

# Time-Causal VAE: Robust Financial Time Series Generator

Beatrice Acciaio\*, Stephan Eckstein† and Songyan Hou\*

November 6, 2024

## Abstract

We build a time-causal variational autoencoder (TC-VAE) for robust generation of financial time series data. Our approach imposes a *causality constraint* on the encoder and decoder networks, ensuring a causal transport from the real market time series to the fake generated time series. Specifically, we prove that the TC-VAE loss provides an upper bound on the causal Wasserstein distance between market distributions and generated distributions. Consequently, the TC-VAE loss controls the discrepancy between optimal values of various dynamic stochastic optimization problems under real and generated distributions. To further enhance the model’s ability to approximate the latent representation of the real market distribution, we integrate a RealNVP prior into the TC-VAE framework. Finally, extensive numerical experiments show that TC-VAE achieves promising results on both synthetic and real market data. This is done by comparing real and generated distributions according to various statistical distances, demonstrating the effectiveness of the generated data for downstream financial optimization tasks, as well as showcasing that the generated data reproduces stylized facts of real financial market data.

*Keywords:* adapted Wasserstein distance, empirical measure, convergence rate, kernel smoothing  
 MSC (2020): 37M10, 68T07

## 1 Introduction

For financial time series, the shortage of samples makes it statistically hard for empirical processes to achieve an acceptable confidence level in describing the underlying market distribution. In practice, it is widely recognized among financial engineers that back-testing exclusively on empirical market data results in significant over-fitting, which leads to unpredictably high risks in decision making based on these tests [Bai+16]. Synthetic data are therefore generated to augment scarce market data, and used to improve back-testing, stress-testing, exploring new scenarios, and in deep learning processes in financial applications; see the overview given in [Ass+20a]. For those purposes, the generated data should look like plausible samples from the underlying market distribution, for example reproducing stylized facts observed in the market. In particular, we want the distribution of the generated data to be close to the underlying market distribution in their performance on decision making problems, such as pricing and hedging, as well as optimal stopping and utility maximization. Notably, these problems are not continuous with respect to widely used distances, such as the Maximum Mean Discrepancy (MMD) and the Wasserstein distances ( $\mathcal{W}$ -distances). On the other hand, these problems are Lipschitz-continuous with respect to stronger metrics, called adapted Wasserstein distances ( $\mathcal{AW}$ -distances) [Bac+20; PP14]. Therefore, for the augmentation of market data with the purpose of e.g. testing performance of different strategies within any of the above problems, it is desirable to find a generated distribution which is close to the underlying market distribution in  $\mathcal{AW}$ -distance.

In that spirit, one natural choice would be to build a generative adversarial network (GAN) employing adapted distances as loss functions. Xu et al. [Xu+20] introduce the COT-GAN, using causal Wasserstein

---

\*Department of Mathematics, ETH Zürich, Switzerland.  
 beatrice.acciaio@math.ethz.ch, songyan.hou@math.ethz.ch

†Department of Mathematics, University of Tübingen, Germany.  
 stephan.eckstein@uni-tuebingen.de

distances ( $CW$ -distances) as a compromise between  $\mathcal{W}$ -distances and  $\mathcal{AW}$ -distances. Since  $CW$ -distances are still considerably more expensive to compute than  $\mathcal{W}$ -distances in a multi-step setting, COT-GAN can only provide satisfactory results for time series with few time steps. This limits the application of  $CW$ -GAN for financial time series generation. Aside from the time-step constraint, recent research has shown that distributions generated via Wasserstein GANs often are far from the source distributions in  $\mathcal{W}$ -distance [Sta+21]. This also explains why GANs show mode collapse [TT20], which refers to a scenario where the generator starts producing a limited variety of outputs, often very similar to each other, instead of a diverse range that represents the real data distribution. Not to mention that the adversarial training to find the saddle point of the min-max problem is notoriously unstable. Lastly, GANs are also usually “data hungry” [Kar+20], and scarcity of market data is the initial problem we started with.

For these reasons, we decided to avoid adversarial minimization, and instead adopt the network structure of variational autoencoders (VAEs) introduced in [KW14]. VAEs are highly expressive models that retain the computational efficiency of fully factorized models [Cin+21] and have found wide applications in generating data for speech, images, and text [Bow+15]. Notably, very deep VAEs generalize autoregressive models and can outperform them on images [Chi20]. Moreover, VAEs frameworks are not only useful in generation, but also able to learn a disentangled latent representation of the data distribution, see [Che+18; Mat+19]. This is in particular true for  $\beta$ -VAE, which we use in the present paper, see [Bur+18; BSR24]. Recently, a series of papers have presented different extensions of VAEs to process sequential data, see for example the summary paper [Gir+20]. In the present paper, we introduce a variation of VAEs, which is able to learn the conditional distribution of financial time series under  $CW$ -distance. Specifically, the encoder maps the market underlying data distribution  $\mu_{\text{data}}$  into a latent distribution  $\mu_{\text{latent}}$  on the latent space, while the decoder maps the latent distribution back to a reconstructed distribution  $\mu_{\text{rec}}$  on the data space. The decoder will be used to generate a distribution  $\mu_{\text{gen}}$  by pushing a prior distribution  $\mu_{\text{prior}}$  defined on the latent space. As it is common for VAEs, we want to achieve two goals at the same time: 1) minimize the reconstruction error  $\mathcal{L}_{\text{rec}}$  between  $\mu_{\text{data}}$  and  $\mu_{\text{rec}}$ ; 2) minimize the latent error  $\mathcal{L}_{\text{latent}}$  between  $\mu_{\text{latent}}$  and the prior distribution  $\mu_{\text{prior}}$ . As a result, the generated distribution  $\mu_{\text{gen}}$  should also be close to the data distribution  $\mu_{\text{data}}$ .

Crucially, we incorporate two modifications to VAEs:

- (i) **Causality constraint:** we impose a causality condition on the encoder and decoder, so that the reconstruction path at time  $t$  depends on the input path only up to time  $t$ . We name the resulting network structure *Time-Causal-VAE* (TC-VAE).
- (ii) **Flexible prior:** we apply a flexible learnable prior distribution  $\mu_{\text{prior}}$ , and specifically the RealNVP introduced in [DSB16a; GST20].

Networks with a time-causal structure are already present in the literature. Those include, for example, recurrent neural networks and causal self-attention networks, both proven highly successful in time series generation; see [Chu+15; YJS19; EHR17; Yan+21].

From the causal optimal transport point of view, our encoder and decoder together transport the market data distribution to the reconstructed distribution in a causal fashion. Consequently, we can prove that the  $CW$ -distance between  $\mu_{\text{data}}$  and  $\mu_{\text{rec}}$  is bounded by the reconstruction error  $\mathcal{L}_{\text{rec}}$ . On the other hand, RealNVP has been proven very successful in approximating distributions, and its density computationally very tractable [DSB16b], which allows an easy computation of the KL-divergence. The flexibility and tractability of RealNVP empowers TC-VAE such that  $\mu_{\text{prior}}$  and  $\mu_{\text{latent}}$  are close enough (in KL-divergence) to control the  $CW$ -distance between  $\mu_{\text{gen}}$  and  $\mu_{\text{rec}}$ . Consequently, the TC-VAE loss controls the  $CW$ -distance between  $\mu_{\text{gen}}$  and  $\mu_{\text{data}}$ , thereby providing one-sided guarantees of such control problems through the  $CW$ -distance.

With these improvements, TC-VAE achieves the goals which we laid out above, generating financial time series data with strong statistical guarantees according to causal Wasserstein distances and showcasing promising numerical results for financial tasks. On synthetic datasets like the Black-Scholes model, Heston model and Path-Dependent-Volatility model, TC-VAE learns the data distribution very well in terms of drift, volatility, marginal distribution, Wasserstein distance [ACB17], Gaussian maximum mean discrepancies [Gre+12], Signature maximum mean discrepancies [Lia+24], adapted Wasserstein distance [Bac+20],

and optimal values of multistage optimization problems, like mean-variance portfolio optimization [FV22], log-utility maximization [Mer75], and optimal stopping [BCJ19]. On real market datasets, such as S&P 500 and VIX, conditional TC-VAE enables us to generate paths, as many as possible and as long as possible. The generated paths reproduce stylized facts of financial time series [Con01] capturing key properties such as gain/loss asymmetry, skewness and kurtosis of returns, heavy-tail returns, no correlation in returns, short time correlation in square returns, long time correlation in absolute returns, and volatility clustering.

**Organization of the paper.** In the rest of Section 1, we give a brief overview of related works and introduce relevant notation. In Section 2, we introduce the architecture of TC-VAE and its loss function. In Section 3 we prove robustness of stochastic optimization problems w.r.t. causal distances. Finally, in Section 4 we show that TC-VAE achieves promising results on both unconditional and conditional financial time series generation on several different metrics and datasets<sup>1</sup>.

**Related Literature.** Numerous methodologies have been explored for generating financial time series. [KS19] were among the first to use restricted Boltzmann machines to generate synthetic foreign exchange rates. Recent advances in deep learning have introduced promising techniques, including Variational Autoencoders (VAEs) introduced in [KW14] and Generative Adversarial Networks (GANs) pioneered in [Goo+14], for generating synthetic financial data.

Variational Autoencoders (VAEs) have recently gained significant popularity in financial data generation. The first contribution in this field is the logSig-VAE, introduced by [Büh+20], which utilizes a log-signature transformation and then applies the VAE in the transformed log-signature space. In parallel, [Des+21] proposed Time-VAE, specifically designed for predicting time series data. In the application of simulating option markets, [Wie+21] combines the autoencoder structure with normalizing flows, while seamlessly integrating a no-arbitrage condition to ensure market consistency. Later, [Cai+23] developed a hybrid VAE to integrate the learning of local patterns and temporal dynamics by variational inference for time series forecasting. Meanwhile, [Liu+22] introduced an innovative VAE variant that bridges temporal convolutional networks and transformers through a layer-wise parallel structure, enhancing the model’s ability to handle temporal sequences. Furthermore, [HCQ24] presents FTS-Diffusion, a novel VAE designed in particular to model irregular and scale-invariant patterns in time series. In addition, [CS24] introduce the SigMMD-VAE, which uses the signature MMD to separate distributions. More recently, [Sch24] addresses the critical balance between model performance and interpretability by connecting ARMA-GARCH with LSTM-based VAE.

Other research mostly follows the GAN approach. This approach is first adapted to financial data generation in [TCT19], and later its variants are explored in [Efi+20; Mee19]. To better address the time dependency of financial time series, GANs with temporal structure are also explored in [Wie+20; FHO22; EHR17; YJS19]. Conditional GANs are also studied in [KFT21; Col+21]. A big family of time series GAN is based on signature transformation, e.g. in [Ni+21; Lia+24; LLN24; BGW24; Iss+24]. In the signature-based GANs, neural SDEs introduced in [Kid21] have been proven successful. The generative expressiveness of neural SDEs is then studied in [Kid21] as infinite-dimensional GANs. Some research particularly focus on the financial quantities, see e.g. [Con+22; Eri+24; VPC24; Riz+23]. Many other approaches have also been explored. For example, [HHP23] suggests a novel time series generation approach using the Schrödinger bridge framework, while [Nag+23] uses autoregressive models to generate limit order book data. In fact, a great amount of literature is dedicated to the generation of limit order book data, e.g. [Hul+23; Con+23; Li+20; Özy21; Hul21; Col+23].

For an extensive overview on synthetic data generation, we refer the reader to [Lu+23] (for general data), [Igl+23] (for time series), [EO21] (GANs for financial time series), and [Ass+20b] (for general data in finance).

**Notation.** All random variables are defined on a fixed probability space  $(\Omega, \mathcal{F}, \mathbb{P})$ , and all equalities and inequalities are intended to hold in  $\mathbb{P}$ -almost sure sense. For  $n \in \mathbb{N}$ , we denote by  $\mathcal{P}(\mathbb{R}^n)$  the set of probability measures on  $\mathbb{R}^n$  and by  $\mathcal{P}_p(\mathbb{R}^n)$  its subset of measures with finite  $p$ -th moment,  $p \in [1, \infty)$ . For  $m \in \mathbb{R}^n$  and

---

<sup>1</sup>The code and data are available at <https://github.com/justinhou95/TimeCausalVAE>.

a positive-semidefinite matrix  $\Sigma \in \mathbb{R}^{n^2}$ , we write  $\mathcal{N}(m, \Sigma) \in \mathcal{P}(\mathbb{R}^n)$  for the normal distribution on  $\mathbb{R}^n$  with mean  $m$  and covariance matrix  $\Sigma$ , and  $\varphi_{m, \Sigma}$  for its density. For simplicity, if  $m = \mathbf{0}$  and  $\Sigma = \mathbf{id}_n$  is the identity matrix, we denote this density by  $\varphi$ . The *entropy* of a measure  $\mu \in \mathcal{P}(\mathbb{R}^n)$  with density  $p_\mu$  is given by

$$\mathbb{H}(\mu) = - \int \log(p_\mu(x)) p_\mu(x) dx.$$

For measures  $\mu, \nu \in \mathcal{P}(\mathbb{R}^n)$  with densities  $p_\mu, p_\nu$ , the *Kullback–Leibler (KL) divergence* between  $\mu$  and  $\nu$  (or *relative entropy* of  $\mu$  w.r.t.  $\nu$ ) is given by

$$\mathcal{D}_{\text{KL}}(\mu|\nu) = \int \log\left(\frac{p_\mu(x)}{p_\nu(x)}\right) p_\mu(x) dx.$$

For  $p \in [1, \infty)$  and  $\mu, \nu \in \mathcal{P}_p(\mathbb{R}^n)$ , the  $p$ -Wasserstein distance between  $\mu$  and  $\nu$  is given by

$$\mathcal{W}_p(\mu, \nu) := \inf_{\pi \in \Pi(\mu, \nu)} \left( \int \|x - y\|^p \pi(dx, dy) \right)^{1/p},$$

where  $\Pi(\mu, \nu)$  denotes the subset of  $\mathcal{P}(\mathbb{R}^n \times \mathbb{R}^n)$  of measures with first marginal  $\mu$  and second marginal  $\nu$ . The elements of  $\Pi(\mu, \nu)$  are called *couplings* of  $\mu$  and  $\nu$ . For  $\pi \in \Pi(\mu, \nu)$ , we denote by  $\pi^x$  its kernel (disintegration) w.r.t.  $\mu$ , so that  $\pi(dx, dy) = \mu(dx) \pi^x(dy)$ . For  $n, N \in \mathbb{N}$  and  $\xi \in \mathcal{P}(\mathbb{R}^n)$ , the push-forward measure of  $\xi$  through a measurable map  $T : \mathbb{R}^n \rightarrow \mathbb{R}^N$ , denoted by  $T_{\#}\xi$ , is the probability measure on  $\mathbb{R}^N$  such that  $T_{\#}\xi(A) = \xi(T^{-1}(A))$  for all Borel sets  $A \in \mathcal{B}(\mathbb{R}^N)$ . For  $d, T \in \mathbb{N}$ , we will look at the space  $\mathbb{R}^{dT}$  as the collection of  $d$ -dimensional paths of length  $T$ . For  $x = (x_1, \dots, x_T) \in \mathbb{R}^{dT}$ , we adopt the notation  $x_{s:t} = (x_s, \dots, x_t)$ , for  $1 \leq s \leq t \leq T$ . Moreover, we denote the up-to-time- $t$  marginal of  $\mu \in \mathcal{P}(\mathbb{R}^{dT})$  by  $\mu_{1:t}$ , and the kernel of  $\mu$  w.r.t. it by  $\mu_{x_{1:t}}$ , so that  $\mu(dx) = \mu_{x_{1:t}}(dx_{t+1:T})$ . Similarly, we denote the up-to-time- $t$  marginal of  $\pi \in \Pi(\mu, \nu)$  by  $\pi_{1:t}$ , and the kernel of  $\pi$  w.r.t. it by  $\pi_{x_{1:t}, y_{1:t}}$ .

## 2 Causal generator

Our goal is to construct a path generator such that the generated paths are close to the observed ones, in the sense that they can be thought of as originating from the same underlying distribution. For  $d, T \in \mathbb{N}$ , we are interested in the space  $\mathcal{P}_1(\mathbb{R}^{dT})$  of distributions on  $d$ -dimensional paths of length  $T$ . We start with the observation of a set of such paths, which we consider to be an i.i.d. sample from an underlying *data distribution*  $\mu_{\text{data}} \in \mathcal{P}_1(\mathbb{R}^{dT})$ . The aim then is to build a generator that produces paths from a *generated distribution*  $\mu_{\text{gen}} \in \mathcal{P}_1(\mathbb{R}^{dT})$  which we want to be as close as possible to  $\mu_{\text{data}}$ . As explained in the introduction, our main motivation is data augmentation for robust decision making. The robustness results presented in Section 3 motivate us to use a modified version of the classical Wasserstein distance, called causal Wasserstein distance, which we recall in the next subsection. Afterwards, we introduce a specific structure of variational autoencoder, where we impose causality constraints on the encoder and decoder maps, and build a generator connected to it; see Figure 1. We will show that, thanks to this causal structure, one can successfully control the causal Wasserstein distance between the data distribution and the generated one.

### 2.1 Causal Distances

When facing stochastic optimization problems in a dynamic setting (that is, when the optimization depends on a process that evolves in time), distances from classical optimal transport (OT), such as Wasserstein distances, have proven unsuitable; see e.g. [PP14]. The key observation is that, in a dynamic context, the value process depends crucially on the conditional distributions forward in time and hence agents take decisions accordingly. This suggests to modify OT-distances by putting additional emphasis on conditional laws. This leads to couplings  $\pi \in \Pi(\mu, \nu)$  such that the conditional law of  $\pi$  is still a coupling of the conditional laws of  $\mu$  and  $\nu$ , that is  $\pi_{x_{1:t}, y_{1:t}} \in \Pi(\mu_{x_{1:t}}, \nu_{y_{1:t}})$ . Such couplings are called *bi-causal* and we denote their collection by  $\Pi_{\text{bc}}(\mu, \nu)$ ; see [Las18; PP12; PP14]. We call the coupling  $\pi$  *causal* if  $\pi_{x_{1:t}, y_{1:t}} \in \Pi(\mu_{x_{1:t}}, \cdot)$ ,

and write  $\pi \in \Pi_c(\mu, \nu)$ , and we call it *anti-causal* if  $\pi_{x_{1:t}, y_{1:t}} \in \Pi(\cdot, \nu_{y_{1:t}})$ , and write  $\pi \in \Pi_{ac}(\mu, \nu)$ . Clearly,  $\Pi_{bc}(\mu, \nu) = \Pi_c(\mu, \nu) \cap \Pi_{ac}(\mu, \nu)$ . The causality constraint can be expressed in several different ways, see e.g. [Bac+17; ABZ20] in the context of transport, and [BY78] in the filtration enlargement framework. Roughly speaking, in a causal transport, for every time  $t$ , only information on the  $x$ -coordinate up to time  $t$  is used to determine the mass transported to the  $y$ -coordinate at time  $t$ . And in a bi-causal transport this holds in both directions, i.e. also when exchanging the role of  $x$  and  $y$ . By means of these concepts, one can introduce constrained optimal transport problems, where the allowed couplings satisfy the causality or bicausality condition. We introduce them directly in the space of interest for us, that is  $\mathcal{P}(\mathbb{R}^{dT})$ , and for a specific cost, that is the sum over time steps of the Euclidean norm on  $\mathbb{R}^d$ .

