JEMS

Daniel Bartl · Mathias Beiglböck · Gudmund Pammer

The Wasserstein space of stochastic processes

Abstract. Wasserstein distance induces a natural Riemannian structure for the probabilities on the Euclidean space. This insight of classical transport theory is fundamental for tremendous applications in various fields of pure and applied mathematics. We believe that an appropriate probabilistic variant, the *adapted Wasserstein distance* $\mathcal{AW}$, can play a similar role for the class FP of *filtered processes*, i.e., stochastic processes together with a filtration. In contrast to other topologies for stochastic processes, probabilistic operations such as the Doob decomposition, optimal stopping and stochastic control are continuous with respect to $\mathcal{AW}$. We also show that (FP, $\mathcal{AW}$) is a geodesic space, isometric to a classical Wasserstein space, and that martingales form a closed geodesically convex subspace.

Keywords: Aldous' extended weak topology, Hoover and Keisler's adapted distribution, Hellwig's information topology, Pflug–Pichler's nested distance, geodesic space, barycenter of stochastic processes.

1. Overview

It is often useful to change the view from considering objects in isolation to studying the space of these objects and specifically their mutual relationship with respect to the *ambient space*. A classic instance is to switch from studying functions to considering the Lebesgue- or Sobolev-spaces of functions in functional analysis; a more contemporary example is the passing from measures to the manifold of measures based on optimal transport theory. The aim of this article is to investigate what should be the appropriate ambient space for the class of stochastic processes.

Daniel Bartl: Faculty of Mathematics, University of Vienna, Oskar-Morgenstern-Platz 1, 1090 Vienna, Austria; daniel.bartl@univie.ac.at

Mathias Beiglböck: Faculty of Mathematics, University of Vienna, Oskar-Morgenstern-Platz 1, 1090 Vienna, Austria; mathias.beiglboeck@univie.ac.at

Gudmund Pammer: Department of Mathematics, ETH Zürich, Rämistrasse 101, 8092 Zürich, Switzerland; Institute of Statistics, Graz University of Technology, Kopernikusgasse 24/III, 8010 Graz, Austria; gudmund.pammer@math.ethz.ch, gudmund.pammer@tugraz.at

1.1. The space of laws on $\mathbb{R}^N$ and its limitations

A natural starting point for this is to represent stochastic processes on the canonical probability space. Specifically, the class of real-valued processes in finite discrete time $\{1, \ldots, N\}$ is naturally represented as the set $\mathcal{P}(\mathbb{R}^N)$ of probability measures on $\mathbb{R}^N$. Importantly, the usual weak topology on $\mathcal{P}(\mathbb{R}^N)$ fails to capture the temporal structure of stochastic processes and is not strong enough to guarantee the continuity of stochastic optimization problems or basic operations like the Doob decomposition.

As a remedy, a number of researchers from different scientific communities have introduced topological structures on the set of stochastic processes with the common goal to adequately capture the temporal structure. We list Aldous' extended weak topology [9] (stochastic analysis), Hellwig's information topology [57, 58] (economics), Bion-Nadal–Talay's version of the Wasserstein distance [35] (stochastic analysis, optimal control), Pflug–Pichler's nested distance [80, 81] (stochastic optimization), Rüschendorf's Markov constructions [89] (optimal transport), Lassalle's causal transport problem [72] (optimal transport), and Nielsen–Sun's chain rule transport [77] (machine learning). Remarkably, in finite discrete time, these seemingly independent approaches define the same topology on $\mathcal{P}(\mathbb{R}^N)$, the *weak adapted topology*, see [14].

A natural compatible metric[1] for the weak adapted topology is the *adapted Wasserstein distance*

$$\mathcal{A}\mathcal{W}_p^p(\mu, \nu) = \inf_{\pi \in \mathrm{Cpl}_{bc}(\mu, \nu)} \mathbb{E}_\pi[\|X - Y\|_p^p], \quad p \in [1, \infty).$$

The difference to the classical Wasserstein distance comes from the fact that one considers only *bicausal* couplings $\mathrm{Cpl}_{bc}(\mu, \nu)$, see Definition 2.1 below. These couplings are non-anticipative and can be viewed as Kantorovich analogues of non-anticipative transport maps, see [34].

While the weak adapted topology/$\mathcal{A}\mathcal{W}_p$ appears canonical and has recently seen a burst of applications (see [3, 11, 13, 43, 52, 64, 69, 76, 82, 83, 85–87, 95, 96] among others), we also highlight two limitations:

(1) The metric space $(\mathcal{P}_p(\mathbb{R}^N), \mathcal{A}\mathcal{W}_p)$ is not complete. In fact, this shortcoming also arises for other natural distances that respect the information structure of stochastic processes.

(2) Following the classical theory of stochastic analysis, one would like to consider processes together with a general filtration, not just the filtration generated by the process itself.

1.2. Filtered processes as the completion of $\mathcal{P}_p(\mathbb{R}^N)$

Rather conveniently, these supposed shortcomings already represent their mutual resolution: a possible interpretation of the incompleteness of $(\mathcal{P}_p(\mathbb{R}^N), \mathcal{A}\mathcal{W}_p)$ is that the

---

[1]Precisely, convergence in $\mathcal{A}\mathcal{W}_p$ is equivalent to convergence in the weak adapted topology plus convergence of the p-th moment, see also [14].

space $\mathcal{P}_p(\mathbb{R}^N)$ is not 'large' enough to represent all processes one would like to consider. In our first main result, we show that the extra information that can be stored in an ambient filtration is precisely what is needed to arrive at the completion of $(\mathcal{P}_p(\mathbb{R}^N), \mathcal{A}\mathcal{W}_p)$. To make this precise, we need the following definition.

Definition 1.1. A five-tuple

$$\mathbb{X} := (\Omega, \mathcal{F}, \mathbb{P}, (\mathcal{F}_t)_{t=1}^N, (X_t)_{t=1}^N),$$

where $(X_t)_{t=1}^N$ is adapted to $(\mathcal{F}_t)_{t=1}^N$, is called a *filtered (stochastic) process*. We write $\mathcal{FP}$ for the class of all filtered processes and $\mathcal{FP}_p$ for the subclass of processes with $\mathbb{E}[\|X\|_p^p] < \infty$.

Clearly, $\mathcal{P}_p(\mathbb{R}^N)$ is embedded in $\mathcal{FP}_p$: for $\mu \in \mathcal{P}(\mathbb{R}^N)$, set

$$\mathbb{X} := (\mathbb{R}^N, \mathcal{B}(\mathbb{R}^N), \mu, (\mathcal{F}_t)_{t=1}^N, (X_t)_{t=1}^N),$$

where $(X_t)_{t=1}^N$ is the canonical process, i.e., $X_t(\omega) = \omega(t)$, and $\mathcal{F}_t = \sigma(X_s : s \leq t)$ for $t \leq N$.

It is relatively straightforward to extend the concept of bicausal couplings to processes with filtrations (see Definition 2.1 for details) and accordingly the notion of adapted Wasserstein distance extends to filtered processes via

$$\mathcal{A}\mathcal{W}_p^p(\mathbb{X}, \mathbb{Y}) := \inf_{\pi \in \mathrm{Cpl}_{\mathrm{bc}}(\mathbb{X}, \mathbb{Y})} \mathbb{E}_\pi[\|X - Y\|_p^p].$$

As in the case of $\mathcal{L}^p/L^p$ spaces (or similar situations), we identify two filtered processes $\mathbb{X}, \mathbb{Y}$ if $\mathcal{A}\mathcal{W}_p(\mathbb{X}, \mathbb{Y}) = 0$ and denote the corresponding set of equivalence classes by FP_p. Our first main result is the following.

Theorem 1.2. *The adapted Wasserstein distance $\mathcal{A}\mathcal{W}_p$ is a metric on FP_p. Moreover, $(\mathrm{FP}_p, \mathcal{A}\mathcal{W}_p)$ is the completion of $(\mathcal{P}_p(\mathbb{R}^N), \mathcal{A}\mathcal{W}_p)$.*

We also show in Theorem 5.4 below that certain simpler classes of processes are dense in FP_p, e.g., filtered processes that can be represented on a finite state space Ω or finite state Markov chains. This seems important in view of numerical applications.

1.3. FP_p as geodesic space

In fact, our proof of Theorem 1.2 reveals more, namely the following result on the metric structure of $(\mathrm{FP}_p, \mathcal{A}\mathcal{W}_p)$.

Theorem 1.3. *There exists a Polish space $(\mathcal{V}, d)$ such that $(\mathrm{FP}_p, \mathcal{A}\mathcal{W}_p)$ is isometric to the classical Wasserstein space $(\mathcal{P}_p(\mathcal{V}), \mathcal{W}_p^{(d)})$.*

Explicitly, $\mathcal{V}$ is constructed by considering Wasserstein spaces of Wasserstein spaces in an iterated fashion as already considered by Vershik [91]. Theorem 1.3 allows us to transfer concepts from optimal transport to the theory of stochastic processes, e.g.,
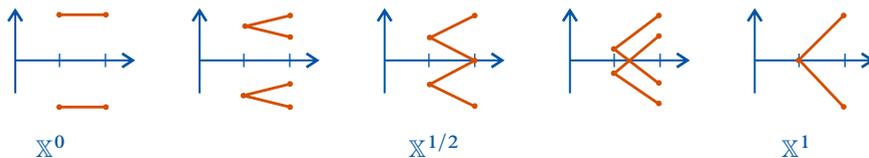
Fig. 1. Interpolation between two simple martingales $\mathbb{X}^0$ and $\mathbb{X}^1$.

it allows us to consider displacement interpolation (see, e.g., Figure 1) and Wasserstein barycenters of filtered processes and to view $(\mathrm{FP}_p, \mathcal{A}\mathcal{W}_p)$ as a (formal) Riemannian manifold. Specifically, using the work of Lisini [74], we obtain the following.

Theorem 1.4. *Assume $p > 1$. Then $(\mathrm{FP}_p, \mathcal{A}\mathcal{W}_p)$ is a geodesic space and the set of martingales forms a closed geodesically convex subspace.*

Famously, McCann introduced the concept of displacement interpolation in his PhD thesis [75], giving a new meaning to the transformation of one probability into another. By analogy, Theorem 1.4 suggests an interpolation between stochastic processes.

We emphasize that the usual Wasserstein interpolation on $\mathcal{P}_p(\mathbb{R}^N)$ is not compatible with concepts one would like to consider for stochastic processes, e.g., stochastic optimization problems are not continuous along geodesics, the set of martingales is not displacement convex, etc.

We also note that the set $\mathcal{P}_p(\mathbb{R}^N)$ is not $\mathcal{A}\mathcal{W}_p$-displacement convex: even if $\mathbb{P}, \mathbb{Q} \in \mathcal{P}_p(\mathbb{R}^N)$ are laws of relatively regular processes, the respective geodesic does in general not lie in $\mathcal{P}_p(\mathbb{R}^N)$, see Example 5.11. This further underlines the importance of considering processes together with their filtration.

1.4. Equivalence of filtered processes

We briefly discuss the equivalence relation induced by

$$\mathcal{A}\mathcal{W}_p(\mathbb{X}, \mathbb{Y}) = 0.$$

Intuitively, one would hope that processes with zero distance are equivalent in the sense that they have identical properties from a probabilistic perspective. In fact, based on formalizing what assertions belong to the 'language of probability', Hoover and Keisler [62] have made precise what it should mean that two processes $\mathbb{X}, \mathbb{Y}$ have the *same probabilistic properties*. Then $\mathbb{X}$ and $\mathbb{Y}$ are called *equivalent in adapted distribution*, in signs $\mathbb{X} \sim_\infty \mathbb{Y}$. We will establish below that equivalence in adapted distribution can be expressed in terms of adapted Wasserstein distance.

Theorem 1.5. *Let $\mathbb{X}, \mathbb{Y} \in \mathcal{F}\mathcal{P}_p$. Then $\mathbb{X} \sim_\infty \mathbb{Y}$ if and only if $\mathcal{A}\mathcal{W}_p(\mathbb{X}, \mathbb{Y}) = 0$.*

Informally, Theorem 1.5 asserts that the equivalence classes in FP_p collect precisely all representatives of a process that should be considered identical from a probabilistic point of view.

Other (more familiar) notions of equivalence on $\mathcal{F}\mathcal{P}$ are *equivalence in law*, in symbols $\sim_0$, and Aldous' notion of *synonymity*, in symbols $\sim_1$. Both of these are strictly coarser than $\sim_\infty$ and may identify processes that have different probabilistic properties. For example, there are filtered processes $\mathbb{X}$, $\mathbb{Y}$ with $\mathbb{X} \sim_0 \mathbb{Y}$, where $\mathbb{X}$ is a martingale while $\mathbb{Y}$ is not. Similarly, we will construct examples of processes $\mathbb{X}$, $\mathbb{Y}$ where $\mathbb{X} \sim_1 \mathbb{Y}$ but optimal stopping problems written on $\mathbb{X}$, $\mathbb{Y}$ lead to different results (see Theorem 7.1).

1.5. Continuity of Doob decomposition, optimal stopping, Snell envelope

In line with Theorem 1.5, $\mathcal{A}\mathcal{W}_p(\mathbb{X}, \mathbb{Y}) = 0$ implies that $\mathbb{X}$ and $\mathbb{Y}$ have the same Doob decomposition that optimal stopping problems of the form

$$\sup_\tau \mathbb{E}[G_\tau(X_1, \ldots, X_\tau)], \tag{1.1}$$

where τ runs through all $(\mathcal{F}_t)_{t=1}^N$-stopping times, yield the same optimal value and have the same Snell envelop. Moreover, we show that the above operations are continuous with respect to the weak adapted topology, indeed we establish the following.

Theorem 1.6. *The mapping that assigns to a filtered process its Doob decomposition is Lipschitz continuous. If $G_t \colon \mathbb{R}^t \to \mathbb{R}$ is bounded and continuous (resp. Lipschitz) for each t, then* (1.1) *is continuous (resp. Lipschitz) in $\mathbb{X}$.*

In Section 6, we collect further statements of similar flavour as Theorem 1.6. It is important to note that comparable results do not hold with respect to other (coarser) topologies for filtered processes. Specifically, convergence in Aldous' extended weak topology is strictly weaker than convergence in $\mathcal{A}\mathcal{W}_p$ and is not strong enough to obtain continuity of optimal stopping problems (see Section 7). This seems remarkable since Aldous [9, p. 105] deliberates the question of which framework is natural to study the continuity of optimal stopping.

1.6. Canonical representatives of filtered processes

A slightly altered variant $\mathcal{Z}$ of the Polish space $\mathcal{V}$ appearing in Theorem 1.3 also plays an important role in finding canonical representatives for the equivalence classes in FP. In Section 3 below, we will show that there exists $(\mathcal{Z}, \mathcal{F}^{\mathcal{Z}}, (\mathcal{F}_t^{\mathcal{Z}})_{t=1}^N, (Z_t)_{t=1}^N)$ with $\mathcal{Z}$ Polish, such that every filtered process $\mathbb{X}$ is represented in a canonical way via a probability on $\mathcal{Z}$, i.e., there exists $\mathbb{Q}^{\mathbb{X}} \in \mathcal{P}(\mathcal{Z})$ such that

$$(\mathcal{Z}, \mathcal{F}^{\mathcal{Z}}, \mathbb{Q}^{\mathbb{X}}, (\mathcal{F}_t^{\mathcal{Z}})_{t=1}^N, (Z_t)_{t=1}^N) \sim_\infty \mathbb{X}.$$

In particular, all information about the process $\mathbb{X}$ is stored in the corresponding measure $\mathbb{Q}^{\mathbb{X}}$, while the underlying probability space Ω, the representing stochastic process Z and the respective filtration do not depend on $\mathbb{X}$. In this sense, the situation is analogous to the canonical representation of stochastic processes via probabilities on the path space. In view of Theorem 1.5, this also implies that one can assume without loss of generality that a given filtered process is defined on a Polish probability space.

1.7. *Prokhorov-type result and barycenter of processes*

An extremely useful property of the usual weak topology is the abundance of (pre-)compact sets based on Prokhorov's theorem. Remarkably, this carries over to 'adapted' topologies. This was first established by Hoover [61, Theorem 4.3], see also [13, Lemma 1.7]. In the present context, this fact can be expressed as follows.

Theorem 1.7. *A set $K \subseteq \mathrm{FP}_p$ is $\mathcal{AW}_p$-precompact if and only if the respective set of laws in $\mathcal{P}_p(\mathbb{R}^N)$ is $\mathcal{W}_p$-precompact.*

Note that by Prokhorov's theorem, $\mathcal{W}_p$-precompactness in $\mathcal{P}_p(\mathbb{R}^N)$ is equivalent to tightness plus uniform p-integrability, see, for instance, [94].

Theorem 1.7 is relevant in several proofs given below and has important consequences for the applications of our results presented in Section 6. For instance, it allows us to establish the existence of barycenters of stochastic processes. Famously, Agueh and Carlier [6] introduced the concept of barycenters with respect to Wasserstein distance which has striking consequences in machine learning (e.g., [42,88]), statistics (e.g., [18,78]) as well as in pure mathematics (e.g., [68,73]). In Theorem 6.7, we show that for filtered processes $\mathbb{X}^1, \ldots, \mathbb{X}^k \in \mathrm{FP}_p$ and convex weights $\lambda_1, \ldots, \lambda_k$ there exists a *barycenter process* (see, e.g., Figure 2), i.e., a filtered process $\mathbb{X}^* \in \mathrm{FP}_p$ which minimizes

$$\inf_{\mathbb{X}} \lambda_1 \mathcal{AW}_p^p(\mathbb{X}^1, \mathbb{X}) + \cdots + \lambda_k \mathcal{AW}_p^p(\mathbb{X}^k, \mathbb{X}).$$

1.8. *Applications and extensions*

As already noted above, the adapted Wasserstein distance improves over the classical weak topology/Wasserstein distance in that it guarantees stability of basic operations such as the Doob decomposition and optimal stopping. Naturally, we expect similar results for other probabilistic problems with inherent time structure. In this line, we describe applications to stability of stochastic optimal control, utility maximization and pricing/hedging, robust finance in the realm of American options, conditional McKean–Vlasov control, and weak optimal transport, see Sections 6.1–6.8 below. In view of applications, it is relevant that adapted Wasserstein distance can be efficiently computed numerically as well as estimated from given data; we comment on this in Section 6.9.

While the focus of the present article lies on stochastic processes in finite discrete time, extensions to more general cases are intriguing. In Appendix B, we consider the case of infinite discrete time, i.e., the set $\mathrm{FP}^{(\infty)}$ of processes $(X_t)_{t=1}^{\infty}$ whose paths lie in $\mathbb{R}^{\infty}$ (or a countable product of Polish spaces). We obtain results very similar to the finite discrete-time case, mainly based on limiting arguments.[2] A notable difference is that the path space $\mathbb{R}^{\infty}$ is not geodesic and hence $\mathrm{FP}^{(\infty)}$ is not geodesic either.

Concerning continuous-time processes $(X_t)_{t \in [0,T]}$, it is known from stochastic analysis that different applications require the use of different topologies/metrics on the path

---

[2]We thank an anonymous referee for pointing us to this direction.
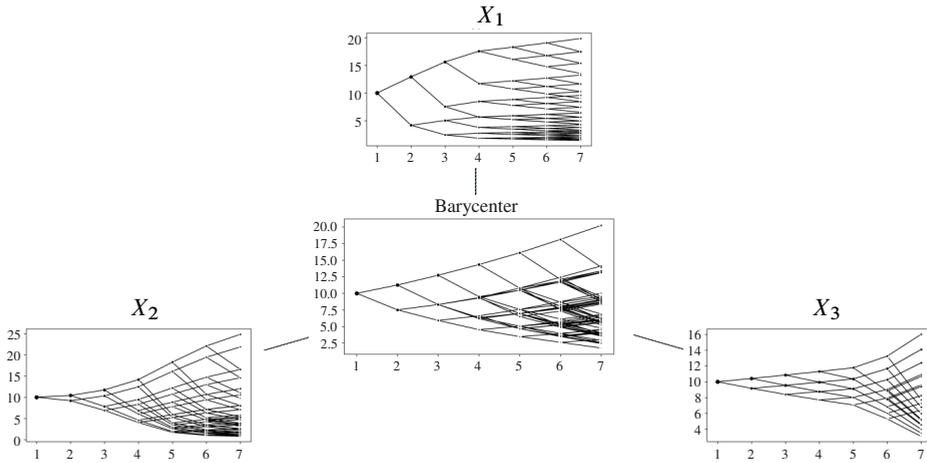
Fig. 2. Three CRR-type models as considered in finance and econometrics and their barycenter (implemented by Stephan Eckstein).

space. This fact appears even more noticeable when also information is taken into account. In Appendix C, we briefly present adapted topologies for continuous-time stochastic processes that have been used in the literature or seem sensible. We describe how $\mathcal{AW}$ needs to be altered to fit the respective choices and comment on some strengths and weaknesses of the emerging theories.

1.9. Remarks on related literature

Imposing a 'causality' constraint on a transport plan between laws of processes seems to go back to the Yamada–Watanabe criterion for stochastic differential equations [99] and is used under the name 'compatibility' by Kurtz [71].

A systematic treatment and use of causality as an interesting property of abstract transport plans between filtered probability spaces and their associated optimal transport problems were initiated by Lassalle [72] and Acciaio–Backhoff–Zalashko [2].

As noted above, different groups of authors have introduced similar 'adapted' variants of the Wasserstein distance, this includes the works of Vershik [91, 92], Rüschendorf [89], Gigli [51, Chapter 4] (see also [10, Section 12.4]), Pflug–Pichler [80], Bion-Nadal–Talay [35], and Nielsen–Sun [77]. Pflug–Pichler's nested distance had a particular impact on multistage programming, see [52, 69, 81, 82, 86] among others.

In addition to these distances, extensions of the weak topology that account for the flow of information were introduced by Aldous in stochastic analysis [9] (based on Knight's prediction process [70]), Hoover and Keisler in mathematical logic [61, 62] and Hellwig in economics [57]. Very recently, an approach using higher rank signatures was given by Bonnier–Liu–Oberhauser [36], in particular providing a metric for convergence in the adapted distribution in the sense of Hoover–Keisler.

The idea to represent information (in the sense of filtrations) using conditional distributions originated in the theory of dynamical systems and Vershik's program to classify filtrations whose time horizons start at $-\infty$, see, e.g., [92] and in particular the survey [93]. For a more probabilistic account of this line of research, we refer to [47]. Independently, Pflug [79] introduced this idea in stochastic optimization and defined the space of 'nested distributions'.

Recently, there has been significant interest in adapted/causal transport problems in discrete time or with a finite number of hierarchical levels. A goal of the present article is to provide the theoretical framework for these emerging lines of research and we briefly indicate some of these directions. In mathematical finance, the use of weak adapted topologies was initiated in the context of game options [43]. Further contributions apply adapted transport and adapted Wasserstein distances to questions of insider trading and enlargement of filtrations [2], stability of pricing/hedging and utility maximization [1, 13, 22, 23, 30, 31, 52], and interest rate uncertainty [3]. Adapted transport is used in [21] to study the sensitivity of multiperiod optimization problems and distributionally robust optimization problems. In [56], it is applied to time-dynamic matching problems [25], and in [63, 90] for the computational resolution of optimal stopping and other filtration-dependent problems. In [38], a connection of adapted transport to the Weisfeiler–Lehman distance is revealed. Machine learning algorithms based on adapted or hierarchical structures are studied in the context of image processing [65], text processing and hierarchical domain translation [46, 100], causal graph learning [7], video prediction and generation [97, 98], and universal approximation [5]. In [39], adapted transport is used as the starting point to develop a framework for more general causal dependence structures.

1.10. Organization of the paper

In Section 2, we introduce some important concepts and in particular the notions of (bi-)causality and adapted Wasserstein distance.

In Section 3, we formally discuss the Wasserstein space of filtered processes FP_p as a preparation to establish Theorems 1.2 and 1.3 subsequently. In Section 3.1, we construct a canonical filtered space that supports for each equivalence class $\mathbb{X} \in \mathrm{FP}_p$ a canonical representative. Building on the foregoing subsections, we establish in Section 3.2 an isometric isomorphism between filtered processes, their canonical counterparts, and a classical Wasserstein space.

Section 4 links the weak adapted topology to the concept of adapted functions and the prediction process by Hoover and Keisler.

Section 5 deals with topological and geometric aspects. We prove a compactness criterion, show that FP_p is the completion of $\mathcal{P}_p(\mathbb{R}^N)$ and prove that finite state Markov processes are dense in FP_p, prove that FP_p is a geodesic space for $1 < p < \infty$, and show that martingales form a closed geodesically convex subset of FP_p.

In Section 6, we discuss applications and comment on numerical aspects related to $\mathcal{A}\mathcal{W}$.

Section 7 is concerned with an example that, among other things, shows that Aldous' extended weak topology fails to guarantee the continuity of optimal stopping problems.

Finally, in Appendix A, we discuss a notion of 'block approximation' of couplings, which is an auxiliary concept required to prove the results in Section 3.2 for probability spaces that are not necessarily Polish. In Appendices B and C, we discuss extensions of the present setting to the case of stochastic processes index by infinite discrete time and continuous time, respectively.

2. Notational conventions

Throughout this article, we fix a time horizon $N \in \mathbb{N}$ and $1 \leq p < \infty$. For each time $1 \leq t \leq N$, let $\mathcal{X}_t$ be a Polish space with a fixed compatible complete metric $d_{\mathcal{X}_t}$. If $\mathcal{X}_t = \mathbb{R}^d$, then $d_t(x, y) = |x - y|$, where $|\cdot| = \|\cdot\|_2$ is the Euclidean norm. Given a finite family of sets $(A_u)_{u=1}^t$ and $1 \leq s \leq t$, we use the following abbreviation for its product:

$$A_{s:t} := A_s \times \cdots \times A_t.$$

The same convention applies to vectors. For $s \leq r \leq t$, the projection onto the r-th coordinate of $A_{s:t}$ is denoted by $\mathrm{pj}_r : A_{s:t} \to A_r$. Using this convention, we are interested in stochastic processes taking values in the path space $\mathcal{X} := \mathcal{X}_{1:N}$. Processes on $\mathcal{X}$ are usually denoted by capital letters, i.e., $X = (X_t)_{t=1}^N$, whereas specific elements of the path space are denoted by lower case, i.e., $(x_t)_{t=1}^N \in \mathcal{X}_{1:N}$.

Distances: For a Polish space $\mathcal{A}$ with fixed compatible complete metric $d_{\mathcal{A}}$, we write $\mathcal{P}(\mathcal{A})$ for the set of Borel probability measures on $\mathcal{A}$, and $\mathcal{P}_p(\mathcal{A})$ for the subset whose elements integrate $d_{\mathcal{A}}^p(\cdot, a_0)$ for some (and hence all) $a_0 \in \mathcal{A}$. If $\mathcal{B}$ is another Polish space and $\mu \in \mathcal{P}(\mathcal{A})$, $\nu \in \mathcal{P}(\mathcal{B})$, we write $\mathrm{Cpl}(\mu, \nu)$ for the set of all couplings with marginals μ, ν, that is, $\pi \in \mathrm{Cpl}(\mu, \nu)$ if $\pi \in \mathcal{P}(\mathcal{A} \times \mathcal{B})$ and its first marginal equals μ and its second equals ν. We equip $\mathcal{P}(\mathcal{A})$ with the topology of weak convergence and $\mathcal{P}_p(\mathcal{A})$ with the (p-th order) Wasserstein distance $\mathcal{W}_{\mathcal{P}_p(\mathcal{A})}$, that is,

$$\mathcal{W}_{\mathcal{P}_p(\mathcal{A})}^p(\mu, \nu) := \inf_{\pi \in \mathrm{Cpl}(\mu, \nu)} \int d_{\mathcal{A}}^p(a, \hat{a}) \pi(\mathrm{d}a, \mathrm{d}\hat{a}).$$

This renders $\mathcal{P}(\mathcal{A})$ and $\mathcal{P}_p(\mathcal{A})$ Polish spaces. Note that for a bounded metric, the weak convergence topology and the one induced by $\mathcal{W}_p$ coincide. Whenever clear from context, we will omit excessive subscripts and simply write d and $\mathcal{W}_p$ for $d_{\mathcal{A}}$ and $\mathcal{W}_{\mathcal{P}_p(\mathcal{A})}$, respectively.

