SIAM J. OPTIM.
Vol. 26, No. 3, pp. 1715–1740

FROM EMPIRICAL OBSERVATIONS TO TREE MODELS FOR STOCHASTIC OPTIMIZATION: CONVERGENCE PROPERTIES*

GEORG CH. PFLUG† AND ALOIS PICHLER‡

Abstract. In multistage stochastic optimization we use stylized processes to model the relevant stochastic data processes. The basis for building these models is empirical observations. It is well known that the determining distance concept for multistage stochastic optimization problems is the nested distance and not the distance in distribution. In this paper we investigate the question of how to generate models out of empirical data, which approximate well the underlying stochastic processes in nested distance. We demonstrate first that the empirical measure, which is built from observed sample paths, does not converge in nested distance to the pertaining distribution if the latter has a density. On the other hand, we show that smoothing convolutions, which are appropriately adapted from classical kernel density estimation, can be employed to modify the empirical measure in order to obtain stochastic processes which converge in nested distance to the underlying process. We employ the results to estimate the conditional densities for each time stage. Finally we construct discrete tree processes from observed empirical paths, which approximate well the original stochastic process as they converge in nested distance to the underlying process.

Key words. decision trees, stochastic optimization, optimal transportation

AMS subject classifications. 90C15, 60B05, 62P05

DOI. 10.1137/15M1043376

1. Introduction.

For stochastic optimization problems, i.e., problems involving random variables, the most widespread numerical solution method is to replace the original probability measure by an appropriate discrete approximation of it, for instance by the empirical distribution or by a sample from it. Reducing the computational complexity is of even higher importance for applications involving stochastic processes, as they are typically more difficult to handle than simple random variables. In this paper, we consider only stochastic processes in discrete time.

An empirical observation of a stochastic process is a single sample path. The empirical measure corresponding to n observations assigns the probability $1/n$ to each of the sample paths. It is evident that the empirical measure cannot capture conditional transition probabilities, given an arbitrarily chosen subpath. Indeed, consider a subpath which is possible but was not observed, from its origin up to some intermediate state. Then, with probability 1, none of the empirical observations coincides with this chosen subpath, and hence the empirical measure cannot reproduce the distribution conditional on this chosen path.

Pagès and coworkers (cf. [21] or [2]) elaborate optimal discrete approximations (often called quantizers) to treat specific problems such as, e.g., option pricing. These simpler models consist of representative paths, which approximate a probability measure in some optimal way (cf. Graf and Luschgy [9]). Although optimal for specific problems, these representative quantizers also do not describe conditional transitions,

*Received by the editors October 12, 2015; accepted for publication (in revised form) May 9, 2016; published electronically September 1, 2016.
http://www.siam.org/journals/siopt/26-3/M104337.html

†Department of Statistics and Operations Research, University of Vienna, 1010 Vienna, Austria, and International Institute for Applied Systems Analysis (IIASA), Laxenburg, Austria (georg.pflug@univie.ac.at).

‡Norwegian University of Science and Technology, NTNU, 7491 Trondheim, Norway (alois.pichler@univie.ac.at). This author's work was partly supported by the Research Council of Norway (grant 207690/E20).

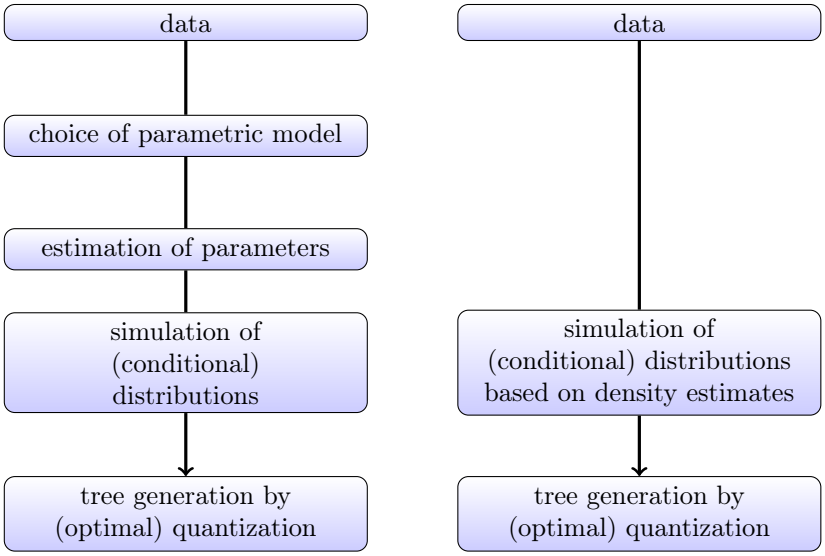


FIG. 1. *The parametric (left) and the nonparametric (right) approaches.*

as they lack a branching structure as well.

The branching structure corresponds to the information gain obtained in time, i.e., the pertaining *filtration*. Taking available information into account is essential for stochastic optimization problems. It is well known that *trees* (scenario or decision trees) constitute an appropriate data structure to model both the stochastic dynamics of the scenario process and the evolution of information, the filtration (cf. Pflug [22]). It is therefore the goal to construct trees that represent the underlying stochastic processes and their information structure in the best way. To this end, there are many heuristic algorithms available, like stochastic approximation [22], clustering (cf. Gülpınar, Rustem, and Settergren [10]), moment matching (cf. Høyland and Wallace [15]), Monte Carlo sampling (cf. Kim [18]), large tree sampling together with tree reduction (cf. Heitsch and Römisch [12, 13, 14]), and more. None of these methods allow quantifying the approximation error, and thus convergence results are not available.

The authors have introduced in [23, 26] a distance for stochastic processes and designed algorithms for minimizing this distance for a given distribution of the underlying process when constructing a finite tree approximation. They use a *parametric* approach by first estimating a parametric model from the data (for instance, a GARCH-model), and then use random number generation from the estimated conditional distributions to construct an approximating tree. In this paper we investigate a *nonparametric* technique, which completely avoids specifying a parametric model, but which works on the raw data, i.e., replications of finitely many observed trajectories—see Figure 1. This approach is new.

The following section reviews the *nested distance* or *process distance* for stochastic processes introduced in Pflug and Pichler [24]. This concept correctly captures these subtle and essential characteristics of conditional transition probabilities and evolution of information, as is relevant for multistage stochastic optimization. We prove that the empirical measure (in general) is inconsistent in nested distance topology if the process has a density. In contrast, we show that one may construct tree models which are consistent in nested distance. To this end, we propose building trees using

multivariate kernel density and conditional density estimation.

We prove that approximations obtained in this way indeed converge in probability to the genuine process if n, the number of observed paths, tends to infinity.

Outline of the paper. The following section (section 2) covers the nested distance, an extension of the Wasserstein distance. We illustrate the inconsistency of the empirical measure in nested distance. We prove further that nonbranching approximations (fans) are not adequate data models for stochastic optimization problems.

Section 3 introduces kernel density estimation and states the results needed to obtain trees from empirical data. Section 4 relates the nested distance and kernel density estimation. Finally, section 5 establishes the main result of this paper, which is convergence of the appropriately smoothed empirical process to the original process in probability and in nested distance. We conclude with an algorithm in section 6, which exploits our results for scenario tree generation. This final section presents some selected examples.

2. Distance concepts for probability measures and stochastic processes. In what follows we introduce the nested distance to measure the distance of stochastic processes in discrete time. By employing the central theorem for multistage stochastic optimization (Theorem 1, below) we prove first that the empirical measure does not in general converge in nested distance to the initial process.

2.1. The nested distance. The nested distance is a distance for filtered probability spaces. It extends the Wasserstein distance, a transportation distance for probability spaces on metric (Polish) spaces (Ξ, d).

DEFINITION 1 (nested distance, also process distance). *Let*

$$\mathbb{P} := \big(\Xi, (\Sigma_t)_{t=0,\dots,T}, P\big) \quad \textit{and} \quad \tilde{\mathbb{P}} := \big(\Xi, (\tilde{\Sigma}_t)_{t=0,\dots,T}, \tilde{P}\big)$$

be filtered probability spaces (a.k.a. stochastic bases*). The* nested distance *(also process, or multistage distance) of order $r \geq 1$ is defined by*

$$\mathsf{dl}_r(\mathbb{P}, \tilde{\mathbb{P}})^r := \inf \iint_{\Xi\times\Xi} \mathsf{d}(x,y)^r \pi(\mathrm{d}x, \mathrm{d}y), \tag{1}$$

where the infimum is among all probability measures π with conditional marginals P and $\tilde{P}$; i.e.,

$$\pi\big(A \times \Xi \big|\, \Sigma_t \otimes \tilde{\Sigma}_t\big) = P\big(A \big|\, \Sigma_t\big) \quad \textit{and} \tag{2}$$

$$\pi\big(\Xi \times B \big|\, \Sigma_t \otimes \tilde{\Sigma}_t\big) = \tilde{P}\big(B \big|\, \tilde{\Sigma}_t\big) \quad \textit{for all } t = 0, \dots, T, \tag{3}$$

whenever $A \in \Sigma_T$ and $B \in \tilde{\Sigma}_T$.

Remark 1. The conditional probability $P\big(A|\,\Sigma_t\big)$ is a random variable on the space Ξ, while $\pi\big(A \times \Xi|\,\Sigma_t \otimes \tilde{\Sigma}_t\big)$ is a random variable on $\Xi \times \Xi$. The identity (2) thus is understood as $\pi\big(A \times \Xi|\,\Sigma_t \otimes \tilde{\Sigma}_t\big) = P\big(A|\,\Sigma_t\big) \circ i_1$, where $i_1(\xi_1, \xi_2) := \xi_1$ is the natural projection. The projection $i_2(\xi_1, \xi_2) := \xi_2$ takes the role of i_1 in (3).

Remark 2. If $T = 1$ and if the filtration consists of just the trivial sigma algebras $\Sigma = (\Sigma_0, \Sigma_1)$ with $\Sigma_0 = \tilde{\Sigma}_0 = \{\emptyset, \Xi\}$ and $\Sigma_1 = \tilde{\Sigma}_1 = \mathscr{B}(\Xi)$, the Borel sets, then the constraints (2) and (3) read

$$\pi(A \times \Xi) = P(A) \quad \text{and} \quad \pi(\Xi \times B) = \tilde{P}(B);$$

i.e., the sigma algebras can be dropped. This is the usual notion of the *Wasserstein distance*, such that the Wasserstein distance of order r ($r \geq 1$) represents a special case of the nested distance of processes with a deterministic ξ_0 and a stochastic ξ_1. We denote the Wasserstein distance of order $r \geq 1$ by d_r to distinguish it from dl_r, the nested distance.

Remark 3. A detailed discussion of the Wasserstein distance can be found in Rachev and Rüschendorf [30, 31], as well as in Villani [38]. Occasionally we shall also write $\mathsf{dl} = \mathsf{dl}_1$ and $\mathsf{d}_1 = \mathsf{d}$ for the distance of order $r = 1$.

The nested distance is designed to capture also the evolution of the information of the stochastic processes over time. It is the crucial and determining distance for stochastic optimization problems. The nested distance was introduced in Pflug [23] for nested distributions. Its dual formulation as well as basic properties are elaborated in Pflug and Pichler [24].

Definition 1 involves a (continuous) distance function d in (1). However, much more general cost functions can be considered here, which are defined, e.g., on different spaces. Beiglböck, Léonard, and Schachermayer [3] consider the Wasserstein distance for general measurable cost functions.

