Higher Order Kernel Mean Embeddings to Capture Filtrations of Stochastic Processes

Anonymous Author(s)
Affiliation
Address
email

Abstract

Stochastic processes are random variables with values in some space of paths. However, reducing a stochastic process to a path-valued random variable ignores its *filtration*, i.e. the flow of information carried by the process through time. By conditioning the process on its filtration, we introduce a family of *higher order kernel mean embeddings* (KMEs) that generalizes the notion of KME and captures additional information related to the filtration. We derive empirical estimators for the associated *higher order maximum mean discrepancies* (MMDs) and prove consistency. We then construct a filtration-sensitive kernel two-sample test able to pick up information that gets missed by the standard MMD test. In addition, leveraging our higher order MMDs we construct a family of universal kernels on stochastic processes that allows to solve real-world calibration and optimal stopping problems in quantitative finance (such as the pricing of American options) via classical kernel-based regression methods. Finally, adapting existing tests for conditional independence to the case of stochastic processes, we design a causal-discovery algorithm to recover the causal graph of structural dependencies among interacting bodies solely from observations of their multidimensional trajectories.

1 Introduction

The idea of embedding probability distributions into a reproducing kernel Hilbert space (RKHS) via kernel mean embeddings (KMEs) has become ubiquitous in many areas of statistics and data science such as hypothesis testing [1, 2], non-linear regression [3, 4], distribution regression [5, 6] etc. Despite strong progress in the study of KMEs, most of the examples considered in the literature tend to focus on random variables supported on some finite (possibly high) dimensional euclidean spaces like $\mathbb{R}^d$. The study of KMEs for function-valued random variables has been largely ignored.

Stochastic processes are random variables with values in some space of paths. However, reducing a stochastic process to a path-valued random variable ignores its *filtration*, which can be informally thought of as the *flow of information carried by the process through time*. A question that naturally emerges from the study of many random, time-evolving systems like financial markets is how does the information available up to present time affect the future evolution of the system?

Formally, this question can be addressed by conditioning a process on its filtration (Sec. 3.1 and 3.2). In this paper we introduce a family of *higher order KMEs* that generalizes the notion of KME to capture additional, filtration-related information (Sec. 3.3 and 3.5). In view of concrete applications, we derive empirical estimators for the associated *higher order MMDs* and use one of them to construct a filtration-sensitive kernel two-sample test (Sec. 3.4) demonstrating with simulated data its ability to capture information that otherwise gets missed by the standard MMD test (Sec. 4.1). Furthermore, we construct a family of universal kernels on stochastic processes (Sec. 3.6) that allows to solve challenging, real-world optimisation problems in quantitative finance such as the pricing of American

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

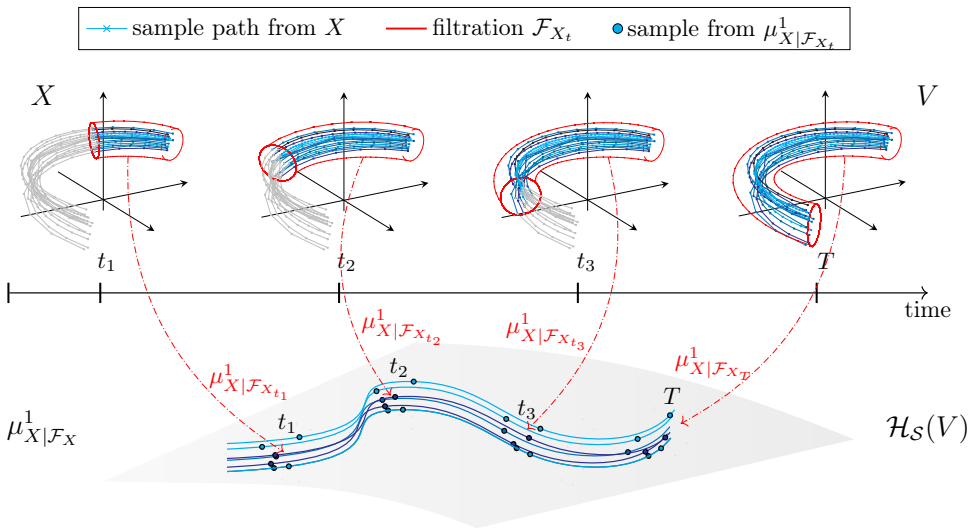


Figure 1: Schematic overview of the construction of the 1st order predictive KME $\mu^1_{X|\mathcal{F}_X}$ (Sec. 3.2). Here X is a stochastic process with sample paths taking their values in V. The red contours indicate the portion of its filtration $\mathcal{F}_X$ upon which the conditioning is applied, i.e. the available information about X from start up to time t. As explained in Sec. 3.2, the 1st order predictive KME $\mu^1_{X|\mathcal{F}_X}$ is a path whose value at time t is a $\mathcal{H}_{\mathcal{S}}(V)$-valued random variable representing the law of X conditioned on its filtration $\mathcal{F}_{X_t}$. Equivalently $\mu^1_{X|\mathcal{F}_X}$ is a stochastic process with sample paths taking their values in $\mathcal{H}_{\mathcal{S}}(V)$.

options via classical kernel-based regression methods (Sec. 4.2). Finally, we adapt existing tests for conditional independence to the case of stochastic processes in order to design a causal-discovery algorithm able to recover the causal graph of structural dependencies among interacting bodies solely from observations of their multidimensional trajectories (Sec. 4.3).

1.1 Related work

The notion of conditioning is a powerful probabilistic tool allowing to understand possibly complex, non-linear interactions between random variables. As their unconditional counterparts, conditional distributions can also be embedded into RKHSs [7]. Recently, conditional KMEs have received increased attention, especially in the context of graphical models [8], state-space models [9], dynamical systems [10], causal inference [11, 12, 13], two-sample and conditional independence hypothesis testing [14, 13, 15] and others. Embeddings of distributions via KMEs have also shown their success in the context of distribution regression (DR), which is the task of learning a function mapping a collection of samples from a probability distribution to scalar targets [16, 17, 18]. More recently, a framework for DR that addresses the setting where inputs are sample paths from an underlying stochastic process is proposed in [6]. The authors make extensive use of the *signature transform* [19, 20] and of the *signature kernel* [21, 22], two well established tools in stochastic analysis.

