Convergence of the Smoothed Empirical Process in Nested Distance

Georg Ch. Pflug* and Alois Pichler†

September 6, 2015

Abstract

The nested distance, also process distance, provides a quantitative measure of distance for stochastic processes. It is the crucial and determining distance for stochastic optimization problems.

In this paper we demonstrate first that the empirical measure, which is built from observed sample paths, does not converge in nested distance to its underlying distribution. We show that smoothing convolutions, which are appropriately adapted from classical density estimation using kernels, 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 transition probabilities at each time moment. Finally we construct processes with discrete sample space from observed empirical paths, which approximate well the original stochastic process as they converge in nested distance.

Keywords: Decision trees, stochastic optimization, optimal transportation
Classification: 90C15, 60B05, 62P05

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. Quite often, the approximation is done by considering the empirical measure based on past observations. Reducing in this way 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 the approximation of stochastic processes with 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 sub-path. Indeed, consider a sub-path which is possible but was not observed, from its origin up to some intermediate state. Then, with probability 1, none of the empirical

*University of Vienna. Department of Statistics and Operations Research.
International Institute for Applied Systems Analysis (IIASA), Laxenburg, Austria.
†Norwegian University of Science and Technology, NTNU. The author gratefully acknowledges support of the Research Council of Norway (grant 207690/E20).

observations coincides with this sub-path chosen and hence the empirical measure cannot reproduce the distribution conditional on this chosen path.

Pagès et al. (cf. [14] or [2]) elaborate optimal discrete approximations (often called quantizers) to treat specific problems 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 do not describe conditional transitions neither, as they lack a branching structure as well.

The branching structure corresponds to the information gain obtained in time, i.e., the pertaining *filtration*. Considering available information 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 [15]).

The following section reviews a distance for stochastic processes, called *nested distance* or *process distance* introduced in Pflug and Pichler [17]. This concept of a distance for stochastic processes 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. In contrast, there are correctly chosen tree models which are consistent in nested distance. To this end, we propose to build 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. Here we illustrate the inconsistency of the empirical measure in nested distance. We prove further that non-branching 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. Section 5 finally 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 also 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 5 below) we prove first that the empirical measure does not converge in nested distance to the initial process.

2.1 The nested distance

The nested distance is a distance for filtered, separable metric probability spaces (Ξ, d). It is an extension of the Wasserstein distance, a transportation distance for probability spaces on metric

(Polish) spaces.

Definition 1 (Nested distance, also process distance)**.** Let

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

be filtered probability spaces (a. k. a. *stochastic basis*). 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 π is a probability measure with conditional marginals P and $\tilde{P}$, i.e.,

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

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

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

Remark 2. If $T = 1$ and if the filtration just consists of 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

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

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 [21], as well as in Villani [28]. 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 and measure the evolution of the information of a stochastic process over time. It is the crucial and determining distance for stochastic optimization problems. The nested distance was introduced in Pflug [16] for nested distributions. Its dual formulation, as well as basic properties are elaborated in [17].

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 et al. [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) resp. $(\tilde{\Omega}, \tilde{P})$ 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 an analogous way as 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 $\tilde{P} \circ \tilde{\xi}^{-1}$ thus is a distance for the distributions of the stochastic processes ξ and $\tilde{\xi}$.

Convention for 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 consists of the sigma algebras

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

generated by process $\xi = (\xi_0, \ldots, \xi_T)$, 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 $\Sigma_0 = \{\emptyset, \Xi\}$ is the trivial sigma algebra, we thus omit the 0^{th}-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 as $(\Xi, (\Sigma_t), P)$, which contain the filtration as intregral part of it, while 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 [17, Theorem 11]) 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 5 (Continuity of stochastic optimization problems)**.** *Let* $\mathbb{P} := \left(\Xi, (\Sigma_t)_{t=0,\ldots T}, P\right)$ *and* $\tilde{\mathbb{P}} := \left(\Xi, (\tilde{\Sigma}_t)_{t=0,\ldots T}, \tilde{P}\right)$ *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*

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

for every $r \geq 1$.

The constraint $x \lhd \sigma(\xi)$ is shorthand for x_t *is measurable with respect to* $\Sigma_t = \sigma(\xi_1, \ldots, \xi_t)$ *for all* $t = 0, \ldots T$, where $x = (x_t)_{t=0}^T$ in (6) is the (stochastic) decision process. By the Doob–Dynkin lemma (cf. Kallenberg [11]), 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 $x_t = x'_t(\xi_0, \ldots, \xi_t)$ (i.e., $x_t(\cdot) = x'_t\big(\xi_0(\cdot), \ldots, \xi_t(\cdot)\big)$).

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

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

is

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

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 6. Discrete measures are — with respect to the Wasserstein distance — dense in the space of measures satisfying an adequate moment constraint (see Bolley [4], e.g., 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:

Proposition 7. *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 discrete measure $P_n := \sum_{i=1}^n w_i^{(n)}\delta_{\xi_i}$ do not converge in nested distance to $\mathbb{P} := (\Xi, (\Sigma_t)_{t=1,\dots,T}, P)$, provided that*

$$\xi_{i,t} \neq \xi_{j,t} \quad \text{for all } t \geq 1 \text{ and } i \neq j. \tag{9}$$

Remark 8 (The empirical measure does not converge). Note that Proposition 7 covers empirical measures, because different samples $i \neq j$ from P satisfy the "non-branching condition" $\xi_{i,t} \neq \xi_{j,t}$ with probability 1 for every $t \geq 1$ (as P has a density). Hence, with probability 1, empirical measures do not converge in nested distance a.s.

Proof. We give a specific counterexample first.

