Scenario Trees – A Process Distance Approach

Raimund M. Kovacevic [a,b], Alois Pichler [a,*]

[a] *University of Vienna, Austria. Department of Statistics and Operations Research .*
[b] *Funded by WWTF.*

Abstract

The approximation of stochastic processes by trees is an important topic in multistage stochastic programming. In this paper we focus on improving the approximation of large trees by smaller (tractable) trees. The quality of the approximation is measured by the nested distance, recently introduced in [Pflug, Pfl09]. The nested distance is derived from the Wasserstein distance. It additionally takes into account the effect of information, which is increasing over time.

After discussing the basic relations between processes and trees and reviewing the nested distance we introduce and analyze an algorithm for finding good approximations. The algorithm, step by step, improves the probabilities on a tree, and also improves the paths. For the important case of quadratic nested distances the algorithm, generalizing multistage, k-means clustering, finds locally best approximating trees in finitely many iterations.

Keywords: Stochastic processes and trees, Wasserstein and Kantorovich Distance, Tree approximation, Optimal transport, Facility location
2010 MSC: 90C15, 60B05, 90-08

1. Introduction

Stochastic programming is an important methodology for decision making under uncertainty. Typical formulations may involve minimizing expected loss or maximizing expected profit and can be extended by risk averse optimization (e.g. using expected utility or coherent risk measures). In addition it is possible to deal with stochastic constraints e.g. by recourse formulations or by using probabilistic constraints. See [SDR09] for a recent and comprehensive treatment of the main techniques.

An important subset of stochastic programs is given by multistage stochastic programs, where decisions have to be taken at multiple points in time. As a simple example we sketch the problem

$$\begin{array}{ll} \text{maximize} \\ \text{(in } x\text{)} & \mathbb{E}\, H\left(\xi, x\right) \\ \text{subject to} & x_t \in \mathbb{X}_t \quad t \in \{0, \dots T\}\,, \\ & x_t \text{ is measurable with respect to } \mathcal{F}_t, \end{array} \tag{1}$$

which uses a real valued profit functions $H\colon \Xi \times \mathscr{X} \to \mathbb{R}$, a stochastic process $(\xi_t)_t$, and expectation utility in the objective. The decision vector $x = (x_0, x_1, \dots, x_T)$ models decisions at points $t \in \{0, \dots T\}$ in time. Typical constraints are expressed by equations and inequalities and e.g. model budgeting, bookkeeping or storage constraints and any lower and upper bounds on the decisions ($x_t \in \mathbb{X}_t$). The measurability constraints refer to a filtration, modeling the increase in information over time, and express the fact that decisions have to be taken without knowing the future (*nonanticipativity*). Because the decisions x also form a stochastic processes, optimization must be done in function spaces. Unfortunately, only in rare cases an analytic solution can be given.

*Corresponding Author
URL: `http://isor.univie.ac.at/` (Alois Pichler)

In stochastic programming the underlying processes are typically replaced by approximations: The process ξ is replaced by a finitely valued stochastic scenario process ξ' and the decisions x_t are replaced by (often high dimensional) vectors x'_t. In order to model nonanticipativity, it is assumed that the decisions x' are adapted to some suitable filtration $\mathcal{F}'$, related to the discretized process ξ'. This filtration usually is modeled by a tree structure, such that all the relevant information (values, probabilities, decisions) is related to nodes in the tree.

In a tree-based framework it is possible to rewrite problem (1) in straightforward manner

$$\begin{array}{c} \underset{(\text{in } x)}{\text{maximize}} \quad \sum_{i>0} p_i H\left(x_{i-}, \xi_i\right) \\ \text{subject to } x_i \in \mathbb{X}_i. \end{array}$$

The notation will be clarified later on (Section 4), but it is important to note that the random process ξ and the random decision process x is now modeled just by real values, sitting on the nodes i of the tree. This means that optimization can done numerically with $x_i \in \mathbb{R}^n$ for some n.

From this sketch it should be clear that the construction of trees – in fact the approximation of a process by a tree – is an important topic in multistage stochastic programming. Different approaches have been used in literature. Besides just simulating trees by Monte Carlo simulation, the most popular approach consists in constructing trees such that the conditional moments (up to some order) of the tree are close to the conditional moments of the real process (moment method). This approach was generalized in [HW01] (cf. also [Kla02]) by minimizing the Euclidean distance between whole collections of statistical properties. Important work also was done on using probability metrics, e.g. based on the Wasserstein/ Kantorovich-distances [DGKR03]. The Wasserstein distance is a transportation distance intending to minimize total costs that have to be taken into account when passing from a given distribution to a desired one. Other concepts of distances for multistage stochastic programming emphasize the role of filtrations and use distances between filtrations, as introduced in [Boy71], see also [Kud74], [HR09]. Recently, [Pfl09] proposed a new type of distance – the nested, or process distance – that extends in a natural way the Wasserstein distance between probability distributions to a distance between processes. Both aspects, the distributional and the filtration, are accounted for by this measure.

The present paper aims at using the process distance for tree construction. We will focus on improving the distance between a given, big scenario tree (constructed e.g. by simulation or any other means) and a smaller scenario tree, suitable for solving a stochastic multistage optimization problem. While this can be done in a relative simple way when using the Wasserstein distance, approximations and iterative approaches for finding local optima have to be used in the tree-case. We introduce suitable algorithms, discuss their properties and provide numerical examples.

In order to clarify the notation the concept of trees and their link to stochastic processes and filtrations are reviewed in Section 2. The Wasserstein distance – as the most similar "ordinary" probability metric – and its key properties as regards the approximation quality are discussed in Section 3. This is the basis to introduce the nested distance in Section 4 and elaborate in Section 5 how to improve the values and the probabilities within a given tree structures in order to improve the approximation quality.

2. Trees and filtrations

When stating stochastic optimization problems it is often advantageous to use the notion of stochastic processes and filtered probability spaces to describe the objects being studied. This is, however, not adequate when implementing concrete realizations in computer models. Typically the models are reformulated in terms of finite state spaces. For multistage stochastic decision problems the basic data structure is given by stochastic trees.

The most important aspect of the nested distance, defined later in Section 4, consists in accounting for the increase of information over time. Usually this is modeled by a filtration. Let $(\Omega, \mathcal{F}_T, P)$ be a

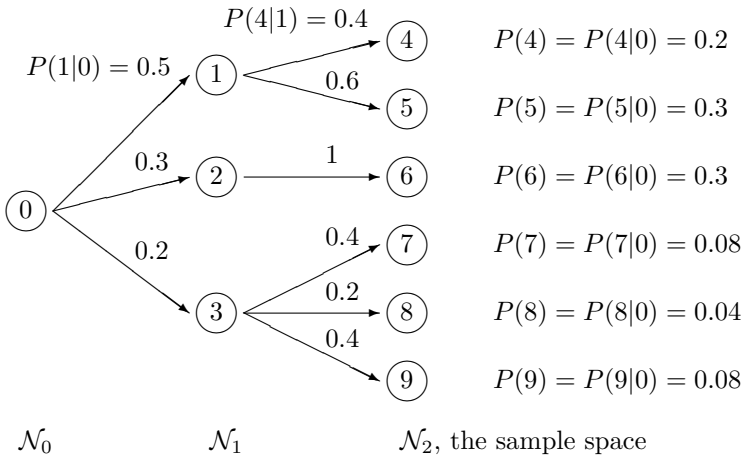


Figure 1: An exemplary finite tree process $\nu = (\nu_0, \nu_1, \nu_2)$ with nodes $\mathcal{N} = \{0, \dots 9\}$ and leaves $\mathcal{N}_2 = \{4, \dots 9\}$ at $T = 2$ stages. The filtrations, generated by the respective atoms, are $\mathcal{F}_2 = \sigma(\{4\}, \{5\}, \dots \{9\})$, $\mathcal{F}_1 = \sigma(\{4,5\}, \{6\}, \{7,8,9\})$ and $\mathcal{F}_0 = \sigma(\{4,5,\dots 9\})$ (cf. [PR07, Section 3.1.1]).

probability space and $\mathcal{F} = (\mathcal{F}_t)_{t\in\{0,1,\dots T\}}$ a family of sigma algebras. Then $\mathcal{F} = (\mathcal{F}_t)_{t\in\{0,1,\dots T\}}$ is a *filtration* provided that $\mathcal{F}_t \subset \mathcal{F}_\tau$ whenever $t \leq \tau$. A triple $(\Omega, \mathcal{F}, P)$, where $\mathcal{F} = (\mathcal{F}_t)_{t\in\{0,1,\dots T\}}$ is a filtration, is called *filtered probability space* or *stochastic basis*.

Because the nested distance is able to compare processes and their approximating scenario trees, it is necessary to review the main properties of scenario trees and to carefully discuss the relations between processes and trees.

Trees

A tree is a directed graph $(\mathcal{N}, A)$ without circles. The vertices $\mathcal{N}$ will be called *nodes* (following [PR07, p. 216]) in the following. A node $m \in \mathcal{N}$ is a *direct predecessor* or *parent* of the node $n \in \mathcal{N}$ if $(m, n) \in A$. The parental relation between m and n is denoted by $m = n_-$. The set of *direct successor*s (or *children*) of a vertex m is m_+, such that $m = n_-$ iff $n \in m_+$. A node $m \in \mathcal{N}$ is said to be a *predecessor* of $n \in \mathcal{N}$ – in symbols $m \supset n$ – if $n_- = n_1$, $n_{1-} = n_2$, and finally $n_{k-} = m$ for some sequence $n_k \in \mathcal{N}$. It holds in particular that $n_- \supset n$.