When it comes to stochastic processes, it was first shown in [23] that weak convergence of random variables does not always account for the information contained in the filtration, as highlighted by means of numerous numerical examples in [24, 25]. This limitation is addressed in [23, 26] through the construction of a sequence of so–called *adapted topologies*[1] $(\tau_n)_{n\geq 1}$ that become progressively finer[2] as n increases (with τ_1 coinciding with the weak topology). In particular, higher order adapted topologies are shown to capture more filtration-related information than their weak counterpart. This characteristic becomes relevant for example in some optimal stopping problems such as the pricing of American options, where the pricing function can be shown to be discontinuous with respect to the weak topology, but is continuous with respect to the second order adapted topology[3] [24, 28, 27]. Leveraging properties of the signature transform, it has been shown that adapted topologies are

[1]We say that a sequence of random variables $\{X_n\}_{n\in\mathbb{N}}$ converges to a random variable X in the topology τ if and only if for every τ-open neighbourhood $\mathbb{U}$ of X there exists $N \in \mathbb{N}$ such that $X_n \in \mathbb{U}$ as soon as $n \geq N$.

[2]A topology τ_1 is said to be finer than a topology τ_2 if every τ_2-open set is also τ_1-open.

[3]The second order adapted topology τ_2 is equivalent to the *adapted Wasserstein distance* [27].

intimately related to a family of higher order MMDs [29]. However, providing empirical estimators for these discrepancies that can be deployed on real-world tasks remains a challenge. In this paper we propose to address this challenge by presenting an alternative construction to this higher order MMDs using the language of kernels and KMEs. The results in [29] serve as a strong theoretical background for the present paper. A more in-depth review of related work can be found in the Appendix.

2 Preliminaries

We begin with a brief summary of tools from stochastic analysis needed to define higher order KMEs. Let $\mathcal{X}(\mathbb{R}^d) = \{x : [0,T] \to \mathbb{R}^d\}$ a compact set of continuous, piecewise linear, $\mathbb{R}^d$-valued paths defined over a common time interval $[0,T]$. obtained for example by linearly interpolating a multivariate time series. For any path $x \in \mathcal{X}(\mathbb{R}^d)$ we denote its kth coordinate by $x^{(k)} : [0,T] \to \mathbb{R}$, for $k \in \{1,\dots,d\}$. More generally we denote by $\mathcal{X}(V) = \{x : [0,T] \to V\}$ a compact set of continuous, piecewise linear paths with values in a Hilbert space V with a countable basis.

2.1 The signature transform

The *signature transform* $\mathcal{S} : \mathcal{X}(V) \to H_{\mathcal{S}}(V)$ is a *feature map* defined for any path $x \in \mathcal{X}(V)$ as the following infinite collection of statistics

$$\mathcal{S}(x) = \left(1, \left\{\mathcal{S}(x)^{(k_1)}\right\}_{k_1=1}^{d}, \left\{\mathcal{S}(x)^{(k_1,k_2)}\right\}_{k_1,k_2=1}^{d}, \dots\right) \tag{1}$$

where each term is a real number equal to the iterated integral

$$\mathcal{S}(x)^{(k_1,\dots,k_j)} = \int\limits_{0<s_1<\dots<s_j<T} \dots \int dx_{s_1}^{(k_1)} \dots dx_{s_j}^{(k_j)} \tag{2}$$

The signature *feature space* $H_{\mathcal{S}}(V)$ is defined as the following direct sum of tensor powers of V

$$H_{\mathcal{S}}(V) = \bigoplus_{k=0}^{\infty} V^{\otimes k} = \mathbb{R} \oplus V \oplus (V)^{\otimes 2} \oplus \dots \tag{3}$$

where $\otimes$ denotes the standard tensor product of vector spaces [30, 19].

2.2 The signature kernel

Because V is Hilbert $H_{\mathcal{S}}(V)$ is also Hilbert [21]. The *signature kernel* $k_{\mathcal{S}} : \mathcal{X}(V) \times \mathcal{X}(V) \to \mathbb{R}$ is a characteristic kernel defined for any pair of paths $x, y \in \mathcal{X}(V)$ as the following inner product

$$k_{\mathcal{S}}(x,y) = \langle \mathcal{S}(x), \mathcal{S}(y) \rangle_{H_{\mathcal{S}}(V)} \tag{4}$$

The recent article [22] establishes a surprising connection between the signature kernel and a certain class of partial differential equations (PDEs), culminating in the following kernel trick for $k_{\mathcal{S}}$.

Theorem 1. *[22, Thm. 2.5] For any $x, y \in \mathcal{X}(V)$ the signature kernel satisfies the equation $k_{\mathcal{S}}(x,y) = u_{x,y}(T,T)$, where $u_{x,y} : [0,T] \times [0,T] \to \mathbb{R}$ is the solution of the hyperbolic PDE*

$$\frac{\partial^2 u_{x,y}}{\partial s \partial t} = \langle \dot{x}_s, \dot{y}_t \rangle_V \, u_{x,y} \tag{5}$$

with boundary conditions $u_{x,y}(0,\cdot) = u_{x,y}(\cdot,0) = 1$ and where $\dot{z}_s = \frac{dz_r}{dr}\big|_{r=s}$.

Hence, evaluating $k_{\mathcal{S}}$ at a pair of paths (x,y) is equivalent to solving the PDE (5); in this paper we solve PDEs numerically via a finite difference scheme (see Appendix for additional details). In what follows, we denote by $\mathcal{H}_{\mathcal{S}}(V) = \{f : \mathcal{X}(V) \to \mathbb{R}\}$ the RKHS associated to $k_{\mathcal{S}}$.