**Definition 2.1** (Causal Wasserstein distance). The (first order) *causal Wasserstein distance*  $\mathcal{CW}_1$  on  $\mathcal{P}_1(\mathbb{R}^{dT})$  is defined by

$$\mathcal{CW}_1(\mu, \nu) = \inf_{\pi \in \Pi_c(\mu, \nu)} \int \sum_{t=1}^T \|x_t - y_t\|_{\mathbb{R}^d} \pi(dx, dy). \quad (2.1)$$

We need to stress that calling  $\mathcal{CW}_1$  a distance is an abuse of terminology, as this is clearly an asymmetric notion. We still adopt it as this is customary in the literature. One way to recover symmetry, is to use bi-causal couplings instead of causal ones.

**Definition 2.2** (Adapted Wasserstein distance / Nested distance). The (first order) *adapted Wasserstein distance*  $\mathcal{AW}_1$  on  $\mathcal{P}_1(\mathbb{R}^{dT})$  is defined by

$$\mathcal{AW}_1(\mu, \nu) = \inf_{\pi \in \Pi_{bc}(\mu, \nu)} \int \sum_{t=1}^T \|x_t - y_t\|_{\mathbb{R}^d} \pi(dx, dy). \quad (2.2)$$

Bi-causal couplings and the corresponding optimal transport problem were considered by Rüschendorf [Rüs85] in so-called ‘Markov-constructions’. This concept was independently introduced by Pflug-Pichler [PP12] in the context of stochastic multistage optimization problems (see also [PP14; PP15; PP16; GPP19; Pic13]), and considered by Bion-Nadal and Talay in [BT19] and by Gigli in [Gig08].

As already mentioned above, adaptedness (or bi-causality) turns out to be the correct constraint to impose on couplings in order to modify the Wasserstein distance so to ensure robustness of a large class of stochastic optimization problems. That is to say, if two measures  $\mu, \nu$  are close w.r.t. this distance, then solving w.r.t.  $\mu$  optimization problems such as optimal stopping, optimal hedging, utility maximization etc, provides an ‘almost optimizer’ for  $\nu$ ; see [Bac+20; PP14]. This is not true for the Wasserstein distance, which is in fact unable to capture fundamental differences in the evolution of time series, see [PP14]. We refer to Section 3 below for robustness results w.r.t. the causal Wasserstein distance.

We conclude this subsection by showing some simple ordering between distances, that will turn out to be useful for our estimates later. For this we recall the concepts of total variation and adapted total variation, for  $\mu, \nu \in \mathcal{P}(\mathbb{R}^{dT})$ :

$$\text{TV}_0(\mu, \nu) = \inf_{\pi \in \Pi(\mu, \nu)} \int \mathbb{1}_{x \neq y} \pi(dx, dy) \quad \text{and} \quad \text{AV}_0(\mu, \nu) = \inf_{\pi \in \Pi_{bc}(\mu, \nu)} \int \mathbb{1}_{x \neq y} \pi(dx, dy),$$

respectively, with  $|\mu - \nu| = \mu + \nu - 2(\mu \wedge \nu)$ .

**Lemma 2.3.** *Let  $B_1 = \{x \in \mathbb{R}^{dT} : \|x\| \leq 1\}$ . Then, for all  $\mu, \nu \in \mathcal{P}_1(B_1)$ ,*

$$\mathcal{CW}_1(\mu, \nu) \leq \mathcal{AW}_1(\mu, \nu) \leq 2\text{AV}_0(\mu, \nu) \leq C\text{TV}_0(\mu, \nu) \leq C\sqrt{\frac{1}{2}\mathcal{D}_{\text{KL}}(\mu|\nu)}, \quad (2.3)$$

where  $C = 2(2^T - 1)$ .

*Proof.* The first inequality is obvious, and the second one is due to the fact that  $\mu, \nu \in \mathcal{P}_1(B_1)$ . The third inequality is Lemma 3.5 in [EP22] and the last inequality follows by Pinsker’s inequality [Tak05].  $\square$

## 2.2 TC-VAE: time-causal variational autoencoder

*Variational Autoencoders* (VAEs), introduced in [KW14], are *deep latent-variable models* (DLVMs) employed to generate new data. Our data space is  $\mathbb{R}^{dT}$ , while as latent space we consider  $\mathbb{R}^{dzT}$  for some fixed  $d_Z \in \mathbb{N}$ . A VAE consists of an encoding map (encoder) and a decoding map (decoder) where the former allows to go from the data space to the latent one, and the latter goes in the opposite direction. The encoder involves two networks  $\mu_\phi: \mathbb{R}^{dT} \rightarrow \mathbb{R}^{dzT}$  and  $\sigma_\phi: \mathbb{R}^{dT} \rightarrow \mathbb{R}^{dzT}$ , parameterized by  $\phi$ . Here,  $\mu_\phi$  encodes data  $x \in \mathbb{R}^{dT}$  to a latent point  $\mu_\phi(x) \in \mathbb{R}^{dzT}$ , and  $\sigma_\phi(x)$  defines the scaling of Gaussian noise that is later added to the latent point. On the other hand, the decoder map consists of a network  $\text{De}_\theta: \mathbb{R}^{dzT} \rightarrow \mathbb{R}^{dT}$ , parameterized by  $\theta$ , simply mapping points from the latent space back to the data space. The input distribution is our data distribution  $\mu_{\text{data}}$ , and we let  $X \sim \mu_{\text{data}}$ . We consider Gaussian noise  $\varepsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{dzT})$  independent of  $X$ , and define

$$\begin{cases} Z = \mu_\phi(X) + \sigma_\phi(X)\varepsilon, \\ Y = \text{De}_\theta(Z). \end{cases} \quad (2.4)$$

We denote by  $\mu_{\text{latent}} = Z_{\#}\mathbb{P}$  the distribution on the latent space  $\mathbb{R}^{dzT}$  resulting from the encoding step, and by  $\mu_{\text{rec}} = Y_{\#}\mathbb{P}$  the reconstructed distribution on the data space  $\mathbb{R}^{dT}$  resulting from the combination of the encoding and decoding steps.

As a DLVM, VAE generates data in the following two steps: (1) sampling  $\hat{z}^{(i)} \in \mathbb{R}^{dzT}$  from a prior distribution  $\mu_{\text{prior}} \in \mathcal{P}_1(\mathbb{R}^{dzT})$ ; (2) generating a sample  $\hat{x}^{(i)} \in \mathbb{R}^{dT}$ , conditioned on  $\hat{z}^{(i)}$ , through the pushforward of the decoder network, i.e. getting  $\hat{x}^{(i)}$  as sample from the generated distribution

$$\mu_{\text{gen}} = \text{De}_\theta_{\#}\mu_{\text{prior}} \in \mathcal{P}_1(\mathbb{R}^{dT}).$$

Our goal is to control the causal Wasserstein distance  $\mathcal{CW}_1(\mu_{\text{data}}, \mu_{\text{gen}})$ , so we aim at training our VAE such that this is minimized. As we discussed already, estimating  $\mathcal{CW}_1$  is difficult and computationally intractable due to the lack of explicit causal couplings. To overcome this difficulty, we use the fact that  $\mathcal{CW}_1$  satisfies the triangle inequality (see [Pam24]), so that we can control  $\mathcal{CW}_1(\mu_{\text{data}}, \mu_{\text{gen}})$  as follows:

$$\mathcal{CW}_1(\mu_{\text{data}}, \mu_{\text{gen}}) \leq \mathcal{CW}_1(\mu_{\text{data}}, \mu_{\text{rec}}) + \mathcal{CW}_1(\mu_{\text{rec}}, \mu_{\text{gen}}). \quad (2.5)$$

Crucially, if we restrict  $\mu_\phi$ ,  $\sigma_\phi$  and  $\text{De}_\theta$  to the class of causal maps introduced below,  $(X, Y)_{\#}\mathbb{P}$  is a causal coupling from  $\mu_{\text{data}}$  to  $\mu_{\text{rec}}$ , i.e.  $(X, Y)_{\#}\mathbb{P} \in \Pi_c(\mu_{\text{data}}, \mu_{\text{rec}})$ . We will show that  $\mathcal{CW}_1(\mu_{\text{data}}, \mu_{\text{rec}})$  can be bounded by the transport cost associated to the coupling  $(X, Y)_{\#}\mathbb{P}$ , while  $\mathcal{CW}_1(\mu_{\text{rec}}, \mu_{\text{gen}})$  can be bounded by a computationally tractable quantity.

**Causal maps.** A map  $\mathcal{T}: \mathbb{R}^{d_1T} \rightarrow \mathbb{R}^{d_2T}$ ,  $d_1, d_2 \in \mathbb{N}$ , is causal if and only if there exist Borel-measurable maps  $\mathcal{T}^t: \mathbb{R}^{d_1t} \rightarrow \mathbb{R}^{d_2t}$ ,  $t = 1, \dots, T$ , such that

$$\mathcal{T}(x) = (\mathcal{T}^1(x_{1:1}), \mathcal{T}^2(x_{1:2}), \dots, \mathcal{T}^T(x)), \quad x \in \mathbb{R}^{d_1T}.$$

Intuitively, the  $t$ -coordinate of  $\mathcal{T}(x) = \mathcal{T}(x_{1:T})$  only depends on  $x_{1:t}$ , i.e. on the values of  $x$  up to time  $t$ .

Let  $X \sim \mu \in \mathcal{P}_1(\mathbb{R}^{d_1T})$ ,  $Y = \mathcal{T}(X)$ , and denote the law of  $Y$  by  $\nu = \mathcal{T}_{\#}\mu \in \mathcal{P}_1(\mathbb{R}^{d_2T})$ , so that  $\pi = (\mathbf{id}, \mathcal{T})_{\#}\mu \in \Pi(\mu, \nu)$ . Note that, for  $\mathcal{T}$  causal, for all  $t = 1, \dots, T-1$ ,

$$\text{Law}(X_{t+1}|Y_{1:t}, X_{1:t}) = \text{Law}(X_{t+1}|\mathcal{T}^t(X_{1:t}), X_{1:t}) = \text{Law}(X_{t+1}|X_{1:t}),$$

which implies that  $\pi_{x_{1:t}, y_{1:t}}(dx_{t+1}) = \mu_{x_{1:t}}(dx_{t+1})$ . Therefore, in this case  $\pi$  is a causal coupling from  $\mu$  to  $\nu$ , i.e.  $\pi \in \Pi_c(\mu, \nu)$ . Moreover, if we have a sequence of causal maps, then their composition is still a causal map by definition. This enables us to build complex maps by using simple causal maps as building blocks.

**Time-causal VAE (TC-VAE).** We let  $\mu_\phi: \mathbb{R}^{dT} \rightarrow \mathbb{R}^{dzT}$ ,  $\sigma_\phi: \mathbb{R}^{dT} \rightarrow \mathbb{R}^{dzT}$  and  $\text{De}_\theta: \mathbb{R}^{dzT} \rightarrow \mathbb{R}^{dT}$  in VAE be *causal maps* parameterized by  $\phi$  and  $\theta$ . Let  $\varepsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{dzT})$  be independent of  $X \sim \mu_{\text{data}}$  and consider the autoencoder structure in (2.4); see Figure 1 for visualization.

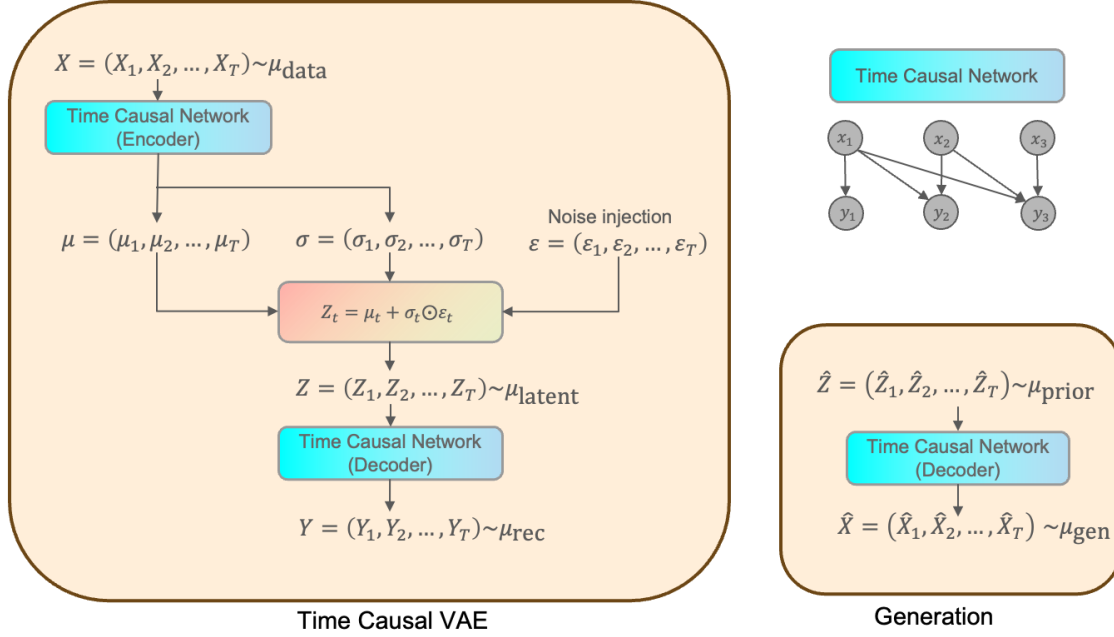


Figure 1: Time-causal variational autoencoder and generation

Since  $X$  and  $\varepsilon$  are independent, we have that, for all  $t = 1, \dots, T - 1$ ,

$$\begin{aligned} \text{Law}(X_{t+1}|Y_{1:t}, X_{1:t}) &= \text{Law}(X_{t+1}|\text{De}_\theta(\mu_\phi(X_{1:t}) + \sigma_\phi(X_{1:t})\varepsilon_{1:t}), X_{1:t}) = \text{Law}(X_{t+1}|X_{1:t}, \varepsilon_{1:t}) \\ &= \text{Law}(X_{t+1}|X_{1:t}), \end{aligned}$$

so that  $(X, Z)_{\#}\mathbb{P}$  is a causal coupling.

Now we define the reconstruction loss by

$$\mathcal{L}_{\text{rec}} := \mathbb{E}_{\mathbb{P}}[\|X - Y\|] = \mathbb{E}_{\mathbb{P}}[\|X - \text{De}_\theta(\mu_\phi(X) + \sigma_\phi(X)\varepsilon)\|].$$

Hereby, we emphasize that both  $Y$  and  $\mathcal{L}_{\text{rec}}$  implicitly depend on both  $\theta$  and  $\phi$ . Intuitively, minimizing  $\mathcal{L}_{\text{rec}}$  over the set of parameters  $\phi, \theta$  corresponds to minimizing a relaxed version of  $\mathcal{CW}_1(\mu_{\text{data}}, \mu_{\text{rec}})$  where causal couplings are restricted to a subset of couplings defined by our network's (causal) structure.

**Lemma 2.4.** *We can estimate the first term in (2.5) via*

$$\mathcal{CW}_1(\mu_{\text{data}}, \mu_{\text{rec}}) \leq \mathcal{L}_{\text{rec}}. \quad (2.6)$$

*Proof.* Similarly as above, we can see that  $\pi = (X, Y)_{\#}\mathbb{P} \in \Pi_{\text{c}}(\mu_{\text{data}}, \mu_{\text{rec}})$ . Then, by the definition of causal Wasserstein distance, we have the estimate

$$\mathcal{CW}_1(\mu_{\text{data}}, \mu_{\text{rec}}) \leq \mathbb{E}_{\mathbb{P}}[\|X - Y\|] = \mathbb{E}_{\mathbb{P}}[\|X - \text{De}_\theta(\mu_\phi(X) + \sigma_\phi(X)\varepsilon)\|].$$

□

Next we focus on the second term in (2.5), that is  $\mathcal{CW}_1(\mu_{\text{rec}}, \mu_{\text{gen}})$ .

**Theorem 2.5.** *Assume  $\mu_{\text{gen}}, \mu_{\text{rec}} \in \mathcal{P}_1(B_1)$ , where  $B_1 = \{x \in \mathbb{R}^{dT} : \|x\| \leq 1\}$ . Then*

$$\mathcal{CW}_1(\mu_{\text{rec}}, \mu_{\text{gen}}) \leq C \sqrt{\frac{1}{2} \mathcal{D}_{\text{KL}}(\mu_{\text{latent}}|\mu_{\text{prior}})}. \quad (2.7)$$

*Proof.* Lemma 2.3 implies

$$\mathcal{CW}_1(\mu_{\text{rec}}, \mu_{\text{gen}}) \leq C \sqrt{\frac{1}{2} \mathcal{D}_{\text{KL}}(\mu_{\text{rec}} | \mu_{\text{gen}})}. \quad (2.8)$$

Now, note that  $\mu_{\text{rec}}$  and  $\mu_{\text{gen}}$  originate from the same pushforward map  $\text{De}_\theta$ , applied to the latent distribution  $\mu_{\text{latent}}$  and to the prior distribution  $\mu_{\text{prior}}$ , respectively, i.e.

$$\mu_{\text{rec}} = \text{De}_\theta \# \mu_{\text{latent}}, \quad \mu_{\text{gen}} = \text{De}_\theta \# \mu_{\text{prior}}.$$

Then, the data processing inequality (see e.g. [Nut21, Lemma 1.6]) yields

$$\mathcal{D}_{\text{KL}}(\mu_{\text{rec}}, \mu_{\text{gen}}) \leq \mathcal{D}_{\text{KL}}(\mu_{\text{latent}}, \mu_{\text{prior}}), \quad (2.9)$$

which concludes the proof.  $\square$

Combining (2.5), Lemma 2.4 and Theorem 2.5, gives the following estimate for the causal Wasserstein distance between market data and generated data.