Consider a pair (ξ_1,ξ_2) which is distributed according to P, the uniform distribution on $[0,1] \times [0,1]$. Let Σ_1 be the σ-algebra generated by ξ_1. We aim at solving the optimal prediction problem

$$v(\mathbb{P}) = \min\left\{\mathbb{E}_P[|\xi_2 - x|] : \ x \lhd \Sigma_1\right\} \tag{10}$$

for the underlying model and for its empirical approximation. Notice that one may solve (10) by decomposing it into the conditional problems

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

which has the optimal decision $x_1(\xi_1) = \frac{1}{2}$ (constant and not depending on ξ_1) with optimal value

$$v(\mathbb{P}) = \int_0^1 \left|u - \frac{1}{2}\right| \mathrm{d}u = \frac{1}{4}.$$

Consider the discrete measure $P_n = \sum_{i=1}^n w_i \cdot \delta_{\xi^{(i)}}$ and recall that all $\xi^{(i)} = (\xi_1^{(i)}, \xi_2^{(i)})$ are different with probability 1. Then problem (10), formulated for the measure P_n, can also be decomposed into the conditional problems

$$\min_{x_1 \lhd \mathcal{F}_1} \mathbb{E}_{P_n}\left[|\xi_2 - x_1| \,\middle|\, \Sigma_1\right],$$

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

and this problem has the optimal solution

$$x_1(\xi_1^{(i)}) = \begin{cases} \xi_2^{(i)} & \text{if } \xi_1 = \xi_1^{(i)}, \\ \text{arbitrary} & \text{else.} \end{cases}$$

Note that $x_1(\cdot)$ is well defined, as all $\xi_1^{(i)}$ are all different by assumption. Obviously, the optimal value of (10) is

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

Now, according to Theorem 5 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

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

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

$$\frac{1}{4} = |v(\mathbb{P}) - v(\mathbb{P}_n)| \le \mathsf{dl}(\mathbb{P}, \mathbb{P}_n)$$

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

The general case follows in the same way as above by considering the support of the measure, which has a density. □

Remark 9. It is well known that the empirical measure converges a.s. weakly to the underlying distribution on separable metric spaces (see Varadarajan [27]). Under the assumption of finite r-th 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 $n \to \infty$. But 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 satisfying the *non-branching conditions* (9) at every stage except the root is 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 to handle approximations of stochastic processes on filtered spaces adequately, have to branch at each stage.

The negative statement contained in Proposition 7 is not a shortfall of the nested distance. To 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 7.

3 Convolution and density estimation

The previous section demonstrates that empirical measures are not adequate models to approximate 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 so, we dilute the original paths $(\xi^{(i)})_{i=1}^n$ in a way which makes differently continuing paths possible. We dilute the observations $(\xi^{(i)})_{i=1}^n$ by convoluting them with a pre-specified kernel, as is known from density estimation. We demonstrate that by introducing an appropriate amount of blur, the paths with a similar past cannot be distinguished any longer. This allows for the possibility of different continuations than 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.,

$$(P * Q)(A) = \iint \mathbb{1}_A(x + y)P(\mathrm{d}x)Q(\mathrm{d}y), \qquad A \text{ measurable.} \tag{11}$$

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 10. 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 11 (Notational convention). We shall write P^f for the measure induced by the Lebesuge 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

$$\tilde{P}_n = \sum_{i=1}^{n} w_i \cdot \delta_{\xi_i} \tag{12}$$

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

$$\sum_{i=1}^{n} w_i \cdot \frac{1}{h^m} k\left(\frac{x-\xi_i}{h}\right). \tag{13}$$

The usual Rosenblatt-Parzen 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

$$\hat{f}_{k_{h_n}}(\cdot) := \frac{1}{n\, h_n^m} \sum_{i=1}^{n} k\left(\frac{\cdot - \xi_i}{h_n}\right), \tag{14}$$

the ususal Nadaraya-Watson estimate, where the bandwidth h_n may depend on n. Employing the notational convention 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

$$\int u_i\, k(u)\, \mathrm{d}u = 0 \tag{15}$$

for all i.

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

$$\mathbb{E}\hat{f}_n(x) = \int k_{h_n}(x-y)\, f(y)\, \mathrm{d}y = f * k_{h_n}(x), \tag{16}$$

where $*$ denotes the usual convolution of densities. It follows from (16) that $\hat{f}_n(x)$ is biased in general. The bias can be stated as

$$\begin{aligned} \mathsf{bias}\hat{f}_n(x) &:= \mathbb{E}\hat{f}_n(x) - f(x) = \frac{1}{h_n^m} \iint k\left(\frac{x-y}{h_n}\right) \big(f(y) - f(x)\big)\, \mathrm{d}y \\ &= \iint k\,(u)\, \big(f(x - h_n \cdot u) - f(x)\big)\, \mathrm{d}u. \end{aligned} \tag{17}$$

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. Indeed, by assuming that f is smooth and employing a Taylor series expansion (17) reduces to

$$\begin{aligned}\mathsf{bias}\hat{f}_n(x) &= \iint k(u)\left(f(x) - f'(x)^\top h_n u + \frac{1}{2}(h_n u)^\top f''(x)(h_n u) - f(x) + o(h_n^2)\right)\mathrm{d}u \\ &= \frac{1}{2}h_n^2 \sum_{i,j=1}^{m}\left(f''_{i,j}(x)\cdot\kappa_{i,j}\right) + o\big(h_n^2\big), \end{aligned} \tag{18}$$

whenever (15) holds and where κ is the matrix with entries $\kappa_{i,j} = \iint u_i u_j k(u)\,\mathrm{d}u$. Note that expression (17), as well as the approximation (18) 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 (17) and (18), provided that $h_n \to 0$.