2.3 Stochastic processes and filtrations

We take $(\Omega, \mathcal{F}, \mathbb{P})$ as the underlying probability space. A *(discrete time) stochastic process* X is a random variable with values on $\mathcal{X}(V)$. We denote by $\mathbb{P}_X = \mathbb{P} \circ X^{-1}$ the *law* of X. Assuming the

integrability condition $\mathbb{E}_X[k_{\mathcal{S}}(X,X)] < \infty$, the 1st *order kernel mean embedding* (KME) of X is defined as[4] the following point in $\mathcal{H}_{\mathcal{S}}(V)$

$$\mu^1_X = \mathbb{E}_X[k_{\mathcal{S}}(X,\cdot)] = \int_{x\in\mathcal{X}(V)} k_{\mathcal{S}}(x,\cdot)\mathbb{P}_X(dx) \tag{6}$$

Accordingly, given two stochastic processes X, Y, their 1st *order maximum mean discrepancy* (MMD) is the standard MMD distance with kernel $k_{\mathcal{S}}$ given by the following expression

$$\mathcal{D}^1_{\mathcal{S}}(X,Y) = \left\|\mu^1_X - \mu^1_Y\right\|_{\mathcal{H}_{\mathcal{S}}(V)} \tag{7}$$

Because the signature kernel $k_{\mathcal{S}}$ is characteristic, it is a classical result [1, 31] that the 1st order MMD is a sufficient statistics to distinguish between the laws of X and Y, in other words

$$\mathcal{D}^1_{\mathcal{S}}(X,Y) = 0 \iff \mathbb{P}_X = \mathbb{P}_Y \tag{8}$$

Despite the fact that stochastic processes are path-valued random variables, they encode a much richer structure compared to standard $\mathbb{R}^d$-valued random variables, that goes well beyond their laws. This additional structure is described mathematically by the concept of *filtration* of a process X, defined as the following family of σ-algebras

$$\mathcal{F}_X = (\mathcal{F}_{X_t})_{t\in[0,T]}, \tag{9}$$

where for any $t \in [0,T]$, $\mathcal{F}_{X_t}$ is the σ-algebra generated by the variables $\{X_s\}_{s\in[0,t]}$. Note that $\mathcal{F}_X$ is totally ordered in the sense that $\mathcal{F}_{X_s} \subset \mathcal{F}_{X_t}$ for all $s < t$, which naturally explains why filtrations are good mathematical descriptions to model the flow information carried by the process X.

In the next section, we will present our main findings and introduce a family of *higher order KMEs* and corresponding *higher order MMDs* as generalizations of the standard KME and MMD respectively. We will do so by conditioning stochastic processes on elements of their filtrations.

3 Higher order kernel mean embeddings

We begin by describing how KMEs can be extended to conditional laws of stochastic processes.

3.1 Conditional kernel mean embeddings for stochastic processes

Let X, Y be two stochastic processes. For a given path $x \in \mathcal{X}(V)$, define the 1st *order conditional kernel mean embeddings* $\mu^1_{Y|X=x} \in \mathcal{H}_{\mathcal{S}}(V)$ and $\mu^1_{Y|X} : \mathcal{H}_{\mathcal{S}}(V) \to \mathcal{H}_{\mathcal{S}}(V)$ as follows

$$\mu^1_{Y|X=x} = \mathbb{E}[k_{\mathcal{S}}(Y,\cdot)|X=x] = \int_{y\in\mathcal{X}(V)} k_{\mathcal{S}}(\cdot,y)\mathbb{P}_{Y|X=x}(dy) \tag{10}$$

$$\mu^1_{Y|X} = \mathbb{E}[k_{\mathcal{S}}(Y,\cdot)|X] = \int_{y\in\mathcal{X}(V)} k_{\mathcal{S}}(\cdot,y)\mathbb{P}_{Y|X}(dy) \tag{11}$$

Note that whilst μ^1_X and $\mu^1_{Y|X=x}$ are both single points on $\mathcal{H}_{\mathcal{S}}(V)$, the 1st order conditional KME $\mu^1_{Y|X}$ describes a cloud of points on $\mathcal{H}_{\mathcal{S}}(V)$, each indexed by a path $x \in \mathcal{X}(V)$. Equivalently, $\mu^1_{Y|X}$ constitutes a $\mathbb{P}_X$-measurable, $\mathcal{H}_{\mathcal{S}}(V)$-valued random variable. The 1st order conditional KMEs of equations (10) and (11) satisfy for any $x \in \mathcal{X}(V)$ the following relation [15]

$$\mu^1_{Y|X=x} = \mu^1_{Y|X} k_{\mathcal{S}}(x,\cdot) \tag{12}$$

These embeddings allow to extend the applications of conditional KMEs to the case where the random variables are (possibly multidimensional) stochastic processes. In particular one can directly obtain conditional independence criterions for stochastic processes (see Appendix), enabling to deploy standard kernel-based causal learning algorithms [13], as we demonstrate in Sec. 4. Next we describe how in the case of stochastic processes, conditioning on filtrations is a an important mathematical operation to model real-world time-evolving systems.

[4] The 1st order KME is the standard KME with the signature kernel $k_{\mathcal{S}}$.

3.2 Conditioning stochastic processes on their filtrations

Financial markets are examples of complex dynamical systems that evolve under the influence of randomness. An important objective for financial practitioners is to determine how actionable information available up to present could affect the future market trajectories. The task of *conditioning on the past to describe the future* of a stochastic process X can be formulated mathematically by conditioning X on its filtration $\mathcal{F}_{X_t}$ for any time $t \in [0, T]$.

More precisely, consider the 1st order KME of the conditional law $\mathbb{P}_{X|\mathcal{F}_{X_t}}$, which is defined as the following $\mathcal{F}_{X_t}$-measurable, $\mathcal{H}_{\mathcal{S}}(\mathbb{R}^d)$-valued random variable

$$\mu^1_{X|\mathcal{F}_{X_t}} = \mathbb{E}[k_{\mathcal{S}}(X, \cdot)|X_{[0,t]}] = \int_{x \in \mathcal{X}(V)} k_{\mathcal{S}}(\cdot, x)\mathbb{P}_{X|\mathcal{F}_{X_t}}(dx) \tag{13}$$

where $X_{[0,t]}$ denotes the stochastic process X restricted to the sub-interval $[0, t] \subset [0, T]$. By varying the time index t, we can form the following ordered collection of 1st order KMEs

$$\mu^1_{X|\mathcal{F}_X} = \left(\mu^1_{X|\mathcal{F}_{X_t}}\right)_{t \in [0,T]} \tag{14}$$

that we term 1st *order predictive KME* of the process X. By construction, $\mu^1_{X|\mathcal{F}_X}$ describes a path taking its values in the space of $\mathcal{H}_{\mathcal{S}}(V)$-valued random variables, in other words it is itself a stochastic process [5] (see Fig. 1). Hence, the law of $\mu^1_{X|\mathcal{F}_X}$ can itself be embedded via KMEs into a "higher-order RKHS" (see next section), making the full procedure iterable, as we shall discuss next.