In addition we assume that any node $n \in \mathcal{N}$ is at a certain *stage* within the tree such that the following properties are fulfilled:

- Nodes at the same stage t are collected in $\mathcal{N}_t$, such that $\mathcal{N}$ is the union of the disjoint subsets $\mathcal{N}_0, \mathcal{N}_1, \dots \mathcal{N}_T$, $\mathcal{N} = \dot{\bigcup}_{t=0}^{T} \mathcal{N}_t$;
- $r \in \mathcal{N}$ is a *root* node if r is a predecessor of all nodes, $r \supset n$ $(n \in \mathcal{N})$, and $\mathcal{N}_0$ (stage 0) contains the root node r. By convention, the unique node of a rooted tree (the root node) is denoted by 0, hence $\mathcal{N}_0 = \{0\}$;
- $i \in \mathcal{N}$ is a *leaf* node if $i_+ = \emptyset$. $\mathcal{N}_T$ collects all *leaf nodes* of the tree, and T is the height of the tree;
- $n \in \mathcal{N}_t$ iff $n_+ \subset \mathcal{N}_{t+1}$.

In concrete implementations all vertices can be numbered consecutively, starting with 0 for the root node (cf. Figure 1).

Any tree induces a filtration

Any tree with height T and finitely many nodes $\mathcal{N}$ naturally induces a filtration $\mathcal{F}$: First use $\mathcal{N}_T$ as sample space. For any $n \in \mathcal{N}$ define the atom[1] $a(n) \subset \mathcal{N}_T$ in a backward recursive way by

$$a(n) := \begin{cases} \{n\} & \text{if } n \in \mathcal{N}_T \\ \bigcup_{j \in n_+} a(j) & \text{else.} \end{cases}$$

Employing these atoms, the related sigma algebra is defined by

$$\mathcal{F}_t := \sigma\left(a(n) : n \in \mathcal{N}_t\right).$$

From the construction of the atoms it is evident that $\mathcal{F}_0 = \{\emptyset, \mathcal{N}_T\}$ for a rooted tree and that $\mathcal{F} = (\mathcal{F}_0, \dots \mathcal{F}_T)$ is a filtration on the sample space $\mathcal{N}_T$, i.e. it holds that $\mathcal{F}_t \subset \mathcal{F}_{t+1}$. Notice that node m is a predecessor of n, i.e. $m \supset n$, if and only if

$$a(m) \supset a(n).$$

This observation suggests the symbol $m \supset n$ introduced in the previous section for the predecessor relation in a tree structure.

Employing the atoms $a(n)$ a *tree process* can be defined by

$$\begin{aligned} \nu : \{0, \dots T\} \times \mathcal{N}_T &\to \mathcal{N} \\ (t,\, i) &\mapsto n \text{ if } i \in a(n) \text{ and } n \in \mathcal{N}_t \text{ (i.e. } n \supset i), \end{aligned}$$

such that each

$$\begin{aligned} \nu_t : \mathcal{N}_T &\to \mathcal{N}_t \\ i &\mapsto \nu(t, i) \end{aligned}$$

is $\mathcal{F}_t-$measurable. Moreover, the process ν is *adapted* to its *natural filtration*, i.e.

$$\mathcal{F}_t = \sigma(\nu_0, \dots \nu_t) = \sigma(\nu_t).$$

It is natural to introduce the notation $i_t := \nu_t(i)$ which denotes the state of the tree process for any final outcome $i \in \mathcal{N}_T$ at stage t. It then holds that $i_T = i$, and moreover that $i_t \supset i_\tau$ whenever $t \leq \tau$, and finally – for a rooted tree – $i_0 = 0$. The *sample path* from the root node 0 to a final node $i \in \mathcal{N}_T$ is

$$(\nu_t(i))_{t=0}^T = (i_t)_{t=0}^T.$$

Any filtration induces a tree

On the other hand, given a filtration $\mathcal{F} = (\mathcal{F}_0, \dots \mathcal{F}_T)$ on a finite sample space Ω it is possible to define a tree, representing the filtration: Just consider the set A_t collecting all atoms generating $\mathcal{F}_t$ ($\mathcal{F}_t = \sigma(A_t)$), and define the nodes

$$\mathcal{N} := \{(a, t) : a \in A_t\}$$

and the arcs

$$A = \{((a, t), (b, t+1)) : a \in A_t,\, a \supset b \in A_{t+1}\}.$$

$(\mathcal{N}, A)$ then is a directed tree respecting the filtration $\mathcal{F}$.

Hence filtrations on a finite sample space and finite trees are equivalent structures up to possibly different labels, and in the following we will not distinguish between them.

[1] A $\mathcal{F}-$measurable set $a \in \mathcal{F}$ is an atom if $b \subsetneq a$ implies that $P(a) = 0$.

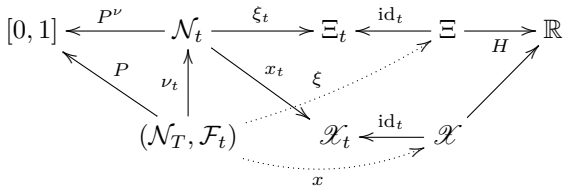


Figure 2: The probabilistic setup: Diagram for the filtered probability space $(\mathcal{N}_T, (\mathcal{F}_t)_t, P)$ (left) , and the value process ξ (including the decision process x) (right).

Measures on trees

Let P be a probability measure on $\mathcal{F}_T$, such that $(\mathcal{N}_T, \mathcal{F}_T, P)$ is a probability space. The notions introduced above allow to *extend* the probability measure to the entire tree via the definition (cf. Figure 1)

$$P^\nu(A) := P\left(\bigcup_{t\in\{0,\dots T\}} \nu_t^{-1}(A \cap \mathcal{N}_t)\right) \qquad (A \subset \mathcal{N}).$$

In particular this definition includes the unconditional probabilities

$$P(\{n\}) =: P(n)$$

for each node. Furthermore it can be used to define conditional probabilities

$$P(\{n\}|\{m\}) =: P(n|m),$$

representing the probability of transition from n to m, if $m \supset n$.

Value and decision processes

In a multi-period, discrete time setup the *outcomes* or *realizations* of a stochastic process are of interest, not the concrete model (the sample space): in focus is the sample space

$$\Xi := \Xi_0 \times \dots \Xi_T$$

of the stochastic process

$$\xi : \{0, \dots T\} \times \mathcal{N}_T \to \Xi.$$

The process is measurable with respect to each $\mathcal{F}_t = \sigma(\nu_t)$, from which follows (cf. [Shi96, Theorem II.4.3]) that ξ can be decomposed as

$$\xi_t = \xi_t \circ \nu_t,$$

(i.e. $\mathrm{id}_t \circ \xi = \xi_t \circ \nu_t$, where $\mathrm{id}_t : \Xi \to \Xi_t$ is the natural projection) as depicted in Figure 1. Notice that $\xi_t \in \Xi_t$ is an observation of the stochastic process at stage t and measurable with respect to $\mathcal{F}_t$ (in symbols $\xi_t \lhd \mathcal{F}_t$), and at this stage t all prior observations

$$\xi_{0:t} := (\xi_0, \dots \xi_t)$$

are $\mathcal{F}_t-$measurable as well.

In *multistage* stochastic programming, a decision maker has the possibility to influence the results to be expected at the very end of the process by making a decision x_t at any stage t of time, having available the information which occurred up to the time when the decision is made, that is $\xi_{0:t}$. The decision has to be taken prior to the next observation ξ_{t+1} (e.g., a decision about a new portfolio allocation has to be made *before* knowing next days security prices).

This *nonanticipativity* property of the decisions is modeled by the assumption that any x_t is measurable with respect to $\mathcal{F}_t$ ($x_t \lhd \mathcal{F}_t$), such that again

$$x_t = x_t \circ \nu_t$$

(i.e. $\mathrm{id}_t \circ x = x_t \circ \nu_t$).

3. The Wasserstein distance for probability measures – definition and computation

Probability metrics are functionals that quantify distances between random objects like random variables, random vectors or even random processes. See e.g. [Rac91] for an encyclopedic treatment or [GS02] for a comprehensive overview of relations between some classical probability metrics. An important group of probability metrics is given by the Wasserstein, or Kantorovich distances.

Definition 1. Given two probability spaces $\mathbb{P} := (\Xi, \Sigma, P)$ and $\mathbb{P}' := (\Xi', \Sigma', P')$ and a convex function $d\colon \Xi \times \Xi' \to \mathbb{R}$, the *Wasserstein distance of order* $r \geq 1$ – denoted $\mathsf{d}_r(P, P')$ – is the optimal value of the optimization problem

$$\begin{array}{ll} \underset{(\text{in } \pi)}{\text{minimize}} & \left(\int d(\xi, \xi')^r \, \pi(\mathrm{d}\xi, \mathrm{d}\xi')\right)^{\frac{1}{r}} \\ \text{subject to} & \pi(M \times \Xi') = P(M) \qquad (M \in \Sigma), \\ & \pi(\Xi \times N) = P'(N) \qquad (N \in \Sigma'), \end{array} \tag{2}$$

where the infimum in (2) is among all bivariate probability measures $\pi \in \mathcal{P}(\Xi \times \Xi')$ which are measures on the product sigma algebra $\Sigma \otimes \Sigma'$. Often $\Xi = \Xi'$ and in typical applications of interest $d : \Xi \times \Xi' \to \mathbb{R}$ is a distance function.