Remark 4. The Wasserstein distance generalizes naturally to a distance of random variables by considering the induced pushforward measures. Indeed, if $\xi : \Omega \to \Xi$ and $\tilde{\xi} : \tilde{\Omega} \to \Xi$ are random variables on (Ω, P) ($(\tilde{\Omega}, \tilde{P})$, resp.) with the same metric state space Ξ, then the pushforward measures $P \circ \xi^{-1}$ and $P \circ \tilde{\xi}^{-1}$ are measures on Ξ. In this way the Wasserstein distance of $P \circ \xi^{-1}$ and $\tilde{P} \circ \tilde{\xi}^{-1}$ provides a distance for the distributions of the random variables ξ and $\tilde{\xi}$.

The nested distance generalizes naturally to a distance of stochastic processes in a way analogous to how the Wasserstein distance generalizes to a distance of random variables (cf. above). For this consider the law $P \circ \xi^{-1}$ ($\tilde{P} \circ \tilde{\xi}^{-1}$, resp.) of the process $\xi : \Omega \to \times_{t=0,\dots,T} \Xi_t$ ($\tilde{\xi} : \tilde{\Omega} \to \times_{t=0,\dots,T} \Xi_t$, resp.). The nested distance of the laws $P \circ \xi^{-1}$ and $\tilde{P} \circ \tilde{\xi}^{-1}$ is thus a distance for the distributions of the stochastic processes ξ and $\tilde{\xi}$.

Convention in this paper. In what follows we restrict ourselves to the filtered probability spaces on

$$\Xi = \mathbb{R}^{m_0} \times \mathbb{R}^{m_1} \times \cdots \times \mathbb{R}^{m_T}, \tag{4}$$

and we set $M := m_0 + \cdots + m_T$ for the entire dimension. The filtrations considered consist of the sigma algebras,

$$\Sigma_t := \sigma\left(\xi_0, \dots, \xi_t\right), \tag{5}$$

generated by process $\xi = (\xi_0, \dots, \xi_T)$, where $\xi_t \in \mathbb{R}^{m_t}$ (and analogously for $\tilde{\Sigma}_t$). Throughout the paper we assume that $\xi_0 = \tilde{\xi}_0$ is deterministic and that $\Sigma_0 = \{\emptyset, \Xi\}$ is the trivial sigma algebra; we thus omit the 0th component occasionally. We shall assume further that the distance on Ξ is induced by some norm, $\mathsf{d}(x, y) = \|y - x\|$.

With double struck letters like $\mathbb{P}$ we denote structures like $(\Xi, (\Sigma_t), P)$, which contain the filtration as an integral part; when ignoring the filtration, we would just write P, the probability measure alone. While the nested distance is defined for objects like $\mathbb{P}$ and $\tilde{\mathbb{P}}$, the ordinary Wasserstein distance is defined for probabilities P and $\tilde{P}$ on the metric space Ξ.

2.2. The empirical measure does not converge. The nested distance is adapted for stochastic optimization problems. Indeed, the following main theorem (contained in [24, Theorem 11]; cf. also [27, 28]) establishes that optimal values of stochastic optimization problems are continuous with respect to the nested distance. We employ this result to demonstrate that the empirical measure is inconsistent.

THEOREM 1 (continuity of stochastic optimization problems). *Let* $\mathbb{P} := \big(\Xi, (\Sigma_t)_{t=0,\dots,T}, P\big)$ *and* $\tilde{\mathbb{P}} := \big(\Xi, (\tilde{\Sigma}_t)_{t=0,\dots,T}, \tilde{P}\big)$ *be filtered probability spaces. Consider the multistage stochastic optimization problem*

$$v(\mathbb{P}) := \inf\left\{\mathbb{E}_P Q(x,\xi) : x \lhd \sigma(\xi)\right\}, \tag{6}$$

where Q *is convex in* x *for any* ξ *fixed, and Lipschitz with constant* L *in* ξ *for any* x *fixed. Then*

$$\left|v(\mathbb{P}) - v(\tilde{\mathbb{P}})\right| \le L \cdot \mathsf{dl}_r(\mathbb{P}, \tilde{\mathbb{P}})$$

for every $r \ge 1$.

The constraint $x \lhd \sigma(\xi)$ is shorthand for x_t *is nonanticipative*, i.e., *measurable with respect to* $\Sigma_t = \sigma(\xi_1, \dots, \xi_t)$ *for all* $t = 0, \dots, T$, where $x = (x_t)_{t=0}^T$ in (6) is the (stochastic) decision process. By the Doob–Dynkin lemma (cf. Kallenberg [17]), the constraint $x \lhd \sigma(\xi)$ forces x to be a function of the process ξ; i.e., there are measurable functions x_t' such that the feasible process x_t in (6) can be written as $x_t = x_t'(\xi_0, \dots, \xi_t)$ (i.e., $x_t(\cdot) = x_t'(\xi_0(\cdot), \dots, \xi_t(\cdot))$).

Discrete measures. The empirical measure of the independent and identically distributed (i.i.d.) observations

$$\begin{aligned} \xi_1 &= (\xi_{1,0}, \dots, \xi_{1,T}) \\ &\vdots \\ \xi_n &= (\xi_{n,0}, \dots, \xi_{n,T}) \end{aligned} \tag{7}$$

is

$$P_n := \frac{1}{n}\sum_{i=1}^n \delta_{\xi_i} = \frac{1}{n}\sum_{i=1}^n \delta_{(\xi_{i,0},\dots,\xi_{i,T})} \tag{8}$$

on $\mathbb{R}^M$, where each $\xi_i = (\xi_{i,0}, \dots, \xi_{i,T})$ is an observation of an entire sample path and δ_x is the point mass at x.[1] The empirical measure is a special case of a random discrete measure.

Remark 5. Discrete measures are—with respect to the Wasserstein distance—dense in the space of measures satisfying an adequate moment constraint (see, e.g., Bolley [4] for details). Also, empirical measures converge a.s. to the underlying measure in the Wasserstein distance. The following proposition outlines that this property is no longer valid for multistage empirical processes and the nested distance. To resolve this issue we will replace the original empirical measures by smoothed versions later.

We have the following negative result.

[1] Notice that all $\xi_{i,0}$ are identical, since the starting value is deterministic.

PROPOSITION 1. *Consider the space $\Xi = \mathbb{R}^M$ (cf. (4)) equipped with its natural filtration Σ_t introduced in (5). Suppose that P has a density on $\mathbb{R}^M$ and $T \geq 2$. Then the filtered spaces $\mathbb{P}_n := (\Xi, (\Sigma_t)_{t=1,\dots,T}, P_n)$, equipped with the empirical measure $P_n := \frac{1}{n}\sum_{i=1}^n \delta_{\xi_i}$, do not converge in nested distance to $\mathbb{P} := (\Xi, (\Sigma_t)_{t=1,\dots,T}, P)$ a.s.*

Proof. Consider a pair (ξ_1, ξ_2) which is distributed according to P, which is assumed to have a density and finite first moments.[2] Let Σ_1 be the σ-algebra generated by ξ_1. We aim at solving the optimal prediction problem

$$v(\mathbb{P}) = \min\{\mathbb{E}_P[|\xi_2 - x|] : x \lhd \Sigma_1\} \tag{9}$$

for the underlying model and for its empirical approximation. Notice that one may solve (9) by decomposing it into the conditional problem

$$\min_{x_1 \lhd \Sigma_1} \mathbb{E}_P\left[|\xi_2 - x_1| \,\middle|\, \Sigma_1\right],$$

which has the optimal decision $x_1(\xi_1) = \mathrm{med}(\xi_2|\xi_1)$, where $\mathrm{med}(\xi_2|\xi_1)$ is the conditional median of ξ_2 given ξ_1, with optimal value

$$v(\mathbb{P}) = \mathbb{E}[\xi_2 - \mathrm{med}\ (\xi_2|\xi_1)|] =: c \qquad \text{(say)},$$

where $c > 0$.

Consider the empirical measure $P_n = \frac{1}{n}\sum_{i=1}^n \delta_{\xi_i}$, and recall that by assumption all $\xi_i = (\xi_{i,1}, \xi_{i,2})$ are different with probability 1. Then problem (9), formulated for the measure P_n, can also be decomposed into the conditional problems

$$\min_{x_1 \lhd \mathcal{F}_1} \mathbb{E}_{P_n}\left[|\xi_{i,2} - x_1| \,\middle|\Sigma_1\right],$$

and this problem has the optimal solution

$$x_1(\xi_{i,1}) = \begin{cases} \xi_{i,2} & \text{if } \xi_1 = \xi_{i,1}, \\ \text{arbitrary} & \text{otherwise.} \end{cases}$$

Note that $x_1(\cdot)$ is well defined, as all $\xi_{i,1}$ are mutually different by assumption. Obviously, the optimal value of (9) is

$$v(\mathbb{P}_n) = 0.$$

Now, according to Theorem 1 and observing that the objective function $(x, \xi_2) \mapsto |\xi_2 - x|$ is Lipschitz 1 in ξ_2 and convex in x, we have that

$$|v(\mathbb{P}) - v(\mathbb{P}_n)| \leq \mathsf{dl}(\mathbb{P}, \mathbb{P}_n),$$

where $\mathbb{P}$ ($\mathbb{P}_n$, resp.) are the filtered spaces pertaining to P and P_n, respectively. Since

$$0 < c = |v(\mathbb{P}) - v(\mathbb{P}_n)| \leq \mathsf{dl}(\mathbb{P}, \mathbb{P}_n)$$

for all n, $\mathbb{P}_n$ does not converge to $\mathbb{P}$ in the nested distance sense. ∎

[2] If the distributions do not have finite first moments, the nested distance is not explained.

Remark 6. It is well known that the empirical measure converges a.s. weakly to the underlying distribution on separable metric spaces (see Varadarajan [37]). Under the assumption of finite rth moments (i.e., that $\int \mathsf{d}(x_0, x)^r \, P(\mathrm{d}x) < \infty$ for some x_0), also the a.s. convergence in Wasserstein distance holds. Define the Wasserstein distance for processes as in (1), but without the constraints (2) and (3),

$$\mathsf{d}_r(\mathbb{P}, \tilde{\mathbb{P}})^r := \inf \iint_{\Xi \times \Xi} \mathsf{d}(x, y)^r \pi(\mathrm{d}x, \mathrm{d}y),$$

where π runs through all joint probability measures with marginals $\mathbb{P}$ and $\tilde{\mathbb{P}}$. Then $\mathsf{d}_r(\mathbb{P}, \tilde{\mathbb{P}})^r \le \mathsf{dl}_r(\mathbb{P}, \tilde{\mathbb{P}})^r$, and for the empirical measure $\hat{\mathbb{P}}_n$ we have that

$$\mathsf{d}_r(\hat{\mathbb{P}}_n, \mathbb{P}) \to 0$$

a.s. for $n \to \infty$. However, convergence in d_r does not imply convergence in dl_r and of the conditional distributions. Even if $\mathsf{d}_r(\mathbb{P}, \tilde{\mathbb{P}})^r = 0$, the information structures (generated filtrations) of $\mathbb{P}$ and $\tilde{\mathbb{P}}$ may be quite different.

Trees versus fans. We call a stochastic process in discrete time and discrete space a *(stochastic) tree*. A tree for which every internal node except possibly the root has only one successor is called a *fan*. The empirical measure based on n samples of the process is a fan (with probability 1).

Notice that the filtration induced by a fan is quite degenerate: as of time 1, the full information is available, and no increase of information takes place later; i.e., $\tilde{\Sigma}_1 = \cdots = \tilde{\Sigma}_T$ in terms of the sigma algebras carrying the information. In contrast, "usual" trees, which are the usual data structures for adequately handling approximations of stochastic processes on filtered spaces, have to branch at each stage.

