Nonparametric generative modeling for time series via Schrödinger bridge

Huyên PHAM

Université Paris Cité (UPC), LPSM





based on joint work with Mohamed HAMDOUCHE (UPC, Qube RT) and Pierre HENRY-LABORDERE (Qube RT)

> INTERNATIONAL SEMINAR ON SDES AND RELATED TOPICS November, 24, 2023

Generative modeling (for time series)

- Given datasets from an (unknown) distribution μ (target) of a time series, e.g.
 - Medical data of a patient
 - Renewable energy production
 - Finance: asset price, volatility surface, ...
- ▶ The goal is to
 - $\bullet \ {\rm learn/estimate} \ \mu$
 - generate new samples of μ :
 - Useful for improving clinical predictions, weather forecast
 - Financial industry: market stress test, market risk measurement, deep hedging, reinforcement learning for optimal trading



Generative modeling techniques

• Generative modeling (GM) has become a classical task in AI and machine learning used notably in image processing with spectacular success (DALL-E, Midjourney, Stable diffusion, etc), and controversies!

- Several competing methods:
 - Likelihood-based models (2011-): energy-based models (EBM), variational auto-encoders (VAE), normalizing flow models, etc
 - Implicit generative models (2014-): generative adversarial network (GAN)
 - Score-based diffusion models (2020-): emergent class of generative AI models that achieved state-of-the-art performance by outperforming GANs.



but mostly for static data/image.

Challenges of GM for time series

- Temporal setting (sequential data) poses new challenges to GM:
 - not enough to learn the time marginals
 - learn the joint distribution without exploiting the sequential structure is also not sufficient
 - capture the potentially complex dynamics of variables across time: conditional distribution over time

State-of-the-art generative methods for time series

- Time series GAN (Yoon et al. 19): combination of an *unsupervised adversarial* loss on real/synthetic data and *supervised* loss for generating sequential data
- \bullet Quant GAN (Wiese et al. 20): adversarial generator using temporal convolutional networks
- Causal optimal transport GAN (Xu et al. 20): adversarial generator using the adapted Wasserstein distance for processes
- PCF-GAN (Lou et al. 23): Path characteristic function into GAN
- Neural SDEs: SDE representation of time series with parametric (e.g. NN) coefficients to be trained for fitting with real samples (Remlinger et al. 21, Kidger et al. 21)
- Signature embedding of time series: Fermanian (19), Ni et al. (20), Buehler et al. (20).
- \blacktriangleright Most of these GM are parametric and require the training of NN

State-of-the-art generative methods for time series

- Time series GAN (Yoon et al. 19): combination of an *unsupervised adversarial* loss on real/synthetic data and *supervised* loss for generating sequential data
- \bullet Quant GAN (Wiese et al. 20): adversarial generator using temporal convolutional networks
- Causal optimal transport GAN (Xu et al. 20): adversarial generator using the adapted Wasserstein distance for processes
- PCF-GAN (Lou et al. 23): Path characteristic function into GAN
- Neural SDEs: SDE representation of time series with parametric (e.g. NN) coefficients to be trained for fitting with real samples (Remlinger et al. 21, Kidger et al. 21)
- Signature embedding of time series: Fermanian (19), Ni et al. (20), Buehler et al. (20).
- ▶ Most of these GM are parametric and require the training of NN
- ► We propose here a nonparametric generative model based on Schrödinger bridge, relying on diffusion, but in contrast with score-based diffusion models, it is over a finite horizon without time reversal, and adapted for time series.

Outline



Numerical experiments with applications

Reminder on the (classical) Schrödinger bridge (SB) problem

• Entropy optimal transport problem of Schrödinger (1932), see survey in Léonard (14): Given:

- reference measure on path spaces (e.g. Wiener \mathbb{W}) over a finite horizon \mathcal{T}
- two distributions μ , ν (e.g. data and prior)

find the closest probability measure \mathbb{P} to the reference w.r.t. Kullback-Leibler divergence, i.e. relative entropy, which admits as marginals: μ at time 0 and ν at time T.

