A DISTANCE FOR MULTISTAGE STOCHASTIC OPTIMIZATION MODELS

GEORG CH. PFLUG [†‡] AND ALOIS PICHLER[†§]

Abstract. We describe multistage stochastic programs in a purely in-distribution setting, i.e. without any reference to a concrete probability space. The concept is based on the notion of nested distributions, which encompass in one mathematical object the scenario values as well as the information structure under which decisions have to be made.

The nested distance between these distributions is introduced, which turns out to be a generalization of the Wasserstein distance for stochastic two-stage problems. We give characterizations of this distance and show its usefulness in examples. The main result states that the difference of the optimal values of two multistage stochastic programs, which are Lipschitz and differ only in the nested distribution of the stochastic parameters, can be bounded by the nested distance of these distributions. This theorem generalizes the well-known Kantorovich-Rubinstein Theorem, which is applicable only in two-stage situations, to multistage. Moreover, a dual characterization for the nested distance is established.

The setup is applicable both for general stochastic processes and for finite scenario trees. In particular, the nested distance between general processes and scenario trees is well defined and becomes the important tool for judging the quality of the scenario tree generation. Minimizing – at least heuristically – this distance is what good scenario tree generation is all about.

Key words. Stochastic Optimization, Quantitative Stability, Transportation Distance, Scenario Approximation

AMS subject classifications. 90C15, 90C31, 90C08

1. Introduction.

Multistage stochastic programming models have been successfully developed for the financial sector (banking [9], insurance [5], pension fund management [18]), the energy sector (electricity production and trading of electricity [15] and gas [1]), the transportation [6] and communication sector [10] and airline revenue management [22] among others. In general, the observable data for a multistage stochastic optimization problem are modeled as a stochastic process $\xi = (\xi_0, \dots, \xi_T)$ (the scenario process) and the decisions may depend on its observed values, making the problem an optimization problem in function spaces. The general problem is only in rare cases solvable in an analytic way and for numerical solution the stochastic process is replaced by a *finite valued* stochastic scenario process $\tilde{\xi} = (\tilde{\xi}_0, \dots, \tilde{\xi}_T)$. By this discretization, the decisions become high dimensional vectors, i.e. are themselves discretizations of the general decision functions. An extension function is then needed to transform optimal solutions of the approximate problem to feasible solutions of the basic underlying problem.

There are several results about the approximation of the discretized problem to the original problem, for instance [24, 19, 21, 14]. All these authors assume that both processes, the original ξ and the approximate $\tilde{\xi}$ are defined on the same probability space. This assumption is quite *un*natural, since the approximate processes are finite trees which do not have any relation to the original stochastic processes. In this paper, we demonstrate how to define a new distance between the (nested) distributions of the two stochastic processes and how this distance relates to the solutions of multistage stochastic optimization problems.

[†] University of Vienna, Austria. Department of Statistics and Operations Research

[‡] International Institute for Applied Systems Analysis (IIASA), Laxenburg, Austria

[§] alois.pichler@univie.ac.at

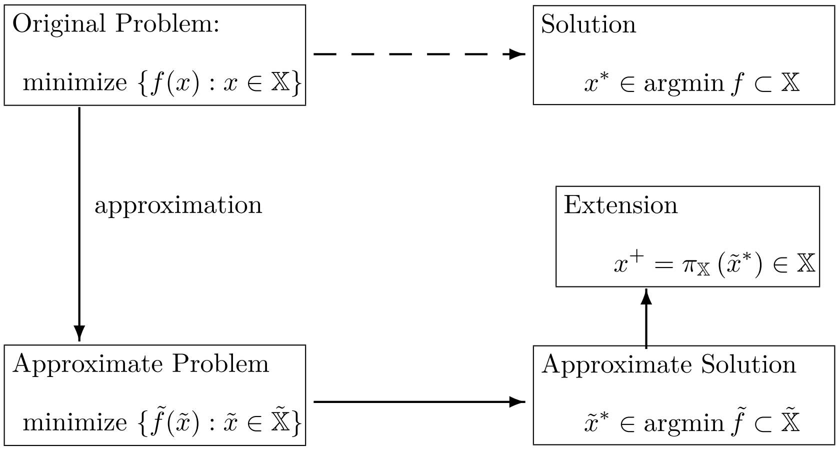
Original Problem:
minimize $\{f(x) : x \in \mathbb{X}\}$
Solution
$x^* \in \operatorname{argmin} f \subset \mathbb{X}$
approximation
Extension
$x^+ = \pi_{\mathbb{X}}(\tilde{x}^*) \in \mathbb{X}$
Approximate Problem
minimize $\{\tilde{f}(\tilde{x}) : \tilde{x} \in \tilde{\mathbb{X}}\}$
Approximate Solution
$\tilde{x}^* \in \operatorname{argmin} \tilde{f} \subset \tilde{\mathbb{X}}$

Fig. 1.1. *The approximation error of the optimization problem is* $f\left(x^+\right) - f\left(x^*\right)$.

Designing approximations to multistage stochastic decision models leads to a dilemma. The approximation should be coarse enough to allow an efficient numerical solution but also fine enough to make the approximation error small. It is therefore of fundamental interest to understand the relation between model complexity and model quality. In Figure 1, f denotes the objective function of the basic problem and $\pi_{\mathbb{X}}$ is the extension of the optimal solution of the approximate problem to a feasible solution of the original problem. Instead of the direct solution (the dashed arrow), one has to go in an indirect way (the solid arrows).

Other concepts of distances for multistage stochastic programming (see [31] for a comprehensive introduction to stochastic programming) use notions of distances of filtrations, as introduced in [3], see also [20] (cf. [12, 14, 13]). The essential progress in this paper is given by the fact that the nested, multistage distance established here naturally incorporates the information, which is gradually increasing in time in a single notion of distance. So a separate concept of a filtration-distance is not needed any longer.

This paper is organized as follows: Section 2 presents a framework for multistage stochastic optimization and develops the terms necessary for a multistage framework. Section 3 introduces a general notion of a tree as probability space to carry increasing information in a multistage situation. The key concept of this paper is the *nested* or *multistage distance*, which is introduced and described in Sections 4 and 5, basic features are being elaborated there as well. The next Section 6 relates the distance to multistage stochastic optimization and contains a main result which states that the new distance introduced is adapted to multistage stochastic optimization in a natural way. Indeed, it turns out that the multistage optimal value is continuous with respect to the nested distance and the nested distance turns out to be the best distance available in the context presented.

As the distance investigated results from a measure which is obtained by an optimization procedure there is a dual characterization as well. We have dedicated Section 7 to elaborate this topic, generalizing the Kantorovich Rubinstein duality theorem for the multistage situation.

Some selected and illustrating examples complete the paper.

2. Definitions and Problem Description. The stochastic structure of two-stage stochastic programs is simple: In the first stage, all decision relevant parameters are deterministic and in the second stage the uncertain parameters follow a known distribution, but no more information is available.

In multistage situations the notion of information is much more crucial: The initially unknown, uncertain parameters are revealed gradually stage-by-stage and this increasing amount of information is the basis for the decisions at later stages.

The following objects are the basic constituents of multistage stochastic optimization problems:

- Stages: Let $\mathbf{T} = \{0, 1, \dots T\}$ be an index set. An element $t \in \mathbf{T}$ is called a *stage* and associated with time. T is the final stage.
- The information process. A stochastic process η_t, $t \in \mathbf{T}$ describes the observable information at all stages $t \in \mathbf{T}$. We assume that the first value η_0 is deterministic, i.e. does not contain probabilistic information. Since information cannot be lost, the information available at time t is the history process $\nu_t = (\eta_0, \dots, \eta_t)$.
- Filtration. Let $\mathcal{F}_t$ be the sigma-algebra generated by ν_t (in symbol $\mathcal{F}_t = \sigma(\nu_t)$). Notice that $\mathcal{F}_0$ is the trivial sigma-algebra, as ν_0 is trivial. The sequence $\mathfrak{F} = (\mathcal{F}_t)_{t=0}^T$ of increasing sigma-algebras[1] is called a *filtration.* We shall write $\nu_t \lhd \mathcal{F}_t$ to express that the function ν_t is $\mathcal{F}_t$ measurable and – following [27] – summarize by writing

  $$\nu \lhd \mathfrak{F}$$

  that $\nu_t \lhd \mathcal{F}_t$ for all $t \in \mathbf{T}$.
- The value process. The process describing the decision relevant quantities is the value process $\xi = (\xi_0, \dots, \xi_T)$. The process ξ is measurable with respect to the filtration $\mathfrak{F}$,

  $$\xi \lhd \mathfrak{F}.$$

  Therefore ξ_t can be viewed as a function of ν_t, i.e. $\xi_t = \xi_t(\nu_t)$ (cf. [32, Theorem II.4.3]) and ξ_0 again is trivial. The values of the process ξ_t lie in a space endowed with a metric d_t. In many situations this is just the linear metric space $(\mathbb{R}^{n_t}, d_t)$, where the metric d_t is any metric, not necessarily the Euclidean one.
- The decision space. At each stage t a decision x_t has to be made, its value is required to lie in a feasible set $\mathbb{X}_t$, which is a linear vector space. The total decision space is $\mathbb{X} := (\mathbb{X}_t)_{t=0}^T$.
- Non-Anticipativity. The decision x_t must be based on the information available at each time $t \in \mathbf{T}$, therefore it must satisfy

  $$x \lhd \mathfrak{F}.$$

  This measurability condition is frequently referred to as *non-anticipativity* in literature.
- The Loss Function is $H(\xi, x)$. In the sequel the cost function may be associated with loss, which is intended to be minimized by choosing an optimal decision x.

[1] $\mathcal{F}_{t_1} \subseteq \mathcal{F}_{t_2}$ whenever $t_1 \leq t_2$.

Multistage stochastic optimization problems with expectation maximization can be framed as

$$\begin{aligned}&\min\{\mathbb{E}H\left(\xi_0, x_0, \dots \xi_T, x_T\right):\, x_t \in \mathbb{X}_t,\, x_t \lhd \mathcal{F}_t,\ t \in \mathbf{T}\}\\&\quad= \min\left\{\mathbb{E}H\left(\xi, x\right):\, x \in \mathbb{X},\, x \lhd \mathfrak{F}\right\}.\end{aligned} \tag{2.1}$$

In applications, problem (2.1) is formulated without reference to a specific probability space, i.e. just the distributions of the stochastic processes are given. Notice that the observable information process ν is determined only up to bijective transformations, since for any $\tilde{\nu}_t$, which is a bijective image of ν_t, the generated filtration $\mathfrak{F}$ is the same. This invariance with respect to bijective transformations becomes in particular evident, if the problem is formulated in form of a scenario tree. Here, typically, the names of the nodes are irrelevant, only the tree's topology and the scenario values ξ sitting on the nodes are of importance.

For this reason we reformulate the setting in a purely in-distribution manner, using the notion of trees in probabilistic sense as outlined in the next section.

3. Trees as the basic probability space. Keeping in mind that the process ν typically represents the history of the information process we may start with directly defining the process ν as a tree process.

- Tree Processes. A stochastic process (ν_t), $t \in \mathbf{T}$ with state spaces in $\mathcal{N}_t$, $t \in \mathbf{T}$ is called *tree process*, if $\sigma\left(\nu_t\right) = \sigma\left(\nu_0, \dots \nu_t\right)$ for all t. A tree process can be equivalently characterized by the fact that the conditional distribution of $(\nu_0, \dots, \nu_{t-1})$ given ν_t is degenerate (i.e. sits on just one value). We denote a typical element of $\mathcal{N}_T$ by ω and its predecessor in $\mathcal{N}_t$ (which is almost surely determined) by ω_t, in symbol $\omega_t = \text{pred}_t\left(\omega\right)$.
- Trees. The tree process induces a probability distribution P on $\mathcal{N}_T$ and we may introduce $\mathcal{N}_T$ as the basic probability space, i.e. we set $\Omega := \mathcal{N}_T$. Ω is a *tree* (of depth T), if there are projections pred_t, $t \in \mathbf{T}$ such that

$$\text{pred}_s \circ \text{pred}_t = \text{pred}_s \qquad (s \le t)\,. \tag{3.1}$$

  Typically a tree is *rooted,* that is pred_0 is one single value. Notice that in this definition a tree does not have to be finite or countable.
  Property (3.1) implies that the sigma algebras $\mathcal{F}_t := \sigma\left(\text{pred}_t\right)$ form a filtration, which is denoted by $\mathfrak{F}\left(\text{pred}\right) := \sigma\left(\text{pred}_t \colon t \in \mathbf{T}\right)$. Without loss of generality we my choose in the following $\left(\Omega, \mathfrak{F}\left(\text{pred}\right), P\right)$ as our basic filtered probability space.
- Value-and-information-structures. As the value process ξ_t is a function of the tree process ν_t, we may view it as a stochastic process on Ω adapted to the filtration $\mathfrak{F}\left(\text{pred}\right)$, i.e. $\xi_t(\omega) = \xi_t(\omega_t)$ with $\omega_t = \text{pred}_t(\omega)$. We call the structure $\left(\Omega, \mathfrak{F}\left(\text{pred}\right), P, \xi\right)$ the *value-and-information-structure.*

It is the purpose of this paper to assign a distance to different value-and-information-structures on the basis of distributional properties only. To this end we introduce the distribution of such a value-and-information-structure as *nested distribution.* This concept is explained in the Appendix.

