Available online at www.sciencedirect.com
ScienceDirect
Stochastic Processes and their Applications 130 (2020) 5575–5591
stochastic processes and their applications
www.elsevier.com/locate/spa

Fundamental properties of process distances

Julio Backhoff Veraguas[a,1], Mathias Beiglböck[a,*,1], Manu Eder[a,1], Alois Pichler[b]

[a] *University of Vienna, Austria*
[b] *University of Technology, Chemnitz, Germany*

Received 12 January 2018; received in revised form 29 August 2019; accepted 27 March 2020
Available online 18 April 2020

Abstract

To quantify the difference of distinct stochastic processes it is not sufficient to consider the distance of their states and corresponding probabilities. Instead, the information, which evolves and accumulates over time and which is mathematically encoded by filtrations, has to be accounted for as well. The *nested distance*, also known as *bicausal Wasserstein distance*, recognizes this component and involves the filtration properly. This distance is of emerging importance due to its applications in stochastic analysis, stochastic programming, mathematical economics and other disciplines.

This paper investigates the basic metric and topological properties of the nested distance on the space of discrete-time processes. In particular we prove that the nested distance generates a Polish topology, although the genuine space is not complete. Moreover we identify its completion to be the space of *nested distributions*, a space of generalized stochastic processes.

MSC: 60G05; 60G99; 90C15

Keywords: Optimal transport; Nested distance; Martingales; Causal Wasserstein distance; Information topology

1. Introduction

A real-valued stochastic process $X = (X_t)_{t=1}^N$ is fully described by its law $\mathcal{L}_X$, which itself is a probability measure on $\mathbb{R}^N$. It is a basic and fundamental question in applications when two processes should be considered as close. Simply endowing $\mathcal{P}(\mathbb{R}^N)$, the set of probability measures on $\mathbb{R}^N$, with the weak topology or Wasserstein distance (say) is not adequate as the

* Corresponding author.
E-mail address: mathias.beiglboeck@univie.ac.at (M. Beiglböck).
[1] The authors acknowledge support by the Austrian Science Fund (FWF) under grant Y782-N25.

https://doi.org/10.1016/j.spa.2020.03.017
0304-4149/

evolution over consecutive time steps $t = 1, \dots, N$ matters. Various groups of researches from different areas ([2] — stochastic analysis; [16] — mathematical economics; [23] – optimization; [17] — logic; [11] — mathematical finance) have introduced different 'adapted' refinements of the usual weak topology. Remarkably all of these concepts lead to the same topology in the present finite discrete time setup, see [7].

It is the main goal of this paper to recognize the basic topological (and metric, resp.) properties of the space $\mathcal{P}(\mathbb{R}^N)$ when equipped with this prevailing and common adapted topology.

1.1. Nested distance and completeness

The nested distance d^{nd} has been introduced by Pflug and Pichler [21,22]. It can be seen as a modification of the celebrated Wasserstein distance: in contrast to this classical distance, the nested distance respects the underlying filtrations, conditioned on the time elapsed, see (3.1). With this additional feature the nested distance is well suited for a number of problems in stochastic optimization, stochastic control, and mathematical finance, where decisions are subject to the time elapsed.

We demonstrate that $(\mathcal{P}(\mathbb{R}^N), d^{\mathrm{nd}})$ is a separable metric space that is necessarily incomplete if $N > 1$, see Theorem 4.6 and Example 4.1. As a first central result, Theorem 4.9 identifies its completion as the space of *nested distributions*. Of course, the completion of a metric space is unique (up to isometry), but the space of nested distributions is constructed explicitly and allows for natural interpretations.

1.2. $(\mathcal{P}(\mathbb{R}^N), d^{\mathrm{nd}})$ is a Polish space

Although not complete with respect to the nested distance d^{nd}, our second main result (Theorem 5.6) demonstrates that $\mathcal{P}(\mathbb{R}^N)$, equipped with the topology generated by the nested distance, is nonetheless Polish. That is, there exists a metric d such that $(\mathcal{P}(\mathbb{R}^N), d)$ is a separable complete metric space and d induces the same topology τ^{nd} as d^{nd}.

Importantly we do not suggest to replace d^{nd} with the complete metric; in fact d^{nd} is much more natural and handy in applications. Rather it is the pure fact that $(\mathcal{P}(\mathbb{R}^N), \tau^{\mathrm{nd}})$ is Polish which has significant consequences. E.g., the Borel sets of $(\mathcal{P}(\mathbb{R}^N), d^{\mathrm{nd}})$ are the usual Borel sets of $\mathcal{P}(\mathbb{R}^N)$ equipped with Wasserstein distance or weak convergence, see Remark 5.7. It implies that familiar techniques for establishing measurability can be applied as usual, e.g. allowing to apply common measurable selection theorems which are often crucial in optimization . In short, Theorem 5.6 guarantees that one may work on $(\mathcal{P}(\mathbb{R}^N), d^{\mathrm{nd}})$ in the same carefree way as often accepted in optimization, or even in analysis and probability.

1.3. Extreme points

For computational reasons, any simplification of a measure or a law of a stochastic process is of crucial importance in stochastic optimization and applications. Brenier maps, e.g., provide concrete transformations of measures which may be used to simplify a given measure. One reason why deterministic maps are useful in applications is that they are dense in the set of all couplings with a given initial marginal and, more importantly, they coincide with the extreme point of such sets of couplings. Section 6 establishes the analogue result when filtrations are considered: deterministic and adapted transformations of an initial measure correspond exactly

to the extreme points of the set of all couplings preserving causality (i.e., the filtrations) and having the same initial measure. This is also relevant for the nested distance, as such sets of couplings constitute building blocks of this transport-based metric.

1.4. Relation to the literature

Among the first authors to consider the role of information in defining a process topology are [2], [17], [16] and [21]. Aldous was concerned with a topology granting the stability of optimal stopping problems and the semimartingale decomposition, Hellwig was interested in the stability of games (see also [8]) and Pflug in the stability of stochastic programs. For the latter purpose, Pflug and Pichler [21–24] introduce the *nested distance*, borrowing from an analogy with optimal transport and Wasserstein distances.[2] The nested distance is a particular case of a (bi)causal transport problem, a concept introduced by [19] and further developed in [1,5]. In fact, the latter articles are in the wider tradition of constrained transport problems and in particular are related to martingale optimal transport (cf. [9,10,14,15,20] among many others). Independently, [11] have considered a continuous-time analogue to the nested distance for the laws of diffusion processes; see also [6] for the case of continuous semimartingale laws and applications to stability in mathematical finance.

Risk averse optimal stopping/stochastic optimization problems and their numerics are considered from a nested distance perspective in [26,27]. The article [7] establishes that the weak nested topology is the coarsest topology that guarantees continuity of optimal stopping problems. Notably we believe that these results represent only a prelude on the interplay between nested distance and optimal stopping/stochastic control which appears as a promising field for future research.

Outline. Section 2 introduces the notation used throughout the paper and describes the mathematical setting. Section 3 discusses elementary properties of the nested distance, while Section 4 is concerned with its metric properties. Next, in Section 5, we introduce the weak nested topology and establish its Polish character. Section 6 discusses extreme points of important sets associated with the nested distance, and Section 7 concludes with a brief summary.

2. Notation and mathematical setup

The ambient set throughout this article is $\mathbb{R}^N$, which we consider as a filtered space endowed with the canonical (i.e., coordinate) filtration $\left(\mathcal{F}_t\right)_{t=1}^N$. More precisely, $\mathcal{F}_t$ is the smallest σ-algebra on $\mathbb{R}^N$ such that the projection $\mathbb{R}^N \ni x \mapsto (x_1, \dots, x_t) \in \mathbb{R}^t$ onto the first t components is Borel-measurable.

We endow $\mathbb{R}^N$ with the ℓ^p-type metric

$$d(x, y) := d_p(x, y) := \sqrt[p]{\sum_{t=1}^{N} \underline{d}(x_i, y_i)^p}, \qquad p \in [1, \infty), \tag{2.1}$$

where $\underline{d}$ is some base metric on $\mathbb{R}$ compatible with the usual topology. We are particularly interested in the cases where $\underline{d}$ is the usual distance or is a compatible bounded metric on $\mathbb{R}$. Throughout this paper we fix $\underline{d}$, p and d as described.

[2] For classical optimal transport and Wasserstein distances we refer to the monographs [3,28,29,31–33].

Remark 2.1. We make the important remark that, mutatis mutandis, our results in Sections 3, 4 and 5 hold true when replacing $\mathbb{R}^N$ by S^N, where S is an abstract Polish space. The reason to work with processes taking values in $\mathbb{R}$ is just to simplify notation and avoid confusion when referring to various Polish spaces of probability measures.