The negative statement contained in Proposition 1 is not a shortfall of the nested distance. On the contrary, the counterexample shows that the nested distance captures a fundamental and characterizing property of stochastic optimization problems by correctly distinguishing between processes with different information structures. Indeed, the standard empirical measure carries the full information already at the very beginning of the process, as the remaining paths are already determined by the first observation. Thus, the empirical process does not gather information over time as the underlying process does.

The nested distance is designed to recognize and quantify the amount of information available for the following decisions. Hence, the nested distance of a process with a density and the empirical process cannot vanish, as is the content of Proposition 1.

3. Convolution and density estimation. The previous section demonstrates that empirical measures are not adequate models for approximating a stochastic process for stochastic optimization. In what follows we construct scenario trees to approximate stochastic processes. However, the scenario trees are constructed from the samples observed without involving additional knowledge. In this way the samples are exploited to find discrete time and discrete space approximations, which are necessary for computation.

To do this, we smooth the observations $(\xi_i)_{i=1}^n$ by convoluting them with a prespecified kernel, as is known from density estimation. We demonstrate that when an appropriate amount of blurring is introduced, the paths with a similar past can no longer be distinguished. This allows for the possibility of continuations different from those associated with a single path. It is exactly this property which is essential for correctly specifying the evolution of information in multistage settings.

This is outlined in the following sections. The next section reviews kernel density estimation first, particularly the estimation of conditional densities, as they turn out to be important to sample conditionally on some specified history.

3.1. Convolution of measures. The density of the sum of two random variables is given by the convolution of the individual densities. Here we introduce the convolution for measures to formulate the results for kernel density estimation.

Recall that the convolution measure of two measures P and Q is the measure $P * Q$, defined as the pushforward of the addition $(+)$ with respect to the product measure; i.e.,

$$(P * Q)(A) = \iint \mathbb{1}_A(x+y)P(\mathrm{d}x)Q(\mathrm{d}y) \qquad (A \text{ measurable}). \tag{10}$$

The convolution of measures is commutative, $P * Q = Q * P$, as the addition commutes. The convolution with a Dirac measure $\delta_x(\cdot)$ is the shifted measure $P * \delta_{x_0}(A) = P(A - x_0)$, where $A - x_0 := \{a - x_0 : a \in A\}$.

DEFINITION 2. *With a density function k on $\mathbb{R}^m$ we associate the parametric family of densities $k_h(x) := \frac{1}{h^m}k(x/h)$ on $\mathbb{R}^m$, $h > 0$. If h is not a positive scalar but a vector with positive entries $h = (h^{(1)}, \dots, h^{(m)})$, then $k_h(x) := \frac{1}{h^{(1)} \cdot \dots \cdot h^{(m)}} k\left(\frac{x_1}{h^{(1)}}, \dots, \frac{x_m}{h^{(m)}}\right)$. k_h again is a density on $\mathbb{R}^m$. However, for the sake of a simpler presentation, we assume that the bandwidth vector is $(h, h, \dots, h)$.*

Remark 7 (notational convention). We shall write P^f for the measure induced by the Lebesgue density f,

$$P^f(A) := \int_A f \mathrm{d}\lambda.$$

The convolution of the measure with density k_h with a (weighted) discrete measure

$$\tilde{P}_n = \sum_{i=1}^n w_i \cdot \delta_{\xi_i} \tag{11}$$

on $\mathbb{R}^m$ has the density

$$\sum_{i=1}^n w_i \cdot \frac{1}{h^m} k\left(\frac{x - \xi_i}{h}\right). \tag{12}$$

The usual Parzen–Rosenblatt kernel density estimator is a particular case with n independent draws $(\xi_i)_{i=1}^n$ from P and equal weights $w_i = \frac{1}{n}$. The density associated with the empirical measure $\hat{P}_n := \frac{1}{n}\sum_{i=1}^n \delta_{\xi_i}$ is

$$\hat{f}_{k_{h_n}}(\cdot) := \frac{1}{n\, h_n^m} \sum_{i=1}^n k\left(\frac{\cdot - \xi_i}{h_n}\right), \tag{13}$$

where the bandwidth h_n may depend on n. Employing the notational convention $h = h_n$, we can write $P^{\hat{f}_{k_h}} = \hat{P}_n * k_h$.

In what follows we shall consider a fixed kernel function k. For this reason we sometimes omit the index k in the notation and write (for instance) $\hat{f}_n$ instead of $\hat{f}_{k_{h_n}}$ if no confusion is possible.

3.2. Multivariate density estimation. We address important convergence theorems from multivariate kernel density estimation first. These results turn out to be essential in extracting scenario trees out of samples. The general assumption for kernels is that

$$\int u_i \, k(u) \, \mathrm{d}u = 0 \tag{14}$$

for all i.

The bias term. The bias of the density estimator $\hat{f}_n$ can be expressed as

$$\mathbb{E}\hat{f}_n(x) = \int k_{h_n}(x-y) \, f(y) \, \mathrm{d}y = f * k_{h_n}(x), \tag{15}$$

where $*$ denotes the convolution of densities. It follows from (15) that $\hat{f}_n(x)$ is biased in general. The bias can be stated as

$$\begin{aligned} \mathsf{bias}\hat{f}_n(x) &:= \mathbb{E}\hat{f}_n(x) - f(x) = \frac{1}{h_n^m} \int k\left(\frac{x-y}{h_n}\right) \big(f(y) - f(x)\big) \, \mathrm{d}y \\ &= \int k\,(u) \, \big(f(x - h_n \cdot u) - f(x)\big) \, \mathrm{d}u. \end{aligned} \tag{16}$$

It is evident that $\mathbb{E}\hat{f}_n(x) \to f(x)$ whenever $h_n \to 0$ and if x is a point of continuity of f. Further, by assuming that f is smooth and employing a Taylor series expansion, (16) reduces to

$$\begin{aligned} \mathsf{bias}\hat{f}_n(x) &= \int k\,(u) \left(f(x) - f'(x)^\top h_n u + \frac{1}{2}(h_n u)^\top f''(x)(h_n u) - f(x) + o(h_n^2) \right) \mathrm{d}u \\ &= \frac{1}{2} h_n^2 \sum_{i,j=1}^{m} \big(f''_{i,j}(x) \cdot \kappa_{i,j}\big) + o\big(h_n^2\big) \end{aligned} \tag{17}$$

whenever (14) holds and where κ is the matrix with entries $\kappa_{i,j} = \int u_i u_j k(u) \, \mathrm{d}u$. Note that expression (16) and the approximation (17) are deterministic quantities; they do not involve any random component. Instead, the bias depends on the density function f and its smoothness or (local) differentiability. Moreover, it should be noted that the bias tends to 0 in (16) and (17), provided that $h_n \to 0$.

Convergence. The variance of the multivariate kernel statistics is

$$\begin{aligned} \mathsf{var}\hat{f}_n(x) &= \mathsf{var} \frac{1}{n\,h^m} \sum_{i=1}^{n} k\left(\frac{x-\xi_i}{h_n}\right) = \frac{1}{n} \mathsf{var} \frac{1}{h^m} k\left(\frac{x-\xi_1}{h_n}\right) \\ &= \frac{1}{n} \int \frac{1}{h^{2m}} k\left(\frac{x-y}{h_n}\right)^2 f(y) \, \mathrm{d}y - \frac{1}{n} \left(\mathbb{E} \frac{1}{h_n^m} k\left(\frac{x-\xi_1}{h_n}\right) \right)^2 \\ &= \frac{1}{n\,h^m} \int k\,(u)^2 \, f(x - h \cdot u) \, \mathrm{d}u - \frac{1}{n} \big(\mathbb{E} f_n(x)\big)^2 \\ &= \frac{f(x)}{n h^m} \int k(u)^2 \mathrm{d}u - \frac{1}{n} \big(\mathbb{E}\hat{f}_n(x)\big)^2 + o\left(\frac{1}{n\,h^m}\right), \end{aligned}$$

and the *mean-squared error* is given by

$$\mathsf{MSE}\,\hat{f}_n(x) := \mathbb{E}\big(\hat{f}_n(x) - f(x)\big)^2 = \mathsf{bias}^2 \hat{f}_n(x) + \mathsf{var}\hat{f}_n(x).$$

To minimize the mean-squared error with respect to the bandwidth h_n it is advantageous to get rid of the mixed terms $h_i h_j$ $(i \neq j)$ in (17) for the bias. This can be accomplished by assuming that k has uncorrelated components; i.e.,

$$\kappa_{i,j} = \int u_i u_j k(u) \mathrm{d}u = 0 \quad \text{whenever } i \neq j. \tag{18}$$

Then the mean-squared error is minimized for

$$h_n^{m+4} \simeq \frac{m}{n} \cdot \frac{f(x) \cdot \int k(u)^2 \, \mathrm{d}u}{\left(\sum_{i=1}^m f_{x_i x_i} \kappa_{i,i}\right)^2}. \tag{19}$$

If, instead of the mean-squared error at a specific point x, the *mean integrated square error*

$$\mathsf{MISE}\, \hat{f}_n := \int \mathsf{MSE}\, \hat{f}_n(x) \mathrm{d}x = \mathbb{E} \int \big(\hat{f}_n(x) - f(x)\big)^2 \mathrm{d}x$$

is to be minimized, then the optimal bandwidth is

$$h_n^{m+4} \simeq \frac{m}{n} \cdot \frac{\int k(u)^2 \, \mathrm{d}u}{\left(\sum_{i=1}^m \kappa_{i,i} \int f_{x_i x_i} \, \mathrm{d}x\right)^2}, \tag{20}$$

which is of the same order as (19).[3]

Remark 8. Assumption (18) is an assumption on the kernel k. Any kernel exhibiting the product form

$$k(u) = k_1(u_1) \cdot k_2(u_2) \cdot \cdots \cdot k_m(u_m) \tag{21}$$

satisfies this assumption. The bias (17) of a product kernel of the particular form (21) reduces to

$$\mathsf{bias} \hat{f}_n(x) = \frac{\kappa_2}{2} \sum_{s=1}^m h_n^2 f_{x_s x_s}(x) + o\big(h_n^2\big),$$

where

$$\kappa^{(2)} := \int u^2 k(u) \mathrm{d}u \tag{22}$$

is the second moment (or variance) of the distribution associated with the kernel.

Remark 9. Both formulae ((19) and (20)) for the asymptotic optimal bandwidth involve f'', the Hessian of the density function f. As the function f is unknown (this is what kernel density estimation intends to estimate), the formulae provide the correct asymptotic order, but the optimal constant remains an oracle (cf. Tsybakov [36]). Different methods of obtaining an optimal bandwidth as cross-validation are designed to overcome this difficulty and outlined in Racine, Li, and Zhu [32], e.g., or the plug-in rules of Sheather [33].

[3]Note, that $\sum_{i=1}^m \kappa_{i,i} f_{x_i x_i} = \operatorname{div}(\kappa \cdot \nabla f)$ and $\sum_{i=1}^m \kappa_{i,i} f_{x_i x_i} = \kappa \, \Delta f$ (the Laplace operator) for constant $\kappa_{i,i} = \kappa$.

Uniform consistency. The previous sections investigate the density f at a fixed point x. It will be important to have a result with uniform convergence at hand as well. This is accomplished by the following theorem, which is presented in a more general form in Giné and Guillou [8, Proposition 3.1] (cf. also Stute [35] and Wied and Weißbach [39, Theorem 2]).