The nested distributions are defined in a pure distributional concept. The relation between the nested distribution $\mathbb{P}$ and the value-and-information-structure $\left(\Omega, \mathfrak{F}\left(\text{pred}\right), P, \xi\right)$ is comparable to the the relation between a probability measure P on $\mathbb{R}^d$ and a $\mathbb{R}^d$-valued random variable ξ with distribution P.

Bearing this in mind, we may alternatively consider either the nested distribution $\mathbb{P}$ or its realization

$$\left(\Omega, \mathfrak{F}\left(\text{pred}\right), P, \left(\xi_t\right)_{t\in\mathbf{T}}\right)$$

with Ω being a tree, $\mathfrak{F}$ the information and ξ the value process. We symbolize this fact by

$$\left(\Omega, \mathfrak{F}\left(\text{pred}\right), P, \xi\right) \sim \mathbb{P}.$$

In the next section the nested distance between two nested distributions $\mathbb{P}$ and $\tilde{\mathbb{P}}$ is defined. Due to the properties just mentioned one may – without loss of generality – assume that they are represented by two probability distributions P on $\left(\Omega, \mathfrak{F}\left(\text{pred}\right)\right)$ and $\tilde{P}$ on $\left(\tilde{\Omega}, \tilde{\mathfrak{F}}\left(\text{pred}\right)\right)$ together with the value processes $\xi_t\left(\omega_t\right)$ and $\tilde{\xi}_t\left(\tilde{\omega}_t\right)$.

4. The Transportation Distance. The distance of two value-and-information-structures, as defined here, is in line with the concept of transportation distances, which have been studied intensively in the recent past. In order to thoroughly introduce the concept recall the usual Wasserstein or Kantorovich distance for distributions.

Kantorovich Distance and Wasserstein Distance. Transportation distances intend to minimize the effort or total costs that have to be taken into account when passing from a given distribution to a desired one. Initial works on the subject include the original work by Monge [23] as well as the seminal work by Kantorovich [17]; a compelling treatment of the topic can be found in Villani's books [34] and [35], as well as in [28]; the Wasserstein distance has been discussed in [30] as well for stochastic two-stage problems.

The cumulative cost is the sum of all respective distances arising from transporting a particle from ω to $\tilde{\omega}$ over the distance $d(\omega, \tilde{\omega})$. The optimal value to accomplish this is called *Wasserstein* or *Kantorovich distance* of order r ($r \geq 1$) and denoted $\mathsf{d}_r\left(P, \tilde{P}\right)$: [2]

$$\mathsf{d}_r\left(P, \tilde{P}\right) = \inf\left(\int d\left(\omega, \tilde{\omega}\right)^r \pi\left[\mathrm{d}\omega, \mathrm{d}\tilde{\omega}\right]\right)^{\frac{1}{r}}, \tag{4.1}$$

where the infimum is over all bivariate probability measures (also called transportation measures) π on the product sigma algebra

$$\mathcal{F}_T \otimes \tilde{\mathcal{F}}_T := \sigma\left(\left\{A \times B \colon A \in \mathcal{F}_T,\ B \in \tilde{\mathcal{F}}_T\right\}\right)$$

having the initial distribution P and final distribution $\tilde{P}$ as marginals, that is

$$\begin{aligned} \pi\left[A \times \tilde{\Omega}\right] &= P\left[A\right] \quad \text{and} \\ \pi\left[\Omega \times B\right] &= \tilde{P}\left[B\right] \end{aligned} \tag{4.2}$$

for all measurable sets $A \in \mathcal{F}_T$ and $B \in \tilde{\mathcal{F}}_T$. The infimum in (4.2) is attained, i.e. the optimal transportation measure π exists.

The solution to the linear problem (4.1) is well investigated and understood, and in many situations (cf. [29] and [4]) allows a particular representation as a transport plan: That is to say there is a function (a transport map) $\tau : \Omega \to \tilde{\Omega}$ such that π is a simple push-forward or image measure $\pi = P^{\mathrm{id}\times\tau} = P \circ \left(\mathrm{id}\times\tau\right)^{-1}$.

[2] In case of costs which are not proportional to the distance a convex cost function c can be employed instead of d, which generalizes the concept to some extend.

5. The Multistage Distance. In light of the introduction, the problem (4.1) just describes the situation where the filtration consists of a single sigma algebra. To generalize for the multistage situation let two nested distributions $\mathbb{P}$ and $\tilde{\mathbb{P}}$ and a particular realization

$$(\Omega, \mathfrak{F}, P, \xi) \sim \mathbb{P} \text{ and } \left(\tilde{\Omega}, \tilde{\mathfrak{F}}, \tilde{P}, \tilde{\xi}\right) \sim \tilde{\mathbb{P}}$$

be given. We intend to minimize the effort or costs that have to be taken into account when passing from one value-and-information-structure to another.

For this purpose a distance on the original sample space $\Omega \times \tilde{\Omega}$ – as in (4.1) – is needed. This is accomplished by the function

$$d\left(\omega, \tilde{\omega}\right) := \sum_{t=0}^{T} d_t\left(\xi_t\left(\omega_t\right), \tilde{\xi}_t\left(\tilde{\omega}_t\right)\right), \tag{5.1}$$

where d_t is the distance available in the state space of the processes ξ and $\tilde{\xi}$ and $\omega_t = \text{pred}_t\left(\omega\right)$ ($\tilde{\omega}_t = \text{pred}_t\left(\tilde{\omega}\right)$, resp.).[3]

Most importantly, one needs to take care of the gradually increasing information provided by the filtrations. In the presence of filtrations the entire, complete information is available at the very final stage T only via $\mathcal{F}_T$ and $\tilde{\mathcal{F}}_T$. So the optimal measure π for (4.1) in general is *not* adapted to the situations of lacking information, which are described by previous σ-algebras $\mathcal{F}_t$ and $\tilde{\mathcal{F}}_t$, $t < T$.

This is respected by the following definition. The new distance – the multistage distance – then is influenced by both, the probability measure P and the entire sequence of increasing information $\mathfrak{F}$, so that the resulting quantity depends on the entire $\mathbb{P}$.

DEFINITION 5.1 (The multistage distance). *The multistage distance of order $r \geq 0$* [4] *of two value-and-information-structures $\mathbb{P}$ and $\tilde{\mathbb{P}}$ is the optimal value of the optimization problem*

$$\begin{array}{lll} \textit{minimize} & (\textit{in } \pi) & \left(\int d\left(\omega, \tilde{\omega}\right)^r \pi\left[\mathrm{d}\omega, \mathrm{d}\tilde{\omega}\right]\right)^{\frac{1}{r}} \\ \textit{subject to} & & \pi\left[A \times \tilde{\Omega} \mid \mathcal{F}_t \otimes \tilde{\mathcal{F}}_t\right]\left(\omega, \tilde{\omega}\right) = P\left[A \mid \mathcal{F}_t\right]\left(\omega\right) \quad \left(A \in \mathcal{F}_T,\, t \in \mathbf{T}\right), \\ & & \pi\left[\Omega \times B \mid \mathcal{F}_t \otimes \tilde{\mathcal{F}}_t\right]\left(\omega, \tilde{\omega}\right) = \tilde{P}\left[B \mid \tilde{\mathcal{F}}_t\right]\left(\tilde{\omega}\right) \quad \left(B \in \tilde{\mathcal{F}}_T,\, t \in \mathbf{T}\right), \end{array} \tag{5.2}$$

where the infimum in (5.2) *is among all bivariate probability measures $\pi \in \mathcal{P}\left(\Omega \times \tilde{\Omega}\right)$ defined on $\mathcal{F}_T \otimes \tilde{\mathcal{F}}_T$. Its optimal value – the nested distance – is denoted by*

$$\mathsf{dl}_r\left(\mathbb{P}, \tilde{\mathbb{P}}\right). \tag{5.3}$$

REMARK 1. *It is an essential observation that the conditional measures in* (5.2) *depend on two variables $\left(\omega, \tilde{\omega}\right)$, although the conditions imposed force them to effectively depend just on a single variable ω ($\tilde{\omega}$, resp.). To ease the notation we introduce*

[3] Alternatively one may use the equivalent distance functions

$$d(\omega, \tilde{\omega}) = \left(\sum_{t=0}^{T} d_t\left(\xi_t\left(\omega_t\right), \tilde{\xi}_t\left(\tilde{\omega}_t\right)\right)\right)^{\frac{1}{p}}$$

or $d(\omega, \tilde{\omega}) = \max_{t=0 \ldots T} d_t\left(\xi_t\left(\omega_t\right), \tilde{\xi}_t\left(\tilde{\omega}_t\right)\right)$ instead.

[4] This turns out to be a distance only for $r \geq 1$.

the natural projections $\mathrm{i}\colon \Omega\times\tilde{\Omega}\to\Omega$ *and* $\tilde{\mathrm{i}}\colon \Omega\times\tilde{\Omega}\to\tilde{\Omega}$*; moreover,* $f\,\mathrm{i}$ *(*$g\,\tilde{\mathrm{i}}$*, resp.) is just shorthand for the composition* $f\circ\mathrm{i}$ *(*$g\circ\tilde{\mathrm{i}}$*, resp.). The symbol* i *was chosen to account for* identity *onto the respective subspace.*

Remark 2. *Problem* (5.2) *may be generalized by replacing* d^r *by a general (convex) cost function* c*. The theory developed below will not be affected by the particular choice* d^r*, it can be repeated for a general cost function* c*.*

Remark 3. *In the Appendix we show that* dl_r *can be seen as the usual Kantorovich distance on the complex Polish space of nested distributions. Hence* dl_r *satisfies the triangle inequality.*

The definition of the multistage distance builds on conditional probabilities. This comes quite natural as it is built on conditional information. The marginal conditions (5.2) intuitively state that the observation $P\left[A\mid\mathcal{F}_t\right]$, at some previous stage $\mathcal{F}_t$, has to be reflected by $\pi\left[A\times\Omega\mid\mathcal{F}_t\otimes\tilde{\mathcal{F}}_t\right]$, irrespective of the current status of the second process $\tilde{\mathbb{P}}$ and irrespective of the previous outcome in $\tilde{\mathcal{F}}_t$, which represents the information available to $\tilde{P}$ at the same earlier stage $t\in\mathbf{T}$.

As for the notion of conditional probabilities involved in Definition 5.1 we recall the basic features.

Conditional Expectation. For g a measurable function

$$\sigma\left(g\right):=\left\{g^{-1}\left(S\right): S\text{ measurable}\right\}\tag{5.4}$$

is a sigma algebra. By the Radon-Nikodym Theorem (cf. [36]) there is a random variable, denoted $\mathbb{E}\left[X|\,g\right]$ on the image set of g, such that

$$\int_{g^{-1}(S)}X\left(\omega\right)\,P\left[\mathrm{d}\omega\right]=\int_S\mathbb{E}\left[X|\,g\right]\left(s\right)\,\mathbb{P}\left[g^{-1}\left(\mathrm{d}s\right)\right]=\int_{g^{-1}(S)}\mathbb{E}\left[X|\,g\right]\left(g\left(\omega\right)\right)\,\mathbb{P}\left[\mathrm{d}\omega\right]$$

for any measurable S; the relation to conditional expectation with respect to the filtration $\sigma\left(g\right)$ thus is

$$\mathbb{E}\left[X|\,g\right]\circ g=\mathbb{E}\left[X|\,\sigma\left(g\right)\right].$$

Conditional Probabilities. Conditional probabilities are defined via conditional expectation, $P\left[A\mid\mathcal{F}_t\right]:=\mathbb{E}_P\left[\mathbb{1}_A\mid\mathcal{F}_t\right]$, where $A\in\mathcal{F}_T$ and $\mathcal{F}_t\subseteq\mathcal{F}_T$. The conditional probability is a function

$$P\left[\cdot\mid\mathcal{F}_t\right]\left(\cdot\right):\mathcal{F}_T\times\Omega\to\left[0,1\right]$$

with the characterizing property

$$\int_B P\left[A\mid\mathcal{F}_t\right]\left(\omega\right)P\left[\mathrm{d}\omega\right]=P\left[A\cap B\right]\qquad\left(A\in\mathcal{F}_T,\,B\in\mathcal{F}_t\right).\tag{5.5}$$

Remark 4 (The constraints in (5.2) are redundant at the final stage $t=T$). *Note that for* $A\in\mathcal{F}_t$*,* $P\left[A\mid\mathcal{F}_t\right]=\mathbb{1}_A$ *and* $\pi\left[A\times\tilde{\Omega}\mid\mathcal{F}_t\otimes\tilde{\mathcal{F}}_t\right]=\mathbb{1}_{A\times\tilde{\Omega}}$*. As obviously*

$$\mathbb{1}_{A\times\tilde{\Omega}}=\mathbb{1}_A\,\mathrm{i}$$

always holds true ($A\in\mathcal{F}_t$*) it follows that*

$$\pi\left[A\times\tilde{\Omega}\mid\mathcal{F}_t\otimes\tilde{\mathcal{F}}_t\right]=P\left[A\mid\mathcal{F}_t\right]\,\mathrm{i};$$

by analogue reasoning

$$\pi\left[\Omega \times B \mid \mathcal{F}_t \otimes \tilde{\mathcal{F}}_t\right] = \tilde{P}\left[B \mid \tilde{\mathcal{F}}_t\right] \tilde{\mathrm{i}}$$

certainly holds for $B \in \tilde{\mathcal{F}}_t$*: Whence the marginal conditions in* (5.2) *are certainly satisfied for all measures on* $\mathcal{F}_T$ $(t \leq T)$ *and thus are redundant for the final stage* $t = T$.