Of particular interest is the Wasserstein distance of order $r = 2$ with a Euclidean norm $d(\xi, \xi') = \|\xi - \xi'\|_2$. We shall refer to this combination as the *quadratic Wasserstein distance*.

The Wasserstein distance was treated first in [Mon81] in an entirely different context. A very comprehensive summary can be found in [Vil03]. In the Russian literature (cf. [Ver06]) the Wasserstein distance is rather known under the name Kantorovich distance. As a matter of fact the Wasserstein distance depends on the sigma algebras Σ and Σ'. This fact is neglected by writing $\mathsf{d}_r(P, P')$.

Basically, (2) can be interpreted as a transportation problem. The resulting functional $\mathsf{d}_r(\cdot, \cdot)$ can be shown to be a full distance. Furthermore, convergence in $\mathsf{d}_r(\cdot, \cdot)$ is equivalent to weak convergence plus convergence of the r-th moment (cf. [Vil03]). It has been shown (see e.g. [DGKR03]) that single stage expected loss minimization problems with objective function $\mathbb{E}_\xi H(\xi, x)$ are (under some regularity conditions on the loss function) Lipschitz continuous with respect to the Wasserstein distance.

Remark 1. It should be noted that the Wasserstein distance is a well-defined distance of probability measures, even if the sample spaces Ξ and Ξ' are entirely different. The link between different spaces is provided by the distance function – or cost function – d.

If $P = \sum_i p_i \delta_{\xi_i}$ and $P' = \sum_j p'_j \delta_{\xi'_j}$ are discrete measures on a space Ξ (Ξ', respectively), then the Wasserstein distance can be computed by the linear program (LP)

$$\begin{array}{ll} \underset{(\text{in } \pi)}{\text{minimize}} & \sum_{i,j} d_{i,j}^r \pi_{i,j} \\ \text{subject to} & \sum_j \pi_{i,j} = p_i, \\ & \sum_i \pi_{i,j} = p'_j, \\ & \pi_{i,j} \geq 0, \end{array} \tag{3}$$

where $d_{i,j}$ is the matrix with entries $d_{i,j} = d\left(\xi_i, \xi'_j\right)$.

Remark 2. It can be derived from the complementary slackness conditions for linear programs that the optimizing transport plan $\pi_{i,j}$ in (3) is sparse, i.e. it has at most $|\Xi| + |\Xi'| - 1$ non-zero entries. This corresponds to the number of entries in one row plus one column of the matrix π or d.

3.1. Scenario approximation with Wasserstein distances

Given a probability measure P one might ask for the best approximating probability measure, with support Q. The following Lemma 1 reveals that the probability measure P_Q^*, which is the best approximation of P located just on Q , i.e.

$$\mathsf{d}_r\left(P, P_Q^*\right) \leq \mathsf{d}_r(P, P') \qquad (P'(Q) = 1), \tag{4}$$

can be computed in a direct way.

Lemma 1 (Lower bounds and best approximation)**.** *Let P and P' be probability measures.*

(i) *The Wasserstein distance has the lower bound*

$$\mathsf{d}_r\left(P,\, P'\right)^r \geq \int \min_{\xi\in\Xi'} d\left(\xi,\xi'\right)^r P\left(\mathrm{d}\xi\right). \tag{5}$$

(ii) *The lower bound in (5) is attained if the transport map $\mathbf{T}:\Xi\to\Xi'$ with $\mathbf{T}\left(\xi\right)\in\operatorname{argmin}_{\xi'} d\left(\xi,\xi'\right)$ is measurable. The pushforward $P^*:=P\circ\mathbf{T}^{-1}$ satisfies* [2]

$$\mathsf{d}_r\left(P,\, P^*\right)^r = \int \min_{\xi'\in\Xi'} d\left(\xi,\xi'\right)^r P\left(\mathrm{d}\xi\right). \tag{6}$$

(iii) *If $\Xi=\Xi'$ is vector space, then*

$$\mathsf{d}_r\left(P,\, P^{\tilde{\mathbf{T}}}\right) \leq \mathsf{d}_r\left(P,\, P^{\mathbf{T}}\right),$$

where $\tilde{\mathbf{T}}\left(\xi\right):=\mathbb{E}_P\left(\tilde{\xi}\middle|\,\mathbf{T}\left(\tilde{\xi}\right)=\mathbf{T}\left(\xi\right)\right)$.

Proof of Lemma 1. Let π have the marginals of P and P'. Then

$$\begin{aligned}\int_{\Xi\times\Xi'} d\left(\xi,\xi'\right)^r \pi\left(\mathrm{d}\xi,\mathrm{d}\xi'\right) &\geq \int_{\Xi}\int_{\Xi'} \min_{q\in\Xi'} d\left(\xi,q\right)^r \pi\left(\mathrm{d}\xi,\mathrm{d}\xi'\right)\\ &= \int_{\Xi} \min_{q\in\Xi'} d\left(\xi,q\right)^r P\left(\mathrm{d}\xi\right).\end{aligned}$$

Taking the infimum leads to (5).

Employing the transport map $\mathbf{T}$, define the transport plan $\pi:=P\circ\left(\mathrm{id}_\Xi\times\mathbf{T}\right)^{-1}$ where id_Ξ is the identity on Ξ, i.e.

$$\pi\left(A\times B\right)=P\left(\{\xi\colon\ \left(\xi,\mathbf{T}\left(\xi\right)\right)\in A\times B\}\right)=P\left(\{\xi\colon \xi\in A,\ \mathbf{T}\left(\xi\right)\in B\}\right).$$

π is feasible, hence it has the marginals $\pi\left(A\times\Xi'\right)=P\left(\{\xi\colon \xi\in A,\ \mathbf{T}\left(\xi\right)\in\Xi'\}\right)=P\left(A\right)$ and $\pi\left(\Xi\times B\right)=P\left(\{\xi\colon \mathbf{T}\left(\xi\right)\in B\}\right)=P^{\mathbf{T}}\left(B\right)$. Thus

$$\int\int_{\Xi\times\Xi'} d\left(\xi,\xi'\right)^r \pi\left(\mathrm{d}\xi,\mathrm{d}\xi'\right)=\int_{\Xi} d\left(\xi,\mathbf{T}\left(\xi\right)\right)^r P\left(\mathrm{d}\xi\right)=\int_{\Xi}\min_{\xi'\in\Xi'} d\left(\xi,\xi'\right)^r P\left(\mathrm{d}\xi\right),$$

which is (6).

For the last assertion apply the conditional Jensen's inequality $\varphi\circ\mathbb{E}\left(X|\mathbf{T}\right)\leq\mathbb{E}\left(\varphi\left(X\right)|\mathbf{T}\right)$ to $\varphi\left(y\right):=d\left(x,y\right)$ and obtain

$$d\left(x,\mathbb{E}\left(\mathrm{id}\,|\mathbf{T}\right)\circ\mathbf{T}\right)\leq\mathbb{E}\left(d\left(x,\mathrm{id}\right)|\mathbf{T}\right)\circ\mathbf{T}.$$

The measure $\tilde{\pi}\left(A\times B\right):=P\left(A\cap\tilde{\mathbf{T}}^{-1}\left(B\right)\right)$ has marginals P and $P^{\tilde{\mathbf{T}}}$, from which follows that

$$\begin{aligned}\mathsf{d}_r\left(P,P^{\tilde{\mathbf{T}}}\right)^r &\leq \int d\left(\xi,\tilde{\mathbf{T}}\left(\xi\right)\right)^r P\left(\mathrm{d}\xi\right)=\int d\left(\xi,\mathbb{E}\left(\mathrm{id}\,|\mathbf{T}\right)\circ\mathbf{T}\left(\xi\right)\right)^r P\left(\mathrm{d}\xi\right)\\ &\leq \int \mathbb{E}\left(d\left(\xi,\mathrm{id}\right)^r|\mathbf{T}\right)\left(\mathbf{T}\left(\xi\right)\right)P\left(\mathrm{d}\xi\right)=\int d\left(\xi,\mathbf{T}\left(\xi\right)\right)^r P\left(\mathrm{d}\xi\right)=\mathsf{d}_r\left(P,P^{\mathbf{T}}\right)^r.\end{aligned}$$

□

It should be noted that the measure P_Q^* does *not depend* on the order r. Moreover, given a probability measure P, Lemma 1 allows to find the best approximation, which is located just on finitely many points $Q=\{q_1\ldots q_n\}$. For this consider $\Xi'=Q$, define $p_j^*:=P\left(\mathbf{T}=q_j\right)$ (the collection of distinct sets $\{\mathbf{T}=q_j\}$ is a Voronoi tessellation; for a comprehensive treatment see [GL00] and the œuvre of Gilles Pagès, e.g. [BPP05]) and $P_Q^*:=\sum_j p_j^*\delta_{q_j}$, as above. Then $\mathsf{d}_r\left(P,\,P_Q^*\right)^r=\int\min_{q\in Q} d\left(\xi,q\right)^r P\left(\mathrm{d}\xi\right)$, and no better approximation is possible by Lemma 1. Usually the q_j are called *quantizers,* which we will adopt in the following.