► Stochastic control formulation by Girsanov's theorem (Dai Pra 1991, Chen et al. 20) Minimize over control process α

$$\mathbb{E}\Big[rac{1}{2}\int_0^{\mathcal{T}} |lpha_t|^2 \mathrm{d}t\Big] \qquad (ext{equal to } \operatorname{KL}(\mathbb{P},\mathbb{W}) \ := \ \int \log rac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{W}} \mathrm{d}\mathbb{P} \)$$

subject to

$$\mathrm{d} X_t \quad = \quad \alpha_t \mathrm{d} t + \mathrm{d} W_t, \ \ 0 \leq t \leq T, \qquad X_0 \sim \mu, \quad X_T \sim \nu.$$

Application of SB to generative modeling

The optimal drift of the SB problem is in feedback form: $\alpha_t^* = a^*(t, X_t)$ with a^* characterized in terms of a Schrödinger system, and the solution can be solved numerically by

- Iterative Proportional Fitting (IPF), a.k.a. Sinkhorn algorithm
- Score-based matching: refinement of IPF

 \rightarrow Generative model for sampling μ_{data} : recent works by De Bortoli et al. (21-), Vargas et al. (21), Wang et al. (21)

Schrödinger bridge time series problem

Let $\mu \in \mathcal{P}((\mathbb{R}^d)^N)$ be the data time series distribution of some \mathbb{R}^d -valued process observed at N dates: target time series measure.

• Entropic interpolation of μ : : Find a diffusion process X on \mathbb{R}^d satisfying

$$\mathrm{d} X_t \quad = \quad \alpha_t \mathrm{d} t + \mathrm{d} W_t, \quad 0 \leq t \leq T, \quad X_0 = 0,$$

with a controlled drift α minimizing

$$J(lpha) \ := \ \mathbb{E}\Big[rac{1}{2}\int_0^T |lpha_t|^2 \mathrm{d}t\Big]$$

and such that $(X_{t_1}, \ldots, X_{t_N}) \sim \mu$ (perfect match of the target time series measure), for some time grid $t_0 = 0 < \ldots < t_i < \ldots < t_N = T$.

Remark: the time grid $\mathcal{T} = \{t_i : i \in [\![1, N]\!]\}$ for the interpolation of the Schrödinger diffusion may be different from the observed times of the time series.

Assumptions

Assume that μ admits a density w.r.t. Lebesgue measure on $(\mathbb{R}^d)^N$, denoted by misuse of notation: $\mu(x_1, \ldots, x_N)$.

Denote by μ_T^W the distribution of Brownian motion W on \mathcal{T} , i.e. of $(W_{t_1}, \ldots, W_{t_N})$, hence with density:

$$\mu_{\mathcal{T}}^{W}(x_{1},\ldots,x_{N}) = \prod_{i=0}^{N-1} \frac{1}{\sqrt{2\pi(t_{i+1}-t_{i})}} \exp\Big(-\frac{|x_{i+1}-x_{i}|^{2}}{2(t_{i+1}-t_{i})}\Big).$$

• We assume that the relative entropy of μ w.r.t. $\mu_{\mathcal{T}}^{\mathsf{W}}$ is finite, i.e.

$$(\mathbf{H}) \qquad \qquad \mathrm{KL}(\mu|\mu_{\mathcal{T}}^{W}) \ := \ \int \log \frac{\mu}{\mu_{\mathcal{T}}^{W}} \mathrm{d}\mu \quad < \ \infty.$$

Remark: Assumption (H) is satisfied whenever μ comes from a process with

- Gaussian noise
- Heavy-tailed distribution but with second moment

Solution to Schrödinger bridge time series (SBTS)

Theorem (Diffusion SBTS)

Under (H), the optimal controlled drift of the SBTS problem is in the **path-dependent** form:

$$lpha^*_t = \mathrm{a}^*(t, X_t; oldsymbol{X}_{t_i}), \quad t_i \leq t < t_{i+1}, \quad i = 0, \dots, N-1,$$

where we set $\boldsymbol{X}_{t_i} := (X_{t_1}, \ldots, X_{t_i})$, and

$$\mathbf{a}^*(t, x; \mathbf{x}_i) = \nabla_x \log \mathbb{E}_{\mathbb{W}}\Big[\frac{\mu}{\mu_T^W}(X_{t_1}, \dots, X_{t_N}) \big| \mathbf{X}_{t_i} = \mathbf{x}_i, X_t = x\Big],$$

for $x_i = (x_1, \ldots, x_i) \in (\mathbb{R}^d)^i$, $x \in \mathbb{R}^d$. Here $\mathbb{E}_{\mathbb{W}}$ denotes the expectation under which X is a Brownian motion by Girsanov's theorem.