3.3 Second order kernel mean embedding and maximum mean discrepancy

The 2nd *order KME* is the point in $\mathcal{H}_{\mathcal{S}}(\mathcal{H}_{\mathcal{S}}(V))$ defined as the KME of the 1st order predictive KME

$$\mu^2_X = \int_{x \in \mathcal{X}(\mathcal{H}_{\mathcal{S}}(V))} k_{\mathcal{S}}(\cdot, x)\mathbb{P}_{\mu^1_{X|\mathcal{F}_X}}(dx) \tag{15}$$

Thus, the 2nd *order MMD* of X, Y is the difference in $\mathcal{H}_{\mathcal{S}}(\mathcal{H}_{\mathcal{S}}(V))$ of their 2nd order KMEs, i.e.

$$\mathcal{D}^2_{\mathcal{S}}(X, Y) = \left\|\mu^2_X - \mu^2_Y\right\|_{\mathcal{H}_{\mathcal{S}}(\mathcal{H}_{\mathcal{S}}(V))} \tag{16}$$

The next theorem states that the 2nd order MMD of two stochastic processes X, Y is a stronger discrepancy measure than the 1st order MMD.

Theorem 2. *Given two stochastic processes X, Y*

$$\mathcal{D}^2_{\mathcal{S}}(X, Y) = 0 \iff \mathbb{P}_{X|\mathcal{F}_X} = \mathbb{P}_{Y|\mathcal{F}_Y} \tag{17}$$

Furthermore

$$\mathcal{D}^2_{\mathcal{S}}(X, Y) = 0 \implies \mathcal{D}^1_{\mathcal{S}}(X, Y) = 0 \tag{18}$$

but the converse is not generally true.

Proof. All proofs are given in the Appendix. □

Next we make use of Thm. 2 in the context of two-sample hypothesis testing [1, 31] for stochastic processes. In Sec. 4 we will show by means of a numerical example that the 2nd order MMD is able to capture filtration-related information otherwise ignored by the 1st order MMD.

3.4 A filtration-sensitive kernel two-sample test

Suppose we are given m sample paths $\{x^i\}_{i=1}^m \sim X$ and n sample paths $\{y^i\}_{i=1}^n \sim Y$. A classical two-sample test [1] for X, Y tests a null-hypothesis

$$H_0 : \mathbb{P}_X = \mathbb{P}_Y \text{ against the alternative } H_A : \mathbb{P}_X \neq \mathbb{P}_Y \tag{19}$$

[5] Because $\mathbb{P}_X = \mathbb{P}_{X|\mathcal{F}_{X_T}}$, all the information about the law of X is contained in just the terminal point of the trajectory traced by $\mu^1_{X|\mathcal{F}_X}$ (Fig. 1).

The probability of falsely rejecting the null is called the *type I error* (and similarly the probability of falsely accepting the null is called the *type II error*). If the type I error can be bounded from above by a constant α, then we say that the test is of level α. In [31, Sec. 8] it is shown that rejecting the null if $\widehat{\mathcal{D}}^1_{\mathcal{S}}(X,Y)^2 > c_\alpha$ (for some c_α that depends on m, n and α) gives a test of level α, where $\widehat{\mathcal{D}}^1_{\mathcal{S}}(X,Y)$ denotes the classical unbiased estimator of the 1st order MMD [1]. This choice of threshold is conservative and can be improved by using data-dependent bounds such as in permutation tests (we refer to the MMD testing literature for extra details [1, 32, 33]).

However, as discussed in Sec. 3.3 comparing the laws $\mathbb{P}_X, \mathbb{P}_Y$ via the estimator above might be insufficient to capture filtration-related information about X, Y. To overcome this limitation we propose instead to test the null-hypothesis

$$H_0 : \mathbb{P}_{X|\mathcal{F}_X} = \mathbb{P}_{Y|\mathcal{F}_Y} \text{ against the alternative } H_A : \mathbb{P}_{X|\mathcal{F}_X} \neq \mathbb{P}_{Y|\mathcal{F}_Y} \tag{20}$$

Using Thm. 2, one can immediately construct a filtration-sensitive kernel two-sample test for (20) provided one can build an empirical estimator of the 2nd order MMD $\mathcal{D}^2_{\mathcal{S}}(X,Y)$. In the rest of this section we explain how to obtain such an estimator and ultimately show its consistency.

Assuming availability of m sample paths $\{\widetilde{x}^i\}_{i=1}^m$ from the stochastic process $\mu^1_{X|\mathcal{F}_X}$ and n sample paths $\{\widetilde{y}^i\}_{i=1}^n$ from $\mu^1_{Y|\mathcal{F}_Y}$, an estimator of the squared 2nd order MMD is given by

$$\widehat{\mathcal{D}}^2_{\mathcal{S}}(X,Y)^2 = \frac{1}{m(m-1)} \sum_{\substack{i,j=1 \\ i\neq j}}^{m} k_{\mathcal{S}}(\widetilde{x}^i, \widetilde{x}^j) - \frac{2}{mn} \sum_{i,j=1}^{m,n} k_{\mathcal{S}}(\widetilde{x}^i, \widetilde{y}^j) + \frac{1}{n(n-1)} \sum_{\substack{i,j=1 \\ i\neq j}}^{n} k_{\mathcal{S}}(\widetilde{y}^i, \widetilde{y}^j)$$

Computing this estimator boils down to evaluating the signature kernel $k_{\mathcal{S}}(\widetilde{x}, \widetilde{y})$ on sample paths $\widetilde{x} \sim \mu^1_{X|\mathcal{F}_X}$ and $\widetilde{y} \sim \mu^1_{Y|\mathcal{F}_Y}$. By Thm. 1, the signature kernel solves the following PDE

$$\frac{\partial^2 u_{\widetilde{x},\widetilde{y}}}{\partial s \partial t} = \left(\left\langle \widetilde{x}_{s-\delta}, \widetilde{y}_{t-\delta} \right\rangle_{\mathcal{H}_{\mathcal{S}}(V)} - \left\langle \widetilde{x}_{s-\delta}, \widetilde{y}_t \right\rangle_{\mathcal{H}_{\mathcal{S}}(V)} - \left\langle \widetilde{x}_s, \widetilde{y}_{t-\delta} \right\rangle_{\mathcal{H}_{\mathcal{S}}(V)} + \left\langle \widetilde{x}_s, \widetilde{y}_t \right\rangle_{\mathcal{H}_{\mathcal{S}}(V)} \right) u_{\widetilde{x},\widetilde{y}}$$