Convergence. The variance of the multivariate kernel statistics is

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

and the *mean square error* is given by

$$\mathsf{MSE}\, f_n(x) := \mathbb{E}\left(f_n(x) - f(x)\right)^2 = \mathsf{bias}^2 f_n(x) + \mathsf{var} f_n(x).$$

To minimize the mean square 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 (18) for the bias. This can be accomplished by assuming that k has uncorrelated components, i.e.,

$$\kappa_{i,j} = \iint u_i u_j k(u)\mathrm{d}u = 0 \text{ whenever } i \neq j. \tag{19}$$

Then the mean square error is minimized for

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

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

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

is to be minimized, then the optimal bandwidth is

$$h_n^{m+4} \simeq \frac{m}{n} \cdot \frac{\iint k(u)^2 \,\mathrm{d}u}{\left(\sum_{i=1}^m \kappa_{i,i} \iint f_{x_i x_i} \,\mathrm{d}x\right)^2}, \tag{21}$$

which is the same order as in (20).[2]

Remark 12. Assumption (19) is an assumption on the kernel k. Any kernel exhibiting the product form

$$k(u) = k_1(u_1) \cdot k_2(u_2) \cdot \ldots k_m(u_m) \tag{22}$$

satisfies this assumption. The bias (18) of a product kernel of the particular form (22) reduces to

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

where

$$\kappa^{(2)} := \int u^2 k(u) \mathrm{d}u \tag{23}$$

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

Remark 13. Both formulae ((20) and (21)) 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 [26]). Different methods to obtain an optimal bandwidth as cross-validation are designed to overcome this difficulty and outlined in Racine et al. [22], e.g., or plug-in rules of Sheather [23].

Asymptotic normality. The kernel density estimator (13) is a sum of independent, identically distributed random variables. Evoking the central limit theorem (CLT) for independent identically distributed random variables, it is expected that after correcting the bias (18), the estimator $\hat{f}_n(x)$ satisfies the CLT

$$\sqrt{n\, h_n^m} \left(\hat{f}_n(x) - f(x) - \frac{\kappa_{(2)}}{2} \sum_{s=1}^m h_n^2 f_{x_s x_s} \right) \xrightarrow{d} N\left(0,\, f(x) \kappa_{(2)}^d \right), \tag{24}$$

where

$$\kappa_{(2)} := \int k(u)^2 \mathrm{d}u$$

(notice the difference to (23)). This is indeed the case, as is shown in Li and Racine [12, Theorem 1.3] under mild regularity conditions by employing Liapunov's central limit theorem for triangular arrays.

Remark 14 (Over- and undersmoothing). Notice that the bias term in (24) cannot be dropped if the bandwidth is chosen as proposed in (20) or (21), because $\sqrt{n\, h_n^m} \cdot h_n^2 \sim 1$ whenever $h_n \sim n^{-1/(m+4)}$. By choosing $h_n \sim n^{-\alpha}$ for some $\alpha > 1/(m+4)$, the bias is asymptotically negligible relative to $\hat{f}_n - f$. This is known as undersmoothing.

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

In case of oversmoothing (for example if $h_n \sim n^{-\alpha}$ and $\alpha < 1/(m+4)$) the normalized term $\sqrt{n\,h_n^m} \cdot \left(\hat{f}_n - f\right)$ in (24) diverges, but $\hat{f}_n - f$ still converges. The following statements are provided in terms of $\hat{f}_n - \mathbb{E}\hat{f}_n$ instead of $\hat{f}_n - f$ to automatically correct for the bias term $\mathsf{bias} = \mathbb{E}\hat{f}_n - f$.

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 [25] and Wied and Weißbach [29, Theorem 2]).

Theorem 15 (Uniform consitency)**.** *Suppose 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*

$$h_n \to 0,\ \frac{nh_n^m}{|\log h_n|} \to \infty,\ \frac{|\log h_n|}{\log\log n} \to \infty \text{ and } nh_n^m \to \infty, \tag{25}$$

then

$$\lim_{n\to\infty} \sqrt{\frac{nh_n^m}{\log h_n^{-m}}} \cdot \left\| \hat{f}_n - \mathbb{E}\hat{f}_n \right\|_D = \|k\|_2 \sqrt{2\,\|f\|_D} \quad a.s., \tag{26}$$

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

Remark 16. Einmahl and Mason outline in [7] that the result of Theorem 15 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 (26) exists *almost everywhere*.

3.3 Conditional density estimation

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

$$f(x|y) = \frac{f(x,y)}{f(y)},\ \text{where } f(y) = \int f(x,y)\,\mathrm{d}x \tag{27}$$

(here Y is the explanatory variable in (27), 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

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

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

$$\begin{aligned}\hat{f}_n(x|y) &:= \frac{\hat{f}_n(x,y)}{\hat{f}_n(y)} = \sum_{i=1}^n \frac{k_{h_n}(y-Y_i)}{\sum_{j=1}^n k_{h_n}(y-Y_j)} \cdot k_{h_n}(x-X_i) \\ &= \sum_{i=1}^n \frac{\frac{1}{h_n^{m_y}} k\left(\frac{y-Y_i}{h_n}\right)}{\sum_{j=1}^n \frac{1}{h_n^{m_y}} k\left(\frac{y-Y_j}{h_n}\right)} \cdot \frac{1}{h_n^{m_x}} k\left(\frac{x-X_i}{h_n}\right) \end{aligned} \tag{28}$$

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 (28) for the conditional density rewrites as

$$\hat{f}_n(x|y) = \sum_{i=1}^n w_i^{(n)}(y) \cdot k_{h_n}(x-X_i), \text{ where } w_i^{(n)}(y) := \frac{k\left(\frac{y-Y_i}{h_n}\right)}{\sum_{j=1}^n k\left(\frac{y-Y_j}{h_n}\right)} \tag{29}$$

are the weights corresponding to the conditioning y. The conditional estimator (29) is of the same type as the kernel estimator (14), except that the weights are $w_i^{(n)}(y)$ instead of $1/n$. Notice that the Nadaraya–Watson estimator (cf. Tsybakov [26]) is of the same type as (29).