 \rightarrow By construction, the diffusion (called SBTS) process

$$\mathrm{d} X_t \quad = \quad \mathrm{a}^*(t, X_t; (X_{t_i})_{t_i \leq t}) \mathrm{d} t + \mathrm{d} W_t, \quad X_0 = 0,$$

satisfies $(X_{t_1}, \ldots, X_{t_N}) \sim \mu$.

Application to generative modeling

- Choice of the time grid $\mathcal{T} = \{t_i : i \in \llbracket 1, N \rrbracket\}$, $\Delta t_i = t_{i+1} t_i$.
 - When d = 1: calibrate Δt_i to the (empirical) variance of μ over [t_i, t_{i+1}] (time-changed Brownian motion):

$$\Delta t_i = \operatorname{Var}_{\mu}(X_{i+1} - X_i).$$

- For d > 1: normalize each component of the random vector of the time series by its Std, and then use $\Delta t_i = 1$.
- Estimate/learn the Schrödinger drift from samples of μ , see next slides
- \bullet Simulate e.g. by Euler scheme the SBTS diffusion \rightarrow
 - New samples of μ with realizations of $(X_{t_1}, \ldots, X_{t_N})$
 - Prediction by computing conditional law of X_{ti+1} | X_{ti}

Schrödinger drift function

From the above theorem and Gaussian properties of BM, the drift function is given by

$$\mathbf{a}^{*}(t,x;\boldsymbol{x}_{i}) = \frac{\nabla_{x}h_{i}(t,x;\boldsymbol{x}_{i})}{h_{i}(t,x;\boldsymbol{x}_{i})}, \quad t \in [t_{i},t_{i+1}), \boldsymbol{x}_{i} \in (\mathbb{R}^{d})^{i}, x \in \mathbb{R}^{d},$$
(1)

with

$$\begin{aligned} h_i(t,x; \mathbf{x}_i) &= \mathbb{E}_{\mathbf{Y}^i \sim \mathcal{N}(0, I_{d \times (N-i)})} \Big[\rho(\mathbf{x}_i, x + \sqrt{t_{i+1} - t} \mathbf{Y}_{i+1}, \dots, \\ & x + \sqrt{t_{i+1} - t} \mathbf{Y}_{i+1} + \sum_{j=i+1}^{N-1} \sqrt{t_{j+1} - t_j} \mathbf{Y}_{j+1}) \Big], \end{aligned}$$

with $\rho := \frac{\mu}{\mu_{\tau}^{W}}$ the density ratio, and $\mathbf{Y}^{i} = (Y_{i+1}, \dots, Y_{N}) \sim \mathcal{N}(0, \mathbb{I}_{N-i}).$

 \rightarrow Following method in Wang et al. (21), one can first derive an estimator $\hat{\rho}$ of ρ by logistic regression, and then get an estimator of the drift function by plugging into (1) and computing the expectation with Monte-Carlo.

▶ But this method is costly as it requires to sample from $\mu_{\mathcal{T}}^W$, and then use Monte-Carlo expectation in the drift expression (1) by sampling from $\mathbf{Y}^i \sim \mathcal{N}(0, \mathbb{I}_{d \times (N-i)})$.

Alternate expression of the Schrödinger drift function

Using Bayes formula, we derive the following expression:

$$a^{*}(t, x; \boldsymbol{x}_{i}) = \frac{1}{t_{i+1} - t} \frac{\mathbb{E}_{\mu} \left[(X_{t_{i+1}} - x) F_{i}(t, x_{i}, x, X_{t_{i+1}}) \big| \boldsymbol{X}_{t_{i}} = \boldsymbol{x}_{i} \right]}{\mathbb{E}_{\mu} \left[F_{i}(t, x_{i}, x, X_{t_{i+1}}) \big| \boldsymbol{X}_{t_{i}} = \boldsymbol{x}_{i} \right]},$$
(2)

for $t \in [t_i, t_{i+1})$, $i = 0, \dots, N-1$, $\mathbf{x}_i \in (\mathbb{R}^d)^i$, $x \in \mathbb{R}^d$, where

$$F_i(t, x_i, x, x_{i+1}) = \exp\left(-\frac{|x_{i+1} - x|^2}{2(t_{i+1} - t)} + \frac{|x_{i+1} - x_i|^2}{2(t_{i+1} - t_i)}\right).$$

Here $\mathbb{E}_{\mu}[\cdot|\cdot]$ is the (conditional) expectation under $\mu \to \text{One can then estimate the drift function by relying directly on samples of data distribution <math>\mu$.