THEOREM 2 (uniform consistency). *Suppose that the kernel k is nonnegative and compactly supported on $\mathbb{R}^m$, the density f is bounded and uniformly continuous, and the bandwidth sequence satisfies*

$$h_n \to 0, \quad \frac{nh_n^m}{|\log h_n|} \to \infty, \quad \frac{|\log h_n|}{\log\log n} \to \infty, \quad \textit{and} \quad nh_n^m \to \infty; \tag{23}$$

then

$$\lim_{n\to\infty} \sqrt{\frac{nh_n^m}{\log h_n^{-m}}} \cdot \left\| \hat{f}_n - \mathbb{E}\hat{f}_n \right\|_D = \|k\|_2 \sqrt{2\,\|f\|_D} \qquad a.s., \tag{24}$$

where $\|f\|_D = \sup_{x\in D} |f(x)|$ is the supremum norm on an open set D.

Remark 10. Einmahl and Mason outline in [7] that the result of Theorem 2 does not even require continuity of f, and asymptotic uniform consistency

$$\left\| \hat{f}_n - \mathbb{E}\hat{f}_n \right\|_D = \mathcal{O}\left(\frac{\log h_n^{-m}}{n\, h_n^m} \right)$$

still holds true whenever f is bounded.

We emphasize as well the fact that the limit in (24) exists *almost everywhere*.

3.3. Conditional density estimation. This section follows Li and Racine [19]. Suppose that the density of the multivariate pair (X, Y) is $f(x, y)$. The conditional density of the random variable $X|Y = y$ is

$$f(x|y) = \frac{f(x,y)}{f(y)}, \quad \text{where } f(y) = \int f(x,y)\,\mathrm{d}x \tag{25}$$

(here Y is the explanatory variable in (25), and X is explained). By employing a product kernel $k(x, y) = k(x) \cdot k(y)$, the density estimator for the multivariate density based on a sample (X_i, Y_i) is found to be

$$\hat{f}_n(x,y) = \frac{1}{n} \sum_{i=1}^{n} k_{h_n}(x - X_i) \cdot k_{h_n}(y - Y_i),$$

and the marginal density estimate has the closed form $\hat{f}_n(y) = \int \hat{f}_n(x,y)\,\mathrm{d}x = \frac{1}{n}\sum_{i=1}^{n} k_{h_n}(y - Y_i)$. It follows that

$$\begin{aligned} \hat{f}_n(x|y) &:= \frac{\hat{f}_n(x,y)}{\hat{f}_n(y)} = \sum_{i=1}^{n} \frac{k_{h_n}(y - Y_i)}{\sum_{j=1}^{n} k_{h_n}(y - Y_j)} \cdot k_{h_n}(x - X_i) \\ &= \sum_{i=1}^{n} \frac{\frac{1}{h_n^{m_y}} k\left(\frac{y - Y_i}{h_n}\right)}{\sum_{j=1}^{n} \frac{1}{h_n^{m_y}} k\left(\frac{y - Y_j}{h_n}\right)} \cdot \frac{1}{h_n^{m_x}} k\left(\frac{x - X_i}{h_n}\right) \end{aligned} \tag{26}$$

is a density again, where h_n is the common bandwidth for the variables $(X_i, Y_i) \in \mathbb{R}^{m_x} \times \mathbb{R}^{m_y}$. The estimator (26) for the conditional density can be rewritten as

$$\hat{f}_n(x|y) = \sum_{i=1}^{n} w_i^{(n)}(y) \cdot k_{h_n}(x - X_i), \quad \text{where } w_i^{(n)}(y) := \frac{k\big(\frac{y-Y_i}{h_n}\big)}{\sum_{j=1}^{n} k\big(\frac{y-Y_j}{h_n}\big)} \tag{27}$$

are the weights corresponding to the conditioning y. The conditional estimator (27) is of the same type as the kernel estimator (13), except that the weights are $w_i^{(n)}(y)$ instead of $1/n$. Notice that the Nadaraya–Watson estimator (cf. Tsybakov [36]) is of the same type as (27).

Asymptotic normality. Note that $\hat{f}_n(x|y)$ is the density of the measure

$$\left(\hat{P}_n * k_h\right)(A|\, y) = \int_A \hat{f}_n(x|y)\, \mathrm{d}x, \quad A \in \mathscr{B}\big(\mathbb{R}^{m_x}\big),$$

with $\hat{P}_n = \sum_{i=1}^{n} \frac{1}{n} \delta_{(X_i, Y_i)}$ (according the disintegration theorem).

Both estimates, $\hat{f}_n(x, y)$ and $\hat{f}_n(x)$ converge in distribution to the respective true values. These ingredients can be combined for the expression

$$\sqrt{n\, h_n^{m_x+m_y}} \left(\hat{f}_n(x|y) - f(x|y) - \frac{\kappa_{(2)}}{2} h_n^2 B(x, y)\right) \xrightarrow{d} \mathcal{N}\left(0,\, \kappa_{(2)}^{m_x+m_y} \frac{f(x|y)}{f(x)}\right) \tag{28}$$

of asymptotic normality of the conditional density. Although the expectation of $\hat{f}_n(x|y)$ does not have a closed form like (15), the bias term in (28) is

$$B(x, y) = \sum_{s=1}^{m_y} \frac{f_{y_s y_s}(x, y) - f(x|y) \cdot f_{y_s y_s}(y)}{f(y)} + \sum_{s=1}^{m_x} \frac{f_{x_s x_s}(x, y)}{f(y)}.$$

Formula (28) and the asymptotic normality of the conditional density (25) are again elaborated in Li and Racine [19, Theorem 5.5] together with the optimal bandwidth selection

$$h_n \simeq \frac{1}{n^{1/(m_x+m_y+4)}}.$$

We may refer to Hyndman, Bashtannyk, and Grunwald [16] for a further discussion on the integrated mean-squared error.

4. Relations of the Wasserstein distance to density estimation. Density estimation recovers a density function from samples at a specified point. In this sense the Parzen–Rosenblatt estimator (13) provides a *local* approximation of the density function, and the uniform result outlined in Theorem 2 measures approximations locally as well.

In contrast, the Wasserstein distance takes notice of the distance of individual samples by involving $\mathsf{d}(x, y)$ in Definition 1. In this sense, the Wasserstein distance relates distant points and does not consider only the approximation quality locally. From this perspective it may seem unnatural to combine density estimation and the Wasserstein distance. However, they have an important point in common: if two densities are close, then the Wasserstein distance will not move the mass located under both densities—this is a consequence of the triangle inequality. We exploit this fact in what follows to establish relationships between density estimation and approximations in the Wasserstein distance.

The following subsection elaborates that convolution is continuous in terms of the Wasserstein distance. We further present bounds for the Parzen–Rosenblatt estimator in terms of the Wasserstein distance.

The reverse inequalities are more delicate. We will require that the probability measure have bounded support (cf. Proposition 4 below).

4.1. The empirical measure and the convolution. We establish first that convolution is a continuous operation in the Wasserstein distance in the following sense.

LEMMA 1. *For a translation invariant distance* d *(i.e.,* $\mathsf{d}(x+z, y+z) = \mathsf{d}(x,y)$*) it holds that*

$$\mathsf{d}_r(\tilde{P} * k_h,\, P) \leq \mathsf{d}_r(\tilde{P}, P) + \kappa_r^{1/r} \cdot \max_{i=1,\ldots,m} h_i,$$

where $\kappa_r = \int \|x\|^r \, k(x)\mathrm{d}x$ *is the* r*th absolute moment of the kernel* k*.*

Proof. We include a proof in Appendix A. □

Bounds for the convolution density. Following Bolley, Guillin, and Villan [5], we have the following relation between the densities and the Wasserstein distance of the measures P and its smoothed empirical measure $\hat{P}_n$. Again, this result gives rise to oversmoothing, as the subsequent remark outlines.

PROPOSITION 2. *Let* P *be a measure on* $\mathbb{R}^m$ *with density* f*. Suppose the kernel is Lipschitz with constant* $\|k\|_{Lip}$ *and supported in the unit ball,* $\{k(\cdot) > 0\} \subseteq \{\|\cdot\| \leq 1\}$*. Then the kernel density estimator* $\hat{f}_n$ *corresponding to* $\hat{P}_n * k_{h_n}$ *satisfies*

$$\left\| \hat{f}_n - f \right\|_\infty \leq \delta_f(h) + \frac{\|k\|_{Lip}}{h^{m+1}} \mathsf{d}_r(P, \hat{P}_n) \tag{29}$$

(i.e., the distance is uniformly small on the support $\mathbb{R}^m$*) for every* $r \geq 1$*. Here*

$$\delta_f(h) := \sup_{\{\|x-y\| \leq h\}} |f(x) - f(y)|$$

is the modulus of continuity of the density f*.*

For the proof of Proposition 2, we refer the reader to Bolley, Guillin, and Villan [5, Proposition 3.1] or to the appendix.

Remark 11 (oversmoothing). Suppose that the density f is Lipschitz continuous as well; then $\delta_f(h) = \|f\|_{Lip} \cdot h$. Suppose further that P_n is chosen such that $\mathsf{d}_r(P, P_n) \sim c \cdot n^{-1/m}$; then the optimal rate in (29) is

$$h_n \sim \left(\frac{c(m+1)}{\|f\|_{Lip}} \right)^{\frac{1}{m+2}} n^{-\frac{1}{m(m+2)}} \tag{30}$$

and

$$\left\| \hat{f}_n - f \right\|_\infty \sim n^{-\frac{1}{m(m+2)}} \to 0$$

such that the density of the smoothed discrete distribution converges. Convergence, however, is slow, particularly for large m.

The traditional bandwidth of the kernel density estimator has order $h_n = n^{-1/(m+4)}$ (cf. (19) and (20) above). As $\frac{1}{m(m+2)} < \frac{1}{m+4}$, the bandwidth (30) oversmoothes the density f.

The following proposition relates the L_2-distance of densities to the Wasserstein distance.

PROPOSITION 3. Let f and g be densities on $\mathbb{R}^m$. Then the squared L_2-distance is bounded by

$$\int \big(f(x)-g(x)\big)^2 \mathrm{d}x \le \|f-g\|_{Lip} \cdot \mathsf{d}_r\left(P^f, P^g\right)$$

for every $r \ge 1$.

Proof. Let X be a random variable with density f, and Y have density g. Then

$$\begin{aligned}
&\int \big(f(x)-g(x)\big)^2 \,\mathrm{d}x \\
&\qquad = \int f(x)f(x)\mathrm{d}x - \int f(x)g(x)\,\mathrm{d}x - \int g(x)f(x)\,\mathrm{d}x + \int g(x)g(x)\,\mathrm{d}x \\
&\qquad = \mathbb{E}f(X) - \mathbb{E}f(Y) - \mathbb{E}g(X) + \mathbb{E}g(Y) \\
&\qquad = \mathbb{E}(f-g)(X) - \mathbb{E}(f-g)(Y) \\
&\qquad \le \|f-g\|_{Lip} \cdot \mathsf{d}_r\left(P^f, P^g\right),
\end{aligned}$$

by the Kantorovich–Rubinstein theorem. ∎

COROLLARY 1. *Let P be a measure on $\mathbb{R}^m$ with density f. Then the kernel density estimator $\hat{f}_n$ corresponding to $\hat{P}_n * k_h$ satisfies*

$$\int \big(f(x)-\hat{f}_n(x)\big)^2 \,\mathrm{d}x \le \left\|f-\hat{f}_n\right\|_{Lip} \cdot \mathsf{d}_r\left(\hat{P}_n * k_h, P\right) \tag{31}$$

for every $r \ge 1$.