[2] see also [DGKR03, Theorem 2]

Optimal probabilities

According to Lemma 1 the best approximating measure for $P = \sum_i p_i \delta_{\xi_i}$, which is located on Q, is $P_Q^* = \sum_j p_j^* \delta_{q_j}$. The respective linear program is

$$\begin{array}{ll} \text{minimize} \\ \text{(in } \pi\text{)} \\ \text{subject to} \end{array} \quad \begin{array}{l} \sum_{i,j} d_{i,j}^r \pi_{i,j} \\ \sum_j \pi_{i,j} = p_i, \\ \pi_{i,j} \geq 0, \end{array}$$

which is solved by the optimal transport plan

$$\pi_{i,j}^* := \begin{cases} p_i & \text{if } d(\xi_i, q_j) = \min_{q \in Q} d(\xi_i, q) \\ 0 & \text{else} \end{cases} \tag{7}$$

such that

$$p_j^* = \sum_i \pi_{i,j}^* \qquad \text{and} \qquad \mathsf{d}_r \left(P, P_Q^*\right)^r = \mathbb{E}_{\pi^*} d^r. \tag{8}$$

Observe as well that the matrix π^* in (7) has just $|\Xi|$ non-zero entries, which is less than in Remark 3: In every row i of π^* there is just one non-zero entry $\pi_{i,j}^*$.

Given the support points Q, it is hence an easy exercise to look up the closest points according to (7), and sum up their probabilities according (8), such that the solution of (4) is immediately obtained by $P_Q^* = \sum_j p_j^* \delta_{q_j}$.

Optimal supporting points – facility location

Given the previous results on optimal probabilities the problem of finding a sufficiently good approximation of P in the Wasserstein distance reduces to the problem of looking up good locations Q, that is to minimize the function

$$\{q_1, \dots q_n\} \mapsto \mathsf{d}_r \left(P,\, P_{\{q_1,\dots q_n\}}^*\right)^r = \int \min_{q \in \{q_1,\dots q_n\}} d(\xi, q)^r \, P(\mathrm{d}\xi). \tag{9}$$

This problem is often referred to as *facility location* [DH02]. It is not convex, and no closed form solution exists in general, it hence has to be handled with adequate numerical algorithms. Moreover the facility location problem is NP-hard.

For the important case of the quadratic Wasserstein distance Lemma 1 and its proof give rise for an adaption of the k-means clustering algorithm (also referred to as Lloyd's algorithm, [Llo82]). The approach is described in Algorithm 1.

Theorem 1. *The measures P^k generated by Algorithm 1 are improved approximations for P, they satisfy*

$$\mathsf{d}_r \left(P, P^{k+1}\right) \leq \mathsf{d}_r \left(P, P^k\right).$$

Algorithm 1 terminates after finitely many iterations.

In the case of the quadratic Wasserstein distance Algorithm 1 terminates at a local minimum $\{q_1, \dots q_n\}$ of (9).

Proof. Algorithm 1 is an iterative refinement technique, which finds the measure

$$P^k = \sum_{j=1}^n P\left(T_j^k\right) \delta_{q_j^k}$$

after k iterations. By construction of (10) it is an improvement due to Lemma 1, (ii) and (iii), and hence

$$\mathsf{d}_r \left(P, P^{k+1}\right) \leq \mathsf{d}_r \left(P, P^k\right).$$

Algorithm 1
Facility location for $P = \sum_i p_i \delta_{\xi_i}$ in the special case of the Euclidean distance and quadratic Wasserstein distance (order $r = 2$).

Initialization ($k = 0$):
Choose n points $\left\{q_i^0 \colon i = 1, \dots n\right\}$, for example by randomly picking n distinct points from $\{\xi_i \colon i\}$.
Assignment Step:
In each step k assign each ξ_i to the cluster with the closest mean,

$$T_j^k := \left\{\xi_i \colon \left\|\xi_i - q_j^k\right\| \leq \left\|\xi_i - q_{j'}^k\right\| \text{ for all } j' = 1, \dots n\right\}$$

and set

$$P^k := \sum_{j=1}^{n} P\left(T_j^k\right) \delta_{q_j^k}.$$

Update Step:
Set

$$q_j^{k+1} := \sum_{\xi_i \in T_j^k} \frac{P\left(\xi_i\right)}{P\left(T_j^k\right)} \xi_i. \tag{10}$$

Iteration:
Set $k \leftarrow k + 1$ and continue with an assignment step until $\left\{q_j^{(k)} : j = 1, \dots n\right\}$ is met again.

The algorithm terminates after finitely many iterations because there are just finitely many Voronoi-combinations T_j.

For the Euclidean distance and $r = 2$ the expectation $\mathbb{E}X = \sum_i p_i x_i$ minimizes the function

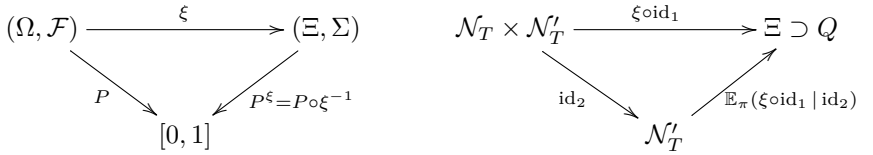
$$q \mapsto \sum_i p_i \cdot \|q - \xi_i\|_2^2 = \mathbb{E}\,\|q - \xi\|_2^2 .$$

In this case P^k thus is a local minimum of (9). □

For other distances than the quadratic Wasserstein distance, P^k is possibly a good starting point, but in general *not* a local (global) minimum of (9).

4. A nested distance for stochastic processes

The Wasserstein distance basically is a distance for random variables. However, it can be used for stochastic processes as well:



(a) Diagram for the push-forward measure P^{ξ}.

(b) The conditional process: The diagram is commutative *in the average*.

Figure 3: Pushforward measure and the projected process.

A random variable

$$\xi \colon (\Omega, \mathcal{F}) \to (\Xi, \Sigma)$$

on a probability space $(\Omega, \mathcal{F}, P)$ naturally induce the push-forward measure

$$P^{\xi} := P \circ \xi^{-1} \colon \Sigma \to [0,\, 1]$$

on the state space (Ξ, Σ) (cf. Figure 3a and Figure 1), such that in particular the distance $\mathtt{d}_r\left(P^{\xi}, P'^{\xi'}\right)$ is available by employing a distance function

$$d \colon \Xi \times \Xi' \to \mathbb{R}$$

defined on the *state spaces* $\Xi \times \Xi'$. If the spaces Ξ and Ξ' are interpreted as containing the sample paths of Stochastic processes, it is possible to consider a process as a random variables, and to apply the Wasserstein distance to the sample paths and their distribution. However, the gradually increasing information, which is the essential ingredient of stochastic processes, is simply ignored by just having a look at the final sigma algebra $\sigma(\xi) \subset \Sigma_T$ instead of the entire filtration $(\Sigma_0, \ldots \Sigma_T)$. Figure 4 shows a situation where similar paths (small ε) lead to a small value of the Wasserstein distance between the first and the second case, which neglects the fact that in the second case perfect information about the final outcome is available already at the intermediary step.

These considerations led to the proposal of a nested distance in [Pfl09, PP12].

4.1. Definition of the nested distance

In the following we will use a multistage (nested) distance concept that shares many properties of the Wasserstein distances but accounts for the effects of filtrations. It was introduced first in [Pfl09] and analyzed in [PP12]. In order to introduce nested distances we have to generalize the distributional concepts used so far from random variables to stochastic processes. For this consider the process $(\xi_t)_{t\in\{0,\ldots T\}}$, where $\xi_t \colon (\Omega, \mathcal{F}) \to (\Xi_t, \Sigma_t)$ are random variables with possibly different state spaces $(\Xi_t,\, \Sigma_t)$. Define the product space $\Xi := \Xi_1 \times \ldots \Xi_T$, which can be equipped itself with the product sigma algebra $\Sigma := \sigma(\Sigma_1 \otimes \ldots \Sigma_T)$. Then

$$\begin{aligned} \xi : (\Omega, \mathcal{F}) \to &\quad (\Xi, \Sigma) \\ \omega \mapsto &\quad (\xi_t(\omega))_{t\in\{0,\ldots T\}} \end{aligned}$$

is a random variable, mapping any outcome $\omega \in \Omega$ to its path $(\xi_t(\omega))_{t=0}^{T}$. The law of of the process ξ,

$$P^{\xi} := P \circ \xi^{-1} \colon \Sigma \to [0,\, 1]\,,$$

is the push-forward measure on $\Xi = \Xi_1 \times \ldots \Xi_T$. The situation for processes is completely analogous to random variables, provided that a distance function

$$d \colon \Xi \times \Xi' \to \mathbb{R}$$

on $\Xi = \Xi_0 \times \ldots \Xi_T$ and $\Xi' = \Xi'_0 \times \ldots \Xi'_T$ is available. For metric spaces (Ξ_t, d_t) the functions $d(\xi, \xi') = \sum_{t=0}^{T} d_t(\xi_t, \xi'_t)$ or $d(\xi, \xi') = \max_{t\in\{0,\ldots T\}} d_t(\xi_t, \xi'_t)$ are immediate candidates. We shall

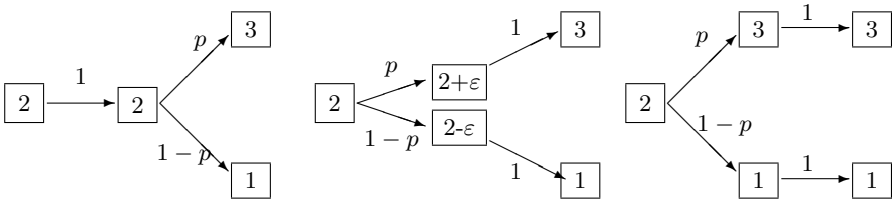


Figure 4: Three tree processes to illustrate three different flows of information: If ε is small the Wasserstein distance will be small too (cf. [HRS06]).

call a metric *(weighted) Euclidean*, if $d(\xi, \xi') = \left(\sum_{t=0}^{T} w_t \left\|\xi_t - \xi'_t\right\|_2^2\right)^{1/2}$, where $w_t > 0$ are positive weights and each norm $\|\cdot\|_2$ satisfies the parallelogram law.