**Corollary 2.6.** *Assume  $\mu_{\text{gen}}, \mu_{\text{rec}} \in \mathcal{P}_1(B_1)$ , where  $B_1 = \{x \in \mathbb{R}^{dT} : \|x\| \leq 1\}$ . Then*

$$\mathcal{CW}_1(\mu_{\text{data}}, \mu_{\text{gen}}) \leq \mathcal{L}_{\text{rec}} + C \sqrt{\frac{1}{2} \mathcal{D}_{\text{KL}}(\mu_{\text{latent}} | \mu_{\text{prior}})}. \quad (2.10)$$

*Remark 2.7.* The compactness assumption in Theorem 2.5 and Corollary 2.6 is not a constraint on our data since, in practice, we typically normalize our training data to reside within a bounded region, typically a ball.

The estimate in (2.10) allows us to obtain one sided estimates for a class of stochastic optimization problems, when employing generated rather than observed data, see Remark 3.2 below.

### 2.3 TC-VAE Loss

Motivated by Corollary 2.6, we aim at training our TC-VAE by minimizing the two terms on the RHS of (2.10):  $\mathcal{L}_{\text{rec}}$  and  $\mathcal{D}_{\text{KL}}(\mu_{\text{latent}} | \mu_{\text{prior}})$ . The first term  $\mathcal{L}_{\text{rec}}$ , defined in (2.6), is the well-known reconstruction loss and can be efficiently estimated by taking batch-wise sample expectations. However, the second term  $\mathcal{D}_{\text{KL}}(\mu_{\text{latent}} | \mu_{\text{prior}})$  involves an intractable term  $\mu_{\text{latent}}$ , also called the *aggregated posterior distribution*. Recall that  $\mu_{\text{latent}} = Z \# \mathbb{P} = (\mu_\phi(X) + \sigma_\phi(X)\varepsilon) \# \mathbb{P}$  where  $X \sim \mu_{\text{data}}$  and  $\varepsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{dT})$  are independent. We denote the density of  $\mu_{\text{latent}}$  and  $\mu_{\text{prior}}$  by  $p_{\text{latent}}$  and  $p_{\text{prior}}$ , respectively. Then we have  $p_{\text{latent}}(\cdot) = \int q_\phi(\cdot | x) \mu_{\text{data}}(dx)$ , where  $q_\phi(\cdot | x)$  is the density of  $\mathcal{N}(\mu_\phi(x), \Sigma_\phi(x))$  and  $\Sigma_\phi(x)$  is the diagonal matrix with diagonal vector  $\sigma_\phi(x)$  for  $x \in \mathbb{R}^{dT}$ . Note that the integral form of  $p_{\text{latent}}$  makes it computationally intractable, and so is  $\mathcal{D}_{\text{KL}}(\mu_{\text{latent}} | \mu_{\text{prior}})$ . Luckily,  $\mathcal{D}_{\text{KL}}(\mu_{\text{latent}} | \mu_{\text{prior}})$  can be bounded by  $\int \mathcal{D}_{\text{KL}}(q_\phi(\cdot | x) | p_{\text{prior}}) \mu_{\text{data}}(dx)$ , following the same arguments used in [KW14] to bound log-likelihood by evidence lower bound:

$$\begin{aligned} \mathcal{D}_{\text{KL}}(\mu_{\text{latent}} | \mu_{\text{prior}}) &= - \int \log(p_{\text{prior}}(z)) p_{\text{latent}}(z) dz - \mathbb{H}(p_{\text{latent}}) \\ &= - \int \log(p_{\text{prior}}(z)) \int q_\phi(z | x) \mu_{\text{data}}(dx) dz - \mathbb{H}(p_{\text{latent}}) \\ &= \int \mathcal{D}_{\text{KL}}(q_\phi(\cdot | x) | p_{\text{prior}}) \mu_{\text{data}}(dx) + \int \mathbb{H}(q_\phi(\cdot | x)) \mu_{\text{data}}(dx) - \mathbb{H}\left(\int q_\phi(\cdot | x) \mu_{\text{data}}(dx)\right) \\ &\leq \int \mathcal{D}_{\text{KL}}(q_\phi(\cdot | x) | p_{\text{prior}}) \mu_{\text{data}}(dx) = \mathbb{E}_{\mathbb{P}}[\mathcal{D}_{\text{KL}}(q_\phi(\cdot | X) | p_{\text{prior}})] =: \mathcal{L}_{\text{latent}}, \end{aligned} \quad (2.11)$$

where last inequality follows by Jensen’s inequality, and  $\mathcal{L}_{\text{latent}}$  stands for “latent loss”. Here again, for simplicity, we omit  $\phi$  in the notation of  $\mathcal{L}_{\text{latent}}$ . Now, instead of minimizing  $\mathcal{D}_{\text{KL}}(\mu_{\text{latent}} | \mu_{\text{prior}})$ , we minimize



$\mathcal{L}_{\text{latent}}$ , which is tractable. More precisely, we follow the loss design in  $\beta$ -VAE in [Hig+16] by setting  $\mathcal{L}_{\theta,\phi} := \mathcal{L}_{\text{rec}} + \beta \mathcal{L}_{\text{latent}}$  and solving

$$\min_{\theta,\phi} \mathcal{L}_{\theta,\phi},$$

where  $\beta$  is an hyper-parameter. We now focus on computing  $\mathcal{L}_{\theta,\phi}$ . Notice that  $q_\phi(\cdot|x)$  is the density of  $\text{Law}(Z|X=x)$ , so we can rewrite

$$\mathcal{L}_{\text{latent}} = \mathbb{E}_{\mathbb{P}} \left[ \mathbb{E}_{Z \sim q_\phi(\cdot|X)} \left[ \log q_\phi(Z|X) - \log p_{\text{prior}}(Z) \right] \right] = \mathbb{E}_{\mathbb{P}} \left[ \log q_\phi(Z|X) - \log p_{\text{prior}}(Z) \right].$$

This yields

$$\mathcal{L}_{\theta,\phi} = \mathbb{E}_{(X,\epsilon) \sim \mu_{\text{data}} \otimes \mathcal{N}(\mathbf{0}, \mathbf{I}_{d_{z^T}}), Z = \mu_\phi(X) + \sigma_\phi(X)\epsilon} \left[ \|X - \text{De}_\theta(Z)\| + \beta \left( \log q_\phi(Z|X) - \log p_{\text{prior}}(Z) \right) \right].$$

**Sample-based loss.** In practice, we have no access to  $\mu_{\text{data}}$ , but only to finitely many samples from observation. We let  $x^{(i)} \in \mathbb{R}^{d^T}, i = 1, \dots, n$ , be i.i.d. samples from  $\mu_{\text{data}} \in \mathcal{P}_1(\mathbb{R}^{d^T})$ , and  $\hat{\mu}_{\text{data}} = \frac{1}{n} \sum_{i=1}^n \delta_{x^{(i)}}$  be the corresponding empirical distribution. So we take expectation under  $\hat{\mu}_{\text{data}}$  instead of  $\mu_{\text{data}}$  and minimize a sample-based version of the loss:

$$\mathcal{L}_{\theta,\phi}^n = \frac{1}{n} \sum_{i=1}^n \|x^{(i)} - \text{De}_\theta(z^{(i)})\| + \beta \frac{1}{n} \sum_{i=1}^n \left( \log q_\phi(z^{(i)}|x^{(i)}) - \log p_{\text{prior}}(z^{(i)}) \right),$$

where  $\{(x^{(i)}, \epsilon^{(i)})\}_{i=1, \dots, n}$  are i.i.d. samples from  $\mu_{\text{data}} \otimes \mathcal{N}(\mathbf{0}, \mathbf{I}_{d_{z^T}})$ , and  $z^{(i)} = \mu_\phi(x^{(i)}) + \sigma_\phi(x^{(i)})\epsilon^{(i)}$  for  $i = 1, \dots, n$ .

## 2.4 Flow-based prior distribution

In the standard VAE, the prior distribution  $\mu_{\text{prior}}$  is a fix standard Gaussian prior, hence non-learnable, and the regularization term pushes the aggregated posterior to match it. In practice, the matching is not satisfied because  $\mathcal{L}_{\text{rec}}$  forces the encoder to be irregular and, in the end, it is almost impossible to precisely match a fixed-shaped prior. As a result, one obtains ‘holes’, namely regions in the latent space where the aggregated posterior assigns low probability while the prior assigns relatively high probability. This is an issue in generation because sampling from the prior, from the hole, may result in a sample that is of extremely low quality [RV18]. It is even more problematic in our application, because we are not only interested in generating samples on the manifold where empirical data lies, but also generating a distribution close to the empirical measure under some strong probabilistic metric. Therefore, in the latent space, we not only require that the aggregated posterior has no holes, but actually need that the aggregated posterior is close to the prior.

Sampling directly from the posterior distribution at first glance seems to solve this issue, but it potentially leads to overfitting of the empirical data. Therefore, we consider a learnable prior to match the posterior distribution. There are many options such as mixture of Gaussians [Dil+16], VampPrior [TW18], generative topographic mapping [BSW98], flow-based prior [Che+16], etc. In our case, we use the flow-based prior because of its high flexibility with time series data. Let  $Z_0$  be a random variable on  $\mathbb{R}^{d_{z^T}}$  with density  $p_0(z_0)$ . Let  $f_1: \mathbb{R}^{d_{z^T}} \rightarrow \mathbb{R}^{d_{z^T}}$  be invertible, set  $Z_1 = f_1(Z_0)$  and denote its density by  $p_1(z_1)$ . Then, for all  $z_1 = f_1(z_0), z_0 \in \mathbb{R}^{d_{z^T}}$ ,

$$p_1(z_1) = p_0(z_0) \left| \det \frac{dz_0}{dz_1} \right| = p_0(z_0) \left| \det \nabla f_1^{-1}(z_1) \right|,$$

where  $\nabla f_1^{-1}(z_1)$  is the Jacobian of  $f_1^{-1}$  at  $z_1$ . Now, let  $f_1, \dots, f_N$  be a sequence of invertible functions, and set  $f = f_1 \circ \dots \circ f_N$ . Let  $Z_j = f_j(Z_{j-1})$  and denote its density by  $p_j(z_j)$ , for  $j = 1, \dots, N$ . Then, for all  $j = 1, \dots, N, z_j = f_j(z_{j-1}), z_0 \in \mathbb{R}^{d_{z^T}}$ , we have

$$p_j(z_j) = p_{j-1}(z_{j-1}) \left| \det \nabla f_j^{-1}(z_j) \right|.$$

Given such a chain of probability density functions, we have

$$\log(p_N(z_N)) = \log\left(p_0(z_0) \prod_{j=1}^N \left|\det \nabla f_j^{-1}(z_j)\right|\right) = \log(p_0(z_0)) + \sum_{j=1}^N \log\left(\left|\det \nabla f_j^{-1}(z_j)\right|\right).$$

Next, we parameterize  $f_1^\lambda, \dots, f_N^\lambda$  by a parameter  $\lambda$  and let  $f^\lambda = f_1^\lambda \circ \dots \circ f_N^\lambda$ . Then  $Z_N = f^\lambda(Z_0)$  has a learnable density  $p_\lambda(z)$ .

**Sample-based loss with flow-based prior distribution.** Setting  $p_{\text{prior}} = p_\lambda$  in our TC-VAE, we end up with a learnable prior, and  $\log p_{\text{prior}}(z)$  in the regularization term  $\mathcal{L}_{\text{latent}}$  becomes

$$\log p_{\text{prior}}(z) = \log(p_0(z_0)) + \sum_{j=1}^N \log\left(\left|\det \nabla f_j^{\lambda^{-1}}(z_j)\right|\right),$$

where  $z = z_N$ ,  $z_j = f_j^\lambda(z_{j-1})$  for all  $j = 1, \dots, N$ ,  $z_0 \in \mathbb{R}^{dz^T}$ . We end up minimizing the following loss:

$$\mathcal{L}_{\theta, \phi, \lambda}^n = \frac{1}{n} \sum_{i=1}^n \|x^{(i)} - \text{De}_\theta(z^{(i)})\| + \beta \frac{1}{n} \sum_{i=1}^n \log q_\phi(z^{(i)} | x^{(i)}) - \beta \frac{1}{n} \sum_{i=1}^n \left( \log(p_0(z_0^{(i)})) + \sum_{j=1}^N \log\left(\left|\det \nabla f_j^{\lambda^{-1}}(z_j^{(i)})\right|\right) \right),$$

where  $z_N^{(i)} = z^{(i)}$ ,  $z_j^{(i)} = f_j^\lambda(z_{j-1}^{(i)})$  for all  $i = 1, \dots, N$ . In practice, we choose the sequence of invertible transformations from the class of real-valued non-volume-preserving (real NVP) transformations; see [DSB16a].

*Remark 2.8.* In the flow-based prior, we generalize the neural SDE [Kid+21] with noise process driven by a richer class of distribution beyond the Gaussian noise.

### 3 Causal Robustness

In this section, we show why causal (resp. adapted) Wasserstein distance is a suitable way to measure closeness when considering a broad class of stochastic optimization problems. This is done by establishing one-sided (resp. two-sided) robustness of controlled problems under this distance. Let  $\mathcal{Q}: \mathbb{R}^{dT} \times \mathbb{R}^{dT} \rightarrow \mathbb{R}$  be a measurable function. For  $\mu \in \mathcal{P}_1(\mathbb{R}^{dT})$ , we consider the following stochastic optimization problem:

$$\mathcal{V}(\mu) := \inf_{H \in \mathcal{H}} V(H, \mu), \quad V(H, \mu) = \int \mathcal{Q}(x, H(x)) \mu(dx) = \mathbb{E}_{X \sim \mu}[\mathcal{Q}(X, H(X))], \quad (3.1)$$

where  $\mathcal{H}$  is a closed convex subset of the set  $\mathcal{K}$  of adapted strategies (controls):

$$\mathcal{K} := \{H = (H_t)_{t=1}^T : H_t(x) = H_t(x_{1:t}), H_t: \mathbb{R}^{dT} \rightarrow \mathbb{R}^d \text{ measurable}\}.$$

Here, with an abuse of notation, we write  $H_t(x_{1:t})$ , meaning that the function  $H_t$  defined on  $\mathbb{R}^{dT}$  only depends on the first  $t$  coordinates of the argument, i.e.  $H$  is adapted.

**Theorem 3.1.** *Let  $L \geq 0$  and  $\mathcal{Q}$  be s.t.  $(x, h) \mapsto \mathcal{Q}(x, h)$  is uniformly  $L$ -Lipschitz in  $x$  and convex in  $h$ . Then, for all  $\mu, \nu \in \mathcal{P}_1(\mathbb{R}^{dT})$ ,*

$$\begin{aligned} \mathcal{V}(\mu) - \mathcal{V}(\nu) &\leq LCW_1(\mu, \nu), \\ |\mathcal{V}(\mu) - \mathcal{V}(\nu)| &\leq LAW_1(\mu, \nu). \end{aligned}$$

*Proof.* Let  $\pi \in \Pi_c(\mu, \nu)$  be an optimal coupling for  $CW_1(\mu, \nu)$ . Notice that, for all  $G \in \mathcal{H}$ , we have

$$\begin{aligned} -V(G, \nu) &= - \int \mathcal{Q}(y, G(y)) \nu(dy) = - \int \mathcal{Q}(y, G(y)) \pi(dx, dy) \\ &= \int [\mathcal{Q}(x, G(y)) - \mathcal{Q}(y, G(y))] \pi(dx, dy) - \int \mathcal{Q}(x, G(y)) \pi(dx, dy). \end{aligned} \quad (3.2)$$

We first estimate the first term on the RHS of (3.2). By the Lipschitz property of  $\mathcal{Q}$ , we have

$$\int [\mathcal{Q}(x, G(y)) - \mathcal{Q}(y, G(y))] \pi(dx, dy) \leq L \int \|x - y\| \pi(dx, dy) = L \cdot \mathcal{CW}_1(\mu, \nu). \quad (3.3)$$

Next, we estimate the second term on the RHS of (3.2). By convexity of  $\mathcal{Q}$  and Jensen's inequality, we have

$$\begin{aligned} - \int \mathcal{Q}(x, G(y)) d\pi &= - \int \int \mathcal{Q}(x, G(y)) \pi_x(dy) \mu(dx) \\ &\leq - \int \mathcal{Q}\left(x, \int G(y) \pi_x(dy)\right) \mu(dx) = - \int \mathcal{Q}\left(x, \tilde{G}(x)\right) \mu(dx), \end{aligned}$$

where  $\tilde{G} = (\tilde{G}_t)_{t=1}^T: \mathbb{R}^{dT} \rightarrow \mathbb{R}^{dT}$  is defined as  $\tilde{G}(x) := \int G(y) \pi_x(dy)$ . Then, since  $\pi \in \Pi_c(\mu, \nu)$ , for all  $x = x_{1:T} \in \mathbb{R}^{dT}$  and  $t = 1, \dots, T$ , we have

$$\tilde{G}_t(x) = \int G_t(y) \pi_x(dy) = \int G_t(y_{1:t}) \pi_x(dy_{1:t}) = \int G_t(y_{1:t}) \pi_{x_{1:t}}(dy_{1:t}) = \tilde{G}_t(x_{1:t}),$$

where the second equality follows by adaptedness of  $G$ , and the third one by causality of  $\pi$ . This implies that  $\tilde{G}$  is also an adapted strategy, and thus  $\tilde{G} \in \mathcal{H}$ . Therefore, for the second term on the RHS of (3.2), we have

$$- \int \mathcal{Q}(x, G(y)) \pi(dx, dy) \leq -V(\tilde{G}, \mu) \leq -\mathcal{V}(\mu). \quad (3.4)$$

Combining (3.2), (3.3) and (3.4), for all  $H, G \in \mathcal{H}$  we have that

$$-V(G, \nu) \leq L \cdot \mathcal{CW}_1(\mu, \nu) - \mathcal{V}(\mu),$$

so that

$$\mathcal{V}(\mu) - V(G, \nu) \leq L \cdot \mathcal{CW}_1(\mu, \nu).$$

Then, by the arbitrariness of  $G \in \mathcal{H}$ , we conclude that

$$\mathcal{V}(\mu) - \mathcal{V}(\nu) \leq L \cdot \mathcal{CW}_1(\mu, \nu).$$

By symmetry, the claimed inequality for  $\mathcal{AW}_1$  holds as well. This completes the proof.  $\square$

*Remark 3.2.* Thanks to the estimate obtained in Corollary 2.6, Theorem 3.1 provides one-sided estimates for stochastic optimization problems as in (3.1), when employing generated rather than observed data. Specifically, for  $\mu_{\text{gen}}, \mu_{\text{rec}} \in \mathcal{P}_1(B_1)$ ,  $B_1 = \{x \in \mathbb{R}^{dT} : \|x\| \leq 1\}$ , and  $\mathcal{Q}$  as in Theorem 3.1, we have

$$\mathcal{V}(\mu_{\text{data}}) \leq \mathcal{V}(\mu_{\text{gen}}) + L \left( \mathcal{L}_{\text{rec}} + C \sqrt{\frac{1}{2} \mathcal{D}_{\text{KL}}(\mu_{\text{latent}} | \mu_{\text{prior}})} \right).$$

This means that, if we solve the minimization problem on the samples we generate, the optimal value, together with the reconstruction loss and the latent loss, gives a conservative upper bound of the optimal value  $\mathcal{V}(\mu_{\text{data}})$  under the data distribution.