The pushforward of a measure γ by a map M is denoted by $M_*\gamma := \gamma \circ M^{-1}$. For a product of sets $\mathcal{X} \times \mathcal{Y}$ we denote by p^1 (p^2, resp.) the projection onto the first (second, resp.) coordinate. We denote by γ^x, γ^y the regular kernels of a measure γ on $\mathcal{X} \times \mathcal{Y}$ with respect to its first and second coordinate, respectively, obtained by disintegration so that $\gamma(A \times B) = \int_A \gamma^{x_1}(B)\,\gamma^1(\mathrm{d}x_1)$ with $\gamma^1(A) := p^1_*\gamma(A) = \gamma(A \times \mathcal{Y})$ (cf. [4]). The notation extends analogously to products of more than two spaces. We convene that for a probability measure η on $\mathbb{R}^N$, $\eta^{x_1,\dots,x_t}$ denotes the one-dimensional measure on x_{t+1} obtained by disintegration of η with respect to $(x_1, \dots, x_t)$. Finally, if $\mathcal{Z}$ is a topological space, we denote by $\mathcal{P}(\mathcal{Z})$ the set of Borel probability measures, $C(\mathcal{Z})$ the set of real-valued continuous functions, and $supp(m)$ refers to the topological support of the measure $m \in \mathcal{P}(\mathcal{Z})$.

A statement like "for η-a.e. $x_1, \dots, x_t$" is meant to denote "almost-everywhere" with respect to the projection of η onto the coordinates $(x_1, \dots, x_t)$. On $\mathbb{R}^N \times \mathbb{R}^N$ we denote by $(x_1, \dots, x_N)$ the first half and by $(y_1, \dots, y_N)$ the second half of the coordinates. Similarly, we use the convention that for a probability measure γ on $\mathbb{R}^N \times \mathbb{R}^N$, $\gamma^{x_1,\dots,x_t,y_1,\dots,y_t}$ denotes the two-dimensional measure on (x_{t+1}, y_{t+1}) given by regular disintegration of γ with respect to $(x_1, \dots, x_t, y_1, \dots, y_t)$, so a statement like "for γ-a.e. $x_1, \dots, x_t, y_1, \dots, y_t$" is meant to denote "almost-everywhere" with respect to the projection of γ onto $x_1, \dots, x_t, y_1, \dots, y_t$.

The probability measures on the product space $\mathbb{R}^N \times \mathbb{R}^N$ with marginals μ and ν constitute the possible *transport plans* or *couplings* between the given marginals. We denote this set by

$$\Pi(\mu, \nu) = \left\{\gamma \in \mathcal{P}(\mathbb{R}^N \times \mathbb{R}^N) \colon \gamma \text{ has marginals } \mu \text{ and } \nu\right\}.$$

We often consider processes $X = (X_t)_{t=1}^N$, $Y = (Y_t)_{t=1}^N$ defined on some probability space. Each pair (X, Y) is a coupling or – abusing notation slightly – a transport plan upon identifying it with its law. For the sake of simplicity, being measurable with respect to a sigma algebra means to be equal to a correspondingly measurable function modulo a null set with respect to the measure relevant in the given context.

Definition 2.2 (*Causality*)**.** A transport plan $\gamma \in \Pi(\mu, \nu) \subset \mathcal{P}(\mathbb{R}^N \times \mathbb{R}^N)$ is called *bicausal* (between μ and ν) if the mappings

$$\mathbb{R}^N \ni x \mapsto \gamma^x(B) \text{ and } \mathbb{R}^N \ni y \mapsto \gamma^y(B)$$

are $\mathcal{F}_t$-measurable for any $B \in \mathcal{F}_t \subset \mathbb{R}^N$ and $t < N$. The collection of all bicausal plans is

$$\Pi_{bc}(\mu, \nu).$$

The product measure $\mu \otimes \nu$ is bi-causal, so $\Pi_{bc}(\mu, \nu)$ is non-empty. In terms of stochastic processes, a coupling is *bicausal* if

$$\gamma\big((Y_1, \dots, Y_t) \in B_t \mid X_1, \dots, X_N\big) = \gamma\big((Y_1, \dots, Y_t) \in B_t \mid X_1, \dots X_t\big) \text{ and}$$
$$\gamma\big((X_1, \dots, X_t) \in B_t \mid Y_1, \dots, Y_N\big) = \gamma\big((X_1, \dots, X_t) \in B_t \mid Y_1, \dots Y_t\big)$$

for all $t = 1, \dots, N$ and $B_t \subset \mathbb{R}^t$ Borel.

Testing whether a coupling or transport plan is bicausal reduces to a property of its transition kernel. Specifically we have the following characterization (see, e.g., [5])

Proposition 2.3 (*Characterization of Bicausality Via Transition Kernels*)**.** *The following are equivalent:*

(i) γ *is a bicausal transport plan on* $\mathbb{R}^N \times \mathbb{R}^N$ *between the measures* μ *and* ν.
(ii) The successive regular kernels $\bar{\gamma}$ *of the decomposition*

$$\begin{aligned}\gamma(\mathrm{d}x_1, \dots, \mathrm{d}x_N, \mathrm{d}y_1, \dots, \mathrm{d}y_N)\\ = \bar{\gamma}(\mathrm{d}x_1, \mathrm{d}y_1)\, \gamma^{x_1,y_1}(\mathrm{d}x_2, \mathrm{d}y_2) \dots \gamma^{x_1,\dots,x_{N-1},y_1,\dots,y_{N-1}}(\mathrm{d}x_N, \mathrm{d}y_N)\end{aligned} \tag{2.2}$$

satisfy

$$\bar{\gamma} \in \Pi(p^1_* \mu,\, p^1_* \nu)$$

and further, for $t < N$ *and* γ*-almost all* $x_1, \dots, x_t, y_1, \dots, y_t$,

$$p^1_* \gamma^{x_1,\dots,x_t,y_1,\dots,y_t} = \mu^{x_1,\dots,x_t} \text{ and } p^2_* \gamma^{x_1,\dots,x_t,y_1,\dots,y_t} = \nu^{y_1,\dots,y_t}. \tag{2.3}$$

3. The nested distance

Following [21–23] we consider for μ, ν as above the *p-nested distance*, or simply *nested distance*, defined by

$$d_p^{\mathrm{nd}}(\mu, \nu) := \inf_{\gamma \in \Pi_{bc}(\mu,\nu)} \left(\iint d^p \, \mathrm{d}\gamma \right)^{1/p} = \inf_{\gamma \in \Pi_{bc}(\mu,\nu)} \left(\iint \sum_{t=1}^{N} \underline{d}(x_t, y_t)^p \, \mathrm{d}\gamma \right)^{1/p}. \tag{3.1}$$

In direct analogy with the classical p-Wasserstein distance it defines a metric on the space

$$\mathcal{P}^p(\mathbb{R}^N) := \{\mu \in \mathcal{P}(\mathbb{R}^N) \colon \int d(x, x_0)^p \, \mu(\mathrm{d}x) < \infty \text{ for some } x_0\}.$$

As noted in [25], the nested distance (3.1) is best suited to separate μ from ν if their information structure differs. In particular, the authors show that empirical measures μ_n^{emp} of a multivariate measure μ with density never converge in nested distance (even though they do converge in Wasserstein distance); the essential point here is that each empirical measure μ_n^{emp} is roughly a tree with non-overlapping branches (commonly a *fan*) and therefore deterministic as soon as the first component is observed. From an information perspective, μ_n^{emp} is radically different from μ. Notably, this is a key property of the nested distance and this is the essential distinctive characteristic and strength in comparison with the Wasserstein distance.

3.1. Recursive computation

A useful comment at this point is that the nested distance can be stated and computed recursively as for Bellman equations: starting with $V_N^p := 0$ define

$$\begin{aligned}&V_t^p(x_1, \dots, x_t, y_1, \dots, y_t) := \\ &\inf_{\gamma^{t+1} \in \Pi(\mu^{x_1,\dots,x_t}, \nu^{y_1,\dots,y_t})} \iint \begin{pmatrix} V_{t+1}^p(x_1, \dots, x_{t+1}, y_1, \dots, y_{t+1}) \\ + \, \underline{d}(x_{t+1}, y_{t+1})^p \end{pmatrix} \gamma^{t+1}(\mathrm{d}x_{t+1}, \mathrm{d}y_{t+1}),\end{aligned} \tag{3.2}$$

so that the nested distance is finally obtained by V_0^p, i.e.,

$$d_p^{\mathrm{nd}}(\mu, \nu)^p = \inf_{\gamma^1 \in \Pi(p^1_* \mu, p^1_* \nu)} \iint \left(V_1^p(x_1, y_1) + \, \underline{d}(x_1, y_1)^p \right) \gamma^1(\mathrm{d}x_1, \, \mathrm{d}y_1). \tag{3.3}$$

3.2. Comparison with weak topology

The Wasserstein distance metrizes the weak topology on probability measures with suitably integrable moments. We recall that the weak topology (also called weak* or vague topology) is characterized by integration on bounded and continuous functions. It is thus natural to ask if there is a class of functions which characterizes the topology generated by the nested distance.