Bounds for the Wasserstein distance. The reverse inequalities, which provide bounds for the Wasserstein distance in terms of the Parzen–Rosenblatt density estimator, are more delicate. To provide results that we can build on for the nested distance, we need to restrict the considerations to spaces with a compact support in $\mathbb{R}^m$.[4]

PROPOSITION 4. *Let K be a compact set and $\beta \ge 1$. Then there is a constant C depending on K, β, and r only such that, for all measures P^{f_1} and P^{f_2} with arbitrary density f_1 and f_2, both supported by K, the inequalities*

$$\mathsf{d}_r\left(P^{f_2}, P^{f_1}\right)^r \le C_{\beta,K} \cdot \|f_2 - f_1\|_\beta$$

hold true. In particular it holds that

$$\mathsf{d}_2\left(P^{f_2}, P^{f_1}\right)^2 \le C \cdot \|f_2 - f_1\|_2$$

and

$$\mathsf{d}_r\left(P^{f_2}, P^{f_1}\right)^r \le C \cdot \|f_2 - f_1\|_\infty .$$

Proof. Without loss of generality we may assume that $f_1 \ne f_2$. Set $g := \min\{f_1, f_2\}$ and $\mu := \int g\,\mathrm{d}\lambda$. As f_1 and f_2 are densities, it is evident that $0 \le \mu < 1$. Define the measures $P_1(A) := \frac{1}{1-\mu}\int_A f_1 - g\,\mathrm{d}\lambda$ and $P_2(B) := \frac{1}{1-\mu}\int_B f_2 - g\,\mathrm{d}\lambda$, and observe that

[4] In fact, for every C there exist f_1 and f_2 with unbounded support such that $\mathsf{d}_r(P^{f_1}, P^{f_2}) > C\|f_1 - f_2\|$.

P_1 and P_2 are probability measures, because $f_1 \geq g$ and $\int f_1 - g \, \mathrm{d}\lambda = 1 - \mu$ (and analogously for f_2). The bivariate probability measure

$$\pi(A \times B) := \int_{A \cap B} g \, \mathrm{d}\lambda + (1 - \mu) \cdot P_1(A) \cdot P_2(B)$$

has the marginal densities f_1 and f_2. Indeed, $\pi(A \times \Omega) = \int_A g \, \mathrm{d}\lambda + \int_A f_1 - g \, \mathrm{d}\lambda = \int_A f_1 \, \mathrm{d}\lambda$, which is the first marginal constraint of the Wasserstein distance in Definition 1. The second follows by analogous reasoning.

Note next that $\mathsf{d}(x, y)^r = \|x - y\|^r \leq (\|x\| + \|y\|)^r \leq 2^{r-1} (\|x\|^r + \|y\|^r)$, so

$$\begin{aligned}
\int \mathsf{d}^r \mathrm{d}\pi &= \int \mathsf{d}(x, x)^r g(x) \, \mathrm{d}x + \frac{1 - \mu}{(1 - \mu)^2} \iint \mathsf{d}(x, y)^r \left(f_1 - g\right)(x) \cdot \left(f_2 - g\right)(y) \, \mathrm{d}x \, \mathrm{d}y \\
&\leq 0 + \frac{1}{1 - \mu} 2^{r-1} \iint \left(\|x\|^r + \|y\|^r\right) \left(f_1 - g\right)(x) \cdot \left(f_2 - g\right)(y) \, \mathrm{d}x \, \mathrm{d}y \\
&= \frac{2^{r-1}}{1 - \mu} \int \|x\|^r \left(f_1 - g\right)(x) \, \mathrm{d}x \cdot \int \left(f_2 - g\right)(y) \, \mathrm{d}y \\
&\quad + \frac{2^{r-1}}{1 - \mu} \int \left(f_1 - g\right)(x) \mathrm{d}x \cdot \int \|y\|^r \left(f_2 - g\right)(y) \, \mathrm{d}y \\
&= 2^{r-1} \int \|x\|^r \left(f_1 - g\right)(x) \mathrm{d}x + 2^{r-1} \int \|y\|^r \left(f_2 - g\right)(y) \, \mathrm{d}y.
\end{aligned}$$

Note next that $0 \leq f_1 - g \leq |f_2 - f_1|$ such that

$$\iint \mathsf{d}^r \mathrm{d}\pi \leq 2^r \int \|x\|^r \cdot |f_2(x) - f_1(x)| \, \mathrm{d}x.$$

By Hölder's inequality on a compact domain K thus

$$\iint \mathsf{d}^r \mathrm{d}\pi \leq 2^r \left(\int_K \|x\|^{r\beta'} \, \mathrm{d}x \right)^{1/\beta'} \cdot \left(\int |f_2(x) - f_1(x)|^\beta \, \mathrm{d}x \right)^{1/\beta} = C \cdot \|f_2 - f_1\|_\beta ,$$

where C depends on r, β, and K and where $1/\beta + 1/\beta' = 1$. The assertion follows. ∎

The following corollary ensures convergence in probability of the convoluted measures; it derives from convergence of the mean integrated squared error for density estimators.

COROLLARY 2. Let P^f be a probability distribution on a compact K, induced by a density f. Then

$$\mathsf{d}_2 \left(P^{\hat{f}_n}, P^f \right) \xrightarrow{p} 0 \quad \text{(in probability)},$$

where $\hat{f}_n$ is the kernel density estimator (13), provided that the mean integrated squared error MISE tends to 0.

Proof. It follows from Proposition 4 and Markov's inequality that

$$\begin{aligned}
P\left(\mathsf{d}_r \left(P^{\hat{f}_n}, P^f \right) > \varepsilon \right) &\leq P\left(C \cdot \left\| \hat{f}_n - f \right\|_2^{1/r} > \varepsilon \right) \\
&\leq P\left(\left\| \hat{f}_n - f \right\|_2^2 > \frac{\varepsilon^{2r}}{C^{2r}} \right) \leq \frac{C^{2r}}{\varepsilon^{2r}} \mathbb{E} \left\| \hat{f}_n - f \right\|_2^2,
\end{aligned}$$

which is the mean integrated squared error. Convergence in probability follows, as the MISE tends to 0 by assumption whenever $n \to \infty$. ∎

5. Convergence of the nested distance in probability. We have seen in Proposition 1 that $\mathsf{dl}\big(\hat{\mathbb{P}}_n, \mathbb{P}\big) \geq c > 0$, so that the empirical measure $\hat{P}_n$ cannot be considered as a useful approximation of P, when the filtration is relevant. In what follows we prove, however, that $\hat{P}_n * k_h$ can be employed as an escape. It holds that $\mathsf{dl}\left(\mathbb{P}_n^{k_h}, \mathbb{P}\right) \to 0$ in probability (cf. Theorem 4 below), where $\mathbb{P}_n^{k_h}$ is based on smoothed measures $\hat{P}_n * k_{h_n}$ instead of the empirical measure $\hat{P}_n$. We shall make use of the following auxiliary result for the rather technical proof.

THEOREM 3. *Suppose that the bandwidth sequence h_n satisfies the conditions of Theorem 2, and that the density f is bounded by $0 < u < f(\cdot) < U < \infty$ (cf. Remark 10) on its support. Suppose further that the support $K = \{f > 0\}$ is convex and compact, and that f is continuous in the interior of K. Then, for a regular kernel k,*

$$P\left(\sup_y \mathsf{d}\left(P(\cdot|y),\, \hat{P}_n * k_{h_n}(\cdot|y)\right) > \varepsilon\right) \to 0 \tag{32}$$

for every $\varepsilon > 0$.

Proof. The conditional measures $P(\cdot|y)$ and $\hat{P}_n * k_{h_n}(\cdot|y)$ have densities $f(\cdot|y)$ and $\hat{f}_n(\cdot|y)$. It follows from Proposition 4, Markov's inequality, and the triangle inequality that

$$\begin{aligned}
P\left(\sup_y \mathsf{d}\left(P(\cdot|y,\, \hat{P}_n * k_{h_n}(\cdot|y)\right) > \varepsilon\right) &\leq P\left(\sup_y \left\|\hat{f}_n(\cdot|y) - f(\cdot|y)\right\|_2 > \frac{\varepsilon}{C}\right) \\
&\leq \frac{C}{\varepsilon}\mathbb{E}\sup_y \left\|\hat{f}_n(\cdot|y) - f(\cdot|y)\right\|_2 \qquad (33) \\
&\leq \frac{C}{\varepsilon}\sup_y \left\|\frac{\mathbb{E}\hat{f}_n(\cdot,y)}{\mathbb{E}\hat{f}_n(y)} - f(\cdot|y)\right\|_2 + \frac{C}{\varepsilon}\mathbb{E}\sup_y \left\|\hat{f}_n(\cdot|y) - \frac{\mathbb{E}\hat{f}_n(\cdot,y)}{\mathbb{E}\hat{f}_n(y)}\right\|_2 \\
&\leq \frac{C}{\varepsilon}\sup_y \left\|\frac{\mathbb{E}\hat{f}_n(\cdot,y)}{\mathbb{E}\hat{f}_n(y)} - f(\cdot|y)\right\|_2 + \frac{C}{\varepsilon}\lambda(K)^{1/2}\,\mathbb{E}\sup_y \left\|\hat{f}_n(\cdot|\xi) - \frac{\mathbb{E}\hat{f}_n(\cdot,y)}{\mathbb{E}\hat{f}_n(y)}\right\|_\infty . \qquad (34)
\end{aligned}$$

The first summand in (34) is deterministic and converges because the density f is almost everywhere smooth.

Note next that $\hat{f}_n(y) > \frac{1}{2}\mathbb{E}\hat{f}_n(y) \geq \frac{c}{2} > 0$ on the support K (in the interior we can choose $c = u$, as $\hat{f}_n(\cdot) \geq u$) almost everywhere for n large enough. It follows that

$$\begin{aligned}
\left|\hat{f}_n(\cdot|y) - \frac{\mathbb{E}\hat{f}_n(\cdot,y)}{\mathbb{E}\hat{f}_n(y)}\right| &= \left|\frac{\hat{f}_n(\cdot,y)}{\hat{f}_n(y)} - \frac{\mathbb{E}\hat{f}_n(\cdot,y)}{\mathbb{E}\hat{f}_n(y)}\right| \\
&= \left|\frac{\hat{f}_n(\cdot,y)\mathbb{E}\hat{f}_n(y) - \hat{f}_n(y)\mathbb{E}f_n(\cdot,y)}{\hat{f}_n(y)\mathbb{E}\hat{f}_n(y)}\right| \\
&\leq \frac{4}{c^2}\left|\hat{f}_n(\cdot,y)\mathbb{E}\hat{f}_n(y) - \hat{f}_n(y)\mathbb{E}\hat{f}_n(\cdot,y)\right| \to 0,
\end{aligned}$$

where the latter converges to 0 because of Theorem 2. Convergence is uniform in y, because the constant $c > 0$ can be chosen uniformly on y, and Theorem 2 (Remark 10) ensures uniform convergence. It follows that (33) tends to 0, and the assertion follows. ∎

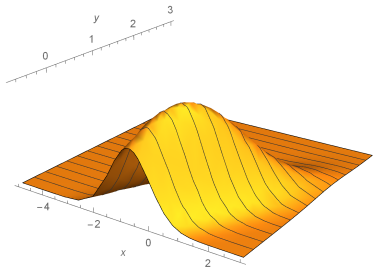


FIG. 2. *Conditional Gaussian distribution.*

Below we formulate the the main theorem. Coming back to the initial setup, we consider a stochastic process $(\xi_0, \dots, \xi_T)$ and introduce the notation $\xi_{0:t} := (\xi_0, \dots, \xi_t)$ for a substring of $(\xi_0, \dots, \xi_T)$.

The empirical observations are as in (7).