In order to provide an example of application of the above theorem, let us recall the definitions of Value at Risk and Average Value at Risk.

**Definition 3.3** (VaR and AVaR). Let  $\alpha \in (0, 1)$  and  $U$  be a real random variable on  $(\Omega, \mathcal{F}, \mathbb{P})$ . The Value-at-Risk (VaR) of  $U$  at confidence level  $\alpha$  is defined as the negative  $\alpha$ -quantile of  $U$  under  $\mathbb{P}$ :

$$\text{VaR}_\alpha(U) = - \inf\{x \in \mathbb{R} : \mathbb{P}(U \leq x) \geq \alpha\}.$$

The Average Value at Risk (or Expected Shortfall) of  $U$  at confidence level  $\alpha$  is defined as the average of the VaR below  $\alpha$ :

$$\text{AVaR}_\alpha(U) = \text{ES}_\alpha(U) = \frac{1}{\alpha} \int_0^\alpha \text{VaR}_u(U) du = \min_{z \in \mathbb{R}} \left\{ \frac{1}{\alpha} E[(z - U)_+] - z \right\}.$$

**Corollary 3.4.** Let  $L \geq 0$  and  $\mathcal{Q}$  be s.t.  $(x, h) \mapsto \mathcal{Q}(x, h)$  is uniformly  $L$ -Lipschitz in  $x$  and concave in  $h$ . For  $\alpha \in (0, 1)$  and  $\mu \in \mathcal{P}_1(\mathbb{R}^{dT})$ , define

$$\mathcal{R}_\alpha(\mu) = \inf_{H \in \mathcal{K}} \text{ES}_\alpha(\mathcal{Q}(X, H(X))), \quad \text{where } X = (X_1, \dots, X_T) \sim \mu.$$

Then, for all  $\mu, \nu \in \mathcal{P}_1(\mathbb{R}^{dT})$ ,

$$\begin{aligned} \mathcal{R}_\alpha(\mu) - \mathcal{R}_\alpha(\nu) &\leq \frac{L}{\alpha} \mathcal{CW}_1(\mu, \nu), \\ |\mathcal{R}_\alpha(\mu) - \mathcal{R}_\alpha(\nu)| &\leq \frac{L}{\alpha} \mathcal{AW}_1(\mu, \nu). \end{aligned} \tag{3.5}$$

*Proof.* By the dual representation of the expected shortfall, we can rewrite  $\mathcal{R}_\alpha(\mu)$  as a stochastic optimization problem on an enlarged space:

$$\mathcal{R}_\alpha(\mu) = \inf_{H \in \mathcal{K}} \inf_{z \in \mathbb{R}} \mathbb{E} \left[ \frac{1}{\alpha} (z - \mathcal{Q}(X, H(X)))_+ - z \right] = \inf_{\tilde{H} \in \tilde{\mathcal{H}}} \mathbb{E}[\tilde{\mathcal{Q}}_\alpha(\tilde{X}, \tilde{H}(\tilde{X}))],$$

with

$$\begin{aligned} \tilde{\mathcal{Q}}_\alpha(\tilde{x}, \tilde{h}) &= \frac{1}{\alpha} (\tilde{h}_1^{(1)} - \mathcal{Q}(\tilde{x}^{(2:d+1)}, \tilde{h}^{(2:d+1)}))_+ - \tilde{h}_1^{(1)}, \\ \tilde{X} &= ([1, X_t]^\top)_{t=1}^T \in \mathbb{R}^{(d+1)T}, \\ \tilde{\mathcal{H}} &= \{\tilde{H}: \tilde{H}_t(\tilde{x}) = [z, H_t(\tilde{x}^{(2:d+1)})]^\top, H \in \mathcal{K}, z \in \mathbb{R}\}, \end{aligned}$$

where, for  $v = [v_1, \dots, v_m]^\top$ , we use the notation  $v^{(i)} = v_i$  and  $v^{(i:j)} = (v_i, \dots, v_j)$ . On the one hand,  $\tilde{\mathcal{Q}}_\alpha$  is uniformly  $\frac{L}{\alpha}$ -Lipschitz in the first argument, because  $u \mapsto \frac{1}{\alpha} u_+$  is  $\frac{1}{\alpha}$ -Lipschitz and  $\mathcal{Q}$  is uniformly  $L$ -Lipschitz in the first argument. On the other hand,  $\tilde{\mathcal{Q}}_\alpha$  is convex in the second argument, because  $u \mapsto \frac{1}{\alpha} u_+$  is convex and  $-\mathcal{Q}$  is convex in the second argument. Therefore, we can apply Theorem 3.1 and complete the proof.  $\square$

A concrete and practical example of a function  $\mathcal{Q}$  satisfying the assumptions in Theorem 3.1 is the profit and loss function for bounded strategies.

**Example 3.5 (P&L).** Consider  $\mathcal{Q}(x, h) = \sum_{t=1}^{T-1} h_t(x_{t+1} - x_t)$ . Then  $\mathcal{Q}$  is linear in  $h$ . Also notice that

$$\left\| \sum_{t=1}^{T-1} h_t(x_{t+1} - x_t) - \sum_{t=1}^{T-1} h_t(x'_{t+1} - x'_t) \right\| \leq \sum_{t=1}^{T-1} \|h\|_\infty (\|x'_{t+1} - x_{t+1}\| + \|x'_t - x_t\|) \leq 2\|h\|_\infty \|x - x'\|.$$

For  $B \geq 0$ , then, over all  $h \in \mathbb{R}^{dT}$  s.t.  $\|h\|_\infty \leq B$ ,  $\mathcal{Q}$  is uniformly  $2B$ -Lipschitz in  $x$ . Therefore, for all  $\alpha \in (0, 1)$ , if we consider the risk minimization problem with  $B$ -bounded strategies:

$$\mathcal{R}_\alpha^B(\mu) = \inf_{H \in \mathcal{H}_B} \text{ES}_\alpha(\mathcal{Q}(X, H(X))) = \inf_{H \in \mathcal{H}_B} \text{ES}_\alpha\left(\sum_{t=1}^{T-1} H_t(X_{1:t})(X_{t+1} - X_t)\right),$$

where  $\mathcal{H}_B = \{H \in \mathcal{K}: \|H\|_\infty \leq B\}$ , then, by Corollary 3.4, the following holds true:

$$\begin{aligned} \mathcal{R}_\alpha^B(\mu) - \mathcal{R}_\alpha^B(\nu) &\leq \frac{2B}{\alpha} \mathcal{CW}_1(\mu, \nu), \\ |\mathcal{R}_\alpha^B(\mu) - \mathcal{R}_\alpha^B(\nu)| &\leq \frac{2B}{\alpha} \mathcal{AW}_1(\mu, \nu). \end{aligned} \tag{3.6}$$

To conclude, while the Wasserstein distance is not strong enough to guarantee closeness in the performance of stochastic optimization problems (see [PP14]), we show that the causal Wasserstein distance can control the closeness in a Lipschitz fashion. The above results can be regarded as asymmetric versions of robustness results known for the adapted Wasserstein distance; see [Bac+20].

## 4 Experiments

In this section, we test the generative capabilities of TC-VAE in three major aspects. First, we compare the generated data and market data under weak metrics [Rac+13]. These include financial statistics [Con01], Wasserstein distance [ACB17], and different maximum mean discrepancies [Gre+12]. Second, we compare the generated data and market data under adapted metrics [Bac+20]. These include adapted Wasserstein distance [Bac+20] and the optimal value of multistage optimization problems, like portfolio optimization [FV22], utility maximization [Mer75], and optimal stopping [BCJ19]. Finally, we evaluate the diversity in generated data compared to market data. This tests how much we enlarge the market dataset with new samples. Throughout this section, we also refer to market data as real paths and to generated data as fake paths.

### 4.1 Synthetic data

#### 4.1.1 Black-Scholes model

The Black-Scholes model [BS73] is the most renowned model in mathematical finance. Various stochastic optimization problems have analytic optimal solutions under this model [LN00; Mer75]. This provides benchmarks for the comparison between market and generated data. Let  $(S_t^{\text{BS}})_{t \geq 0}$  be defined by  $S_0^{\text{BS}} = 1$  and  $dS_t^{\text{BS}} = S_t^{\text{BS}}(\mu dt + \sigma dW_t)$  for all  $t \geq 0$ , with drift  $\mu = 0.1$ , volatility  $\sigma = 0.2$ , and where  $(W_t)_{t \geq 0}$  is the Wiener process. Since we work in discrete time, we let  $\Delta t = 1/12$ ,  $T = 5$ ,  $N_T = T/\Delta t$  to model monthly prices over an investment horizon of 5 years. We choose  $N = 1000$  to be the number of samples in market data, reflecting the scarcity of data in practice [Büh+20].

#### Evaluation under weak metrics

First, we visually compare real paths and fake paths to provide a proof-of-concept for TC-VAE. Figure 2 illustrates that fake paths are visually indistinguishable from real paths.

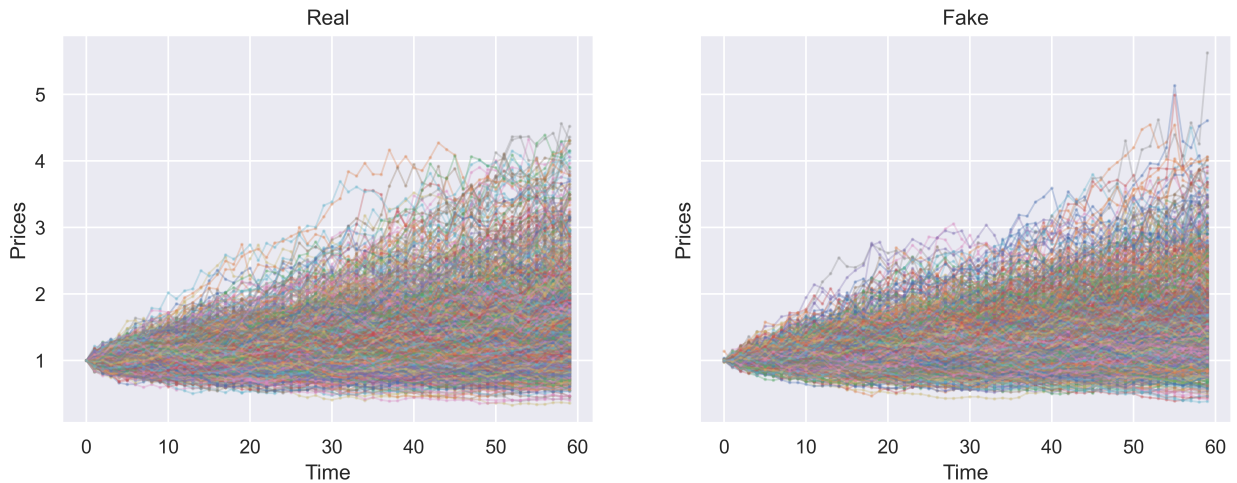


Figure 2: Illustration of real paths from a discretized Black-Scholes model (left) compared to fake paths generated from the TC-VAE model (right).

Then we compare the marginal histograms between real and fake paths. Figure 3 shows that real and fake paths are close in one-dimensional marginal distributions.

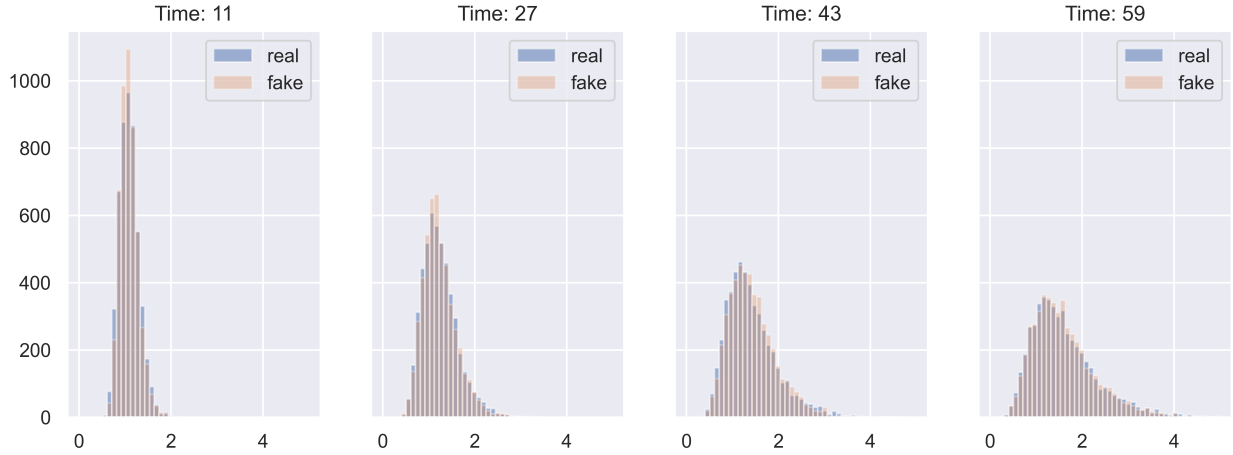


Figure 3: Visualization of marginal distributions at different time slices for real paths from a discretized Black-Scholes model (blue) compared to fake paths generated from the TC-VAE model (orange).

Furthermore, we compare drift and volatility between real and fake paths. For sample paths  $(S_{j\Delta t}^{(n)})_{j=0, \dots, N_T}$ ,  $n=1, \dots, N$ , we compute the log-paths and the volatility:

$$\left( \log S_{j\Delta t}^{(n)} \right)_{\substack{j=0, \dots, N_T \\ n=1, \dots, N}}, \quad \left( \sqrt{\frac{1}{N_T \Delta t} \sum_{j=1}^{N_T} \left( \log S_{j\Delta t}^{(n)} - \log S_{(j-1)\Delta t}^{(n)} \right)^2} \right)_{n=1, \dots, N}.$$

We compare the log-path distribution and the volatility distribution for both real and fake paths; see Figure 4. Overall real and fake paths are close in drift and volatility.

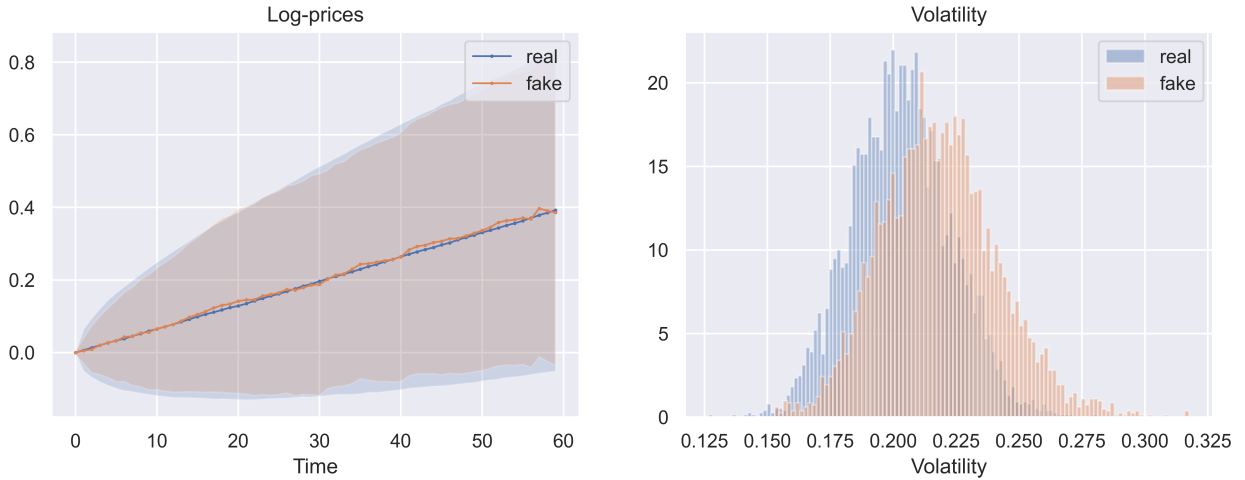


Figure 4: The left-hand side visualizes the log-paths for real prices (blue) compared to fake prices (orange). The solid lines represent the mean and the widths of shadow areas represent the standard deviation. The right-hand side visualizes the histogram of the volatility of real paths (blue) compared to fake paths (orange).

Next, we compute the sliced Wasserstein distance, which is a principled method to simultaneously compare all one-dimensional projected distributions between two measures (see [Kol+19]). We compute the sliced Wasserstein distance between real and fake paths with 10 realization. To benchmark, we compute the

sliced Wasserstein distance between real and control paths, where control paths are discretized Black-Scholes paths different from real paths only in volatility. Moreover, we conduct the same evaluation under Gaussian kernels MMD [Gre+12] and signature MMD [CO18], which are widely used to evaluate the discrepancy between probability measures [Li+17; Ni+21]. In Figure 5, we show that real and fake paths are relatively close in sliced Wasserstein distance, Gaussian MMD, and signature MMD. As a comparison, we train the unconditional Sig-VAE introduced in [CS24] using the same Black-Scholes distribution. We utilize the code from the authors’ repository available at <https://github.com/luchungi/Generative-Model-Signature-MMD>.