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

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

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

Notice that both, $\hat{f}_n(x,y)$ and $\hat{f}_n(x)$ converge in distribution to the respective true values according (24). These ingredients can be combined for the expression

$$\sqrt{n\,h_n^{m_x+m_y}}\left(\hat{f}_n(x|y) - f(x|y) - \frac{\kappa_{(2)}}{2}h_n^2 B(x,y)\right) \xrightarrow{d} N\left(0,\, \kappa_{(2)}^{m_x+m_y}\frac{f(x|y)}{f(x)}\right) \tag{30}$$

on asymptotic normality of the conditional density. Although the expectation of $\hat{f}_n(x|y)$ does not have a closed form as (16) the bias term in (30) 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 (30) and the asymptotic normality of the conditional density (27) are again elaborated in Li and Racine [12, 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 et al. [10] for a further discussion on the integrated mean square 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 (14) provides a *local* approximation of the density function, and the uniform result outlined in Theorem 15 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 only consider 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 (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 has bounded support (cf. Proposition 22 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 17. *For a translation invariant distance* d *(i.e.* $\mathsf{d}(x + z, y + z) = \mathsf{d}(x, y)$*) it holds that*

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

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 et al. [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 for oversmooting, as the subsequent remark outlines.

Proposition 18. *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*

$$\left\| \hat{f}_n - f \right\|_\infty \leq \delta_f(h) + \frac{\|k\|_{Lip}}{h^{m+1}} \mathsf{d}_r(P, \hat{P}_n) \tag{31}$$

(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*.*

Proof. Observe first that

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

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

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

and the assertion is immediate by the triangle inequality. □

Remark 19 (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 (31) is

$$h_n \sim \left(\frac{c(m+1)}{\|f\|_{Lip}} \right)^{\frac{1}{m+2}} n^{-\frac{1}{m(m+2)}} \tag{32}$$

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. (20) and (21) above). As $\frac{1}{m(m+2)} < \frac{1}{m+4}$, the bandwidth (32) oversmoothes the density f.

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

Proposition 20. *Let f and g be densities on $\mathbb{R}^m$. Then the squared L_2-distance is bounded by*

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

for every $r \geq 1$.

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

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

by the Kantorovich–Rubinstein theorem. □

Corollary 21. *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*

$$\int \left(f(x) - \hat{f}_n(x)\right)^2 \mathrm{d}x \leq \left\|f - \hat{f}_n\right\|_{Lip} \cdot \mathsf{d}_r\left(\hat{P}_n * k_h, P\right). \tag{33}$$

for every $r \geq 1$.

Bounds for the Wasserstein distance. The reverse inequalities, which provide bounds of the Wasserstein distance in terms of the Parzen–Rosenblatt density estimator are more delicate. To provide results where we can build on for the nested distance we need to restrict the considerations to spaces with a compact support in $\mathbb{R}^m$.[3]

Proposition 22. *Let K be a compact set and $\beta \geq 1$. 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*

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

hold true. In particular it holds that

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

and

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

Proof. Without loss of generality we may assume that $f_1 \neq f_2$. 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 $0 \leq \mu < 1$. 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 P_1 and P_2 are probability measures, because $f_1 \geq g$ and $\int f_1 - g \,\mathrm{d}\lambda = 1 - \mu$ (and the same for f_2, resp). 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 $\mathsf{d}(x,y)^r = \|x - y\|^r \leq (\|x\| + \|y\|)^r \leq 2^{r-1}\left(\|x\|^r + \|y\|^r\right)$, so

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

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

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

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

By Hölder's inequality on a compact domain K thus

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

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

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

Corollary 23. *Let P^f be a probability distribution on a compact K, induced by a density f. Then*

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

where $\hat{f}_n$ is the kernel density estimator (14), provided that the mean integrated square error MISE tends to 0.

Proof. It follows from Proposition 22 and Markov's inequality that

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

which is the mean integrated square 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 7 and Remark 8 that $\mathsf{dl}\left(\hat{\mathbb{P}}_n, \mathbb{P} \right) > c > 0$ 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 25 below), where $\mathbb{P}_n^{k_h}$ is based on smoothed measures $\hat{P}_n * k_{h_n}$ instead of the empirical measure $\hat{P}_n$. The proof is rather technical. We need the following auxiliary result.

Theorem 24. *Suppose the bandwidth sequence h_n satisfies the conditions of Theorem 15 and the density f is bounded by $0 < u < f(\cdot) < U < \infty$ (cf. Remark 16) on its support. Suppose further that the support $K = \{f > 0\}$ is convex and compact, and f is continuous in the interior of K. Then, for a regular kernel k,*

$$P\Big(\sup_y \mathsf{d}\big(P(\cdot|y), \, \hat{P}_n * k_{h_n}(\cdot|y) \big) > \varepsilon \Big) \to 0 \tag{34}$$

for ever $\varepsilon > 0$.

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 22, Markov's inequality and the triangle inequality that

$$P\Big(\sup_y \mathsf{d}\big(P(\cdot|y),\, \hat{P}_n * k_{h_n}(\cdot|y)\big) > \varepsilon\Big) \le P\left(\sup_y \left\|\hat{f}_n(\cdot|y) - f(\cdot|y)\right\|_2 > \frac{\varepsilon}{C}\right) \tag{35}$$

$$\le \frac{C}{\varepsilon}\mathbb{E}\sup_y \left\|\hat{f}_n(\cdot|y) - f(\cdot|y)\right\|_2$$

$$\le \frac{C}{\varepsilon}\sup_y \left\|\frac{\mathbb{E}\hat{f}_n(\cdot,y)}{\mathbb{E}\hat{f}_n(y)} - f(\cdot|y)\right\|_2 + \frac{C}{\varepsilon}\mathbb{E}\sup_y \left\|\hat{f}_n(\cdot|y) - \frac{\mathbb{E}\hat{f}_n(\cdot,y)}{\mathbb{E}\hat{f}_n(y)}\right\|_2$$

$$\le \frac{C}{\varepsilon}\sup_y \left\|\frac{\mathbb{E}\hat{f}_n(\cdot,y)}{\mathbb{E}\hat{f}_n(y)} - f(\cdot|y)\right\|_2 + \frac{C}{\varepsilon}\lambda(K)^{1/2}\,\mathbb{E}\sup_y \left\|\hat{f}_n(\cdot|\xi) - \frac{\mathbb{E}\hat{f}_n(\cdot,y)}{\mathbb{E}\hat{f}_n(y)}\right\|_\infty. \tag{36}$$

The first summand in (36) is deterministic and converges because the density f is almost everywhere smooth.