With these preparations the nested distances can be defined as follows:

Definition 2. For two filtered probability spaces $\mathbb{P} := (\Xi, \Sigma, P)$, $\mathbb{P}' := (\Xi', \Sigma', P')$ and a real-valued, convex function $d\colon \Xi \times \Xi' \to \mathbb{R}$ the *nested distance of order* $r \geq 1$ – denoted $\mathsf{dl}_r(\mathbb{P}, \mathbb{Q})$ – is the optimal value of the optimization problem

$$\begin{array}{ll} \text{minimize} \\ \text{(in } \pi\text{)} \\ \text{subject to} \end{array} \quad \begin{array}{ll} \left(\int d(\xi, \xi')^r \, \pi(\mathrm{d}\xi, \mathrm{d}\xi')\right)^{\frac{1}{r}} & \\ \pi(M \times \Xi' \mid \Sigma_t \otimes \Sigma'_t) = P(M \mid \Sigma_t) & (M \in \Sigma_T,\, t \in \{0, \dots T\}), \\ \pi(\Xi \times N \mid \Sigma_t \otimes \Sigma'_t) = P'(N \mid \Sigma'_t) & (N \in \Sigma'_T,\, t \in \{0, \dots T\}), \end{array} \tag{11}$$

where the infimum in (11) is among all bivariate probability measures $\pi \in \mathcal{P}(\Xi \times \Xi')$, which are measures on the product sigma algebra $\Sigma_T \otimes \Sigma'_T$. We will refer to the nested distance also as *process distance*, or *multistage distance*. The nested distance dl_2 (order $r = 2$), with d a weighted Euclidean distance is referred to as *quadratic nested distance*.

Note that the minimization (2) for the Wasserstein distance $\mathsf{d}_r(P, P')$ is a relaxation of (11). Hence the Wasserstein distance is always less or equal to the nested distance,

$$\mathsf{d}_r(P, P') \leq \mathsf{dl}_r(\mathbb{P}, \mathbb{P}').$$

It is possible therefore to decompose the nested distance into the Wasserstein and the effect $\mathsf{dl}_r(\mathbb{P}, \mathbb{P}') - \mathsf{d}_r(P, P')$ caused by the filtration related to the additional constraints in (11).

The multistage distance $\mathsf{dl}_r(\cdot, \cdot)$ also preserves important regularity properties (Lipschitz and Hölder continuity) of the objective function of multistage stochastic programs (see [PP12, Section 6]).

4.2. The nested distance for trees.

The Wasserstein distance between discrete probability measures can be calculated by solving the linear program (3). To establish a similar linear program for the nested distance we use trees that model the whole filtration. Then problem (11) reads

$$\begin{array}{ll} \text{minimize} \\ \text{(in } \pi\text{)} \\ \text{subject to} \end{array} \quad \begin{array}{ll} \sum_{i,j} \pi_{i,j} \cdot d_{i,j}^r & \\ \sum_{j \subset n} \pi(i, j \mid m, n) = P(i \mid m) & (m \supset i,\, n), \\ \sum_{i \subset m} \pi(i, j \mid m, n) = P'(j \mid n) & (n \supset j,\, m), \\ \pi_{i,j} \geq 0 \text{ and } \sum_{i,j} \pi_{i,j} = 1, & \end{array} \tag{12}$$

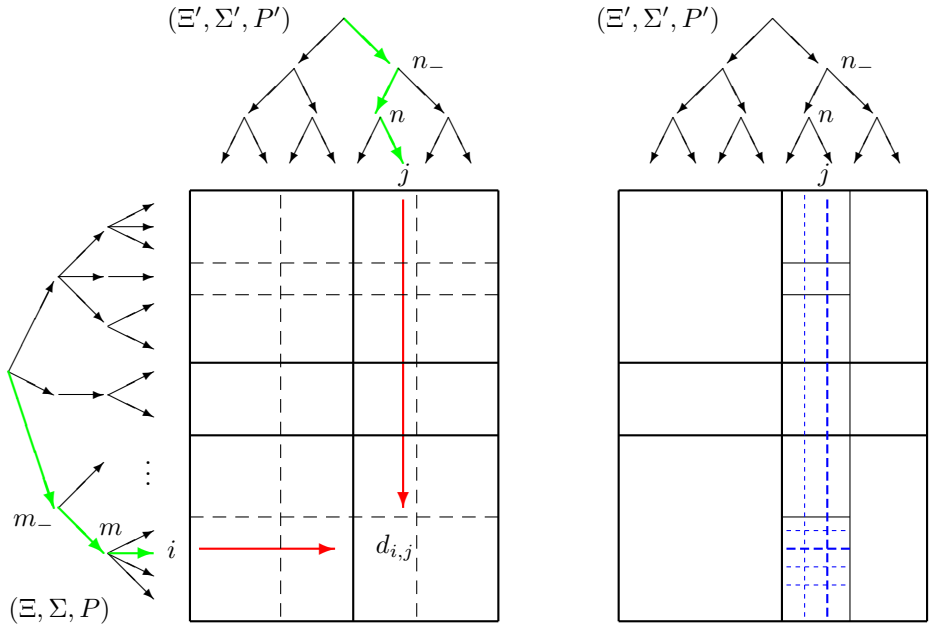
where again $\pi_{i,j}$ is a matrix defined on the samples ($i \in \mathcal{N}_T$, $j \in \mathcal{N}'_T$) and $m \in \mathcal{N}_t$, $n \in \mathcal{N}'_t$ are arbitrary nodes. The conditional probabilities $\pi(i, j \mid m, n)$ are given by

$$\pi(i, j \mid m, n) = \frac{\pi_{i,j}}{\sum_{i' \subset m,\, j' \subset n} \pi_{i',j'}}.$$

The nested structure of the transportation plan π, which is induced by the trees, is schematically depicted in Figure 5.

The constraints in (12) can be written in more detail

$$\begin{aligned} P(i) \cdot \sum_{i' \subset m,\, j' \subset n} \pi_{i',j'} &= P(m) \cdot \sum_{j' \subset n} \pi_{i,j'} \qquad (m \supset i,\, n) \qquad \text{and} \\ P'(j) \cdot \sum_{i' \subset m,\, j' \subset n} \pi_{i',j'} &= P'(n) \cdot \sum_{i' \subset m} \pi_{i',j} \qquad (m,\, n \supset j). \end{aligned}$$



(a) Structure of the transport matrix π for two trees, each of height $T = 3$. m and n are nodes, i and j are leaves.

(b) Structure of the constraints in Algorithm 2

Figure 5: Schematic structure of the distance matrix d and the transport matrix π, as it is imposed by the structures of the trees and the respective constraints.

As P and P' are given, this shows that (12) is equivalent to the linear program

$$\begin{array}{ll} \text{minimize} \\ \text{(in } \pi\text{)} & \sum_{i,j} \pi_{i,j} \cdot d^r_{i,j} \\ \text{subject to} & P(i) \cdot \sum_{i' \subset m,\, j' \subset n} \pi_{i',j'} = P(m) \cdot \sum_{j' \subset n} \pi_{i,j'} \qquad (m \supset i), \\ & P'(j) \cdot \sum_{i' \subset m,\, j' \subset n} \pi_{i',j'} = P'(n) \cdot \sum_{i' \subset m} \pi_{i',j} \qquad (n \supset j), \\ & \pi_{i,j} \geq 0 \text{ and } \sum_{i,j} \pi_{i,j} = 1. \end{array}$$

Remark 3. As a matter of fact many constraints in (12) are linearly dependent. For computational reasons (loss of significance during numerical evaluations, which can impact linear dependencies and the feasibility) it is advisable to remove linear dependencies. This is partially accomplished by the simpler program

$$\begin{array}{ll} \text{minimize} \\ \text{(in } \pi\text{)} & \sum_{i,j} \pi_{i,j} \cdot d^r_{i,j} \\ \text{subject to} & \sum_{j \in (n_-)_+} \pi(m, j \,|\, m_-, n_-) = P(m \,|\, m_-) \qquad (m \in \mathcal{N} \backslash \mathcal{N}_0), \\ & \sum_{i \in (m_-)_+} \pi(i, n \,|\, m_-, n_-) = P'(n \,|\, n_-) \qquad (n \in \mathcal{N}' \backslash \mathcal{N}'_0), \\ & \pi_{i,j} \geq 0 \text{ and } \sum_{i,j} \pi_{i,j} = 1, \end{array} \tag{13}$$

which by [PP12, Lemma 10] is equivalent to (12) and can be reformulated as an LP as well. Further constraints can be removed from (13) by taking into account that $\sum_{i_- = m_-} \frac{P(i)}{P(m_-)} = 1$. Hence, for each node m it is possible to drop one constraint out of all $\left|(m_-)_+\right|$ related equations.