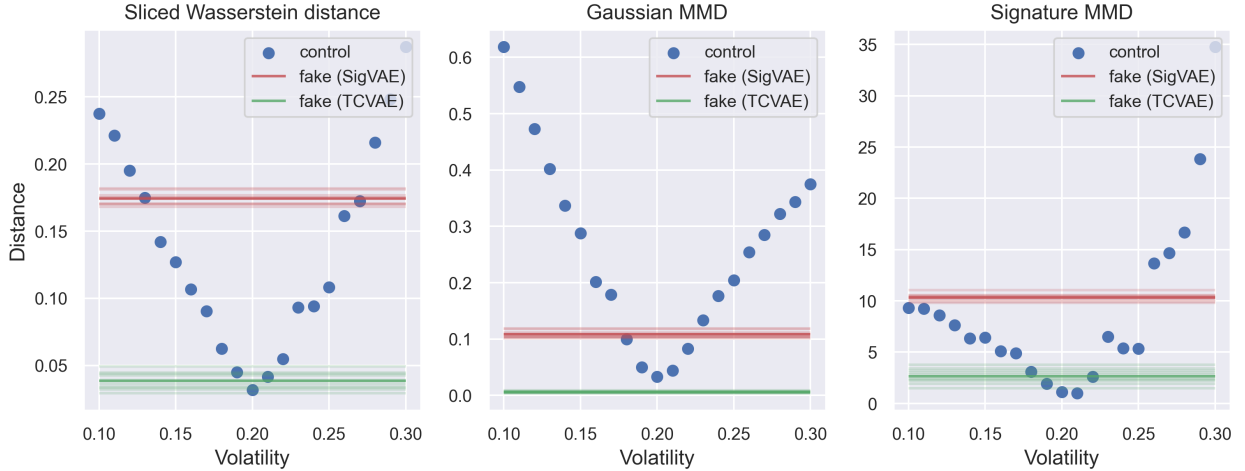


Figure 5: From left to right, we visualize the sliced Wasserstein distance, Gaussian MMD, and signature MMD. The green (resp. red) lines illustrate distances between real paths of the Black-Scholes model and fake paths generated from TC-VAE (resp. Sig-VAE); each line from a different random seed. The blue dots show the distances between real paths and control paths under different volatility levels.

Next, we compare real and fake paths using adapted metrics. In particular, we study the mean-variance portfolio optimization problem, the log-utility maximization problem, the optimal stopping problem and the adapted Wasserstein distance.

### Mean-variance portfolio optimization problem

The mean-variance portfolio optimization problem is one of the most classical and frequently used portfolio selection rules [Rub02]. The investor aims at maximizing the expected return from terminal wealth while minimizing the variance of the portfolio:

$$\mathcal{V}(\mu) = \sup_{\theta} \mathbb{E}_{\mu}[V_T^{\theta}] - \kappa \text{Var}_{\mu}[V_T^{\theta}].$$

The supremum is taken over the set of self-financing trading strategies (see [FS11, Definition 5.4]) and  $\kappa$  is the hyper-parameter.  $V_T^{\theta}$  is the corresponding terminal value of the value process (see [FS11, Definition 5.6]), where the market consists of one risky asset (either real, fake or control paths in our case) and a risk-free asset  $S_t^{\text{bond}} = e^{rt}$ ,  $t \geq 0$ , with risk-free rate  $r = 0.01$ . In Section 3, we showed that this problem is a multistage optimal control problem, for which the optimal values are Lipschitz w.r.t. the adapted Wasserstein distance.

As a first comparison, we consider constant proportional trading strategies. For each constant strategy, we calculate mean and variance of the corresponding terminal wealth. By varying proportions, we obtain the efficient frontier of constant proportional trading strategies. We compare such efficient frontiers for real, fake and control paths. Control paths, as before, are discretized Black-Scholes paths different from real paths only in volatility. The efficient frontier by fake paths matches extremely well to the one by real paths, see the left-hand side of Figure 6.

Next we compare the efficient frontiers under optimal strategies. For Black-Scholes paths (real and control paths), we know the analytical optimal strategies [LN00; FV22], so that we can directly calculate the corresponding efficient frontiers. For fake paths, we restrict the optimization to optimal strategies of the Black-Scholes model. So we compute efficient frontiers with the optimal strategies of Black-Scholes under different volatility  $\sigma$ , denoted by  $\theta_\sigma$ . Then we take the supremum of all efficient frontiers as the efficient frontier of fake paths. The efficient frontier of fake paths matches again extremely well to the one by real paths, see the right-hand side of Figure 6. This implies that the performance of mean-variance portfolio optimization problem are close between real and fake paths.

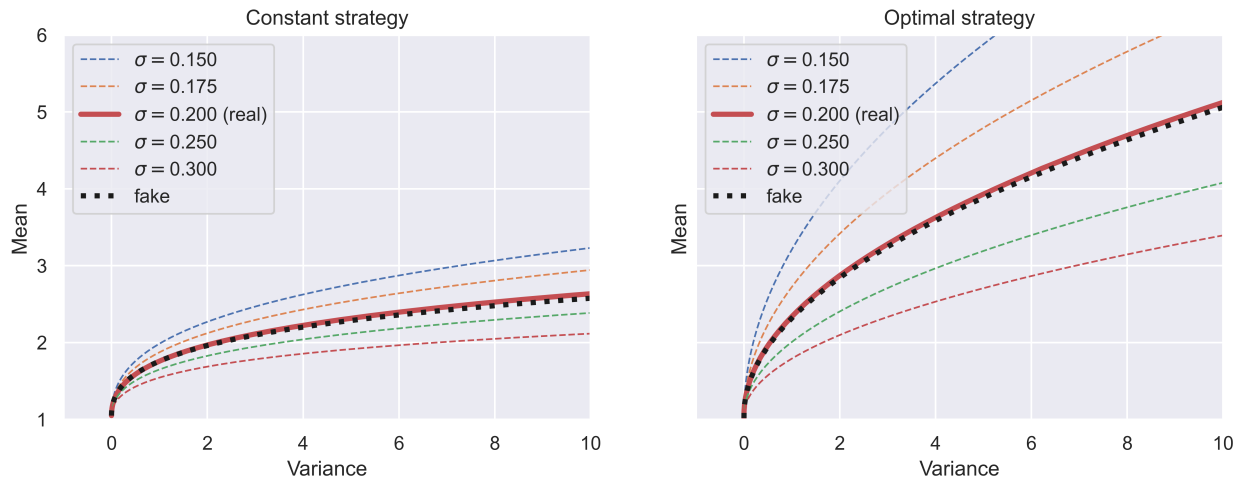


Figure 6: Visualization of efficient mean-variance portfolio frontiers for real (red bold solid line), fake (black dotted lines) and control (dashed lines) paths. On the left-hand side, we show efficient frontiers under constant proportional trading strategies. On the right-hand side, we show efficient frontiers under optimal strategies.

### Log-utility maximization

Next we consider the log-utility maximization problem (Merton's problem [Mer75]):

$$\mathcal{V}(\mu) = \sup_{\theta} \mathbb{E}_{\mu}[\log(V_T^{\theta})],$$

under the same setting as the mean-variance problem above. For both real and fake paths, we solve the log-utility maximization problem numerically among constant proportional trading strategies. We denote  $v^*$  the theoretical optimal utility,  $\mu_{\text{real}}$  the empirical measure of real paths with 50000 samples,  $\mu_{\text{fake}}$  the empirical measure of fake paths with 50000 samples,  $G^*$  the optimal strategy,  $G_{\text{real}}$  the optimal constant proportional trading strategy under  $\mu_{\text{real}}$ ,  $G_{\text{fake}}$  the optimal constant proportional trading strategy under  $\mu_{\text{fake}}$ , and  $V(\mu, G)$  the expected log-utility which arises from using the trading strategy  $G$  in the market  $\mu$ . We compute  $V(\mu_{\text{real}}, G_{\text{real}})$ ,  $V(\mu_{\text{real}}, G_{\text{fake}})$ ,  $V(\mu_{\text{fake}}, G_{\text{real}})$  and  $V(\mu_{\text{fake}}, G_{\text{fake}})$ , and compare them on the left-hand side of Figure 7. To benchmark, we compute the numerically optimal expected log-utility of real paths with 1000 samples (same size as training data). Then we calculate the utilities with 200 random realizations and plot the histogram also on the left-hand side of Figure 7. First we notice that  $V(\mu_{\text{real}}, G_{\text{real}})$  is close to  $v^*$ . This justifies reliability of the numerical solver. Since  $V(\mu_{\text{real}}, G_{\text{real}})$ ,  $V(\mu_{\text{real}}, G_{\text{fake}})$ ,  $V(\mu_{\text{fake}}, G_{\text{real}})$  and  $V(\mu_{\text{fake}}, G_{\text{fake}})$  are all within the support of the histogram, they are considered relatively close to  $v^*$ . In particular, applying  $G_{\text{fake}}$  to  $\mu_{\text{real}}$  yields a very close estimate of the optimal utility.

On the right-hand side of Figure 7, we plot the curve of theoretical optimal log-utility vs volatility. We can interoperate  $V(\mu_{\text{fake}}, G_{\text{fake}})$  as the optimal log-utility of Black-Scholes distribution with a different volatility level, which is very close to the real volatility level  $\sigma = 0.2$ .



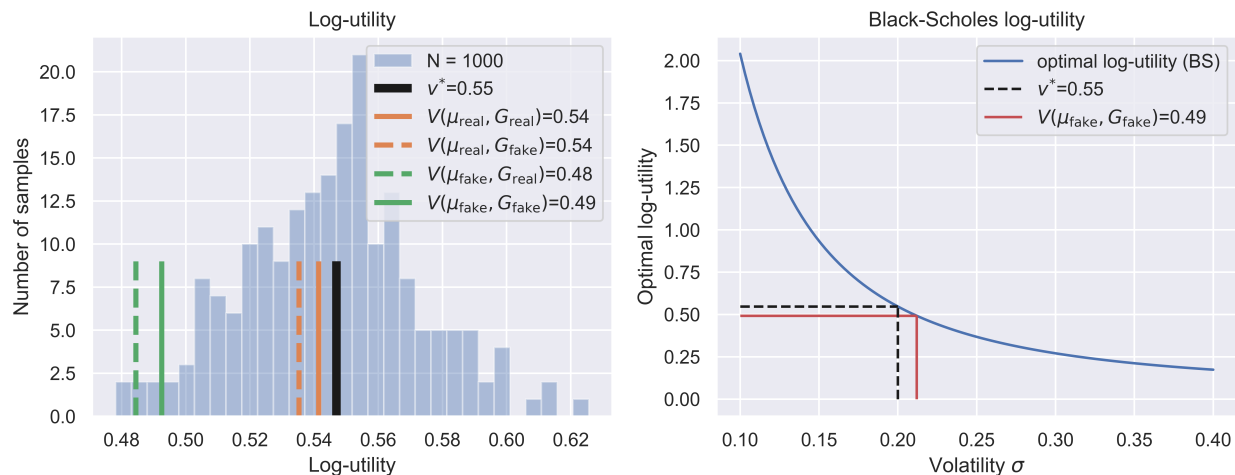


Figure 7: On the left-hand side, we compare the expected log-utility of  $V(\mu_{\text{real}}, G_{\text{real}})$ ,  $V(\mu_{\text{real}}, G_{\text{fake}})$ ,  $V(\mu_{\text{fake}}, G_{\text{real}})$  and  $V(\mu_{\text{fake}}, G_{\text{fake}})$ . To benchmark, the blue histogram showcases numerical optimal utilities values calculated with 1000 samples, using 200 different random seeds for the histogram. On the right-hand side, the blue curve illustrates theoretical optimal log-utility vs volatility.

### Optimal Stopping problem

Now we consider the optimal stopping problem for an American put option written on an underlying asset  $S$  (given by either real, fake, or control paths). The stochastic optimization problem is given by

$$\sup_{\tau \in \mathcal{T}} \mathbb{E}[g(\tau, S_\tau)], \quad (4.1)$$

where  $\mathcal{T}$  is the set of stopping times and  $g(t, S) = e^{-rt} \max(S - K)^+$ ,  $r = 0.1$ ,  $S_0 = 100$ ,  $K = 100$ . Control paths, as before, are discretized Black-Scholes paths different from real paths only in volatility. For real, fake and control paths, we compute the optimal stopping values with the deep optimal stopping solver introduced in [BCJ19]. We compute the optimal stopping values for real and fake paths, where each computation is repeated with 10 different random seeds, and compare them in Figure 8. The optimal stopping values are almost indistinguishable.

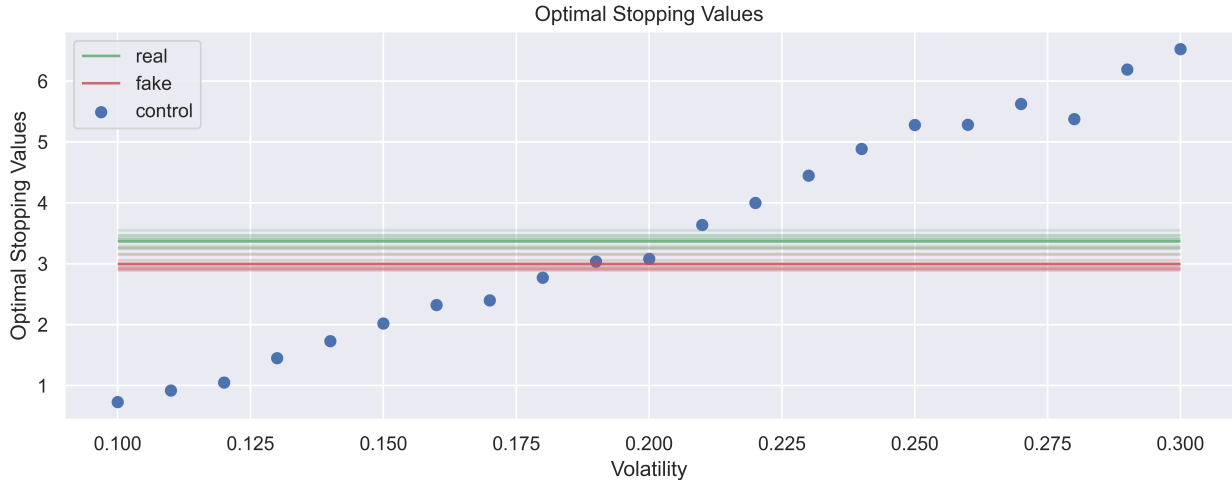


Figure 8: Optimal stopping values under real (Black-Scholes model), fake (TC-VAE) and control distributions (varying volatility for the Black-Scholes model).

### Sliced adapted Wasserstein distance

Ideally, we would like to compute the adapted Wasserstein distance between real and fake paths. However, the adapted Wasserstein distance is computationally heavy, and hence we instead evaluate a sliced version of it. Hereby, we draw inspiration from the growing literature on sliced Wasserstein distances (see e.g. [Des+19; Kol+19; Nie+22]). In general, slicing is a technique to reduce computational burden by considering low-dimensional projections of the distributions of interest. For comparing time series distributions, we interpret slicing in the sense that we only compute the adapted Wasserstein distance over subsets of time, and then average over those subsets. To this end, for a set  $I \subset \{1, \dots, T\}$ , denote by  $\mu_I$  the marginal distribution of  $\mu$  on the subset of times indexed by  $I$ , and by  $|I|$  the number of elements in  $I$ . Given a distribution  $\gamma$  over such subsets  $I$ , we define

$$\mathcal{SAW}_1(\mu, \nu) := \int \mathcal{AW}_1(\mu_I, \nu_I) \gamma(dI).$$

In practice, we use as  $\gamma$  the uniform distribution over subsets of a certain size, called  $n_{\text{len}}$ . The  $\mathcal{AW}$  distances are evaluated using adapted empirical measure transformations (see [AH24; Bac+22; Hou24]) and then computed using backward induction (cf. [Bac+17; PW22]). To be precise, computation of adapted empirical measures requires clustering of the support points of the considered distributions. While [Bac+22] introduced the adapted empirical measure using clusters based on a predefined grid, we instead use K-means clustering as in [BCJ24], thus effectively adjusting the grids to the particular distributional shapes at hand. For the numerical implementation of the backward induction, we use the POT package [Fla+21] for solving the inner optimal transport problems, i.e., for calculating each value in the dynamic programming formulation (see [Bac+17, equation (5.1)]). We believe that evaluating  $\mathcal{SAW}_1$  gives a good indicator of temporal similarity between two measures while significantly reducing computational burden compared to  $\mathcal{AW}_1$ . However, one must be clear that certain information is lost through slicing, like complex long-range temporal dependencies.

We configure the size of time slice  $n_{\text{len}} = 5$ , the number of subsets  $n_{\text{slice}} = 100$ , the number of samples  $n_{\text{sample}} = 500$ , and the number of random seeds  $n_{\text{seed}} = 100$ . Under this configuration, we compute  $\mathcal{SAW}_1$  between real vs real paths (with different random seed), real vs fake paths, and real vs control paths. Control paths, as before, are discretized Black-Scholes paths different from real paths only in volatility, with  $\sigma = 0.3$  instead of  $\sigma = 0.2$  for real paths. We compare the average distance across  $n_{\text{seed}} = 100$  realizations and the standard deviation in Table 1. We find that  $\mathcal{SAW}_1(\mu_{\text{real}}, \mu'_{\text{real}})$  and  $\mathcal{SAW}_1(\mu_{\text{real}}, \mu_{\text{fake}})$  are almost indistinguishable, indicating that real and fake paths are relatively close under  $\mathcal{SAW}_1$ . Hereby, it is worth

noting that the relatively high mean difference  $\mathcal{SAW}_1(\mu_{\text{real}}, \mu'_{\text{real}})$  can be explained by the fact that our models lie in a very high-dimensional space, and fine-grained high-dimensional similarity is difficult to obtain with only 500 samples. The fact that  $\mathcal{SAW}_1(\mu_{\text{real}}, \mu'_{\text{real}})$  is however relatively stable (low standard deviation) and much lower than  $\mathcal{SAW}_1(\mu_{\text{real}}, \mu_{\text{control}})$  indicates that there are certain low-dimensional features which most sample paths share. In fact, these features may be learned by the generator model as well, which could explain the similarly low values of  $\mathcal{SAW}_1(\mu_{\text{real}}, \mu_{\text{fake}})$ .

Description	Distance	Mean difference	Standard deviation
real-real (different samples)	$\mathcal{SAW}_1(\mu_{\text{real}}, \mu'_{\text{real}})$	0.367	0.032
real-fake	$\mathcal{SAW}_1(\mu_{\text{real}}, \mu_{\text{fake}})$	0.382	0.028
real-control	$\mathcal{SAW}_1(\mu_{\text{real}}, \mu_{\text{control}})$	0.670	0.066

Table 1: Sliced adapted Wasserstein distances between different measures.

#### 4.1.2 Heston model

Next, we consider a more complicated temporal dynamic given by the Heston model:  $(S_t^H, V_t^H)_{t \geq 0}$  s.t.

$$\begin{aligned} \frac{dS_t^H}{S_t^H} &= \mu dt + \sqrt{V_t^H} dW_t^S, \quad S_0^H = 1, \\ dV_t^H &= \kappa(\theta - V_t^H)dt + \xi \sqrt{V_t^H} dW_t^V, \quad V_0^H = \theta, \end{aligned}$$

where  $(W_t^S, W_t^V)_{t \geq 0}$  are Wiener processes with correlation  $\rho = -0.9$ ,  $\mu = 0.02$ ,  $\kappa = 1$ ,  $\theta = 0.2$ ,  $\xi = 0.5$ . We aim at learning the distribution of  $(S_t^H)_{t > 0}$ . Under the same time discretization and sample size as in the Black-Scholes case of Section 4.1.1, we conduct the same tests (omitting the visualization of paths and marginal distribution for the sake of space) and observe a similar performance. For this, see sliced Wasserstein distance, Gaussian kernel MMD, and signature MMD compared in Figure 9; optimal stopping values compared in Figure 10; sliced adapted Wasserstein distance compared in Table 2.