Filtrations: For a filtered process $\mathbb{X} = (\Omega^{\mathbb{X}}, \mathcal{F}^{\mathbb{X}}, \mathbb{P}^{\mathbb{X}}, (\mathcal{F}_t^{\mathbb{X}})_{t=1}^N, (X_t)_{t=1}^N)$, we use the convention that $\mathcal{F}_0^{\mathbb{X}} := \{\emptyset, \Omega^{\mathbb{X}}\}$. It is important to note that since the processes start at time $t = 1$, this convention is only notational and does not imply that the initial σ-algebra is trivial. Frequently, we consider multiple products of σ-algebras. To prevent notation from getting out of hand, we write

$$\mathcal{F}_{t,s}^{\mathbb{X}, \mathbb{Y}} := \mathcal{F}_t^{\mathbb{X}} \otimes \mathcal{F}_s^{\mathbb{Y}} \quad \text{for } 0 \leq s, t \leq N$$

for two filtered processes $\mathbb{X}$ and $\mathbb{Y}$. Moreover, we will often identify $\mathcal{F}_{0,t}^{\mathbb{X},\mathbb{Y}}$ with $\mathcal{F}_t^{\mathbb{Y}}$; e.g., an $\mathcal{F}_{0,t}^{\mathbb{X},\mathbb{Y}}$-measurable functions is naturally associated on $\Omega^{\mathbb{X}} \times \Omega^{\mathbb{Y}}$ with an $\mathcal{F}_t^{\mathbb{Y}}$-measurable function depending only on the second coordinate, and vice versa. In a similar manner, for a function $f: A_t \to \mathcal{Y}$, we continue to write $f: A_{1:n} \to \mathcal{Y}$ for the function $f \circ \mathrm{pr}_t$.

Couplings: In optimal transport, couplings are the central tool for comparing probability measures. For filtered processes, this role is taken by bicausal couplings, i.e., couplings which respect the information structure of the underlying filtered probability spaces.

Definition 2.1 (Causal couplings). Let $\mathbb{X}$, $\mathbb{Y}$ be filtered processes. A probability π on $(\Omega^{\mathbb{X}} \times \Omega^{\mathbb{Y}}, \mathcal{F}^{\mathbb{X}} \otimes \mathcal{F}^{\mathbb{Y}})$ is called *coupling* between $\mathbb{X}$ and $\mathbb{Y}$ if its marginals are $\mathbb{P}^{\mathbb{X}}$ and $\mathbb{P}^{\mathbb{Y}}$. We call π

(a) *causal* (or causal from $\mathbb{X}$ to $\mathbb{Y}$) if, for every $1 \le t \le N$, conditionally on $\mathcal{F}_{t,0}^{\mathbb{X},\mathbb{Y}}$ we have that $\mathcal{F}_{N,0}^{\mathbb{X},\mathbb{Y}}$ and $\mathcal{F}_{0,t}^{\mathbb{X},\mathbb{Y}}$ are independent,

(b) *anticausal* (or causal from $\mathbb{Y}$ to $\mathbb{X}$) if, for every $1 \le t \le N$, conditionally on $\mathcal{F}_{0,t}^{\mathbb{X},\mathbb{Y}}$ we have that $\mathcal{F}_{0,N}^{\mathbb{X},\mathbb{Y}}$ and $\mathcal{F}_{t,0}^{\mathbb{X},\mathbb{Y}}$ are independent,

(c) *bicausal* if it is both, causal and anticausal.

We write $\mathrm{Cpl}(\mathbb{X}, \mathbb{Y})$, $\mathrm{Cpl}_c(\mathbb{X}, \mathbb{Y})$ and $\mathrm{Cpl}_{bc}(\mathbb{X}, \mathbb{Y})$ for the set of couplings, causal couplings and bicausal couplings, respectively.

In case that the underlying spaces are path spaces equipped with canonical filtration/processes, these definitions correspond precisely to the classical definitions given in the literature, see [12, 14, 15, 80, 81] among others. In the context of space with more general filtrations, causality and causal transport are considered in [2, 72].

The following lemma provides useful characterizations of causality which we will frequently use throughout the article.

Lemma 2.2 (Causality). *Let π be a coupling between two filtered processes $\mathbb{X}$ and $\mathbb{Y}$. Then the following are equivalent:*

(i)   *π is causal (from $\mathbb{X}$ to $\mathbb{Y}$).*

(ii)   *$\mathbb{E}_\pi[U|\mathcal{F}_{t,t}^{\mathbb{X},\mathbb{Y}}] = \mathbb{E}_\pi[U|\mathcal{F}_{t,0}^{\mathbb{X},\mathbb{Y}}]$ for all $1 \le t \le N$ and bounded $\mathcal{F}_N^{\mathbb{X}}$-mb. U.*

(iii)   *$\mathbb{E}_\pi[V|\mathcal{F}_{N,0}^{\mathbb{X},\mathbb{Y}}] = \mathbb{E}_\pi[V|\mathcal{F}_{t,0}^{\mathbb{X},\mathbb{Y}}]$ for all $1 \le t \le N$ and bounded $\mathcal{F}_t^{\mathbb{Y}}$-mb. V.*

Moreover, U in (ii) can be allowed to be $\mathcal{F}_{t,N}^{\mathbb{X},\mathbb{Y}}$-measurable and V in (iii) can be allowed to be $\mathcal{F}_{t,t}^{\mathbb{X},\mathbb{Y}}$-measurable.

In words, (ii) says that given the past of $\mathbb{X}$, the past of $\mathbb{Y}$ does not provide additional information about the future of $\mathbb{X}$, and (iii) says that given the past of $\mathbb{X}$, the future of $\mathbb{X}$ does not provide additional information about the past of $\mathbb{Y}$.

Proof of Lemma 2.2. The equivalence between (i)–(iii) is a consequence of [66, Proposition 5.6]. The second statement follows from a standard application of the monotone class theorem.   ∎

The adapted Wasserstein distance between filtered processes $\mathbb{X}$ and $\mathbb{Y}$ taking values in $\mathcal{X}$ is then defined as

$$\mathcal{A}\mathcal{W}_p^p(\mathbb{X}, \mathbb{Y}) := \inf_{\pi \in \mathrm{Cpl}_{\mathrm{bc}}(\mathbb{X}, \mathbb{Y})} \mathbb{E}_\pi[d_{\mathcal{X},p}^p(X, Y)],$$

where

$$d_{\mathcal{X},p}(x, y) = \left(\sum_{t=1}^N d_{\mathcal{X}_t}^p(x_t, y_t) \right)^{1/p}.$$

When clear from context, we write d instead of $d_{\mathcal{X},p}$. Similarly, we write $\mathbb{E}[f(X)]$ instead of $\mathbb{E}_{\mathbb{P}^{\mathbb{X}}}[f(X)]$ etc.

Kernels and product measures: For two Polish spaces $\mathcal{A}$ and $\mathcal{B}$, the term kernel refers to a Borel measurable mapping $k \colon \mathcal{A} \to \mathcal{P}_p(\mathcal{B})$. For $\mu \in \mathcal{P}_p(\mathcal{A})$ and a kernel k, we write $\mu \otimes k \in \mathcal{P}_p(\mathcal{A} \times \mathcal{B})$ for the measure given by $\mu \otimes k(A \times B) = \int_A k^a(B)\,\mu(\mathrm{d}a)$. If $\nu \in \mathcal{P}_b(\mathcal{B})$, we write $\mu \otimes \nu \in \mathcal{P}_p(\mathcal{A} \times \mathcal{B})$ for the product measure. For a measure $\mu \in \mathcal{P}(\mathcal{A})$ and a Borel measurable mapping $f \colon \mathcal{A} \to \mathcal{B}$, the push-forward of μ under f is denoted by $f_*\mu$.

3. The Wasserstein space of stochastic processes

3.1. The canonical filtered space

In order to prove Theorem 1.2, we introduce the canonical space of filtered processes. The classical canonical space of a stochastic process X is the triplet consisting of path space, Borel σ-algebra and its induced law. Clearly, this triplet is adequate if one is interested solely in trajectorial properties of the process. However, the filtration is a major part of filtered processes $\mathbb{X} \in \mathcal{FP}$ and therefore we need to capture the information contained in its filtration $\mathcal{F}^{\mathbb{X}}$ in a canonical way. Thus, we need to define a canonical space which is capable to carry, besides the path properties, also the relevant informational properties of $\mathbb{X}$.

As an instructional example, consider two 1-step filtered processes $\mathbb{X}$ and $\mathbb{Y}$ taking values in $\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2 = \{0\} \times \{-1, 1\}$. We write $\mathcal{G}_1$ for the trivial σ-algebra, $\mathcal{G}_2 = \mathcal{B}$ for the Borel σ-algebra on $\mathcal{X}$, and $X = (X_t)_{t=1}^2$ for the coordinate process on $\mathcal{X}$. Let $\mathbb{P} := (\delta_{(0,-1)} + \delta_{(0,1)})/2$, and define

$$\mathbb{X} := (\mathcal{X}, \mathcal{B}, \mathbb{P}, (\mathcal{G}_1, \mathcal{G}_2), X) \quad \text{and} \quad \mathbb{Y} := (\mathcal{X}, \mathcal{B}, \mathbb{P}, (\mathcal{G}_2, \mathcal{G}_2), X).$$

Even though the laws of the paths of the two processes coincide, their probabilistic behaviour is very different due to their different filtrations, specifically we have

$$\mathcal{L}(X_2 | \mathcal{F}_1^{\mathbb{X}})(\omega) = \frac{1}{2}(\delta_{-1} + \delta_1) \quad \text{whereas} \quad \mathcal{L}(X_2 | \mathcal{F}_1^{\mathbb{Y}})(\omega) = \delta_{X_2(\omega)}.$$

From this perspective, the product $\mathcal{X}_1 \times \mathcal{P}(\mathcal{X}_2)$ is adequate to capture X_1, and additionally the information on X_2 we can witness at time 1 based on the filtration, i.e., $(X_1, \mathcal{L}(X_2 | \mathcal{F}_1^{\mathbb{X}}))$.

This observation leads to the definition of what we call the canonical filtered space, the information process and the canonical filtered process.

Definition 3.1 (Canonical space). Fix $p \in [1, \infty)$. We iteratively define a sequence of nested spaces. We write $(\mathcal{Z}_N, d_{\mathcal{Z}_N}) := (\mathcal{X}_N, d_{\mathcal{X}_N})$ and recursively for $t = N - 1, \dots, 1$,

$$\mathcal{Z}_t := \mathcal{Z}_t^- \times \mathcal{Z}_t^+ := \mathcal{X}_t \times \mathcal{P}_p(\mathcal{Z}_{t+1})$$

with metric $d_{\mathcal{Z}_t}^p := d_{\mathcal{X}_t}^p + \mathcal{W}_{p,\mathcal{P}(\mathcal{Z}_{t+1})}^p$. The elements of $\mathcal{Z}_t$ are denoted by

$$z_t = (z_t^-, z_t^+) \in \mathcal{Z}_t^- \times \mathcal{Z}_t^+.$$

The *canonical filtered space* is given by the triplet

$$(\mathcal{Z}, \mathcal{F}^{\mathcal{Z}}, (\mathcal{F}_t^{\mathcal{Z}})_{t=1}^N), \tag{3.1}$$

where $\mathcal{Z}_{1:N}$ is denoted by $\mathcal{Z}$, elements of $\mathcal{Z}$ by $z = z_{1:N}$, the Borel σ-algebra on $\mathcal{Z}$ by $\mathcal{F}^{\mathcal{Z}}$, and $\sigma(z \mapsto z_{1:t})$ by $\mathcal{F}_t^{\mathcal{Z}}$. In the context of the canonical filtered space, the map $Z^-: \mathcal{Z} \to \mathcal{X}_{1:N}$ denotes the evaluation map

$$Z^-(z) := (Z_t^-(z))_{t=1}^N := (z_t^-)_{t=1}^N. \tag{3.2}$$

The spaces introduced in Definition 3.1 are Polish as all operations involved in their definition preserve this property. The next definition associates to a filtered process $\mathbb{X}$ its canonical counterpart – the information process $\mathrm{ip}(\mathbb{X})$ defined on $\Omega^{\mathbb{X}}$ and taking values in $\mathcal{Z}$. We shall later see that the information process selects all information contained in the original filtration relevant for the process; hence its name.

Definition 3.2 (The information process). To each $\mathbb{X} \in \mathcal{FP}$, we associate its *information process* $\mathrm{ip}(\mathbb{X}) = (\mathrm{ip}_t(\mathbb{X}))_{t=1}^N$ defined by setting

$$\mathrm{ip}_N(\mathbb{X}): \Omega^{\mathbb{X}} \to \mathcal{Z}_N, \quad \omega \mapsto X_N(\omega)$$

and, recursively for $t = N - 1, \dots, 1$,

$$\mathrm{ip}_t(\mathbb{X}) = (\mathrm{ip}_t^-(\mathbb{X}), \mathrm{ip}_t^+(\mathbb{X})): \Omega^{\mathbb{X}} \to \mathcal{Z}_t, \quad \omega \mapsto (X_t(\omega), \mathcal{L}(\mathrm{ip}_{t+1}(\mathbb{X}) | \mathcal{F}_t^{\mathbb{X}})(\omega)).$$

Note that $\mathrm{ip}_t^-(\mathbb{X})$ is $\mathcal{Z}_t^-$-valued and $\mathrm{ip}_t^+(\mathbb{X})$ is $\mathcal{Z}_t^+$-valued. In particular, this implies that the information process is well defined: as $(\mathcal{Z}_t)_{t=1}^N$ consists of Polish spaces, the conditional probabilities appearing in the recursive definition of $\mathrm{ip}(\mathbb{X})$ above exist.

The information processes will be an essential ingredient when we define canonical representatives of filtered processes (see Definition 3.8 below). The next lemma can be seen as a first justification of its name.

Lemma 3.3 (The information process is self-aware). *For every bounded Borel (continuous) function $f: \mathcal{Z} \to \mathbb{R}$ and every $1 \le t \le N$, there is a bounded Borel (continuous) function $g: \mathcal{Z}_{1:t} \to \mathbb{R}$ such that*

$$\mathbb{E}[f(\mathrm{ip}(\mathbb{X})) | \mathcal{F}_t^{\mathbb{X}}] = g(\mathrm{ip}_{1:t}(\mathbb{X})) \quad \text{for all } \mathbb{X} \in \mathcal{FP}.$$

More generally, let $\mathcal{A}$ be a Polish space. For every Borel (continuous) function $f\colon \mathcal{Z} \to \mathcal{A}$ and every $1 \leq t \leq N$, there is a Borel (continuous) function $g\colon \mathcal{Z}_{1:t} \to \mathcal{P}(\mathcal{A})$ such that

$$\mathcal{L}(f(\mathrm{ip}(\mathbb{X}))|\mathcal{F}_t^{\mathbb{X}}) = g(\mathrm{ip}_{1:t}(\mathbb{X})) \quad \text{for all } \mathbb{X} \in \mathcal{FP}.$$

This lemma is an easy consequence of properties of the unfold operator introduced below, see Lemma 3.5. Nevertheless, to boost the reader's intuition, we want to include a direct proof for the first statement in the notationally lighter case $N = 2$.

Sketch of proof for $N = 2$. For $t = 2$, there is nothing to do, as $\mathrm{ip}(\mathbb{X})$ is $\mathcal{F}_2^{\mathbb{X}}$-measurable. Let $t = 1$. For simplicity, we assume first that f is a product

$$f(z_1, z_2) = f_1(z_1) f_2(z_2)$$

for suitable f_1 and f_2. Then we can write

$$\mathbb{E}[f(\mathrm{ip}(\mathbb{X}))|\mathcal{F}_1^{\mathbb{X}}] = f_1(\mathrm{ip}_1(\mathbb{X}))\mathbb{E}[f_2(\mathrm{ip}_2(\mathbb{X}))|\mathcal{F}_1^{\mathbb{X}}]$$
$$= f_1(\mathrm{ip}_1(\mathbb{X})) \int f_2(z_2)\mathcal{L}(\mathrm{ip}_2(\mathbb{X})|\mathcal{F}_1^{\mathbb{X}})(\mathrm{d}z_2),$$

where we used that $\mathrm{ip}_1(\mathbb{X})$ is $\mathcal{F}_1^{\mathbb{X}}$-measurable. By definition,

$$\mathrm{ip}_1^+(\mathbb{X}) = \mathcal{L}(\mathrm{ip}_2(\mathbb{X})|\mathcal{F}_1^{\mathbb{X}}),$$

whence

$$\mathbb{E}[f(\mathrm{ip}(\mathbb{X}))|\mathcal{F}_1^{\mathbb{X}}] = g(\mathrm{ip}_1(\mathbb{X})) \quad \text{for } g(z_1) := f_1(z_1) \int f_2(z_2)z_1^+(\mathrm{d}z_2).$$

For general f not necessarily of product form, a straightforward application of the monotone class theorem concludes the proof. ∎

In what follows, we are often dealing with mappings between nested spaces of probability measures with different algebraic structures. The unfold operator, introduced below, is an essential tool in reducing bookkeeping to a comprehensible level.

Definition 3.4 (Unfold). For every $1 \leq t \leq N - 1$, we define

$$\mathrm{uf}_t\colon \mathcal{P}_p(\mathcal{Z}_t) \to \mathcal{P}_p(\mathcal{Z}_{t:N}),$$
$$\mu \mapsto \mu(\mathrm{d}z_t)z_t^+(\mathrm{d}z_{t+1})z_{t+1}^+(\mathrm{d}z_{t+2})\cdots z_{N-1}^+(\mathrm{d}z_N), \tag{3.3}$$

and call uf_t the *unfold* operator (at time t). For $t = N$, we define uf_N to be the identity map.

Recall the definition of $\mathrm{ip}(\mathbb{X})$, see Definition 3.2. As $\mathrm{ip}_t^+(\mathbb{X})$ is a random variable taking values in $\mathcal{P}(\mathcal{Z}_{t+1})$, the unfold operator can be applied pointwise, i.e., we may consider $\mathrm{uf}_{t+1}(\mathrm{ip}_t^+(\mathbb{X}))$. The following lemma explores properties of uf_t in view of the information process $\mathrm{ip}(\mathbb{X})$.

Lemma 3.5. *For $1 \leq t \leq N - 1$, the following hold:*

(i)   *For $\mu \in \mathcal{P}_p(\mathbb{Z}_t)$, we have*

$$\mathrm{uf}_t(\mu)(dz_{t:N}) = \mu(dz_t)\mathrm{uf}_{t+1}(z_t^+)(dz_{t+1:N}). \tag{3.4}$$

(ii)   *For $\mathbb{X} \in \mathcal{FP}_p$, we have*

$$\mathcal{L}(\mathrm{ip}(\mathbb{X})|\mathcal{F}_t^{\mathbb{X}}) = \delta_{\mathrm{ip}_{1:t}(\mathbb{X})} \otimes \mathrm{uf}_{t+1}(\mathrm{ip}_t^+(\mathbb{X})). \tag{3.5}$$

In other words, for all bounded Borel functions $f: \mathbb{Z} \to \mathbb{R}$, we have

$$\mathbb{E}[f(\mathrm{ip}(\mathbb{X})) \mid \mathcal{F}_t^{\mathbb{X}}] = \int f(\mathrm{ip}_{1:t}(\mathbb{X}), z_{t+1:N})\mathrm{uf}_{t+1}(\mathrm{ip}_t^+(\mathbb{X}))(dz_{t+1:N}). \tag{3.6}$$

(iii) uf_t *is Lipschitz continuous from $\mathcal{P}_p(\mathbb{Z}_t)$ to $\mathcal{P}_p(\mathbb{Z}_{t:N})$.*

Proof. (i) Equation (3.4) is another way of expressing (3.3).

(ii) The statements in (3.5) and (3.6) are clearly equivalent and we shall therefore only prove the latter one via backward induction.

Since there is nothing to do for $t = N - 1$, we assume that the statement is true for $2 \leq t \leq N - 1$. Let $g: \mathbb{Z}_{1:t} \to \mathbb{R}$ be the Borel function defined as

$$g(z_{1:t}) := \int f(z_{1:t}, z_{t+1:N})\mathrm{uf}_{t+1}(z_t^+)(dz_{t+1:N}), \tag{3.7}$$

whereby the inductive hypothesis now reads $\mathbb{E}[f(\mathrm{ip}(\mathbb{X}))|\mathcal{F}_t^{\mathbb{X}}] = g(\mathrm{ip}_{1:t}(\mathbb{X}))$.

Then $\mathrm{ip}_{t-1}^+(\mathbb{X}) = \mathcal{L}(\mathrm{ip}_t(\mathbb{X})|\mathcal{F}_{t-1}^{\mathbb{X}})$ and the tower property implies

$$\mathbb{E}[f(\mathbb{X})|\mathcal{F}_{t-1}^{\mathbb{X}}] = \mathbb{E}[\mathbb{E}[f(\mathbb{X})|\mathcal{F}_t^{\mathbb{X}}]|\mathcal{F}_{t-1}^{\mathbb{X}}] = \mathbb{E}[g(\mathrm{ip}_{1:t}(\mathbb{X}))|\mathcal{F}_{t-1}^{\mathbb{X}}]$$

$$= \int g(\mathrm{ip}_{1:t-1}(\mathbb{X}), z_t)\mathrm{ip}_{t-1}^+(\mathbb{X})(dz_t). \tag{3.8}$$

Recalling (3.4) and (3.7), we conclude that the last term in (3.8) equals

$$\int \left(\int f(\mathrm{ip}_{1:t-1}(\mathbb{X}), z_t, z_{t+1:N})\mathrm{uf}_{t+1}(z_t^+)(dz_{t+1:N}) \right)\mathrm{ip}_{t-1}^+(\mathbb{X})(dz_t)$$

$$= \int f(\mathrm{ip}_{1:t-1}(\mathbb{X}), z_{t:N})\mathrm{uf}_t(\mathrm{ip}_{t-1}^+(\mathbb{X}))(dz_{t:N}),$$

which completes the proof of (ii).

(iii) The assertion is shown via a backward induction over t.

For $t = N$, the unfold operator is given as the identity map, which is in particular Lipschitz continuous. Assume that the claim is true for some $2 \leq t + 1 \leq N$. By (3.4), we can write

$$\mathrm{uf}_t(\mu) = \mu \otimes k_t,$$

where $k_t: \mathbb{Z}_t \to \mathcal{P}_p(\mathbb{Z}_{t+1:N})$. This map is explicitly given by $k_t(z_t) := \mathrm{uf}_{t+1}(z_t^+)$. An application of Lemma 3.6 below implies that uf_t is again Lipschitz continuous with a new constant.   ∎

Lemma 3.6. *Let $\mathcal{A}$ and $\mathcal{B}$ be Polish spaces and let $k\colon \mathcal{A} \to \mathcal{P}_p(\mathcal{B})$ be L-Lipschitz. Then the map*

$$\mathcal{P}_p(\mathcal{A}) \to \mathcal{P}_p(\mathcal{A} \times \mathcal{B}), \quad \mu \mapsto \mu \otimes k$$

is $(1 + L^p)^{1/p}$-Lipschitz.

Proof. A short computation shows that there are $a_0 \in \mathcal{A}$, $b_0 \in \mathcal{B}$ and a constant c such that $\int d^p(b, b_0)\, k^a(\mathrm{d}b) \le c(1 + d^p(a, a_0))$ for all $a \in \mathcal{A}$; in particular, $\mu \otimes k \in \mathcal{P}_p(\mathcal{A} \times \mathcal{B})$ is well defined for every $\mu \in \mathcal{P}_p(\mathcal{A})$.

Let $\mu, \widehat{\mu} \in \mathcal{P}_p(\mathcal{A})$ and denote by π the $\mathcal{W}_p$-optimal coupling between them. For every pair $(a, \widehat{a}) \in \mathcal{A} \times \mathcal{A}$, let $\gamma^{a, \widehat{a}} \in \mathrm{Cpl}(k^a, k^{\widehat{a}})$ be optimal for $\mathcal{W}_p(k^a, k^{\widehat{a}})$. Using the Jankov–von Neumann theorem [67, Theorem 18.1], standard arguments show that $a, \widehat{a} \mapsto \gamma^{a, \widehat{a}}$ can be chosen universally measurably. Then

$$\Pi(\mathrm{d}a, \mathrm{d}b, \mathrm{d}\widehat{a}, \mathrm{d}\widehat{b}) := \pi(\mathrm{d}a, \mathrm{d}\widehat{a})\gamma^{a, \widehat{a}}(\mathrm{d}b, \mathrm{d}\widehat{b})$$

defines a coupling between $\mu \otimes k$ and $\widehat{\mu} \otimes k$. Thus, we can estimate

$$\mathcal{W}_p^p(\mu \otimes k, \widehat{\mu} \otimes k) \le \int d^p(a, \widehat{a}) + d^p(b, \widehat{b})\, \Pi(\mathrm{d}a, \mathrm{d}b, \mathrm{d}\widehat{a}, \mathrm{d}\widehat{b})$$

$$= \int d^p(a, \widehat{a}) + \mathcal{W}_p^p(k^a, k^{\widehat{a}})\pi(\mathrm{d}a, \mathrm{d}\widehat{a}) \le (1 + L^p)\mathcal{W}_p^p(\mu, \widehat{\mu}),$$

where the last inequality holds by L-Lipschitz continuity of k. ∎

Proof of Lemma 3.3. Let $f\colon \mathcal{Z} \to \mathcal{A}$ be bounded and Borel measurable (continuous), and $1 \le t \le N - 1$. As uf_{t+1} is continuous by Lemma 3.5 (iii), we obtain continuity of

$$F\colon \mathcal{Z}_{1:t} \to \mathcal{P}_p(\mathcal{Z}), \quad z_{1:t} \mapsto \delta_{z_{1:t}} \otimes \mathrm{uf}_{t+1}(z_t^+).$$

Define the map $G\colon \mathcal{Z}_{1:t} \to \mathcal{P}(\mathcal{A})$ as the push-forward $G(z_{1:t}) := f_* F(z_{1:t})$. Obviously, when f is continuous, G is also continuous as the composition of continuous functions. By Lemma 3.5 (ii), we obtain

$$\mathcal{L}(f(\mathrm{ip}(\mathbb{X}))|\mathcal{F}_t^{\mathbb{X}}) = f_* \mathcal{L}(\mathrm{ip}(\mathbb{X})|\mathcal{F}_t^{\mathbb{X}}) = f_* F(\mathrm{ip}_{1:t}(\mathbb{X})) = G(\mathrm{ip}_{t+1}(\mathbb{X})).$$

We have thus proved the second assertion.

To obtain the first assertion, let $\mathcal{A} = \mathbb{R}$. Define $g\colon \mathcal{Z}_{1:t} \to \mathbb{R}$ by

$$g(z_{1:t}) := \int a\, G(z_{1:t})(\mathrm{d}a),$$

which is well defined since f is bounded with values in a compact set, say, $K \subset \mathbb{R}$. Hence, we can view G as a function mapping into $\mathcal{P}(K)$. By the first part of the proof, we obtain

$$\mathbb{E}[f(\mathrm{ip}(\mathbb{X}))|\mathcal{F}_t^{\mathbb{X}}] = \int a\, G(\mathrm{ip}_{t+1}(\mathbb{X}))(\mathrm{d}a) = g(\mathrm{ip}_{t+1}(\mathbb{X})).$$

The map $p \mapsto \int a p(\mathrm{d}a)$ is continuous on $\mathcal{P}(K)$, and we conclude that g is continuous if f is. ∎

Equipped with the unfold operator, we define our real object of interest.

Definition 3.7 (Canonical filtered processes). We call $\mathbb{X} \in \mathscr{FP}$ a *canonical filtered process*, in symbols $\mathbb{X} \in \mathrm{CFP}$, if

$$\mathbb{X} = (\mathcal{Z}, \mathscr{F}^{\mathcal{Z}}, \mathrm{uf}_1(\bar{\mu}), (\mathscr{F}_t^{\mathcal{Z}})_{t=1}^N, Z^-), \tag{3.9}$$

where $(\mathcal{Z}, \mathscr{F}^{\mathcal{Z}}, (\mathscr{F}_t^{\mathcal{Z}})_{t=1}^N)$ is the canonical filtered space, see (3.1), Z^- is the evaluation map (3.2), and $\bar{\mu} \in \mathcal{P}(\mathcal{Z}_1)$. As usual, we write CFP_p for the subset of processes whose laws have finite p-th moment.

Using the concept of information process, we can associate to an arbitrary filtered process a unique element in CFP.

Definition 3.8 (Associated canonical filtered processes). Let $\mathbb{X} \in \mathscr{FP}_p$, and let $\bar{\mathbb{X}} \in \mathrm{CFP}_p$ be given by (3.9) with $\bar{\mu} := \mathcal{L}(\mathrm{ip}_1(\mathbb{X}))$, or, according to Lemma 3.5, equivalently

$$\bar{\mathbb{X}} = (\mathcal{Z}, \mathscr{F}^{\mathcal{Z}}, \mathcal{L}(\mathrm{ip}(\mathbb{X})), (\mathscr{F}_t^{\mathcal{Z}})_{t=1}^N, Z^-). \tag{3.10}$$

We call $\bar{\mathbb{X}}$ the *canonical filtered process associated to* $\mathbb{X}$.

We want to stress at this point that, as stated in (3.10), all information of $\mathcal{L}(\mathrm{ip}(\mathbb{X})) \in \mathcal{P}(\mathcal{Z})$ is already contained in $\mathcal{L}(\mathrm{ip}_1(\mathbb{X})) \in \mathcal{P}(\mathcal{Z}_1)$ by Lemma 3.5. The relations between CFP_p and $\mathcal{P}_p(\mathcal{Z}_1)$ become apparent in Theorem 3.10 below, and the relation of filtered processes to their canonical counterparts becomes apparent through Lemma 3.9.