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

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

where the latter converges to 0 because of Theorem 15. Convergence is uniform in y, because the constant $c > 0$ can be chosen uniformly on y, and Theorem 15 (Remark 16) ensures uniform convergence. It follows that (35) tends to 0, and the assertion follows. □

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 25 (The nested distance of the convoluted empirical measure converges)**.** *Suppose that*

(i) the conditions of Theorem 24 hold, and

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

Then the nested distance between the filtered spaces $\mathbb{P}_n^k = \left(\Xi, (\Sigma_t)_{t=0,\dots T},\, \hat{P}_n * k_{h_n}\right)$ *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 $n \to \infty$.

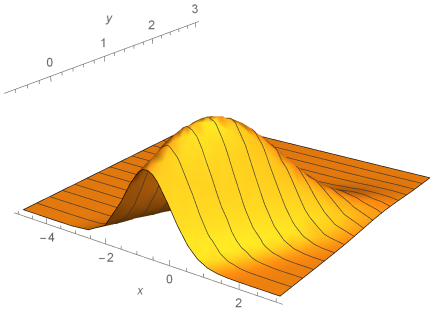


Figure 1: Conditional Gaussian distribution

Example 26. A simple example of a probability satisfying condition (ii) is the Gaussian distribution. Consider a multivariate Gaussian random variable $\begin{pmatrix} Y \\ X \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_Y \\ \mu_X \end{pmatrix}, \begin{pmatrix} \Sigma_{YY} & \Sigma_{YX} \\ \Sigma_{YX} & \Sigma_{XX} \end{pmatrix}\right)$ with regular covariance matrix. Then the distribution of X conditional on $\{Y = y\}$ is a Gaussian variable again with distribution

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

as outlined in Liptser and Shiryaev [13, Theorem 13.1] (cf. Figure 1). 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 25 with

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

Proof of Theorem 25 (cf. also Chapter 4.2 in [18]). 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 a transportation plan π for $\xi_{0:T}$ resp. $\tilde{\xi}_{0:T}$. We get

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

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

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

By assumption one may conclude from (34) 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,

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

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

$$\mathsf{dl}\big(\mathbb{P},\, \mathbb{P}_n^k\big) \le \varepsilon + \varepsilon\,(1+\gamma_T) + \varepsilon\,(1+\gamma_T)\,(1+\gamma_{T-1}) + \dots. \tag{37}$$

The probability of the set, where (37) holds true, is not less than $1 - \varepsilon \cdot T$ whenever $n \ge \max\{n_1, n_2, \dots n_T\}$. Hence

$$P\left(\mathsf{dl}\big(\mathbb{P},\, \mathbb{P}_n^k\big) > C \cdot \varepsilon\right) < \varepsilon \cdot T, \tag{38}$$

where $C := 1 + (1+\gamma_T) + (1+\gamma_T)(1+\gamma_{T-1}) + \dots < \infty$ 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 (38). □

6 Estimating scenario trees based on observed trajectories

The basic data structure for stochastic optimization problems are trees. In order to estimate trees, which are close in nested distance, the distribution conditional on some past $(\xi_0, \dots \xi_t)$ has to be known. The previous sections justify conditional density estimation to estimate these distributions and it follows from the results that the filtered space with measure $P_n * k_h$ converges in nested distance to the genuine model $\mathbb{P}$.

The tree generator algorithm we propose here (Algorithm 1) replaces the probability distribution at the first stage $t = 1$ by the discrete measure $\sum_{i=1}^{b_t} p_i \delta_{\xi_1^{(i)}}$. This can be accomplished based

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 a 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 k be the node to be considered next and let $t < T$ be its stage. Let $\xi_0, \xi_1, \dots \xi_t$ be the already fixed values at node k 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
  $$f(x_{t+1} | \xi_0, \dots \xi_t) \sim \hat{f}_n(x_{t+1} | \xi_0, \dots \xi_t). \tag{39}$$
  (ii) Store the b_t successor nodes and assign to the tree 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.
- **Stopping Criterion.** If all nodes at stage $T-1$ have been considered as parent nodes, the generation of the tree is finished.

on optimal quantizers, cf. Graf and Luschgy [9], or by algorithms outlined in [19]. Recursively, given that the tree is already established for t stages, each path $(\xi_0, \dots \xi_t)$ from the tree already constructed is being considered again. The conditional distribution is estimated from the samples by

$$f(x_{t+1} | \xi_0, \dots \xi_t) \sim \hat{f}_n(x_{t+1} | \xi_0, \dots \xi_t),$$

where f_n is the conditional density. This distribution is again replaced by a discrete probability measure. Algorithm 1 summarizes this procedure.

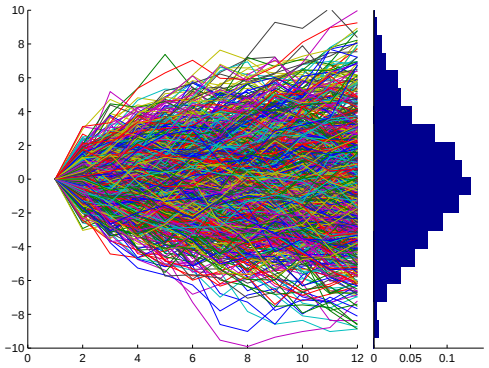
Examples. We demonstrate the behavior of Algorithm 1 by the following 3 examples.