Proposition 3.1. *Let $N \geq 2$. There does* not *exist a family $\mathbb{F}$ of functions on $\mathbb{R}^N$ which determines convergence of a sequence $(\mu_n)_{n=1}^{\infty}$ towards $\mu \in \mathcal{P}^p$ with respect to the nested distance d_p^{nd}. I.e., there is* no *family $\mathbb{F}$ so that*

$$d_p^{\mathrm{nd}}(\mu_n, \mu) \to 0 \quad \Longleftrightarrow \quad \int f \,\mathrm{d}\mu_n \to \int f \,\mathrm{d}\mu \text{ for all } f \in \mathbb{F}.$$

In fact, such a convergence determining family does not even exist if the support of the measures is restricted to a bounded region $[-K, K]^N$, $K > 0$.

Proof. Assume that such a family exists. Without loss of generality we can further assume that the integral of $f \in \mathbb{F}$ against all measures in $\mathcal{P}^p$ is well-defined. By considering $\delta_{(x_1^n,\dots,x_N^n)}$, which converge in nested distance to $\delta_{(x_1,\dots,x_N)}$ if their supports do in $\mathbb{R}^N$, we conclude that $\mathbb{F} \subset C(\mathbb{R}^N)$. Set

$$\mu_\epsilon := \frac{1}{2}\left[\delta_{(\epsilon,\dots,\epsilon,1)} + \delta_{(-\epsilon,\dots,-\epsilon,-1)}\right] \quad \text{and} \quad \mu := \frac{1}{2}\left[\delta_{(0,\dots,0,1)} + \delta_{(0,\dots,0,-1)}\right].$$

By continuity we find that

$$\int f \,\mathrm{d}(\mu_\epsilon - \mu) \to 0 \text{ as } \epsilon \to 0 \tag{3.4}$$

for every $f \in \mathbb{F}$. Taking $\underline{d}$ to be the usual distance on $\mathbb{R}$ we find $d_p^{\mathrm{nd}}(\mu_\epsilon, \mu) \geq 2^{1-1/p}$. Indeed, we have from (3.2) that $V_t^p \geq \frac{1}{2}2^p$ for all $t = 1, \dots, N$ and from (3.3) that $d_p^{\mathrm{nd}}(\mu, \nu) \geq 2^{1-1/p}$. In general we find that $d_p^{\mathrm{nd}}(\mu_\epsilon, \mu) \geq 1$ and this contradicts (3.4). The initial assumption thus is false and $\mathbb{F}$ cannot determine convergence in nested distance. □

Remark 3.2 (*Separating evaluations*)**.** The nested distance was initially introduced with the intention to compare stochastic programs and the question addressed by the preceding Proposition 3.1 was initially posed by Pflug. Indeed, Corollary 2 in [22] demonstrates that there are stochastic optimization programs with differing objective values whenever the nested distance differs.

The separating objects are thus entire stochastic programs which, in view of the preceding Proposition 3.1, cannot be replaced by a set of functions on $\mathbb{R}^N$. The proposition further emphasizes the intrinsic relation between stochastic programs, the nested distance and the role of information.

We will see in Example 4.1 in the next section that d_p^{nd} is *not* complete. This further demonstrates how differing the nested distance and the usual Wasserstein distance are.

Remark 3.3 (*Equivalence with Respect to the Genuine Distance*)**.** We emphasize that the metric results in this and the following section and the topological results in Section 5 are

also applicable if we based the p-nested distance on an ℓ^q-type product norm in $\mathbb{R}$. Indeed, for each $q \in [1, \infty)$ we easily find $c, C > 0$ such that

$$c\, d(x, y) \leq d_q(x, y) \leq C\, d(x, y);$$

see (2.1) for notation. In particular, if we base the p-nested distance (3.1) in terms of d_q instead of $d = d_p$, we obtain a strongly equivalent metric on $\mathcal{P}^p$ (with the same constants c and C). By the form of the metric d, we obtained a convenient and amenable expression for d_p^{nd}, as seen on the right hand side of (3.1), which we would not have under d_q for $q \neq p$. For these reasons, we may and will continue to work with d_p^{nd} defined in terms of $d = d_p$ keeping in mind that the forthcoming results would generalize trivially.

4. Completeness and completion

The space $\mathcal{P}^p(\mathbb{R}^N)$, endowed with the p-Wasserstein distance, is complete. This is not the case for the nested distance, as the following example reveals.

Example 4.1 (*The Nested Distance is Not Complete*)**.** We observe that d_p^{nd} is not a complete metric as soon as the number of time steps N is greater than or equal than 2. For the sake of the argument we take $N = 2$, $\underline{d}$ the usual distance on $\mathbb{R}$ and consider $\mu_n = 1/2\left(\delta_{(1/n,1)} + \delta_{(-1/n,-1)}\right)$. One verifies that $d_p^{\mathrm{nd}}(\mu_n, \mu_m) \leq |1/n - 1/m|$, so the sequence is Cauchy. The only possible limit of this sequence is the limit based on the Wasserstein distance, that is $\mu = 1/2\left(\delta_{(0,1)} + \delta_{(0,-1)}\right)$. But in nested distance we have $d_p^{\mathrm{nd}}(\mu_n, \mu) = (2^{p-1} + n^{-p})^{1/p} > 1$, in particular this sequence does not tend to zero and we conclude that the nested distance is not complete for $N > 1$.

The distinguishing point is that μ is a real tree with coinciding states at the first stage, whereas the μ_n's are not. The nested distance is designed to capture this distinction, which is ignored by the Wasserstein distance.

To identify the completion of $\mathcal{P}^p(\mathbb{R}^N)$ with respect to the p-nested distance we consider the *nested distributions* introduced in [21].

Definition 4.2 (*Nested Distribution*)**.** Consider the sequence of metric spaces

$$\begin{aligned}
R_{N:N} &:= (\mathbb{R}, d_{N:N}) \text{ equipped with the distance } d_{N:N} := \underline{d} = [\underline{d}^p]^{1/p},\\
R_{N-1:N} &:= \left(\mathbb{R} \times \mathcal{P}^p(R_{N:N}), d_{N-1:N}\right) \text{ with } d_{N-1:N} := \left[\underline{d}^p + W^p_{d_{N:N},p}\right]^{1/p},\\
&\vdots\\
R_{1:N} &:= \left(\mathbb{R} \times \mathcal{P}^p(R_{2:N}), d_{1:N}\right) \text{ with } d_{1:N} := \left[\underline{d}^p + W^p_{d_{2:N},p}\right]^{1/p},
\end{aligned}$$

where at each stage t, the space $\mathcal{P}^p(R_{t:N})$ is endowed with the p-Wasserstein distance with respect to the metric $d_{t:N}$ on $R_{t:N}$, which we denote $W_{d_{t:N},p}$. The set of *nested distributions* (of depth N) with pth moment is defined as $\mathcal{P}^p(R_{1:N})$.

Each of the spaces $R_{t:N}$ ($t = 1, \ldots N$) is a Polish space. Indeed, a complete metric is given explicitly and the spaces are separable since $\mathcal{P}(R)$ is complete and separable whenever (R, ρ) is complete and separable (cf. [12]). We endow $\mathcal{P}^p(R_{1:N})$ with the complete metric $W_{d_{1:N},p}$.

Example 4.3. When $N = 2$, we have that $R_{1:2} = \mathbb{R} \times \mathcal{P}^p(\mathbb{R})$ and for $P, Q \in \mathcal{P}^p(R_{1:2})$ the distance is

$$W_{d_{1:2},p}(P, Q) = \left\{ \inf_{\Gamma \in \Pi(P,Q)} \iint \left(\underline{d}(x, y)^p + W_p^p(\mu, \nu) \right) \Gamma(\mathrm{d}x, \mathrm{d}\mu, \mathrm{d}y, \mathrm{d}\nu) \right\}^{1/p} \tag{4.1}$$

with W_p the classical p-Wasserstein distance for measures on the line and with respect to the metric $\underline{d}$. The formulation (4.1) notably exactly corresponds to the recursive descriptions (3.2) and (3.3).