Lemma 3.9. *Let* $\mathbb{X}, \mathbb{Y} \in \mathscr{FP}_p$, *and let* $\bar{\mathbb{X}}, \bar{\mathbb{Y}} \in \mathrm{CFP}_p$ *be their associated canonical processes. The following hold:*

(i) $(\mathrm{id}, \mathrm{ip}(\mathbb{X}))_* \mathbb{P}^{\mathbb{X}} \in \mathrm{Cpl}_{\mathrm{bc}}(\mathbb{X}, \bar{\mathbb{X}})$.

(ii)   *If* $\pi \in \mathrm{Cpl}_{\mathrm{c}}(\mathbb{X}, \mathbb{Y})$, *then* $(\mathrm{ip}(\mathbb{X}), \mathrm{ip}(\mathbb{Y}))_* \pi \in \mathrm{Cpl}_{\mathrm{c}}(\bar{\mathbb{X}}, \bar{\mathbb{Y}})$.

(iii) *If* $\pi \in \mathrm{Cpl}_{\mathrm{bc}}(\mathbb{X}, \mathbb{Y})$, *then* $(\mathrm{ip}(\mathbb{X}), \mathrm{ip}(\mathbb{Y}))_* \pi \in \mathrm{Cpl}_{\mathrm{bc}}(\bar{\mathbb{X}}, \bar{\mathbb{Y}})$.

Proof. To show (i), we write $\gamma := (\mathrm{id}, \mathrm{ip}(\mathbb{X}))_* \mathbb{P}^{\mathbb{X}}$ and first check causality of γ using (iii) of the characterization of causality given in Lemma 2.2. To that end, let $V: \mathcal{Z} \to \mathbb{R}$ be bounded and $\mathscr{F}_t^{\mathcal{Z}}$-measurable. From the definition of γ, we see that γ-almost surely $V = V(\mathrm{ip}(\mathbb{X}))$. Thus, $V(\mathrm{ip}(\mathbb{X}))$ is $\mathscr{F}_t^{\mathbb{X}}$-measurable and causality of γ from $\mathbb{X}$ to $\bar{\mathbb{X}}$ follows from

$$\mathbb{E}_\gamma[V \mid \mathscr{F}_{N,0}^{\mathbb{X},\mathcal{Z}}] = \mathbb{E}_\gamma[V(\mathrm{ip}(\mathbb{X})) \mid \mathscr{F}_{N,0}^{\mathbb{X},\mathcal{Z}}] = V(\mathrm{ip}(\mathbb{X}))$$
$$= \mathbb{E}_\gamma[V(\mathrm{ip}(\mathbb{X})) \mid \mathscr{F}_{t,0}^{\mathbb{X},\mathcal{Z}}] = \mathbb{E}_\gamma[V \mid \mathscr{F}_{t,0}^{\mathbb{X},\mathcal{Z}}].$$

To see the causality of γ from $\bar{\mathbb{X}}$ to $\mathbb{X}$, we will again use Lemma 2.2, this time item (ii). Let $U: \mathcal{Z} \to \mathbb{R}$ be bounded and $\mathscr{F}_N^{\mathcal{Z}}$-measurable. Again, due to the structure of γ it is readily verified that γ-almost surely $U = U(\mathrm{ip}(\mathbb{X}))$ and

$$\mathbb{E}_\gamma[U \mid \mathscr{F}_{t,t}^{\mathbb{X},\mathcal{Z}}] = \mathbb{E}_\gamma[U \mid \mathscr{F}_{t,0}^{\mathbb{X},\mathcal{Z}}], \tag{3.11}$$

$$\mathbb{E}_\gamma[U \mid \mathscr{F}_{0,t}^{\mathbb{X},\mathcal{Z}}] = \mathbb{E}_\gamma[U \mid \mathrm{ip}_{1:t}(\mathbb{X})]. \tag{3.12}$$

By the self-awareness property of the information process, see Lemma 3.3, (3.11) and (3.12), we find

$$\mathbb{E}_\gamma[U \mid \mathcal{F}_{t,t}^{\mathbb{X},\mathbb{Z}}] = \mathbb{E}_\gamma[U(\mathrm{ip}(\mathbb{X})) \mid \mathcal{F}_{t,0}^{\mathbb{X},\mathbb{Z}}] = \mathbb{E}_\gamma[U(\mathrm{ip}(\mathbb{X})) \mid \mathrm{ip}_{1:t}(\mathbb{X})]$$
$$= \mathbb{E}_\gamma[U(\mathrm{ip}(\mathbb{X})) \mid \mathcal{F}_{0,t}^{\mathbb{X},\mathbb{Z}}] = \mathbb{E}_\gamma[U \mid \mathcal{F}_{0,t}^{\mathbb{X},\mathbb{Z}}],$$

which completes the proof of item (i).

To verify (ii), we write $\bar{\pi} := (\mathrm{ip}(\mathbb{X}), \mathrm{ip}(\mathbb{Y}))_* \pi$, $\eta := (\mathrm{id}, \mathrm{ip}(\mathbb{X}), \mathrm{ip}(\mathbb{Y}))_* \pi$, and let $U : \mathbb{Z} \to \mathbb{R}$ be as above. By Lemmas 2.2 and 3.3, we have η-almost surely

$$\mathbb{E}_\pi[U(\mathrm{ip}(\mathbb{X})) \mid \mathcal{F}_{t,t}^{\mathbb{X},\mathbb{Y}}] = \mathbb{E}_\pi[U(\mathrm{ip}(\mathbb{X})) \mid \mathcal{F}_{t,0}^{\mathbb{X},\mathbb{Y}}] = \mathbb{E}_\pi[U(\mathrm{ip}(\mathbb{X})) \mid \mathrm{ip}_{1:t}(\mathbb{X})]. \quad (3.13)$$

Using (3.13) yields η-almost surely

$$\mathbb{E}_{\bar{\pi}}[U \mid \mathcal{F}_{t,0}^{\mathbb{Z},\mathbb{Z}}] = \mathbb{E}_\pi[U(\mathrm{ip}(\mathbb{X})) \mid \mathcal{F}_{t,t}^{\mathbb{X},\mathbb{Y}}].$$

We conclude by the tower property η-almost surely

$$\mathbb{E}_{\bar{\pi}}[U \mid \mathcal{F}_{t,0}^{\mathbb{Z},\mathbb{Z}}] = \mathbb{E}_\pi[\mathbb{E}_\pi[U(\mathrm{ip}(\mathbb{X})) \mid \mathcal{F}_{t,t}^{\mathbb{X},\mathbb{Y}}] \mid \mathrm{ip}_{1:t}(\mathbb{X}), \mathrm{ip}_{1:t}(\mathbb{Y})]$$
$$= \mathbb{E}_\pi[U(\mathrm{ip}(\mathbb{X})) \mid \mathrm{ip}_{1:t}(\mathbb{X}), \mathrm{ip}_{1:t}(\mathbb{Y})] = \mathbb{E}_{\bar{\pi}}[U \mid \mathcal{F}_{t,t}^{\mathbb{Z},\mathbb{Z}}],$$

which completes the proof of (ii).

Finally, for symmetry reasons (ii) implies (iii).   ∎

3.2. The isometry

Based on the preparatory work from the preceding subsection, we are able to establish Theorem 3.10. From this, we derive that $\mathcal{A}\mathcal{W}_p$ naturally induces a complete metric on the factor space FP_p, and that FP_p is isometrically isomorphic to the (classical) Wasserstein space $(\mathcal{Z}_1, \mathcal{W}_p)$, thereby establishing Theorem 1.3.

Theorem 3.10. *Let $\mathbb{X}, \mathbb{Y} \in \mathcal{F}\mathcal{P}_p$, and let $\bar{\mathbb{X}}, \bar{\mathbb{Y}} \in \mathrm{CFP}_p$ be the associated canonical processes. Then*

$$\mathcal{A}\mathcal{W}_p(\mathbb{X}, \mathbb{Y}) = \mathcal{A}\mathcal{W}_p(\bar{\mathbb{X}}, \bar{\mathbb{Y}}) = \mathcal{W}_p(\mathcal{L}(\mathrm{ip}_1(\mathbb{X})), \mathcal{L}(\mathrm{ip}_1(\mathbb{Y}))). \quad (3.14)$$

In particular, $\mathcal{A}\mathcal{W}_p$ is a pseudo-metric on $\mathcal{F}\mathcal{P}_p$ and the embedding $\bar{\mathbb{X}} \mapsto \mathcal{L}(\mathrm{ip}_1(\bar{\mathbb{X}}))$ is an isometric isomorphism of CFP_p and $\mathcal{P}_p(\mathcal{Z}_1)$.

Proof. The first equality in (3.14) is a direct consequence of Lemma 3.9 and Theorem A.4 in the appendix. For the convenience of the reader, we present an alternative proof of the first equality under the assumption that the probability spaces of $\mathbb{X}$ and $\mathbb{Y}$ are Polish, thereby omitting the technical result in Theorem A.4.

By Lemma 3.9 (iii), we find

$$\mathcal{A}\mathcal{W}_p(\mathbb{X}, \mathbb{Y}) \geq \mathcal{A}\mathcal{W}_p(\bar{\mathbb{X}}, \bar{\mathbb{Y}}).$$

To see the reverse inequality, let $\bar{\pi} \in \mathrm{Cpl}_{\mathrm{bc}}(\overline{\mathbb{X}}, \overline{\mathbb{Y}})$ and write

$$\gamma := (\mathrm{id}, \mathrm{ip}(\mathbb{X}))_* \mathbb{P}^{\mathbb{X}} \quad \text{and} \quad \hat{\gamma} := (\mathrm{id}, \mathrm{ip}(\mathbb{Y}))_* \mathbb{P}^{\mathbb{Y}}.$$

These couplings are bicausal by Lemma 3.9 (i) and admit disintegrations $(\gamma_z)_{z \in \mathbb{Z}}$ and $(\hat{\gamma}_{\hat{z}})_{\hat{z} \in \mathbb{Z}}$ since the considered probability spaces are Polish by assumption. Consider the probability

$$\pi(\mathrm{d}\omega, \mathrm{d}\hat{\omega}) := \int \gamma_z(\mathrm{d}\omega) \hat{\gamma}_{\hat{z}}(\mathrm{d}\hat{\omega}) \bar{\pi}(\mathrm{d}z, \mathrm{d}\hat{z}).$$

For symmetry reasons, we will only show that π is causal from $\mathbb{X}$ to $\mathbb{Y}$. By Lemma 2.2, it suffices to show that for any bounded $\mathcal{F}_t^{\mathbb{Y}}$-measurable V, we have

$$\mathbb{E}_\pi[V \mid \mathcal{F}_{N,0}^{\mathbb{X},\mathbb{Y}}] = \mathbb{E}_\pi[V \mid \mathcal{F}_{t,0}^{\mathbb{X},\mathbb{Y}}].$$

As $\hat{\gamma}$ is bicausal, Lemma 2.2 asserts that $\hat{z} \mapsto \int V(\hat{\omega})\hat{\gamma}_{\hat{z}}(\mathrm{d}\hat{\omega})$ is $\mathcal{F}_t^{\overline{\mathbb{Y}}}$-measurable. By the same reasoning, we obtain $\mathcal{F}_t^{\mathbb{X}}$-measurability of

$$\omega \mapsto W(\omega) := \iiint V(\hat{\omega})\hat{\gamma}_{\hat{z}}(\mathrm{d}\hat{\omega})\bar{\pi}_z(\mathrm{d}\hat{z})\gamma_\omega(\mathrm{d}z),$$

where $\gamma_\omega(\mathrm{d}z) := \mathcal{L}(\mathrm{ip}(\mathbb{X})|\mathcal{F}_N^{\mathbb{X}})(\omega)$. Hence, by the definition of π and the tower property, we get

$$\mathbb{E}_\pi[V \mid \mathcal{F}_{N,0}^{\mathbb{X},\mathbb{Y}}] = W = \mathbb{E}_\pi[V \mid \mathcal{F}_{t,0}^{\mathbb{X},\mathbb{Y}}].$$

As V is arbitrary, this yields $\pi \in \mathrm{Cpl}_{\mathrm{c}}(\mathbb{X}, \mathbb{Y})$ and by symmetry, $\pi \in \mathrm{Cpl}_{\mathrm{bc}}(\mathbb{X}, \mathbb{Y})$. Moreover, we have $\mathbb{E}_\pi[d^p(X, Y)] = \mathbb{E}_{\bar{\pi}}[d^p(\overline{X}, \overline{Y})]$ and conclude that

$$\mathcal{A}\mathcal{W}_p(\mathbb{X}, \mathbb{Y}) = \mathcal{A}\mathcal{W}_p(\overline{\mathbb{X}}, \overline{\mathbb{Y}}).$$

It remains to show the second equality. Write $\mu := \mathcal{L}(\mathrm{ip}_1(\mathbb{X}))$ and $\nu := \mathcal{L}(\mathrm{ip}_1(\mathbb{Y}))$. By Lemma A.1, we have that

$$\mathcal{A}\mathcal{W}_p^p(\overline{\mathbb{X}}, \overline{\mathbb{Y}}) = \inf_{\pi_1 \in \mathrm{Cpl}(\mu,\nu)} \inf_{(k_t)_{t=1}^{N-1}} \int \sum_{t=1}^N d^p(z_t^-, \hat{z}_t^-)(\pi_1 \otimes k_1 \otimes \cdots \otimes k_{N-1})(\mathrm{d}z, \mathrm{d}\hat{z}),$$

where the second infimum is taken over all kernels

$$k_t \colon \mathbb{Z}_{1:t} \times \mathbb{Z}_{1:t} \to \mathcal{P}_p(\mathbb{Z}_{t+1} \times \mathbb{Z}_{t+1}) \quad \text{with } k_t^{z_{1:t}, \hat{z}_{1:t}} \in \mathrm{Cpl}(z_t^+, \hat{z}_t^+). \tag{3.15}$$

Now, for every $1 \le t \le N - 1$, let k_t^* be a kernel as in (3.15), that is, an optimal coupling (with respect to $\mathcal{W}_p$) between its marginals. Their existence follows from a standard measurable selection argument. Then, for every $1 \le t \le N - 1$, $z_{1:t}, \hat{z}_{1:t} \in \mathbb{Z}_{1:t}$, and every kernel k_t as in (3.15), we have that

$$
\begin{aligned}
d^p(z_t, \hat{z}_t) &= d^p(z_t^-, \hat{z}_t^-) + \mathcal{W}_p^p(z_t^+, \hat{z}_t^+) \\
&\le d^p(z_t^-, \hat{z}_t^-) + \int d^p(z_{t+1}, \hat{z}_{t+1}) k_t^{z_{1:t}, \hat{z}_{1:t}}(\mathrm{d}z_{t+1}, \mathrm{d}\hat{z}_{t+1})
\end{aligned}
\tag{3.16}
$$

with equality if $k_t = k_t^*$. In particular, for every $\pi_1 \in \mathrm{Cpl}(\mu, \nu)$, an iterative application of (3.16) shows that

$$\int d^p(z_1, \hat{z}_1) \pi_1(\mathrm{d}z_1, \mathrm{d}\hat{z}_1) \leq \int \sum_{t=1}^{N} d^p(z_t^-, \hat{z}_t^-)(\pi_1 \otimes k_1 \otimes \cdots \otimes k_{N-1})(\mathrm{d}z, \mathrm{d}\hat{z})$$

with equality if $k_t = k_t^*$ for every $1 \leq t \leq N - 1$. Optimizing over $\pi_1 \in \mathrm{Cpl}(\mu, \nu)$ yields the claim. ∎

Definition 3.11 (Wasserstein space of stochastic processes). We call the quotient space

$$\mathrm{FP}_p := \mathcal{FP}_p / {}_{\mathcal{A}} \mathcal{W}_p$$

the Wasserstein space of stochastic processes. The space FP_p is equipped with $\mathcal{A}\mathcal{W}_p$ (which is by Theorem 3.10 well defined on FP_p and independent of the choice of representative).

Our canonical choice of a representative of $\mathbb{X} \in \mathrm{FP}_p$ is the associated canonical process $\overline{\mathbb{X}} \in \mathrm{CFP}_p$. From now on, whenever we use the probability space of (the equivalence class of filtered processes) $\mathbb{X}$, we refer to the filtered probability space provided by $\overline{\mathbb{X}}$ if not stated otherwise.

Corollary 3.12. *The map* $(\mathbb{X}, \mathbb{Y}) \mapsto \mathcal{A}\mathcal{W}_p(\mathbb{X}, \mathbb{Y})$ *is lower semicontinuous with respect to the weak adapted topology*[3] *on* $\mathrm{FP}_p \times \mathrm{FP}_p$.

Proof. The result of the corollary follows from combining [94, Corollary 6.11], that is the 'non-adapted' analogue of Corollary 3.12 from the classical optimal transport theory, with Theorem 3.10. ∎

4. Adapted functions and the prediction process

This section relates the adapted Wasserstein distance to the existing concepts of prediction processes and adapted functions introduced by Knight [70], Aldous [9], and Hoover and Keisler [62]. The main result of this section, Theorem 4.11 below, shows that all concepts induce the same relation on filtered processes.

Before recalling the definition of adapted functions from [60] (see also [62]), let us say that, intuitively, an adapted function is an operation that takes a filtered process as argument and returns a random variable defined on the underlying probability space of this filtered process. Simple examples of adapted functions are $\mathbb{X} \mapsto \sin(X_1)$ and $\mathbb{X} \mapsto \mathbb{E}[\min\{\exp(X_2 X_4), 1\} | \mathcal{F}_2^{\mathbb{X}}]$.

---

[3]The weak adapted topology relates to the adapted Wasserstein distance the same way the weak topology of measures relates to the Wasserstein distance: The weak adapted topology is induced by $\mathcal{A}\mathcal{W}_p$ when, for each $1 \leq t \leq N$, the metric $d_{\mathcal{X}_t}$ is replaced by the bounded metric $d_{\mathcal{X}_t} \wedge 1$.

Definition 4.1 (Adapted functions). We call f an *adapted function – we write $f \in \mathrm{AF}$ –
if it can be built using the following three operations:*

(AF1) If $\Phi \colon \mathcal{X} \to \mathbb{R}$ is continuous bounded, then $\Phi \in \mathrm{AF}$; we set $\Phi(\mathbb{X}) := \Phi(X)$.

(AF2) If $m \in \mathbb{N}$, $f_1, \ldots, f_m \in \mathrm{AF}$, and $\varphi \in C_b(\mathbb{R}^m)$, then $\varphi(f_1, \ldots, f_m) \in \mathrm{AF}$; we set
$\varphi(f_1, \ldots, f_m)(\mathbb{X}) := \varphi(f_1(\mathbb{X}), \ldots, f_m(\mathbb{X}))$.

(AF3) If $1 \leq t \leq N$ and $g \in \mathrm{AF}$, then $(g|t) \in \mathrm{AF}$; we set $(g|t)(\mathbb{X}) := \mathbb{E}[g(\mathbb{X})|\mathcal{F}_t^{\mathbb{X}}]$.

Further, define the *rank* of an adapted function inductively as follows: the rank of Φ is 0;
the rank of $\varphi(f_1, \ldots, f_m)$ is the maximal rank of $f_1, \ldots, f_m$; and the rank of $(g|t)$ is the
rank of g plus 1.

The set of all adapted functions of rank at most $n \in \mathbb{N} \cup \{0\}$ is denoted by $\mathrm{AF}[n]$.

Moreover, we can naturally embed $\mathrm{AF}[n]$ into $\mathrm{AF}[n+1]$ by identifying $f \in \mathrm{AF}[n]$
with $(f|N)$, since $f(\mathbb{X}) = (f|N)(\mathbb{X})$ for all $\mathbb{X} \in \mathcal{FP}_p$. Consequently, we may assume
without loss of generality in item (AF2) that $f_1, \ldots, f_m$ all have the same rank.

Adapted functions were defined in [62] in a continuous-time setting. The present
discrete-time setting permits to give the following, perhaps clearer, representation.

Lemma 4.2. *Let $f \in \mathrm{AF}$ and $n \in \mathbb{N}$. Then $f \in \mathrm{AF}[n]$ if and only if for every $k = 1, \ldots, N$,
there are $m_k \in \mathbb{N}$ and an m_k-dimensional vector $\vec{g}_k$ consisting of elements in $\mathrm{AF}[n-1]$,
and there is $F \in C_b(\mathbb{R}^{\sum_{k=1}^N m_k})$ such that*

$$f(\mathbb{X}) = F(\mathbb{E}[\vec{g}_1(\mathbb{X})|\mathcal{F}_1^{\mathbb{X}}], \ldots, \mathbb{E}[\vec{g}_N(\mathbb{X})|\mathcal{F}_N^{\mathbb{X}}]) \quad \text{for all } \mathbb{X} \in \mathcal{FP}_p. \tag{4.1}$$

Proof. It turns out to be useful to keep track of the depth of an adapted function, a notion
that we now introduce. Loosely speaking, for $f \in \mathrm{AF}[n]$, its depth is the number of
times (AF2) was applied to a *base element* of the form $(g|t)$ with $g \in \mathrm{AF}[n-1]$. The
depth (at rank n) of a 'base element' $(g|t)$ with $g \in \mathrm{AF}[n-1]$ is defined as 0, i.e.,

$$\mathrm{depth}((g|t)) := 0 \quad \text{for all } g \in \mathrm{AF}[n-1] \text{ and } 1 \leq t \leq N. \tag{4.2}$$

Recursively, we assign to $\varphi(f_1, \ldots, f_m) = f \in \mathrm{AF}[n]$ with $\varphi \in C_b(\mathbb{R}^m)$ and $f_i \in \mathrm{AF}[n]$,
$i = 1, \ldots, m$, its depth

$$\mathrm{depth}(f) := \max_{i=1,\ldots,m} \mathrm{depth}(f_i) + 1. \tag{4.3}$$

Note that by the iterative construction of any formation $f \in \mathrm{AF}[n]$, f is either a base
element or of the form detailed in (AF2), and its depth is well defined by (4.2) and (4.3).

Let $f \in \mathrm{AF}[n]$. We begin the induction at depth 0. Then f has $\mathrm{depth}(f) = 0$ if
and only if it is a base element, in which case (4.1) holds true. Now assume that $k :=
\mathrm{depth}(f) > 0$ and that (4.1) applies to all $g \in \mathrm{AF}[n]$ with $\mathrm{depth}(g) < k$. We write $f =
\varphi(f_1, \ldots, f_m)$, where $\varphi \in C_b(\mathbb{R}^m)$ and all $f_i \in \mathrm{AF}[n]$ have depth less than k. By the
inductive hypothesis, for every $1 \leq i \leq m$ there are vectors $\vec{g}_1^i, \ldots, \vec{g}_N^i$ consisting of ele-
ments in $\mathrm{AF}[n-1]$, and $F^i \in C_b(\mathbb{R}^{m^i})$ such that

$$f_i(\mathbb{X}) = F^i(\mathbb{E}[\vec{g}_1^i(\mathbb{X})|\mathcal{F}_1^{\mathbb{X}}], \ldots, \mathbb{E}[\vec{g}_N^i(\mathbb{X})|\mathcal{F}_N^{\mathbb{X}}]) \quad \text{for all } \mathbb{X} \in \mathcal{FP}_p.$$

Collecting and sorting all the terms of the vectors $\vec{g}_t^i$ for $1 \leq t \leq N$ gives

$$\vec{g}_t := \vec{g}_t^{1:m} = (\vec{g}_t^1, \ldots, \vec{g}_t^m).$$

Finally, let σ be the permutation with the property

$$\sigma(\vec{g}_1, \ldots, \vec{g}_N) = (\vec{g}^1, \ldots, \vec{g}^m),$$

then $F = \varphi \circ (F^1, \ldots, F^k) \circ \sigma$ together with $(\vec{g}_1, \ldots, \vec{g}_N)$ satisfies (4.1). ∎

Definition 4.3 (Adapted distribution). Two filtered processes $\mathbb{X}, \mathbb{Y} \in \mathcal{FP}_p$ have the same *adapted distribution (of rank $n \geq 0$)* if $\mathbb{E}[f(\mathbb{X})] = \mathbb{E}[f(\mathbb{Y})]$ for every adapted function $f \in \mathrm{AF}$ (resp. $f \in \mathrm{AF}[n]$); we write $\mathbb{X} \sim_\infty \mathbb{Y}$ (resp. $\mathbb{X} \sim_n \mathbb{Y}$).

Remark 4.4. In the definition of adapted functions, we started in (AF1) with the base set of continuous and bounded functions from $\mathcal{X}$ to $\mathbb{R}$. It is possible to vary this base set, without changing the induced equivalence relations $\sim_\infty$ and $\sim_n$, see Definition 4.3. One may replace (AF1) by any of the following choices:

(AF1a) if $\Phi \colon \mathcal{X} \to \mathbb{R}$ is bounded and Borel measurable, then $\Phi \in \mathrm{AF}$;

(AF1b) if $\Phi \colon \mathcal{X} \to \mathbb{R}$ is bounded and Lipschitz continuous, then $\Phi \in \mathrm{AF}$;

(AF1c) if $1 \leq t \leq N$ and $\Phi \colon \mathcal{X}_t \to \mathbb{R}$ is bounded and continuous, then $\Phi \circ \mathrm{pj}_t \in \mathrm{AF}$.

In a similar manner, we may consider in (AF2) solely Lipschitz continuous/Borel measurable and bounded φ, and still preserve the equivalence relations introduced in Definition 4.3. We shall prove this further down below.

The purpose of the next example is twofold: first, to show where adapted distributions and adapted functions naturally appear, and also to familiarize the reader with the latter.

Example 4.5 (Martingales and optimal stopping). Let $\mathbb{X}, \mathbb{Y} \in \mathcal{FP}$.

(a) If $\mathbb{X}$ is a martingale $\mathbb{Y} \sim_1 \mathbb{X}$, then so is $\mathbb{Y}$ as already observed in [9]. Indeed, for $1 \leq t \leq N$,

$$f_t := |\mathrm{pj}_t - (\mathrm{pj}_{t+1}|t)|$$

is an element of $\mathrm{AF}[1]$.[4] Its evaluation yields

$$\mathbb{E}[|Y_t - \mathbb{E}[Y_{t+1}|\mathcal{F}_t^{\mathbb{Y}}]|] = f_t(\mathbb{Y}) = f_t(\mathbb{X}) = \mathbb{E}[|X_t - \mathbb{E}[X_{t+1}|\mathcal{F}_t^{\mathbb{X}}]|] = 0,$$

that is, the martingale property of $\mathbb{Y}$.

(b) Another important property preserved by $\sim_1$ is Markovianity. This also holds for the important property of being 'plain' defined in (5.1) below.

---

[4]For demonstrative purposes, we disregard that only bounded functions are allowed in the definition of adapted functions. Indeed, this is only a technical issue and all terms are well defined by standard approximation arguments.

(c) Let $c: \mathcal{X} \times \{1, \ldots, N\} \to \mathbb{R}$ be non-anticipative.[5] By the Snell envelope theorem,

$$v_c(\mathbb{X}) := \inf_{\tau \text{ is } (\mathcal{F}_t^{\mathbb{X}})_{t=1}^N\text{-stopping time}} \mathbb{E}[c_\tau(X)] = \mathbb{E}[S_1],$$

where S_1 is defined by backward induction starting with $S_N := c_N(X)$ and

$$S_t := c_t(X) \wedge \mathbb{E}[S_{t+1}|\mathcal{F}_t^{\mathbb{X}}] \quad \text{for } t = N-1, \ldots, 1.$$

Thus, each S_t equals the value of an adapted function of rank $N - t$, from where it follows that $\mathbb{X} \sim_{N-1} \mathbb{Y}$ implies $v_c(\mathbb{X}) = v_c(\mathbb{Y})$.

We will come back to (a) and (c) in Section 5.

Closely related to adapted functions is the *prediction process*.

Definition 4.6 (Prediction process). For $\mathbb{X} \in \mathcal{FP}_p$, the *first-order prediction processes* is given by

$$\mathrm{pp}^1(\mathbb{X}): \Omega^{\mathbb{X}} \to \mathcal{M}_1 := \mathcal{P}_p(\mathcal{X})^N, \quad \omega \mapsto (\mathcal{L}(X|\mathcal{F}_t^{\mathbb{X}})(\omega))_{t=1}^N.$$

Iteratively, the *n-th order prediction process* is given by

$$\mathrm{pp}^n(\mathbb{X}): \Omega^{\mathbb{X}} \to \mathcal{M}_n := \mathcal{P}_p(\mathcal{M}_{n-1})^N, \quad \omega \mapsto (\mathcal{L}(\mathrm{pp}^{n-1}(\mathbb{X})|\mathcal{F}_t^{\mathbb{X}})(\omega))_{t=1}^N.$$

Finally, the *prediction process* is defined as the $\bigtimes_{n \in \mathbb{N}} \mathcal{M}_n =: \mathcal{M}$-valued random variable

$$\mathrm{pp}(\mathbb{X}) := (\mathrm{pp}^n(\mathbb{X}))_{n \in \mathbb{N}}.$$

For convenience, we set the zero-order prediction process $\mathrm{pp}^0(\mathbb{X}) := X$ and $\mathcal{M}_0 := \mathcal{X}$, so that the iterative scheme of Definition 4.6 is valid for $n \geq 0$.