Example 27. Figure 2a displays 1 000 sample paths from a Gaussian walk in 12 stages. A binary tree with 4 095 nodes was extracted (cf. Figure 2b) by employing Algorithm 1. Note that the extracted tree has $2^{11} = 2048$ leaves, which is *more*, even more than twice the size of the original sample ($n = 1000$). Nevertheless, the approximating tree is apparently a useful approximation of the Gaussian process.

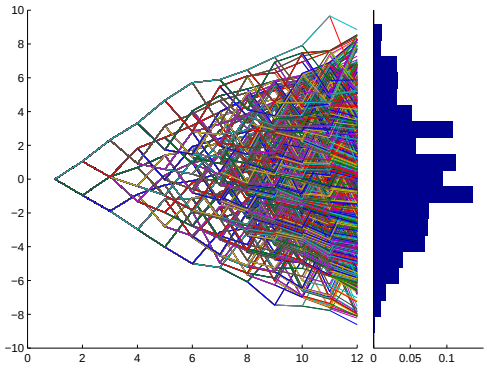
The Figures 2c and 2d display the results of Algorithm 1 for the (non Markovian) running maximum process derived from the samples of a Gaussian walk.

Example 28 (Consistency)**.** This example considers a tree as a starting process. Figure 3 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 still is able to recover the initial tree with reduced branches.

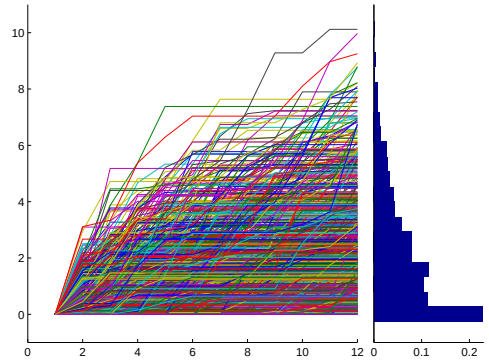
Figure 3 displays the result of the algorithm for a binary tree.



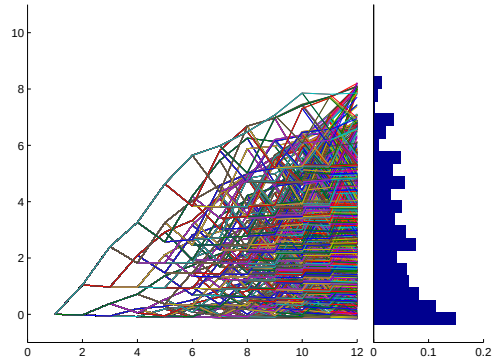
(a) 1 000 sample paths from a (modified) Gaussian random walk



(b) Binary tree of height 12 with 4 095 nodes, approximating the random walk from Figure 2a

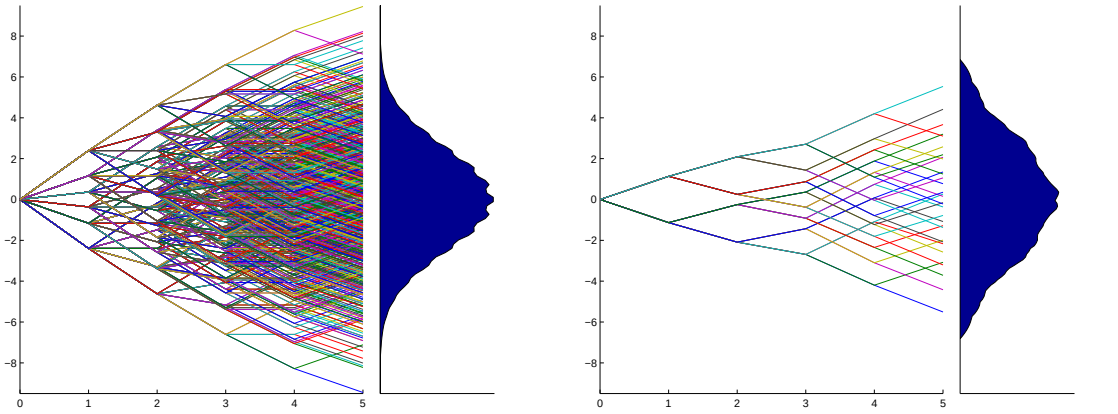


(c) The running maximum process from Figure 2a



(d) Binary tree, extracted from the running maximum process in Figure 2c

Figure 2: 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 3a

Figure 3: Reconstruction of a tree processes

Markovian processes. The transition of a Markovian processes can be described based on the current state only, the entire history is not necessary to do that. 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., Eq. (39) can be replaced by

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

As as consequence the estimator (29) to estimate the conditional density is simplifies significantly, as further dimensions do not have to be included.

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 exposes.