REMARK 5. *It follows from the previous remark that the multistage distance and the Wasserstein distance coincide, i.e.* $\mathsf{dl}_r\left(\mathbb{P}, \tilde{\mathbb{P}}\right) = \mathsf{d}_r\left(P, \tilde{P}\right)$ *for the filtrations* $\mathcal{F} = (\mathcal{F}_0, \mathcal{F}_T, \dots \mathcal{F}_T)$ *and* $\tilde{\mathcal{F}} = \left(\tilde{\mathcal{F}}_0, \tilde{\mathcal{F}}_T, \dots \tilde{\mathcal{F}}_T\right)$*. The same, however, holds true for the more general situation of filtrations* $\mathcal{F} = (\mathcal{F}_0, \dots \mathcal{F}_0, \mathcal{F}_T, \dots \mathcal{F}_T)$ *and* $\tilde{\mathcal{F}} = \left(\tilde{\mathcal{F}}_0, \dots \tilde{\mathcal{F}}_0, \tilde{\mathcal{F}}_T, \dots, \tilde{\mathcal{F}}_T\right)$ *where the information available increases all of a sudden for both at the same stage.*

LEMMA 5.2. *The multistage distance* (5.3) *is well defined, the product measure* $\pi := P \otimes \tilde{P}$ *is feasible for all conditions in* (5.2).

Proof. The product measure satisfies all above conditions:

$$\begin{aligned}
&\int_{C\times D} P\left[A \mid \mathcal{F}_t\right] \mathrm{i}\cdot \tilde{P}\left[B \mid \tilde{\mathcal{F}}_t\right] \tilde{\mathrm{i}}\,\mathrm{d}\pi = \int_{C\times D} P\left[A \mid \mathcal{F}_t\right] \mathrm{i}\cdot \tilde{P}\left[B \mid \mathcal{F}_t\right] \tilde{\mathrm{i}}\,\mathrm{d}P \otimes \tilde{P}\\
&= \int_C P\left[A \mid \mathcal{F}_t\right] \mathrm{d}P \cdot \int_D \tilde{P}\left[B \mid \tilde{\mathcal{F}}_t\right] \mathrm{d}\tilde{P}\\
&= P\left[A \cap C\right] \cdot \tilde{P}\left[B \cap D\right] = \pi\left[(A \cap C) \times (B \cap D)\right]\\
&= \pi\left[(A \times B) \cap (C \times D)\right] = \int_{C\times D} \pi\left[A \times B \mid \mathcal{F}_t \otimes \tilde{\mathcal{F}}_t\right].
\end{aligned}$$

As these equations hold true for any sets $C \in \mathcal{F}_t$ and $D \in \tilde{\mathcal{F}}_t$, and as moreover both, $P\left[A \mid \mathcal{F}_t\right] \mathrm{i}\cdot \tilde{P}\left[B \mid \mathcal{F}_t\right] \tilde{\mathrm{i}}$ and $\pi\left[A \times B \mid \mathcal{F}_t \otimes \tilde{\mathcal{F}}_t\right]$ are $\mathcal{F}_t \otimes \tilde{\mathcal{F}}_t$ measurable, it follows that they are just versions of each other, so they coincide

$$P\left[A \mid \mathcal{F}_t\right] \cdot \tilde{P}\left[B \mid \tilde{\mathcal{F}}_t\right] = \pi\left[A \times B \mid \mathcal{F}_t \otimes \tilde{\mathcal{F}}_t\right]$$

π-almost everywhere. For the particular choices $A = \Omega$ or $B = \tilde{\Omega}$ we finally get the conditions in the primal problem (5.2). $\square$

It is an immediate consequence of Hölder's inequality that continuity with respect to order 1 immediately implies continuity for other orders as well:

LEMMA 5.3 (Hölder inequality). *Suppose that* $0 < r_1 \leq r_2$*, then*

$$\mathsf{dl}_{r_1}\left(\mathbb{P}, \tilde{\mathbb{P}}\right) \leq \mathsf{dl}_{r_2}\left(\mathbb{P}, \mathbb{P}\right).$$

Proof. Observe that $\frac{1}{\frac{r_2}{r_2 - r_1}} + \frac{1}{\frac{r_2}{r_1}} = 1$. By the generalized Hölder inequality[5] thus

$$\int d^{r_1} \mathrm{d}\pi = \int 1 \cdot d^{r_1} \mathrm{d}\pi \leq \left(\int 1^{\frac{r_2}{r_2 - r_1}} \mathrm{d}\pi\right)^{\frac{r_2 - r_1}{r_2}} \cdot \left(\int d^{r_1 \frac{r_2}{r_1}} \mathrm{d}\pi\right)^{\frac{r_1}{r_2}} = \left(\int d^{r_2} \mathrm{d}\pi\right)^{\frac{r_1}{r_2}}.$$

Taking the infimum over all feasible probability measures reveals the assertion.$\square$

As $\pi = P \otimes \tilde{P}$ is feasible we may further conclude that $\mathsf{dl}_r\left(\mathbb{P}, \tilde{\mathbb{P}}\right)^r \leq \mathbb{E}_{P \otimes \tilde{P}} d^r$ for any filtrations.

[5] The generalized Hölder inequality applies for indices smaller than 1 as well, cf. [37] or [11].

Theorem 5.4. *For P and $\tilde P$ probability measures and irrespective of the filtrations $\mathfrak{F}$ and $\tilde{\mathfrak{F}}$ there are uniform lower and upper bounds*

$$\mathsf{d}_r\left(P,\tilde P\right)^r \le \mathsf{dl}_r\left(\mathbb{P},\tilde{\mathbb{P}}\right)^r \le \mathbb{E}_{P\otimes\tilde P}\, d^r.$$

Proof. The upper bound was established in Lemma 5.2. As for the lower bound notice first that $\mathcal{F}_0\otimes\tilde{\mathcal{F}}_0=\left\{\emptyset,\,\Omega\times\tilde\Omega\right\}$ is the trivial sigma algebra on $\Omega\times\tilde\Omega$.

For the trivial sigma algebra the conditional probabilities are constant functions, thus

$$P\left[A\right]\equiv P\left[A\mid\mathcal{F}_0\right]\text{ i}=\pi\left[A\times\Omega\mid\mathcal{F}_0\otimes\tilde{\mathcal{F}}_0\right]\equiv\pi\left[A\times\Omega\right],$$

so the first marginal conditions hold. As

$$\tilde P\left[B\right]\equiv\tilde P\left[B\mid\tilde{\mathcal{F}}_0\right]\ \tilde{\text{i}}=\pi\left[\Omega\times B\mid\mathcal{F}_0\otimes\tilde{\mathcal{F}}_0\right]\equiv\pi\left[\Omega\times B\right],$$

the second marginal conditions hold as well. Together they are just the marginal conditions for the Wasserstein distance d_r in (4.2) and since this constraint is contained in the constraints (5.2), it is obvious that $\mathsf{d}_r\left(P,\tilde P\right)\le\mathsf{dl}_r\left(\mathbb{P},\tilde{\mathbb{P}}\right)$. □

It is important to note that the conditions (5.2) in Definition 5.1 can be relaxed: The equations do not have to hold for all sets $A\in\mathcal{F}_T$ and $B\in\tilde{\mathcal{F}}_T$, it is sufficient to require that those conditions just hold for sets taken from the next stage. The precise statement will be of importance in the sequel and reads as follows:

Lemma 5.5 (Tower property). *In (5.2), the conditions*

$$\pi\left[A\times\tilde\Omega\mid\mathcal{F}_t\otimes\tilde{\mathcal{F}}_t\right]=P\left[A\mid\mathcal{F}_t\right]\text{ i}\qquad\left(A\in\mathcal{F}_T\right)\tag{5.6}$$

may be replaced by

$$\pi\left[A\times\tilde\Omega\mid\mathcal{F}_t\otimes\tilde{\mathcal{F}}_t\right]=P\left[A\mid\mathcal{F}_t\right]\text{ i}\qquad\left(A\in\mathcal{F}_{t+1}\right).\tag{5.7}$$

Proof. To verify this observe first that for $A\in\mathcal{F}_T$

$$\begin{aligned}&\mathbb{E}_\pi\left[\mathbb{1}_A\text{ i}\mid\mathcal{F}_t\otimes\tilde{\mathcal{F}}_t\right]=\mathbb{E}_\pi\left[\mathbb{1}_{A\times\tilde\Omega}\mid\mathcal{F}_t\otimes\tilde{\mathcal{F}}_t\right]=\pi\left[A\times\tilde\Omega\mid\mathcal{F}_t\otimes\tilde{\mathcal{F}}_t\right]\\&=P\left[A\mid\mathcal{F}_t\right]\text{ i}=\mathbb{E}_P\left[\mathbb{1}_A\mid\mathcal{F}_t\right]\text{ i},\end{aligned}$$

and by linearity thus

$$\mathbb{E}_\pi\left[\lambda\text{ i}\mid\mathcal{F}_t\otimes\tilde{\mathcal{F}}_t\right]=\mathbb{E}_P\left[\lambda\mid\mathcal{F}_t\right]\text{ i}$$

for every integrable $\lambda\lhd\mathcal{F}_T$.

Assume now that (5.7) holds true and let $A\in\mathcal{F}_T$. The assertion follows from the tower property of conditional expectation, for

$$\begin{aligned}&\pi\left[A\times\tilde\Omega\mid\mathcal{F}_t\otimes\tilde{\mathcal{F}}_t\right]=\mathbb{E}_\pi\left[\mathbb{1}_A\text{ i}\mid\mathcal{F}_t\otimes\tilde{\mathcal{F}}_t\right]\\&=\mathbb{E}_\pi\left[\mathbb{E}_\pi\left[\mathbb{1}_A\text{ i}\mid\mathcal{F}_{T-1}\otimes\tilde{\mathcal{F}}_{T-1}\right]\mid\mathcal{F}_t\otimes\tilde{\mathcal{F}}_t\right]=\mathbb{E}_\pi\left[\mathbb{E}_P\left[\mathbb{1}_A\mid\mathcal{F}_{T-1}\right]\text{ i}\mid\mathcal{F}_t\otimes\tilde{\mathcal{F}}_t\right].\end{aligned}$$