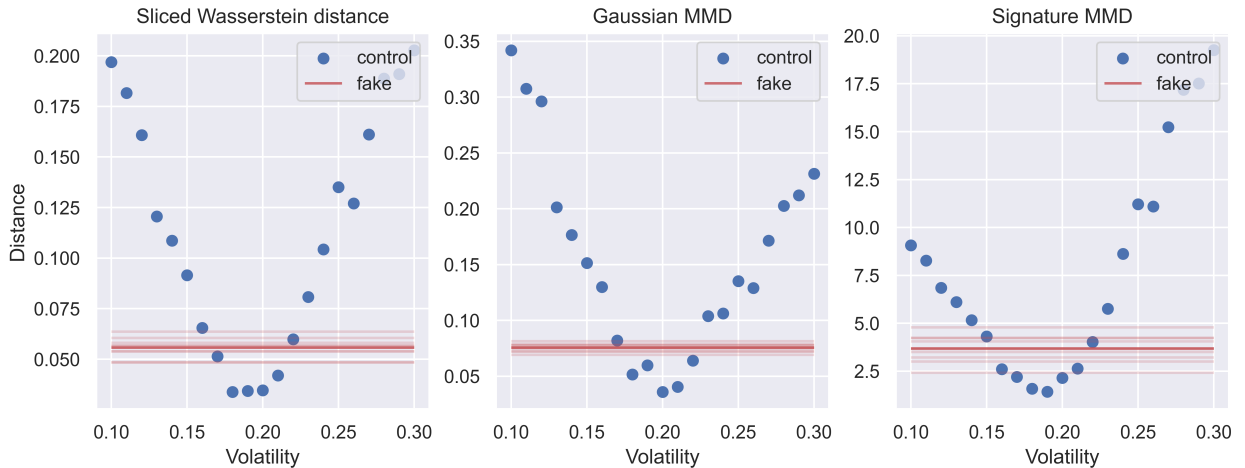


Figure 9: From left to right, we visualize the sliced Wasserstein distance, Gaussian MMD, and signature MMD. The red lines illustrate distances between real paths of the Heston model and fake paths generated from the TC-VAE model (each line is a different random seed). The blue dots show the distances between real paths and control paths. Control paths are discretized Heston paths different from real paths only in  $\theta$ .

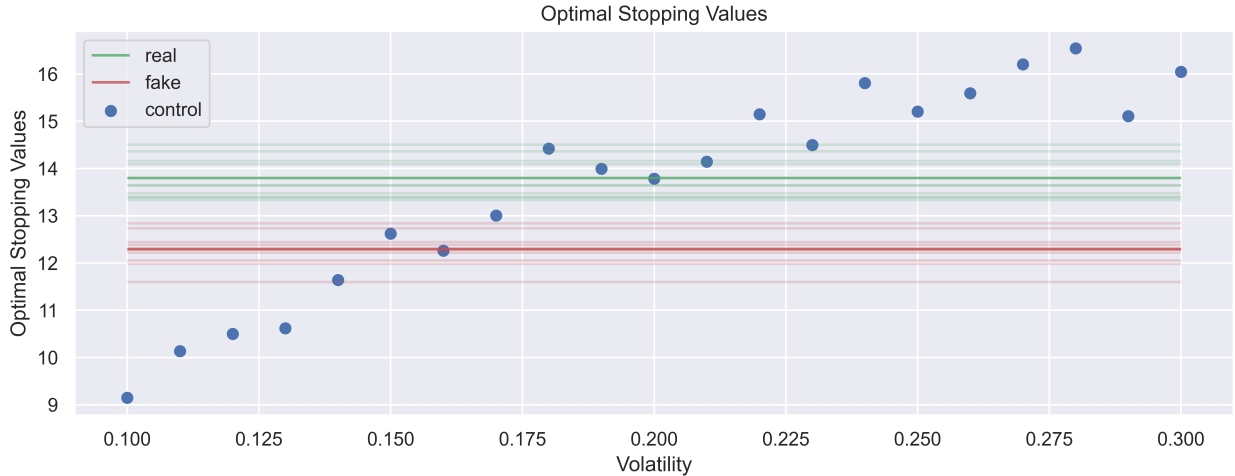


Figure 10: Optimal stopping values under real, fake and control distributions in case of synthetic data using the Heston model. Control paths are discretized Heston paths different from real paths only in  $\theta$ .

Description	Formula	Mean difference	Standard deviation
real-real (different samples)	$\mathcal{SAW}_1(\mu_{\text{real}}, \mu'_{\text{real}})$	0.505	0.027
real-fake	$\mathcal{SAW}_1(\mu_{\text{real}}, \mu_{\text{fake}})$	0.549	0.039
real-control	$\mathcal{SAW}_1(\mu_{\text{real}}, \mu_{\text{control}})$	0.700	0.052

Table 2: Sliced adapted Wasserstein distances between different measures arising from the Heston model. Control paths are discretized Heston paths different from real paths only in  $\theta$  (0.15 vs 0.2).

### 4.1.3 Path dependent volatility model

In a financial market, we only observe a single realization of a path, such as stock prices, rather than i.i.d. paths. Although one can use rolling windows to sample many sub-paths and assume them to be i.i.d. samples, this clearly leads to several risks. First of all, the sub-paths are in fact not independent and this causes severe over-fitting. Even worse, when the observed path is short, in order to extract more sub-paths for training, the rolling windows greatly overlap and causes even higher correlation. Although non-overlapping windows can alleviate correlation, this requires much longer observed paths, which is sometimes not possible. Even when this is possible, this exposes the model to distributional shift over time, which is related to another issue: non-stationarity. The sub-paths are also not identically distributed if the observed path is not stationary. This is common in financial data where the distribution shifts swiftly over time. Thus, the sub-paths might only be correlated samples coming from a different distribution. In the end, we are learning an average distribution over time, which is meaningless for forecasts about possible future evolution. To tackle this issue, we deploy the temporal nature of financial time series by generating future paths from real historical paths. To do so, we further develop a conditional version of TC-VAE. The only difference compared to TC-VAE is that we concatenate the latent variable with an additional conditional variable, see Figure 11.

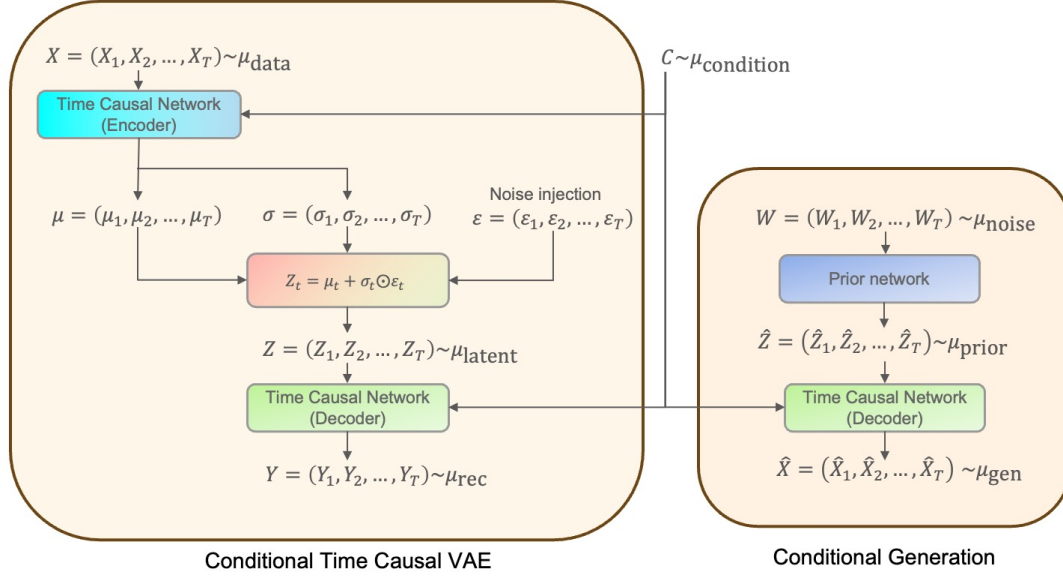


Figure 11: Conditional time-causal variational autoencoder and generation

With the *conditional TC-VAE*, we can generate the distribution of future paths conditional on historical paths. In financial modelling, path dependent models have shown to be able to successfully capture the price dynamics. For example, consider the path dependent volatility model where the prices  $(S_t)_{t \geq 0}$  satisfy

$$dS_t = S_t \sigma(S_{\leq t}) dW_t,$$

where  $\sigma$  is measurable and  $(W_t)_{t \geq 0}$  is a Wiener process. Given the time-homogeneous dynamic, for every  $\tau > 0$ , the law of  $S_{[t, t+\tau]}$  conditional on  $\sigma(S_{\leq t})$  is the same for all  $t \geq 0$ . As an example, we consider the 4-factor Markovian path dependent volatility model (PDV4) introduced in [GL23], where the volatility function is constructed by exponential kernels to produce the Markovian model  $(S_t^{\text{PDV}})_{t \geq 0}$  s.t.

$$\begin{cases} \frac{dS_t^{\text{PDV}}}{S_t^{\text{PDV}}} = \mu dt + \sigma_t dW_t, & \sigma_t = \sigma(R_{1,t}, R_{2,t}), \quad \sigma(R_1, R_2) = \beta_0 + \beta_1 R_1 + \beta_2 \sqrt{R_2}, \\ R_{1,t} = (1 - \theta_1)R_{1,1,t} + \theta_1 R_{1,2,t}, & R_{2,t} = (1 - \theta_2)R_{2,1,t} + \theta_2 R_{2,2,t}, \\ dR_{1,j,t} = \lambda_{1,j} (\sigma_t dW_t - R_{1,j,t} dt), & dR_{2,j,t} = \lambda_{2,j} (\sigma_t^2 dW_t - R_{1,j,t}), \quad j = 1, 2 \end{cases}$$

where  $\mu = 0.1, \beta_0 = 0.04, \beta_1 = -0.13, \beta_2 = 0.65, \lambda_{1,1} = 55, \lambda_{1,2} = 10, \theta_1 = 0.25, \lambda_{2,1} = 20, \lambda_{2,2} = 3, \theta_2 = 0.5$ . PDV4 captures important stylized facts of volatility, produces very realistic price and volatility paths (see Figure 12), and jointly fits SPX and VIX smiles remarkably well [GL23].

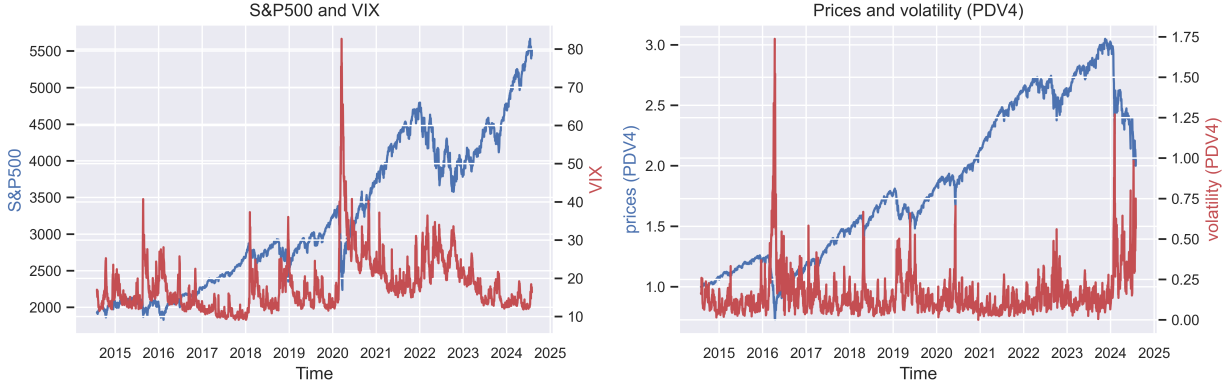


Figure 12: On the left, we plot daily S&P500 (blue) and VIX (red) from 2014-08-01 to 2024-08-01. On the right, we plot prices (blue) and volatility (red) of the 4-factor Markovian path dependent volatility model.

Since we work in discrete time, we let  $dt = 1/365$ ,  $N_T = 60$  to model daily prices. We choose  $N_{\text{sample}} = 2560$  to be the number of samples in market data. Let  $S_{t=1}^N$  be a discretized price path sampled from the PDV4 model. We extract the latest sub-paths  $S^{(i)} = S_{i:i+N_T}$ ,  $i \in \mathcal{I} = \{N - N_T - N_{\text{sample}}, \dots, N - N_T\}$ , and normalize them by dividing each sub-path by its starting price. This gives the return paths  $X^{(i)} = S_{i:i+N_T}/S_i$ ,  $i \in \mathcal{I}$ , which are our real paths. We denote the weighted historical volatility by

$$\Sigma^{(i)} = \sqrt{\sum_{j \leq i} K_2(i-j) \left(\frac{S_j - S_{j-1}}{S_{j-1}}\right)^2}, \quad i \in \mathcal{I},$$

where  $K_2(k) = Z_{\alpha, \delta}^{-1} (k + \delta)^{-\alpha}$ ,  $\alpha > 1$ ,  $\delta > 0$  and  $Z_{\alpha, \delta}^{-1}$  is chosen s.t.  $\sum_{k=0}^{\infty} K_2(k) = 1$ . Given the observed sample  $(X^{(i)}, \Sigma^{(i)})$ ,  $i \in \mathcal{I}$ , we apply the conditional TC-VAE to learn the distribution of future returns given the weighted historical volatility.

First, we visualize the fake paths under different conditions, see Figure 13. The generator indeed generates different distributions conditional on different conditions. Moreover, the generated paths show gain/loss asymmetry and volatility clustering, which are stylized facts of financial time series [Con01].

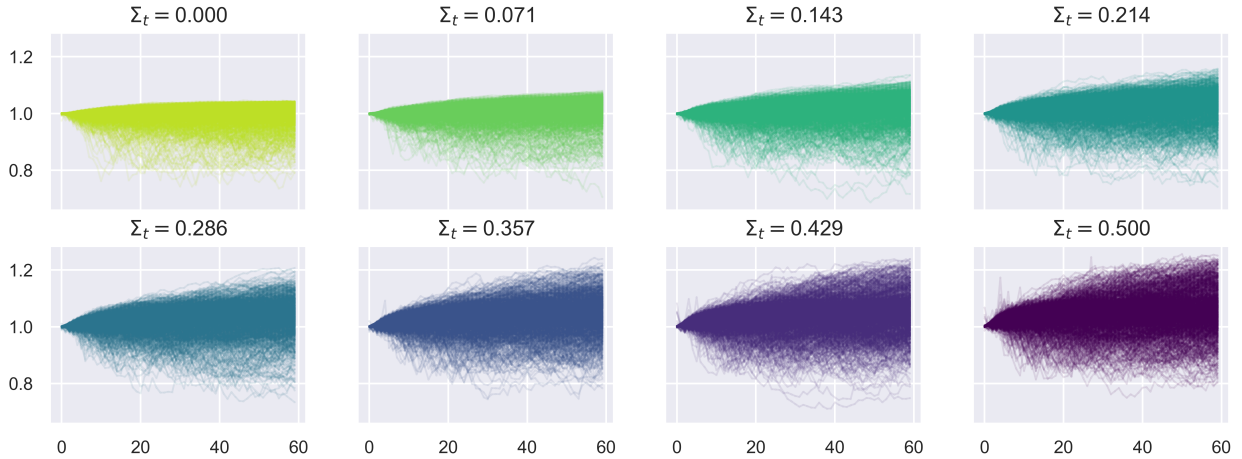


Figure 13: Visualization of generated returns conditional on weighted historical volatility

Then we compare fake paths and true paths distributions conditional on the same history path. To benchmark, we also sample a Black-Scholes distribution with drift and volatility estimated from the past

100 time steps. We quantitatively evaluate the conditional distributions under sliced Wasserstein distance, Gaussian MMD, signature MMD, and sliced adapted Wasserstein distance. For each historical path, we generate real, fake and control paths and compute distances between real vs real paths (with different random seed), real vs fake paths, and real vs control paths. For control paths, we use Black-Scholes paths with drift and volatility estimated from the historical paths, which serves as a benchmark. In Figure 14, we compare sliced Wasserstein distance, Gaussian MMD, signature MMD, and sliced adapted Wasserstein distance.

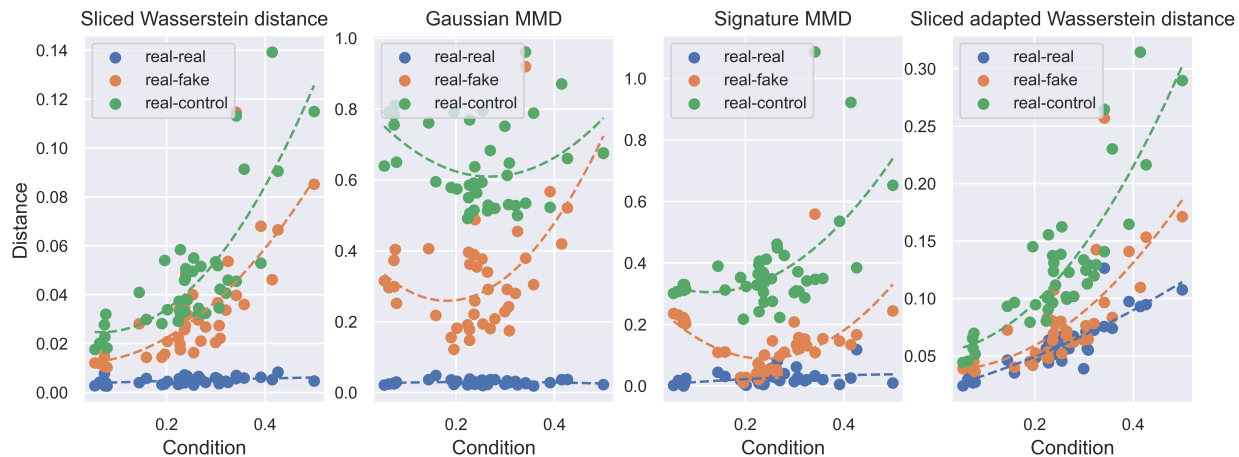


Figure 14: From left to right, we visualize the sliced Wasserstein distance, Gaussian MMD, signature MMD, sliced adapted Wasserstein distance. The dots are the distances between paths and the dash lines are quadratic polynomials fitted to the distances.