THEOREM 4 (the nested distance of the convoluted empirical measure converges). *Suppose that*

(i) *the conditions of Theorem* 3 *hold, and*

(ii) *the measure P is conditionally Lipschitz; i.e.,* $\mathsf{d}\big(P(\cdot|\xi_{0:t}), P(\cdot|\tilde{\xi}_{0:t})\big) \le \gamma_t \cdot \big\|\xi_{0:t} - \tilde{\xi}_{0:t}\big\|$.

Then the nested distance between the filtered spaces $\mathbb{P}_n^k = \big(\Xi, (\Sigma_t)_{t=0,\dots,T}, \hat{P}_n * k_{h_n}\big)$, equipped with the convolution measure $\hat{P}_n * k_{h_n}$, and the true model $\mathbb{P} = (\Xi, (\Sigma_t)_{t=0,\dots,T}, P)$ converges to zero in probability; i.e.,

$$P\left(\mathsf{dl}\big(\mathbb{P}, \mathbb{P}_n^k\big) > \varepsilon\right) \to 0$$

as $n \to \infty$.

Example 1. A simple example of a probability satisfying condition (ii) is the Gaussian distribution. Consider a multivariate Gaussian random variable $\binom{Y}{X} \sim \mathcal{N}\left(\binom{\mu_Y}{\mu_X}, \left(\begin{smallmatrix}\Sigma_{YY} & \Sigma_{YX}\\ \Sigma_{YX} & \Sigma_{XX}\end{smallmatrix}\right)\right)$ with regular covariance matrix. Then the distribution of X conditional on $\{Y = y\}$ is a Gaussian variable again with distribution

$$X|\,Y = y \sim \mathcal{N}\left(\mu_X + \Sigma_{XY}\Sigma_{YY}^{-1}(y - \mu_X),\, \Sigma_{XX} - \Sigma_{XY}\Sigma_{YY}^{-1}\Sigma_{YX}^{\top}\right),$$

as outlined in Liptser and Shiryaev [20, Theorem 13.1] (cf. Figure 2). Importantly, the conditional covariance matrix does not depend on y. Hence the Wasserstein distance of the corresponding probability measure can be obtained by shifting. The Wasserstein distance thus satisfies condition (ii) of Theorem 4 with

$$\mathsf{d}\big(P(\cdot|y), P(\cdot|\tilde{y})\big) \le \big\|\Sigma_{XY}\Sigma_{YY}^{-1}\big\| \cdot \|y - \tilde{y}\|\,.$$

Proof of Theorem 4 *(cf. also Chapter* 4.2 *in* [25]*).* Without loss of generality we may assume that the norm on the product space $\mathbb{R}^m$ is $d(x, \tilde{x}) = \sum_{t=1}^{T} \|x_t - \tilde{x}_t\|$ in (ii) and further that $r = 1$. We shall proceed by backward induction from $t = T$ down to $t = 0$.

Choose an optimal collection of transport plans $\pi^{T-1}(\cdot, \cdot|\xi_{0:T-1}, \tilde{\xi}_{0,T-1})$ for the conditional distributions $P(\cdot|\xi_{0:T-1})$ and $\hat{P}_n * k_{h_n}(\cdot|\tilde{\xi}_{0:T-1})$ at stage T, and an optimal transportation plan $\pi_{T-1}(\cdot, \cdot)$ for the unconditional distributions of $P|_{\xi_{0:T-1}}$ and $(\hat{P}_n *$

$k_{h_n})|_{\tilde\xi_{0:T-1}}$ of $\xi_{0:T-1}$ (resp., $\tilde\xi_{0:T-1}$) up to stage $T-1$. Glue them together to obtain a transportation plan π for $\xi_{0:T}$ (resp., $\tilde\xi_{0:T}$). We get

$$\begin{aligned}
\mathsf{dl}(P,\, \hat P_n * k_{h_n}) &\le \iint \sum_{t=1}^{T} \left\|\xi_t - \tilde\xi_t\right\| \pi(\mathrm{d}\xi, \mathrm{d}\tilde\xi)\\
&= \iint \left(\sum_{t=1}^{T-1} \left\|\xi_t - \tilde\xi_t\right\| + \left\|\xi_T - \tilde\xi_T\right\| \right) \pi^{T-1}\big(\mathrm{d}\xi_T, \mathrm{d}\tilde\xi_T \big|\, \xi_{0:T-1}, \tilde\xi_{0:T-1}\big) \pi_{T-1}(\mathrm{d}\xi_{0:T-1}, \mathrm{d}\tilde\xi_{0:T-1})\\
&= \iint \left(\sum_{t=1}^{T-1} \left\|\xi_t - \tilde\xi_t\right\| + \iint \left\|\xi_T - \tilde\xi_T\right\| \pi(\mathrm{d}\xi_T, \mathrm{d}\tilde\xi_T \big|\, \xi_{0:T-1}, \tilde\xi_{0:T-1}) \right) \pi_{T-1}(\mathrm{d}\xi_{0:T-1}, \mathrm{d}\tilde\xi_{0:T-1})\\
&= \iint \left(\sum_{t=1}^{T-1} \left\|\xi_t - \tilde\xi_t\right\| + \mathsf{d}\big(P(\cdot|\xi_{0:T-1}),\, \hat P_n * k_{h_n}(\cdot|\tilde\xi_{0:T-1})\big) \right) \pi_{T-1}(\mathrm{d}\xi_{0:T-1}, \mathrm{d}\tilde\xi_{0:T-1}).
\end{aligned}$$

By the triangle inequality for the Wasserstein distance and assumption (ii) on conditional Lipschitz continuity,

$$\begin{aligned}
&\mathsf{d}\big(P(\cdot|\xi_{0:T-1}), \hat P_n * k_{h_n}(\cdot|\tilde\xi_{0:T-1})\big)\\
&\qquad \le \mathsf{d}\big(P(\cdot|\xi_{0:T-1}),\, P(\cdot|\tilde\xi_{0:T-1})\big) + \mathsf{d}\big(P(\cdot|\tilde\xi_{0:T-1}),\, P * k_{h_n}(\cdot|\tilde\xi_{0:T-1})\big)\\
&\qquad \le \gamma_T \cdot \mathsf{d}\left(\xi_{0:T-1}, \tilde\xi_{0:T-1}\right) + \mathsf{d}\left(P(\cdot|\tilde\xi_{0:T-1}),\, P * k_{h_n}(\cdot|\tilde\xi_{0:T-1})\right).
\end{aligned}$$

By assumption one may conclude from (32) that one can choose n_t big enough such that $\mathsf{d}\big(P(\cdot|\tilde\xi),\, \hat P_{n_t} * k_{h_{n_t}}(\cdot|\tilde\xi)\big) < \varepsilon$ on a set of probability at least $1-\varepsilon$ (here, the probability is in $P^{\mathbb{N}}$). On this set,

$$\begin{aligned}
\mathsf{dl}\big(\mathbb{P},\, \mathbb{P}_n^k\big) &\le \iint \left(\sum_{t=1}^{T-1} \left\|\xi_t - \tilde\xi_t\right\| + \varepsilon + \gamma_T \cdot \mathsf{d}\left(\xi_{0:T-1}, \xi_{0:T-1}\right) \right) \pi(\mathrm{d}\xi_{0:T-1}, \mathrm{d}\tilde\xi_{0:T-1})\\
&= \iint \left(\sum_{t=1}^{T-1} \left\|\xi_t - \tilde\xi_t\right\| + \varepsilon + \gamma_T \cdot \left\|\xi_{0:T-1} - \tilde\xi_{0:T-1}\right\| \right) \pi(\mathrm{d}\xi_{0:T-1}, \mathrm{d}\tilde\xi_{0:T-1})\\
&= \varepsilon + (1+\gamma_T) \iint \left(\sum_{t=1}^{T-1} \left\|\xi_t - \tilde\xi_t\right\| \right) \pi(\mathrm{d}\xi_{0:T-1}, \mathrm{d}\tilde\xi_{0:T-1}).
\end{aligned}$$

By repeating the same arguments successively on each stage (i.e., T times) and collecting terms, it follows that

$$\mathsf{dl}\big(\mathbb{P},\, \mathbb{P}_n^k\big) \le \varepsilon + \varepsilon\,(1+\gamma_T) + \varepsilon\,(1+\gamma_T)\,(1+\gamma_{T-1}) + \cdots. \tag{35}$$

The probability of the set where (35) holds true is not less than $1 - \varepsilon \cdot T$ whenever $n \ge \max\{n_1, n_2, \ldots, n_T\}$. Hence

$$P\left(\mathsf{dl}\big(\mathbb{P},\, \mathbb{P}_n^k\big) > C \cdot \varepsilon\right) < \varepsilon \cdot T, \tag{36}$$

where $C := 1 + (1+\gamma_T) + (1+\gamma_T)(1+\gamma_{T-1}) + \cdots < \infty$ is a constant, depending solely on the conditional Lipschitz constants. Convergence in probability of the nested distance,

$$\mathsf{dl}\big(\mathbb{P},\, \mathbb{P}_n^k\big) \xrightarrow[n\to\infty]{p} 0,$$

is a restatement of (36). □

6. Estimating scenario trees based on observed trajectories. The construction of the discrete approximating tree proceeds in two steps:

(i) In step one, the model of densities and conditional densities $\mathbb{P}_n^k$ is obtained by the density and conditional density estimates.

(ii) In step two, this density model is used to construct the approximating tree.

While step (i) has its justification in Theorem 4, step (ii) is similar to the construction of a tree from simulated conditional distributions.

The tree generator algorithm (Algorithm 1) first estimates the probability density $\hat{f}_n(x_1)$ at the first stage $t = 1$ and discretizes it by the discrete measure $\sum_{i=1}^{b_1} p_i \delta_{\tilde{\xi}_{1,i}}$ sitting on b_1 points. This can be accomplished based on optimal quantizers (cf. Graf and Luschgy [9]) or by stochastic approximation algorithms outlined in [26]. Recursively, given that the tree is already established for t stages, each path $(\tilde{\xi}_1, \ldots, \tilde{\xi}_t)$ from the tree already constructed is considered again. The conditional distribution is estimated by sampling from the conditional density,

$$\hat{f}_n(x_{t+1} | \tilde{\xi}_0, \ldots, \tilde{\xi}_t).$$

This sampled distribution is again approximated by a discrete probability measure sitting on b_{t+1} points.

Remark 12 (the composition method). The quantization algorithm needs a larger number (e.g., 500) of independent random samples from the estimated conditional density. Suppose that the density is estimated by

$$\hat{f}_n(x) = \sum_{i=1}^{n} w_i k\left(\frac{x - \xi_i}{h}\right).$$

A random deviate X from the density $\hat{f}_n$ is generated as follows: first, a random index i^* is generated according to the weight distribution $(w_1, \ldots, w_n)$. Second, X is sampled by

$$X = \xi_{i^*} + h \cdot Z,$$

where Z is a random deviate from density k. The bandwidth h may be determined by cross-validation to prevent over- and undersmoothing (cf. Hall [11]).

In this paper, we do not repeat in detail the construction mechanism for this step, but refer to the book by Pflug and Pichler [25]. Algorithm 1 summarizes this procedure.

Remark 13. Algorithm 1 is formulated for a fixed branching structure. However, the algorithm may also adapt the branching structure in such a way that a prespecified distance between the discrete approximation and the corresponding conditional density at each node is not exceeded. If the number of discretization points is not large enough to meet the distance goal, this branching number is increased, and the optimal discretization algorithm is repeated until this goal is reached. This is called *dynamic tree construction*. For further details about this construction, see Pflug and Pichler [26]. By making the threshold of this distance go to zero, the constructed trees will converge in nested distance to the density model, and if in turn this density model converges to the true model, we have full convergence of the tree model.