As $\mathbb{E}_P\left[\mathbb{1}_A\mid\mathcal{F}_{T-1}\right]\lhd\mathcal{F}_{T-1}$ the steps above may be repeated to give

$$\begin{aligned}&\pi\left[A\times\tilde\Omega\mid\mathcal{F}_t\otimes\tilde{\mathcal{F}}_t\right]=\mathbb{E}_\pi\left[\mathbb{E}_P\left[\mathbb{1}_A\mid\mathcal{F}_{T-1}\right]\text{ i}\mid\mathcal{F}_t\otimes\tilde{\mathcal{F}}_t\right]\\&=\mathbb{E}_\pi\left[\mathbb{E}_\pi\left[\mathbb{E}_P\left[\mathbb{1}_A\mid\mathcal{F}_{T-1}\right]\text{ i}\mid\mathcal{F}_{T-2}\otimes\tilde{\mathcal{F}}_{T-2}\right]\mid\mathcal{F}_t\otimes\tilde{\mathcal{F}}_t\right]\\&=\mathbb{E}_\pi\left[\mathbb{E}_P\left[\mathbb{E}_P\left[\mathbb{1}_A\mid\mathcal{F}_{T-1}\right]\mid\mathcal{F}_{T-2}\right]\text{ i}\mid\mathcal{F}_t\otimes\tilde{\mathcal{F}}_t\right]=\mathbb{E}_\pi\left[\mathbb{E}_P\left[\mathbb{1}_A\mid\mathcal{F}_{T-2}\right]\text{ i}\mid\mathcal{F}_t\otimes\tilde{\mathcal{F}}_t\right],\end{aligned}$$

and a repeated application gives further

$$\begin{aligned}
&\pi\left[A\times\tilde{\Omega}\mid\mathcal{F}_t\otimes\tilde{\mathcal{F}}_t\right]=\mathbb{E}_\pi\left[\mathbb{E}_P\left[\mathbb{1}_A\mid\mathcal{F}_{T-2}\right]\mathrm{i}\mid\mathcal{F}_t\otimes\tilde{\mathcal{F}}_t\right]\\
&=\mathbb{E}_\pi\left[\mathbb{E}_P\left[\mathbb{1}_A\mid\mathcal{F}_{T-3}\right]\mathrm{i}\mid\mathcal{F}_t\otimes\tilde{\mathcal{F}}_t\right]=\dots\\
&=\mathbb{E}_\pi\left[\mathbb{E}_P\left[\mathbb{1}_A\mid\mathcal{F}_{t}\right]\mathrm{i}\mid\mathcal{F}_t\otimes\tilde{\mathcal{F}}_t\right].
\end{aligned}$$

Finally, as $\mathbb{E}_P\left[\mathbb{1}_A\mid\mathcal{F}_t\right]\mathrm{i}$ is $\mathcal{F}_t$ measurable,

$$\begin{aligned}
&\pi\left[A\times\tilde{\Omega}\mid\mathcal{F}_t\otimes\tilde{\mathcal{F}}_t\right]=\mathbb{E}_\pi\left[\mathbb{E}_P\left[\mathbb{1}_A\mid\mathcal{F}_{t}\right]\mathrm{i}\mid\mathcal{F}_t\otimes\tilde{\mathcal{F}}_t\right]\\
&=\mathbb{E}_P\left[\mathbb{E}_P\left[\mathbb{1}_A\mid\mathcal{F}_{t}\right]\mid\mathcal{F}_t\right]\mathrm{i}=\mathbb{E}_P\left[\mathbb{1}_A\mid\mathcal{F}_{t}\right]\mathrm{i}=P\left[A\mid\mathcal{F}_t\right]\mathrm{i},
\end{aligned}$$

which is the general condition (5.6). □

6. Relation to Multistage Stochastic Optimization. As already addressed in the introduction the multistage distance is a suitable distance for multistage stochastic optimization problems. To elaborate the relation consider the value function $v(\mathbb{P})$ of stochastic optimization problem

$$\begin{aligned}
v\left(\mathbb{P}\right)&=\inf\left\{\mathbb{E}_P H\left(\xi,x\right):x\in\mathbb{X},\ x\lhd\mathfrak{F}\right\}\\
&=\inf\left\{\int H\left(\xi,x\right)\mathrm{d}P\colon x\in\mathbb{X},\ x\lhd\mathfrak{F}\right\}
\end{aligned}\tag{6.1}$$

of the expectation-maximization type.

The following theorem is the main theorem to bound stochastic optimization problems by the nested distance, it links smoothness properties of the loss function H with smoothness of the value function v with respect to the multistage distance.

Theorem 6.1 (Lipschitz property of the value function). *Let $\mathbb{P}$, $\tilde{\mathbb{P}}$ be two nested distributions. Assume that $\mathbb{X}$ is convex, and the profit function H is convex in x for any ξ fixed,*

$$H\left(\xi,\left(1-\lambda\right)x_0+\lambda x_1\right)\leq\left(1-\lambda\right)H\left(\xi,x_0\right)+\lambda H\left(\xi,x_1\right).$$

Moreover let H be uniformly Hölder continuous ($\beta\leq1$) with constant L_β, that is

$$\left|H\left(\xi,x\right)-H\left(\tilde{\xi},x\right)\right|\leq L_\beta\cdot\left(\sum_{t\in\mathbf{T}}d_t\left(\xi_t,\tilde{\xi}_t\right)\right)^\beta$$

for all $x\in\mathbb{X}$.

Then the value function v (6.1) inherits the Hölder constant with respect to the multistage distance, that is

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

for any $r\geq1$.[6] In addition we have the following corollary.

Corollary 1 (Best possible bound). *Assuming that the distance may be represented by a norm, the Lipschitz constant for the situation $\beta=1$ cannot be improved.*

Proof. (Proof of Theorem 6.1) Let $x\lhd\mathfrak{F}$ be a decision vector for problem (6.1) and nested distribution $\mathbb{P}$ and let π be a bivariate probability measure on $\mathcal{F}_T\otimes\tilde{\mathcal{F}}_T$ which satisfies the conditions (5.2), i.e. which is an optimal transportation measure.

[6] For $\beta=1$ Hölder continuity is just Lipschitz continuity.

Note that x is a vector of non-anticipative decisions for any $t \in \mathbf{T}$, and whence

$$\begin{aligned}\mathbb{E}_P H\left(\xi, x\right) &= \int H\left(\xi\left(\omega\right), x\left(\omega\right)\right) P\left[\mathrm{d}\omega\right] \\ &= \int H\left(\xi\left(\omega\right), x\left(\omega\right)\right) \pi\left[\mathrm{d}\omega, \mathrm{d}\tilde{\omega}\right] = \mathbb{E}_\pi H\left(\xi\,\mathrm{i}, x\,\mathrm{i}\right). \qquad (6.2)\end{aligned}$$

Define next the new decision function

$$\tilde{x} : = \mathbb{E}_\pi\left[x\,\mathrm{i} \mid \tilde{\mathrm{i}}\right],$$

which has the desired measurability, that is $\tilde{x} \lhd \tilde{\mathcal{F}}$ due its definition[7] and the conditions (5.2) imposed on π. With this choice, by convexity,

$$H\left(\tilde{\xi}\left(\tilde{\omega}\right), \tilde{x}\left(\tilde{\omega}\right)\right) = H\left(\tilde{\xi}\left(\tilde{\omega}\right), \mathbb{E}_\pi\left[x\,\mathrm{i} \mid \tilde{\mathrm{i}}\right]\left(\tilde{\omega}\right)\right) \leq \mathbb{E}_\pi\left[H\left(\tilde{\xi}\left(\tilde{\omega}\right), x\,\mathrm{i}\right) \mid \tilde{\mathrm{i}}\right]\left(\tilde{\omega}\right)$$

by Jensen's inequality: Again, Jensen's inequality applies for all $t \in \mathbf{T}$ jointly due to the joint restrictions (5.2) on π. Integrating with respect to $\tilde{P}$ one obtains

$$\begin{aligned}&\mathbb{E}_{\tilde{P}} H\left(\tilde{\xi}, \tilde{x}\right) = \int_{\tilde{\Omega}} H\left(\tilde{\xi}\left(\tilde{\omega}\right), \tilde{x}\left(\tilde{\omega}\right)\right) \tilde{P}\left[\mathrm{d}\tilde{\omega}\right] \leq \int_{\tilde{\Omega}} \mathbb{E}_\pi\left[H\left(\tilde{\xi}\,\tilde{\mathrm{i}}, x\,\mathrm{i}\right) \mid \tilde{\mathrm{i}}\right] \tilde{\mathrm{i}}\,\mathrm{d}\tilde{P} \\ &= \mathbb{E}_\pi\left[\mathbb{E}_\pi\left[H\left(\tilde{\xi}\,\tilde{\mathrm{i}}, x\,\mathrm{i}\right) \mid \tilde{\mathrm{i}}\right] \tilde{\mathrm{i}}\right] = \mathbb{E}_\pi H\left(\tilde{\xi}\,\tilde{\mathrm{i}}, x\,\mathrm{i}\right),\end{aligned}$$

and together with (6.2) it follows that

$$\begin{aligned}&\mathbb{E}_{\tilde{P}} H\left(\tilde{\xi}, \tilde{x}\right) - \mathbb{E}_P H\left(\xi, x\right) \leq \mathbb{E}_\pi\left[H\left(\tilde{\xi}\,\tilde{\mathrm{i}}, x\,\mathrm{i}\right)\right] - \mathbb{E}_\pi\left[H\left(\xi\,\mathrm{i}, x\,\mathrm{i}\right)\right] \\ &= \mathbb{E}_\pi\left[H\left(\tilde{\xi}\,\tilde{\mathrm{i}}, x\,\mathrm{i}\right) - H\left(\xi\,\mathrm{i}, x\,\mathrm{i}\right)\right] \leq L_\beta \cdot \mathbb{E}_\pi d\left(\mathrm{i}, \tilde{\mathrm{i}}\right)^\beta.\end{aligned}$$

Now let x be an ε-optimal decision for $v\left(\mathbb{P}\right)$, that is

$$\int H\left(\xi, x\right) \mathrm{d}P < v\left(\mathbb{P}\right) + \varepsilon.$$

It follows that

$$\begin{aligned}\mathbb{E}_{\tilde{P}} H\left(\tilde{\xi}, \tilde{x}\right) - v\left(\mathbb{P}\right) <& \mathbb{E}_{\tilde{P}} H\left(\tilde{\xi}, \tilde{x}\right) - \mathbb{E}_{\mathbb{P}} H\left(\xi, x\right) + \varepsilon \\ \leq& \varepsilon + L_\beta \cdot \mathbb{E}_\pi d^\beta\end{aligned}$$

and whence

$$v\left(\tilde{\mathbb{P}}\right) - v\left(\mathbb{P}\right) < \varepsilon + L_\beta \cdot \mathbb{E}_\pi d^\beta.$$

Letting $\varepsilon \to 0$, taking the infimum over all π and interchanging the roles of $\mathbb{P}$ and $\tilde{\mathbb{P}}$ gives that

$$\left|v\left(\mathbb{P}\right) - v\left(\tilde{\mathbb{P}}\right)\right| \leq L_\beta \cdot \mathsf{dl}_\beta\left(\mathbb{P}, \tilde{\mathbb{P}}\right)^\beta.$$

The statement for general index $r \geq 1 \geq \beta$ finally follows by applying Lemma 5.3. □

[7] $\tilde{x}$ may be defined alternatively for any component separately as $\tilde{x}_t\left(\tilde{\omega}\right) := \int_\Omega x_t\left(\omega\right) \pi\left[\mathrm{d}\omega \mid \tilde{\omega}\right]$; the conditional probabilities are available by disintegration (cf. [7], [8] or [2]) and they satisfy $\pi\left[A \times B\right] = \int_B \pi\left[A|\tilde{\omega}\right] \pi\left[\tilde{\mathrm{i}}^{-1}\left(\mathrm{d}\tilde{\omega}\right)\right] = \int_B \pi\left[A|\tilde{\omega}\right] \tilde{P}\left(\mathrm{d}\tilde{\omega}\right)$.