Notably, the path generation is not constrained by the length of training paths, which means that we can extend a path as long as desired. We iterate the following two steps: 1) extending the path by conditional generation; 2) calculating conditions of the extended path. See Figure 15 for path extension of  $600 = 10 * N_T$  time steps after 10 iterative path extensions. The generation shows long time stability without blowing up or vanishing.

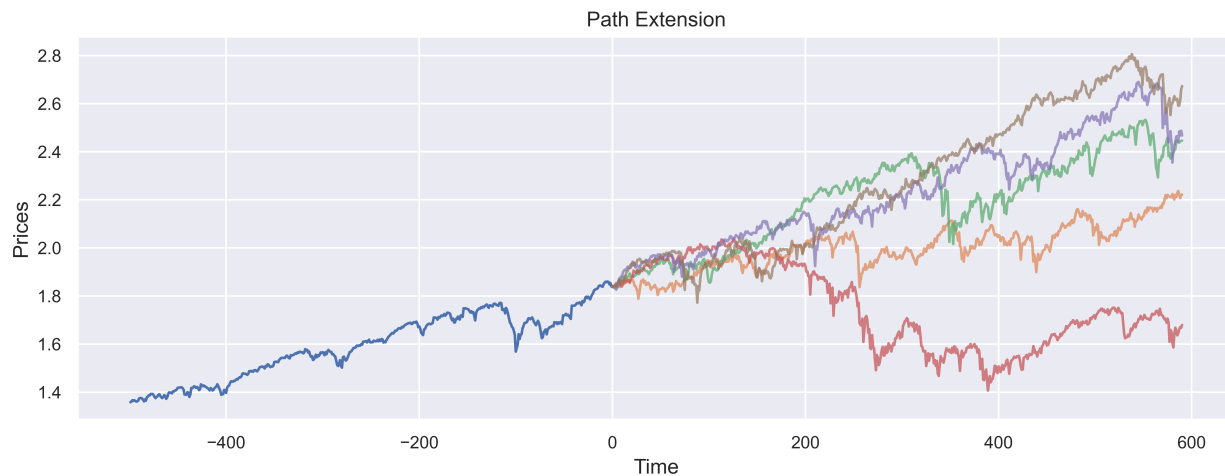


Figure 15: Extending a real path under the PDV4 model by generated conditional paths using TC-VAE.

## 4.2 Market data

### 4.2.1 S&P500 and VIX

Encouraged by the promising conditional generation from the path-dependent model considered above, we now apply our generator to data from S&P500 and VIX. We take daily S&P500 and VIX from 2014-08-01 to 2024-08-01, for a total of  $N = 2516$  trading days; see Figure 12. We denote by  $(S_t)_{t=1}^N$  the S&P500 path and by  $(V_t)_{t=1}^N$  the VIX path. As before, we consider subpaths of length  $N_T = 60$  and extract the sub-paths  $S^{(i)} = S_{i:i+N_T}$ ,  $i \in \mathcal{I} = \{1, \dots, N - N_T\}$  and normalize them by dividing each sub-path by its starting price. This gives the return paths  $X^{(i)} = S_{i:i+N_T}/S_i$ ,  $i \in \mathcal{I}$ , which are our real paths. As before, conditional on the VIX, we apply the conditional TC-VAE to learn S&P500 prices in the future. With the data and condition pair  $(X^{(i)}, \Sigma^{(i)})$ , we apply the conditional TC-VAE to learn the distribution of  $(X^{(i)})_{i \in \mathcal{I}}$  given  $(\Sigma^{(i)})_{i \in \mathcal{I}}$ .

First, we visualize the fake paths under different conditions, see Figure 16. The conditional generation shows skewness. Moreover, the volatility of generated paths is increasing with the VIX, which is in line with the financial interpretation of the condition.

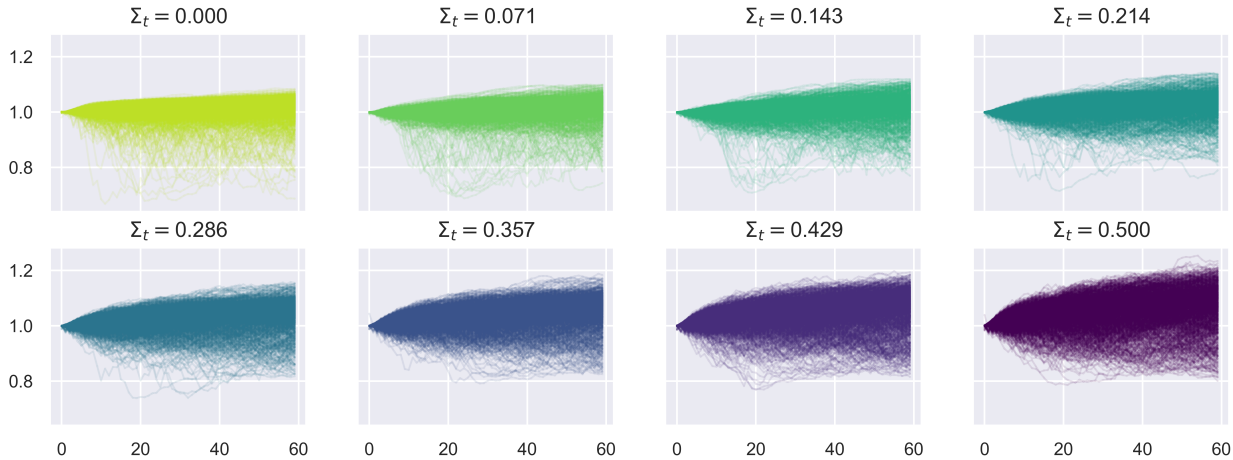


Figure 16: Visualization of generated returns conditional on VIX.

VIX can be well estimated by the weighted historical volatility, see [GL23] for a through analysis. Thus, similar to the PDV4 case, we can extend the path as long as desired. See Figure 17 for path extension of  $600 = 10 * N_T$  time steps after 10 iterative path extensions.





Figure 17: Path extension of normalized S&P500 prices using TC-VAE.

Finally, we generate a long fake path by path extension, and compare it with S&P 500 prices in terms of stylized facts of financial time series; see [Con01]. This includes: 1) heavy tail of returns, 2) volatility clustering, 3) zero auto-correlation of returns, 4) short-time auto-correlation of square returns, 5) long-time auto-correlation of absolute returns, and 6) negative skewness of returns. Returns of both S&P500 prices and fake prices display power-law or Pareto-like distribution; see Figure 18. The high-volatility returns of both S&P500 prices and fake prices tend to cluster, which is known as volatility clustering.

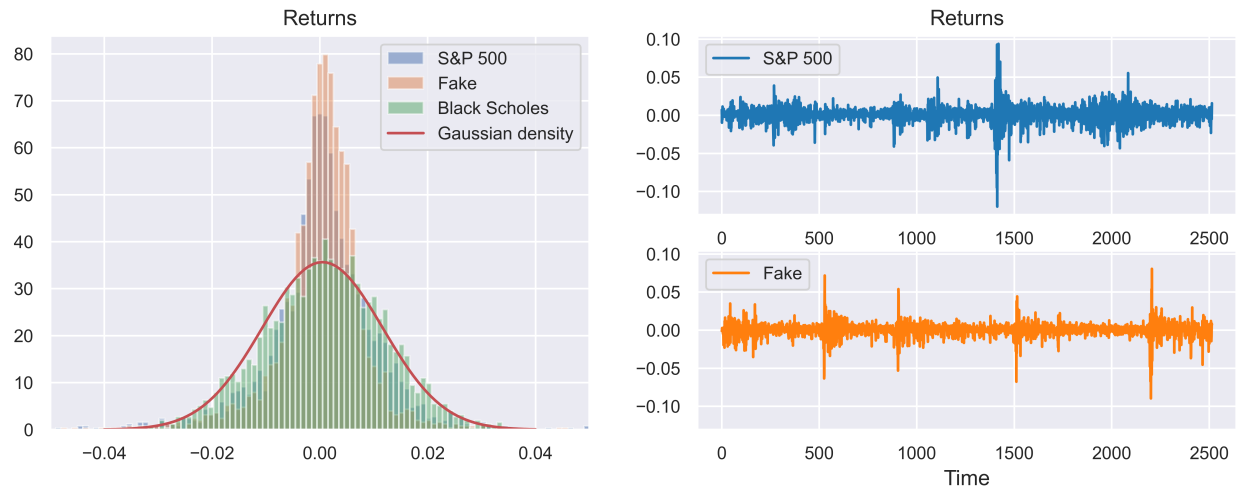


Figure 18: On the left, we visualize returns histogram of S&P 500 (blue), fake prices (orange), Black Scholes prices (green), and the Gaussian density (red). On the right we plot returns of S&P 500 (blue) and fake prices (orange) along the time horizon.

Furthermore, we inspect the auto-correlation of returns. Returns of S&P500 prices and fake prices both show no correlation in returns, short time correlation in square returns, and long time correlation in absolute returns; see Figure 19.

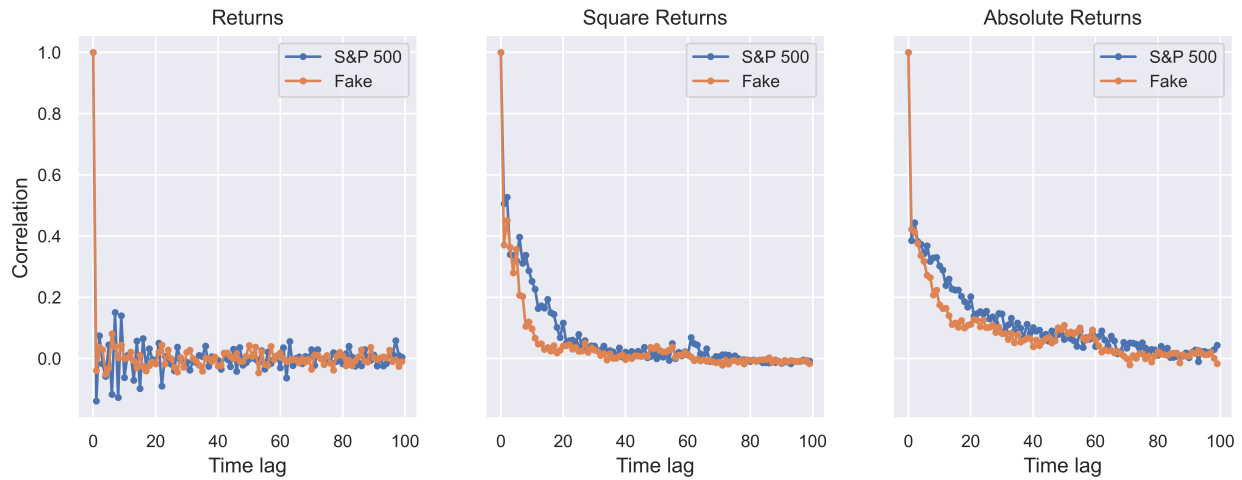


Figure 19: From left to right, we visualize the auto-correlation of returns, square returns, and absolute returns for both S&P 500 prices (blue) and fake prices (orange).

Lastly, we compare the skewness and kurtosis of returns for both S&P500 from 2014-08-01 to 2024-08-01 and 1000 fake paths; see Figure 20. Overall, S&P 500 prices and fake prices are close in skewness and kurtosis of returns.

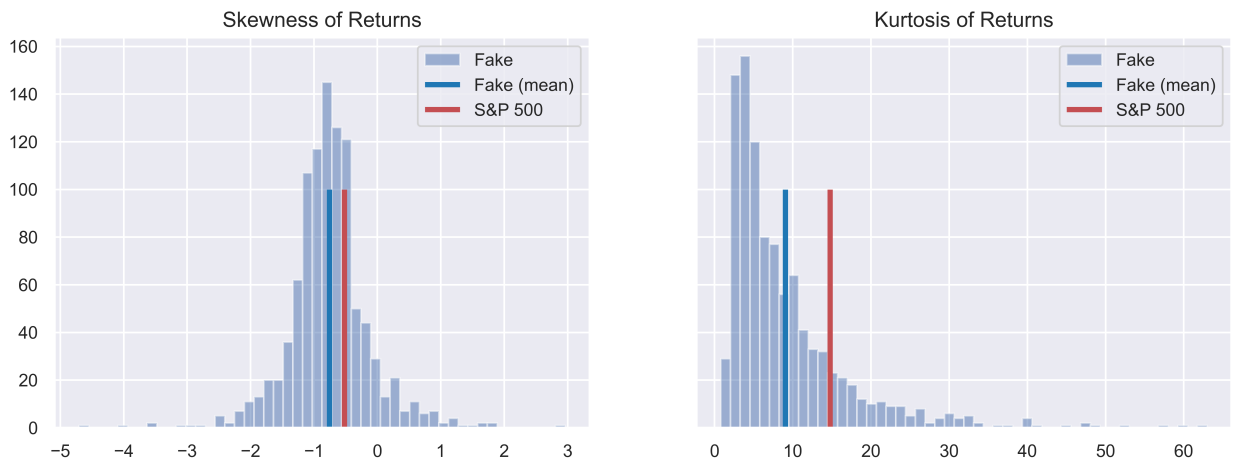


Figure 20: On the left, we plot the histogram of skewness of returns for 1000 fake paths, their mean (blue), and the skewness of returns for S&P 500. On the right, we plot the histogram of kurtosis of returns for 1000 fake paths, their mean (blue), and the kurtosis of returns for S&P 500.

## References

- [ABZ20] Beatrice Acciaio, Julio Backhoff-Veraguas, and Anastasiia Zalashko. “Causal optimal transport and its links to enlargement of filtrations and continuous-time stochastic optimization”. In: *Stochastic Processes and their Applications* 130.5 (2020), pp. 2918–2953.
- [AH24] Beatrice Acciaio and Songyan Hou. “Convergence of adapted empirical measures on  $\mathbb{R}^d$ ”. In: *Ann. Appl. Probab.* 34.5 (2024), pp. 4799–4835. ISSN: 1050-5164. DOI: 10.1214/24-AAP2082.

- [ACB17] Martin Arjovsky, Soumith Chintala, and Léon Bottou. “Wasserstein Generative Adversarial Networks”. In: *Proceedings of the 34th International Conference on Machine Learning*. Ed. by Doina Precup and Yee Whye Teh. Vol. 70. PMLR. Aug. 2017, pp. 214–223.
- [Ass+20a] 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*. 2020, pp. 1–8.
- [Ass+20b] 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*. 2020, pp. 1–8.
- [Bac+20] Julio Backhoff-Veraguas, Daniel Bartl, Mathias Beiglböck, and Manu Eder. “Adapted Wasserstein distances and stability in mathematical finance”. In: *Finance and Stochastics* 24 (2020), pp. 601–632.
- [Bac+22] Julio Backhoff-Veraguas, Daniel Bartl, Mathias Beiglböck, and Johannes Wiesel. “Estimating processes in adapted Wasserstein distance”. In: *The Annals of Applied Probability* 32.1 (2022), pp. 529–550.
- [Bac+17] Julio Backhoff-Veraguas, Mathias Beiglbock, Yiqing Lin, and Anastasiia Zalashko. “Causal transport in discrete time and applications”. In: *SIAM Journal on Optimization* 27.4 (2017), pp. 2528–2562.
- [Bai+16] David H Bailey, Jonathan Borwein, Marcos Lopez de Prado, and Qiji Jim Zhu. “The probability of backtest overfitting”. In: *Journal of Computational Finance, forthcoming* (2016).
- [BCJ19] Sebastian Becker, Patrick Cheridito, and Arnulf Jentzen. “Deep optimal stopping”. In: *The Journal of Machine Learning Research* 20.1 (2019), pp. 2712–2736.
- [BCJ24] Cyril Bénézet, Ziteng Cheng, and Sebastian Jaimungal. “Learning conditional distributions on continuous spaces”. In: *arXiv preprint arXiv:2406.09375* (2024).
- [BSR24] Pratik Bhowal, Achint Soni, and Sirisha Rambhatla. “Why do Variational Autoencoders Really Promote Disentanglement?” In: *Forty-first International Conference on Machine Learning*. 2024.
- [BGW24] Francesca Biagini, Lukas Gonon, and Niklas Walter. “Universal randomised signatures for generative time series modelling”. In: *arXiv preprint arXiv:2406.10214* (2024).
- [BT19] Jocelyne Bion-Nadal and Denis Talay. “On a Wasserstein-type distance between solutions to stochastic differential equations”. In: *The Annals of Applied Probability* 29.3 (2019), pp. 1609–1639.
- [BSW98] Christopher M Bishop, Markus Svensén, and Christopher KI Williams. “GTM: The generative topographic mapping”. In: *Neural computation* 10.1 (1998), pp. 215–234.
- [BS73] Fischer Black and Myron Scholes. “The pricing of options and corporate liabilities”. In: *Journal of political economy* 81.3 (1973), pp. 637–654.
- [Bow+15] Samuel R Bowman, Luke Vilnis, Oriol Vinyals, Andrew M Dai, Rafal Jozefowicz, and Samy Bengio. “Generating sentences from a continuous space”. In: *arXiv preprint arXiv:1511.06349* (2015).
- [BY78] Pierre Brémaud and Marc Yor. “Changes of filtrations and of probability measures”. In: *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete* 45.4 (1978), pp. 269–295.
- [Büh+20] Hans Bühler, Blanka Horvath, Terry Lyons, Imanol Perez Arribas, and Ben Wood. “A data-driven market simulator for small data environments”. In: *ERN: Neural Networks & Related Topics (Topic)* (2020).
- [Bur+18] Christopher P Burgess, Irina Higgins, Arka Pal, Loic Matthey, Nick Watters, Guillaume Desjardins, and Alexander Lerchner. “Understanding disentangling in beta-VAE”. In: *arXiv preprint arXiv:1804.03599* (2018).