Lemma 4.7. *For every $n \in \mathbb{N}$, there is a continuous function $F^n: \mathcal{Z} \to \mathcal{M}_n$ such that $\mathrm{pp}^n(\mathbb{X}) = F^n(\mathrm{ip}(\mathbb{X}))$ for all $\mathbb{X} \in \mathcal{FP}_p$.*

Proof. Let $F^0: \mathcal{Z} \to \mathcal{M}_0 = \mathcal{X}$ be the corresponding projection, that is, $F^0(z) = z_{1:N}^-$. Therefore, F^0 is continuous with $\mathrm{pp}^0(\mathbb{X}) = X = F^0(\mathrm{ip}(\mathbb{X}))$.

Let $n \in \mathbb{N}$, and consider the inductive hypothesis that there is a continuous map $F^{n-1}: \mathcal{Z} \to \mathcal{M}_{n-1}$ such that

$$\mathrm{pp}^{n-1}(\mathbb{X}) = F^{n-1}(\mathrm{ip}(\mathbb{X})) \quad \text{for all } \mathbb{X} \in \mathcal{FP}_p. \tag{4.4}$$

By definition of the n-th order prediction process and hypothesis (4.4), we have

$$\mathrm{pp}^n(\mathbb{X}) = (\mathcal{L}(F^{n-1}(\mathrm{ip}(\mathbb{X}))|\mathcal{F}_t^{\mathbb{X}}))_{t=1}^N \quad \text{for all } \mathbb{X} \in \mathcal{FP}_p.$$

By Lemma 3.3, for every $1 \leq t \leq N$, there is a continuous map $G_t^n: \mathcal{Z} \to \mathcal{P}_p(\mathcal{M}_{n-1})$ such that

$$\mathcal{L}(F^{n-1}(\mathrm{ip}(\mathbb{X}))|\mathcal{F}_t^{\mathbb{X}}) = G_t^n(\mathrm{ip}(\mathbb{X})) \quad \text{for all } \mathbb{X} \in \mathcal{FP}_p.$$

---

[5]That is, $c_t(x) = c(x, t)$ depends only on $x_{1:t}$ when $1 \leq t \leq N$.

The proof is completed be setting $F^n \colon \mathcal{Z} \to \mathcal{M}_n = \mathcal{P}_p(\mathcal{M}_{n-1})^N$ to be

$$F^n := (G_1^n, \ldots, G_N^n). \qquad \blacksquare$$

It is worth pointing out that pp^n contains 'at least as much information' as its predecessor pp^{n-1}. In fact, we shall later see in Proposition 7.2 that for $n < N - 1$, it contains strictly more information in general.

Lemma 4.8. *Let $1 \le k \le n$. There is a 1-Lipschitz function $F \colon \mathcal{M}_n \to \mathcal{M}_k$ such that*

$$F(\mathrm{pp}^n(\mathbb{X})) = \mathrm{pp}^k(\mathbb{X}) \quad \text{for all } \mathbb{X} \in \mathcal{F}\mathcal{P}_p.$$

In particular, if $\mathbb{X}, \mathbb{Y} \in \mathcal{F}\mathcal{P}_p$ are such that $\mathrm{pp}^n(\mathbb{X})$ and $\mathrm{pp}^n(\mathbb{Y})$ have the same distribution, then $\mathrm{pp}^k(\mathbb{X})$ and $\mathrm{pp}^k(\mathbb{Y})$ have the same distribution as well.

Proof. For $m \ge 0$, consider the isometric injections

$$\iota_m \colon \mathcal{M}_m \to \mathcal{P}_p(\mathcal{M}_m), \quad p \mapsto \delta_p.$$

For $m \ge 1$, since $\mathrm{pp}_N^m(\mathbb{X}) = \delta_{\mathrm{pp}^{m-1}(\mathbb{X})}$, we may apply $\iota_{m-1}^{-1} \circ \mathrm{pj}_N$ to pp^m and obtain $\mathrm{pp}^{m-1}(\mathbb{X}) = \iota_{m-1}^{-1}(\mathrm{pp}_N^m(\mathbb{X}))$ for all $\mathbb{X} \in \mathcal{F}\mathcal{P}_p$. Moreover, ι_m^{-1} admits a 1-Lipschitz extension $I_m^{-1} \colon \mathcal{P}_p(\mathcal{M}_m) \to \mathcal{M}_m$ given by

$$I_m^{-1}(P) := \left(\int p_1 \, P(\mathrm{d}p), \ldots, \int p_N \, P(\mathrm{d}p) \right),$$

where we write $p = (p_1, \ldots, p_N) \in \mathcal{M}_m = \mathcal{P}_p(\mathcal{M}_{m-1})^N$. Indeed, I_m^{-1} is 1-Lipschitz as, for $P, Q \in \mathcal{P}_p(\mathcal{M}_m)$, we have by Jensen's inequality

$$\begin{aligned}
\mathcal{W}_{\mathcal{P}_p(\mathcal{M}_m)}^p(P, Q) &= \int \sum_{t=1}^N \mathcal{W}_{\mathcal{P}_p(\mathcal{M}_{m-1})}^p(p_t, q_t) \pi^*(\mathrm{d}p, \mathrm{d}q) \\
&\ge \sum_{t=1}^N \mathcal{W}_{\mathcal{P}_p(\mathcal{M}_{m-1})}^p \left(\int p_t \pi^*(\mathrm{d}p, \mathrm{d}q), \int q_t \pi^*(\mathrm{d}p, \mathrm{d}q) \right) \\
&= d_{\mathcal{M}_m}^p(I_m^{-1}(P), I_m^{-1}(Q)),
\end{aligned}$$

where π^* is a $\mathcal{W}_{\mathcal{P}_p(\mathcal{M}_m)}$-optimal coupling of P and Q. We denote by

$$F^m \colon \mathcal{M}_m = \mathcal{P}_p(\mathcal{M}_{m-1})^N \to \mathcal{M}_{m-1}$$

the composition $I_{m-1}^{-1} \circ \mathrm{pj}_N$.

Finally, the mapping

$$F := F^{k+1} \circ \cdots \circ F^n \colon \mathcal{M}_n \to \mathcal{M}_k$$

is 1-Lipschitz and satisfies

$$F(\mathrm{pp}^n(\mathbb{X})) = F^{k+1} \circ \cdots \circ F^{n-1}(\mathrm{pp}^{n-1}(\mathbb{X})) = \cdots = \mathrm{pp}^k(\mathbb{X})$$

for all $\mathbb{X} \in \mathcal{F}\mathcal{P}_p$. This completes the proof. $\qquad \blacksquare$

Let us remark that, when $n \geq 1$, the process $\mathrm{pp}^n(\mathbb{X})$ is a measure-valued martingale with respect to $(\mathcal{F}_t^{\mathbb{X}})_{t=1}^N$ which is terminating at $\delta_{\mathrm{pp}^{n-1}(\mathbb{X})}$.

A version of the next lemma can be found in [62], though the proof is different due to differences in the definition of adapted functions (as multi-time stochastic processes).

Lemma 4.9. *Let $\mathbb{X}, \mathbb{Y} \in \mathcal{FP}_p$ and let $n \in \mathbb{N}$. Then the following are equivalent:*

(i) $\mathbb{X} \sim_n \mathbb{Y}$;

(ii) $\mathrm{pp}^n(\mathbb{X})$ *and* $\mathrm{pp}^n(\mathbb{Y})$ *have the same distribution.*

Before proving the lemma, we want to point out that the same proof with obvious modifications also works to obtain Remark 4.4. For example, replace at every instance 'continuous' with 'Borel measurable' for (AF1a).

Proof of Lemma 4.9. We start with the easier direction that (ii) implies (i). Clearly, it is sufficient to show the following.

Claim. *For every $f \in \mathrm{AF}[n]$, there is $F \in C_b(\mathcal{M}_n)$ such that*

$$f(\mathbb{X}) = F(\mathrm{pp}^n(\mathbb{X})) \quad \text{for all } \mathbb{X} \in \mathcal{FP}_p. \tag{4.5}$$

For $n = 0$, the claim is trivially true as $\mathcal{M}_0 = \mathcal{X}$ and due to item (AF1). Assume now that the claim is true for $n \in \mathbb{N} \cup \{0\}$, and let $f \in \mathrm{AF}[n+1]$. Using Lemma 4.2, we may represent f as in (4.1). Thus, it suffices to show (4.5) for $f = (g|t)$, where $g \in \mathrm{AF}[n]$ and $1 \leq t \leq N$. By the inductive hypothesis, there is $G \in C_b(\mathcal{M}_n)$ such that $g(\mathbb{X}) = G(\mathrm{pp}^n(\mathbb{X}))$ for all $\mathbb{X} \in \mathcal{FP}_p$. Therefore,

$$f(\mathbb{X}) = \mathbb{E}[G(\mathrm{pp}^n(\mathbb{X}))|\mathcal{F}_t^{\mathbb{X}}] = \int G \, d\mathrm{pp}_t^{n+1}(\mathbb{X}) = H(\mathrm{pp}^{n+1}(\mathbb{X}))$$

for all $\mathbb{X} \in \mathcal{FP}_p$, where $H \colon \mathcal{M}_{n+1} \to \mathbb{R}$, $p \mapsto \int G \, dp_t$ is continuous and bounded. This shows (4.5) and thus that (ii) implies (i).

We proceed to show that (i) implies (ii). To that end, we interject two preliminary statements. Define $\mathcal{S}_0 := C_b(\mathcal{M}_0)$ and inductively define $\mathcal{S}_n \subset C_b(\mathcal{M}_n)$ as the set of all functions of the form

$$p \mapsto \psi\left(\int \vec{G}_1 \, dp_1, \ldots, \int \vec{G}_N \, dp_N\right), \tag{4.6}$$

where, for every $1 \leq t \leq N$, $\vec{G}_t$ is a vector of functions in $\mathcal{S}_{n-1}$ and $\psi \colon \mathbb{R}^m \to \mathbb{R}$ (with adequate m) is continuous and bounded.

Claim. *The set $\mathcal{S}_n$ is an algebra which separates points in $\mathcal{M}_n$.*

As usual, we proceed by induction. The claim follows trivially for $n = 0$ since $\mathcal{S}_0 = C_b(\mathcal{X})$. Assume now that the claim is true for n. Clearly, $\mathcal{S}_{n+1}$ is an algebra. To see the second part of the claim, namely that it separates points, let $p = (p_t)_{t=1}^N$ and $q = (q_t)_{t=1}^N$ be two distinct elements of $\mathcal{M}_{n+1}$, that is, $p_{t_0} \neq q_{t_0}$ for some t_0. By the inductive

hypothesis, $\mathcal{S}_n$ is an algebra which separates points in $\mathcal{M}_n$, therefore [48, Theorem 4.5] provides $G \in \mathcal{S}_n$ with

$$\int G \, \mathrm{d}p_t \neq \int G \, \mathrm{d}q_t,$$

whence, $\mathcal{S}_{n+1}$ separates points in $\mathcal{M}_{n+1}$.

Claim. *For $F \in \mathcal{S}_n$, there is $f \in \mathrm{AF}[n]$ with*

$$F(\mathrm{pp}^n(\mathbb{X})) = f(\mathbb{X}) \quad \text{for all } \mathbb{X} \in \mathcal{FP}_p. \tag{4.7}$$

Again, the assertion is trivial for $n = 0$ as $\mathrm{pp}^0(\mathbb{X}) = X$. Assume that the claim holds for n, and let $F \in \mathcal{S}_{n+1}$ be represented by ψ and $\vec{G}_1, \ldots, \vec{G}_N$ as in (4.6). By definition of the prediction process pp^n, we have

$$F(\mathrm{pp}^{n+1}(\mathbb{X})) = \psi(\mathbb{E}[\vec{G}_1(\mathrm{pp}^n(\mathbb{X}))|\mathcal{F}_1^{\mathbb{X}}], \ldots, \mathbb{E}[\vec{G}_N(\mathrm{pp}^n(\mathbb{X}))|\mathcal{F}_N^{\mathbb{X}}])$$

for all $\mathbb{X} \in \mathcal{FP}_p$. By assumption, there are vectors $\vec{g}_t$ of adapted functions in $\mathrm{AF}[n]$ with

$$\vec{G}_t(\mathrm{pp}^n(\mathbb{X})) = \vec{g}_t(\mathbb{X}) \quad \text{for all } \mathbb{X} \in \mathcal{FP}.$$

Similar to the proof of Lemma 4.2, we collect all terms and obtain some $f \in \mathrm{AF}[n+1]$ with $F(\mathrm{pp}^{n+1}(\mathbb{X})) = f(\mathbb{X})$, which shows the claim.

With our two preliminary claims already established, we are ready to show that (i) implies (ii). By (4.7), we have that

$$\mathbb{E}[F(\mathrm{pp}^n(\mathbb{X}))] = \mathbb{E}[F(\mathrm{pp}^n(\mathbb{Y}))] \quad \text{for all } F \in \mathcal{S}_n.$$

As $\mathcal{S}_n$ is an algebra which separates points, it follows, e.g., from [48, Theorem 4.5] that $\mathrm{pp}^n(\mathbb{X})$ and $\mathrm{pp}^n(\mathbb{Y})$ have the same distribution. ∎

Lemma 4.10. *There is a 1-Lipschitz map $F: \mathcal{M}_{N-1} \to \mathcal{Z}$ such that*

$$F(\mathrm{pp}^{N-1}(\mathbb{X})) = \mathrm{ip}(\mathbb{X}) \quad \text{for all } \mathbb{X} \in \mathcal{FP}_p.$$

Proof. By Lemma 4.8, there are 1-Lipschitz maps $G^{k,n}: \mathcal{M}_n \to \mathcal{M}_k$, $k < n$ with $G^{k,n} \circ \mathrm{pp}^n = \mathrm{pp}^k$.

Claim. *For $1 \leq t \leq N$ there is a 1-Lipschitz map $F^t: \mathcal{M}_{N-t} \to \mathcal{Z}_t$ with*

$$F^t(\mathrm{pp}^{N-t}(\mathbb{X})) = \mathrm{ip}_t(\mathbb{X}) \quad \text{for all } \mathbb{X} \in \mathcal{FP}_p. \tag{4.8}$$

Clearly, (4.8) is satisfied when $t = N$. Indeed, $\mathrm{ip}_N = \mathrm{pp}_N^0$ whereby $F^N := \mathrm{pj}_N$ fulfills (4.8).

To establish (4.8) for general t, we proceed by induction. Assuming that the claim holds true for $2 \leq t \leq N$, we find by the definition of the information process, see Definition 5.9, for $\mathbb{X} \in \mathcal{FP}_p$ that

$$\begin{aligned}
\mathrm{ip}_{t-1}(\mathbb{X}) &= (X_{t-1}, \mathcal{L}(\mathrm{ip}_t(\mathbb{X})|\mathcal{F}_{t-1}^{\mathbb{X}})) = (\mathrm{pp}_{t-1}^0(\mathbb{X}), \mathcal{L}(F^t(\mathrm{pp}^{N-t}(\mathbb{X}))|\mathcal{F}_{t-1}^{\mathbb{X}})) \\
&= (\mathrm{pp}_{t-1}^0(\mathbb{X}), (F^t)_*(\mathrm{pp}_{t-1}^{N-t+1}(\mathbb{X}))) \\
&= (\mathrm{pj}_{t-1} \circ G^{0,N-t+1}, \mathrm{pj}_{t-1} \circ H^{t-1}) \circ \mathrm{pp}^{N-t+1}(\mathbb{X}),
\end{aligned}$$

where $H^{t-1}: \mathcal{M}_{N-t+1} \to \mathscr{P}_p(\mathbb{Z}_t)$ is given by

$$H^{t-1}(p) := (F_*^t p_1, \ldots, F_*^t p_N).$$

Since F^t is 1-Lipschitz by assumption, the same holds true for H^{t-1}. Therefore, $F^{t-1} := (\mathrm{pj}_{t-1} \circ G^{0,N-t+1}, \mathrm{pj}_{t-1} \circ H^{t-1})$ is 1-Lipschitz and satisfies (4.8), which yields the claim.

Finally, by the previously shown claim, the map

$$F := (F^1, F^2 \circ G^{N-2,N-1}, \ldots, F^N \circ G^{0,N-1})$$

has the desired properties. ∎

Theorem 4.11. *Let* $\mathbb{X}, \mathbb{Y} \in \mathcal{F}\mathscr{P}_p$. *All of the following are equivalent:*

(i) $\mathbb{X} \sim_\infty \mathbb{Y}$.

(ii) $\mathbb{X} \sim_{N-1} \mathbb{Y}$.

(iii) $\mathrm{pp}(\mathbb{X})$ *and* $\mathrm{pp}(\mathbb{Y})$ *have the same distribution.*

(iv) $\mathrm{pp}^{N-1}(\mathbb{X})$ *and* $\mathrm{pp}^{N-1}(\mathbb{Y})$ *have the same distribution.*

(v) $\mathrm{ip}(\mathbb{X})$ *and* $\mathrm{ip}(\mathbb{Y})$ *have the same distribution.*

(vi) $\mathrm{ip}_1(\mathbb{X})$ *and* $\mathrm{ip}_1(\mathbb{Y})$ *have the same distribution.*

(vii) $\mathcal{A}\mathcal{W}_p(\mathbb{X}, \mathbb{Y}) = 0$.

In Proposition 7.2, we shall further prove that for every $1 \le n \le N-1$, the relation $\sim_n$ strictly refines $\sim_{n-1}$: there are $\mathbb{X}, \mathbb{Y} \in \mathcal{F}\mathscr{P}_p$ with $\mathbb{X} \sim_{n-1} \mathbb{Y}$ but $\mathbb{X} \not\sim_n \mathbb{Y}$ (and especially $\mathcal{A}\mathcal{W}_p(\mathbb{X}, \mathbb{Y}) > 0$). Importantly, these refinements are essential even for seemingly simple applications as we shall show in Theorem 7.1: Only the relation $\mathbb{X} \sim_{N-1} \mathbb{Y}$ guarantees that two processes $\mathbb{X}$ and $\mathbb{Y}$ have the same values for optimal stopping problems.

Proof of Theorem 4.11. In a first step, note that (i) implies (ii); that (iii) implies (iv); and that (v) implies (vi). Further, Lemma 4.9 shows that (i) and (iii) are equivalent and that (ii) and (iv) are equivalent. Theorem 3.10 shows that (v) and (vii) are equivalent. Lemma 4.7 shows that (v) implies (iii). Finally, Lemma 4.10 shows that (iv) implies (v). This concludes the proof. ∎

5. Topological and geometric properties of FP$_p$

5.1. Compactness in FP$_p$

To develop a comprehensive understanding of a topology, it is essential to get a hold on compact sets. For the weak topology, this is bestowed on us by Prokhorov's theorem which gives an easy-to-check tightness-criterion for relative compactness. Theorem 5.1 implies that, perhaps surprisingly, the very same tightness-criterion also implies relative compactness for stochastic processes in FP$_p$.

Theorem 5.1 (Prokhorov's theorem). *For a subset* $\Pi \subseteq \mathrm{FP}_p$, *the following are equivalent:*

(i) Π *is relatively compact in* FP_p.

(ii) $\{\mathcal{L}(X) : \mathbb{X} \in \Pi\}$ *is relatively compact in* $\mathcal{P}_p(\mathcal{X})$.

It is worthwhile to recall that condition (ii) is equivalent to tightness plus uniform integrability (see, e.g., [94]), that is, for every $x_0 \in \mathcal{X}$ and $\varepsilon > 0$, there is a compact set $K \subset \mathcal{X}$ such that

$$\sup_{\mathbb{X} \in \Pi} \mathbb{E}[(1 + d^p(x_0, X))1_{\{X \notin K\}}] \leq \varepsilon.$$

As a consequence of the nested structure of $\mathcal{Z}_1$, the following *intensity operator* plays an important role in the proof of Theorem 5.1. For two Polish spaces $\mathcal{A}$ and $\mathcal{B}$, we define $\hat{I}: \mathcal{P}_p(\mathcal{A} \times \mathcal{P}_p(\mathcal{B})) \to \mathcal{P}_p(\mathcal{A} \times \mathcal{B})$ via

$$\int f(a, b)\hat{I}(\pi)(\mathrm{d}a, \mathrm{d}b) = \iint f(a, b)p(\mathrm{d}b)\pi(\mathrm{d}a, \mathrm{d}p)$$

for $f \in C_b(\mathcal{A} \times \mathcal{B})$. The intensity map $\hat{I}$ closely relates relatively compact sets of its domain and its range in the sense of the subsequent lemma.

Lemma 5.2 (cf. [31, Lemma 5.7]). *Let $\mathcal{A}$ and $\mathcal{B}$ be two Polish spaces. For $\Pi \subseteq \mathcal{P}_p(\mathcal{A} \times \mathcal{P}_p(\mathcal{B}))$, the following are equivalent:*

(i) $\Pi \subseteq \mathcal{P}_p(\mathcal{A} \times \mathcal{P}_p(\mathcal{B}))$ *is relatively compact.*

(ii) $\hat{I}(\Pi) \subseteq \mathcal{P}_p(\mathcal{A} \times \mathcal{B})$ *is relatively compact.*

Proof of Theorem 5.1. As we know by Theorem 3.10 that FP_p is isometrically isomorphic to $\mathcal{P}_p(\mathcal{Z}_1)$, we obtain that Π is relatively compact if and only if the set $\{\mathcal{L}(\mathrm{ip}_1(\mathbb{X})) : \mathbb{X} \in \Pi\}$ is relatively compact. On the other hand, note that for $1 \leq t < N$, $\mathcal{A} := \mathcal{X}_{1:t}$ and $\mathcal{B} := \mathcal{P}_p(\mathcal{Z}_{t+1})$, we have by the nested definition of ip_t that for all $\mathbb{X} \in \mathrm{FP}_p$,

$$\hat{I}(\mathcal{L}(X_{1:t-1}, \mathrm{ip}_t(\mathbb{X}))) = \mathcal{L}(X_{1:t}, \mathrm{ip}_{t+1}(\mathbb{X})).$$

Applying Lemma 5.2 yields equivalence of the following statements:

- $\{\mathcal{L}(X_{1:t-1}, \mathrm{ip}_t(\mathbb{X})) : \mathbb{X} \in \Pi\}$ is relatively compact;
- $\{\mathcal{L}(X_{1:t}, \mathrm{ip}_t(\mathbb{X})) : \mathbb{X} \in \Pi\}$ is relatively compact.

Hence, by applying this argument iteratively, we find that $\{\mathcal{L}(X) : \mathbb{X} \in \Pi\}$ is relatively compact if and only if $\{\mathcal{L}(\mathrm{ip}_1(\mathbb{X})) : \mathbb{X} \in \Pi\}$ is relatively compact, which we wanted to show. The final assertion is a direct consequence of the classical Prokhorov's theorem and the characterization of Wasserstein convergence in $\mathcal{P}_p(\mathcal{A} \times \mathcal{B})$, see [94, Definition 5.8]. ∎

5.2. Denseness of simple processes

A canonical way of embedding $\mathcal{P}_p(\mathcal{X})$ into FP_p is the following: we can associate to each law $\mathbb{P} \in \mathcal{P}_p(\mathcal{X})$ the processes

$$\mathbb{X} \equiv (\mathcal{X}, \mathcal{B}(\mathcal{X}), \mathbb{P}, (\sigma(X_{1:t}))_{t=1}^N, X), \tag{5.1}$$

where X denotes the coordinate process on $\mathcal{X}$, and call this type of process *plain*. The set of all plain processes is denoted by $\Lambda_{\text{plain}} \subseteq \text{FP}_p$.

Proposition 5.3. *Let $(\Omega, \mathcal{F}, \mathbb{Q})$ be an arbitrary probability space, let $Y \colon \Omega \to \mathcal{X}$ be an $\mathcal{F}$-measurable map such that $\mathcal{L}(Y) \in \mathcal{P}_p(\mathcal{X})$, and denote by $\mathbb{X}$ the plain process associated to $\mathcal{L}(Y)$; i.e., $\mathbb{X}$ is given by* (5.1) *with $\mathbb{P} = \mathcal{L}(Y)$. Then*

$$\widetilde{\mathbb{X}} := (\Omega, \mathcal{F}, \mathbb{Q}, (\sigma(Y_{1:t}))_{t=1}^N, Y)$$

satisfies $\mathcal{A}W_p(\mathbb{X}, \widetilde{\mathbb{X}}) = 0$. In particular, if $\hat{\mathbb{X}}, \hat{\mathbb{Y}} \in \text{FP}_p$ are plain and $\mathcal{L}(\hat{X}) = \mathcal{L}(\hat{Y})$, then $\hat{\mathbb{X}} = \hat{\mathbb{Y}}$.

Proof. The coupling $(\text{id}_\Omega, Y)_* \mathbb{Q}$ is bicausal between $\widetilde{\mathbb{X}}$ and $\mathbb{X}$ as well as $\widetilde{\mathbb{Y}}$ and $\mathbb{Y}$. Thus, $\mathcal{A}W_p(\mathbb{X}, \widetilde{\mathbb{X}}) = 0 = \mathcal{A}W_p(\mathbb{Y}, \widetilde{\mathbb{Y}})$.   ∎

Among others, the purpose of this subsection is to show that the space of filtered processes FP_p naturally appears as the completion of all plain processes Λ_{plain}.

We call a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ *finite* if Ω consists of finitely many elements.

Theorem 5.4. *If $\mathcal{X}$ has no isolated points, then the set*

$$\{\mathbb{X} \in \text{FP}_p : \mathbb{X} \text{ is Markov and has a representative on a finite probability space}\}$$

is dense in FP_p. In particular, the plain processes Λ_{plain} form a dense subset.

This theorem follows from the next proposition.

Proposition 5.5. *Let $\mathbb{X} \in \mathcal{F}\mathcal{P}_p$ and let $\varepsilon > 0$. Then there is*

$$\mathbb{Y} := (\Omega^{\mathbb{X}}, \mathcal{F}^{\mathbb{X}}, \mathbb{P}^{\mathbb{X}}, (\mathcal{F}_t^{\mathbb{Y}})_{t=1}^N, Y) \in \mathcal{F}\mathcal{P}_p, \tag{5.2}$$

where, for every $1 \le t \le N$, Y_t is a function of $\text{ip}_t(\mathbb{X})$ and $\mathcal{F}_t^{\mathbb{Y}}$ is a finite subset of $\mathcal{F}_t^{\mathbb{X}}$, such that

$$\mathcal{A}W_p(\mathbb{X}, \mathbb{Y}) < \varepsilon.$$

In particular, if $\mathcal{X}$ has no isolated points, then $\mathbb{Y}$ can be chosen Markovian.

The proof relies on the following result, which is essentially shown in [32, Lemma 4.8]. Recall here that for $\mu, \nu \in \mathcal{P}_p(\mathcal{X})$, the term $\mathcal{A}W_p(\mu, \nu)$ refers to $\mathcal{A}W_p(\mathbb{X}, \mathbb{Y})$, where $\mathbb{X}$ and $\mathbb{Y}$ are plain processes, see (5.1), distributed according to μ and ν, respectively. In a similar fashion, we will use here $\text{Cpl}_{\text{bc}}(\mu, \nu)$ to denote the set of all bicausal couplings between the corresponding plain processes.

Lemma 5.6. *For each $1 \le t \le N$, let $(\mathcal{Y}_t^m)_{m \in \mathbb{N}}$ be an increasing family of finite subsets of Polish spaces $\mathcal{Y}_t$ such that $\bigcup_{m \in \mathbb{N}} \mathcal{Y}_t^m$ is dense in $\mathcal{Y}_t$, and let $\varphi_t^m \colon \mathcal{Y}_t \to \mathcal{Y}_t^m$ be the map which assigns each point its nearest point in $\mathcal{Y}_t^m$ (with ties broken arbitrarily but measurably). Then, for any $\mu \in \mathcal{P}_p(\mathcal{Y}_{1:N})$, we have*

$$\lim_{m \to \infty} \mathcal{A}W_p(\mu, (\varphi_{1:N}^m)_* \mu) = 0. \tag{5.3}$$

Proof. For every m, set

$$\mu^m := (\varphi^m_{1:N})_* \mu = (x_{1:N} \mapsto (\varphi^m_1(x_1), \ldots, \varphi^m_N(x_N)))_* \mu.$$

By [14, Lemma 1.4], (5.3) is equivalent to

$$\inf_{\pi \in \mathrm{Cpl}_{\mathrm{bc}}(\mu, \mu^m)} \int d(x, y) \wedge 1 \, \pi(\mathrm{d}x, \mathrm{d}y) \to 0 \tag{5.4}$$

and

$$\int d^p(x, x^0) \mu^m(\mathrm{d}x) \to \int d^p(x, x^0) \mu(\mathrm{d}x), \tag{5.5}$$

as $m \to \infty$, where $x^0 \in \mathcal{Y}_{1:N}$ is some arbitrary but fixed element. For convenience, we choose $x^0 \in \mathcal{Y}^0_{1:N}$.