Proof. (Proof of the Corollary) Let $\mathbb{X}$ be the convex set of Markov kernels $x\left[B|\omega,\omega'\right]$ $(\omega,\omega'\in\Omega,\ B\in\tilde{\mathcal{F}}_T)$ such that

$$\pi\left[A\times B\right]:=\int_A x\left[B|\omega,\omega'\right]\mathbb{P}\left[\mathrm{d}\omega'\right]=\int_{A\times B}x\left[\mathrm{d}\omega''|\omega,\omega'\right]\mathbb{P}\left[\mathrm{d}\omega'\right]\tag{6.3}$$

satisfies the marginal conditions (5.2) for $\tilde{\mathbb{P}}$ and all ω, and abbreviate $x\left[B|\omega'\right]:=\int x\left[B|\omega,\omega'\right]\mathbb{P}\left[\mathrm{d}\omega\right]$. Define

$$H\left(\omega,x\right):=\int\left\|\omega-\omega''\right\|x\left[\mathrm{d}\omega''|\omega'\right]\mathbb{P}\left[\mathrm{d}\omega'\right]$$

and observe that

$$\begin{aligned}H\left(\omega,x\right)&=\int\left\|\omega-\omega''\right\|x\left[\mathrm{d}\omega''|\omega,\omega'\right]\mathbb{P}\left[\mathrm{d}\omega'\right]\\&=\int\left\|\omega-\omega''\right\|x\left[\mathrm{d}\omega''|\tilde{\omega},\omega'\right]\mathbb{P}\left[\mathrm{d}\omega'\right]\mathbb{P}\left[\mathrm{d}\tilde{\omega}\right];\end{aligned}$$

moreover H is Lipschitz-1, as

$$\begin{aligned}H\left(\omega_1,\pi\right)-H\left(\omega_2,\pi\right)&=\int\left\|\omega_1-\omega''\right\|-\left\|\omega_2-\omega''\right\|x\left[\mathrm{d}\omega''|\omega'\right]\mathbb{P}\left[\mathrm{d}\omega'\right]\\&\le\int\left\|\omega_1-\omega_2\right\|x\left[\mathrm{d}\omega''|\omega'\right]\mathbb{P}\left[\omega'\right]=\left\|\omega_1-\omega_2\right\|\end{aligned}$$

by the reverse triangle inequality and the fact that the role of ω_1 and ω_2 may be interchanged. Then consider the value function

$$v\left(\mathbb{P}\right):=\inf_{x\in\mathbb{X}}\mathbb{E}_{\mathbb{P}}H\left(\cdot,x\right)=\inf_{x\in\mathbb{X}}\int H\left(\omega,x\right)\mathbb{P}\left[\mathrm{d}\omega\right],$$

where the infimum is among all Markov kernels $x\in\mathbb{X}$ with marginal condition $\pi_2=\tilde{\mathbb{P}}$ as above. With this choice

$$\begin{aligned}v\left(\mathbb{P}\right)&=\inf_{x\in\mathbb{X}}\mathbb{E}_{\mathbb{P}}H\left(\cdot,x\right)=\inf_{x\in\mathbb{X}}\int\int\left\|\omega-\omega''\right\|x\left[\mathrm{d}\omega''|\omega,\omega'\right]\mathbb{P}\left[\mathrm{d}\omega'\right]\mathbb{P}\left[\mathrm{d}\omega\right]\\&=\inf_{x\in\mathbb{X}}\int\int\left\|\omega-\omega''\right\|x\left[\mathrm{d}\omega''|\omega\right]\mathbb{P}\left[\mathrm{d}\omega\right]=\inf_{\pi}\int\int\left\|\omega-\omega''\right\|\pi\left[\mathrm{d}\omega'',\mathrm{d}\omega\right]=\mathsf{dl}_1\left(\mathbb{P},\tilde{\mathbb{P}}\right)\end{aligned}$$

by the marginal conditions imposed on $\mathbb{X}$. Moreover

$$v\left(\tilde{\mathbb{P}}\right)=\inf_{x\in\mathbb{X}}\mathbb{E}_{\tilde{\mathbb{P}}}H\left(\cdot,x\right)=\inf_{x\in\mathbb{X}}\int\int\left\|\omega-\omega''\right\|x\left[\mathrm{d}\omega''|\omega,\omega'\right]\mathbb{P}\left[\mathrm{d}\omega'\right]\tilde{\mathbb{P}}\left[\mathrm{d}\omega\right].$$

Employing the Dirac-measure $x\left[A|\omega,\omega'\right]:=\mathbb{1}_A\left(\omega\right)$ we find that

$$\int\left\|\omega-\omega''\right\|x\left[\mathrm{d}\omega''|\omega',\omega\right]=0,$$

and thus

$$v\left(\tilde{\mathbb{P}}\right)=0.$$

Whence

$$v\left(\mathbb{P}\right)-v\left(\tilde{\mathbb{P}}\right)=\mathsf{dl}_1\left(\mathbb{P},\tilde{\mathbb{P}}\right),$$

and the Lipschitz constant cannot be improved. As for a general Lipschitz constant L other than 1 consider the function $L\cdot H$ instead, which completes the proof. □

7. The Dual Representation of the Multistage Distance. As the problem (4.2) to compute the Kantorovich or Wasserstein distance is formulated as an optimization problem it has a dual formulation. Its dual is well-known and well investigated and may be stated as follows (cf. [2]):

Theorem 7.1 (Duality formula). *The minimum of the Kantorovich Problem* (4.2) *equals*

$$\sup_{\mu,\tilde{\mu}} \mathbb{E}_P \mu + \mathbb{E}_{\tilde{P}} \tilde{\mu}$$

where the supremum runs among all pairs $(\mu, \tilde{\mu})$ *such that*

$$\mu(\omega) + \tilde{\mu}(\tilde{\omega}) \leq d(\omega, \tilde{\omega})^r .$$

For the optimal measure π*, in addition,*

$$\mu(\omega) + \tilde{\mu}(\tilde{\omega}) = d(\omega, \tilde{\omega})^r \qquad \pi\text{-}a.e.\ \ in\ \Omega \times \tilde{\Omega}. \tag{7.1}$$

The duality formula in Theorem 7.1 can be given as optimal value for the alternative representation

$$\begin{array}{ll} \text{maximize (in } M_0,\ \lambda,\ \tilde{\lambda}) & M_0 \\ \text{subject to} & M_0 + \lambda(\omega) + \tilde{\lambda}(\tilde{\omega}) \leq d(\omega, \tilde{\omega})^r , \\ & \mathbb{E}_{\mathbb{P}} \lambda = 0 \text{ and } \mathbb{E}_{\tilde{\mathbb{P}}} \tilde{\lambda} = 0; \end{array}$$

to accept the latter statement just shift the dual functions by their respective means and choose $\lambda := \mu - \mathbb{E}_{\mathbb{P}} \mu$, $\tilde{\lambda} := \tilde{\mu} - \mathbb{E}_{\tilde{\mathbb{P}}} \tilde{\mu}$ and $M_0 := \mathbb{E}_{\mathbb{P}} \mu + \mathbb{E}_{\tilde{\mathbb{P}}} \tilde{\mu}$.

It is natural to ask for the dual of the problem of the multistage distance, that is to say the dual of problem (5.2). The dual allows a characterization as well, which is the content of the next theorem, its formulation is in line with the preceding remark.

Denote the set of $\mathcal{F}_T \otimes \tilde{\mathcal{F}}_T$ measurable functions as $\mathbb{L}^0\left(\mathcal{F}_T \otimes \tilde{\mathcal{F}}_T\right)$ and define the projections

$$\begin{aligned} \mathrm{Pr}_t \colon \mathbb{L}^0\left(\mathcal{F}_T \otimes \tilde{\mathcal{F}}_T\right) &\to \mathbb{L}^0\left(\mathcal{F}_t \otimes \tilde{\mathcal{F}}_T\right) \\ \mu\,\mathrm{i} \cdot \tilde{\mu}\tilde{\mathrm{i}} &\mapsto \mathbb{E}_{\mathbb{P}}\left(\mu | \mathcal{F}_t\right) \mathrm{i} \cdot \tilde{\mu}\tilde{\mathrm{i}} \end{aligned}$$

and

$$\begin{aligned} \tilde{\mathrm{P}}\mathrm{r}_t \colon \mathbb{L}^0\left(\mathcal{F}_T \otimes \tilde{\mathcal{F}}_T\right) &\to \mathbb{L}^0\left(\mathcal{F}_T \otimes \tilde{\mathcal{F}}_t\right) \\ \mu\,\mathrm{i} \cdot \tilde{\mu}\tilde{\mathrm{i}} &\mapsto \mu\,\mathrm{i} \cdot \mathbb{E}_{\tilde{\mathbb{P}}}\left(\tilde{\mu} | \tilde{\mathcal{F}}_t\right) \tilde{\mathrm{i}}; \end{aligned}$$

as the functions $\mathbb{1}_A\,\mathrm{i} \cdot \mathbb{1}_B \tilde{\mathrm{i}}$ form a basis of $\mathbb{L}^0$ the projections Pr_t and $\tilde{\mathrm{P}}\mathrm{r}_t$ are well defined.

Theorem 7.2 (Duality for the multistage distance). *The minimum of the multistage problem* (5.2) *equals the supremum of all numbers* M_0 *such that*

$$M_T(\omega, \tilde{\omega}) \leq d(\omega, \tilde{\omega})^r \qquad (\omega, \tilde{\omega}) \in \Omega \times \tilde{\Omega},$$

where M_t *is a* $\mathbb{R}-$*valued process on* $\Omega \times \tilde{\Omega}$ *of the form*

$$M_t := M_0 + \sum_{s=1}^{t} \left(\lambda_s + \tilde{\lambda}_s\right)$$

and the measurable functions $\lambda_t \lhd \mathcal{F}_t \otimes \tilde{\mathcal{F}}_{t-1}$ *and* $\tilde{\lambda}_t \lhd \mathcal{F}_{t-1} \otimes \tilde{\mathcal{F}}_t$ *are chosen such that* $\operatorname{Pr}_{t-1} \lambda_t = 0$ *and* $\tilde{\operatorname{Pr}}_{t-1} \tilde{\lambda}_t = 0$.

For the optimal measure π *of the primal problem and the optimal process* M_t *of the dual problem, in addition,*

$$M_t = \mathbb{E}_\pi \left[d^r \mid \mathcal{F}_t \otimes \tilde{\mathcal{F}}_t \right] \quad \pi\text{-}a.e.\ in\ \Omega \times \tilde{\Omega}$$

and M_t *is a* $\pi-$*martingale for the filtration* $\left(\mathcal{F}_t \otimes \tilde{\mathcal{F}}_t \right)_{t \in \mathbf{T}}$ *such that*

$$\mathsf{dl}_r \left(\mathbb{P}, \tilde{\mathbb{P}} \right)^r = M_0 = \mathbb{E}_\pi M_t$$

for all stages $t \in \mathbf{T}$.

Remark 6. *It should be noted that particularly equality is attained* π*-a.e. at the final stage* T*, i.e.*

$$M_T = d^r \quad \pi\text{-}a.e.\ in\ \Omega \times \tilde{\Omega}$$

holds true almost everywhere. This is in line with (7.1) *for the two stage problem,* $T = 1$*.* In order to prove the latter theorem let us start with the following observation.

Proposition 7.3 (Encoding). *The measure* π *satisfies the conditions*

$$\pi \left[A \times \tilde{\Omega} \mid \mathcal{F}_t \otimes \tilde{\mathcal{F}}_t \right] = \mathbb{P} \left[A \mid \mathcal{F}_t \right] \mathrm{i} \tag{7.2}$$

for all $A \in \mathcal{F}_T$ *if and only if*

$$\mathbb{E}_\pi \lambda = \mathbb{E}_\pi \operatorname{Pr}_t \lambda \tag{7.3}$$

holds for all integrable functions $\lambda \lhd \mathcal{F}_T \otimes \tilde{\mathcal{F}}_t$.

Proof. To prove assertion (7.3) note first that it is enough to prove the claim just for functions $\lambda = \mu\, \mathrm{i} \cdot \tilde{\mu} \tilde{\mathrm{i}}$, as these functions form a basis of all $\mathcal{F}_T \otimes \tilde{\mathcal{F}}_t$-integrable functions whenever $\mu \lhd \mathcal{F}_T$ and $\tilde{\mu} \lhd \tilde{\mathcal{F}}_t$.