4.1. Embedding

We demonstrate that the nested distributions of depth N introduced in Definition 4.2 extend the notion of probability measures in $\mathbb{R}^N$ in a metrically meaningful way. Let us introduce the following function, already present in [21], which associates $\mu \in \mathcal{P}^p(\mathbb{R}^N)$ with the nested distribution $I[\mu] \in \mathcal{P}^p(R_{1:N})$ given by

$$I[\mu] := \mathcal{L}\left(X_1\,,\, \mathcal{L}^{X_1}\left(X_2\,,\, \ldots\,,\, \mathcal{L}^{X_{1:N-2}}\left(X_{N-1}\,,\, \mathcal{L}^{X_{1:N-1}}(X_N)\right)\right)\right), \tag{4.2}$$

where $(X_1, \ldots, X_N)$ is a vector with law μ. We used the shorthand $\mathcal{L}^{X_{1:k}}$ for the *conditional law* given $(X_1, \ldots, X_k)$ (and missing superscripts indicate unconditional law).

Remark 4.4. To provide an example, $\mathcal{L}^{X_{1:N-1}}(X_N)$ is the law of X_N given the past up to time $N-1$, then $\mathcal{L}^{X_{1:N-2}}\left(X_{N-1}\,,\, \mathcal{L}^{X_{1:N-1}}(X_N)\right)$ is the joint law of $\mathcal{L}^{X_{1:N-1}}(X_N)$ and X_{N-1} given the past up to time $N-2$. The nested distribution $I[\mu]$ is obtained by repeating this procedure backwards in time.

Remark 4.5 (*Motivation and Relation to Stochastic Optimization*)**.** The embedding (4.2) naturally appears in stochastic optimization. Indeed, suppose that a stochastic process $X_{1:t}$ has already materialized up to time t, then the remaining distribution follows the conditional measure $\mathcal{L}^{X_{1:t}}$. The embedding $I[\mu]$ in (4.2) provides the holistic perspective for the entire time horizon from $t = 1$ up to $t = N$.

Theorem 4.6 (*Isometric Embedding*)**.** *Let $d = d_p$. Then the classical Wasserstein distance of nested distributions extends the nested distance of classical distributions. More precisely, the mapping I defined in (4.2) embeds the metric space $(\mathcal{P}^p(\mathbb{R}^N), d_p^{\mathrm{nd}})$ defined via (3.1)* isometrically *into the separable complete metric space $(\mathcal{P}^p(R_{1:N}), W_{d_{1:N},p})$. In particular $(\mathcal{P}^p(\mathbb{R}^N), d_p^{\mathrm{nd}})$ is a separable metric space.*

Proof. It is enough to consider $N = 2$. For a probability measure μ on $\mathbb{R}^2$ consider its disintegration measure

$$\mu(A \times B) = \textstyle\int_A \mu^{x_1}(B)\, p^1_* \mu(\mathrm{d}x_1),$$

where p^1 is the projection onto the first coordinate. An embedding of μ in the space $\mathcal{P}\big(\mathbb{R} \times \mathcal{P}(\mathbb{R})\big)$ is given by the probability measure generated uniquely by (here A, B are Borel sets of $\mathbb{R}$ and $\mathcal{P}(\mathbb{R})$ respectively)

$$I[\mu](A \times B) := \mu\big(A \cap T_\mu^{-1}(B)\big) = p^1_* \mu\big(A \cap T_\mu^{-1}(B)\big),$$

where T_μ is the Borel measurable function

$$\begin{aligned} T_\mu &: \mathbb{R} \to \mathcal{P}(\mathbb{R}) \\ &\quad x_1 \mapsto \mu^{x_1}(\mathrm{d}x_2). \end{aligned}$$

In this way we find that $I[\mu]$ is the μ-law of $x_1 \mapsto (x_1, \mu^{x_1}(\mathrm{d}x_2))$. For $\mu \in \mathcal{P}^p(\mathbb{R}^2)$ we also have

$$\begin{aligned} \int \left\{\underline{d}(x,0)^p + W_p^p(\nu, \delta_0)\right\} I[\mu](\mathrm{d}x, \mathrm{d}\nu) &= \int \left\{\underline{d}(x_1,0)^p + W_p^p(\mu^{x_1}, \delta_0)\right\} p_*^1\mu(\mathrm{d}x_1) \\ &= \int \left\{\underline{d}(x_1,0)^p + \underline{d}(x_2,0)^p\right\} \mu(\mathrm{d}x_1, \mathrm{d}x_2) < \infty \end{aligned}$$

and thus $I[\mu] \in \mathcal{P}^p(R_{1:2})$.

We now observe that the embedding $\mu \mapsto I[\mu]$ is actually an isometry between $(\mathcal{P}^p(\mathbb{R}^N), d_p^{\mathrm{nd}})$ and $(\mathcal{P}^p(R_{1:N}), W_{d_{1:N},p})$. To this end, first note that every coupling between $I[\mu]$ and $I[\nu]$ (i.e., every $\Gamma \in \Pi(I[\mu], I[\nu])$) is of the form $\bar\gamma(\mathrm{d}x_1, \mathrm{d}y_1)\, \delta_{T_\mu(x_1)}(\mathrm{d}M)\, \delta_{T_\nu(y_1)}(\mathrm{d}N)$ for some $\bar\gamma \in \Pi(p_*^1\mu, p_*^1\nu)$ and vice-versa. Hence from (4.1) and (3.2) we have that

$$\begin{aligned} W_{d_{1:2},p}\big(I[\mu], I[\nu]\big)^p &= \inf_{\bar\gamma \in \Pi(p_*^1\mu, p_*^1\nu)} \int \left\{\underline{d}(x_1,y_1)^p + W_p^p(\mu^{x_1}, \nu^{y_1})\right\} \bar\gamma(\mathrm{d}x_1, \mathrm{d}y_1) \\ &= d_p^{\mathrm{nd}}(\mu,\nu)^p, \end{aligned} \tag{4.3}$$

and hence the isometry by (3.3). Finally, since the image of I is a subspace of the separable metric space $(\mathcal{P}^p(R_{1:N}), W_{d_{1:N},p})$, it is separable itself. We conclude that $(\mathcal{P}^p(\mathbb{R}^N), d_p^{\mathrm{nd}})$ is separable too. □

Remark 4.7 (*Surjectivity*)**.** From the preceding arguments follows that the embedding I in (4.2) is onto if and only if $N = 1$.

Remark 4.8 (*Duality*)**.** Returning to Example 4.3 and applying (4.3) we find for $\mu, \nu \in \mathcal{P}^1(\mathbb{R}^2)$ that

$$\begin{aligned} d_1^{\mathrm{nd}}(\mu,\nu) &= W_{d_{1:2},1}\big(I[\mu], I[\nu]\big) \\ &= \sup\left\{\int F(x,\mu^x)\, p_*^1\mu(\mathrm{d}x) - \int F(x,\nu^x)\, p_*^1\nu(\mathrm{d}x)\right\}, \end{aligned}$$

where the supremum is among all (bounded) functions $F\colon \mathbb{R} \times \mathcal{P}^1(\mathbb{R}) \to \mathbb{R}$ with Lipschitz constant at most one (with respect to the metric $\underline{d} + W_1$). Indeed, this is a consequence of the Kantorovich–Rubinstein Theorem ([32, Theorem 1.14]) for the 1-Wasserstein metric on $\mathcal{P}^1(\mathbb{R} \times \mathcal{P}^1(\mathbb{R}))$. Similar results apply for $\mu, \nu \in \mathcal{P}^1(\mathbb{R}^N)$ by using $R_{1:N}$ instead.

4.2. The completion

In what follows we identify the space of nested distributions as the completion of the space $(\mathcal{P}^p(\mathbb{R}^N), d_p^{\mathrm{nd}})$. This result thus provides the natural link between these two separate mathematical objects.

Theorem 4.9 (*Completion*)**.** *The space* $\big(\mathcal{P}^p(R_{1:N}), W_{d_{1:N},p}\big)$ *of nested distributions is the completion of* $\big(\mathcal{P}^p(\mathbb{R}^N), d_p^{\mathrm{nd}}\big)$.

Proof. We provide an isometry J from $(\mathcal{P}^p(\mathbb{R}^N), d_p^{\mathrm{nd}})$ into $(\mathcal{P}^p(R_{1:N}), W_{d_{1:N},p})$ which has a dense range. We shall prove that $J := I$ defined in (4.2) does this task. This can be done for arbitrary N at notational costs, but already the case $N = 2$ is representative for the general situation. We thus assume $N = 2$ in what follows.