If each $\mathcal{Y}_t$ were compact, (5.4) would follow directly from [32, Lemma 4.8]. However, compactness in [32, Lemma 4.8] was only used to additionally obtain the rate of convergence, and the same proof shows that in the present setting (5.4) holds true. As for (5.5), note that the reverse triangle inequality shows that

$$|d(\varphi^m_{1:N}(x), x^0) - d(x, x^0)| \leq d(\varphi^m_{1:N}(x), x)$$

for every $x \in \mathcal{Y}_{1:N}$. Thus, $d(\varphi^m_{1:N}(x), x)$ decreases to zero as $m \to \infty$ and is bounded by $d(x^0, x)$ due to the definition of φ^m and $x^0 \in \mathcal{Y}^0_{1:N}$. Dominated convergence shows that (5.5) holds and completes the proof. ∎

Proof of Proposition 5.5. Consider $\mu := \mathcal{L}(\mathrm{ip}(\mathbb{X})) \in \mathcal{P}_p(\mathcal{Z})$ and interpret $\mathcal{Z} = \mathcal{Z}_{1:N}$ as a path space. By Corollary 5.6, there is a family of laws $(\mu^m)_{m \in \mathbb{N}}$ on $\mathcal{P}_p(\mathcal{Z})$ with

$$\lim_{m \to \infty} \mathcal{A}W_p(\mu, \mu^m) = 0$$

and for every $m \in \mathbb{N}$, μ^m is finitely supported.

Fix $\varepsilon > 0$ and $m \in \mathbb{N}$ such that $\mathcal{A}W_p(\mu, \mu^m) < \varepsilon$. Let $\varphi^m_t \colon \mathcal{Z}_t \to \mathcal{Z}_t$, $1 \leq t \leq N$, be the family of maps introduced in Corollary 5.6. We write $\bar{Y}_t := \mathrm{pj}_{\mathcal{X}_t} \circ \varphi^m_{1:N}$. The maps $\varphi^m_t(\mathrm{ip}_t(\mathbb{X}))$ and $Y_t := \bar{Y}_t(\mathrm{ip}_t(\mathbb{X}))$ are both $\mathcal{F}^{\mathbb{X}}_t$-measurable and

$$\mathcal{F}^{\mathbb{Y}}_t := \sigma(\varphi^m_s(\mathrm{ip}_s(\mathbb{X})) : 1 \leq s \leq t)$$

is a finite sub-σ-algebra of $\mathcal{F}^{\mathbb{X}}_t$. The filtered process $\mathbb{Y}$ is given as in (5.2) with process Y and filtration $(\mathcal{F}^{\mathbb{Y}}_t)^N_{t=1}$. By virtue of Lemma 2.2, it is readily verified that the coupling $(\mathrm{id}_{\Omega^{\mathbb{X}}}, \varphi^m_t(\mathrm{ip}_t(\mathbb{X}))^N_{t=1})_* P^{\mathbb{X}}$ is bicausal between $\mathbb{Y}$ and

$$\bar{\mathbb{Y}} := (\mathcal{Z}, \mathcal{F}^{\mathcal{Z}}, \mu, (\sigma(\varphi^m_{1:t}))^N_{t=1}, \bar{Y}_{1:N}),$$

whence $\mathcal{A}W_p(\bar{\mathbb{Y}}, \mathbb{Y}) = 0$. Similarly, the coupling $(\mathrm{id}_{\mathcal{Z}}, \varphi^m_{1:N})_* \mu$ is bicausal between $\bar{\mathbb{Y}}$ and

$$\mathbb{Z} := (\mathcal{Z}, \mathcal{F}^{\mathcal{Z}}, \mu^m, (\mathcal{F}^{\mathcal{Z}}_t)^N_{t=1}, Z), \quad \text{where } Z_t = \mathrm{pj}_{\mathcal{X}_t},$$

thus, $\mathcal{A}\mathcal{W}_p(\overline{\mathbb{Y}}, \mathbb{Z}) = 0$ and $\mathcal{A}\mathcal{W}_p(\mathbb{Y}, \mathbb{Z}) = 0$. By Theorem 3.10, we conclude

$$\mathcal{A}\mathcal{W}_p(\mathbb{X}, \mathbb{Y}) = \mathcal{A}\mathcal{W}_p(\mathbb{X}, \mathbb{Z}) = \left(\inf_{\pi \in \mathrm{Cpl}_{\mathrm{bc}}(\mu, \mu^m)} \int d^p(z^-, \hat{z}^-) \pi(\mathrm{d}z, \mathrm{d}\hat{z}) \right)^{1/p}$$

$$\leq \left(\inf_{\pi \in \mathrm{Cpl}_{\mathrm{bc}}(\mu, \mu^m)} \int d^p(z, \hat{z}) \pi(\mathrm{d}z, \mathrm{d}\hat{z}) \right)^{1/p} = \mathcal{A}\mathcal{W}_p(\mu, \mu^m) < \varepsilon. \qquad \blacksquare$$

5.3. Martingales

Assume for this subsection that $\mathcal{X}_t = \mathbb{R}^d$ for each $1 \leq t \leq N$.

Proposition 5.7 (Martingales). *The set of all martingales*

$$\mathrm{M}_p := \{\mathbb{X} \in \mathrm{FP}_p : \mathbb{X} \text{ is a martingale}\}$$

is closed with respect to $\mathcal{A}\mathcal{W}_p$.

There is a multitude of ways how to prove Proposition 5.7:

(i)   as a consequence of the continuity of Doob decomposition (Proposition 6.8 below),

(ii)   as a consequence of the continuity of optimal stopping,

(iii) as a consequence of Example 4.5,

(iv)   by characterizing martingales as those processes $\mathbb{X} \in \mathrm{FP}_1$ for which $\mathcal{L}(\mathrm{ip}_1(\mathbb{X}))$ is concentrated on a particular closed subset of $\mathcal{Z}_1$,

(v)   or directly by coupling arguments.

We will present the last variant.

Proof of Proposition 5.7. Let $(\mathbb{X}^n)_{n \in \mathbb{N}}$ be a sequence in M_p converging to $\mathbb{X}$. Fix $n \in \mathbb{N}$, let $\pi \in \mathrm{Cpl}_{\mathrm{bc}}(\mathbb{X}^n, \mathbb{X})$ and $1 \leq t \leq s \leq N$. By Lemma 2.2 and the martingale property of $\mathbb{X}^n$, we have

$$\mathbb{E}[X_s|\mathcal{F}_t^{\mathbb{X}}] = \mathbb{E}_\pi[X_s|\mathcal{F}_{t,t}^{\mathbb{X},\mathbb{X}^n}] \quad \text{and} \quad X_t^n = \mathbb{E}_\pi[X_s^n|\mathcal{F}_{t,t}^{\mathbb{X},\mathbb{X}^n}].$$

Thus, letting $\Delta^n := (X_t^n - X_t) + (X_s^n - X_s)$, Jensen's inequality yields that

$$\mathbb{E}[|X_t - \mathbb{E}[X_s|\mathcal{F}_t^{\mathbb{X}}]|] = \mathbb{E}_\pi[|X_t^n - \mathbb{E}_\pi[X_s^n + \Delta^n|\mathcal{F}_{t,t}^{\mathbb{X},\mathbb{X}^n}]|]$$

$$\leq \mathbb{E}_\pi[|\Delta^n|] \leq \mathbb{E}_\pi[|X^n - X|].$$

As $\pi \in \mathrm{Cpl}_{\mathrm{bc}}(\mathbb{X}^n, \mathbb{X})$ was arbitrary, $\mathbb{E}[|X_t - \mathbb{E}[X_s|\mathcal{F}_t^{\mathbb{X}}]|]$ is dominated by $\mathcal{A}\mathcal{W}_1(\mathbb{X}^n, \mathbb{X})$ which is arbitrarily small for n sufficiently large. Hence, $\mathbb{E}[|X_t - \mathbb{E}[X_s|\mathcal{F}_t^{\mathbb{X}}]|] = 0$ showing that $\mathbb{X}$ is a martingale. $\qquad \blacksquare$

5.4. $(\mathrm{FP}_p, \mathcal{A}\mathcal{W}_p)$ is a geodesic space

The purpose of this section is to show Theorem 5.10, and in particular that $(\mathrm{FP}_p, \mathcal{A}\mathcal{W}_p)$ is a geodesic space. For concise notation, we shall assume throughout that $\mathcal{X}_t = \mathbb{R}^d$ for every $1 \leq t \leq N$ (but see also Remark 5.13).

Definition 5.8 (Constant speed geodesic). A family $(\mathbb{Z}^u)_{u\in[0,1]}$ in FP_p is said to be a *constant speed geodesic* connecting $\mathbb{X}, \mathbb{Y} \in \mathrm{FP}_p$ if

(a) $\mathbb{Z}^0 = \mathbb{X}$ and $\mathbb{Z}^1 = \mathbb{Y}$,

(b) $\mathcal{AW}_p(\mathbb{Z}^u, \mathbb{Z}^v) = |u - v|\mathcal{AW}_p(\mathbb{X}, \mathbb{Y})$ for all $u, v \in [0, 1]$.

A tangible way of defining constant speed geodesics is – in analogy to the classical $\mathcal{W}_p$-displacement interpolation – by means of geodesics on the state space and optimal couplings. To that end, recall that $\mathbb{Z}$ is the canonical space defined in Definition 3.1.

Definition 5.9 (Interpolation process). Let $\mathbb{X}, \mathbb{Y} \in \mathrm{FP}_p$ and let $\pi \in \mathrm{Cpl}_{\mathrm{bc}}(\mathbb{X}, \mathbb{Y})$. We call the family $(\mathbb{Z}^{\pi,u})_{u\in[0,1]}$ given by

$$\mathbb{Z}^{\pi,u} := (\mathcal{Z} \times \mathcal{Z}, \mathcal{F}^{\mathcal{Z}} \otimes \mathcal{F}^{\mathcal{Z}}, \pi, (\mathcal{F}_{t,t}^{\mathcal{Z},\mathcal{Z}})_{t=1}^N, ((1-u)X_t + uY_t)_{t=1}^N)$$

the *interpolation process* between $\mathbb{X}$ and $\mathbb{Y}$ (with respect to the coupling π).

The following is the main result of this section.

Theorem 5.10 (Filtered processes form a geodesic space). *Let $p \in (1, \infty)$.*

(i) *The space $(\mathrm{FP}_p, \mathcal{AW}_p)$ is a geodesic space, that is, for every $\mathbb{X}, \mathbb{Y} \in \mathrm{FP}_p$ there is a constant speed geodesic connecting them.*

(ii) *A family $(\mathbb{Z}^u)_{u\in[0,1]}$ in FP_p is a constant speed geodesic between $\mathbb{X}, \mathbb{Y} \in \mathrm{FP}_p$ if and only if the family*

$$(\gamma^u)_{u\in[0,1]} := (\mathcal{L}(\mathrm{ip}_1(\mathbb{Z}^u)))_{u\in[0,1]}$$

is a $\mathcal{W}_p$-constant speed geodesic[6] between $\mathcal{L}(\mathrm{ip}_1(\mathbb{X}))$ and $\mathcal{L}(\mathrm{ip}_1(\mathbb{Y}))$.

(iii) *If π is an optimal bicausal coupling for $\mathcal{AW}_p(\mathbb{X}, \mathbb{Y})$, then the interpolation process $(\mathbb{Z}^{\pi,u})_{u\in[0,1]}$ is a constant speed geodesic between $\mathbb{X}$ and $\mathbb{Y}$.*

In a forthcoming paper, it will be shown that not only does the interpolation process constitute a geodesic, but actually *all* geodesics can be described as interpolation processes, at least once one is willing to allow for external randomization, that is, extending the probability space by independent randomness.

Proof of Theorem 5.10. We start by proving (i) and (ii). As $\mathcal{Z}_N = \mathcal{X}_N$ is a geodesic space, it follows from [74] that $\mathcal{P}_p(\mathcal{X}_N)$ is a geodesic space, too. Moreover, the product (endowed with the l^p-norm) of two geodesic spaces remains geodesic, hence

$$\mathcal{Z}_{N-1} = \mathcal{X}_{N-1} \times \mathcal{P}_p(\mathcal{Z}_N)$$

is a geodesic space. Repeating this argument inductively shows that $\mathcal{P}_p(\mathcal{Z}_1)$ is a geodesic space. Claims (i) and (ii) now follow from the isometry between FP_p and $\mathcal{P}_p(\mathcal{Z}_1)$ given in Theorem 3.10.

---

[6]A family $(\gamma^u)_{u\in[0,1]}$ in $\mathcal{P}_p(\mathcal{Z}_1)$ is said to be a $\mathcal{W}_p$-constant speed geodesic connecting $\mu, \nu \in \mathcal{P}(\mathcal{Z}_1)$ if $\gamma^0 = \mu$, $\gamma^1 = \nu$, and $\mathcal{W}_p(\gamma^u, \gamma^v) = |u - v|\mathcal{W}_p(\mu, \nu)$ for every $u, v \in [0, 1]$.

We now show (iii). Bicausality of π and Lemma 2.2 immediately show part (a) of Definition 5.8, and it remains to deal with part (b). To that end, note that the coupling $\Pi := (\mathrm{id}, \mathrm{id})_* \pi$ is bicausal between $\mathbb{Z}^{\pi,u}$ and $\mathbb{Z}^{\pi,v}$. Then, for every $u, v \in [0, 1]$, as $Z^u - Z^v = (u - v)(X - Y)$, we compute

$$\mathcal{A}\mathcal{W}_p(\mathbb{Z}^{\pi,u}, \mathbb{Z}^{\pi,v}) \leq \mathbb{E}_\Pi[\|Z^u - Z^v\|_p^p]^{1/p} = |u - v|\mathbb{E}_\pi[\|X - Y\|_p^p]^{1/p}$$
$$= |u - v|\mathcal{A}\mathcal{W}_p(\mathbb{X}, \mathbb{Y}), \tag{5.6}$$

where the last equality holds by optimality of π. A straightforward application of the triangle inequality shows that there cannot be a strict inequality in (5.6), and whence the claim follows. ∎

The next example illustrates that even in the case of geodesics between plain processes (cf. (5.1)) it is necessary to consider general filtrations instead of the filtrations generated by the processes (see also Figure 3).
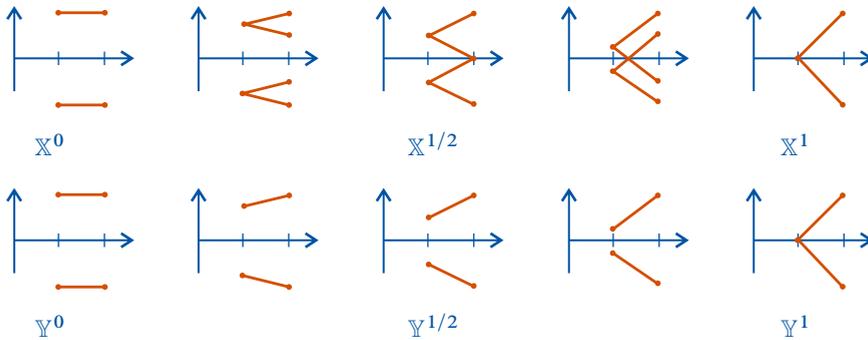
Fig. 3. Comparison of the adapted Wasserstein interpolation $(\mathbb{X}^u)_{u \in [0,1]}$ and a classical Wasserstein interpolation $(\mathbb{Y}^u)_{u \in [0,1]}$ where the geodesic between martingales does not consist of martingales.

Example 5.11. Let $N = 2$. We consider two processes with paths in $\mathbb{R}^2$: $\mathbb{X}$ and $\mathbb{Y}$ are the plain process associated to the laws (on the path space $\mathbb{R}^2$)

$$\mu := \frac{1}{2}(\delta_{(1,-2)} + \delta_{(-1,2)}) \quad \text{and} \quad \nu := \frac{1}{2}(\delta_{(1,1)} + \delta_{(-1,-1)}),$$

respectively. It is then easy to verify that there exists a unique constant speed geodesic and that it is given by the interpolation process of $\mathbb{X}$ and $\mathbb{Y}$ with respect to the unique $\mathcal{A}\mathcal{W}_p$-optimal coupling $\pi \in \mathrm{Cpl}_{\mathrm{bc}}(\mathbb{X}, \mathbb{Y})$. This coupling sends the mass from $(1, -2)$ to $(-1, -1)$ and the mass from $(-1, 2)$ to $(1, 1)$. Therefore, at time $1/2$, $\mathbb{Z}^{\pi,1/2}$ is not a plain process, since

$$\mathcal{L}(\mathrm{ip}_1(\mathbb{Z}^{\pi,1/2})) = \frac{1}{2}(\delta_{(0,\delta_{3/2})} + \delta_{(0,\delta_{-3/2})});$$

or put differently, even though $Z_1^{\pi,1/2} = 0$ we know at $t = 1$ already precisely where we will end up at $t = 2$, thus, $\mathbb{Z}^{\pi,1/2}$ is not plain.

Theorem 5.12. *For* $p \in (1, \infty)$*, the set* M_p *(of martingales in* FP_p*) forms a closed geodesically convex subset of* FP_p*.*

For ease of exposition, we will only check that the set of martingales is geodesically convex when restricting to geodesics given by interpolation processes. Knowing that all geodesics can be characterized this way modulo an external randomization, this assumption can in fact be made without loss of generality. Alternatively, a proof via our description of geodesic processes as geodesics on $\mathcal{P}_p(\mathcal{Z}_1)$ in Theorem 5.10 is possible too, but less informative, and therefore left to the ambitious reader.

Simplified proof of Theorem 5.12. According to Proposition 5.7, it remains to show that M_p is geodesically convex. To that end, let $\mathbb{X}, \mathbb{Y} \in M_p$ and let $(\mathbb{Z}^u)_{u \in [0,1]}$ be a constant speed geodesic connecting them, which, as already explained, is assumed to be given as the interpolation processes with respect to some $\pi \in \mathrm{Cpl}_{bc}(\mathbb{X}, \mathbb{Y})$.

We need to show that, for given fixed $u \in [0, 1]$, the processes $\mathbb{Z}^u$ is a martingale, too. To that end, let $1 \leq s \leq t \leq N$ and write

$$\mathbb{E}[Z_t^u | \mathcal{F}_s^{\mathbb{Z}}] = \mathbb{E}_\pi[(1-u)X_t + uY_t | \mathcal{F}_{s,s}^{\mathbb{X},\mathbb{Y}}] = (1-u)\mathbb{E}[X_t | \mathcal{F}_s^{\mathbb{X}}] + u\mathbb{E}[Y_t | \mathcal{F}_s^{\mathbb{Y}}] = Z_s^u,$$

where we use bicausality of π in the form of Lemma 2.2, and the martingale property of both $\mathbb{X}$ and $\mathbb{Y}$. Hence, $\mathbb{Z}^u$ is a martingale which completes the proof. ∎

Remark 5.13. We chose $\mathcal{X}_t = \mathbb{R}^d$ in this section to lighten notation. However, Theorem 5.10 remains valid if all of the $\mathcal{X}_t$'s are geodesic space, with the obvious modifications, such as replacing $(1 - u)X_t + uY_t$ in Definition 5.9 by (appropriately measurable selections of) geodesics between X_t and Y_t.

6. Continuity with respect to $\mathcal{A}\mathcal{W}_p$ and applications

We have argued in the introduction that the weak adapted distribution governs 'all' probabilistic aspects of a stochastic process. In this section, we highlight this claim, and further show that several (optimization) problems involving stochastic processes *continuously* depend on the adapted distribution. And moreover, that *quantitative* estimates with respect to $\mathcal{A}\mathcal{W}_p$ are possible. Throughout this section, we assume that $\mathcal{X}_t = \mathbb{R}^d$ for every $1 \leq t \leq N$.

6.1. Optimal stopping

Fix a non-anticipative function $c \colon (\mathbb{R}^d)^N \times \{1, \ldots, N\} \to \mathbb{R}$ and set

$$v_c(\mathbb{X}) := \inf_{\tau \in \mathrm{ST}(\mathbb{X})} \mathbb{E}[c_\tau(X)] \tag{6.1}$$

for $\mathbb{X} \in \mathcal{F}\mathcal{P}$, where $\mathrm{ST}(\mathbb{X})$ denotes the set of all $(\mathcal{F}_t^{\mathbb{X}})_{t=1}^N$-stopping times taking values in $\{1, \ldots, N\}$. In Example 4.5, we have already seen that v_c is well defined on FP_p in the sense that the value of v_c does not depend on the choice of representative.

Proposition 6.1 (Optimal stopping). *Let $\mathbb{X}, \mathbb{Y} \in$ FP be such that $c_{1:N}(X)$ and $c_{1:N}(Y)$ are integrable. Then we have*

$$v_c(\mathbb{X}) - v_c(\mathbb{Y}) \leq \inf_{\pi \in \mathrm{Cpl}_c(\mathbb{X}, \mathbb{Y})} \mathbb{E}_\pi [\max_{1 \leq t \leq N} |c_t(X) - c_t(Y)|].$$

In particular, the following hold:

(i) *If c_t is continuous bounded for every t, then $\mathbb{X} \mapsto v_c(\mathbb{X})$ is continuous on FP with respect to the weak adapted topology.*

(ii) *If c_t is continuous and satisfies that $|c_t(\cdot)| \leq \alpha(1 + |\cdot|^p)$ for every t and some $\alpha > 0$, then $\mathbb{X} \mapsto v_c(\mathbb{X})$ is continuous on FP_p with respect to $\mathcal{AW}_p$.*

(iii) *If c_t is Lipschitz for every t, then $\mathbb{X} \mapsto v_c(\mathbb{X})$ is $\mathcal{AW}_1$-Lipschitz on FP_1.*

In Theorem 7.1, we will construct an example showing that the whole adapted distribution is required to control optimal stopping problems. Note that the non-quantitative version of Proposition 6.1 (i.e., continuity of optimal stopping for continuous c that satisfies an adequate growth condition) already follows from Example 4.5.

Proof of Proposition 6.1. The 'in particular' statement follows by symmetry from the first statement, so we shall only prove the first one. Its proof is similar to [14] and is included here for the convenience of the reader. Let $\varepsilon > 0$ and let $\tau^* \in \mathrm{ST}(\mathbb{Y})$ be a stopping time such that $\mathbb{E}[c_{\tau^*}(Y)] \leq v_c(\mathbb{Y}) + \varepsilon$. Further, let $\pi \in \mathrm{Cpl}_c(\mathbb{X}, \mathbb{Y})$ and, for every $u \in [0, 1]$, define

$$\sigma_u := \min\{t \in \{1, \ldots, N\} : \pi(\tau^* \leq t | \mathcal{F}_{N,0}^{\mathbb{X},\mathbb{Y}}) \geq u\}.$$

By causality, cf. Lemma 2.2, we have $\sigma_u \in \mathrm{ST}(\mathbb{X})$, hence

$$v_c(\mathbb{X}) \leq \inf_{u \in [0,1]} \mathbb{E}_\pi[c_{\sigma_u}(X)] \leq \int_{[0,1]} \mathbb{E}_\pi[c_{\sigma_u}(X)] \, du$$

$$= \sum_{t=0}^{N} \int_{[0,1]} \mathbb{E}_\pi[c_t(X) \mathbf{1}_{\pi(\tau^* \leq t | \mathcal{F}_{N,0}^{\mathbb{X},\mathbb{Y}}) \geq u > \pi(\tau^* \leq t-1 | \mathcal{F}_{N,0}^{\mathbb{X},\mathbb{Y}})}] \, du = \mathbb{E}_\pi[c_{\tau^*}(X)].$$

In conclusion, we obtain

$$v_c(\mathbb{X}) - v_c(\mathbb{Y}) \leq \mathbb{E}_\pi[c_{\tau^*}(X) - c_{\tau^*}(Y)] + \varepsilon \leq \mathbb{E}_\pi[\max_{1 \leq t \leq N} |c_t(X) - c_t(Y)|] + \varepsilon,$$

which, as $\varepsilon > 0$ and $\pi \in \mathrm{Cpl}_c(\mathbb{X}, \mathbb{Y})$ were arbitrary, yields the assertion. ∎

6.2. American options and robust pricing

Working in the setup of, e.g., [29], $X_t \in \mathbb{R}_+$ stands for the (discounted) price of a financial asset at a time $t \in \{1, \ldots, N\}$ and $x_1 = X_1 \in \mathbb{R}_+$ denotes the current price. One assumes that there exists a family of *European* derivatives, described by a (continuous and linearly bounded) family of functions $\varphi_i : (\mathbb{R}_+)^N \to \mathbb{R}$, $i \in I$, which are liquidly traded in the market, meaning that the respective prices p_i, $i \in I$, are specified from externally given

data. The set of all *calibrated models* consists of all martingales with mean x_1 which correctly reproduce the prices given by the market, i.e., in mathematical terms

$$\mathrm{M}_I := \{\mathbb{X} \in \mathrm{M}_1 : X_1 = x_1, \, \mathbb{E}[\varphi_i(X)] = p_i, \, i \in I\}.$$

A common assumption is

$$\{\varphi_i : i \in I\} \text{ contains all call options written on } X_N \text{ and } \mathrm{M}_I \neq \emptyset. \quad \text{(CC)}$$

Going back to a famous observation of Breeden–Litzenberger [37], this implies the following basic fundamental fact.

Proposition 6.2. *Under assumption* (CC)*, the set* $\{\mathcal{L}(\mathbb{X}) : \mathbb{X} \in \mathrm{M}_I\}$ *is* $\mathcal{W}_1$*-compact.*

A direct consequence is that for a further (continuous, linearly bounded) European derivative $\Phi : (\mathbb{R}_+)^N \to \mathbb{R}$, $i \in I$, the set of possible arbitrage free prices consists of the interval $[\inf_{\mathbb{X} \in \mathrm{M}_I} \mathbb{E}[\Phi(X)], \sup_{\mathbb{X} \in \mathrm{M}_I} \mathbb{E}[\Phi(X)]]$, where the endpoints are attained for 'extremal models'. Proposition 6.2 as well as its various extensions play a crucial role for the duality theory as well as the characterization of extremal models in robust finance, see [27, 33, 40, 54] among many others.

A particular limitation of Proposition 6.2 is that it allows us only to consider derivatives with a European payoff structure, but neglects derivatives with an American exercise structure, where the buyer may choose when to exercise and the buyers price in a model $\mathbb{X}$ equals

$$\sup\{\mathbb{E}[\Phi(X)] : \tau \in \mathrm{ST}(\mathbb{X})\}. \quad (6.2)$$

Apart from a few important exceptions (see [8, 26, 59]) the robust finance literature is focused on the case of European derivatives. The simple reason is that going beyond the standard European case requires an adequate topology on processes with a *non-trivial* filtration, which was hitherto unavailable. As derivatives with American exercise structure are *more* common than European derivatives it is highly desirable to extend the existing theory to this case. As a consequence of our results we obtain.

Proposition 6.3. *Assume that* $\{\varphi_i : i \in I\}$ *is a family of derivatives with linearly bounded continuous payoffs and European or American exercise structures. Under assumption* (CC)*, the set* $\{\mathbb{X} : \mathbb{X} \in \mathrm{M}_I\}$ *is* $\mathcal{AW}_1$*-compact. Moreover, for a European or American derivative with continuous linearly bounded payoff* $\Phi : (\mathbb{R}_+)^N \to \mathbb{R}$*, the lower/upper pricing bounds for the derivative* Φ *are attained.*

Proof. This is a direct consequence of Prokhorov's theorem in our setting (Theorem 5.1), Proposition 6.2, and the continuity of optimal stopping (Proposition 6.1). ∎

6.3. Utility maximization

Let $U : \mathbb{R} \to \mathbb{R}$ be an increasing concave (utility) function, and denote by U' the left-continuous version of the derivative. Denote by $\mathcal{H}(\mathbb{X})$ the set of all $(\mathcal{F}_t^{\mathbb{X}})_{t=1}^N$-predictable

processes H that are bounded by 1, and by $(H \cdot X)_t := \sum_{s=1}^{t-1} H_{s+1}(X_{s+1} - X_s)$ the discrete-time stochastic integral of H with respect to X. For $C : (\mathbb{R}^d)^N \to \mathbb{R}$ and $\mathbb{X} \in \mathrm{FP}_p$, denote by $u(\mathbb{X})$ the value of the utility maximization problem with random endowment C, that is,

$$u(\mathbb{X}) := \sup_{H \in \mathcal{H}(\mathbb{X})} \mathbb{E}[U(C(X) + (H \cdot X)_N)].$$

In [13, Theorem 1.8], it is shown that $u(\mathbb{X})$ depends continuously on $\mathbb{X}$ (with respect to $\mathcal{A}\mathcal{W}_p$) when restricting to plain processes (i.e., to $\mathbb{X}$ whose filtration is generated only by their paths). In the context of utility maximization, however, it is of central importance to also understand the effect that changes of the information/filtration of $\mathbb{X}$ have to $u(\mathbb{X})$, and the following result extends [13, Theorem 1.8] to the general setting.

Theorem 6.4. *Let $C : \mathbb{R}^N \to \mathbb{R}$ be Lipschitz continuous, and assume that there exists α such that $U'(\cdot) \le \alpha(1 + |\cdot|^{p-1})$. Then, for every $R > 0$ there is a constant K (depending only on R, α and the Lipschitz constant of C) such that*

$$|u(\mathbb{X}) - u(\mathbb{Y})| \le K \cdot \mathcal{A}\mathcal{W}_p(\mathbb{X}, \mathbb{Y})$$

for every $\mathbb{X}, \mathbb{Y} \in \mathrm{FP}_p$ with $\mathcal{A}\mathcal{W}_p(\mathbb{X}, 0), \mathcal{A}\mathcal{W}_p(\mathbb{Y}, 0) \le R$.