- [Cai+23] Borui Cai, Shuiqiao Yang, Longxiang Gao, and Yong Xiang. “Hybrid variational autoencoder for time series forecasting”. In: *Knowledge-Based Systems* 281 (2023), p. 111079.
- [Che+18] Ricky TQ Chen, Xuechen Li, Roger B Grosse, and David K Duvenaud. “Isolating sources of disentanglement in variational autoencoders”. In: *Advances in neural information processing systems* 31 (2018).
- [Che+16] Xi Chen, Diederik P Kingma, Tim Salimans, Yan Duan, Prafulla Dhariwal, John Schulman, Ilya Sutskever, and Pieter Abbeel. “Variational lossy autoencoder”. In: *arXiv preprint arXiv:1611.02731* (2016).
- [CO18] Ilya Chevyrev and Harald Oberhauser. “Signature moments to characterize laws of stochastic processes”. In: *arXiv preprint arXiv:1810.10971* (2018).
- [Chi20] Rewon Child. “Very deep vaes generalize autoregressive models and can outperform them on images”. In: *arXiv preprint arXiv:2011.10650* (2020).
- [Chu+15] Junyoung Chung, Kyle Kastner, Laurent Dinh, Kratarth Goel, Aaron C Courville, and Yoshua Bengio. “A recurrent latent variable model for sequential data”. In: *Advances in neural information processing systems* 28 (2015).
- [CS24] Lu Chung I and Julian Sester. “Generative model for financial time series trained with MMD using a signature kernel”. In: *arXiv preprint arXiv:2407.19848* (2024).
- [Cin+21] Lucas Pinheiro Cinelli, Matheus Araújo Marins, Eduardo Antonio Barros Da Silva, and Sérgio Lima Netto. *Variational methods for machine learning with applications to deep networks*. Vol. 15. Springer, 2021.
- [Col+23] Andrea Coletta, Joseph Jerome, Rahul Savani, and Svitlana Vyetrenko. “Conditional generators for limit order book environments: Explainability, challenges, and robustness”. In: *Proceedings of the Fourth ACM International Conference on AI in Finance*. 2023, pp. 27–35.
- [Col+21] Andrea Coletta, Matteo Prata, Michele Conti, Emanuele Mercanti, Novella Bartolini, Aymeric Moulin, Svitlana Vyetrenko, and Tucker Balch. “Towards realistic market simulations: a generative adversarial networks approach”. In: *Proceedings of the Second ACM International Conference on AI in Finance*. 2021, pp. 1–9.
- [Con01] Rama Cont. “Empirical properties of asset returns: stylized facts and statistical issues”. In: *Quantitative finance* 1.2 (2001), p. 223.
- [Con+23] Rama Cont, Mihai Cucuringu, Jonathan Kochems, and Felix Prezel. “Limit Order Book Simulation with Generative Adversarial Networks”. In: *Available at SSRN 4512356* (2023).
- [Con+22] Rama Cont, Mihai Cucuringu, Renyuan Xu, and Chao Zhang. “Tail-gan: Learning to simulate tail risk scenarios”. In: *arXiv preprint arXiv:2203.01664* (2022).
- [Des+21] Abhyuday Desai, Cynthia Freeman, Zuhui Wang, and Ian Beaver. “Timevae: A variational autoencoder for multivariate time series generation”. In: *arXiv preprint arXiv:2111.08095* (2021).
- [Des+19] Ishan Deshpande, Yuan-Ting Hu, Ruoyu Sun, Ayis Pyrros, Nasir Siddiqui, Sanmi Koyejo, Zhizhen Zhao, David Forsyth, and Alexander G Schwing. “Max-sliced wasserstein distance and its use for gans”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pp. 10648–10656.
- [Dil+16] Nat Dilokthanakul, Pedro AM Mediano, Marta Garnelo, Matthew CH Lee, Hugh Salimbeni, Kai Arulkumaran, and Murray Shanahan. “Deep unsupervised clustering with gaussian mixture variational autoencoders”. In: *arXiv preprint arXiv:1611.02648* (2016).
- [DSB16a] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. “Density estimation using real nvp”. In: *arXiv preprint arXiv:1605.08803* (2016).
- [DSB16b] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. “Density estimation using real nvp”. In: *arXiv preprint arXiv:1605.08803* (2016).

- [EO21] Florian Eckerli and Joerg Osterrieder. “Generative adversarial networks in finance: an overview”. In: *arXiv preprint arXiv:2106.06364* (2021).
- [EP22] Stephan Eckstein and Gudmund Pammer. “Computational methods for adapted optimal transport”. In: *arXiv preprint arXiv:2203.05005* (2022).
- [Efi+20] Dmitry Efimov, Di Xu, Luyang Kong, Alexey Nefedov, and Archana Anandakrishnan. “Using generative adversarial networks to synthesize artificial financial datasets”. In: *arXiv preprint arXiv:2002.02271* (2020).
- [Eri+24] Lars Ericson, Xuejun Zhu, Xusi Han, Rao Fu, Shuang Li, Steve Guo, and Ping Hu. “Deep Generative Modeling for Financial Time Series with Application in VaR: A Comparative Review”. In: *arXiv preprint arXiv:2401.10370* (2024).
- [EHR17] Cristobal Esteban, Stephanie Hyland, and Gunnar Rätsch. “Real-valued (Medical) Time Series Generation with Recurrent Conditional GANs”. In: *ArXiv* 1706.02633 (Aug. 2017).
- [Fla+21] Rémi Flamary, Nicolas Courty, Alexandre Gramfort, Mokhtar Z. Alaya, Aurélie Boisbunon, Stanislas Chambon, Laetitia Chapel, Adrien Corenflos, Kilian Fatras, Nemo Fournier, Léo Gautheron, Nathalie T.H. Gayraud, Hicham Janati, Alain Rakotomamonjy, Ievgen Redko, Antoine Rolet, Antony Schutz, Vivien Seguy, Danica J. Sutherland, Romain Tavenard, Alexander Tong, and Titouan Vayer. “POT: Python Optimal Transport”. In: *Journal of Machine Learning Research* 22.78 (2021), pp. 1–8.
- [FS11] Hans Föllmer and Alexander Schied. *Stochastic finance: an introduction in discrete time*. Walter de Gruyter, 2011.
- [FV22] Peter A Forsyth and Kenneth R Vetzal. “Multi-Period Mean Expected-Shortfall Strategies: ‘Cut Your Losses and Ride Your Gains’”. In: *Applied Mathematical Finance* (2022), pp. 1–37.
- [FHO22] Weilong Fu, Ali Hirsra, and Jörg Osterrieder. “Simulating financial time series using attention”. In: *ArXiv* 2207.00493 (July 2022).
- [GST20] Ioannis Gatopoulos, Maarten Stol, and Jakub M Tomczak. “Super-resolution variational autoencoders”. In: *arXiv preprint arXiv:2006.05218* (2020).
- [Gig08] Nicola Gigli. “On the geometry of the space of probability measures in  $\mathbb{R}^n$  endowed with the quadratic optimal transport distance”. PhD thesis. Thesis (Ph. D.)–Scuola Normale Superiore, 2008.
- [Gir+20] Laurent Girin, Simon Leglaive, Xiaoyu Bie, Julien Diard, Thomas Hueber, and Xavier Alamedd-Pineda. “Dynamical variational autoencoders: A comprehensive review”. In: *arXiv preprint arXiv:2008.12595* (2020).
- [GPP19] Martin Glanzer, Georg Ch Pflug, and Alois Pichler. “Incorporating statistical model error into the calculation of acceptability prices of contingent claims”. In: *Mathematical Programming* 174.1 (2019), pp. 499–524.
- [Goo+14] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. “Generative Adversarial Nets”. In: *Advances in Neural Information Processing Systems* 27 (2014). Ed. by Z Ghahramani, M Welling, C Cortes, N Lawrence, and K Q Weinberger.
- [Gre+12] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. “A kernel two-sample test”. In: *The Journal of Machine Learning Research* 13.1 (2012), pp. 723–773.
- [GL23] Julien Guyon and Jordan Lekeufack. “Volatility is (mostly) path-dependent”. In: *Quantitative Finance* 23.9 (2023), pp. 1221–1258.
- [HHP23] Mohamed Hamdouche, Pierre Henry-Labordere, and Huyen Pham. “Generative Modeling for Time Series Via Schrödinger Bridge”. In: *SSRN Electronic Journal* (2023). ISSN: 1556-5068.

- [Hig+16] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. “beta-vae: Learning basic visual concepts with a constrained variational framework”. In: *International conference on learning representations*. 2016.
- [Hou24] Songyan Hou. “Convergence of the Adapted Smoothed Empirical Measures”. In: *arXiv preprint arXiv:2401.14883* (2024).
- [HCQ24] Hongbin Huang, Minghua Chen, and Xiao Qiao. “Generative Learning for Financial Time Series with Irregular and Scale-Invariant Patterns”. In: *The Twelfth International Conference on Learning Representations*. 2024.
- [Hul21] Hanna Hultin. “Generative models of limit order books”. PhD thesis. KTH Royal Institute of Technology, 2021.
- [Hul+23] Hanna Hultin, Henrik Hult, Alexandre Proutiere, Samuel Samama, and Ala Tarighati. “A generative model of a limit order book using recurrent neural networks”. In: *Quantitative Finance* 23.6 (2023), pp. 931–958.
- [Igl+23] Guillermo Iglesias, Edgar Talavera, Ángel González-Prieto, Alberto Mozo, and Sandra Gómez-Canaval. “Data augmentation techniques in time series domain: a survey and taxonomy”. In: *Neural Computing and Applications* 35.14 (2023), pp. 10123–10145.
- [Iss+24] Zacharia Issa, Blanka Horvath, Maud Lemerrier, and Cristopher Salvi. “Non-adversarial training of Neural SDEs with signature kernel scores”. In: *Advances in Neural Information Processing Systems* 36 (2024).
- [Kar+20] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. “Training generative adversarial networks with limited data”. In: *Advances in neural information processing systems* 33 (2020), pp. 12104–12114.
- [Kid21] P. Kidger. *On neural differential equations (PhD Thesis)*. 2021.
- [Kid+21] 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.
- [KW14] Diederik P Kingma and Max Welling. “Auto-encoding variational bayes”. In: *2nd International Conference on Learning Representations, ICLR 2014* (2014).
- [Kol+19] Soheil Kolouri, Kimia Nadjahi, Umut Simsekli, Roland Badeau, and Gustavo Rohde. “Generalized sliced wasserstein distances”. In: *Advances in neural information processing systems* 32 (2019).
- [KS19] Alex Kondratyev and Christian Schwarz. “The Market Generator”. In: *Econometrics: Econometric & Statistical Methods - Special Topics eJournal* (2019).
- [KFT21] Adriano Koshiyama, Nick Firoozye, and Philip Treleaven. “Generative adversarial networks for financial trading strategies fine-tuning and combination”. In: *Quantitative Finance* 21 (5 May 2021), pp. 797–813.
- [Las18] Rémi Lassalle. “Causal transference plans and their Monge-Kantorovich problems”. In: *Stochastic Processes and their Applications* 36.3 (2018), pp. 452–484.
- [Li+17] Chun-Liang Li, Wei-Cheng Chang, Yu Cheng, Yiming Yang, and Barnabás Póczos. “Mmd gan: Towards deeper understanding of moment matching network”. In: *Advances in neural information processing systems* 30 (2017).
- [LN00] Duan Li and Wan-Lung Ng. “Optimal dynamic portfolio selection: Multiperiod mean-variance formulation”. In: *Mathematical finance* 10.3 (2000), pp. 387–406.
- [Li+20] Junyi Li, Xintong Wang, Yaoyang Lin, Arunesh Sinha, and Michael Wellman. “Generating realistic stock market order streams”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34. 01. 2020, pp. 727–734.

- [Lia+24] Shujian Liao, Hao Ni, Marc Sabate-Vidales, Lukasz Szpruch, Magnus Wiese, and Baoren Xiao. “Sig-Wasserstein GANs for conditional time series generation”. In: *Mathematical Finance* 34.2 (2024), pp. 622–670.
- [Liu+22] Yuansan Liu, Sudanthi Wijewickrema, Ang Li, and James Bailey. “Time-Transformer AAE: Connecting Temporal Convolutional Networks and Transformer for Time Series Generation”. In: (2022).
- [LLN24] Hang Lou, Siran Li, and Hao Ni. “PCF-GAN: generating sequential data via the characteristic function of measures on the path space”. In: *Advances in Neural Information Processing Systems* 36 (2024).
- [Lu+23] Yingzhou Lu, Minjie Shen, Huazheng Wang, Xiao Wang, Capucine van Rechem, and Wenqi Wei. “Machine learning for synthetic data generation: a review”. In: *arXiv preprint arXiv:2302.04062* (2023).
- [Mat+19] Emile Mathieu, Tom Rainforth, Nana Siddharth, and Yee Whye Teh. “Disentangling disentanglement in variational autoencoders”. In: *International conference on machine learning*. PMLR, 2019, pp. 4402–4412.
- [Mee19] Fernando de Meer Pardo. “Enriching financial datasets with generative adversarial networks”. In: *MS thesis, Delft University of Technology, The Netherlands* (2019).
- [Mer75] Robert C Merton. “Optimum consumption and portfolio rules in a continuous-time model”. In: *Stochastic optimization models in finance*. Elsevier, 1975, pp. 621–661.
- [Nag+23] Peer Nagy, Sascha Frey, Silvia Saporá, Kang Li, Anisoara Calinescu, Stefan Zohren, and Jakob Foerster. “Generative ai for end-to-end limit order book modelling: A token-level autoregressive generative model of message flow using a deep state space network”. In: *Proceedings of the Fourth ACM International Conference on AI in Finance*. 2023, pp. 91–99.
- [Ni+21] Hao Ni, Lukasz Szpruch, Marc Sabate-Vidales, Baoren Xiao, Magnus Wiese, and Shujian Liao. “Sig-Wasserstein GANs for time series generation”. In: *Proceedings of the Second ACM International Conference on AI in Finance*. 2021, pp. 1–8.
- [Nie+22] Sloan Nietert, Ziv Goldfeld, Ritwik Sadhu, and Kengo Kato. “Statistical, robustness, and computational guarantees for sliced wasserstein distances”. In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 28179–28193.
- [Nut21] Marcel Nutz. “Introduction to entropic optimal transport”. In: *Lecture notes, Columbia University* (2021).
- [Özy21] Muhammed Imran Özyar. “Learning the Limit Order Book: a comprehensive mix between stochastic and machine learning models for generation and prediction”. In: (2021).
- [Pam24] Gudmund Pammer. “A note on the adapted weak topology in discrete time”. In: *Electronic Communications in Probability* 29 (2024), pp. 1–13.
- [PP12] Georg Ch Pflug and Alois Pichler. “A distance for multistage stochastic optimization models”. In: *SIAM Journal on Optimization* 22.1 (2012), pp. 1–23.
- [PP14] Georg Ch Pflug and Alois Pichler. *Multistage stochastic optimization*. Vol. 1104. Springer, 2014.
- [PP15] Georg Ch Pflug and Alois Pichler. “Dynamic generation of scenario trees”. In: *Computational Optimization and Applications* 62.3 (2015), pp. 641–668.
- [PP16] Georg Ch Pflug and Alois Pichler. “From empirical observations to tree models for stochastic optimization: convergence properties”. In: *SIAM Journal on Optimization* 26.3 (2016), pp. 1715–1740.
- [Pic13] Alois Pichler. “Evaluations of risk measures for different probability measures”. In: *SIAM Journal on Optimization* 23.1 (2013), pp. 530–551.
- [PW22] Alois Pichler and Michael Weinhardt. “The nested Sinkhorn divergence to learn the nested distance”. In: *Computational Management Science* 19.2 (2022), pp. 269–293.

- [Rac+13] Svetlozar T Rachev, Lev B Klebanov, Stoyan V Stoyanov, and Frank Fabozzi. *The methods of distances in the theory of probability and statistics*. Vol. 10. Springer, 2013.
- [RV18] Danilo Jimenez Rezende and Fabio Viola. “Taming vaes”. In: *arXiv preprint arXiv:1810.00597* (2018).
- [Riz+23] Matteo Rizzato, Julien Wallart, Christophe Geissler, Nicolas Morizet, and Nouredine Boumlaik. “Generative Adversarial Networks applied to synthetic financial scenarios generation”. In: *Physica A: Statistical Mechanics and its Applications* 623 (2023), p. 128899.
- [Rub02] Mark Rubinstein. “Markowitz’s” portfolio selection”: A fifty-year retrospective”. In: *The Journal of finance* 57.3 (2002), pp. 1041–1045.
- [Rüs85] Ludger Rüschendorf. “The Wasserstein distance and approximation theorems”. In: *Probability Theory and Related Fields* 70.1 (1985), pp. 117–129.
- [Sch24] Christian Schwarz. “Interpretable GenAI: Synthetic Financial Time Series Generation with Probabilistic LSTM”. In: *Available at SSRN 4877007* (2024).
- [Sta+21] Jan Stanczuk, Christian Etmann, Lisa Maria Kreusser, and Carola-Bibiane Schönlieb. “Wasserstein GANs work because they fail (to approximate the Wasserstein distance)”. In: *arXiv preprint arXiv:2103.01678* (2021).
- [TCT19] Shuntaro Takahashi, Yu Chen, and Kumiko Tanaka-Ishii. “Modeling financial time-series with generative adversarial networks”. In: *Physica A: Statistical Mechanics and its Applications* 527 (Aug. 2019), p. 121261.
- [Tak05] Kunio Takezawa. *Introduction to nonparametric regression*. John Wiley & Sons, 2005.
- [TT20] Hoang Thanh-Tung and Truyen Tran. “Catastrophic forgetting and mode collapse in GANs”. In: *2020 international joint conference on neural networks (ijcnn)*. IEEE. 2020, pp. 1–10.
- [TW18] Jakub Tomczak and Max Welling. “VAE with a VampPrior”. In: *International conference on artificial intelligence and statistics*. PMLR. 2018, pp. 1214–1223.
- [VPC24] Milena Vuletić, Felix Prenzel, and Mihai Cucuringu. “Fin-gan: Forecasting and classifying financial time series via generative adversarial networks”. In: *Quantitative Finance* 24.2 (2024), pp. 175–199.
- [Wie+20] Magnus Wiese, Robert Knobloch, Ralf Korn, and Peter Kretschmer. “Quant GANs: deep generation of financial time series”. In: *Quantitative Finance* 20 (9 Sept. 2020), pp. 1419–1440.
- [Wie+21] Magnus Wiese, Ben Wood, Alexandre Pachoud, Ralf Korn, Hans Buehler, Murray Phillip, and Lianjun Bai. “Multi-Asset Spot and Option Market Simulation”. In: *SSRN Electronic Journal* (2021).
- [Xu+20] Tianlin Xu, Li Kevin Wenliang, Michael Munn, and Beatrice Acciaio. “Cot-gan: Generating sequential data via causal optimal transport”. In: *Advances in neural information processing systems* 33 (2020), pp. 8798–8809.
- [Yan+21] Xu Yang, Hanwang Zhang, Guojun Qi, and Jianfei Cai. “Causal attention for vision-language tasks”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2021, pp. 9847–9857.
- [YJS19] Jinsung Yoon, Daniel Jarrett, and Mihaela van der Schaar. “Time-series Generative Adversarial Networks”. In: *Advances in Neural Information Processing Systems* 32 (2019). Ed. by H Wallach, H Larochelle, A Beygelzimer, F d Alché-Buc, E Fox, and R Garnett.