The set of convex combinations of Dirac measures is dense in $\mathcal{P}^p(R_{1:N})$ with respect to the metric $W_{d_{1:N},p}$. This is actually true for any Wasserstein metric (cf. [12]) and thus particularly for $W_{d_{1:2},p}$, which in itself is a Wasserstein metric (see also Example 4.3 for concreteness). So it is sufficient to prove that convex combinations of Dirac measures lie in the closure of the range of I.

Let $A := (a_1, \dots, a_k)$ be a k-tuple of points in $\mathbb{R}$ and $m_1, \dots, m_k$ be measures on the line with finite pth moment. Given weighs $\{\lambda_i\}_{i=1}^k$ we are interested in the measure

$$P(\mathrm{d}x, \mathrm{d}m) = \sum_{i=1}^{k} \lambda_i \, \delta_{(a_i, m_i)}(\mathrm{d}x, \mathrm{d}m)$$

over $R_{1:2}$. Now take any sequence $A^n := \{a_1^n, \dots, a_k^n\}$ such that componentwise $A^n \to A$ as $n \to \infty$ and, for each n fixed, all coordinates of A^n are distinct. We now define $\mu_n \in \mathcal{P}^p(\mathbb{R}^2)$ as the measure whose first marginal is $\sum_{i=1}^k \lambda_j \, \delta_{a_j^n}$ and such that $\mu_n(\mathrm{d}x_2 \mid x_1 = a_j^n) = m_j(\mathrm{d}x_2)$. It is elementary, and this is the main point of having made the a_j^n's distinct for a fixed n, that

$$I[\mu_n] = \sum \lambda_j \, \delta_{(a_j^n, m_j)}.$$

Consequently we get that $I[\mu_n] \to P$ with respect to $W_{d_{1:2},p}$ when $n \to \infty$, as desired. □

5. The weak nested topology

It was demonstrated in the previous section that

$$\left(\mathcal{P}^p(\mathbb{R}^N),\ d_p^{\mathrm{nd}}\right) \tag{5.1}$$

is not complete (Example 4.1); its completion was identified to be the space of nested distributions (Theorem 4.9).

At this point recall that the interval $(0, 1)$ is not complete as a subspace of $[0, 1]$, if equipped with the usual distance. Nonetheless the open interval $(0, 1)$ is Polish, as it is homeomorphic to the real line $\mathbb{R}$.

The situation for the space (5.1) is similar. In what follows we shall demonstrate that the topology of the space (5.1) is Polish, although the genuine distance d_p^{nd} does not reveal this property.

5.1. The weak nested topology

We introduce the space $\mathcal{P}(R_{1:N})$ just as we did for $\mathcal{P}^p(R_{1:N})$, but now denoting $R_{t-1:N} := \mathbb{R} \times \mathcal{P}(R_{t:N})$ at each step of the recursive definition and equipping $R_{t-1:N}$ with the product topology of Euclidean distance in the first component and the usual weak topology in the second one. Doing so, we conclude that $\mathcal{P}(R_{1:N})$ is a Polish space of measures on the likewise Polish space $R_{1:N}$. Inspired by the isometric embedding in Theorem 4.6, which we denoted I in (4.2), a mapping $I\colon \mathcal{P}(\mathbb{R}^N) \to \mathcal{P}(R_{1:N})$ can be obtained by direct generalization.

Definition 5.1 (*Weakly Nested Convergence*)**.** We say that a net $\{\mu_\alpha\}_\alpha$ in $\mathcal{P}(\mathbb{R}^N)$ converges *weakly nested* to $\mu \in \mathcal{P}(\mathbb{R}^N)$, if and only if $I[\mu_\alpha]$ converges weakly in $\mathcal{P}(R_{1:N})$ to $I[\mu]$. We call the corresponding topology *weak nested topology*.

By definition, the weak nested topology is the initial topology for the embedding map I.

Remark 5.2 (*Cf. Remark 7.13(iii) in [32]*)**.** Suppose that X is Polish and that ρ is a compatible *bounded* complete metric. Then the corresponding p-Wasserstein topology is precisely the weak topology. Likewise, we obtain that the weak nested topology is generated by the nested distance d_p^{nd} as soon as we choose $\underline{d}$ as a compatible bounded metric for the usual topology on $\mathbb{R}$. For instance, we may choose the metric $\underline{d}(a,b) = |a-b| \wedge 1$, i.e.,

$$d(x,y) = \sum_{t=1}^{N} |x_t - y_t| \wedge 1.$$

In this way we obtain that the weak nested topology coincides with a p-nested topology of the form we have already treated.

Although there are more direct ways to prove it, the previous remark implies the following:

Lemma 5.3. *The weak nested topology is separable and metrizable.*

5.2. The weak nested topology is Polish

We will establish that the weak nested topology (and actually the nested distance topologies) is Polish.

We recall that a set of a topological space is a G_δ if it is the countable intersection of open sets. Recall also that every separable metrizable space is homeomorphic to a subspace of the Hilbert cube $[0,1]^{\mathbb{N}}$, the latter equipped with the product topology; see [18, Theorem 4.14]. A compatible metric on the Hilbert cube is

$$D\big((x_n),(y_n)\big) := \sum_{n=1} 2^{-n}|x_n - y_n|.$$

Lemma 5.4. *Suppose that $m \in \mathcal{P}(X \times Y)$ with X Polish and (Y,ρ) a separable metric space. Let $\iota\colon Y \to [0,1]^{\mathbb{N}}$ denote the embedding of Y into the Hilbert cube. Then the following are equivalent:*

(i) $m\big(\mathrm{Graph}(f)\big) = 1$ *for* $f\colon X \to Y$ *Borel;*
(ii) $\inf\big\{\int_{X\times Y} \rho\big(f(x),y\big)\, m(\mathrm{d}x,\mathrm{d}y) :\ f\colon X \to Y \text{ Borel}\big\} = 0$*;*
(iii) $\inf\big\{\int_{X\times Y} D\big(F(x),\iota(y)\big)\, m(\mathrm{d}x,\mathrm{d}y) :\ F\colon X \to [0,1]^{\mathbb{N}} \text{ Borel}\big\} = 0$*;*
(iv) $\inf\big\{\int_{X\times Y} D\big(F(x),\iota(y)\big)\, m(\mathrm{d}x,\mathrm{d}y) :\ F\colon X \to [0,1]^{\mathbb{N}} \text{ continuous}\big\} = 0.$

Proof. Clearly (i) $\Longrightarrow$ (ii) $\Longrightarrow$ (iii). Denote μ the first marginal of m. Given $F\colon X \to [0,1]^{\mathbb{N}}$ Borel, $F(x) = (F_n(x))_n$, we can approximate it in $L^1(X,\mu;[0,1]^{\mathbb{N}})$ by continuous functions. This follows since coordinate-wise we can approximate $F_n \in L^1(X,\mu;[0,1])$ by continuous functions. So also (iii) $\Longrightarrow$ (iv).

To establish (iv) $\Longrightarrow$ (i) let $\{F_n\}$ be a sequence of continuous functions approximating the infimum in (iv) and denote $G_n(x) := \int D\big(F_n(x),\iota(y)\big)m^x(\mathrm{d}y)$ so G_n is Borel, non-negative and $\|G_n\|_{L^1(X,\mu;\mathbb{R})} \to 0$ by definition. It follows that $G_n \to 0$ in $L^1(X,\mu;\mathbb{R})$ so up to a subsequence $G_n(x) \to 0$ for μ-a.e. x. From now on we work on such a full measure set, on which we can further assume that $m^x \in \mathcal{P}(Y)$. Assume that we had that $|supp(m^x)| > 1$. Then there would exist disjoint compact sets K_x^1, $K_x^2 \subset Y$ with $M_x = \min\{m^x(K_x^1), m^x(K_x^2)\} > 0$.