Proof. For simplicity, we assume that $C = 0$ and focus on $p = 1$ (i.e., U is Lipschitz); the modifications needed for the general case are minimal and follow, e.g., as detailed in [13]. Let $H^* \in \mathcal{H}(\mathbb{X})$ be (almost) optimal for $u(\mathbb{X})$, and let $\pi \in \mathrm{Cpl}_{\mathrm{bc}}(\mathbb{X}, \mathbb{Y})$ be (almost) optimal for $\mathcal{A}\mathcal{W}_1(\mathbb{X}, \mathbb{Y})$. Define $G_t := \mathbb{E}_\pi[H_t^* | \mathcal{F}_{0,N}^{\mathbb{X},\mathbb{Y}}]$ for every t. Clearly, G is bounded by 1, and by bicausality of π, G_t is $\mathcal{F}_t^{\mathbb{Y}}$-measurable (see Lemma 2.2); hence $G \in \mathcal{H}(\mathbb{Y})$. It follows that

$$u(\mathbb{Y}) \ge \mathbb{E}[U((G \cdot Y)_N)] = \mathbb{E}_\pi[U(\mathbb{E}_\pi[(H^* \cdot Y)_N \mid |\mathcal{F}_{0,N}^{\mathbb{X},\mathbb{Y}}])].$$

By concavity of U and Jensen's inequality, followed by Lipschitz continuity of U,

$$\begin{aligned}
u(\mathbb{Y}) &\ge \mathbb{E}_\pi[U((H^* \cdot Y)_N)] \\
&= \mathbb{E}_\pi[U((H^* \cdot X)_N + (H^* \cdot (Y - X))_N)] \\
&\ge \mathbb{E}_\pi[U((H^* \cdot X)_N)] - L\mathbb{E}_\pi[|(H^* \cdot (Y - X))_N|],
\end{aligned}$$

where L is the Lipschitz constant of U. It remains to note that $|(H^* \cdot (Y - X))_N| \le 2\sum_{t=1}^{N} |X_t - Y_t|$, hence $u(\mathbb{Y}) \ge u(\mathbb{X}) - 2L\mathcal{A}\mathcal{W}_1(\mathbb{X}, \mathbb{Y})$. Reversing the roles of $\mathbb{X}$ and $\mathbb{Y}$ proves the claim. ∎

Remark 6.5. As explained in [13, Section 3.3], in the context of optimization problems involving stochastic integrals and (semi-)martingales, it is more natural to define $\mathcal{A}\mathcal{W}_p$ with a different cost function instead of $\mathbb{E}[d^p(X, Y)]$, namely with $\mathbb{E}[|\langle M^{\mathbb{X}} - M^{\mathbb{Y}}\rangle|^{p/2} + |A^{\mathbb{X}} - A^{\mathbb{Y}}|_{\mathrm{var}}^p]$, where $\mathbb{X} = M^{\mathbb{X}} + A^{\mathbb{X}}$ denotes the Doob decomposition, $\langle\cdot\rangle$ the quadratic variation, and $|\cdot|_{\mathrm{var}}$ the first variation norm. Clearly, this modification of $\mathcal{A}\mathcal{W}_p$ is also possible (and reasonable) in the present setting.

6.4. Stochastic control

Optimal stopping and utility maximization are basic stochastic control problems involving processes, and we have shown in Proposition 6.1 and Theorem 6.4 that their values are continuous with respect to the adapted Wasserstein distance. As it happens, this is the general principle for stochastic control problems.

For example, if $J: (\mathbb{R}^d)^N \times \mathbb{R}^{N-1} \times (\mathbb{R}^d)^N \to \mathbb{R}$ is convex in the second and third argument, a similar reasoning as used for the proof of Theorem 6.4 shows that under suitable continuity and growth assumptions on J,

$$\mathbb{X} \mapsto \inf_{H \in \mathcal{H}(\mathbb{X})} \mathbb{E}[J(X, ((H \cdot X)_t)_{t=2}^N, H)],$$

is (locally Lipschitz) continuous with respect to $\mathcal{A}\mathcal{W}_p$.

We refer to [13] for a more elaborate analysis of such problems in a mathematical finance context. In particular, using the results of the present paper, the results of [13] for the stability of superhedging, risk-based hedging and utility indifference pricing (which are formulated only for processes endowed with their raw filtration) extend to processes with arbitrary filtrations.

6.5. Conditional McKean–Vlasov control

Denote by $\mathcal{A}(\mathbb{X})$ the set of all $(\mathcal{F}_t^{\mathbb{X}})_{t=1}^N$-adapted processes that are bounded by 1, and let $\mathbb{B}$ be a discrete-time Brownian motion.[7] For $\alpha \in \mathcal{A}(\mathbb{B})$, consider the controlled process $\mathbb{X}^{\mathbb{B},\alpha}$, defined recursively via $X_1^{\mathbb{B},\alpha} = 0$ and

$$X_{t+1}^{\mathbb{B},\alpha} := G_{t+1}(X_t^{\mathbb{B},\alpha}, \alpha_t, \mathcal{L}(X_t^{\mathbb{B},\alpha}), B_{t+1} - B_t),$$

where $G_{t+1}: \mathbb{R}^d \times \mathbb{R}^d \times \mathcal{P}_p(\mathbb{R}^d) \times \mathbb{R}^d \to [0, 1]$ is some fixed continuous function prescribing the dynamics of $X^{\mathbb{B},\alpha}$. Further, let $J: (\mathbb{R}^d)^T \times (\mathbb{R}^d)^T \times \mathcal{P}_p((\mathbb{R}^d)^T) \to [0, 1]$ be continuous and consider the McKean–Vlasov control problem in its weak formulation:

$$\inf_{\mathbb{B} \text{ is Brownian motion and } \alpha \in \mathcal{A}(\mathbb{B})} \mathbb{E}[J(X^{\mathbb{B},\alpha}, \alpha, \mathcal{L}(X^{\mathbb{B},\alpha}|\mathcal{F}_t^{\mathbb{B}})_{t=1}^N)]; \tag{6.3}$$

we refer to [84] for more background on problems of type (6.3). Based on Prokhorov's theorem for filtered processes, it is straightforward to show that a solution to (6.3) exists.

Proposition 6.6. *The infimum over all Brownian motions $\mathbb{B}$ and controls $\alpha \in \mathcal{A}(\mathbb{B})$ in* (6.3) *is attained.*

Sketch of proof. Let $(\mathbb{B}^k, \alpha^k)_k$ be a minimizing sequence for (6.3) (in particular, $\alpha^k \in \mathcal{A}(\mathbb{B}^k)$), and set $\mathbb{Y}^k := (\Omega^{\mathbb{B}^k}, \mathcal{F}^{\mathbb{B}^k}, \mathbb{P}^{\mathbb{B}^k}, (\mathcal{F}_t^{\mathbb{B}^k})_{t=1}^N, (B^k, \alpha^k))$. By Theorem 5.1, the set $\{\mathbb{Y}^k : k \in \mathbb{N}\}$ is relatively compact, hence there exists

$$\mathbb{Y} = (\Omega, \mathcal{F}, \mathbb{P}, (\mathcal{F}_t)_{t=1}^N, (B, \alpha)) \in \mathrm{FP}_p$$

---

[7]That is, $B_1 = 0$, and for every t the increments $(B_{t+1} - B_t)$ are standard normal and independent of $\mathcal{F}_t^{\mathbb{B}}$.

such that (potentially after passing to a subsequence) $\mathcal{A}\mathcal{W}_p(\mathbb{Y}^k, \mathbb{Y}) \to 0$. It is straightforward to verify that

$$\mathbb{B} = (\Omega, \mathcal{F}, \mathbb{P}, (\mathcal{F}_t)_{t=1}^N, B)$$

is a Brownian motion and that $\alpha \in \mathcal{A}(\mathbb{B})$. Further, since each G_t is continuous,

$$(X^{\mathbb{B}^k, \alpha^k}, \alpha^k, (\mathcal{L}(X^{\mathbb{B}^k, \alpha^k} | \mathcal{F}_t^{\mathbb{B}^k}))_{t=1}^N) \to (X^{\mathbb{B}, \alpha}, \alpha, (\mathcal{L}(X^{\mathbb{B}, \alpha} | \mathcal{F}_t))_{t=1}^N) \quad \text{in distribution,}$$

as $k \to \infty$. The claim now readily follows from continuity of J. ∎

6.6. Weak optimal transport

Motivated by applications to functional inequalities, Gozlan et al. [53] introduced the *weak optimal transport problem*, which extends classical transport to 'non-linear' costs $c: \mathbb{R}^d \times \mathcal{P}(\mathbb{R}^d) \to [0, \infty)$. Given marginal distributions μ_1, μ_2 on $\mathbb{R}^d$, the task is to solve

$$\min \mathbb{E}[c(X_1, \mathcal{L}(X_2|X_1))] \text{ over couplings } (X_1, X_2)$$
$$\text{such that } \mathcal{L}(X_1) = \mu_1, \ \mathcal{L}(X_2) = \mu_2. \tag{WOT}$$

Weak transport preserves enough structure from the classical case to allow for a useful theory while being sufficiently general to capture many problems that lie outside the scope of transport theory, see [20] for an overview. Cornerstone results in weak optimal transport (existence of optimizers, duality, geometric characterization of optimal couplings, stability in the data) were shown in the generality of classical transport only in [3, 17, 31], relying on two-period results of the current article and the two-period version of Theorem 1.7. In analogy to classical transport, the key idea is to relax (WOT) and consider

$$\inf\{\mathbb{E}[c(X_1, \mathcal{L}(X_2|\mathcal{F}_1))] : \mathbb{X} \in \mathcal{FP} \text{ satisfies } \mathcal{L}(X_1) = \mu_1, \mathcal{L}(X_2) = \mu_2\}. \tag{6.4}$$

By Theorem 1.7, this is an optimization over a compact set which of course admits minimizers under the usual assumption of lower semicontinuity. Indeed, (6.4) can be viewed as a classical linear optimization problem over probabilities on $\mathbb{Z}_1$.

In several weak transport problems, it is natural to consider not just two marginal constraints, but arbitrarily many (relaxed martingale transport [54], robust pricing of VIX-futures [55], model-independence in fixed income markets following [3], multi-marginal Skorokhod embedding [28, 41]). In view of this, we propose the N-marginal weak transport problem

$$\inf\{\mathbb{E}[c(\mathrm{ip}(\mathbb{X}))] : \mathbb{X} \in \mathcal{FP} \text{ satisfies } \mathcal{L}(X_1) = \mu_1, \ldots, \mathcal{L}(X_N) = \mu_N\}.$$

As above, this is an optimization problem over a compact set, which corresponds to a linear optimization problem of probabilities on $\mathbb{Z}_1$, susceptible to classical convex analysis.

6.7. $\mathcal{A}\mathcal{W}_p$-barycenters

The $\mathcal{A}\mathcal{W}_p$-barycenter of two stochastic processes $\mathbb{Z}^0, \mathbb{Z}^1 \in \mathrm{FP}_p$ is the minimizer of

$$\inf_{\mathbb{X} \in \mathrm{FP}_p} \frac{1}{2}(\mathcal{A}\mathcal{W}_p^p(\mathbb{Z}^0, \mathbb{X}) + \mathcal{A}\mathcal{W}_p^p(\mathbb{Z}^1, \mathbb{X})).$$

Clearly, this minimization problem is attained by the connecting constant speed geodesic at time $1/2$ (which exists thanks to Theorem 5.10). More generally, one can ask whether a distribution on the stochastic processes $\gamma \in \mathcal{P}_p(\mathrm{FP}_p)$ has an $\mathcal{A}\mathcal{W}_p$-barycenter $\mathbb{X}^* \in \mathrm{FP}_p$, that is, a minimizer of

$$\inf_{\mathbb{X} \in \mathrm{FP}_p} \int \mathcal{A}\mathcal{W}_p^p(\mathbb{Z}, \mathbb{X})\gamma(d\mathbb{Z}). \tag{6.5}$$

Theorem 6.7. *Let* $\gamma \in \mathcal{P}_p(\mathrm{FP}_p)$. *Then there exists* $\mathbb{X}^* \in \mathrm{FP}_p$ *minimizing* (6.5).

Proof. Corollary 3.12 implies that $(\mathbb{X}, \mathbb{Z}) \mapsto \mathcal{A}\mathcal{W}_p(\mathbb{X}, \mathbb{Z})$ is lower semicontinuity with respect to the weak adapted topology. Hence, it follows that

$$F(\mathbb{X}) := \int \mathcal{A}\mathcal{W}_p^p(\mathbb{Z}, \mathbb{X})\gamma(d\mathbb{Z})$$

is lower semicontinuity on FP_p with respect to the weak adapted topology. As $F(\mathbb{X}) < \infty$ for $\mathbb{X} \in \mathrm{FP}$ if and only if $\mathbb{X} \in \mathrm{FP}_p$, it suffices to show relative compactness of the sub-level set

$$\{\mathbb{X} \in \mathrm{FP} : F(\mathbb{X}) \leq \inf_{\mathbb{Y} \in \mathrm{FP}} F(\mathbb{Y}) + \varepsilon\}, \tag{6.6}$$

where $\varepsilon > 0$, in the weak adapted topology. Note that the moments of $\mathbb{X} \in \mathrm{FP}_p$ can be controlled as follows:

$$\sum_{t=1}^{N} \mathbb{E}[|X_t|^p] = \mathcal{A}\mathcal{W}_p^p(\mathbb{X}, 0) \leq 2^{p-1} \int \mathcal{A}\mathcal{W}_p^p(\mathbb{Z}, \mathbb{X}) + \mathcal{A}\mathcal{W}_p^p(0, \mathbb{Z})\gamma(dz)$$

$$= 2^{p-1}(F(\mathbb{X}) + F(0)). \tag{6.7}$$

By Theorem 5.1, relative compactness of the set (6.6) is equivalent to tightness of the laws. Tightness of (6.6) follows from standard arguments since the p-moments are uniformly bounded by (6.7) and $\{x \in \mathbb{R}^d : |x| \leq K\}$ is compact for $K > 0$. ∎

6.8. The Doob decomposition

Our final example deals with continuity of the Doob decomposition. Recall that the Doob decomposition $\mathbb{D}^{\mathbb{X}}$ of a filtered process $\mathbb{X}$ is given by

$$\mathbb{D}^{\mathbb{X}} := (\Omega^{\mathbb{X}}, \mathcal{F}^{\mathbb{X}}, (\mathcal{F}_t^{\mathbb{X}})_{t=1}^{N}, P^{\mathbb{X}}, (M_t, A_t)_{t=1}^{N}),$$

where $M + A = X$ is the unique decomposition of X such that $A_1 = 0$, $M = M^{\mathbb{X}}$ is a martingale and $A = A^{\mathbb{X}}$ is $(\mathcal{F}_t^{\mathbb{X}})_{t=1}^{N}$-predictable. (Of course, X is a sub-martingale if and only if the process A is increasing.)

Proposition 6.8. *The following chain of inequalities holds:*[8]

$$2^{(1-p)/p} \mathcal{A} \mathcal{W}_p(\mathbb{X}, \mathbb{Y}) \leq \mathcal{A} \mathcal{W}_p(\mathbb{D}^{\mathbb{X}}, \mathbb{D}^{\mathbb{Y}}) \leq c \cdot \mathcal{A} \mathcal{W}_p(\mathbb{X}, \mathbb{Y}) \tag{6.8}$$

for all $\mathbb{X}, \mathbb{Y} \in \mathrm{FP}_p$, *where* $c = c(p, N)$ *depends on* p *and* N.

Recall that the predictable process $A^{\mathbb{X}}$ of the Doob decomposition of $\mathbb{X} \in \mathrm{FP}_p$ is given by

$$A_t^{\mathbb{X}} := \sum_{s=1}^{t-1} \mathbb{E}[X_s - X_{s-1} | \mathcal{F}_s^{\mathbb{X}}]$$

for $2 \leq t \leq N$ and $M^{\mathbb{X}} := X - A^{\mathbb{X}}$. Thus, A_t can be viewed as an adapted function of rank 1, that is, $A_t \in \mathrm{AF}[1]$, and similarly $M_t \in \mathrm{AF}[1]$. Building upon this observation, it is not hard to deduce $\mathcal{A} \mathcal{W}_p$-continuity of $\mathbb{X} \mapsto \mathbb{D}^{\mathbb{X}}$.

Proof of Proposition 6.8. Fix $\mathbb{X}, \mathbb{Y} \in \mathrm{FP}_p$ and note that, as the filtration of a filtered process and its Doob decomposition coincide by definition, we have that

$$\mathrm{Cpl}_{\mathrm{bc}}(\mathbb{X}, \mathbb{Y}) = \mathrm{Cpl}_{\mathrm{bc}}(\mathbb{D}^{\mathbb{X}}, \mathbb{D}^{\mathbb{Y}}).$$

The first inequality in (6.8) is immediate from Jensen's inequality. The triangle inequality together with Jensen's inequality shows

$$|(M^{\mathbb{X}} - M^{\mathbb{Y}}, A^{\mathbb{X}} - A^{\mathbb{Y}})|_p^p \leq 2^{p-1} |X - Y|^p + (2^{p-1} + 1)|A^{\mathbb{X}} - A^{\mathbb{Y}}|^p. \tag{6.9}$$

By definition of $A^{\mathbb{X}}$ and $A^{\mathbb{Y}}$, we have for $2 \leq t \leq N$ that

$$A_t^{\mathbb{X}} - A_t^{\mathbb{Y}} = \sum_{s=2}^{t} \mathbb{E}_\pi[X_s - X_{s-1} - (Y_s - Y_{s-1}) | \mathcal{F}_{s,s}^{\mathbb{X}, \mathbb{Y}}].$$

Again, Jensen's inequality implies

$$\mathbb{E}_\pi[|A^{\mathbb{X}} - A^{\mathbb{Y}}|^p]^{1/p} \leq \left(\sum_{t=2}^{N} t^{p-1} \sum_{s=2}^{t} \mathbb{E}_\pi[|X_s - X_{s-1} - (Y_s - Y_{s-1})|^p] \right)^{1/p}$$

$$\leq (N 2^p N^{p-1} \mathbb{E}_\pi[|X - Y|^p])^{1/p} \tag{6.10}$$

for every $\pi \in \mathrm{Cpl}_{\mathrm{bc}}(\mathbb{X}, \mathbb{Y})$. In conclusion, suitably combining (6.9) and (6.10) leads to the second inequality in (6.8). ∎

---

[8]The processes (M, A) takes values in $\mathbb{R}^d \times \mathbb{R}^d$ which we endow with the norm $|(x, y)|_p^p := |x|^p + |y|^p$.

6.9. Numerical and statistical aspects

In view of applications, the numerical computation of the adapted Wasserstein distance is of crucial importance and has been recently studied in [24, 44, 87]. In fact, these papers consider more general adapted transport problems between stochastic processes $\mathbb{X}$ and $\mathbb{Y}$ of the form

$$\inf_{\pi \in \mathrm{Cpl}_{\mathrm{bc}}(\mathbb{X}, \mathbb{Y})} \mathbb{E}_\pi[c(X, Y)], \tag{6.11}$$

where c is a continuous function. In order to briefly explain the methodology, suppose that the processes $\mathbb{X}$ and $\mathbb{Y}$ are both discrete, i.e., plain and only take finitely many values. (We remark that for general processes $\mathbb{X}$ and $\mathbb{Y}$ that are not necessarily discrete, one may replace them first by discrete approximations, e.g., as in Proposition 5.5). In this case, solving (6.11) is equivalent to solving a linear program (see, e.g., [80, Section 8] and [44, Section 3.4] for more details). The recursive structure of bicausal couplings allows us to solve the linear program via a dynamic programming principle (DPP) involving one-step classical optimal transport problems (see [81, Chapter 2.10.3]). In the current setting, the value functions $V_t \colon \mathcal{Z}_{1:t} \times \mathcal{Z}_{1:t} \to \mathbb{R}$ are recursively given by

$$V_N(\hat{z}_{1:N}, \check{z}_{1:N}) := c(\hat{z}_{1:N}^-, \check{z}_{1:N}^-),$$

$$V_t(\hat{z}_{1:t}, \check{z}_{1:t}) := \inf_{\pi \in \mathrm{Cpl}(\hat{z}_t^+, \check{z}_t^+)} \mathbb{E}_\pi[V_{t+1}[\hat{z}_{1:t}, \cdot, \check{z}_{1:t}^-, \cdot]]$$

for $t = N - 1, \dots, 0$ with the convention that

$$\hat{z}_0^+ := \mathcal{L}(\mathrm{ip}_1(\mathbb{X})) \quad \text{and} \quad \check{z}_0^+ := \mathcal{L}(\mathrm{ip}_1(\mathbb{Y})).$$

Then V_0 is precisely the optimal value (6.11) and the values of V_t can be computed efficiently, e.g., via Sinkhorn's algorithm (see [87]). Based on the DPP formulation, a fitted value iteration method is proposed in [24]. There, the value function of the DPP is iteratively empirically estimated and then approximated by neural networks that minimize an empirical loss. Another recent approach is built on establishing a bicausal version of the Sinkhorn algorithm (see [44]). For comprehensive comparisons of these methods, we refer to [44] and [24]. Furthermore, in [98], a minimax reformulation is proposed where the infimum is taken over all couplings and the supremum is taken over an adequate class of functions penalizing the causality constraint. Both families of functions are then parametrized by neural networks and trained by an adversarial algorithm. In a continuous-time framework, [35] establishes a Hamilton–Jacobi–Bellman equation for the value function.

The estimation of processes with respect to $\mathcal{A}\mathcal{W}_p$ from statistical data was studied in [11] under the assumption of bounded values and later extended in [4] (see also [52, 83]). Roughly put, while the classical empirical measure does not converge to its population counterpart in the weak adapted topology (nor do, e.g., the values of the corresponding optimal stopping problems) one can construct a *modified empirical measure* (based on clustering ideas) which does converge in the weak adapted topology. In fact, the rate of convergence of that estimator (with respect to $\mathcal{A}\mathcal{W}_p$) is similar to the (optimal) rate of convergence of the classical empirical measure (with respect to $\mathcal{W}_p$).

7. An important example

Fix $N \geq 2$ and recall that, by Theorem 4.11, the relation $\sim_{N-1}$ is equal to the relation $\sim_\infty$, which again coincides with the relation induced by $\mathcal{AW}_p$.

A natural question left open in Sections 4 and 5 is what rank of the adapted distribution is necessary to govern probabilistic properties of stochastic processes. For instance, we have seen in Example 4.5 that the adapted distribution of rank 1 suffices for preserving the notion of being a martingale. However, we will see that rank 1 equivalence is not sufficient in general. Rather Theorem 4.11 is sharp: $\sim_{N-2}$ does not imply $\sim_{N-1}$. In fact, we show that there exist filtered processes that are $\sim_{N-2}$ equivalent but lead to different values for an optimal stopping problem.

Theorem 7.1. *Let $N \geq 2$. There exist $\mathbb{X}, \mathbb{Y} \in \mathrm{FP}_p$ with $\mathbb{X} \sim_{N-2} \mathbb{Y}$ and a bounded continuous non-anticipative function $c\colon \mathcal{X} \times \{1, \ldots, N\} \to \mathbb{R}$ such that $v_c(\mathbb{X}) \neq v_c(\mathbb{Y})$ (compare (6.1) for the definition of v_c).*

In particular, using Proposition 6.1 and Theorem 4.11, this implies that $\mathbb{X} \sim_{N-1} \mathbb{Y}$. More generally, we will show in the following that as long as $k \leq N - 1$, the relation $\sim_k$ strictly refines the relation $\sim_{k-1}$.

Proposition 7.2. *Let $N \geq 2$ and let $1 \leq k \leq N - 1$. Then there are $\mathbb{X}, \mathbb{Y} \in \mathrm{FP}_p$ with representatives defined on the same filtered probability space such that*

(i) *$f(\mathbb{X}) = f(\mathbb{Y})$ for every $f \in \mathrm{AF}[k - 1]$,*

(ii) *$\mathbb{E}[f(\mathbb{X})] \neq \mathbb{E}[f(\mathbb{Y})]$ for some $f \in \mathrm{AF}[k]$.*

The processes $\mathbb{X}$ and $\mathbb{Y}$ will be constructed on a common probability space consisting of 2^{N-1} atoms, satisfy $X = Y$ and $X_t = Y_t = 0$ for $t \leq N - 1$, and only their respective filtrations will differ.

7.1. Proof of Proposition 7.2

This section is devoted to the proof of Proposition 7.2 which will then be used to establish Theorem 7.1. In fact, we shall first concentrate on Proposition 7.2 with $k = N - 1$, and later conclude for general k via a simple argument.

The construction of $\mathbb{X}$ and $\mathbb{Y}$ is recursive, and we shall start with $N = 2$. Consider the probability space consisting of two elements, say $\{0, 1\}$, with the uniform measure, and let U be the identity map, that is, $\mathbb{P}(U = 1) = \mathbb{P}(U = 0) = 1/2$. Now define the filtered processes $\mathbb{X}$ and $\mathbb{Y}$ via

$$X := Y := (0, U), \quad \mathcal{F}_1^{\mathbb{X}} := \{\emptyset, \Omega\} \quad \text{and} \quad \mathcal{F}_1^{\mathbb{Y}} := \mathcal{F}_2^{\mathbb{Y}} := \mathcal{F}_2^{\mathbb{X}} := \sigma(U).$$

Then the following holds.

Lemma 7.3. *The processes $\mathbb{X}$ and $\mathbb{Y}$ satisfy the claim in Proposition 7.2 for $N = 2$.*

Proof. Adapted functions of rank 0 depend only on the process itself and not on the filtration, hence $f(\mathbb{X}) = f(\mathbb{Y})$ for all $f \in \mathrm{AF}[0]$. That is, part (i) of Proposition 7.2 is true. On the other hand, one computes

$$\mathbb{E}[g(X_2)|\mathcal{F}_1^{\mathbb{X}}] = \mathbb{E}[g(U)], \quad \mathbb{E}[g(X_2)|\mathcal{F}_1^{\mathbb{Y}}] = g(U)$$

for every bounded measurable function $g\colon \mathbb{R} \to \mathbb{R}$. Now define $f \in \mathrm{AF}$ by

$$f := \varphi((g|1)), \quad \text{where } \varphi = t \mapsto t^2 \wedge 1 \text{ and } g := x \mapsto (0 \vee x) \wedge 1,$$

so that f has rank 1. Then

$$f(\mathbb{X}) = \mathbb{E}[X_2|\mathcal{F}_1^{\mathbb{X}}]^2 = \frac{1}{4}, \quad f(\mathbb{Y}) = \mathbb{E}[X_2|\mathcal{F}_1^{\mathbb{Y}}]^2 = X_2.$$

As $\mathbb{E}[f(\mathbb{Y})] = 1/2$, this proves the second part of Proposition 7.2. ∎

Now assume that, in an $(N-1)$-time step framework, we have constructed filtered processes $\mathbb{X}$ and $\mathbb{Y}$ which satisfy Proposition 7.2 (where $N \geq 3$). We shall construct new filtered processes $\mathbb{X}^{\mathrm{n}}$ and $\mathbb{Y}^{\mathrm{n}}$ in an N-time step framework for which the statement of the lemma remains true. At this particular instance, the superscript 'n' stands for 'new'.