It should be noted that instead of solving (13) the nested distance can be calculated in a recursive way. First define

$$\mathsf{dl}_r(i, j) := d\left(\xi_i, \xi'_j\right) \tag{14}$$

for $i \in \mathcal{N}_T$, $j \in \mathcal{N}'_T$. Given $\mathsf{dl}_r(i,j)$ for $i \in \mathcal{N}_{t+1}$ and $j \in \mathcal{N}'_{t+1}$ set

$$\mathsf{dl}_r(m,n)^r := \sum_{i \in m_+, j \in n_+} \pi(i,j \,|\, m,n) \cdot \mathsf{dl}_r(i,j)^r \qquad (m \in \mathcal{N}_t,\, n \in \mathcal{N}'_t) \tag{15}$$

for $m \in \mathcal{N}_t$, $n \in \mathcal{N}'_t$, where the conditional probabilities $\pi(\cdot,\cdot \,|\, m,n)$ solve

$$\begin{array}{ll} \text{minimize} & \sum_{i \in m+, j \in n+} \pi(i,j \,|\, m,n) \cdot \mathsf{dl}_r(i,j)^r \\ \text{in } \pi(.,. \,|\, m,n) & \\ \text{subject to} & \sum_{j \in n_+} \pi(i,j \,|\, m,n) = P(i \,|\, m) \qquad (i \in m_+), \\ & \sum_{i \in m_+} \pi(i,j \,|\, m,n) = P'(j \,|\, n) \qquad (j \in n_+), \\ & \pi(i,j \,|\, m,n) \geq 0. \end{array}$$

The values $\mathsf{dl}_r(i,j)$ can be interpreted as conditional nested distances for the trees starting in nodes i (j, resp.). Finally the transport plan π on the leaves is recomposed by

$$\pi(i,j) = \pi(i,j \,|\, i_{T-1}, j_{T-1}) \cdot \pi(i_{T-1}, j_{T-1} \,|\, i_{T-2}, j_{T-2}) \cdot \ldots \pi(i_1, j_1 \,|\, 0,0)$$

and the nested distance is given by $\mathsf{dl}_r(\mathbb{P}, \mathbb{P}')^r = \mathsf{dl}_r(0,0)$.

5. Improving an approximating tree

Lemma 1 and the succeeding remark explain how to approximate a probability measure P by a measure P^*_Q, which is located just on the points $Q = \{q_1 \ldots q_n\}$: the measure $P^*_Q = \sum_j P(\mathbf{T} = q_j) \cdot \delta_{q_j}$ (cf. (7)) was found to be the best choice with respect to the Wasserstein distance, irrespective of the order $r \geq 1$. In this section we address the question of looking up processes, i.e. trees, which are close in nested distance to a given tree.

As in Section 3 we split the problem in two parts:

(i) Find *probabilities* on a given tree structure, which improve the nested distance to a given tree;
(ii) facility location: Improve the *locations*, which are the scenarios of the tree, to again improve the approximation overall.

5.1. Optimal probabilities

In a multistage context we have to answer the question which probability measure P^*_Q is *best* to approximate $\mathbb{P} = (\Xi, \Sigma, P)$, provided that the *states* $Q \subset \Xi'$ *and filtration* Σ' of the stochastic processes are given: Knowing the branching structure of the tree, we seek for the best probabilities such that the multistage distance to $\mathbb{P}$ is as small as possible. The best approximation, P^*_Q, satisfies

$$\mathsf{dl}_r\left(\mathbb{P},\, \left(\Xi', \Sigma', P^*_Q\right)\right) \leq \mathsf{dl}_r\left(\mathbb{P},\, (\Xi', \Sigma', P')\right) \quad (P'(Q) = 1),$$

where $Q = \{q_1, \ldots q_n\} \subset \Xi'$.

Compared to the Wasserstein distances it is considerably more difficult to find those optimal probabilities. From (13) it follows that the corresponding transport plan π^* necessarily satisfies

$$\begin{array}{ll} \text{minimize} & \sum_{i,j} \pi^*_{i,j} \cdot d^r_{i,j} \\ (\text{in } \pi^*) & \\ \text{subject to} & \sum_j \pi^*(m,j \,|\, m_-, n_-) = P(m \,|\, m_-), \\ & \sum_i \pi^*(i,n \,|\, m_-, n_-) = \sum_i \pi^*(i,n \,|\, \tilde{m}_-, n_-), \quad m, \tilde{m} \in \mathcal{N}_t \\ & \pi^*_{i,j} \geq 0 \text{ and } \sum_{i,j} \pi^*_{i,j} = 1. \end{array} \tag{16}$$

The constraint

$$\sum_i \pi^*(i,n \,|\, m_-, n_-) = \sum_i \pi^*(i,n \,|\, \tilde{m}_-, n_-) \qquad (m, \tilde{m} \in \mathcal{N}_t) \tag{17}$$

for nodes m and $\tilde{m}$ at the same stage t in (16) ensures that

$$P_Q^*(n|\, n_-) := \sum_i \pi^*(i, n|\, m_-, n_-)$$

is well defined (as it is independent of m), allowing thus to reconstruct a measure P_Q^* by $P_Q^* = \sum_j \delta_{q_j} \cdot \sum_i \pi_{i,j}^*$.

Unfortunately, problem (16) does not allow an immediate solution in general. Moreover the constraints (17), for

$$\frac{\sum_{i\subset m_-} \pi_{i,n}^*}{\sum_{i\subset m_-, j\subset n_-} \pi_{i,j}^*} = \pi^*(n|\, m_-, n_-) = \pi^*(n|\, \tilde{m}_-, n_-) = \frac{\sum_{i\subset \tilde{m}_-} \pi_{i,n}^*}{\sum_{i\subset \tilde{m}_-, j\subset n_-} \pi_{i,j}^*},$$

are not linear any more – in fact they are multilinear in π^*.

Recursive computation of the nested distance

Formulation (16) and the fact that the nested distance can be calculated in a recursive way (see (14) and (15)) leads to the idea of calculating improved probabilities in a recursive way too:

Assume that π is feasible for given quantizers Q. Define

$$\mathsf{dl}_r(i, j) := d(\xi_i, q_j) \tag{18}$$

for $i \in \mathcal{N}_T$, $j \in \mathcal{N}_T'$ and, given $\mathsf{dl}_r(i, j)$ for $i \in \mathcal{N}_{t+1}$ and $j \in \mathcal{N}_{t+1}'$, recursively compute

$$\mathsf{dl}_r(m, n)^r := \sum_{i\in m_+, j\in n_+} \pi^*(i, j|\, m, n) \cdot \mathsf{dl}_r(i, j)^r \qquad (m \in \mathcal{N}_t) \tag{19}$$

for $m \in \mathcal{N}_t$, $n \in \mathcal{N}_t'$, where the conditional probabilities $\pi^*(\cdot, \cdot|\, m, n)$ solve

$$\begin{array}{ll} \text{minimize} & \\ \text{in } \tilde{\pi}(.,.|\, m, n) & \sum_{m\in\mathcal{N}_t} \pi(m, n) \cdot \sum_{i\in m_+, j\in n_+} \tilde{\pi}(i, j|\, m, n) \cdot \mathsf{dl}_r(i, j)^r \\ \text{subject to} & \sum_{j\in n_+} \tilde{\pi}(i, j|\, m, n) = P(i|\, m) \qquad (i \in m_+), \\ & \sum_{i\in m_+} \tilde{\pi}(i, j|\, m, n) = \sum_{i\in \tilde{m}_+} \tilde{\pi}(i, j|\, \tilde{m}, n) \qquad (j \in n_+), \\ & \tilde{\pi}(i, j|\, m, n) \geq 0. \end{array} \tag{20}$$

Recomposing the transport plan π^* on the leaves $i \in \mathcal{N}_T$ and $j \in \mathcal{N}_T'$ by

$$\pi^*(i, j) = \pi^*(i_T, j_T|\, i_{T-1}, j_{T-1}) \cdot \pi^*(i_{T-1}, j_{T-1}|\, i_{T-2}, j_{T-2}) \cdot \ldots \pi^*(i_1, j_1|\, 0, 0) \tag{21}$$

leads to improved probabilities:

Theorem 2. *Let P' be the measure related to the feasible transport probabilities π and P'^* be related to the probabilities π^* by*

$$P'^* := \sum_j \delta_{q_j} \cdot \sum_i \pi^*(i, j).$$

Then $\mathsf{dl}_r(\mathbb{P}, \mathbb{P}'^*) \leq \mathsf{dl}_r(\mathbb{P}, \mathbb{P}')$ and the improved distance is given by

$$\mathsf{dl}_r(\mathbb{P}, \mathbb{P}'^*) = \mathsf{dl}_r(0, 0).$$

Proof. Observe that the measures π and π^* have the iterative decomposition

$$\begin{aligned} \pi(i, j) &= \pi(i_T, j_T) \\ &= \pi(i_T, j_T|\, i_{T-1}, j_{T-1}) \cdot \pi(i_{T-1}, j_{T-1}|\, i_{T-2}, j_{T-2}) \cdot \ldots \pi(i_1, j_1|\, 0, 0), \end{aligned}$$

for all leafs $i \in \mathcal{N}_T$ and $i \in \mathcal{N}_T'$ (cf. [Dur04, Chapter 4, Theorem 1.6]). The terminal distance ($t = T$), given the entire history up to (i, j), is $\mathsf{dl}_{T;r}(i, j) := d(i, j)$, which serves as a starting value for the

iterative procedure. To improve a given transport plan π the algorithm in (20) fixes the conditional probabilities $\pi(m,n)$ in an iterative step at stage t.