Examples. We demonstrate the behavior of Algorithm 1 by the following three examples.

Algorithm 1 Generation of a scenario tree with fixed bushiness from a sample of paths.

Parameters. Let T be the desired height of the tree, and let $(b_1, \dots, b_T)$ be the given bushiness parameters per stage.

- **Determining the root.** The value of the process at the root is ξ_0. Its stage is 0. Set the root as the current open node.
- **Successor generation.** Enumerate the tree stagewise from the root to the leaves.
  (i) Let n_0 be the node to be considered next, and let $t < T$ be its stage. Let $\tilde{\xi}_0, \tilde{\xi}_1, \dots, \tilde{\xi}_t$ be the already fixed values at node n_0 and all its predecessors. Find an approximation of the form $\sum_{i=1}^{b_t} p_i \delta_{x^{(i)}}$, which is close in the Wasserstein distance to the distribution with density

$$\hat{f}_n(x_{t+1} | \tilde{\xi}_0, \dots, \tilde{\xi}_t). \tag{37}$$

  (ii) Store the b_t successor nodes of n_0, say with node numbers $(n_1, \dots, n_{b_t})$, and assign to them the values $\xi(n_1) = x^{(1)}, \dots, \xi(n_{b_t}) = x^{(n_{b_t})}$ as well as their conditional probabilities $q(n_i) = p_i$ in the new tree subtree.
- **Stopping criterion.** If all nodes at stage $T - 1$ have been considered as parent nodes, the generation of the tree is finished.

Example 2. Figure 3a displays 1 000 sample paths from a Gaussian walk in 12 stages. A binary tree with 4 095 nodes was extracted (cf. Figure 3b) by employing Algorithm 1. Note that the extracted tree has $2^{11} = 2\,048$ leaves, which is *more* than, even more than twice the size of, the original sample ($n = 1\,000$). Nevertheless, the approximating tree is apparently a useful approximation of the Gaussian process.

Figures 3c and 3d display the results of Algorithm 1 for the (non-Markovian) running maximum process derived from the samples of a Gaussian walk.

Example 3 (consistency). This example considers a tree as a starting process. Figure 4 depicts 10 000 samples from a tree process with 1 237 nodes. Algorithm 1 recovers the initial tree from the samples. Notice that a tree process does not have a density. Nevertheless, the algorithm is still able to recover the initial tree with reduced branches. Notice further that, by this procedure, one may reduce a larger tree to one of smaller size. Tree reduction algorithms described in the literature only skip or merge subtrees, but with our method, a new but similar smaller tree is constructed.

Figure 4 displays the result of the algorithm for a binary tree.

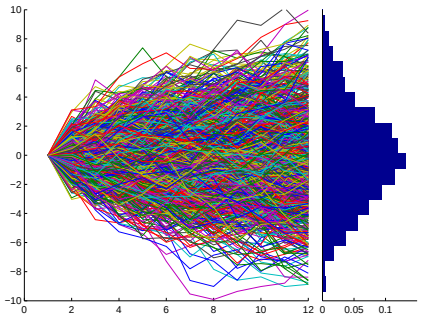
Markovian processes. The transition of a Markovian process can be described based on the current state only; the entire history is not necessary. Expressed in terms of the sigma algebra (cf. (5)),

$$\Sigma_t = \sigma(\xi_t).$$

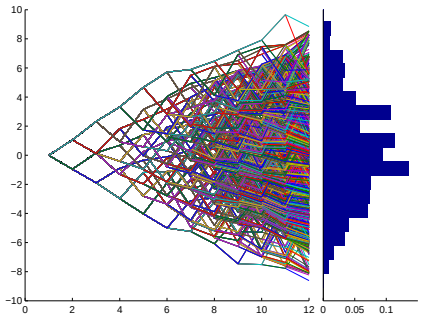
With this assumption the tree generation algorithm (Algorithm 1) simplifies significantly, as the conditional density depends on the previous state solely; i.e., the density (37) can be replaced by

$$\hat{f}_n(x_{t+1} | \xi_t).$$

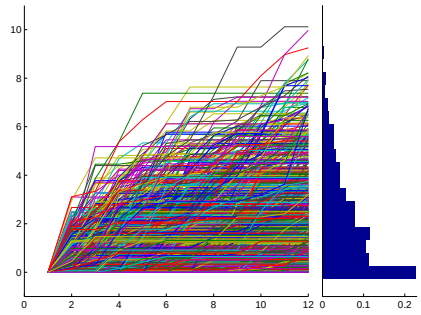
As as consequence, the estimator (27) for estimating the conditional density is simplified significantly, as further dimensions do not have to be included.



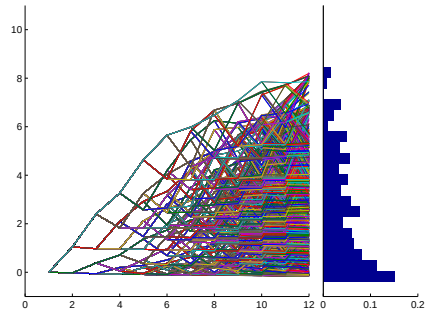
(a) 1 000 sample paths from a (modified) Gaussian random walk.



(b) Binary tree of height 12 with 4 095 nodes, approximating the random walk from Figure 3a.

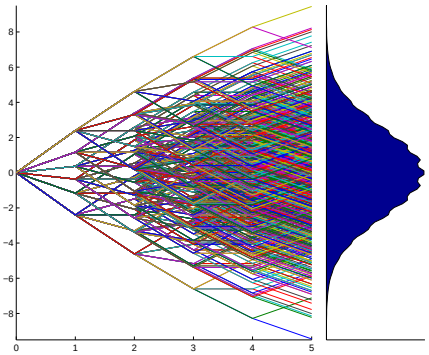


(c) The running maximum process from Figure 3a.

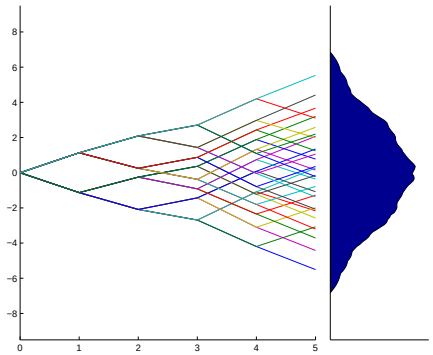


(d) Binary tree, extracted from the running maximum process in Figure 3c.

FIG. 3. *Sample paths (left) and extracted trees (right) of a Markovian (above) and non-Markovian (below) process based on Algorithm* 1.



(a) 10 000 samples, taken from a tree.



(b) Binary tree, constructed from the samples in Figure 4a.

FIG. 4. *Reconstruction of a tree processes.*

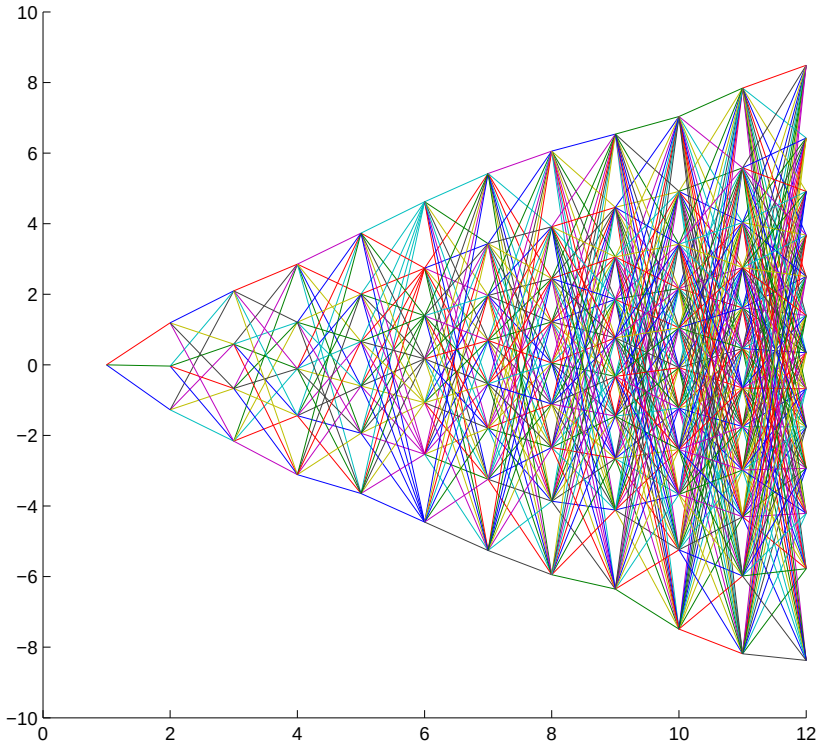


FIG. 5. *A lattice constructed from* 10 000 *empirical observations of sample paths.*

Further computational accelerations are obtained by considering identical children for all nodes of the tree, which are at the same stage. The resulting tree gets the shape of a lattice, as the following example demonstrates.

Example 4. The lattice in Figure 5 is generated from 10 000 sample paths of a Gaussian walk. The lattice was chosen to have $t+1$ states at stage t.

Choice of the parameters. For kernel density estimation, the Epanechnikov kernel $k(u) = \max\left\{\frac{3}{4}\left(1-u^2\right), 0\right\}$ is often proposed, as its shape is most efficient (for a specific criterion; cf. Tsybakov [36] for details). In the present situation the Epanechnikov kernel is not the favorable choice, as a division by zero has to be avoided in (27) (this might be an issue for a small sample size n). This can be avoided by employing, e.g., the logistic kernel

$$k(u) = \frac{1}{e^u + 2 + e^{-u}} = \frac{1}{4}\frac{1}{\left(\cosh\frac{u}{2}\right)^2},$$

which is strictly positive for all $u \in \mathbb{R}$.

As for the optimal bandwidth, we recall from Caillerie et al. [6] that

$$\mathbb{E}\,\mathsf{d}_2\left(P, \hat{P}_n\right)^2 \le \frac{C}{n^{2/(m+4)}}, \tag{38}$$

where $\hat{P}_n$ is the measure with density $\frac{1}{n\,h_n^m}\sum_{i=1}^n k\big(\frac{\cdot-\xi_i}{h_n}\big)$. The rate (38) is the same rate as that obtained by Silverman's rule of thumb (cf. Silverman [34]) or Scott's rule, which suggests using

$$h_n \simeq \mathsf{std}(\xi)\cdot\left(\frac{4}{n\,(m+2)}\right)^{1/(m+4)} \simeq \mathsf{std}(\xi)\cdot n^{-1/(m+4)}.$$

The estimate (38) does not require that the measure P have a density. Slight improvements of the rate of convergence are known in the case that a density is available—cf. Rachev [29] for a discussion.

7. Summary. This paper discusses the nested distance, which is a distance for stochastic processes. The distance is adapted for stochastic optimization problems, as it exactly describes the information structure and the continuity properties of optimization problems of this type.

Empirical observations, which are available, are sample paths. We demonstrate that the empirical measure, which is associated with these sample paths observed, does *not* converge in nested distance.

If the underlying process has a density, then—by employing a convolution, as is known from kernel density estimation—it is possible to obtain smoothed processes also with density, which converge in nested distance. Importantly, the underlying process is built solely from sample paths. The convergence result is stated by employing convergence in probability.

Trees constitute representative points (quantizers) of processes. Trees are adequate finite-space data structures to model processes which are arbitrarily close in nested distance to a process with density. As an application, we illustrate the algorithm for constructing representative trees from samples. The methods employed are nonparametric; i.e., we do not make parametric assumptions on the underlying process.

Appendix A. Proofs.

