comprise mean absolute errors (MAE) for estimating the mean ( $m_{X_t}$ ) and standard deviation ( $\sigma_{X_t}$ ) for the Crypto dataset. For the ECL dataset, we employed mean relative errors (MRE) to scale the results appropriately. These results are reported in Table 2 and compared against CTFP.

## 5. CONCLUSION

We introduced a generalized approach to modeling stochastic processes and SDEs via dynamic NFs and time change. By transforming the Wiener process through time, we generate various Gaussian processes, which are then mapped to the observed process via a bijection. The time change combined with the bijection enabled us to model processes that are otherwise challenging to derive due to calculus constraints. Importantly, this extension retains the advantages of dynamic NFs, such as exact density estimation and efficient sampling.

Experiments showed that our model produces better performance and possesses a generalization capacity, on both toy and real datasets. On the numerical side, we identified that the training of the time-change function can produce somewhat different solutions having, however, similar likelihood. Therefore, further investigation must be carried out on regularization criteria and inductive biases, in order to stabilize training.

## 6. REFERENCES

- [1] Bernt Oksendal, *Stochastic differential equations: an introduction with applications*, Springer Science & Business Media, 2013.
- [2] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, “Generative adversarial nets,” *Advances in neurips*, vol. 27, 2014.
- [3] Diederik P Kingma and Max Welling, “Auto-encoding variational bayes,” *arXiv:1312.6114*, 2013.
- [4] George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan, “Normalizing Flows for Probabilistic Modeling and Inference,” 12 2019.
- [5] Ivan Kobyzev, Simon J. D. Prince, and Marcus A. Brubaker, “Normalizing Flows: An Introduction and Review of Current Methods,” 8 2019.
- [6] Yang Song and Stefano Ermon, “Generative modeling by estimating gradients of the data distribution,” *Advances in neural information processing systems*, vol. 32, 2019.
- [7] Jinsung Yoon, Daniel Jarrett, and Mihaela Van der Schaar, “Time-series generative adversarial networks,” *Advances in neural information processing systems*, vol. 32, 2019.
- [8] Patrick Kidger, James Foster, Xuechen Li, and Terry J Lyons, “Neural sdes as infinite-dimensional gans,” in *International conference on machine learning*. PMLR, 2021, pp. 5453–5463.
- [9] Xuechen Li, Ting-Kam Leonard Wong, Ricky TQ Chen, and David K Duvenaud, “Scalable gradients and variational inference for stochastic differential equations,” in *Symposium on Advances in Approximate Bayesian Inference*. PMLR, 2020, pp. 1–28.
- [10] Sebastian Zeng, Florian Graf, and Roland Kwitt, “Latent sdes on homogeneous spaces,” *arXiv preprint arXiv:2306.16248*, 2023.
- [11] Nazanin Mehrasa, Ruizhi Deng, Mohamed Osama Ahmed, Bo Chang, Jiawei He, Thibaut Durand, Marcus Brubaker, and Greg Mori, “Point process flows,” *arXiv preprint arXiv:1910.08281*, 2019.
- [12] Oleksandr Shchur, Marin Biloš, and Stephan Günnemann, “Intensity-free learning of temporal point processes,” *arXiv preprint arXiv:1909.12127*, 2019.
- [13] Will Grathwohl, Ricky T. Q. Chen, Jesse Bettencourt, Ilya Sutskever, and David Duvenaud, “FFJORD: Free-form Continuous Dynamics for Scalable Reversible Generative Models,” 10 2018.
- [14] Ruizhi Deng, Bo Chang, Marcus A Brubaker, Greg Mori, and Andreas Lehrmann, “Modeling continuous stochastic processes with dynamic normalizing flows,” *Advances in Neural Information Processing Systems*, vol. 33, pp. 7805–7815, 2020.
- [15] Ruizhi Deng, Marcus A Brubaker, Greg Mori, and Andreas Lehrmann, “Continuous latent process flows,” *Advances in Neural Information Processing Systems*, vol. 34, pp. 5162–5173, 2021.
- [16] Belinda Tzen and Maxim Raginsky, “Neural stochastic differential equations: Deep latent gaussian models in the diffusion limit,” *arXiv preprint arXiv:1905.09883*, 2019.
- [17] Xuanqing Liu, Tesi Xiao, Si Si, Qin Cao, Sanjiv Kumar, and Cho-Jui Hsieh, “Neural sde: Stabilizing neural ode networks with stochastic noise,” *arXiv:1906.02355*, 2019.
- [18] Daniel Revuz and Marc Yor, *Continuous martingales and Brownian motion*, vol. 293, Springer Science & Business Media, 2013.
- [19] Shreyas Chaudhari, Srinivasa Pranav, and José MF Moura, “Learning gradients of convex functions with monotone gradient networks,” in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.
- [20] Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Yingxia Shao, Wentao Zhang, Bin Cui, and Ming-Hsuan Yang, “Diffusion models: A comprehensive survey of methods and applications,” *arXiv preprint arXiv:2209.00796*, 2022.
- [21] Alessandro Ticchi, Andrew Scherer, Carla McIntyre, Carlos Stein N Brito, Derek Snow, Develra, dstern, James Colless, Kieran Garvey, Maggie, Maria Perez Ortiz, Ryan Lynch, and Sohier Dane, “G-research crypto forecasting,” <https://kaggle.com/competitions/g-research-crypto-forecasting>, 2021, Kaggle.
- [22] Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang, “Informer: Beyond efficient transformer for long sequence time-series forecasting,” in *Proceedings of the AAAI conference on artificial intelligence*, 2021, vol. 35, pp. 11106–11115.