The constraints in (20) ensure, for

$$\sum_{i\in m_+}\sum_{j\in n_+}\pi^*(i,j|\,m,n)=\sum_{i\in m_+}P(i|\,m)=1,$$

that π^* again is a probability measure for each $m\in\mathcal{N}_t'$, and hence, by (21), π^* is a probability measure on $\mathcal{N}_T\times\mathcal{N}_T'$. Furthermore the constraints ensure that π^* respects the tree structures of both trees: π^* is feasible for (7). Finally it holds that

$$\sum_{i,j}\pi^*_{i,j}d(i,j)^r=\mathsf{dl}_r(0,0)^r$$

due to the recursive construction.

As the initial π is feasible as well for all equations in (20) it follows from the construction that

$$\mathsf{dl}_r(\mathbb{P},\mathbb{P}'^*)^r=\mathsf{dl}_r(0,0)^r=\mathbb{E}_{\pi^*}d^r\le\mathbb{E}_\pi d^r.$$

As π was chosen arbitrarily it follows that

$$\mathsf{dl}_r(\mathbb{P},\mathbb{P}'^*)\le\mathsf{dl}_r(\mathbb{P},\mathbb{P}'),$$

which shows that P'^* is an improvement. □

5.2. Optimal scenarios – facility location

Consider quantizers

$$Q=\{q_1,\dots q_n\}$$

where each $q_j=(q_{j,0},\dots q_{j,T})$ is a path in the tree. Given a fixed, feasible measure π define

$$D_\pi(\{q_1,\dots q_n\})^r:=\mathbb{E}_\pi d^r=\sum_{i,j}\pi_{i,j}d(\xi_i,q_j)^r. \tag{22}$$

The problem of finding optimal quantizers consists in solving the minimization problem

$$\min_{q_1,\dots q_n}D_\pi(\{q_1,\dots q_n\}). \tag{23}$$

Again it is difficult to solve (23), which can be considered as a facility location problem. However, in an iterative procedure as proposed in the following, a few steps of significant descent in each iteration will be sufficient to considerably improve the overall approximation.

In many applications the gradient of function (22) is available as an analytic expression, for example if $d(\xi_i,\xi_j')=\left(\sum_t d_t(\xi_i,\xi_j')^p\right)^{1/p}$. In this situation the derivative of $D_\pi(\{q_1,\dots q_n\})^r$ is given by

$$\nabla_{\xi_{j,t}}D(\xi')=D_\pi(\xi')^{1-r}\cdot\sum_i\pi_{i,j}d(\xi_i,\xi_j')^{r-p}\cdot d_t(\xi_{i,t},\xi_{j,t}')^{p-1}\cdot\nabla_{\xi_j'}d_t(\xi_{i,t},\xi_{j,t}')\quad(j\in\mathcal{N}_t).$$

If in addition the metric at stage t is a norm, $d_t(\xi_{i,t},\xi_{j,t}')=\|\xi_{i,t}-\xi_{j,t}'\|_s$, then it holds that

$$\nabla_{\xi_{j,t}'}d_t(\xi_{i,t},\xi_{j,t}')=d_t(\xi_{i,t},\xi_{j,t}')^{1-s}\cdot|\xi_{i,t}-\xi_{j,t}'|^{s-2}\cdot(\xi_{i,t}-\xi_{j,t}')$$

which can be obtained by direct computation.

To compute the minimum in (23) a few steps by the steepest descent method will ensure some successive improvements. Another possible method is the limited memory BFGS (Broyden-Fletcher-Goldfarb-Shanno) method, cf [Noc80].

In the special case of the quadratic nested distance the facility location problem can be accomplished by explicit evaluations. This is by far the fastest procedure, and summarized in Algorithm 2, Step 3.

Theorem 3. *For a quadratic nested distance the scenarios*

$$q_t(n_t) := \sum_{m_t \in \mathcal{N}_t} \frac{\pi(m_t, n_t)}{\sum_{m_t \in \mathcal{N}_t} \pi(m_t, n_t)} \cdot \xi_t(m_t)$$

(cf. (24)) are the best possible choice to solve the facility location problem (23).

Proof. The explicit decomposition of the nested distance allows for the re-arrangement

$$\begin{aligned} \mathsf{dl}_2(\mathbb{P}, \mathbb{P}')^2 &= \sum_{i,j} \pi_{i,j} d(\xi_i, q_j)^2 \\ &= \sum_{i,j} \pi_{i,j} \sum_{t=0}^{T} w_t \cdot \|\xi_{i_t} - q_{j_t}\|_2^2 \\ &= \sum_{t=0}^{T} w_t \cdot \sum_{n_t \in \mathcal{N}'_t} \left(\sum_{m_t \in \mathcal{N}_t} \pi(m_t, n_t) \|\xi(m_t) - q_t(n_t)\|_2^2 \right). \end{aligned}$$

By the same reasoning as in the proof of Theorem 1 the assertion follows for every $n_t \in \mathcal{N}_t$ by considering and minimizing every map

$$q \mapsto \sum_{m_t \in \mathcal{N}_t} \pi(m_t, n_t) \cdot \|\xi(m_t) - q\|_2^2$$

separately. □

5.3. The overall algorithm

As it is not possible to improve the probabilities and solve the facility location problem in one single step Algorithm 2 describes the course of action. Starting with an initial guess for the quantizers (resp. the scenario paths) and using the related transport probabilities π^0 the algorithm iterates between improving the quantizers (Step 2) and improving the transport probabilities (Step 3). Step 2 goes backward in time and uses conditional versions $\mathsf{dl}_r^{k+1}(m, n)$ of the nested distance, which are related to nodes m and n, in order to resemble an approximation of the full nested distance. To improve the locations q, Step 3 either uses classical optimization algorithms for the general case, or a version of the k-means algorithm in the important case of the quadratic nested distance.

The algorithm leads to an improvement in each iteration step (Theorem 2 and Theorem 3) and converges in finitely many steps.

Theorem 4. *Provided that the minimization (23) can be done exactly – as is the case for the quadratic nested distance – Algorithm 2 terminates at a stationary* $\mathsf{dl}_r(P, P^{k^*})$ *after finitely many iterations* k^*.

Proof. It is possible – although very inadvisable for computational purposes – to rewrite the computation of $\mathsf{dl}_r^{k+1}(0, 0)$ in Algorithm 2 as a single linear program of the form

$$\begin{array}{ll} \text{minimize} \\ \text{in } \pi^{k+1} & c\left(\pi^{k+1} \middle| \pi^k\right) \\ \text{subject to} & A\pi^{k+1} = b, \\ & \pi^{k+1} \geq 0, \end{array}$$

where the matrix A and the vector b collect all linear conditions from (20), and $\pi \mapsto c(\pi | \tilde{\pi})$ is multilinear. Note that the constraints $\Pi := \{\pi \colon A\pi = b, \pi \geq 0\}$ form a convex polytope, which is independent of the iterate π^k . Without loss of generality one may assume that π^k is an edge of the polytope Π. Because Π has finitely many edges and each edge $\pi \in \Pi$ can be associated with a unique quantization scenario $q(\pi)$, by assumption it is clear that the decreasing sequence

$$\mathsf{dl}_r^{k+2}\left(\mathbb{P}, \mathbb{P}^{k+2}\right) = c\left(\pi^{k+2} \middle| \pi^{k+1}\right) \leq c\left(\pi^{k+1} \middle| \pi^k\right) = \mathsf{dl}_r^{k+1}\left(\mathbb{P}, \mathbb{P}^{k+1}\right)$$

cannot improve further whenever the stationary point is met. □

Algorithm 2
Sequential improvement of the measure P^k to approximate $P = \sum_i p_i \delta_{\xi_i}$ in the nested distance on the trees $(\mathcal{F}_t)_{t\in\{0,\dots T\}}$ ($(\mathcal{F}'_t)_{t\in\{0,\dots T\}}$, resp.).

Step 1– Initialization
Set $k \leftarrow 0$, and let q^0 be process quantizers with related transport probabilities $\pi^0(i,j)$ between scenario i of the original $\mathbb{P}$-tree and scenario q_j^0 of the approximating $\mathbb{P}'$-tree; $\mathbb{P}^0 := \mathbb{P}'$.

Step 2 – Improve the quantizers
Find improved quantizers q_j^{k+1}:

- In case of the quadratic Wasserstein distance (Euclidean distance and Wasserstein of order $r = 2$) set
$$q^{k+1}(n_t) := \sum_{m_t\in\mathcal{N}_t} \frac{\pi^k(m_t,n_t)}{\sum_{m_t\in\mathcal{N}_t}\pi^k(m_t,n_t)} \cdot \xi_t(m_t), \tag{24}$$
- or solve (23), for example by applying the steepest descent method, or the limited memory BFGS method.

Step 3 – Improve the probabilities
Setting $\pi \leftarrow \pi^k$ and $q \leftarrow q^{k+1}$ use (18), (19), (20) and (21) to calculate all conditional probabilities $\pi^{k+1}(\cdot,\cdot|m,n) = \pi^*(\cdot,\cdot|m,n)$, the unconditional transport probabilities $\pi^{k+1}(\cdot,\cdot)$ and the distance $\mathsf{dl}_r^{k+1}(0,0) = \mathsf{dl}_r(0,0)$.