For the function $\mathbb{1}_A$ $(A \in \mathcal{F}_T)$

$$\begin{aligned} &\mathbb{E}_P \left[\mathbb{1}_A \mid \mathcal{F}_t \right] \mathrm{i} = P \left[A \mid \mathcal{F}_t \right] \mathrm{i} = \pi \left[A \times \tilde{\Omega} \mid \mathcal{F}_t \otimes \tilde{\mathcal{F}}_t \right] \\ &= \mathbb{E}_\pi \left[\mathbb{1}_{A \times \tilde{\Omega}} \mid \mathcal{F}_t \otimes \tilde{\mathcal{F}}_t \right] = \mathbb{E}_\pi \left[\mathbb{1}_A\ \mathrm{i} \mid \mathcal{F}_t \otimes \tilde{\mathcal{F}}_t \right], \end{aligned}$$

and it follows by linearity that

$$\mathbb{E}_P \left[\mu \mid \mathcal{F}_t \right] \mathrm{i} = \mathbb{E}_\pi \left[\mu\ \mathrm{i} \mid \mathcal{F}_t \otimes \tilde{\mathcal{F}}_t \right]$$

π a.e. for any integrable $\mu \lhd \mathcal{F}_T$. The left hand side, multiplied by $\tilde{\mu}\, \tilde{\mathrm{i}}$, reads

$$\mathbb{E}_P \left[\mu \mid \mathcal{F}_t \right] \mathrm{i} \cdot \tilde{\mu}\, \tilde{\mathrm{i}};$$

the right hand side, multiplied by the same quantity, gives

$$\tilde{\mu} \tilde{\mathrm{i}} \cdot \mathbb{E}_\pi \left[\mu\ \mathrm{i} \mid \mathcal{F}_t \otimes \tilde{\mathcal{F}}_t \right] = \mathbb{E}_\pi \left[\mu\ \ \mathrm{i} \cdot \tilde{\mu}\, \tilde{\mathrm{i}} \mid \mathcal{F}_t \otimes \tilde{\mathcal{F}}_t \right]$$

as $\tilde{\mu}$ is $\tilde{\mathcal{F}}_t$-measurable and thus $\tilde{\mu}\, \tilde{\mathrm{i}} \lhd \mathcal{F}_t \otimes \tilde{\mathcal{F}}_t$, that is

$$\mathbb{E}_P \left[\mu \mid \mathcal{F}_t \right] \mathrm{i} \cdot \tilde{\mu}\, \tilde{\mathrm{i}} = \mathbb{E}_\pi \left[\mu\ \mathrm{i} \cdot \tilde{\mu}\, \tilde{\mathrm{i}} \mid \mathcal{F}_t \otimes \tilde{\mathcal{F}}_t \right]. \tag{7.4}$$

Taking expectation for both with respect to π whence gives that

$$\mathbb{E}_\pi \operatorname{Pr}_t \left(\mu \mathrm{~i} \cdot \tilde{\mu} \tilde{\mathrm{i}}\right) = \mathbb{E}_\pi \mathbb{E}_P \left[\mu \mid \mathcal{F}_t\right] \mathrm{i} \cdot \tilde{\mu} \tilde{\mathrm{i}} = \mathbb{E}_\pi \mathbb{E}_\pi \left[\mu \mathrm{~i} \cdot \tilde{\mu} \tilde{\mathrm{i}} \mid \mathcal{F}_t \otimes \tilde{\mathcal{F}}_t\right] = \mathbb{E}_\pi \left[\mu \mathrm{~i} \cdot \tilde{\mu} \tilde{\mathrm{i}}\right],$$

which is the desired assertion. To prove the converse we need to show (7.2), i. e. that

$$\pi\left[A \times \tilde{\Omega} \mid \mathcal{F}_t \otimes \tilde{\mathcal{F}}_t\right] = P\left[A \mid \mathcal{F}_t\right] \mathrm{i}$$

holds true for every set $A \in \mathcal{F}_T$. As both, the left hand side and the right hand side are $\mathcal{F}_t \otimes \tilde{\mathcal{F}}_t$-measurable densities (random variables) it is sufficient to show that

$$\int_{C \times D} \pi\left[A \times \tilde{\Omega} \mid \mathcal{F}_t \otimes \tilde{\mathcal{F}}_t\right] \mathrm{d}\pi = \int_{C \times D} P\left[A \mid \mathcal{F}_t\right] \mathrm{i}\, \mathrm{d}\pi \tag{7.5}$$

for all sets $C \in \mathcal{F}_t$ and $D \in \mathcal{G}_t$.

To this end observe first that

$$\begin{aligned}
&\int_{C \times D} \pi\left[A \times \tilde{\Omega} \mid \mathcal{F}_t \otimes \tilde{\mathcal{F}}_t\right] \mathrm{d}\pi = \mathbb{E}_\pi\left[\mathbb{1}_{C \times D} \cdot \mathbb{E}_\pi\left[\mathbb{1}_{A \times \tilde{\Omega}} \mid \mathcal{F}_t \otimes \tilde{\mathcal{F}}_t\right]\right] \\
&= \mathbb{E}_\pi\left[\mathbb{E}_\pi\left[\mathbb{1}_{C \times D} \cdot \mathbb{1}_{A \times \tilde{\Omega}} \mid \mathcal{F}_t \otimes \tilde{\mathcal{F}}_t\right]\right] = \mathbb{E}_\pi\left[\mathbb{1}_{(C \times D) \cap \left(A \times \tilde{\Omega}\right)}\right] = \pi\left[(A \cap C) \times D\right];
\end{aligned}$$

secondly

$$\begin{aligned}
&\int_{C \times D} \mathbb{P}\left[A \mid \mathcal{F}_t\right] \mathrm{i}\, \mathrm{d}\pi = \mathbb{E}_\pi\left[\mathbb{1}_{C \times D} \cdot \mathbb{E}_P\left[\mathbb{1}_A \mid \mathcal{F}_t\right] \mathrm{i}\right] \\
&= \mathbb{E}_\pi\left[\mathbb{1}_C \mathrm{~i} \cdot \mathbb{1}_D \tilde{\mathrm{i}} \cdot \mathbb{E}_P\left[\mathbb{1}_A \mid \mathcal{F}_t\right] \mathrm{i}\right] = \mathbb{E}_\pi\left[\mathbb{E}_P\left[\mathbb{1}_C \cdot \mathbb{1}_A \mid \mathcal{F}_t\right] \mathrm{i} \cdot \mathbb{1}_D \tilde{\mathrm{i}}\right] \\
&= \mathbb{E}_\pi\left[\mathbb{E}_P\left[\mathbb{1}_{C \cap A} \mid \mathcal{F}_t\right] \mathrm{i} \cdot \mathbb{1}_D \tilde{\mathrm{i}}\right] = \mathbb{E}_\pi \operatorname{Pr}_t\left(\mathbb{1}_{C \cap A} \mathrm{~i} \cdot \mathbb{1}_D \tilde{\mathrm{i}}\right)
\end{aligned}$$

and by the assertion (7.3) thus

$$\int_{C \times D} P\left[A \mid \mathcal{F}_t\right] \mathrm{i}\, \mathrm{d}\pi = \mathbb{E}_\pi\left[\mathbb{1}_{A \cap C} \mathrm{~i} \cdot \mathbb{1}_D \tilde{\mathrm{i}}\right] = \mathbb{E}_\pi\left[\mathbb{1}_{(A \cap C) \times D}\right] = \pi\left[(A \cap C) \times D\right].$$

The quantities in (7.5) thus coincide and we conclude that (7.2) holds true, which is the desired assertion. □

The proof of the dual characterization can be arranged as follows.

Proof. (Proof of the Duality Theorem 7.2) Using the observation from Lemma 5.5 and the latter Proposition 7.3 to encode the primal conditions (5.2) in the Lagrangian, the primal problem (5.2) rewrites

$$\begin{aligned}
\inf_{\pi \geq 0} \sup_{M_0, \mu_t, \tilde{\mu}_t} & \int d^r \mathrm{d}\pi + M_0 \cdot \left(1 - \mathbb{E}_\pi \mathbb{1}\right) \\
& - \sum_{s=0}^{T-1} \left(\mathbb{E}_\pi \mu_{s+1} - \mathbb{E}_\pi \operatorname{Pr}_s \mu_{s+1}\right) \\
& - \sum_{s=0}^{T-1} \left(\mathbb{E}_\pi \tilde{\mu}_{s+1} - \mathbb{E}_\pi \tilde{\operatorname{Pr}}_s \tilde{\mu}_{s+1}\right),
\end{aligned}$$

where the inf is among all *positive* measures $\pi \geq 0$ (so not only probability measures), and the sup among numbers M_0 and functions $\mu_t \lhd \mathcal{F}_t \otimes \tilde{\mathcal{F}}_{t-1}$ and $\tilde{\mu}_t \lhd \mathcal{F}_{t-1} \otimes \tilde{\mathcal{F}}_t$.

According to Sion's mini-max theorem (cf. [33]; the Lagrangian is linear in π and linear in μ_s(resp.), thus convex and concave (resp.)) this saddle point has the same objective value as

$$\sup_{M_0;\mu_t,\tilde{\mu}_t}\ \inf_{\pi\geq 0}\quad M_0+\mathbb{E}_\pi\left[\begin{array}{l} d^r-M_0\cdot\mathbb{1}\\ \quad-\sum_{s=0}^{T-1}\mu_{s+1}-\Pr_s\mu_{s+1}\\ \quad-\sum_{s=0}^{T-1}\tilde{\mu}_{s+1}-\tilde{\Pr}_s\tilde{\mu}_{s+1}\end{array}\right]. \tag{7.6}$$

Now notice that the $\inf_{\pi\geq 0}$ certainly is $-\infty$ unless the integrand is positive for every measure $\pi\geq 0$, which means that

$$M_0+\sum_{s=0}^{T-1}\mu_{s+1}-\Pr_s\mu_{s+1}+\sum_{s=0}^{T-1}\tilde{\mu}_{s+1}-\tilde{\Pr}_s\tilde{\mu}_{s+1}\leq d^r$$

has to hold. For a positive integrand, however, the $\inf_{\pi\geq 0}$ over all expectations $\mathbb{E}_\pi$ in (7.6) is 0. We thus get rid of the primal variable π and the problem to be solved reads

$$\begin{array}{lll} \text{maximize} & (\text{in } M_0,\mu_t,\tilde{\mu}_t) & M_0\\ \text{subject to} & & M_0+\sum_{s=1}^{T}\left(\mu_s-\Pr_{s-1}\mu_s\right)+\left(\tilde{\mu}_s-\tilde{\Pr}_{s-1}\tilde{\mu}_s\right)\leq d^r\\ & & \mu_t\lhd\mathcal{F}_t\otimes\tilde{\mathcal{F}}_{t-1},\tilde{\mu}_t\lhd\mathcal{F}_{t-1}\otimes\tilde{\mathcal{F}}_t \end{array}$$

Now just define $\lambda_s:=\mu_s-\Pr_{s-1}\mu_s$ and $\tilde{\lambda}_s:=\tilde{\mu}_s-\tilde{\Pr}_{s-1}\tilde{\mu}_s$, so that the latter problem rewrites

$$\begin{array}{ll} \text{maximize} & M_0\\ \text{subject to} & M_0+\sum_{s=1}^{T}\left(\lambda_s+\tilde{\lambda}_s\right)\leq d^r,\\ & \lambda_t\lhd\mathcal{F}_t\otimes\tilde{\mathcal{F}}_{t-1},\tilde{\lambda}_t\lhd\mathcal{F}_{t-1}\otimes\tilde{\mathcal{F}}_t,\\ & \Pr_{t-1}\lambda_t=0,\ \tilde{\Pr}_{t-1}\tilde{\lambda}_t=0, \end{array}$$

which is the desired assertion of the Theorem.

To prove the martingale assertion for the process

$$M_t=M_0+\sum_{s=1}^{t}\left(\lambda_s+\tilde{\lambda}_s\right)$$

observe that

$$\begin{aligned}\mathbb{E}_\pi\left[M_T\mid\mathcal{F}_t\otimes\tilde{\mathcal{F}}_t\right]&=\mathbb{E}_\pi\left[M_0+\sum_{s=1}^{T}\left(\lambda_s+\tilde{\lambda}_s\right)\mid\mathcal{F}_t\otimes\tilde{\mathcal{F}}_t\right]\\ &=M_0+\sum_{s=1}^{t}\left(\lambda_s+\tilde{\lambda}_s\right),\end{aligned}$$

because for $s>t$ by (7.4), and using again the fact that the functions $\mu_s\ \mathrm{i}\cdot\tilde{\mu}_{s-1}\,\tilde{\mathrm{i}}$ form a basis of $\mathbb{L}^0\left(\mathcal{F}_s\otimes\tilde{\mathcal{F}}_{s-1}\right)$

$$\begin{aligned}\mathbb{E}_\pi\left[\mu_s\ \mathrm{i}\cdot\tilde{\mu}_{s-1}\,\tilde{\mathrm{i}}\mid\mathcal{F}_t\otimes\tilde{\mathcal{F}}_t\right]&=\mathbb{E}_\pi\left[\mathbb{E}_P\left[\mu_s\mid\mathcal{F}_t\right]\mathrm{i}\cdot\tilde{\mu}_{s-1}\,\tilde{\mathrm{i}}\mid\mathcal{F}_t\otimes\tilde{\mathcal{F}}_t\right]\\ &=\mathbb{E}_\pi\left[\mathbb{E}_P\left[\mathbb{E}_P\left[\mu_s\mid\mathcal{F}_{s-1}\right]\mid\mathcal{F}_t\right]\mathrm{i}\cdot\tilde{\mu}_{s-1}\,\tilde{\mathrm{i}}\mid\mathcal{F}_t\otimes\tilde{\mathcal{F}}_t\right]\\ &=\mathbb{E}_\pi\left[\mathbb{E}_P\left[0\mid\mathcal{F}_t\right]\mathrm{i}\cdot\tilde{\mu}_{s-1}\,\tilde{\mathrm{i}}\mid\mathcal{F}_t\otimes\tilde{\mathcal{F}}_t\right]=0;\end{aligned}$$

similarly

$$\mathbb{E}_\pi \left[\mu_{s-1} \cdot \tilde{\mu}_s \mid \mathcal{F}_t \otimes \tilde{\mathcal{F}}_t\right] = 0.$$

Moreover we find that

$$M_t = \mathbb{E}_\pi \left[M_T \mid \mathcal{F}_t \otimes \tilde{\mathcal{F}}_t\right] \leq \mathbb{E}_\pi \left[d^r \mid \mathcal{F}_t \otimes \tilde{\mathcal{F}}_t\right].$$

However, as

$$M_0 = \mathbb{E}_\pi d^r$$

because of the vanishing duality gap we may finally conclude that

$$M_t = \mathbb{E}_\pi \left[d^r \mid \mathcal{F}_t \otimes \tilde{\mathcal{F}}_t\right] \qquad \pi\text{-a.e.},$$

completing the proof. □

8. Implementation for Finite Trees. Numerical experiments have been performed with the concept of the nested distance as elaborated above for two processes with finitely many outcomes.