At this point, we enrich the filtered probability space of $\mathbb{X}$ and $\mathbb{Y}$ by an independent random variable V (i.e., V is independent of $\mathcal{F}_{N-1}^{\mathbb{X}} \vee \mathcal{F}_{N-1}^{\mathbb{Y}}$) in the following manner: When $(\Omega, \mathcal{F}, \mathbb{P})$ denotes the probability space of $\mathbb{X}$ and $\mathbb{Y}$, we consider the new probability space

$$\left(\Omega^{\mathrm{n}} := \{0, 1\} \times \Omega, \ \mathcal{F}^{\mathrm{n}} := \mathcal{B}(\{0, 1\}) \otimes \mathcal{F}, \ \mathbb{P}^{\mathrm{n}} := \frac{1}{2}(\delta_0 \otimes \mathbb{P} + \delta_1 \otimes \mathbb{P})\right). \quad (7.1)$$

We write $V\colon \Omega^{\mathrm{n}} \to \{0, 1\}$ for the projection onto the first coordinate. In other words, we can think of V as a coin-flip which is independent of everything that was constructed so far. Further, we view $(\mathcal{G}_t^{\mathbb{X}})_{t=1}^N := (\mathcal{F}_{t-1}^{\mathbb{X}})_{t=1}^{N-1}$ as a filtration on (7.1); similarly for $(\mathcal{G}_t^{\mathbb{Y}})_{t=1}^N$ (recall here that $\mathcal{F}_0^{\mathbb{X}} = \mathcal{F}_0^{\mathbb{Y}}$ are the trivial σ-algebras by convention). Then, by slight abuse of notation, we can view $\mathbb{X}$ and $\mathbb{Y}$ as processes defined on Ω^{n}, namely by identifying $\mathbb{X}$ with $(0, X)$ and filtration $(\mathcal{G}_t^{\mathbb{X}})_{t=1}^N$ and similarly for $\mathbb{Y}$. By independence of V, this clearly does not affect any properties of the processes. Define the processes $\mathbb{X}^{\mathrm{n}}$ and $\mathbb{Y}^{\mathrm{n}}$ on the probability space (7.1) via

$$X^{\mathrm{n}} := Y^{\mathrm{n}} := (0, X),$$

$$\mathcal{F}_1^{\mathbb{X}^{\mathrm{n}}} := \{\emptyset, \Omega^{\mathrm{n}}\} \quad \text{and} \quad \mathcal{F}_1^{\mathbb{Y}^{\mathrm{n}}} := \sigma(V),$$

$$\mathcal{F}_t^{\mathbb{X}^{\mathrm{n}}} := \mathcal{F}_t^{\mathbb{Y}^{\mathrm{n}}} := \{(\{0\} \times A) \cup (\{1\} \times B) : A \in \mathcal{G}_t^{\mathbb{X}}, B \in \mathcal{G}_t^{\mathbb{Y}}\}$$

$$= \{(A \cap \{V = 0\}) \cup (B \cap \{V = 1\}) : A \in \mathcal{G}_t^{\mathbb{X}}, B \in \mathcal{G}_t^{\mathbb{Y}}\}$$

for $2 \leq t \leq N$. One can check that $(\mathcal{F}_t^{\mathbb{X}^{\mathrm{n}}})_{t=1}^N$ and $(\mathcal{F}_t^{\mathbb{Y}^{\mathrm{n}}})_{t=1}^N$ are indeed filtrations and that, for every $\mathbb{P}^{\mathrm{n}}$-integrable random variable Z which is independent of V, we have that

$$\mathbb{E}[Z|\mathcal{F}_t^{\mathbb{X}^{\mathrm{n}}}] = \mathbb{E}[Z|\mathcal{G}_t^{\mathbb{X}}]\mathbb{1}_{\{V=0\}} + \mathbb{E}[Z|\mathcal{G}_t^{\mathbb{Y}}]\mathbb{1}_{\{V=1\}} = \mathbb{E}[Z|\mathcal{F}_t^{\mathbb{Y}^{\mathrm{n}}}] \quad (7.2)$$

for every $1 \leq t \leq N$. We spare the elementary proof hereof.

Lemma 7.4. *For $0 \le k \le N - 3$ and $f \in \mathrm{AF}[k]$, we have*

$$f(\mathbb{X}^{\mathrm{n}}) = f(\mathbb{X}) = f(\mathbb{Y}) = f(\mathbb{Y}^{\mathrm{n}}). \tag{7.3}$$

In particular, $\mathbb{X}^{\mathrm{n}} \sim_{N-3} \mathbb{Y}^{\mathrm{n}}$.

Proof. The proof is by induction over k: for $k = 0$, the statement is trivially true (recalling that $X^{\mathrm{n}} = (0, X) = Y^{\mathrm{n}}$).

Now assume that it is true for $k - 1$ (with $k \le N - 3$), and let $f \in \mathrm{AF}[k]$. First assume that $f = (g|t)$ is formed by (AF3) only. By assumption, we have

$$g(\mathbb{X}^{\mathrm{n}}) = g(\mathbb{X}) = g(\mathbb{Y}) = g(\mathbb{Y}^{\mathrm{n}}). \tag{7.4}$$

Now use independence of V and $\mathcal{F}_N^{\mathbb{X}} \vee \mathcal{F}_N^{\mathbb{Y}}$ and (7.4) to compute

$$f(\mathbb{X}^{\mathrm{n}}) = \mathbb{E}[g(\mathbb{X})|\mathcal{F}_t^{\mathbb{X}^{\mathrm{n}}}] = \mathbb{E}[g(\mathbb{X})|\mathcal{G}_t^{\mathbb{X}}]1_{\{V=0\}} + \mathbb{E}[g(\mathbb{Y})|\mathcal{G}_t^{\mathbb{Y}}]1_{\{V=1\}}$$
$$= f(\mathbb{X})1_{\{V=0\}} + f(\mathbb{Y})1_{\{V=1\}} = f(\mathbb{X}) = f(\mathbb{Y}),$$

where the last three equalities are due to the fact that $\mathbb{X}$ and $\mathbb{Y}$ satisfy Proposition 7.2, that is, $f(\mathbb{X}) = f(\mathbb{Y})$. The same computation is valid for $\mathbb{Y}^{\mathrm{n}}$, whence we have (7.3) for this particular choice of f. Finally, by Lemma 4.2, this extends to all $f \in \mathrm{AF}[k]$, not necessarily those formed solely by (AF3). ∎

Lemma 7.5. *For $f \in \mathrm{AF}[N - 2]$, we have*

$$f(\mathbb{X}^{\mathrm{n}}) = f(\mathbb{X})1_{\{V=0\}} + f(\mathbb{Y})1_{\{V=1\}} = f(\mathbb{Y}^{\mathrm{n}}). \tag{7.5}$$

Proof. First, consider $g \in \mathrm{AF}[N - 3]$, $1 \le t \le N$, $f = (g|t)$. We compute $f(\mathbb{X}^{\mathrm{n}})$ by applying (7.2) and Lemma 7.4

$$f(\mathbb{X}^{\mathrm{n}}) = \mathbb{E}[g(\mathbb{X}^{\mathrm{n}})|\mathcal{F}_t^{\mathbb{X}^{\mathrm{n}}}] = \mathbb{E}[g(\mathbb{X})|\mathcal{F}_t^{\mathbb{X}^{\mathrm{n}}}] = \mathbb{E}[g(\mathbb{X})|\mathcal{G}_t^{\mathbb{X}}]1_{\{V=0\}} + \mathbb{E}[g(\mathbb{Y})|\mathcal{G}_t^{\mathbb{Y}}]1_{\{V=1\}}$$
$$= f(\mathbb{X})1_{\{V=0\}} + f(\mathbb{Y})1_{\{V=1\}}.$$

We conclude by noticing that the same computation holds true when replacing $\mathbb{X}^{\mathrm{n}}$ by $\mathbb{Y}^{\mathrm{n}}$. ∎

Lemma 7.6. *The processes $\mathbb{X}^{\mathrm{n}}$ and $\mathbb{Y}^{\mathrm{n}}$ satisfy the claim in Proposition 7.2 with $k = N - 1$.*

Proof. Part (i) of Proposition 7.2 is a consequence of Lemma 7.5. We proceed to prove part (ii) of Proposition 7.2, that is, we construct $f \in \mathrm{AF}[N - 1]$ such that

$$\mathbb{E}[f(\mathbb{X}^{\mathrm{n}})] \ne \mathbb{E}[f(\mathbb{Y}^{\mathrm{n}})].$$

Since $\mathbb{X} \sim_{N-2} \mathbb{Y}$, there is a function $g \in \mathrm{AF}[N - 2]$ such that $\mathbb{E}[g(\mathbb{X})] \ne \mathbb{E}[g(\mathbb{Y})]$. Set $f := \min\{(g|1)^2, 1\} \in \mathrm{AF}[N - 1]$. Using identity (7.5) from Lemma 7.5 and (7.3), we compute

$$f(\mathbb{X}^{\mathrm{n}}) = \mathbb{E}[g(\mathbb{X})1_{\{V=0\}} + g(\mathbb{Y})1_{\{V=1\}}]^2,$$
$$f(\mathbb{Y}^{\mathrm{n}}) = \mathbb{E}[g(\mathbb{X})1_{\{V=0\}} + g(\mathbb{Y})1_{\{V=1\}}|V]^2 = (\mathbb{E}[g(\mathbb{X})]1_{\{V=0\}} + \mathbb{E}[g(\mathbb{Y})]1_{\{V=1\}})^2.$$

Recalling that $\mathbb{E}[g(\mathbb{X})] \ne \mathbb{E}[g(\mathbb{Y})]$, this implies $\mathbb{E}[f(\mathbb{X}^{\mathrm{n}})] < \mathbb{E}[f(\mathbb{Y}^{\mathrm{n}})]$. ∎

At this point, we have completed the proof of Proposition 7.2 under the additional assumption that $k = N - 1$. For general $k \leq N - 1$, construct $\mathbb{X}$ and $\mathbb{Y}$ for the $(k - 1)$-time step framework as above. Recursively repeating the argument detailed below (7.1), we can append $(N - k)$ trivial time steps to the processes $\mathbb{X}$ and $\mathbb{Y}$, and thereby obtain $(N - 1)$-time step processes with the desired properties. This proves Proposition 7.2 for arbitrary $k \leq N - 1$.

7.2. Proof of Theorem 7.1

As already announced, we use the processes $\mathbb{X}$ and $\mathbb{Y}$ (and the notation specific to these processes) constructed for the proof of Proposition 7.2

The proof will be inductive, starting with $N = 2$. Consider the cost function $c \colon \mathbb{R}^2 \times \{1, 2\} \to \mathbb{R}$,

$$c_1(x_1, x_2) := \frac{1}{2} \quad \text{and} \quad c_2(x_1, x_2) := (0 \vee x_2) \wedge 1.$$

The dynamic programming principle for the optimal stopping problem (also called 'Snell envelope theorem') tells us that the optimal stopping values are

$$v_c(\mathbb{X}) = \mathbb{E}[c_1(X) \wedge \mathbb{E}[c_2(X)|\mathcal{F}_1^{\mathbb{X}}]] = \frac{1}{2} \quad \text{and} \quad v_c(\mathbb{Y}) = \frac{1}{4}.$$

This proves the claim for $N = 2$.

For the case of general N, recall that $\mathbb{X}^{\mathrm{n}}$ and $\mathbb{Y}^{\mathrm{n}}$ are the processes obtained through the $(N - 1)$-step processes $\mathbb{X}$ and $\mathbb{Y}$ and an independent coin-flip V.

By the previous step, we assume that in an $(N - 1)$-step framework we have constructed a cost function c such that $v_c(\mathbb{X}) > v_c(\mathbb{Y})$. Defining the new cost function c^{n} in the N-step framework by

$$c_1^{\mathrm{n}}(x_{1:N}) := \frac{v_c(\mathbb{X}) + v_c(\mathbb{Y})}{2},$$
$$c_t^{\mathrm{n}}(x_{1:N}) := c_{t-1}(x_{2:N}) \quad \text{for } 2 \leq t \leq N,$$

we claim that $v_{c^{\mathrm{n}}}(\mathbb{X}^{\mathrm{n}}) > v_{c^{\mathrm{n}}}(\mathbb{Y}^{\mathrm{n}})$.

To that end, we once more rely on the dynamic programming principle to obtain

$$v_{c^{\mathrm{n}}}(\mathbb{Y}^{\mathrm{n}}) = \mathbb{E}[c_1^{\mathrm{n}} \wedge \inf_{\tau \in \mathrm{ST}(\mathbb{Y}^{\mathrm{n}}), \, \tau \geq 2} \mathbb{E}[c_\tau^{\mathrm{n}}(Y^{\mathrm{n}})|\mathcal{F}_1^{\mathbb{Y}^{\mathrm{n}}}]]. \tag{7.6}$$

Similarly, additionally using that $\mathcal{F}_1^{\mathbb{X}^{\mathrm{n}}}$ is trivial and c_1^{n} is deterministic, we get

$$v_{c^{\mathrm{n}}}(\mathbb{X}^{\mathrm{n}}) = c_1^{\mathrm{n}} \wedge \inf_{\tau \in \mathrm{ST}(\mathbb{X}^{\mathrm{n}}), \, \tau \geq 2} \mathbb{E}[c_\tau^{\mathrm{n}}(X^{\mathrm{n}})]. \tag{7.7}$$

In order to relate the conditional stopping problems after time 2 appearing in (7.6) and (7.7), recall that we view $\mathbb{X}$ and $\mathbb{Y}$ as $(N - 1)$-time step processes by appending a trivial initial time step; see the explanation right after (7.1). The same convention is applied here to c as well.

In a first step, we show the following decomposition of stopping times:

$$\mathrm{ST}(\mathbb{Y}^{\mathrm{n}}) = \{\alpha 1_{\{V=0\}} + \beta 1_{\{V=1\}} : \alpha \in \mathrm{ST}(\mathbb{X}),\ \beta \in \mathrm{ST}(\mathbb{Y})\}. \tag{7.8}$$

The right-hand side is clearly contained in the left-hand side. For the reverse inclusion, pick $\tau \in \mathrm{ST}(\mathbb{Y}^{\mathrm{n}})$. By definition of $\mathcal{F}_t^{\mathbb{Y}^{\mathrm{n}}}$, there are sets $A_t \in \mathcal{G}_t^{\mathbb{X}}$ and $B_t \in \mathcal{G}_t^{\mathbb{Y}}$ such that

$$\{\tau = t\} = A_t \cap \{V = 0\} \cup B_t \cap \{V = 1\}$$

for every $1 \le t \le N$. One can then check that

$$\alpha := \min\{1 \le t \le N : 1_{A_t} = 1\}, \quad \beta := \min\{1 \le t \le N : 1_{B_t} = 1\}$$

define stopping times in $\mathrm{ST}(\mathbb{X})$ and $\mathrm{ST}(\mathbb{Y})$, respectively, and that $\tau = \alpha 1_{\{V=0\}} + \beta 1_{\{V=1\}}$. This shows (7.8).

We are now ready to finish the proof. Recalling that $\mathcal{F}_1^{\mathbb{X}^{\mathrm{n}}}$ is the trivial σ-algebra and that $\mathcal{F}_t^{\mathbb{X}^{\mathrm{n}}} = \mathcal{F}_t^{\mathbb{Y}^{\mathrm{n}}}$ for $2 \le t \le N$, independence of V and $\mathcal{G}_N^{\mathbb{X}} \vee \mathcal{G}_N^{\mathbb{Y}}$ and the decomposition of stopping times (7.8) shows that

$$\inf_{\tau \in \mathrm{ST}(\mathbb{X}^{\mathrm{n}}),\, \tau \ge 2} \mathbb{E}[c_\tau^{\mathrm{n}}(X^{\mathrm{n}})] = \inf_{(\alpha,\beta) \in \mathrm{ST}(\mathbb{X}) \times \mathrm{ST}(\mathbb{Y})} \mathbb{E}[c_\alpha(X) 1_{\{V=0\}} + c_\beta(Y) 1_{\{V=1\}}]$$

$$= \frac{v_c(\mathbb{X}) + v_c(\mathbb{Y})}{2} = c_1^{\mathrm{n}}.$$

In a similar manner,

$$\inf_{\tau \in \mathrm{ST}(\mathbb{Y}^{\mathrm{n}}),\, \tau \ge 2} \mathbb{E}[c_\tau^{\mathrm{n}}(X^{\mathrm{n}}) | \mathcal{F}_1^{\mathbb{Y}^{\mathrm{n}}}] = \inf_{\alpha \in \mathrm{ST}(\mathbb{X})} \mathbb{E}[c_\alpha(X)] 1_{\{V=0\}} + \inf_{\beta \in \mathrm{ST}(\mathbb{Y})} \mathbb{E}[c_\beta(Y)] 1_{\{V=1\}}$$

$$= v_c(\mathbb{X}) 1_{\{V=0\}} + v_c(\mathbb{Y}) 1_{\{V=1\}}.$$

As $v_c(\mathbb{Y}) < c_1^{\mathrm{n}} < v_c(\mathbb{X})$, plugging the above equalities in (7.6) and (7.7) readily shows $v_{c^{\mathrm{n}}}(\mathbb{Y}^{\mathrm{n}}) < v_{c^{\mathrm{n}}}(\mathbb{X}^{\mathrm{n}})$ and thus completes the proof.

Appendix A. The adapted block approximation

We have seen in Lemma 3.9 that bicausal couplings π on arbitrary filtered probability spaces induce bicausal couplings in the canonical setting. In order to establish that the association of $\mathbb{X} \in \mathcal{F}\mathcal{P}_p$ with its canonical representative $\bar{\mathbb{X}} \in \mathrm{CFP}_p$ is an isometry with respect to $\mathcal{A}\mathcal{W}_p$ (see Section 3.2), we have to find for $\bar{\pi} \in \mathrm{Cpl}_{\mathrm{bc}}(\bar{\mathbb{X}}, \bar{\mathbb{Y}})$ a similar coupling $\pi \in \mathrm{Cpl}_{\mathrm{bc}}(\mathbb{X}, \mathbb{Y})$. For this reason, we introduce in this section what we call the *adapted block approximation* in Proposition A.3 (for the canonical filtered setting). The main result of this section is Theorem A.4, which allows us then to pull back block approximations of elements in $\mathrm{Cpl}_{\mathrm{bc}}(\bar{\mathbb{X}}, \bar{\mathbb{Y}})$ to $\mathrm{Cpl}_{\mathrm{bc}}(\mathbb{X}, \mathbb{Y})$.

We start with a characterization of bicausality in terms of kernels for canonical filtered processes.

Lemma A.1. *Let* $\mathbb{X}, \mathbb{Y} \in \mathrm{CFP}_p$, *let* $\pi \in \mathcal{P}_p(\mathbb{Z} \times \mathbb{Z})$, *and set* $\pi_1 := (\mathrm{pj}_{\mathbb{Z}_1 \times \mathbb{Z}_1})_* \pi$. *Then* $\pi \in \mathrm{Cpl}_{\mathrm{bc}}(\mathbb{X}, \mathbb{Y})$ *if and only if*

$$\pi_1 \in \mathrm{Cpl}(\mathcal{L}(\mathrm{ip}_1(\mathbb{X})), \mathcal{L}(\mathrm{ip}_1(\mathbb{Y})))$$

and, for $1 \le t \le N - 1$, *there are kernels*

$$k_t \colon \mathbb{Z}_{1:t} \times \mathbb{Z}_{1:t} \to \mathcal{P}_p(\mathbb{Z}_{t+1} \times \mathbb{Z}_{t+1}) \quad \text{with } k_t^{z_{1:t}, \hat{z}_{1:t}} \in \mathrm{Cpl}(z_t^+, \hat{z}_t^+)$$

such that

$$\pi = \pi_1 \otimes k_1 \otimes \cdots \otimes k_{N-1}. \tag{A.1}$$

Proof. It is clear that any coupling can be represented by a family of measurable kernels $(k_t)_{t=1}^{N-1}$ as in (A.1) with

$$k_t = \mathcal{L}_\pi(z_{t+1}, \hat{z}_{t+1} | \mathcal{F}_{t,t}^{\mathbb{Z}, \mathbb{Z}}).$$

The only thing we need to show is that $k_t^{z_{1:t}, \hat{z}_{1:t}} \in \mathrm{Cpl}(z_t^+, \hat{z}_t^+)$ for every $1 \le t \le N - 1$ if and only if π is bicausal. But this follows from Lemmas 2.2 and 3.5. Indeed, by these lemmas, we have that π-almost surely

$$\mathcal{L}_\pi(z_{t+1} | \mathcal{F}_{t,t}^{\mathbb{Z}, \mathbb{Z}}) = \mathcal{L}_{\mathcal{L}(\mathrm{ip}(\mathbb{X}))}(z_{t+1} | \mathcal{F}_t^{\mathbb{Z}}) = z_t^+,$$
$$\mathcal{L}_\pi(\hat{z}_{t+1} | \mathcal{F}_{t,t}^{\mathbb{Z}, \mathbb{Z}}) = \mathcal{L}_{\mathcal{L}(\mathrm{ip}(\mathbb{Y}))}(\hat{z}_{t+1} | \mathcal{F}_t^{\mathbb{Z}}) = \hat{z}_t^+$$

for all $1 \le t \le N - 1$, if and only if $\pi \in \mathrm{Cpl}_{\mathrm{bc}}(\mathbb{X}, \mathbb{Y})$. This completes the proof. ∎

Lemma A.2. *Let* $1 \le t \le N - 1$, *let* $\pi \in \mathcal{P}_p(\mathbb{Z}_{1:t} \times \mathbb{Z}_{1:t})$, *and let*

$$k \colon \mathbb{Z}_{1:t} \times \mathbb{Z}_{1:t} \to \mathcal{P}_p(\mathbb{Z}_{t+1} \times \mathbb{Z}_{t+1}) \quad \text{with } k^{z_{1:t}, \hat{z}_{1:t}} \in \mathrm{Cpl}(z_t^+, \hat{z}_t^+).$$

Further, let $(\pi^n)_{n \in \mathbb{N}}$ *in* $\mathcal{P}_p(\mathbb{Z}_{1:t} \times \mathbb{Z}_{1:t})$ *such that* $\mathcal{W}_p(\pi, \pi^n) \to 0$. *Then there are kernels*

$$k^n \colon \mathbb{Z}_{1:t} \times \mathbb{Z}_{1:t} \to \mathcal{P}_p(\mathbb{Z}_{t+1} \times \mathbb{Z}_{t+1}) \quad \text{with } k^{n, z_{1:t}, \hat{z}_{1:t}} \in \mathrm{Cpl}(z_t^+, \hat{z}_t^+)$$

such that

$$\mathcal{W}_p(\pi \otimes k, \pi^n \otimes k^n) \to 0. \tag{A.2}$$

Proof. In this proof, we deal with the spaces $(\mathbb{Z}_{1:t} \times \mathbb{Z}_{1:t}) \times (\mathbb{Z}_{1:t} \times \mathbb{Z}_{1:t})$ and $(\mathbb{Z}_{t+1} \times \mathbb{Z}_{t+1}) \times (\mathbb{Z}_{t+1} \times \mathbb{Z}_{t+1})$. For the sake of a clearer presentation, we denote the first product space by $\mathcal{A} \times \mathcal{B}$ and the second one by $\mathcal{C} \times \mathcal{E}$. We write $a = (\hat{a}, \check{a}) = (\hat{a}_{1:t}, \check{a}_{1:t})$ for elements in $\mathcal{A}$ (the space in the first bracket) and $\hat{a}_t = (\hat{a}_t^-, \hat{a}_t^+)$ as well as $\check{a}_t = (\check{a}_t^-, \check{a}_t^+)$. Similar conventions apply to elements in the spaces $\mathcal{B}, \mathcal{C}, \mathcal{E}$.

For every n, let Π^n be an optimal coupling for $\mathcal{W}_p(\pi, \pi^n)$ and denote by K^n its disintegration with respect to π^n, that is,

$$\Pi^n(\mathrm{d}a, \mathrm{d}b) = \pi^n(\mathrm{d}b) K^{n,b}(\mathrm{d}a).$$

In a first step, we define an auxiliary kernel $\tilde{k}^n \colon \mathcal{B} \to \mathcal{P}_p(\mathcal{C})$, which is π^n-almost surely well defined, for $b \in \mathcal{B}$ by

$$\tilde{k}^{n,b}(\cdot) := \int k^a(\cdot) K^{n,b}(\mathrm{d}a) \in \mathcal{P}_p(\mathcal{C}).$$

In general, $\tilde{k}^{n,b}$ needs not to be an element of $\mathrm{Cpl}(\hat{b}_t^+, \check{b}_t^+)$. To amend this, we pick measurable kernels $K^1 \colon \mathcal{B} \to \mathcal{P}_p(\mathcal{C})$ and $K^2 \colon \mathcal{B} \to \mathcal{P}_p(\mathcal{E})$, where

$$K^{1,b} \text{ is an optimal coupling for } \mathcal{W}_p((\mathrm{pj}_1)_* \tilde{k}^{n,b}, \hat{b}_t^+),$$

$$K^{2,b} \text{ is an optimal coupling for } \mathcal{W}_p(\check{b}_t^+, (\mathrm{pj}_2)_* \tilde{k}^{n,b}).$$

A disintegration of $K^{1,b}$ with respect to the first coordinate is denoted by $(K^{1,b,\hat{c}})_{\hat{c} \in \mathcal{Z}_{t+1}}$, and similarly a disintegration of $K^{2,b}$ with respect to the second coordinate is called $(K^{2,b,\check{e}})_{\check{e} \in \mathcal{Z}_{t+1}}$. Proceeding from this, we can define the kernel $k^n \colon \mathcal{B} \to \mathcal{P}_p(\mathcal{C})$ with $k^{n,b} \in \mathrm{Cpl}(\hat{b}_t^+, \check{b}_t^+)$

$$k^{n,b}(\cdot) := \int K^{1,b,\hat{c}}(\cdot) \otimes K^{2,b,\check{e}}(\cdot) \, \mathrm{d}\tilde{k}^{n,b}(\mathrm{d}\hat{c}, \mathrm{d}\check{e}).$$

From here, we get

$$\mathcal{W}_p^p(k^{n,b}, \tilde{k}^{n,b}) \le \iint d^p(\hat{c}, \check{c}) + d^p(\hat{e}, \check{e}) \, \mathrm{d}K^{1,b,\hat{c}} \otimes K^{2,b,\check{e}} \tilde{k}^{n,b}(\mathrm{d}\hat{c}, \mathrm{d}\check{e})$$

$$= \int d^p(\hat{c}, \check{c}) \, \mathrm{d}K^{1,b} + \int d^p(\hat{e}, \check{e}) \, \mathrm{d}K^{2,b}$$

$$= \mathcal{W}_p^p(\mathrm{pj}_1 \tilde{k}^{n,b}, \hat{b}_t^+) + \mathcal{W}_p^p(\check{b}_t^+, \mathrm{pj}_2 \tilde{k}^{n,b}). \tag{A.3}$$

Due to Jensen's inequality, we obtain

$$\left(\int \mathcal{W}_p^p(\mathrm{pj}_1 \tilde{k}^{n,b}, \hat{b}_t^+) + \mathcal{W}_p^p(\mathrm{pj}_2 \tilde{k}^{n,b}, \check{b}_t^+) \pi^n(\mathrm{d}b) \right)^{1/p}$$

$$\le \left(\int \mathcal{W}_p^p(\hat{a}_t^+, \hat{b}_t^+) + \mathcal{W}_p^p(\check{a}_t^+, \check{b}_t^+) K^n(\mathrm{d}a, \mathrm{d}b) \right)^{1/p} \le \mathcal{W}_p(\pi^n, \pi). \tag{A.4}$$

It remains to verify that the sequence $(k^n)_{n \in \mathbb{N}}$ satisfies (A.2). We have by Minkowski's inequality

$$\mathcal{W}_p(\pi^n \otimes k^n, \pi \otimes k) \le \mathcal{W}_p(\pi^n, \pi) + \left(\int \mathcal{W}_p^p(k^a, k^{n,b}) K^n(\mathrm{d}a, \mathrm{d}b) \right)^{1/p}.$$

Using (A.3), (A.4) and Minkowski's inequality, we bound the last term of the right-hand side by

$$\left(\int \mathcal{W}_p^p(k^a, \tilde{k}^{n,b}) K^n(\mathrm{d}a, \mathrm{d}b) \right)^{1/p} + \left(\int \mathcal{W}_p^p(\tilde{k}^{n,b}, k^{n,b}) \pi^n(\mathrm{d}b) \right)^{1/p}$$

$$\le \mathcal{W}_p(\pi^n, \pi) + \left(\iint \mathcal{W}_p(k^a, k^{a'}) K^{n,b}(\mathrm{d}a') K^n(\mathrm{d}a, \mathrm{d}b) \right)^{1/p}.$$

The last term vanishes by [45, Lemma 2.7], since $\int K^{n,b}(\cdot) \otimes K^{n,b}(\cdot)\,\pi^n(\mathrm{d}b) \in \mathrm{Cpl}(\pi, \pi)$ and

$$\left(\iint d^p(a, a') K^{n,b}(\mathrm{d}a') K^n(\mathrm{d}a, \mathrm{d}b) \right)^{1/p}$$

$$\leq \left(\iint (\mathrm{d}(a, b) + d(b, a'))^p K^{n,b}(\mathrm{d}a') K^n(\mathrm{d}a, \mathrm{d}b) \right)^{1/p} \leq 2\mathcal{W}_p(\pi^n, \pi). \qquad \blacksquare$$

Proposition A.3 (Adapted block approximation). *Let* $\mathbb{X}, \mathbb{Y} \in \mathrm{CFP}_p$, *let* $\pi \in \mathrm{Cpl}_{\mathrm{bc}}(\mathbb{X}, \mathbb{Y})$, *and let* $\varepsilon > 0$. *Then, for every* $1 \leq t \leq N$, *there are countable partitions* P_t *of* $\mathcal{Z}_t$ *and families of measurable functions* $(w_t^{A,B})_{(A,B) \in P_t \times P_t}$ *mapping from* $\mathcal{Z}_{1:t-1} \times \mathcal{Z}_{1:t-1}$ *to* $[0, 1]$ *with*

$$\sum_{A \in P_t} w_t^{A,B}(z_{1:t-1}, \widehat{z}_{1:t-1}) = \widehat{z}_{t-1}^+(B),$$

$$\sum_{B \in P_t} w_t^{A,B}(z_{1:t-1}, \widehat{z}_{1:t-1}) = z_{t-1}^+(A) \qquad (\mathrm{A.5})$$

such that the coupling $\pi^\varepsilon \in \mathrm{Cpl}_{\mathrm{bc}}(\mathbb{X}, \mathbb{Y})$,

$$\pi^\varepsilon(\mathrm{d}z, \mathrm{d}\widehat{z})$$

$$:= \prod_{s=1}^N \sum_{(A_s, B_s) \in P_s \times P_s} w_s^{A_s, B_s}(z_{1:s-1}, \widehat{z}_{1:s-1}) z_{s-1}^+(\mathrm{d}z_s | A_s) \widehat{z}_{s-1}^+(\mathrm{d}\widehat{z}_s | B_s), \qquad (\mathrm{A.6})$$

satisfies $\mathcal{W}_p(\pi, \pi^\varepsilon) < \varepsilon$, *where* $z_0^+ := \mathcal{L}(\mathrm{ip}_1(\overline{\mathbb{X}}))$ *and* $z_0^+ := \mathcal{L}(\mathrm{ip}_1(\overline{\mathbb{Y}}))$.