Step 4
Set $k \leftarrow k+1$ and continue with **Step 2** if
$$\mathsf{dl}_r^{k+1}(0,0) < \mathsf{dl}_r^k(0,0) - \varepsilon,$$
where $\varepsilon > 0$ is the desired improvement in each cycle k.
Otherwise, set $q^* \leftarrow q^k$, define the measure
$$P^{k+1} := \sum_j \delta_{q_j^{k+1}} \cdot \sum_i \pi^{k+1}(i,j),$$
for which $\mathsf{dl}_r\left(\mathbb{P},\mathbb{P}^{k+1}\right) = \mathsf{dl}_r^{k+1}(0,0)$ and stop.

Remark. In case of the quadratic nested distance ($r = 2$) and the Euclidean distance the choice $\varepsilon = 0$ is possible.

The same statement as for the Wasserstein distance holds true here for the nested distance: For other distances than the quadratic ones P^k can be used as a starting point, but in general *is not* even a local minimum.

5.4. A numerical example and derived applications

To illustrate the results we have implemented all steps of the discussed algorithms in MATLAB®. All LPs were solved using the function linprog. It is a central observation that optimization for Euclidean norms and the quadratic Wasserstein distance is fastest. This is because the facility location problem can be avoided and replaced by computing the conditional expectation in a direct way. Moreover, when applying the methods, it was a repeated pattern that the first few iteration steps improve the distance significantly, whereas following steps just give minor improvements of the objective.

The computation times collected in Table 1 have been noted for an iteration step in Algorithm 2 on a customary, standard laptop.

Stages	4	5	5	6	* 7	7
Nodes of the initial tree	53	309	188	1,365	1,093	2,426
Nodes of the approximating tree	15	15	31	63	127	127
Time/ sec.	1	10	4	160	157	1,044

Table 1: Time to perform an iteration in Algorithm 2.
The example indicated by the asterisk (*) corresponds to Figure 6.

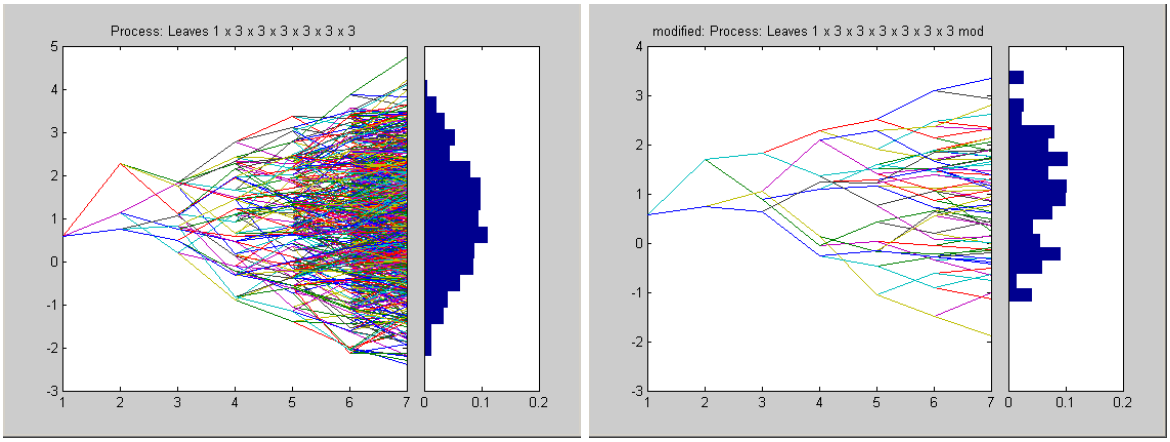


Figure 6: The initial tree with 1093 nodes at 7 stages (left) and a binary, approximating tree, which has 127 nodes (right). Their nested distance is 2.32. The tree structure is depicted, annotated is the histogram of the paths.

Figure 6 exemplary depicts the situation of the latter example with 7 stages. The computed distance of 2.32 allows the rough interpretation, that the scenarios of the initial tree – on average – can be squeezed into a "pipe" of radius 2.32/ 7= 0.3 along a branch of the approximating tree.

6. Summary and outlook

In this paper we address the problem of approximating stochastic processes in discrete time by trees, which are discrete stochastic processes. For this purpose we build on the recently introduced nested distances, generalizations of the well known Wasserstein or Kantorovich distances. In addition to their properties as classical probability metrics they are able to account for the effects of filtrations related to stochastic processes.

In particular we use the nested distance to compare trees, which are important tools for discretizing stochastic optimization problems. The aim is to reduce the distance between a given – usually large – tree, and a smaller tree supposed to approximate the given tree. This problem is of fundamental interest in stochastic programming, as the number of variables of the initial process can be reduced significantly by the techniques and algorithms proposed.

The paper analyzes the relations between processes and trees, reviews the main properties of Wasserstein distances and nested distances and finally proposes and analyzes an iterative algorithm for improving the nested distance between trees. For the important special case of nested distances of order 2 based on Euclidean distances the algorithm can be enhanced by using k-means clustering in order to improve calculation speed.

While first numerical experiences are encouraging, some interesting issues have to be approached in future research: As an example the speed of the algorithm could be further increased by parallelization, as in its Step 3 many conditional distances can be calculated independently and in parallel for each

stage. Furthermore, we will aim at extending the algorithm to improve distances directly between stochastic processes and an approximating tree.

7. Acknowledgment

We wish to express our gratitude to Prof. Georg Ch. Pflug for his continual advice. We thank the referees for their constructive criticism.

8. Bibliography

[Boy71] Edward S. Boylan. Epiconvergence of martingales. *Ann. Math. Statist.*, 42:552–559, 1971. 2

[BPP05] 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. 7

[DGKR03] Jitka Dupačová, Nicole Gröwe-Kuska, and Werner Römisch. Scenario reduction in stochastic programming. *Mathematical Programming, Ser. A*, 95(3):493–511, 2003. 2, 6, 7

[DH02] Z. Drezner and H. W. Hamacher. *Facility Location: Applications and Theory*. Springer, New York, NY, 2002. 8

[Dur04] Richard A. Durrett. *Probability. Theory and Examples*. Duxbury Press, Belmont, CA, second edition, 2004. 14

[GL00] Siegfried Graf and Harald Luschgy. *Foundations of Quantization for Probability Distributions*, volume 1730 of *Lecture Notes in Mathematics*. Springer-Verlag Berlin Heidelberg, 2000. 7

[GS02] A. L. Gibbs and F. E. Su. On choosing and bounding probability metrics. *International Statistical Review*, 70(3):419–435, 2002. 6

[HR09] Holger Heitsch and Werner Römisch. Scenario tree modeling for multistage stochastic programs. *Math. Program. Ser. A*, 118:371–406, 2009. 2

[HRS06] Holger Heitsch, Werner Römisch, and Cyrille Strugarek. Stability of multistage stochastic programs. *SIAM J. Optimization*, 17(2):511–525, 2006. 10

[HW01] Kjetil Høyland and Stein W. Wallace. Generating scenario trees for multistage decision problems. *Management Science*, 47:295–307, 2001. 2

[Kla02] Pieter Klaassen. Comment on "generating scenario trees for multistage decision problems". *Management Science*, 45(11):1512–1516, Nov. 2002. 2

[Kud74] Hirokichi Kudō. A note on the strong convergence of σ-algebras. *Ann. Probability*, 2:76–83, 1974. 2

[Llo82] Stuart P. Lloyd. Least square quantization in PCM. *IEEE Transactions of Information Theory*, 28(2):129–137, 1982. 8

[Mon81] Gaspard Monge. Mémoire sue la théorie des déblais et de remblais. *Histoire de l'Académie Royale des Sciences de Paris, avec les Mémoires de Mathématique et de Physique pour la même année*, pages 666–704, 1781. 6

[Noc80] Jorge Nocedal. Updating quasi-Newton matrices with limited storage. *Mathematics of Computation*, 35(151):773–782, 1980. 15

[Pfl09] Georg Ch. Pflug. Version-independence and nested distribution in multistage stochastic optimization. *SIAM Journal on Optimization*, 20:1406–1420, 2009. 1, 2, 10

[PP12] Georg Ch. Pflug and Alois Pichler. A distance for multistage stochastic optimization models. *SIAM Journal on Optimization*, 22(1):1–23, 2012. 10, 11, 12

[PR07] Georg Ch. Pflug and Werner Römisch. *Modeling, Measuring and Managing Risk*. World Scientific, River Edge, NJ, 2007. 3

[Rac91] Svetlozar T. Rachev. *Probability metrics and the stability of stochastic models*. John Wiley and Sons Ltd., West Sussex PO19, 1UD, England, 1991. 6

[SDR09] Alexander Shapiro, Darinka Dentcheva, and Andrzej Ruszczyński. *Lectures on Stochastic Programming*. MPS-SIAM Series on Optimization 9, 2009. 1

[Shi96] Albert Nikolayevich Shiryaev. *Probability*. Springer, New York, 1996. 5

[Ver06] Anatoly M. Vershik. Kantorovich metric: Initial history and little-known applications. *Journal of Mathematical Sciences*, 133(4):1410–1417, 2006. 6

[Vil03] Cédric Villani. *Topics in Optimal Transportation*, volume 58 of *Graduate Studies in Mathematics*. American Mathematical Society, 2003. 6