Obviously $\iota(K_x^1), \iota(K_x^2)$ are also disjoint compact sets, so $D_x := D\big(\iota(K_x^1), \iota(K_x^2)\big) > 0$. By the triangle inequality, $\max\{D\big(F_n(x), \iota(K_x^1)\big), D\big(F_n(x), \iota(K_x^2)\big)\} \geq D_x/2$, thus $G_n(x) \geq M_x D_x/2$, yielding a contradiction. We conclude that μ-a.s. $|supp(m^x)| = 1$ and therefore we must have $\iota\big(f(x)\big) := \lim_n f_n(x)$ exists, for some $f\colon X \to Y$ Borel. Thus $m^x(\mathrm{d}y) = \delta_{f(x)}(\mathrm{d}y)$, $\mu - a.s.$, which proves (i). □

Observe that it is crucial for point (iv) in Lemma 5.4 to embed Y in the Hilbert cube. Indeed, if X is connected and Y discrete, then the only continuous functions $f\colon X \to Y$ are the constants. The following result is interesting in its own:

Proposition 5.5. *Let X and Y be Polish spaces. Then*

$$S := \{m \in \mathcal{P}(X \times Y)\colon m\big(\mathrm{Graph}(f)\big) = 1, \text{ some Borel } f\colon X \to Y\},$$

with the relative topology inherited from $\mathcal{P}(X \times Y)$, is Polish too.

Proof. Let ρ be a compatible metric for Y, which we may assume bounded. By Lemma 5.4 we have

$$S = \bigcap_{n\in\mathbb{N}} \bigcup_{\substack{F\colon X\to[0,1]^{\mathbb{N}},\\ F \text{ continuous}}} \left\{m \in \mathcal{P}(X \times Y)\colon \int D\big(F(x), \iota(y)\big)\, m(\mathrm{d}x, \mathrm{d}y) < \frac{1}{n}\right\}, \tag{5.2}$$

where $\iota\colon Y \to [0,1]^{\mathbb{N}}$ is an embedding. Since $(x, y) \mapsto D\big(F(x), \iota(y)\big)$ is continuous bounded if F is continuous, the set in curly brackets is open in the weak topology. Thus the union of these is open too and we get that S is a G_δ subset. We conclude by employing [18, Theorem 3.11]. □

Theorem 5.6. *The weak nested topology on $\mathcal{P}(\mathbb{R}^N)$ is Polish.*

Proof. For $N = 2$ we have $\mathcal{P}(R_{1:2}) = \mathcal{P}(\mathbb{R} \times \mathcal{P}(\mathbb{R}))$ and, by definition, $\mathcal{P}(\mathbb{R}^2)$ equipped with the weak nested topology is homeomorphic to $I[\mathcal{P}(\mathbb{R}^2)]$ equipped with the relative topology inherited from $\mathcal{P}(R_{1:2})$. We have

$$I[\mathcal{P}(\mathbb{R}^2)] = \left\{P \in \mathcal{P}(\mathbb{R} \times \mathcal{P}(\mathbb{R}))\colon \ P\big(\mathrm{Graph}(f)\big) = 1, \text{ some Borel } f\colon \mathbb{R} \to \mathcal{P}(\mathbb{R})\right\}.$$

To wit, if $P \in I[\mathcal{P}(\mathbb{R}^2)]$, then by definition of the embedding I we have $P = (\mathrm{id}, T)_*(p^1_*\mu)$ for some $\mu \in \mathcal{P}(\mathbb{R}^2)$ and $T(x) = \mu^x$ (see also the proof of Theorem 4.6). Taking $f = T$ we then get that P belongs to the right hand side above. Conversely, given P on the right hand side, we denote by μ_1 its first marginal and define $\mu^x(dy) := f(x)(dy)$. The measure $\mu(dx, dy) := \mu_1(dx)\mu^x(dy) \in \mathcal{P}(\mathbb{R}^2)$ satisfies $I[\mu] = P$.

By Proposition 5.5 we conclude that $I[\mathcal{P}(\mathbb{R}^2)]$ is Polish and then so is $\mathcal{P}(\mathbb{R}^2)$, as desired. The case for general N is identical; one observes by reverse induction that if $\mathcal{P}(R_{t:N})$ is Polish, then so is $\mathcal{P}(R_{t-1:N})$ using the above arguments. □

Remark 5.7. An immediate consequence of Theorem 5.6 is that the weak nested topology and the weak topology on $\mathcal{P}(\mathbb{R}^N)$ generate the same Borel sets. This follows from a result of Lusin and Suslin (cf. [18, Theorem 15.1]) concerning the measurability of the inverse of a continuous injective function between Polish spaces.

Proposition 5.5, and more specifically (5.2), permit to actually find a compatible complete metric for the weak nested topology. The embedding into the Hilbert cube would make such

metric look more complicated than necessary. As we argue now, there is a way to identify a slightly less abstract compatible complete metric. For simplicity of notation we just consider $N = 2$ here:

Corollary 5.8. *Let ρ be a bounded metric compatible with the weak topology on $\mathcal{P}(\mathbb{R})$ and d^w a complete metric compatible with the weak topology on $\mathcal{P}(\mathbb{R} \times \mathcal{P}(\mathbb{R}))$. Then the weak nested topology on $\mathcal{P}(\mathbb{R}^2)$ is generated by the complete metric*

$$d^{wnt}(P, Q) := d^w(I[P], I[Q]) + \sum_{n\in\mathbb{N}} 2^{-n} \wedge \left| \frac{1}{d^w(I[P], A_n)} - \frac{1}{d^w(I[Q], A_n)} \right|, \tag{5.3}$$

where

$$A_n := \{m \in \mathcal{P}\big(\mathbb{R} \times \mathcal{P}(\mathbb{R})\big) \colon \int \rho(F(x), y)\, m(\mathrm{d}x, \mathrm{d}y) \geq 1/n,\ \forall\, F \colon \mathbb{R} \to \mathcal{P}(\mathbb{R}) \text{ continuous}\},$$

with the embedding I as in (4.2) and A_n,

$$d^w(\cdot, A_n) := \inf_{m\in A_n} d^w(\cdot, m),$$

the distance to the set

Proof. We first observe that for Lemma 5.4 and $Y = \mathcal{P}(\mathbb{R})$, we may bypass the embedding into the Hilbert cube. One way to do this is to follow the "Tietze extension" argument in the proof of [13, Proposition C.1], establishing the equivalence of (i) and (iv) in Lemma 5.4 where now the continuous functions map from $X = \mathbb{R}$ to $\mathcal{P}(\mathbb{R})$. We can thus write (5.2), in the case $Y = \mathcal{P}(\mathbb{R})$, without the embedding ι. Using this and following the proof of [18, Theorem 311] we find a compatible complete metric for $I[\mathcal{P}(\mathbb{R}^2)]$ with the relative topology inherited from $\mathcal{P}(\mathbb{R} \times \mathcal{P}(\mathbb{R}))$, via

$$I[\mathcal{P}(\mathbb{R}^2)] \ni (\bar{P},\ \bar{Q}) \mapsto d^w(\bar{P}, \bar{Q}) + \sum_n 2^{-n} \wedge \left| d^w(\bar{P}, A_n)^{-1} - d^w(\bar{Q}, A_n)^{-1} \right|.$$

This is then transformed into a complete metric for $\mathcal{P}(\mathbb{R}^2)$ via the homeomorphism I, yielding (5.3). □

Remark 5.9. Notice that Example 4.1 shows that the weak nested topology is strictly stronger than the weak topology for $N \geq 2$. In this case, it also shows that even if a sequence of measures has their support contained in a common compact, there need not exist a convergent subsequence, unlike in the weak topology.

Analogous considerations show that $\mathcal{P}^p(R_{1:N})$ with the p-nested distance is Polish as well. Having established the completion and the Polish character of the p-nested distance, it remains an open question, whether there is a more amenable compatible complete metric than the one identified in Corollary 5.8.

6. Extreme points of related sets

Employing the notation as in the definition of bicausal transport plans (Definition 2.2) we say that $\gamma \in \Pi(\mu, \nu)$ is *causal* if the mappings

$$\mathbb{R}^N \ni x \mapsto \gamma^x(B)$$

are $\mathcal{F}_t$-measurable for any $B \in \mathcal{F}_t$, $t < N$. This is a weaker condition than *bi*causality. The set of such couplings is denoted

$$\Pi_c(\mu, \nu).$$

We will write $\Pi(\mu, \cdot)$ meaning that the second marginal of these couplings is left unspecified, with similar notation for the causal and bicausal case.