where the two derivatives in eq. (5) have been approximated by finite difference with time increment δ. It remains to explain how to estimate, for any $s, t \in [0,T]$, the inner product $\left\langle \widetilde{x}_s, \widetilde{y}_t \right\rangle_{\mathcal{H}_{\mathcal{S}}(V)}$ from sample paths of X and Y. This can be achieved using the formalism of *cross-covariance operators* [34] as thoroughly explained in the Appendix, which yields to the following approximation

$$\left\langle \widetilde{x}_s, \widetilde{y}_t \right\rangle_{\mathcal{H}_{\mathcal{S}}(V)} \approx \mathbf{k}^x_s (\mathbf{K}^{x,x}_{s,s} + m\lambda I_m)^{-1} \mathbf{K}^{x,y}_{s,t} (\mathbf{K}^{y,y}_{t,t} + n\lambda I_n)^{-1} \mathbf{k}^y_t \tag{21}$$

where $\mathbf{k}^x_s \in \mathbb{R}^m, \mathbf{k}^y_t \in \mathbb{R}^n$ are the vectors[6]

$$[\mathbf{k}^x_s]_i = k_{\mathcal{S}}(x^i_{[0,s]}, x^i_{[0,s]}), \quad [\mathbf{k}^y_t]_i = k_{\mathcal{S}}(y^i_{[0,t]}, y^i_{[0,t]})$$

and $\mathbf{K}^{x,x}_{s,s} \in \mathbb{R}^{m\times m}$, $\mathbf{K}^{x,y}_{s,t} \in \mathbb{R}^{m\times n}$, $\mathbf{K}^{y,y}_{t,t} \in \mathbb{R}^{n\times n}$ are the matrices

$$[\mathbf{K}^{x,x}_{s,s}]_{i,j} = k_{\mathcal{S}}(x^i_{[0,s]}, x^j_{[0,s]}), \quad [\mathbf{K}^{x,y}_{s,t}]_{i,j} = k_{\mathcal{S}}(x^i_{[0,s]}, y^j_{[0,t]}), \quad [\mathbf{K}^{y,y}_{t,t}]_{i,j} = k_{\mathcal{S}}(y^i_{[0,t]}, y^j_{[0,t]})$$

and where I_m (resp. I_n) is the $m \times m$ (resp. $n \times n$) identity matrix. The corresponding algorithm and its complexity analysis are provided in Appendix.

The next theorem ensures that the estimator $\widehat{\mathcal{D}}^2_{\mathcal{S}}(X,Y)$ is consistent for the 2nd order MMD.

Theorem 3. *$\widehat{\mathcal{D}}^2_{\mathcal{S}}(X,Y)$ is a consistent estimator for the 2nd order MMD, i.e.*

$$|\widehat{\mathcal{D}}^2_{\mathcal{S}}(X,Y) - \mathcal{D}^2_{\mathcal{S}}(X,Y)| \xrightarrow{p} 0 \quad \textit{as } m, n \to \infty \tag{22}$$

with $\{x^i\}_{i=1}^m \sim X$, $\{y^i\}_{i=1}^n \sim Y$ and where convergence is in probability.

We now iterate the procedure presented so far to define higher order KMEs and MMDs.

[6] Here we use the notation $x_{[0,s]}$ to denote the restriction of the path x to the sub-interval $[0,s] \subset [0,T]$.

3.5 Higher order kernel mean embeddings and maximum mean discrepancies

One can iterate the procedure described in Sec. 3.3 and recursively define, for any $n \in \mathbb{N}_{>1}$, the nth *order KME of* X as the following point in $\mathcal{H}_{\mathcal{S}}^n(\mathbb{R}^d)$

$$\mu_X^n = \int_{x \in \mathcal{X}(\mathcal{H}_{\mathcal{S}}^n(\mathbb{R}^d))} k_{\mathcal{S}}(\cdot, x) \mathbb{P}_{\mu_{X|\mathcal{F}_X}^{n-1}}(dx) \tag{23}$$

where $\mu_{X|\mathcal{F}_X}^{n-1}$ is the $(n-1)^{\text{st}}$ predictive KME of X and

$$\mathcal{H}_{\mathcal{S}}^n(\mathbb{R}^d) = \underbrace{\mathcal{H}_{\mathcal{S}}(\mathcal{H}_{\mathcal{S}}(\ldots \mathcal{H}_{\mathcal{S}}}_{n \text{ times}}(\mathbb{R}^d) \ldots)) \tag{24}$$

The associated n^{th} *order MMD* between two processes X, Y is then defined as the difference in $\mathcal{H}_{\mathcal{S}}^n(\mathbb{R}^d)$ of the two n^{th} order KMEs

$$\mathcal{D}_{\mathcal{S}}^n(X, Y) = \|\mu_X^n - \mu_Y^n\|_{\mathcal{H}_{\mathcal{S}}^n(\mathbb{R}^d)} \tag{25}$$

The following result generalizes Thm. 2 in that it shows that the nth order MMD is a stronger (i.e. finer) discrepancy measure than all the kth order MMDs of lower order $1 < k < n$.

Theorem 4. *Given two stochastic processes* X, Y

$$\mathcal{D}_{\mathcal{S}}^n(X, Y) = 0 \implies \mathcal{D}_{\mathcal{S}}^k(X, Y) = 0 \quad \textit{for any } \ 1 < k < n \tag{26}$$

but the converse is not generally true.

Other than hypothesis testing, another important application relying on the ability of distinguishing between random variables is *distribution regression* (DR) [5]. In the next section we make use of the nth order MMD in the setting of DR on path-valued random variables presented in [6] and propose a family of kernels on stochastic processes whose RKHSs contains richer classes of functions than the RKHS associated to the universal kernel proposed in [6].

3.6 Higher order distribution regression

DR on stochastic processes describes the supervised learning problem where the input is a collection of sample paths and the output is a vector of scalars [6]. Denote by $\mathcal{P}(\mathcal{X}(\mathbb{R}^d))$ the set of stochastic processes with sample paths on $\mathcal{X}(\mathbb{R}^d)$. Following the setup in [6], the goal is to learn a function $F : \mathcal{P}(\mathcal{X}(\mathbb{R}^d)) \to \mathbb{R}$ from a training set of input-output pairs $\{(X_i, y_i)\}$ with $X_i \in \mathcal{P}(\mathcal{X}(\mathbb{R}^d))$ and $y_i \in \mathbb{R}$, by means of a classical two-step procedure [16, 17, 18].