Remark: When μ is the distribution arising from a Markov chain, then the conditional expectations in (2) (and so the drift function) will depend on the past values $X_{t_i} = (X_{t_1}, \ldots, X_{t_i})$ only via the last value X_{t_i} .

In practice, we can test the Markov property of μ , and see to what order we need to condition on the past.

Kernel estimation of the drift

• Approximate the conditional expectation under μ by **nonparametric regression** methods, e.g. kernel:

From data samples $X_N^{(m)} = (X_1^{(m)}, \dots, X_N^{(m)})$, $m = 1, \dots, M$ from μ , the Nadaraya-Watson estimator of the drift function in (2) is given by

$$\hat{a}(t,x;\boldsymbol{x}_{i}) = \frac{1}{t_{i+1}-t} \frac{\sum_{m=1}^{M} (X_{i+1}^{(m)}-x)F_{i}(t,X_{i}^{(m)},x,X_{i+1}^{(m)})\boldsymbol{\kappa}^{i}\left(\frac{\boldsymbol{X}_{i}^{(m)}-\boldsymbol{x}_{i}}{h}\right)}{\sum_{m=1}^{M} F_{i}(t,X_{i}^{(m)},x,X_{i+1}^{(m)})\boldsymbol{\kappa}^{i}\left(\frac{\boldsymbol{X}_{i}^{(m)}-\boldsymbol{x}_{i}}{h}\right)},$$

for $\mathbf{x}_i = (x_1, \dots, x_i)$, where \mathbf{K}^i is a kernel function on $(\mathbb{R}^d)^i$, e.g. in multiplicative form: $\mathbf{K}^i(\mathbf{x}_i) = \prod_{i=1}^i K(x_i)$, with K kernel function on \mathbb{R}^d , h > 0 is the bandwith parameter.

Remarks:

- Choice of kernel is not crucial: we take the quartic kernel $K(x) \propto (1-|x|^2)^2 \mathbb{1}_{|x|\leq 1}$
- Choice of bandwith h is more crucial: tradeoff between bias and variance.
- (Rate) of convergence of \hat{a} towards a^* under current investigation

SBTS Algorithm

 N_{π} : number of uniform steps in Euler scheme between two interpolation dates t_i , t_{i+1} :

$$t_{k,i}^{\pi} = t_i + k/N_{\pi}, \quad k = 0, \dots, N_{\pi} - 1.$$

Algorithm 1: SBTS Simulation

```
Input: data samples of time series (X_1^{(m)}, \dots, X_N^{(m)}), m = 1, \dots, M, and N_{\pi}.

Initialization: initial state x_0 = 0;

for i = 0, \dots, N - 1 do

Initialize state y_0 = x_i;

for k = 0, \dots, N_{\pi} - 1 do

Compute \hat{a}(t_{k,i}^{\pi}, y_k; x_i) by kernel estimator;

Sample \varepsilon_k \in \mathcal{N}(0, \mathbb{I}_d) and compute: y_{k+1} = y_k + \frac{1}{N_{\pi}} \hat{a}(t_{k,i}^{\pi}, y_k; x_i) + \frac{1}{\sqrt{N_{\pi}}} \varepsilon_k;

end

Set x_{i+1} = y_{N_{\pi}}.

end

Return: x_1, \dots, x_N
```

 \rightarrow Complexity of order: $O(MNN_{\pi})$.