Proof of Lemma 1. We shall prove that

$$\mathsf{d}_r(\tilde{P}*k_h, P*k_h) \le \mathsf{d}_r(\tilde{P}, P) \tag{39}$$

and

$$\mathsf{d}_r(P*k_h, P) \le \kappa_r^{1/r}\cdot h, \tag{40}$$

from which the assertion follows by employing the triangle inequality for the distance d_r (cf. Ambrosio, Gigli, and Savaré [1]).

Let π be the optimal transportation measure between P and $\tilde{P}$. Define the measure

$$\tilde{\pi}(A\times B) := \int \mathbb{1}_{A\times B}(x+x', y+x')k(dx')\pi(dx,dy).$$

Note that $\tilde{\pi}$ has the marginal distribution

$$\tilde{\pi}(A\times\Omega) = \int \mathbb{1}_A(x+x')k(dx')\pi(dx,dy) = \int \mathbb{1}_A(x+x')k(dx')P(dx) = (P*k)(A)$$

by (10); the second marginal $\tilde{\pi}(\Omega\times B) = (\tilde{P}*k)(B)$ equality holds by symmetric

reasoning. It follows that

$$\int \mathsf{d}(x,y)^r \tilde{\pi}(dx,dy) = \int \mathsf{d}(x+x',y+x')^r k(dx')\pi(dx,dy)$$
$$= \int \mathsf{d}(x,y)^r k(dx')\pi(dx,dy) = \int \mathsf{d}(x,y)^r \pi(dx,dy),$$

as the distance is translation invariant. The inequality (39) follows by taking the infimum over all appropriate $\tilde{\pi}$ on the left-hand side. ∎

Proof of Proposition 2. As for (40), define the measure

$$\tilde{\pi}(A\times B) := \int \mathbb{1}_{A\times B}(x,x+x')\, k(dx')\, P(dx)$$

with marginals $\tilde{\pi}(A\times\Omega) = P(A)$ and $\tilde{\pi}(\Omega\times B) = P * k_h(B)$. It follows that

$$\mathsf{d}(P*K_h,P)^r \le \int d(x,x+x')^r k_h(dx')\, P(dx) = \int \|(x-h\,x')-x\|^r\, k(dx')\, P(dx)$$
$$= \int \|x'\|^r h^r k(dx')\, P(dx) = \kappa_r \cdot h^r,$$

which is the assertion.

Observe first that

$$|f*k_h(x) - f(x)| = \left| \int_{\mathbb{R}^m} k_h(x-y)\big(f(y)-f(x)\big)\,\mathrm{d}y \right|$$
$$\le \int_{\mathbb{R}^m} k_h(x-y)\cdot |f(y)-f(x)|\,\mathrm{d}y$$
$$\le \int_{\{\|x-y\|\le h\}} k_h(x-y)\,|f(y)-f(x)|\,\mathrm{d}y \le \delta_f(h).$$

Moreover, as k is Lipschitz continuous, it follows that $k_h(\cdot) = \frac{1}{h^m} k\big(\frac{\cdot}{h}\big)$ has Lipschitz constant $\|k_h\|_{Lip} = \frac{\|k\|_{Lip}}{h^{m+1}}$. Hence

$$\left|\hat{f}_n(x) - f*k_h(x)\right| = \int k_h(x-y)\,\big(\hat{P}_n(\mathrm{d}y) - P(\mathrm{d}y)\big) \le \|k_h\|_{Lip}\, \mathsf{d}_1(\hat{P}_n,P)$$
$$= \frac{\|k\|_{Lip}}{h^{m+1}} \mathsf{d}_r(\hat{P}_n,P),$$

and the assertion is immediate by the triangle inequality. ∎

REFERENCES

[1] L. Ambrosio, N. Gigli, and G. Savaré, *Gradient Flows in Metric Spaces and in the Space of Probability Measures*, 2nd ed., Birkhäuser-Verlag, Basel, Switzerland, 2005, doi:10.1007/978-3-7643-8722-8

[2] V. Bally, G. Pagès, and J. Printems, *A quantization tree method for pricing and hedging multidimensional American options*, Math. Finance, 15 (2005), pp. 119–168.

[3] M. Beiglböck, C. Léonard, and W. Schachermayer, *A general duality theorem for the Monge-Kantorovich transport problem*, Stud. Math., 209 (2012), pp. 151–167, doi:10.4064/sm209-2-4.

[4] F. Bolley, *Separability and completeness for the Wasserstein distance*, in Séminaire de Probabilités XLI, C. Donati-Martin, M. Émery, A. Rouault, and C. Stricker, eds., Lecture Notes in Math. 1934, Springer, Berlin, Heidelberg, 2008, pp. 371–377.

[5] F. Bolley, A. Guillin, and C. Villani, *Quantitative concentration inequalities for empirical measures on non-compact spaces*, Probab. Theory Rel. Fields, 137 (2007), pp. 541–593, doi:10.1007/s00440-006-0004-7.
[6] C. Caillerie, F. Chazal, J. Dedecker, and B. Michel, *Deconvolution for the Wasserstein metric and geometric inference*, Electronic J. Statist., 5 (2011), pp. 1394–1423, doi:10.1214/11-EJS646.
[7] U. Einmahl and D. M. Mason, *Uniform in bandwidth consistency of kernel-type function estimators*, Ann. Statist., 33 (2005), pp. 1380–1403, doi:10.1214/009053605000000129.
[8] E. Giné and A. Guillou, *Rates of strong uniform consistency for multivariate kernel density estimators*, Ann. Inst. H. Poincaré Probab. Statist., 38 (2002), pp. 907–921, doi:10.1016/S0246-0203(02)01128-7.
[9] S. Graf and H. Luschgy, *Foundations of Quantization for Probability Distributions*, Lecture Notes in Math. 1730, Springer-Verlag, Berlin, 2000, doi:10.1007/BFb0103945.
[10] N. Gülpinar, B. Rustem, and R. Settergren, *Simulation and optimization approaches to scenario tree generation*, J. Econom. Dyn. Control, 28 (2004), pp. 1291–1315.
[11] P. Hall, *Cross-validation in density estimation*, Biometrika, 69 (1982), pp. 383–390.
[12] H. Heitsch and W. Römisch, *Scenario reduction algorithms in stochastic programming*, Comput. Optim. Appl. Stoch. Program., 24 (2003), pp. 187–206, doi:10.1023/A:1021805924152.
[13] H. Heitsch and W. Römisch, *Scenario tree reduction for multistage stochastic programs*, Comput. Management Sci., 2 (2009), pp. 117–133, doi:10.1007/s10287-008-0087-y.
[14] H. Heitsch and W. Römisch, *Scenario tree modeling for multistage stochastic programs*, Math. Program. Ser. A, 118 (2009), pp. 371–406.
[15] K. Høyland and S. W. Wallace, *Generating scenario trees for multistage decision problems*, Management Sci., 47 (2001), pp. 295–307, doi:10.1287/mnsc.47.2.295.9834.
[16] R. J. Hyndman, D. M. Bashtannyk, and G. K. Grunwald, *Estimating and visualizing conditional densities*, J. Comput. Graph. Statist., 5 (1996), pp. 315–336, doi:10.2307/1390887.
[17] O. Kallenberg, *Foundations of Modern Probability*, Springer, New York, 2002.
[18] J. Kim, *Event tree based sampling*, Comput. Oper. Res., 33 (2006), pp. 1184–1199, doi:10.1016/j.cor.2004.09.008.
[19] Q. Li and J. S. Racine, *Nonparametric Econometrics: Theory and Practice*, Princeton University Press, Princeton, NJ, 2006; http://books.google.com.au/books?id=Zsa7ofamTIUC.
[20] R. S. Liptser and A. N. Shiryaev, *Statistics of Random Processes*, Stoch. Model. Appl. Probab., Springer, New York, 2001; http://books.google.com/books?id=gKtK0CjxOaIC.
[21] G. Pagès and J. Printems, *Functional quantization for numerics with an application to option pricing*, Monte Carlo Methods Appl., 11 (2005), pp. 407–446.
[22] G. Ch. Pflug, *Scenario tree generation for multiperiod financial optimization by optimal discretization*, Mathematical Programming, 89 (2001), pp. 251–271, doi:10.1007/s101070000202.
[23] G. Ch. Pflug, *Version-independence and nested distributions in multistage stochastic optimization*, SIAM J. Optim., 20 (2009), pp. 1406–1420, doi:10.1137/080718401.
[24] G. Ch. Pflug and A. Pichler, *A distance for multistage stochastic optimization models*, SIAM J. Optim., 22 (2012), pp. 1–23, doi:10.1137/110825054.
[25] G. Ch. Pflug and A. Pichler, *Multistage Stochastic Optimization*, Springer Ser. Oper. Res. Financial Engrg., Springer, New York, 2014, doi:10.1007/978-3-319-08843-3; https://books.google.com/books?id=q_VWBQAAQBAJ.
[26] G. Ch. Pflug and A. Pichler, *Dynamic generation of scenario trees*, Comput. Optim. Appl., 62 (2015), pp. 641–668, doi:10.1007/s10589-015-9758-0.
[27] G. Ch. Pflug and A. Pichler, *Time-inconsistent multistage stochastic programs: Martingale bounds*, European J. Oper. Res., 249 (2015), pp. 155–163, doi:10.1016/j.ejor.2015.02.033.
[28] G. Ch. Pflug and A. Pichler, *Time-consistent decisions and temporal decomposition of coherent risk functionals*, Math. Oper. Res., 41 (2015), pp. 682–699, doi:10.1287/moor.2015.0747.
[29] S. T. Rachev, *Probability Metrics and the Stability of Stochastic Models*, John Wiley and Sons, West Sussex, England, 1991; http://books.google.com/books?id=5grvAAAAMAAJ.
[30] S. T. Rachev and L. Rüschendorf, *Mass Transportation Problems Volume* I: *Theory*, Probab. Appl. 25, Springer, New York, 1998, doi:10.1007/b98893.
[31] S. T. Rachev and L. Rüschendorf, *Mass Transportation Problems Volume* II: *Applications*, Probab. Appl. 26, Springer, New York, 1998, doi:10.1007/b98894.
[32] J. S. Racine, Q. Li, and X. Zhu, *Kernel estimation of multivariate conditional distributions*, Ann. Econom. Finance, 5 (2004), pp. 211–235.

[33] S. J. Sheather, *An improved data-based algorithm for choosing the window width when estimating the density at a point*, Comput. Statist. Data Anal., 4 (1986), pp. 61–65, doi:10.1016/0167-9473(86)90026-5.

[34] B. W. Silverman, *Density Estimation for Statistics and Data Analysis*, Chapman & Hall/CRC Press, London/Boca Raton, 1998.

[35] W. Stute, *The oscillation behavior of empirical processes: The multivariate case*, Ann. Probab., 12 (1984), pp. 361–379, doi:10.1214/aop/1176993295.

[36] A. B. Tsybakov, *Introduction to Nonparametric Estimation*, Springer, New York, 2008, doi:10.1007/b13794.

[37] V. S. Varadarajan, *Weak convergence of measures on separable metric spaces*, Sankhyā: Indian J. Statist., 19 (1958), pp. 15–22; http://www.jstor.org/stable/25048364.

[38] C. Villani, *Topics in Optimal Transportation*, Grad. Stud. Math. 58, American Mathematical Society, Providence, RI, 2003; http://books.google.com/books?id=GqRXYFxe0l0C.

[39] D. Wied and R. Weissbach, *Consistency of the kernel density estimator: A survey*, Statist. Papers, 53 (2012), pp. 1–21, doi:10.1007/s00362-010-0338-1.