Firstly, a stochastic process $X \in \mathcal{P}(\mathcal{X}(\mathbb{R}^d))$ is embedded into its KME μ_X^1 via the signature kernel $k_{\mathcal{S}}$. Secondly, another function $G : \mathcal{H}_{\mathcal{S}}(V) \to \mathbb{R}$ is learnt by solving the minimization $\arg\min_{G \in \mathcal{H}_{\text{RBF}}} \sum_i \mathcal{L}(g(\mu_{X_i}^1), y_i)$, where $\mathcal{L}$ is a loss function, and $\mathcal{H}_{\text{RBF}}$ is the RKHS associated to the classical Gaussian kernel $k_{\text{RBF}} : \mathcal{H}_{\mathcal{S}}(V) \times \mathcal{H}_{\mathcal{S}}(V) \to \mathbb{R}$. This procedure materialises into a kernel on stochastic processes whose RKHS is shown to be dense in the space of functions $F : \mathcal{P}(\mathcal{X}(\mathbb{R}^d)) \to \mathbb{R}$ that are continuous with respect to the weak topology [6, Thm. 3.3].

However, a class of approximators that is universal with respect to some topology is not guaranteed to well approximate functions that are discontinous with respect to that topology (but potentially continuous with respect to a finer topology). For example, financial practitioners are often interested in calibrating financial models to market data or pricing financial instruments from observations of market dynamics. These tasks can be formulated as DR problems on stochastic processes (see experiments in Sec. 4.2), but the resulting learnable functions are discontinuous with respect to the 1st order MMD whilst being continuous with respect to the 2nd order MMD [27]. This motivates the need to extend the kernel-based DR technique proposed in [6] to situations where the target functions are not weakly continuous, which is what Thm. 5 addresses. A function $f : \mathbb{R} \to \mathbb{R}$ is called *globally analytic with non-negative coefficients* if admits everywhere a Taylor expansion where all the coefficients are strictly positive, i.e. for any $x \in \mathbb{R}$ we have $f(x) = \sum_{i=0}^{\infty} a_i x^i$ with $a_i > 0$.

Theorem 5. *Let* $f : \mathbb{R} \to \mathbb{R}$ *be a globally analytic function with non-negative coefficients. Define the family of kernels* $K_{\mathcal{S}}^n : \mathcal{P}(\mathcal{X}(\mathbb{R}^d)) \times \mathcal{P}(\mathcal{X}(\mathbb{R}^d)) \to \mathbb{R}$ *as follows*

$$K_{\mathcal{S}}^n(X, Y) = f(\mathcal{D}_{\mathcal{S}}^n(X, Y)), \quad n \in \mathbb{N}_{\geq 1} \tag{27}$$

Then the RKHS associated to $K_{\mathcal{S}}^n$ *is dense in the space of functions from* $\mathcal{P}(\mathcal{X}(\mathbb{R}^d))$ *to* $\mathbb{R}$ *which are continuous with respect to the* k^{th} *order MMD for any* $1 < k \leq n$.

In Sec. 4 we will take $f(x) = \exp(-x^2/\sigma)$ with $\sigma > 0$. This result marks the end of our analysis. Next we apply our theoretical results in the contexts of two-sample testing, DR and causal inference.

4 Applications

Here we demonstrate the practical advantage of using 1st order kernel mean embeddings, and evaluate the conditional kernel mean embedding for stochastic processes on a causal discovery task. Additional experimental details can be found in Appendix and the code will be made publicly available.

4.1 Hypothesis testing on filtrations

We start by considering two processes X^n and X with transition probabilities depicted in Fig. 2. Although the laws $\mathbb{P}_n$ and $\mathbb{P}$ get arbitrarily close for large n, their filtrations are very different. Indeed, the two processes have different information structures available before time $t = 1$. Indeed, for any $0 < t \leq 1$, the trajectory of X^n is deterministic, whilst the progression of X remains random until $t = 1$. Being able to distinguish two such stochastic processes is crucial in quantitative finance: if

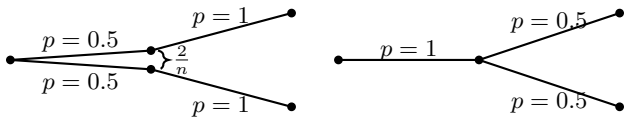


Figure 2: The supports of $\mathbb{P}_n$ (left) and $\mathbb{P}$ (right).

$\mathbb{P}_n$ and $\mathbb{P}$ are the laws of two traded assets, $\mathbb{P}_n$ gives an arbitrage opportunity. As shown in Fig. 3, the 2nd order MMD can distinguish these two processes with similar laws ($n = 10^5$) but different filtrations, while the 1st order MMD fails to do so.

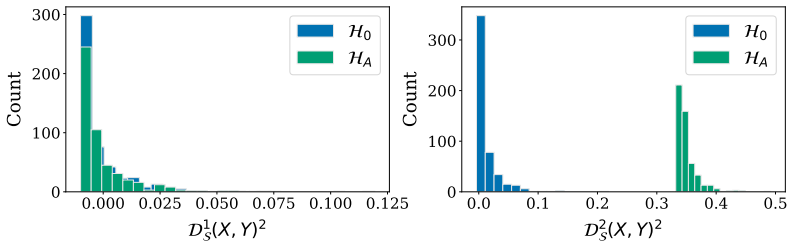


Figure 3: *Left:* Empirical distribution of the 1st order MMD. Under $\mathcal{H}_0$ the two measures are both equal to $\mathbb{P}$ and we use 500 samples from each. Under $\mathcal{H}_A$ with $\mathbb{P}$ and $\mathbb{P}_n$ where $n = 10^5$, and we use 500 samples. *Right:* Same for the 2nd order MMD. Histograms are obtained by computing 500 independent instances of the MMD.

4.2 Applications of higher order distribution regression to quantitative finance

In this section we use kernel Ridge regression and support vector machine (SVM) classification equipped with the kernel $K^2_{\mathcal{S}}$ from Thm. 5 to address two real-world problems arising in quantitative finance, notably the calibration of the *rough Bergomi model* [35] and the pricing of *American options* [36]. We benchmark our filtration-sensitive kernel $K^2_{\mathcal{S}}$ against a range of kernels, including $K^1_{\mathcal{S}}$.