Outline





Evaluation metrics

In addition to visual plot of data vs generator samples path, we use some metrics to evaluate the accuracy of our generators:

- Statistical metrics on
 - *Marginal*: Kolmogorov-Smirnov test with *p*-value: if $p > \alpha$ (usually 5%), we do not reject the null-hypothesis (generator came from data of reference distribution)
 - Temporal dynamics: we compute the empirical distribution of the quadratic variation: $\sum_{i} |X_{t_{i+1}} X_{t_i}|^2$ along the time grid \mathcal{T}
 - \bullet Correlation structure: Comparison of empirical covariance or correlation matrix along the time grid ${\cal T}$
 - Predictive score: measure ability to capture conditional distribution over time.
- Tests on real metric of interest to industry:
 - Compare deep hedging in risk management, pricing, etc on historical data sets vs synthetic samples

Autoregressive (AR) model

$$egin{cases} X_{t_1} &= b + arepsilon_1, \ X_{t_2} &= -X_{t_1} + arepsilon_2, \ X_{t_3} &= -X_{t_2} + \sqrt{|X_{t_1}|} + arepsilon_3, \end{cases}$$

where ε_i are independent noises, $\varepsilon_i \sim \mathcal{N}(0, \sigma_i^2)$, with $\sigma_1 = 0.1$, $\sigma_2 = \sigma_3 = 0.05$, b = 0.7.

• Parameters: M = 1000, bandwith h = 0.05, $N_{\pi} = 100$ Runtime for 500 generated samples = 4s.



Figure: Comparison between distribution from reference AR model (data) and generated distribution for each couple (X_{t_i}, X_{t_i}) with $i, j \in [\![1, 3]\!]$ with $i \neq j$

Metrics table for SBST generator vs AR model

	p-value	q 5	\widetilde{q}_5	q 95	\widetilde{q}_{95}
X_{t_1}	0.98	0.535	0.528	0.855	0.861
X_{t_2}	0.74	-0.873	-0.861	-0.516	-0.514
X_{t_3}	0.90	1.243	1.251	1.808	1.793

Table: Marginal metrics for AR model and generator (\tilde{q} for percentile)

	X_{t_1}	X_{t_2}	X_{t_3}
X_{t_1}	0	0.014	-0.01
X_{t_2}	0.014	0	0.013
X_{t_3}	-0.01	0.013	0

Table: Difference between empirical correlation from generated samples and reference samples

A multivariate AR Gaussian model

$$X_{t_{i+1}} = \phi X_{t_i} + \varepsilon_{t_{i+1}}, \quad ext{ with } \quad \varepsilon_{t_i} \sim \mathcal{N}(\mathbf{0}, \sigma \mathbf{1}_d + (1 - \sigma)\mathbb{I}_d),$$

 $\phi \in [0,1]$: correlation across time steps, $\sigma \in [-1,1]$: correlation across the components.

▶ We compute the predictive score: Mean absolute error between conditional mean (from generated model) and the true value: $\mathbb{E}[X_{t_{i+1}}|X_{t_i}] = \phi X_{t_i}$.

	Temporal correlation (fixing $\sigma = 0.8$)			Feature correlation (fixing $\phi = 0.8$)			
Settings	$\phi = 0.2$	$\phi = 0.5$	$\phi = 0.8$	$\sigma = 0.2$	$\sigma = 0.5$	$\sigma = 0.8$	
Predictive score (lower the better)							
SBTS	$.161 \pm .016$	$.180\pm.026$	$.244\pm.014$	$.325 \pm .052$	$.295 \pm .038$	$ $.244 \pm .014	
TimeGAN	$.640\pm0.003$	0.412 ± 0.002	$.251\pm.002$	$.282\pm.005$	$.261\pm.002$	$.251 \pm .002$	

Table: Predictive score for SBTS vs TimeGan

GARCH model

$$\begin{cases} X_{t_{i+1}} = \sigma_{t_{i+1}} \varepsilon_{t_{i+1}} \\ \sigma_{t_{i+1}}^2 = \alpha_0 + \alpha_1 X_{t_i}^2 + \alpha_2 X_{t_{i-1}}^2, \quad i = 1, \dots, N, \end{cases}$$

where $\alpha_0 = 5$, $\alpha_1 = 0.4$, $\alpha_2 = 0.1$, $\varepsilon_{t_i} \sim \mathcal{N}(0, 0.1)$, N = 60.

• Parameters: M = 1000, bandwith h = 0.2, $N_{\pi} = 100$ Runtime for 1000 generated paths = 100s.