Example 29. The lattice in Figure 4 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 [26] for details). In the present situation the Epanechnikov kernel is not the favorable choice, as a division by zero has to be avoided in (29) (this might be an issue for a small sample size n). This can be avoided by employing, e.g., the logistic kernel

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

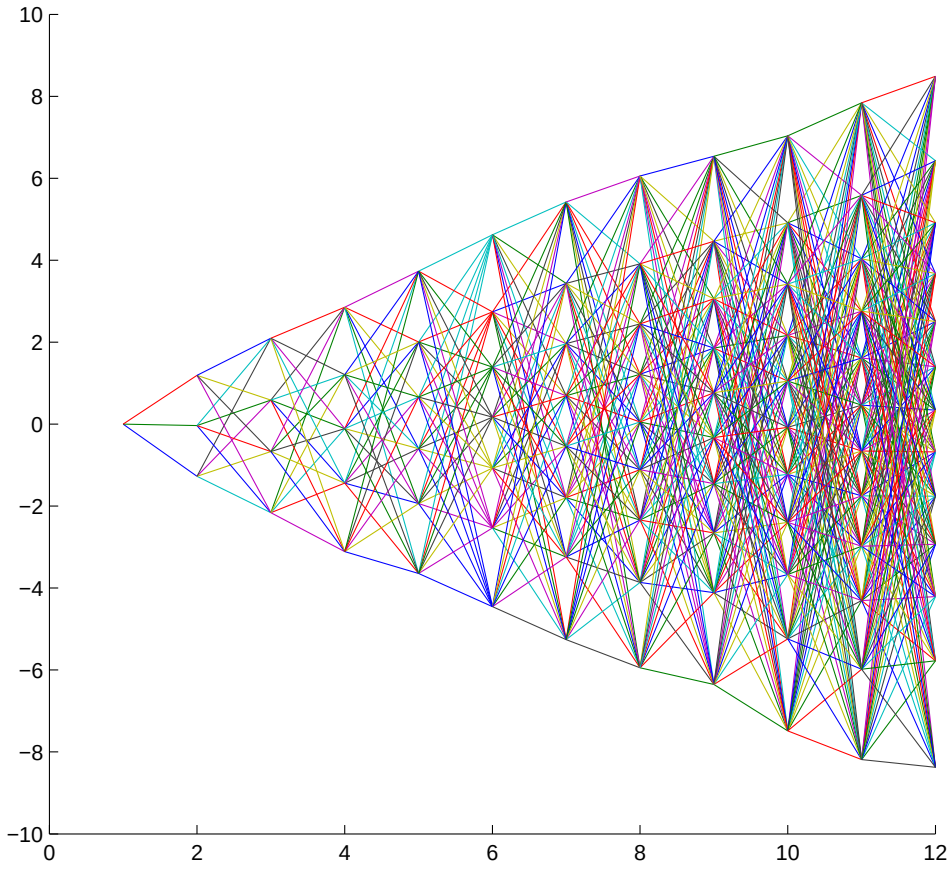


Figure 4: A lattice, constructed from 10 000 empirical observations of sample paths.

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

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

$$\mathbb{E}\, \mathsf{d}_2 \left(P,\, \hat{P}_n\right)^2 \leq \frac{C}{n^{2/(m+4)}}, \tag{40}$$

where $\hat{P}_n$ is the measure with density $\frac{1}{n\, h_n^m} \sum_{i=1}^n k\left(\frac{\cdot - \xi_i}{h_n}\right)$. The rate (40) is the same rate as obtained by Silverman's rule of thumb (cf. Silverman [24]) or Scott's rule, which suggests to use

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

The estimate (40) does not require that the measure P has a density. Slight improvements of the rate of convergence are known in the case that a density is available — cf. Rachev [20] for a discussion.

Remark 30 (Sampling from the kernel estimator). The *compositon method* is a method to provide samples from (14) or (29). To this end one may choose a sample from a uniform random distribution U and pick the index i^* for which

$$\sum_{i=1}^{i^*-1} w_i(\xi) \leq U \leq \sum_{i=1}^{i^*} w_i(\xi).$$

Then the sample from $k_h\,(\cdot - \xi_{i^*})$ follows the conditional distribution (29). Note that this procedure requires only samples from k and thus is very fast, sampling from the smoothed empirical measure thus is computationally cheap. This is easily exploited in implementations.

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 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.

By employing a convolution, as is known from kernel density estimation, it is possible to obtain a process – built solely from sample paths – which converges in nested distribution. The convergence result is stated by employing convergence in probability.

It is further demonstrated that trees constitute representative points (quantizers) of processes. Trees are an adequate finite-space data structures to model processes, which are arbitrarily close in nested distance to a genuine process. As an application we provide an algorithm to construct representative trees from samples. The methods employed are nonparametric, i.e., we do not make parametric assumptions on the underlying process.

References

[1] L. Ambrosio, N. Gigli, and G. Savaré. *Gradient Flows in Metric Spaces and in the Space of Probability Measures.* Birkhäuser Verlag AG, Basel, Switzerland, 2nd edition, 2005. 27

[2] V. Bally, G. Pagès, and J. Printems. A quantization tree method for pricing and hedging multidimensional american options. *Mathematical Finance*, 15(1):119–168, 2005. 2

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

[4] F. Bolley. Separability and completeness for the Wasserstein distance. In C. Donati-Martin, M. Émery, A. Rouault, and C. Stricker, editors, *Séminaire de Probabilités XLI*, volume 1934 of *Lecture Notes in Mathematics*, pages 371–377. Springer Berlin Heidelberg, 2008. 5

[5] F. Bolley, A. Guillin, and C. Villani. Quantitative concentration inequalities for empirical measures on non-compact spaces. *Probability Theory and Related Fields*, 137(3-4):541–593, 2007. ISSN 0178-8051. doi: 10.1007/s00440-006-0004-7. 13

[6] C. Caillerie, F. Chazal, J. Dedecker, and B. Michel. Deconvolution for the Wasserstein metric and geometric inference. *Electronic Journal of Statistics*, 5:1394–1423, 2011. doi: 10.1214/11-EJS646. 24

[7] U. Einmahl and D. M. Mason. Uniform in bandwidth consistency of kernel-type function estimators. *Ann. Statist.*, 33(3):1380–1403, 06 2005. doi: 10.1214/009053605000000129. 11

[8] E. Giné and A. Guillou. Rates of strong uniform consistency for multivariate kernel density estimators. *Annales de l'Institut Henri Poincaré (B) Probability and Statistics*, 38(6):907–921, 2002. ISSN 0246-0203. doi: 10.1016/S0246-0203(02)01128-7. 11