The rough Bergomi model is a rough volatility model [37] satisfying the following stochastic dynamics

$$dS_t = \sqrt{V_t} S_t dW_t, \quad V_t = \int_0^t K(s,t) dZ_s, \quad Z_t = \rho W_t + \sqrt{1-\rho^2} W'_t \tag{28}$$

where W, W' are two independent Brownian motions and $K(s,t) = (t-s)^{h-0.5}$ where here we take $h = 0.2$. The model in eq. (28) is non-Markovian in the sense that the conditional law of $S \mid \mathcal{F}_{S_t}$ depends pathwise on the past history of the process. Of particular importance is the correct retrieval of the sign of the correlation parameter ρ [38]. We consider 50 parameter values $\{\rho_i\}_{i=1}^{50}$ chosen uniformly at random from $[-0.4, 0.4]$. Each ρ_i is regressed on a collection of $m = 200$ sample trajectories. We use an SVM classifier endowed with different kernels (Table 1).

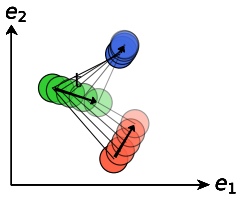
One of the most studied optimal stopping problems is the pricing of an American option with a non-negative payoff function $g : \mathbb{R}^d \to \mathbb{R}$. Stock prices are assumed to follow a d-dimensional stochastic process X. The price of the corresponding option is the solution of the optimal stopping problem $\sup_\tau \mathbb{E}[g(X_\tau) \mid X_0]$, where the supremum is taken over stopping times τ. Despite significant advances, pricing American options remains one of the most computationally challenging problems in financial optimization, in particular when the underlying process X is non-Markovian. This is the setting we consider, modelling stock prices as sample paths from fractional Brownian motion (fBm) [39] with different Hurst exponents $h \in (0, 1)$. True target prices are obtained via expensive Monte Carlo simulations [40]. We consider 25 values of $\{h_i\}_{i=1}^{25}$ sampled uniformly at random in $[0.2, 0.8]$ and use 500 samples from each fBm. As shown in Table 1 our kernel $K_{\mathcal{S}}^2$ yields the best results on both tasks (rough Bergomi model calibration and American option pricing), systematically outperforming other classical kernels as well as the kernel $K_{\mathcal{S}}^1$ introduced in [6].

Table 1: Quantitative finance examples. Average performances with standard errors in parenthesis.

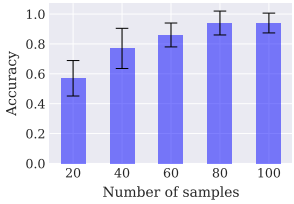
Kernel	Rough Bergomi model calibration (Acc.)	American option pricing (MSE $\times 10^{-3}$)
RBF	80% (5%)	1.07 (0.75)
Matérn	87% (3%)	2.75 (3.05)
$K_{\mathcal{S}}^1$	88% (3%)	0.90 (0.34)
$K_{\mathcal{S}}^2$	**91%** (3%)	**0.52** (0.07)

4.3 Inferring causal graph for interacting bodies

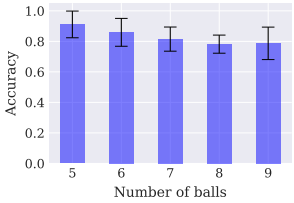
Finally, we consider the task of recovering the causal relationships between interacting bodies solely from observations of their multidimensional trajectories. We employ the multi-body interaction simulator from [41] in order to simulate an environment where N balls are connected by invisible physical relations (e.g. a spring) and describe 2D trajectories (see Fig. 4a with $N = 3$ and 2 springs). At the beginning of a simulated episode, the initial positions of the balls are generated at random, and during the episode, the balls are subject to forces with random intensity and direction. By simulating m episodes we end up with m sample trajectories for each of the N balls. We use the kPC algorithm [13]—which relies on conditional independence testing— with the signature kernel and evaluate its ability to recover whether any two balls are connected or not. We vary m and N and report the results in Figs. 4b and 4c. Each experiment is run 15 times, 30% of the runs are used to chose the hyperparameters, and the reported results have been obtained on the remaining runs.



(a) 3 interacting balls describing trajectories in the 2D plane over time.



(b) Accuracy on binary classification of edges with a varying number of sample episodes (5 balls)



(c) Accuracy on binary classification of edges with a varying number of balls (100 samples)

5 Conclusion

In this paper, we introduced a family of higher order KMEs by conditioning a stochastic process on its filtration, generalizing the classical notion of KME. We derived an empirical estimator for the 2nd order MMD and proved its consistency. We then proposed a filtration-sensitive kernel two-sample test and showed with simulations its ability to capture information that gets missed by the standard MMD test. In addition, we constructed a family of universal kernels on stochastic processes that allows to solve real-world calibration and optimal stopping problems in quantitative finance via Ridge regression. Finally, we designed a causal-discovery algorithm using conditional independence tests to recover the causal graph of structural dependencies among interacting bodies solely from observations of their multidimensional trajectories.

References

[1] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012.

[2] Kun Zhang, Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. Kernel-based conditional independence test and application in causal discovery. *arXiv preprint arXiv:1202.3775*, 2012.

[3] Bernhard Schölkopf, Alexander J Smola, Francis Bach, et al. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2002.

[4] Thomas Hofmann, Bernhard Schölkopf, and Alexander J Smola. Kernel methods in machine learning. *The annals of statistics*, pages 1171–1220, 2008.

[5] Zoltán Szabó, Bharath K Sriperumbudur, Barnabás Póczos, and Arthur Gretton. Learning theory for distribution regression. *The Journal of Machine Learning Research*, 17(1):5272–5311, 2016.

[6] Maud Lemercier, Cristopher Salvi, Theodoros Damoulas, Edwin Bonilla, and Terry Lyons. Distribution regression for sequential data. In *International Conference on Artificial Intelligence and Statistics*, pages 3754–3762. PMLR, 2021.

[7] Le Song, Kenji Fukumizu, and Arthur Gretton. Kernel embeddings of conditional distributions: A unified kernel framework for nonparametric inference in graphical models. *IEEE Signal Processing Magazine*, 30(4):98–111, 2013.

[8] Le Song, Arthur Gretton, and Carlos Guestrin. Nonparametric tree graphical models. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 765–772. JMLR Workshop and Conference Proceedings, 2010.

[9] Kenji Fukumizu, Le Song, and Arthur Gretton. Kernel bayes' rule: Bayesian inference with positive definite kernels. *The Journal of Machine Learning Research*, 14(1):3753–3783, 2013.

[10] Le Song, Jonathan Huang, Alex Smola, and Kenji Fukumizu. Hilbert space embeddings of conditional distributions with applications to dynamical systems. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 961–968, 2009.