Figure: Samples path of reference GARCH (left) and generator SBTS (right)

Metrics for SBST generator vs GARCH model



Figure: Top left: p-value for the marginals X_{t_i} . Top right: samples plot of the joint distribution (X_{t_1}, X_{t_N}) .

Fractional Brownian motion

Fractional Brownian motion (FBM) with Hurst index H = 0.1.

• Parameters: M = 1000, N = 60, $N_{\pi} = 100$, bandwith h = 0.05. Runtime for 1000 generated paths = 100s.



Figure: Four samples path of reference FBM (left) and generator SBTS (right)

Metrics for SBST generator vs FBM



Figure: Top: Quadratic variation distribution $\sum_{i=1}^{N} |X_{t_{i+1}} - X_{t_i}|^2$ for N = 60. Bottom: Covariance matrix for reference FBM and SBTS

Estimation of Hurst index

Standard estimator of Hurst index:

$$\hat{H} = rac{1}{2} \Bigg[1 - rac{\log \Big(\sum_{i=0}^{N-1} |X_{t_{i+1}} - X_{t_i}|^2 \Big)}{\log N} \Bigg].$$

From our generated SBTS with N = 60, we get:

$$\hat{H} = 0.102$$
, Std = 0.003.

Application to deep hedging on real data sets (Apple)

Data: stock prices of Apple from jan. 1, 2010 to jan. 30, 2020, with sliding window of N = 60 days, $\rightarrow M = 2500$ samples.

• Consider a ATM call option on Apple: $g(X_T) = (X_T - K)_+$, and we search for a price p^* and hedging strategy Δ^* minimizing the quadratic criterion (loss function):

$$(p, \Delta) \mapsto \mathbb{E} \left| \underbrace{p + \sum_{i=0}^{N-1} \Delta_{t_i}(X_{t_{i+1}} - X_{t_i}) - g(X_T)}_{\operatorname{PnI}} \right|^2 = \operatorname{replication error}$$

Application to deep hedging on real data sets (Apple)

Data: stock prices of Apple from jan. 1, 2010 to jan. 30, 2020, with sliding window of N = 60 days, $\rightarrow M = 2500$ samples.

• Consider a ATM call option on Apple: $g(X_T) = (X_T - K)_+$, and we search for a price p^* and hedging strategy Δ^* minimizing the quadratic criterion (loss function):

$$(
ho,\Delta) \quad \mapsto \quad \mathbb{E} \Big| \underbrace{p + \sum_{i=0}^{N-1} \Delta_{t_i}(X_{t_{i+1}} - X_{t_i}) - g(X_T)}_{\mathrm{PnL}} \Big|^2 = \mathrm{replication \ error}$$

▶ We parametrize Δ by a LSTM network that is trained from synthetic data sets produced by SBTS, and we compare the results with real-data sets.



Figure: Procedure of backtest for deep hedging

Comparison of the PnL and replication error with real-data and generative SBTS



Figure: Deep hedging PnL distribution from test set

		Training Set		Test Set	
	Price	Mean	Std	Mean	Std
Data	0.0488	0.000165	0.011	-0.015	0.015
SBTS	0.0506	0.0004	0.0109	-0.012	0.013

Table: Mean of PnL and its Std (replication error).

Concluding remarks

- Novel generative model for time series based on Schrödinger bridge (SB) approach:
 - Solution described by a forward stochastic differential equation (SDE) over a finite period, which matchs perfectly the data distribution
 - Drift estimated by nonparametric regression, e.g. kernel method: practical and low-cost computationally (does not require training of neural networks as in GAN type methods)
- Series of numerical experiments, including financial applications with real-data, to illustrate the performance and accuracy of our generative SBTS.
- Further developments:
 - SBTS model can be enriched to fit more accurately with data time series:
 - diffusion coefficient
 - jump-diffusion process
 - Kernel method suffer from curse of dimensionality. Alternately, the drift function can be approximated by neural networks, and more precisely with a LSTM architecture.

Reference

M. Hamdouche, P. Henry-Labordère, H. Pham. Generative modeling for time series via Schrödinger bridge. SSRN 4412434, arXiv:2304.05093

Code available on Github: https://github.com/hamdouchm/SBTimeSeries

THANK YOU FOR YOUR ATTENTION