In order to compute the multistage distance let us start as above and compute the Wasserstein distance for two measures $P = \sum_i P_i \delta_{\omega_i}$ and $\tilde{P} = \sum_j \tilde{P}_j \delta_{\tilde{\omega}_j}$ first – this is the two stage situation, $\mathbf{T} = \{0, 1\}$. The linear program corresponding to (4.1) is

$$\begin{array}{lll} \text{minimize} & (\text{in } \pi) & \sum_{i,j} \pi_{i,j} d_{i,j}^r \\ \text{subject to} & & \sum_j \pi_{i,j} = P_i \\ & & \sum_i \pi_{i,j} = \tilde{P}_j \\ & & \pi_{i,j} \geq 0, \end{array} \tag{8.1}$$

where the matrix $d_{i,j}$ carries the distances $d_{i,j} = d\left(\omega_i, \tilde{\omega}_j\right)$.

For multistage distances let the process be represented by a tree process sitting on some nodes. Any such node has a given depth described by the function depth $(n) \in \mathbf{T}$ (the depth of the root node is 0, all terminal nodes (or leaves) have depth T). Any node (m, say) may have some successor nodes (the direct children), which are collected in the set $m+$. The binary relation $m \leq \omega$ indicates that $m = \text{pred}_t \omega$ for some $t \in \mathbf{T}$. The conditional transition from a given node m to a successor $m' \in m+$ is described by a given conditional probability $p_{m:m'}$; for a tree, however, it is enough to give the conditional probability for the immediate successors.

Combining these observations one may formulate the multistage distance (5.2) as a linear program, just in line with (8.1), as

$$\begin{array}{lll} \text{minimize} & (\text{in } \pi) & \sum_{i,j} \pi_{i,j} d_{i,j}^r \\ \text{subject to} & & \sum_{i,j} \pi_{i,j} = 1, \\ & & p_{m:\omega} = \frac{\sum_{j \geq n} \pi_{\omega,j}}{\sum_{i \geq m,\, j \geq n} \pi_{i,j}} \qquad \left(m \leq \omega,\ \text{depth}\,(n) = \text{depth}\,(m)\right), \\ & & \tilde{p}_{n:\omega'} = \frac{\sum_{i \geq n,} \pi_{i,\omega'}}{\sum_{i \geq m,\, j \geq n} \pi_{i,j}} \qquad \left(n \leq \omega',\ \text{depth}\,(n) = \text{depth}\,(m)\right), \\ & & \pi_{i,j} \geq 0,\ \omega \in \left\{\omega_i : i\right\}, \left\{\tilde{\omega}_j : j\right\}, \end{array} \tag{8.2}$$

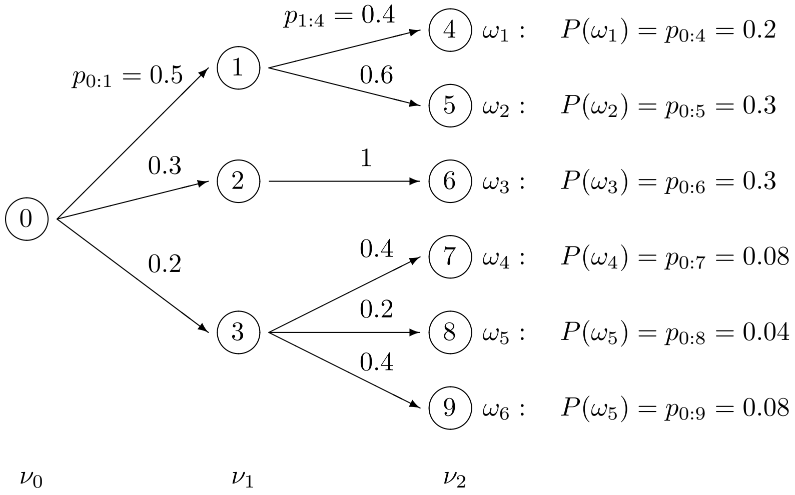
$p_{1:4} = 0.4$
$p_{0:1} = 0.5$
0.6
0.3
1
0.2
0.4
0.2
0.4
0
1
2
3
4
5
6
7
8
9
ω_1 : $P(\omega_1) = p_{0:4} = 0.2$
ω_2 : $P(\omega_2) = p_{0:5} = 0.3$
ω_3 : $P(\omega_3) = p_{0:6} = 0.3$
ω_4 : $P(\omega_4) = p_{0:7} = 0.08$
ω_5 : $P(\omega_5) = p_{0:8} = 0.04$
ω_6 : $P(\omega_5) = p_{0:9} = 0.08$
ν_0
ν_1
ν_2

FIG. 8.1. *An Example of a finite tree process $\nu = (\nu_0, \nu_1, \nu_2)$ with 10 nodes and 6 leaves.*

where $p_{m:\omega}$ is the probability to finally reach ω, conditional on the node m. For a concrete numerical implementation it should be noted that $\sum_{m\leq\omega} p_{m:\omega} = 1$. So one of the $\#\{\omega\colon m \leq \omega\}$ conditions in (8.2) for $p_{m:\omega}$ ($p_{n:\omega'}$, resp.) can (and should) be dropped, as they turn out to be linearly dependent, which impacts the numerical solver to find a solution.

EXAMPLE 1 (cf. [14] for a similar example). *The trees in Figure 8.2 represent three 2-stage processes. They have been chosen similar, but they significantly differ in their information structure (think of ε as a number between zero and one):* The optimal transport measure π for the multistage distance does not depend on the particular distance function chosen, it is

- $\pi = \begin{pmatrix} p \cdot p' & p\,(1-p') \\ (1-p)\,p' & (1-p)\,(1-p') \end{pmatrix}$ for the first and the second tree in the display, and
- similar (just replace p' by p'') for the distance of the first and third tree.
- As for the distance between the second and third tree the optimal transport measure is

$$\pi = \begin{pmatrix} p' \wedge p'' & (p''-p') \vee 0 \\ (p'-p'') \vee 0 & 1-(p' \vee p'') \end{pmatrix}.$$

All these trees differ – in the nested distance – by a number almost one: Indeed, for the particular choice $p = p' = p'' = \frac{1}{2}$ and employing the distance (5.1) for the different paths

- the nested distance of the first and second trees is $1+\varepsilon$: Note that, neglecting the information structure, the distance of these trees just would be ϵ, so the nested distance is able to distinguish the different information available at stage one;
- for the first and third it is 2 and
- the distance of the second and third tree is $1-\varepsilon$; again, for $\varepsilon \nearrow 1$, the nested distance approaches 0, just in line with the information available.

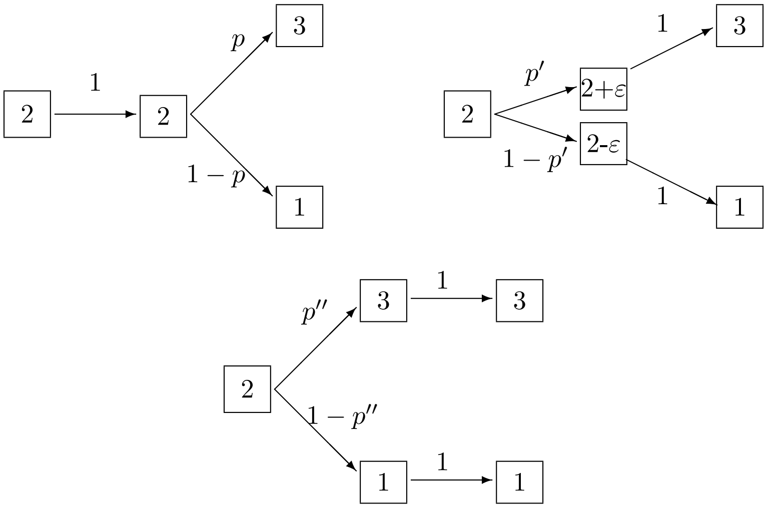

Fig. 8.2. *The states of three tree processes to illustrate three different flows of information.*

9. Summary and Outlook. In this paper we construct a space which is rich enough to carry the outcomes of a process and the evolving information as well. On top of that we elaborate a notion of distance, the *nested* or *multistage distance*, which is a refined and enhanced concept, based on the initial work of the first author in [25].

The crucial and additional observation here is that the nested distance is adapted for problems arising in stochastic optimization. For this reason the nested distance can be used to quantitatively and qualitatively study stochastic optimization problems from a very new perspective, because now there is a general notion of distance, respecting the flow of information, available.

Future Research

Computational Efficiency. A speedy computation of the nested distance is of interest. For this reasons different implementations have to be considered and compared. Moreover, computational efficiency has to be built on the dual, reducing the number of comparisons of sub-trees to a minimal extend. For example we did not address a backwards recursive computation of the nested distance here although this is available.

Risk Measures. We have elaborated in this paper the importance of the nested distance for stochastic optimization with the aim to minimize the expected loss of a given value function. However, in this context risk is not incorporated at all: To incorporate risk measures in the stochastic programming framework seems to be a task worth the effort; this has been addressed in [26] too but an extension for nested distances is necessary.

Scenario Reduction. An obvious application is provided by the fact that the nested distance can be used to reduce the number of scenarios in a given scenario tree: The nested distance can be employed to reduce this given scenario tree in such way that the objective of the reduced problem is close in the sense desired. Reducing the scenario tree is of crucial interest, this may even turn an intractable problem computable.

Scenario Generation. As regards scenario generation the nested distance can be employed to decide where to add scenarios in order to better approximate the situation.

Applications. Stochastic programming offers a huge field for applications, we just pick out energy (cf. [16]) to stand representative for all of them. The nested distance gives a first bound at hand to relate a numerical result to the precise, although numerically intractable objective value.

REFERENCES

[1] E. Allevi, M. Bertocchi, M.T. Vespucci, and M. Innorta, *A stochastic optimization model for a gas sale company*, IMA J Management Math, 19 (2008), pp. 403–416.

[2] Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré, *Gradient Flows in Metric Spaces and in the Space of Probability Measures.*, Birkhäuser, Basel, Switzerland, 2005.

[3] Edward S. Boylan, *Epiconvergence of martingales*, Ann. Math. Statist., 42 (1971), pp. 552–559.

[4] Yann Brenier, *Polar factorization and monotone rearrangement of vector-valued functions*, Comm. Pure Appl. Math., 44 (1991), pp. 375–417.

[5] Abraham I. Brodt, *Min-mad life: A multi-period optimization model for life insurance company investment decisions*, Insurance: Mathematics and Economics, 2 (1983), pp. 91–102.

[6] Raymond K. Cheung and Warren B. Powell, *An algorithm for multistage dynamic networks with random arc capacities, with an application to dynamic fleet management*, Operations Research, 44 (1996), pp. 951–963.

[7] Claude Dellacherie and Paul-André Meyer, *Probabilities and Potential.*, North-Holland Publishing Co., Amsterdam, The Netherlands., 1988.

[8] Richard A. Durrett, *Probability. Theory and Examples*, Duxbury Press, Belmont, CA, second ed., 2004.

[9] N. C. P. Edirisinghe, *Multiperiod portfolio optimization with terminal liability: Bounds for the convex case.*, Computational Optimization and Applications, 32 (2005), pp. 29–59.

[10] Alexei A. Gaivoronski, *Stochastic Optimization Problems in Telecommunications*, in Applications of Stochastic Programming, Stein W. Wallace and William T. Ziemba, eds., vol. 5, MPS-SIAM Series in Optimization, 2005, ch. 32.

[11] Richard J. Gardner, *The Brunn-Minkowski inequality*, Bulletin of the American Mathematical Society, 39 (2002), pp. 355–405.

[12] Holger Heitsch and Werner Römisch, *Scenario tree modeling for multistage stochastic programs*, Math. Program. Ser. A, 118 (2009), pp. 371–406.

[13] Holger Heitsch and Werner Römisch, *Scenario tree reduction for multistage stochastic programs.*, Computational Management Science, 2 (2009), pp. 117–133.