[11] Jovana Mitrovic, Dino Sejdinovic, and Yee Whye Teh. Causal inference via kernel deviance measures. *arXiv preprint arXiv:1804.04622*, 2018.

[12] Robert E Tillman, Arthur Gretton, and Peter Spirtes. Nonlinear directed acyclic structure learning with weakly additive noise models. In *NIPS*, pages 1847–1855, 2009.

[13] Xiaohai Sun, Dominik Janzing, Bernhard Schölkopf, and Kenji Fukumizu. A kernel-based causal learning algorithm. In *Proceedings of the 24th international conference on Machine learning*, pages 855–862, 2007.

[14] Kenji Fukumizu, Arthur Gretton, Xiaohai Sun, and Bernhard Schölkopf. Kernel measures of conditional dependence. In *NIPS*, volume 20, pages 489–496, 2007.

[15] Junhyung Park and Krikamol Muandet. A measure-theoretic approach to kernel conditional mean embeddings. *Advances in Neural Information Processing Systems*, 33, 2020.

[16] Ho Chung Leon Law, Dougal Sutherland, Dino Sejdinovic, and Seth Flaxman. Bayesian approaches to distribution regression. In *International Conference on Artificial Intelligence and Statistics*, pages 1167–1176. PMLR, 2018.

[17] Krikamol Muandet, Kenji Fukumizu, Francesco Dinuzzo, and Bernhard Schölkopf. Learning from distributions via support measure machines. In *Advances in neural information processing systems*, pages 10–18, 2012.

[18] Alex Smola, Arthur Gretton, Le Song, and Bernhard Schölkopf. A hilbert space embedding for distributions. In *International Conference on Algorithmic Learning Theory*, pages 13–31. Springer, 2007.

[19] Terry Lyons. Rough paths, signatures and the modelling of functions on streams. *Proceedings of the International Congress of Mathematicians, Korea*, 2014.

[20] Patric Bonnier, Patrick Kidger, I Perez Arribas, Cristopher Salvi, and Terry Lyons. Deep signature transforms. *Advances in Neural Information Processing Systems*, 2019.

[21] Franz J Király and Harald Oberhauser. Kernels for sequentially ordered data. *Journal of Machine Learning Research*, 20(31):1–45, 2019.

[22] Thomas Cass, Terry Lyons, Cristopher Salvi, and Weixin Yang. The signature kernel is the solution of a goursat pde. *arXiv preprint arXiv:2006.14794*, 2020.

[23] D. J. Aldous. Weak convergence and general theory of processes. *Unpublished draft of monograph*, 1981.

[24] G. C. Pflug and A. Pichler. A distance for multistage stochastic optimization models. *SIAM J. Optim.*, 22(1):1–23, 2012.

[25] J Backhoff-Veraguas, D Bartl, M Beiglböck, and J Wiesel. Estimating processes in adapted wasserstein distance. *Annals of applied probability*, 2021.

[26] D. Hoover and J. Keisler. Adapted probability distributions. *Trans. Amer. Math. Soc.*, 1984.

[27] J Backhoff-Veraguas, D Bartl, M Beiglböck, and M Eder. Adapted wasserstein distances and stability in mathematical finance. *Finance and Stochastics*, 2019.

[28] Julio Backhoff-Veraguas, Daniel Bartl, Mathias Beiglböck, and Manu Eder. All adapted topologies are equal. *Probability Theory and Related Fields*, 178(3):1125–1172, 2020.

[29] Patric Bonnier, Chong Liu, and Harald Oberhauser. Adapted topologies and higher rank signatures. *arXiv preprint arXiv:2005.08897*, 2020.

[30] Terry J Lyons. Differential equations driven by rough signals. *Revista Matemática Iberoamericana*, 14(2):215–310, 1998.

[31] Ilya Chevyrev and Harald Oberhauser. Signature moments to characterize laws of stochastic processes. *arXiv preprint arXiv:1810.10971*, 2018.

[32] Bharath K Sriperumbudur, Arthur Gretton, Kenji Fukumizu, Bernhard Schölkopf, and Gert RG Lanckriet. Hilbert space embeddings and metrics on probability measures. *The Journal of Machine Learning Research*, 11:1517–1561, 2010.

[33] Dino Sejdinovic, Bharath Sriperumbudur, Arthur Gretton, and Kenji Fukumizu. Equivalence of distance-based and rkhs-based statistics in hypothesis testing. *The Annals of Statistics*, pages 2263–2291, 2013.

[34] Krikamol Muandet, Kenji Fukumizu, Bharath Sriperumbudur, and Bernhard Schölkopf. Kernel mean embedding of distributions: A review and beyond. *arXiv preprint arXiv:1605.09522*, 2016.

[35] Christian Bayer, Peter Friz, and Jim Gatheral. Pricing under rough volatility. *Quantitative Finance*, 16(6):887–904, 2016.

[36] H Buehler, L Gonon, J Teichmann, and B Wood. Deep hedging. *Quantitative Finance*, 19(8):1271–1291, 2019.

[37] Jim Gatheral, Thibault Jaisson, and Mathieu Rosenbaum. Volatility is rough. *Quantitative Finance*, 18(6):933–949, 2018.

[38] Paul Gassiat et al. On the martingale property in the rough bergomi model. *Electronic Communications in Probability*, 24, 2019.

[39] Tyrone E Duncan, Yaozhong Hu, and Bozenna Pasik-Duncan. Stochastic calculus for fractional brownian motion i. theory. *SIAM Journal on Control and Optimization*, 38(2):582–612, 2000.

[40] Francis A Longstaff and Eduardo S Schwartz. Valuing american options by simulation: a simple least-squares approach. *The review of financial studies*, 14(1):113–147, 2001.

[41] Yunzhu Li, Antonio Torralba, Anima Anandkumar, Dieter Fox, and Animesh Garg. Causal discovery in physical systems from videos. *Advances in Neural Information Processing Systems*, 33, 2020.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] See Sec. 3 and 4.
   (b) Did you describe the limitations of your work? [Yes] We discuss the complexities of our algorithm in Appendix.
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes] All proofs are provided in the Appendix.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] The code to reproduce all numerical experiments is included in the supplementary material.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Sec. 4 and Appendix.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] See Sec. 4.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] Complexities of the proposed algorithms can be found in the Appendix.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] See Sec. 4.3
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] We included an anonymized GitHub repository
   (d) Did you discuss whether and how consent was obtain tod from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]