We are interested in determining the extreme points of the convex sets

$$\Pi_c(\mu, \cdot) \quad \text{and} \quad \Pi_{bc}(\mu, \cdot).$$

Such extreme points are expected to play an important role when one is interested in "simplifying" the process law μ without changing its information structure. To investigate this question let

$$\Pi_c^{\text{Monge}}(\mu, \cdot) := \left\{\gamma = (\text{id}, T)_*\mu \colon \ T : \mathbb{R}^N \to \mathbb{R}^N \text{ is Borel and adapted}\right\} \tag{6.1}$$

and

$$\Pi_{bc}^{\text{Monge}}(\mu, \cdot) := \left\{\gamma = (\text{id}, T)_*\mu \colon \ T : \mathbb{R}^N \to \mathbb{R}^N \text{ is } \begin{smallmatrix}\text{Borel, adapted and} \\ \mu\text{-a.s. invertible}\end{smallmatrix}\right\}, \tag{6.2}$$

where 'T is μ-a.s. invertible' means that there is a Borel adapted map $R : \mathbb{R}^N \to \mathbb{R}^N$ such that

$$R \circ T = \text{id } (\mu\text{-a.s.}) \quad \text{and} \quad T \circ R = \text{id } (T_*\mu\text{-a.s.}).$$

We recall that $T = (T_1, \dots, T_N)$ is *adapted* if $T_i(x_1, \dots, x_N) = T_i(x_1, \dots, x_i)$ for each i. Mappings having the properties specified in (6.2) have been called "isomorphism of filtered probability spaces" in the literature.

We use the notation ext and conv to denote the extreme points and the convex hull of a set. We can now state the main result of this section.

Theorem 6.1 (*Extreme Points*)**.** *It holds that*

$$\Pi_c^{Monge}(\mu, \cdot) = \text{ext } \Pi_c(\mu, \cdot),$$

and

$$\Pi_{bc}^{Monge}(\mu, \cdot) \subset \Pi_c^{Monge}(\mu, \cdot) \bigcap \Pi_{bc}(\mu, \cdot) = \text{ext } \Pi_{bc}(\mu, \cdot).$$

Proof. It is clear that $\Pi_c^{\text{Monge}}(\mu, \cdot) \subset \Pi_c(\mu, \cdot)$. From this one sees that $\Pi_c^{\text{Monge}}(\mu, \cdot) \subset \text{ext } \Pi_c(\mu, \cdot)$, since a coupling supported on the graph of a function cannot arise as the combination of two couplings without this property. Similarly, we have $\Pi_{bc}^{\text{Monge}}(\mu, \cdot) \subset \Pi_c^{\text{Monge}}(\mu, \cdot) \bigcap \Pi_{bc}(\mu, \cdot) \subset \text{ext } \Pi_{bc}(\mu, \cdot)$.

We first prove that $\Pi_c^{\text{Monge}}(\mu, \cdot) \supset \text{ext } \Pi_c(\mu, \cdot)$. It is easy to see from [5, Proposition 2.4], especially part 4. therein, that $\gamma \in \Pi_c(\mu, \cdot)$ is equivalent to

$$\int F \, \mathrm{d}\gamma = 0$$

for all F of either of the two following forms:

(i) $F = \phi(x_1, \dots, x_N) - \int \phi(\bar{x}_1, \dots, \bar{x}_N)\, \mu(\mathrm{d}\bar{x}_1, \dots, \mathrm{d}\bar{x}_N)$, with $\phi =: \phi^x$ bounded measurable,

(ii) $F = \sum_{t<N} H_t^y[M_{t+1}^x - M_t^x]$, with H^y a bounded continuous process adapted to the y-variables, M^x a bounded μ-martingale adapted to the x-variables (namely $H_t^y = H_t(y_1, \dots, y_t)$ and $M_t^x = M_t(x_1, \dots, x_t)$).

We consider now the vector space V generated by the constant 1 and the functions on $\mathbb{R}^N \times \mathbb{R}^N$ of the form (i) and (ii). Explicitly, we have

$$V = \left\{c + \phi^x + \textstyle\sum_t H_t^y[M_{t+1}^x - M_t^x] \colon \text{ with } \phi^x, H^y, M^x \text{ as described in (i)–(ii)}, c \in \mathbb{R}\right\}.$$

By Douglas' Theorem [30, Ch. V, (4.4)] we have that $\gamma \in \operatorname{ext} \Pi_c(\mu, \cdot)$ is equivalent to V being dense in $L^1(\gamma)$. We take γ such an extreme point, an arbitrary $i \in \{1, \dots, N\}$ and h a Borel bounded function. We will show that $h(y_i) = \int h(\bar{y}_i)\gamma^{x_1,\dots,x_i}(\mathrm{d}\bar{y}_i)$ holds γ-a.s. This would immediately imply the existence of measurable functions T^i such that for all $i \colon y_i = T^i(x_1, \dots, x_i)$ holds γ-a.s. This then implies $\gamma \in \Pi_c^{\text{Monge}}(\mu, \cdot)$.

We start by observing that

$$\int h(y_i)\phi^x \,\mathrm{d}\gamma = \int [h(\bar{y}_i)\gamma^{x_1,\dots,x_N}(\mathrm{d}\bar{y}_i)]\,\phi^x \,\mathrm{d}\mu = \int [h(\bar{y}_i)\gamma^{x_1,\dots,x_i}(\mathrm{d}\bar{y}_i)]\,\phi^x \,\mathrm{d}\mu,$$

since by causality $x_{i+1}, \dots, x_N$ are γ-independent of y_i given $x_1, \dots, x_i$. Now we prove the desired

$$h(y_i) = \int h(\bar{y}_i)\gamma^{x_1,\dots,x_i}(\mathrm{d}\bar{y}_i),$$

which we do by induction in i. For $i = 1$, we get

$$\int h(y_1)H_t^y[M_{t+1}^x - M_t^x]\mathrm{d}\gamma = \int [h(\bar{y}_1)\gamma^{x_1}(\mathrm{d}\bar{y}_i)]\, H_t^y[M_{t+1}^x - M_t^x]\mathrm{d}\gamma,$$

since by the same conditional independence argument both sides are equal to 0 (indeed, M^x must be by causality a martingale in the filtration of the x and y variables). It follows that

$$\int \left\{h(y_1) - \int h(\bar{y}_1)\gamma^{x_1}(\mathrm{d}\bar{y}_i)\right\} v(x_1, \dots, x_N, y_1, \dots, y_N)\,\mathrm{d}\gamma = 0, \ \forall v \in V.$$

Since V is dense in $L^1(\gamma)$, we obtain the claim for $i = 1$. Now let us suppose this has been proved for all indices $i \le j$. In order to establish the result for $j + 1$, the key is to prove that

$$\int h(y_{j+1})H_t^y[M_{t+1}^x - M_t^x]\mathrm{d}\gamma = \int \left[h(\bar{y}_{j+1})\gamma^{x_1,\dots,x_{j+1}}(\mathrm{d}\bar{y}_{j+1})\right] H_t^y[M_{t+1}^x - M_t^x]\mathrm{d}\gamma.$$

But this is true by the same argument as above if $t > j$ (one verifies that both sides are equal to 0). In case $t \le j$, by induction we have that $H_t^y = \tilde{H}_t(x_1, \dots, x_t)$ γ-a.s. so we obtain that the left hand side is equal to $\int \left[h(\bar{y}_{j+1})\gamma^{x_1,\dots,x_N}(\mathrm{d}\bar{y}_{j+1})\right] H_t^y[M_{t+1}^x - M_t^x]\mathrm{d}\gamma$, and this is the equal to the right hand side by causality.

We now prove $\Pi_c^{\text{Monge}}(\mu, \cdot) \supset \operatorname{ext} \Pi_{bc}(\mu, \cdot)$. Here V must be replaced by

$$\tilde{V} = \left\{\ c + \phi^x + \textstyle\sum_t H_t^y[M_{t+1}^x - M_t^x] + \sum_t G_t^x[N_{t+1}^y - N_t^y]\ \right\},$$

with the obvious extension of the notation used so far. By essentially the same arguments as above one obtains that for $\gamma \in \operatorname{ext} \Pi_{bc}(\mu, \cdot)$ we have $y_t = T^t(x_1, \dots, x_t)$ γ-a.s. Indeed the only thing that need be observed, is that under γ any martingale with respect to the y-filtration remains a martingale if we adjoin the x-filtration. This shows $\Pi_c^{\text{Monge}}(\mu, \cdot) \supset \operatorname{ext} \Pi_{bc}(\mu, \cdot)$ and hence $\Pi_c^{\text{Monge}}(\mu, \cdot) \bigcap \Pi_{bc}(\mu, \cdot) = \operatorname{ext} \Pi_{bc}(\mu, \cdot)$. □

Remark 6.2. The inclusion in Theorem 6.1 is strict unless μ is concentrated in a single point. Indeed, for $a \in \mathbb{R}^N$ denote $T^a(x) := a$ and observe that $\gamma_a := (\mathrm{id}, T^a)_*\mu \in \Pi_c^{\text{Monge}}(\mu, \cdot) \cap \Pi_{bc}(\mu, \cdot)$. However, $\gamma_a \in \Pi_{bc}^{\text{Monge}}(\mu, \cdot)$ if and only if μ is concentrated in a point.