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

[15] Ronald Hochreiter, Georg Ch. Pflug, and David Wozabal, *Multi-stage stochastic electricity portfolio optimization in liberalized energy markets.*, in System Modeling and Optimization, vol. 199, Springer, New York, 2006, pp. 219–226.

[16] Ronald Hochreiter and David Wozabal, *A multi-stage stochastic programming model for managing risk-optimal electicity portfolios*, Handbook Power Systems II, 4 (2010), pp. 383–404.

[17] Leonid Kantorovich, *On the translocation of masses*, C.R. Acad. Sci. URSS, 37 (1942), pp. 199–201.

[18] Willem K. Klein-Haneveld, Matthijs H. Streutker, and Maarten H. van der Vlerk, *Indexation of Dutch pension rights in multistage recourse ALM models*, IMA Journal of Management Mathematics, 21 (2010), pp. 131–148.

[19] Lisa A. Korf and Roger J.-B. Wets, *Random LSC functions: An ergodic theorem*, Mathematics of Operations Research, 26 (2001), pp. 421–445.

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

[21] Petr Lachout, Eckhard Liebscher, and Silvia Vogel, *Strong convergence of estimators as ϵ_n-minimisers of optimisation problems*, Ann. Inst. Statist. Math., 57 (2005), pp. 291–313.

[22] Andris Möller, Werner Römisch, and Klaus Weber, *Airline network revenue management by multistage stochastic programming*, Computational Management Science, 5 (2008), pp. 355–377.

[23] 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, (1781), pp. 666–704.

[24] Paul Olsen, *Discretizations of multistage stochastic programming problems.*, Math. Programming Stud., 6 (1976), pp. 111–124.

[25] Georg Ch. Pflug, *Version-independence and nested distribution in multistage stochastic optimization*, SIAM Journal on Optimization, 20 (2009), pp. 1406–1420.

[26] Georg Ch. Pflug and Alois Pichler, *Approximations for Probability Distributions and Stochastic Optimization Problems*, vol. 163 of International Series in Operations Research & Management Science, Springer New York, 2011, ch. 15, pp. 343–387.

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

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

[29] Svetlozar T. Rachev and Ludger Rüschendorf, *Mass Transportation Problems Vol. I: Theory, Vol. II: Applications*, vol. XXV of Probability and its applications, Springer, New York, 1998.

[30] Werner Römisch and Rüdiger Schultz, *Stability analysis for stochastic programs*, Annals of Operations Research, 30 (1991), pp. 241–266.

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

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

[33] Maurice Sion, *On general minimax theorems.*, Pacific Journal of Mathematics, 8 (1958), pp. 171–176.

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

[35] Cédric Villani, *Optimal transport, old and new*, vol. 338 of Grundlehren der Mathematischen Wissenschaften, Springer, New York, 2009.

[36] David Williams, *Probability with Martingales*, Cambridge University Press, Cambridge, 1991.

[37] Przemysław Wojtaszczyk, *Banach Spaces for Analysts*, Cambridge University Press, Cambridge, 1991.

Appendix A. In this Appendix we demonstrate how the multistage distance for $r = 1$ can be seen as the usual transportation distance in the Polish space of *nested distributions*.

Recall the following well-known facts. For a given Polish space (Ξ, d) let $\mathcal{P}_1(\Xi, d)$ be the family of Borel probabilities on (Ξ, d) such that $y \mapsto d_t(y, y_0)$ is integrable for some $y_0 \in \Xi_t$.

For two Borel probabilities P and $\tilde{P}$ in $\mathcal{P}_1(\Xi, d)$, the Wasserstein distance $d\left(P, \tilde{P}\right)$ is given by

$$\begin{aligned} \mathsf{d}_1\left(P, \tilde{P}\right) &= \inf\left\{\mathbb{E} d\left(X, \tilde{X}\right) : X \sim P, \tilde{X} \sim \tilde{P}\right\} \\ &= \sup\left\{\int h(y)\, P(\mathrm{d}y) - \int h(y)\, \tilde{P}(\mathrm{d}y) : |h(y) - h(z)| \leq d\,(y, z)\right\}. \end{aligned}$$

Here $X \sim P$ means that the (marginal) distribution of X is P (cf. (4.1)).

Notice that we denote the original metric and the induced Kantorovich metric with the same symbol. This is justified, since (Ξ, d) is isometrically embedded in $\mathcal{P}_1\left(\Xi, d\right)$ by

$$d\left(\delta_y, \delta_z\right) = d\left(y, z\right)$$

where δ_y are Dirac probability measures. $\mathcal{P}_1$ is a complete separable metric space under d. If (Ξ_1, d_1) and (Ξ_2, d_2) are Polish spaces then so is the Cartesian product $(\Xi_1 \times \Xi_2, d_{1,2})$ with metric

$$d_{1,2}((y_1, y_2), (z_1, z_2)) := d_1(y_1, z_1) + d_2(y_2, z_2).$$

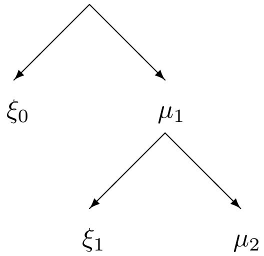

FIG. A.1. *Nested Distribution.*

We define the following Polish metric spaces in a (backward) recursive way.

$$\begin{aligned}
(\Xi_T, \mathsf{dl}_T) &= (\mathbb{R}^{n_T}, d_T) \\
(\Xi_{T-1}, \mathsf{dl}_{T-1}) &= \left(\mathbb{R}^{n_{T-1}} \times \mathcal{P}_1\left(\Xi_{T:T}, \mathsf{dl}_T\right),\, d_{T-1}(\cdot,\cdot) + \mathsf{dl}_T(\cdot,\cdot)\right) \\
(\Xi_{T-2}, \mathsf{dl}_{T-2}) &= \left(\mathbb{R}^{n_{T-2}} \times \mathcal{P}_1\left(\Xi_{T-1:T}, \mathsf{dl}_{T-1}\right),\, d_{T-2}(\cdot,\cdot) + \mathsf{dl}_{T-1}(\cdot,\cdot)\right) \\
&\vdots \\
(\Xi_0, \mathsf{dl}_0) &= \left(\mathbb{R}^{n_0} \times \mathcal{P}_1\left(\Xi_{1:T}, \mathsf{dl}_1\right),\, d_0(\cdot,\cdot) + \mathsf{dl}_1(\cdot,\cdot)\right)
\end{aligned}$$

For simplicity, we write (Ξ, dl) instead of (Ξ_0, dl_0).

DEFINITION A.1. *A Borel probability distribution* $\mathbb{P} \in (\Xi_0, \mathsf{dl})$ *is called a* nested distribution *(of depth T).*

We illustrate the situation for depth $T = 3$. A nested distribution on $\Xi = \Xi_0$, has components ξ_0 (a value in $\mathbb{R}^{n_0}$) and μ_1 (a nested distribution on Ξ_1), which means that it has in turn components ξ_1 (a $\mathbb{R}^{n_1}$-random variable) and μ_2, a random distribution on $\mathbb{R}^{n_2}$). One may visualize the situation as in Figure A.1.

It has been shown in [25] how the nested distribution $\mathbb{P}$ is related to the value-and-information structure $(\Omega, \mathfrak{F}(\text{pred}), P, \xi)$. We summarize here the basic facts.

- If $(\Omega, \mathfrak{F}(\text{pred}), P, \xi)$ is a value-and-information structure with $\omega_t = \text{pred}_t(\omega)$, it induces a nested distribution. For $T = 1$, the nested distribution is $\mathcal{L}(\xi_0 \times \mathcal{L}(\xi_1|\omega_0))$, which is the joint law of ξ_0 and the conditional law of ξ_1 given ω_0 (the root of the tree) [8]. For general T, the induced nested distribution is

$$\mathcal{L}\left(\xi_0 \times \mathcal{L}\left(\xi_1 \times \ldots \mathcal{L}\left(\xi_{T-2} \times \mathcal{L}\left(\xi_{T-1} \times \mathcal{L}\left(\xi_T|\nu_{T-1}\right)|\nu_{T-2}\right)|\nu_{T-1}\right)\ldots|\nu_1\right)\right).$$

- Conversely, to every nested distribution one may associate a standard tree process ω_t and a value process $\xi_t = \xi_t(\omega_t)$ such that its nested distribution is $\mathbb{P}$.

THEOREM A.2. *Let $\mathbb{P}$ resp. $\tilde{\mathbb{P}}$ be two nested distributions in Ξ and let $(\Omega, \mathfrak{F}(pred, P, \xi)$ resp. $(\tilde{\Omega}, \tilde{\mathfrak{F}}(pred, \tilde{P}, \tilde{\xi})$ be the pertaining value-and-information structures. Then the distance $\mathsf{dl}(\mathbb{P}, \tilde{\mathbb{P}})$ equals to the solution of the optimization problem* (5.2).

REMARK 7. *The theorem is formulated for the situation $r = 1$ only, but can be generalized with obvious modifications to arbitrary r.*

Proof. The proof goes in two steps. First, we show that that we may assume that the two scenario processes ξ and $\tilde{\xi}$ are *final*, i.e. that nonzero values appear only

[8] $\mathcal{L}(\xi)$ is the probability law of ξ

in the last stage. To this end, define for each nested distribution the final nested distribution $\mathbb{P}^f$ in the following way:

$$\begin{aligned}\xi_t^f(\omega) &:= 0 \qquad\qquad \text{for } t=0,\dots T-1\\ \xi_T^f(\omega) &:= \left(\xi_T(\omega), \xi_{T-1}\left(\text{pred}_{T-1}(\omega)\right), \dots \xi_1\left(\text{pred}_1(\omega)\right), \xi_0\right).\end{aligned}$$

Notice that ξ_T^f takes values in the metric space

$$\left(\mathbb{R}^{n_0} \times \mathbb{R}^{n_1} \times \dots \mathbb{R}^{n_T},\, d_0 + d_1, + \dots + d_T\right).$$

We show that

$$\mathsf{dl}\left(\mathbb{P}, \tilde{\mathbb{P}}\right) = \mathsf{dl}\left(\mathbb{P}^f, \tilde{\mathbb{P}}^f\right).$$

The proof of this assertion goes by induction. It is trivial for $T = 1$. Suppose it has been proved for depth T. By the recursive character of the nested distance, to show the assertion for $T+1$ it suffices to show it for $T = 2$. The nested distance of two distributions with depth 2 equals $d_0(\xi_0, \tilde{\xi}_0) + d_1\left(\mathcal{L}(\xi_1), \mathcal{L}(\tilde{\xi}_1)\right)$. If $d(P, \tilde{P})$ is the Wasserstein distance of two probability measures with respect to the distance d and we define a new "distance" $d^a\left(v_1, v_2\right) = d\left(v_1, v_2\right) + a$, then $d^a\left(P, \tilde{P}\right) = d\left(P, \tilde{P}\right) + a$. Thus with $a = d_0\left(\xi_0, \tilde{\xi}_0\right)$,

$$d_0\left(\xi_0, \tilde{\xi}_0\right) + d_1\left(\mathcal{L}(\xi_1), \mathcal{L}(\tilde{\xi}_1)\right) = d_1^a\left(\mathcal{L}(\xi_1), \mathcal{L}(\tilde{\xi}_1)\right) = d_1\left(\mathcal{L}(\xi_0, \xi_1), \mathcal{L}(\tilde{\xi}_0, \tilde{\xi}_1)\right),$$

which is the asserted equality.

For a final nested distribution, the distances dl_t are given by the Wasserstein distances in $\mathcal{P}_1\left(\Xi_{t+1}\right)$, implying that $\Xi = \mathcal{P}_1\left(\mathcal{P}_1\left(\dots\left(\mathcal{P}_1(\Xi_T)\right)\dots\right)\right)$. For two final nested distributions a correct transportation measure π must only respect the nested structure of the conditional distributions. At each level one should find an optimal transportation plan for every pair of sub-trees. When combining them to an overall transportation plan π the new plan then should satisfy the conditions

$$\begin{aligned}\pi\left[A \times \tilde{\Omega} \mid \mathcal{F}_t \otimes \tilde{\mathcal{F}}_t\right](\omega, \tilde{\omega}) &= P\left[A \mid \mathcal{F}_t\right](\omega) \quad \left(A \in \mathcal{F}_T,\, t \in \mathbf{T}\right),\\ \pi\left[\Omega \times B \mid \mathcal{F}_t \otimes \tilde{\mathcal{F}}_t\right](\omega, \tilde{\omega}) &= \tilde{P}\left[B \mid \tilde{\mathcal{F}}_t\right](\tilde{\omega}) \quad \left(B \in \tilde{\mathcal{F}}_T,\, t \in \mathbf{T}\right)\end{aligned}$$

which is identical to (5.2) in Definition 5.1.□