[9] S. Graf and H. Luschgy. *Foundations of Quantization for Probability Distributions*, volume 1730 of *Lecture Notes in Mathematics*. Springer-Verlag Berlin Heidelberg, 2000. 2, 20

[10] R. J. Hyndman, D. M. Bashtannyk, and G. K. Grunwald. Estimating and visualizing conditional densities. *Journal of Computational and Graphical Statistics*, 5(4):315–336, 1996. ISSN 10618600. doi: 10.2307/1390887. 12

[11] O. Kallenberg. *Foundations of modern probability*. Springer, 2002. 4

[12] Q. Li and J. S. Racine. *Nonparametric Econometrics: Theory and Practice*. Princeton University Press, 2006. URL http://books.google.com.au/books?id=Zsa7ofamTIUC. 10, 12

[13] R. S. Liptser and A. N. Shiryaev. *Statistics of Random Processes*. Stochastic Modelling and applied Probability. Springer, 2001. URL http://books.google.com/books?id=gKtK0CjxOaIC. 18

[14] G. Pagès and J. Printems. Functional quantization for numerics with an application to option pricing. *Monte Carlo Methods and Appl.*, 11(11):407–446, 2005. 2

[15] G. Ch. Pflug. Scenario tree generation for multiperiod financial optimization by optimal discretization. *Mathematical Programming*, 89:251–271, 2001. doi: 10.1007/s101070000202. 2

[16] G. Ch. Pflug. Version-independence and nested distribution in multistage stochastic optimization. *SIAM Journal on Optimization*, 20:1406–1420, 2009. doi: 10.1137/080718401. 3

[17] G. Ch. Pflug and A. Pichler. A distance for multistage stochastic optimization models. *SIAM Journal on Optimization*, 22(1):1–23, 2012. doi: 10.1137/110825054. 2, 3, 4

[18] G. Ch. Pflug and A. Pichler. *Multistage Stochastic Optimization*. Springer Series in Operations Research and Financial Engineering. Springer, 2014. doi: 10.1007/978-3-319-08843-3. URL http://books.google.com/books?id=e96ZoAEACAAJ. 18

[19] G. Ch. Pflug and A. Pichler. Dynamic generation of scenario trees. *Computational Optimization and Applications*, 2015. doi: 10.1007/s10589-015-9758-0. 20

[20] S. T. Rachev. *Probability metrics and the stability of stochastic models*. John Wiley and Sons Ltd., West Sussex PO19, 1UD, England, 1991. URL http://books.google.com/books?id=5grvAAAAMAAJ. 24

[21] S. T. Rachev and L. Rüschendorf. *Mass Transportation Problems Vol. I: Theory, Vol. II: Applications*, volume XXV of *Probability and its applications*. Springer, New York, 1998. 3

[22] J. S. Racine, Q. Li, and X. Zhu. Kernel estimation of multivariate conditional distributions. *Annals of Economics and Finance*, 5:211–235, 2004. 10

[23] S. J. Sheather. An improved data-based algorithm for choosing the window width when estimating the density at a point. *Computational Statistics & Data Analysis*, 4(1):61–65, 1986. ISSN 0167-9473. doi: 10.1016/0167-9473(86)90026-5. 10

[24] B. W. Silverman. *Density Estimation for Statistics and Data Analysis*. London: Chapman & Hall/CRC, 1998. 24

[25] W. Stute. The oscillation behavior of empirical processes: The multivariate case. *The Annals of Probability*, 12(2):361–379, 05 1984. doi: 10.1214/aop/1176993295. 11

[26] A. B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer, 2008. 10, 12, 22

[27] V. S. Varadarajan. Weak convergence of measures on separable metric spaces. *Sankhyā: The Indian Journal of Statistics (1933-1960)*, 19(1/2):15–22, 1958. ISSN 00364452. URL http://www.jstor.org/stable/25048364. 6

[28] C. Villani. *Topics in Optimal Transportation*, volume 58 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2003. ISBN 0-821-83312-X. URL http://books.google.com/books?id=GqRXYFxe0l0C. 3

[29] D. Wied and R. Weißbach. Consistency of the kernel density estimator: a survey. *Statistical Papers*, 53(1):1–21, 2012. ISSN 0932-5026. doi: 10.1007/s00362-010-0338-1. 11

A Appendix

Proof of Lemma 17. We shall prove that

$$\mathsf{d}_r(\tilde{P} * k_h, P * k_h) \leq \mathsf{d}_r(\tilde{P}, P) \tag{41}$$

and

$$\mathsf{d}_r(P * k_h, P) \leq \kappa_r^{1/r} \cdot h, \tag{42}$$

from which the assertion follows by employing the triangle inequality for the distance d_r (cf. Ambrosio et al. [1]).

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

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

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

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

by (11), the second marginal $\tilde{\pi}(\Omega \times B) = (\tilde{P} * k)(B)$ equality holds by a symmetric reasoning. It follows that

$$\begin{aligned} \iint \mathsf{d}(x, y)^r \tilde{\pi}(dx, dy) &= \iiint \mathsf{d}(x + x', y + x')^r k(dx')\pi(dx, dy) \\ &= \iint \mathsf{d}(x, y)^r k(dx')\pi(dx, dy) = \iint \mathsf{d}(x, y)^r \pi(dx, dy), \end{aligned}$$

as the distance is translation invariant. The inequality (41) follows by taking the infimum over all appropriate $\tilde{\pi}$ on the left hand side.

As for (42) define the measure

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

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

$$\begin{aligned} \mathsf{d}(P * K_h, P)^r &\leq \iint d(x, x + x')^r k_h(dx')\, P(dx) = \iint \|(x - h\, x') - x\|^r\, k(dx')\, P(dx) \\ &= \iint \|x'\|^r h^r k(dx')\, P(dx) = \kappa_r \cdot h^r, \end{aligned}$$

which is the assertion. □