7. Summary

This article investigates fundamental topological and metric properties of the nested distance. In contrast to classical Wasserstein distances, for example, its topology cannot be characterized via integration on test functions, so that complete stochastic programs appear as the natural distinguishing element of the topology induced by the nested distance. The nested distance is also not complete, which again is in contrast to the classical Wasserstein distance.

We obtain two main results. First, we demonstrate that the metric completion of the nested distance is the space of nested distributions with their classical Wasserstein metric, as introduced in [21]. This provides a connection between two hitherto unrelated mathematical objects. Second, we established that the topology generated by the nested distance is Polish, which we hope opens the way for future applications. Along these lines, this article starts the study of extreme points of sets of measures relevant for stochastic optimization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] B. Acciaio, J. Backhoff Veraguas, A. Zalashko, Causal optimal transport and its links to enlargement of filtrations and continuous-time stochastic optimization, 2016, ArXiv e-prints.
[2] D.J. Aldous, Weak Convergence and General Theory of Processes, Unpublished draft of monograph, Department of Statistics, University of California, Berkeley, CA 94720, 1981.
[3] L. Ambrosio, N. Gigli, A user's guide to optimal transport, in: Modelling and Optimisation of Flows on Networks, in: Lecture Notes in Math., vol. 2062, Springer, Heidelberg, 2013, pp. 1–155, http://dx.doi.org/10.1007/978-3-642-32160-3_1.
[4] L. Ambrosio, N. Gigli, G. Savaré, Gradient Flows in Metric Spaces and in the Space of Probability Measures, second ed., Birkhäuser Verlag, Basel, Switzerland, 2005, http://dx.doi.org/10.1007/978-3-7643-8722-8,
[5] J. Backhoff, M. Beiglböck, Y. Lin, A. Zalashko, Causal transport in discrete time and applications, SIAM J. Optim. (ISSN: 1052-6234) 27 (4) (2017) 2528–2562, http://dx.doi.org/10.1137/16M1080197.
[6] J. Backhoff-Veraguas, D. Bartl, M. Beiglböck, M. Eder, Adapted Wasserstein distances and stability in mathematical finance, 2019, arXiv e-prints, arXiv:1901.07450.
[7] J. Backhoff-Veraguas, D. Bartl, M. Beiglböck, M. Eder, All adapted topologies are equal, 2019, arXiv e-prints, arXiv:1905.00368.
[8] M. Barbie, A. Gupta, The topology of information on the space of probability measures over Polish spaces, J. Math. Econom. (ISSN: 0304-4068) 52 (2014) 98–111, http://dx.doi.org/10.1016/j.jmateco.2014.04.003.
[9] M. Beiglböck, A. Cox, M. Huesmann, Optimal transport and Skorokhod embedding, Invent. Math. 208 (2) (2017) 327–400.
[10] M. Beiglböck, P. Henry-Labordère, F. Penkner, Model-independent bounds for option prices—a mass transport approach, Finance Stoch. (ISSN: 0949-2984) 17 (3) (2013) 477–501.
[11] J. Bion-Nadal, D. Talay, On a Wasserstein-type distance between solutions to stochastic differential equations, Ann. Appl. Probab. (ISSN: 1050-5164) 29 (3) (2019) 1609–1639, http://dx.doi.org/10.1214/18-AAP1423.
[12] F. Bolley, Separability and completeness for the Wasserstein distance, in: C. Donati-Martin, M. Émery, A. Rouault, C. Stricker (Eds.), Séminaire de Probabilités XLI, in: Lecture Notes in Mathematics, vol. 1934, Springer, Berlin, Heidelberg, 2008, pp. 371–377, http://dx.doi.org/10.1007/978-3-540-77913-1.
[13] R. Carmona, F. Delarue, D. Lacker, Mean field games with common noise, Ann. Probab. (ISSN: 0091-1798) 44 (6) (2016) 3740–3803, http://dx.doi.org/10.1214/15-AOP1060.
[14] Y. Dolinsky, H.M. Soner, Martingale optimal transport and robust hedging in continuous time, Probab. Theory Related Fields (ISSN: 0178-8051) 160 (1–2) (2014) 391–427, http://dx.doi.org/10.1007/s00440-013-0531-y.
[15] A. Galichon, P. Henry-Labordère, N. Touzi, A stochastic control approach to no-arbitrage bounds given marginals, with an application to lookback options, Ann. Appl. Probab. (ISSN: 1050-5164) 24 (1) (2014) 312–336, http://dx.doi.org/10.1214/13-AAP925.

[16] M.F. Hellwig, Sequential decisions under uncertainty and the maximum theorem, J. Math. Econom. (ISSN: 0304-4068) 25 (4) (1996) 443–464, http://dx.doi.org/10.1016/0304-4068(95)00739-3.
[17] D.N. Hoover, H.J. Keisler, Adapted probability distributions, Trans. Amer. Math. Soc. 286 (1) (1984) 159–201.
[18] A.S. Kechris, Classical Descriptive Set Theory, in: Graduate Texts in Mathematics, vol. 156, Springer-Verlag, New York, ISBN: 0-387-94374-9, 1995, p. xviii+402, http://dx.doi.org/10.1007/978-1-4612-4190-4.
[19] R. Lassalle, Causal transference plans and their Monge–Kantorovich problems, 2013, URL http://arxiv.org/pdf/1303.6925.
[20] M. Nutz, F. Stebegg, Canonical supermartingale couplings, Ann. Probab. 46 (6) (2018) 3351–3398.
[21] G.C. Pflug, Version-independence and nested distributions in multistage stochastic optimization, SIAM J. Optim. 20 (2009) 1406–1420, http://dx.doi.org/10.1137/080718401.
[22] G.C. Pflug, A. Pichler, A distance for multistage stochastic optimization models, SIAM J. Optim. 22 (1) (2012) 1–23, http://dx.doi.org/10.1137/110825054.
[23] G.C. Pflug, A. Pichler, Multistage Stochastic Optimization, in: Springer Series in Operations Research and Financial Engineering, Springer, ISBN: 978-3-319-08842-6, 2014, http://dx.doi.org/10.1007/978-3-319-08843-3, URL https://books.google.com/books?id=q_VWBQAAQBAJ.
[24] G.C. Pflug, A. Pichler, Dynamic generation of scenario trees, Comput. Optim. Appl. 62 (3) (2015) 641–668, http://dx.doi.org/10.1007/s10589-015-9758-0.
[25] G.C. Pflug, A. Pichler, From empirical observations to tree models for stochastic optimization: convergence properties, SIAM J. Control Optim. 26 (3) (2016) 1715–1740, http://dx.doi.org/10.1137/15M1043376.
[26] A. Pichler, R. Schlotter, Martingale characterizations of risk-averse stochastic optimization problems, Math. Program. (ISSN: 1436-4646) (2019) http://dx.doi.org/10.1007/s10107-019-01391-2.
[27] A. Pichler, A. Shapiro, Risk averse stochastic programming: time consistency and optimal stopping, 2018, arXiv e-prints, arXiv:1808.10807.
[28] S.T. Rachev, L. Rüschendorf, Mass Transportation Problems. Vol. I, in: Probability and its Applications (New York), Springer-Verlag, New York, ISBN: 0-387-98350-3, 1998, p. xxvi+508.
[29] S.T. Rachev, L. Rüschendorf, Mass Transportation Problems. Vol. II, in: Probability and its Applications (New York), Springer-Verlag, New York, ISBN: 0-387-98352-X, 1998, p. xxvi+430, Applications.
[30] D. Revuz, M. Yor, Continuous Martingales and Brownian Motion, third ed., in: Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences], vol. 293, Springer-Verlag, Berlin, ISBN: 3-540-64325-7, 1999, p. xiv+602.
[31] F. Santambrogio, Optimal Transport for Applied Mathematicians, in: Progress in Nonlinear Differential Equations and their Applications, vol. 87, Birkhäuser/Springer, Cham, 2015, p. xxvii+353, http://dx.doi.org/10.1007/978-3-319-20828-2, ISBN: 978-3-319-20827-5; 978-3-319-20828-2, Calculus of variations, PDEs, and modeling.
[32] C. Villani, Topics in Optimal Transportation, in: Graduate Studies in Mathematics, vol. 58, American Mathematical Society, Providence, RI, ISBN: 0-821-83312-X, 2003, http://dx.doi.org/10.1090/gsm/058, URL http://books.google.com/books?id=GqRXYFxe0l0C.
[33] C. Villani, Optimal transport. Old and new, in: Grundlehren der mathematischen Wissenschaften, vol. 338, Springer, 2009.