Here we used the notation $\nu(\cdot | A) := \nu(\cdot \cap A)/\nu(A)$ if $\nu(A) > 0$ with an arbitrary convention otherwise.

Proof of Proposition A.3. By Lemma A.1, a coupling of form (A.5) is bicausal between $\mathbb{X}$ and $\mathbb{Y}$. For $1 \leq t \leq N$, define

$$\pi_{1:t} := (\mathrm{pj}_{\mathcal{Z}_{1:t} \times \mathcal{Z}_{1:t}})_* \pi.$$

The proof uses an induction over t, and we claim the following: For given t and for all $\varepsilon > 0$ and $1 \leq s \leq t$, there are partitions P_s of $\mathcal{Z}_s$ and mappings w_s as in the statement of the proposition such that $\pi_{1:t}^\varepsilon$ satisfies $\mathcal{W}_p(\pi_{1:t}, \pi_{1:t}^\varepsilon) < \varepsilon$. Here $\pi_{1:t}^\varepsilon$ is the adapted block approximation up to time t, i.e., defined as in (A.6) but with the product taken over $1 \leq s \leq t$ instead of all $1 \leq s \leq N$.

We start with $t = 1$. To that end, fix ε and let P_1 be a countable partition of $\mathcal{Z}_t$ into measurable sets of diameter at most ε. Denote by

$$w_1^{A,B} := \pi_{1:1}(A \times B)$$

for $(A, B) \in P_1 \times P_1$. Since the diameter of A (resp. B) is smaller than ε, we clearly have that $\mathcal{W}_p(\pi_{1:1}, \pi_{1:1}^\varepsilon) < 2\varepsilon$, where $\pi_{1:1}^\varepsilon$ is the adapted block approximation up to time $t = 1$. Further, a straightforward calculation shows that w_1 satisfies (A.5).

Assuming that our induction claim is true for $1 \leq t \leq N$, fix some $\varepsilon > 0$. By Lemma A.2, there is a measurable kernel

$$k_t \colon \mathcal{Z}_{1:t} \times \mathcal{Z}_{1:t} \to \mathcal{P}_p(\mathcal{Z}_{t+1} \times \mathcal{Z}_{t+1}) \quad \text{with } k_t^{z_{1:t}, \hat{z}_{1:t}} \in \mathrm{Cpl}(z_t^+, \hat{z}_t^+)$$

and $\delta > 0$ small enough such that

$$\mathcal{W}_p(\pi_{1:t+1}, \pi_{1:t}^{\delta} \otimes k_t) < \varepsilon.$$

Now let P_{t+1} be a countable partition of $\mathcal{Z}_{t+1}$ into measurable sets with diameter at most ε. For $(A, B) \in P_{t+1} \times P_{t+1}$, define

$$w_{t+1}^{A,B}(z_{1:t}, \hat{z}_{1:t}) := k_t^{z_{1:t}, \hat{z}_{1:t}}(A \times B)$$

and set

$$\tilde{k}_t^{z_{1:t}, \hat{z}_{1:t}} := \sum_{(A,B) \in P_{t+1} \times P_{t+1}} w_{t+1}^{A,B}(z_{1:t}, \hat{z}_{1:t}) z_t^+ \otimes \hat{z}_t^+(\cdot \, | A \times B).$$

As the sets in P_{t+1} have diameter at most ε, it follows that

$$\mathcal{W}_p(\tilde{k}_t^{z_{1:t}, \hat{z}_{1:t}}, k_t^{z_{1:t}, \hat{z}_{1:t}}) < 2\varepsilon$$

for every $z_{1:t}, \hat{z}_{1:t} \in \mathcal{Z}_{1:t}$. Further, recalling that $k_t^{z_{1:t}, \hat{z}_{1:t}} \in \mathrm{Cpl}(z_t^+, \hat{z}_t^+)$, it follows that $w_{t+1}^{A,B}$ satisfies (A.5). Finally, set

$$\pi_{1:t+1}^{\varepsilon} := \pi_{1:t}^{\delta} \otimes \tilde{k}_t.$$

A straightforward calculation shows that $\mathcal{W}_p(\pi_{1:t+1}, \pi_{1:t+1}^{\varepsilon}) < 3\varepsilon$. It remains to note that $\pi_{1:t+1}^{\varepsilon}$ has the form as claimed in our induction statement, which completes the proof. ∎

Theorem A.4. *Let* $\mathbb{X}, \mathbb{Y} \in \mathcal{FP}_p$, *let* $\bar{\mathbb{X}}, \bar{\mathbb{Y}}$ *be their associated canonical processes, and let* $\pi \in \mathrm{Cpl}_{\mathrm{bc}}(\bar{\mathbb{X}}, \bar{\mathbb{Y}})$. *Then, for every* $\varepsilon > 0$, *there is* $\Pi^{\varepsilon} \in \mathrm{Cpl}_{\mathrm{bc}}(\mathbb{X}, \mathbb{Y})$ *such that*

$$\mathcal{W}_p(\pi, (\mathrm{ip}(\mathbb{X}), \mathrm{ip}(\mathbb{Y}))_* \Pi^{\varepsilon}) < \varepsilon.$$

Proof. Let π^{ε} be the adapted block approximation of π given in Proposition A.3. Then we can express π^{ε} as

$$\pi^{\varepsilon}(\mathrm{d}z, \mathrm{d}\hat{z}) = \prod_{t=1}^{N} \sum_{(A_t, B_t) \in P_t \times P_t} \mathbf{1}_{A_t \times B_t}(z_t, \hat{z}_t) \frac{w_t^{A_t, B_t}(z_{1:t-1}, \hat{z}_{1:t-1})}{z_{t-1}^+(A_t) \hat{z}_{t-1}^+(B_t)} (\mu \otimes \nu)(\mathrm{d}z, \mathrm{d}\hat{z}).$$

With this representation of π^{ε} in mind, by slight abuse of notation, write $\mathbf{1}_A = \mathbf{1}_A(\mathrm{ip}_t(\mathbb{X}))$ for $A \in P_t$ and similarly for $\mathbb{Y}$, and define

$$D_t := \sum_{(A,B) \in P_t \times P_t} \mathbf{1}_{A \times B} \frac{w_t^{A,B}(\mathrm{ip}_{1:t-1}(\mathbb{X}), \mathrm{ip}_{1:t-1}(\mathbb{Y}))}{\mathrm{ip}_{t-1}^+(\mathbb{X})(A) \mathrm{ip}_{t-1}^+(\mathbb{Y})(B)}$$

for every $1 \leq t \leq N$. Now let Π^ε be the measure absolutely continuous with respect to $\mathbb{P}^{\mathbb{X}} \otimes \mathbb{P}^{\mathbb{Y}}$ with density

$$\frac{d\Pi^\varepsilon}{d\mathbb{P}^{\mathbb{X}} \otimes \mathbb{P}^{\mathbb{Y}}} := \prod_{t=1}^{N} D_t.$$

In particular, $\pi^\varepsilon = (\mathrm{ip}(\mathbb{X}), \mathrm{ip}(\mathbb{Y}))_* \Pi^\varepsilon$. Before proving the theorem, we interject the following claim.

Claim (Auxiliary claim). *For every $1 \leq t \leq N - 1$ and every U that is $\mathcal{F}_N^{\mathbb{X}}$-measurable and bounded, we have that*

$$\mathbb{E}_{\mathbb{P}^{\mathbb{X}} \otimes \mathbb{P}^{\mathbb{Y}}}[D_{t+1} U | \mathcal{F}_{t,t}^{\mathbb{X},\mathbb{Y}}] = \mathbb{E}_{\mathbb{P}^{\mathbb{X}}}[U | \mathcal{F}_t^{\mathbb{X}}].$$

To see that this claim is true, note that by definition of D_{t+1}, the left-hand side equals

$$\sum_{(A,B) \in P_{t+1} \times P_{t+1}} \frac{w_{t+1}^{A,B}(\mathrm{ip}_{1:t}(\mathbb{X}), \mathrm{ip}_{1:t}(\mathbb{Y}))}{\mathrm{ip}_t^+(\mathbb{X})(A)\mathrm{ip}_t^+(\mathbb{Y})(B)} \mathbb{E}_{\mathbb{P}^{\mathbb{X}}}[U 1_A | \mathcal{F}_t^{\mathbb{X}}] \mathbb{P}(\mathrm{ip}_{t+1}(\mathbb{Y}) \in B | \mathcal{F}_t^{\mathbb{Y}})$$

$$= \sum_{A \in P_{t+1}} \mathbb{E}_{\mathbb{P}^{\mathbb{X}}}[U 1_A | \mathcal{F}_t^{\mathbb{X}}] \sum_{B \in P_{t+1}} \frac{w_{t+1}^{A,B}(\mathrm{ip}_{1:t}(\mathbb{X}), \mathrm{ip}_{1:t}(\mathbb{Y}))}{\mathrm{ip}_t^+(\mathbb{X})(A)}.$$

Property (A.5) of Proposition A.3 implies that the sum over $B \in P_{t+1}$ equals 1. Further, as P_{t+1} is a partition, our claim follows.

We are now ready to prove the theorem. In the first step, note that Π^ε is indeed a coupling of $\mathbb{X}$ and $\mathbb{Y}$, which follows from property (A.5) of w. Further, for symmetry reasons, it suffices to show that Π^ε is causal from $\mathbb{X}$ to $\mathbb{Y}$. By Lemma 2.2, the latter holds true if

$$\mathbb{E}_{\Pi^\varepsilon}[UV] = \mathbb{E}_{\Pi^\varepsilon}[V \mathbb{E}_{\mathbb{P}^{\mathbb{X}}}[U | \mathcal{F}_t^{\mathbb{X}}]]$$

for every $1 \leq t \leq N - 1$ and every U and V that are bounded and $\mathcal{F}_N^{\mathbb{X}}$ and $\mathcal{F}_{t,t}^{\mathbb{X},\mathbb{Y}}$-measurable, respectively. Now fix such t, U and V.

The definition of Π^ε, iteratively applying the tower property, and our auxiliary claim imply that

$$\mathbb{E}_{\Pi^\varepsilon}[UV] = \mathbb{E}_{\mathbb{P}^{\mathbb{X}} \otimes \mathbb{P}^{\mathbb{Y}}}\left[\prod_{s=1}^{N-1} D_s D_N U V\right]$$

$$= \mathbb{E}_{\mathbb{P}^{\mathbb{X}} \otimes \mathbb{P}^{\mathbb{Y}}}\left[\prod_{s=1}^{N-1} D_s V \mathbb{E}_{\mathbb{P}^{\mathbb{X}} \otimes \mathbb{P}^{\mathbb{Y}}}[D_N U | \mathcal{F}_{N-1,N-1}^{\mathbb{X},\mathbb{Y}}]\right]$$

$$= \mathbb{E}_{\mathbb{P}^{\mathbb{X}} \otimes \mathbb{P}^{\mathbb{Y}}}\left[\prod_{s=1}^{N-1} D_s V \mathbb{E}_{\mathbb{P}^{\mathbb{X}}}[U | \mathcal{F}_{N-1}^{\mathbb{X}}]\right] = \cdots$$

$$= \mathbb{E}_{\mathbb{P}^{\mathbb{X}} \otimes \mathbb{P}^{\mathbb{Y}}}\left[\prod_{s=1}^{t} D_s V \mathbb{E}_{\mathbb{P}^{\mathbb{X}}}[U | \mathcal{F}_t^{\mathbb{X}}]\right].$$

Finally, note that the auxiliary claim (applied with $U = 1$) also shows that, for every $1 \leq s \leq N - 1$, we have $\mathbb{E}_{\mathbb{P}^{\mathbb{X}} \otimes \mathbb{P}^{\mathbb{Y}}}[D_{s+1} | \mathcal{F}_{s,s}^{\mathbb{X}, \mathbb{Y}}] = 1$. Hence, another application of the tower property gives

$$\mathbb{E}_{\Pi^{\varepsilon}}[UV] = \mathbb{E}_{\mathbb{P}^{\mathbb{X}} \otimes \mathbb{P}^{\mathbb{Y}}} \left[\prod_{s=1}^{t} D_s V \mathbb{E}_{\mathbb{P}^{\mathbb{X}}}[U | \mathcal{F}_t^{\mathbb{X}}] \right] = \mathbb{E}_{\Pi^{\varepsilon}}[V \mathbb{E}_{\mathbb{P}^{\mathbb{X}}}[U | \mathcal{F}_t^{\mathbb{X}}]].$$

This concludes the proof. ∎

Appendix B. Infinite discrete time

The main focus of this article is on stochastic processes in a finite discrete-time framework. In this section, we consider the class of stochastic processes with discrete but infinite time horizon. In fact, many of our results carry over to this instance, by simple limit arguments.

We consider the class $\mathcal{FP}^{(\infty)}$ of processes of the form

$$\mathbb{X} = (\Omega, \mathcal{F}, \mathbb{P}, (\mathcal{F}_t)_{t=1}^{\infty}, (X_t)_{t=1}^{\infty}),$$

where $X = (X_t)_{t=1}^{\infty}$ takes values[9] in $\mathbb{R}^{\infty}$. To turn $\mathbb{R}^{\infty}$ into a Polish space, we equip it with the product topology and, more specifically, with the distance

$$d(x, y) = d_p(x, y) := \left(\sum_{t=1}^{\infty} \frac{1}{2^t} (|x_t - y_t|^p \wedge 1) \right)^{1/p}.$$

In complete analogy with Definition 2.1 above, we can then consider bicausality as well as $\mathcal{AW}_p^{(\infty)}$ on $\mathcal{P}_p(\mathbb{R}^{\infty})$ and $\mathcal{FP}_p^{(\infty)}$. As above, we write $\mathrm{FP}_p^{(\infty)}$ for the class obtained after identifying $\mathbb{X}, \mathbb{Y}$ with $\mathcal{AW}_p(\mathbb{X}, \mathbb{Y}) = 0$. Note that the metric d is bounded, consequently $\mathcal{FP}_p^{(\infty)} / \mathrm{FP}_p^{(\infty)}$ do not depend on the choice of p and $\mathcal{P}_p(\mathbb{R}^{\infty})$ carries the usual weak topology. In analogy to the results stated in the introduction, we then obtain the following.

Theorem B.1. *The following hold:*

(i)　$\mathcal{AW}_p^{(\infty)}$ *is a metric on* $\mathrm{FP}_p^{(\infty)}$.

(ii)　$(\mathrm{FP}_p^{(\infty)}, \mathcal{AW}_p^{(\infty)})$ *is the completion of* $(\mathcal{P}_p(\mathbb{R}^{\infty}), \mathcal{AW}_p^{(\infty)})$, *where we identify laws in* $\mathcal{P}_p(\mathbb{R}^{\infty})$ *with corresponding plain processes.*

(iii)　*The set*

$$\{\mathbb{X} \in \mathrm{FP}_p^{(\infty)} : \mathbb{X} \text{ is Markov and has a representative on a finite probability space}\}$$

is dense in $\mathrm{FP}_p^{(\infty)}$.

---

[9] As above, the results and arguments of this section are valid in the case of an arbitrary Polish state space.

(iv) *If a sequence $(\mathbb{X}^n)_n \subset \mathrm{FP}_p^{(\infty)}$ of martingales converges to $\mathbb{X} \in \mathrm{FP}_p^{(\infty)}$ with respect to $\mathcal{A}\mathcal{W}_p^{(\infty)}$ and $\{X^n : n \in \mathbb{N}\}$ is uniformly integrable, then $\mathbb{X}$ is a martingale.*

(v) *If a sequence $(\mathbb{X}^n)_n \subset \mathrm{FP}_p^{(\infty)}$ converges to $\mathbb{X} \in \mathrm{FP}_p^{(\infty)}$ with respect to $\mathcal{A}\mathcal{W}_p^{(\infty)}$ and $\{X^n : n \in \mathbb{N}\}$ is uniformly integrable, then the Doob decomposition $\mathbb{D}^{\mathbb{X}^n}$ of $\mathbb{X}^n$ converge to the Doob decomposition $\mathbb{D}^{\mathbb{X}}$ of $\mathbb{X}$ with respect to $\mathcal{A}\mathcal{W}_p$.*

(vi) *If $G_t \colon \mathbb{R}^t \to \mathbb{R}$ is bounded continuous for each $t \in \mathbb{N}$, and $\lim_{t \to \infty} \|G_t\|_\infty = 0$, then*

$$\sup\{\mathbb{E}[G_\tau(X_1, \ldots, X_\tau)] : \tau \text{ is finite stopping time}\}$$

is continuous in $\mathbb{X}$.

(vii) *('Prokhorov') A set $K \subseteq \mathrm{FP}_p^{(\infty)}$ is precompact if and only if the respective set of laws in $\mathcal{P}_p(\mathbb{R}^\infty)$ is precompact.*

To establish Theorem B.1, we need some notations. We write $\mathcal{F}\mathcal{P}_p^{(N)}$ for the set of N-step filtered processes, $\mathcal{A}\mathcal{W}_p^{(N)}$ for the adapted Wasserstein distance with respect to $d^{(N)}$ given by

$$d^{(N)}(x, y) = d_p^{(N)}(x, y) = \left(\sum_{t=1}^N \frac{1}{2^t}(|x_t - y_t|^p \wedge 1) \right)^{1/p},$$

and we write $\mathrm{FP}_p^{(N)}$ for the space obtained after identifying equivalent processes. For $N \in \mathbb{N}$, we consider the function $r_N \colon \bigcup_{M \in \{N, N+1, \ldots, \infty\}} \mathcal{F}\mathcal{P}_p^{(M)} \to \mathcal{F}\mathcal{P}_p^{(N)}$ given by

$$r_N((\Omega, \mathcal{F}, \mathbb{P}, (\mathcal{F}_t)_{t=1}^M, (X_t)_{t=1}^M)) = (\Omega, \mathcal{F}, \mathbb{P}, (\mathcal{F}_t)_{t=1}^N, (X_t)_{t=1}^N).$$

The next two lemmas will allow us to derive Theorem B.1 from the respective results in the finite time horizon case.

Lemma B.2. *For every $\mathbb{X}, \mathbb{Y} \in \mathcal{F}\mathcal{P}_p^{(\infty)}$ and every $N \in \mathbb{N}$, we have*

$$\mathcal{A}\mathcal{W}_p^{(N)}(r_N(\mathbb{X}), r_N(\mathbb{Y})) \leq \mathcal{A}\mathcal{W}_p^{(\infty)}(\mathbb{X}, \mathbb{Y}) \leq \mathcal{A}\mathcal{W}_p^{(N)}(r_N(\mathbb{X}), r_N(\mathbb{Y})) + \frac{1}{2^N}.$$

In particular, $\mathcal{A}\mathcal{W}_p^{(\infty)}(\mathbb{X}, \mathbb{Y}) = \lim_{N \to \infty} \mathcal{A}\mathcal{W}_p^{(N)}(r_N(\mathbb{X}), r_N(\mathbb{Y}))$ for $\mathbb{X}, \mathbb{Y} \in \mathcal{F}\mathcal{P}_p^{(\infty)}$.

Proof. The first inequality is trivial. Similar to Theorem 3.10, the other inequality has a short proof under the additional assumption that $\mathbb{X}, \mathbb{Y}$ are supported by Polish probability spaces. Indeed, let $\pi \in \mathrm{Cpl}_{\mathrm{bc}}(r_N(\mathbb{X}), r_N(\mathbb{Y}))$ and note that

$$\bar{\pi} := (\mathrm{ip}(r_N(\mathbb{X})), \mathrm{ip}(r_N(\mathbb{Y})))_* \pi \in \mathrm{Cpl}_{\mathrm{bc}}(\overline{r_N(\mathbb{X})}, \overline{r_N(\mathbb{Y})}).$$

We then consider

$$\gamma := (\mathrm{id}, \mathrm{ip}(r_N(\mathbb{X})))_* \mathbb{P}^{\mathbb{X}} \quad \text{and} \quad \hat{\gamma} := (\mathrm{id}, \mathrm{ip}(r_N(\mathbb{Y})))_* \mathbb{P}^{\mathbb{Y}}.$$

As in the proof of Theorem 3.10, these couplings admit disintegrations $(\gamma_z)_z$ and $(\widehat{\gamma}_{\widehat{z}})_{\widehat{z}}$ with respect to the second variable (since the considered probability spaces are Polish by assumption) and we may consider $\Pi(d\omega, d\widehat{\omega}) := \int \gamma_z(d\omega)\widehat{\gamma}_{\widehat{z}}(d\widehat{\omega})\overline{\pi}(dz, d\widehat{z})$. Arguing similar to the proof of Theorem 3.10, we then obtain that $\Pi \in \mathrm{Cpl}_{\mathrm{bc}}(\mathbb{X}, \mathbb{Y})$. Since π is arbitrary, it is straightforward to verify that

$$\mathcal{A}\mathcal{W}_p^{(\infty)}(\mathbb{X}, \mathbb{Y}) \leq \mathcal{A}\mathcal{W}_p^{(N)}(r_N(\mathbb{X}), r_N(\mathbb{Y})) + \frac{1}{2^N}.$$

We now drop the assumption that the underlying probability spaces are Polish. Let $\pi \in \mathrm{Cpl}_{\mathrm{bc}}(\overline{r_N(\mathbb{X})}, \overline{r_N(\mathbb{Y})})$. We first claim that for every $\varepsilon > 0$, there exists $\Pi^\varepsilon \in \mathrm{Cpl}_{\mathrm{bc}}(\mathbb{X}, \mathbb{Y})$ such that

$$\mathcal{W}_p(\pi, (\mathrm{ip}(r_N(\mathbb{X})), \mathrm{ip}(r_N(\mathbb{Y})))_* \Pi^\varepsilon) < \varepsilon. \tag{B.1}$$

Indeed, let $\Pi^\varepsilon \in \mathrm{Cpl}_{\mathrm{bc}}(r_N(\mathbb{X}), r_N(\mathbb{Y}))$ be the coupling constructed in (the proof of) Theorem A.4; hence

$$\frac{d\Pi^\varepsilon}{d\mathbb{P}^{\mathbb{X}} \otimes \mathbb{P}^{\mathbb{Y}}} = \prod_{t=1}^{N} D_t$$

for some $(\mathcal{F}_{t,t}^{\mathbb{X},\mathbb{Y}})_{t=1}^N$-adapted process $D = (D_t)_{t=1}^N$. To see that $\Pi^\varepsilon \in \mathrm{Cpl}_{\mathrm{bc}}(\mathbb{X}, \mathbb{Y})$, let V be bounded and $\mathcal{F}^{\mathbb{X}}$-measurable. Recall the 'chain rule' for conditional expectations: if α and β are two probability measures such that $d\alpha/d\beta = Z$, then $\mathbb{E}_\alpha[\cdot | \mathcal{H}] = \mathbb{E}_\beta[Z \cdot | \mathcal{H}]/\mathbb{E}_\beta[Z | \mathcal{H}]$. Hence,

$$\mathbb{E}_{\Pi^\varepsilon}[V | \mathcal{F}_{t,t}^{\mathbb{X},\mathbb{Y}}] = \mathbb{E}_{\mathbb{P}^{\mathbb{X}} \otimes \mathbb{P}^{\mathbb{Y}}}\left[\prod_{s=t+1}^{N} D_s V | \mathcal{F}_{t,t}^{\mathbb{X},\mathbb{Y}} \right]$$

$$= \mathbb{E}_{\mathbb{P}^{\mathbb{X}} \otimes \mathbb{P}^{\mathbb{Y}}}\left[\prod_{s=t+1}^{N} D_s \mathbb{E}_{\mathbb{P}^{\mathbb{X}} \otimes \mathbb{P}^{\mathbb{Y}}}[V | \mathcal{F}_{N,N}^{\mathbb{X},\mathbb{Y}}] | \mathcal{F}_{t,t}^{\mathbb{X},\mathbb{Y}} \right],$$

where the second inequality follows from the tower property. Moreover,

$$\mathbb{E}_{\mathbb{P}^{\mathbb{X}} \otimes \mathbb{P}^{\mathbb{Y}}}[V | \mathcal{F}_{N,N}^{\mathbb{X},\mathbb{Y}}] = \mathbb{E}_{\mathbb{P}^{\mathbb{X}}}[V | \mathcal{F}_N^{\mathbb{X}}]$$

because $\mathbb{P}^{\mathbb{X}} \otimes \mathbb{P}^{\mathbb{Y}} \in \mathrm{Cpl}_{\mathrm{bc}}(\mathbb{X}, \mathbb{Y})$.

Thus, by the chain rule and because $\Pi^\varepsilon \in \mathrm{Cpl}_{\mathrm{bc}}(r_N(\mathbb{X}), r_N(\mathbb{Y}))$,

$$\mathbb{E}_{\mathbb{P}^{\mathbb{X}} \otimes \mathbb{P}^{\mathbb{Y}}}\left[\prod_{s=1}^{N} D_s \mathbb{E}_{\mathbb{P}^{\mathbb{X}}}[V | \mathcal{F}_N^{\mathbb{X}}] | \mathcal{F}_{t,t}^{\mathbb{X},\mathbb{Y}} \right] = \mathbb{E}_{\Pi^\varepsilon}[\mathbb{E}_{\mathbb{P}^{\mathbb{X}}}[V | \mathcal{F}_N^{\mathbb{X}}] | \mathcal{F}_{t,t}^{\mathbb{X},\mathbb{Y}}] = \mathbb{E}_{\mathbb{P}^{\mathbb{X}}}[V | \mathcal{F}_t^{\mathbb{X}}].$$

Hence, Π^ε is causal and thus, by symmetry, bicausal.

It follows that

$$\mathbb{E}_\pi[d_p^{(N)}(X, Y)^p]^{1/p} + \frac{1}{2^N} + \varepsilon \geq \mathbb{E}_{\Pi^\varepsilon}[d_p^{(N)}(X, Y)^p]^{1/p} + \frac{1}{2^N} \geq \mathbb{E}_{\Pi^\varepsilon}[d_p(X, Y)^p]^{1/p},$$

where we used (B.1) for the first inequality. As $\varepsilon > 0$ and $\pi \in \mathrm{Cpl}_{\mathrm{bc}}(r_N(\mathbb{X}), r_N(\mathbb{Y}))$ were fixed but arbitrary, this concludes the proof. ∎

The following lemma is a version of the Kolmogorov extension theorem.

Lemma B.3. *Let $\mathbb{X}^{(N)} \in \mathcal{F}\mathcal{P}_p^{(N)}$, $N \in \mathbb{N}$, be a sequence of filtered processes such that $r_N(\mathbb{X}^{(N+1)}) = \mathbb{X}^{(N)}$, $N \in \mathbb{N}$. Then there exists a stochastic process $\mathbb{X} \in \mathcal{F}\mathcal{P}_p^{(\infty)}$ defined on a Polish probability space such that*

$$\mathcal{A}\mathcal{W}_p^{(N)}(r_N(\mathbb{X}), \mathbb{X}^{(N)}) = 0 \quad \text{for all } N \in \mathbb{N}.$$

Proof. For every $N \in \mathbb{N}$ and $1 \le k \le N$, we denote by $I^{k,N}$ the information processes of $r_k(\mathbb{X}^{(N)})$, that is, $(I_t^{k,N})_{t=1}^k := (\mathrm{ip}_t(r_k(\mathbb{X}^{(N)})))_{t=1}^k$. Hence, by assumption, for every $N \in \mathbb{N}$ and $M \ge N$,

$$\mathcal{L}((I_t^{k,N})_{1 \le t \le k \le N}) = \mathcal{L}((I_t^{k,M})_{1 \le t \le k \le N}).$$

By the Kolmogorov extension theorem, there exists a Polish probability space $(\Omega, \mathcal{F}, \mathbb{P})$ supporting random variables $\{J_t^k : k \in \mathbb{N}, 1 \le t \le k\}$ such that for all $N \ge k$,

$$\mathcal{L}((I_t^{k,N})_{1 \le t \le k \le N}) = \mathcal{L}((J_t^k)_{1 \le t \le k \le N}). \tag{B.2}$$

Note that (for $t < k$) J_t^k has two components, $J_t^k = (J_t^{k,-}, J_t^{k,+})$, and define the filtered process $\mathbb{X}$ as the tuplet

$$\mathbb{X} := (\Omega, \mathcal{F}, \mathbb{P}, (\mathcal{F}_t)_{t=1}^\infty, X := (J_t^{t,-})_{t=1}^\infty),$$

where $\mathcal{F}_t := \sigma(J_s^k : k \in \mathbb{N}, 1 \le s \le t \wedge k)$. Note that $J_t^{k,-} = J_t^{t,-}$ $\mathbb{P}$-a.s. (for $t \le k$).

We claim that for every $1 \le t \le N$, we have $\mathbb{P}$-a.s.

$$\mathrm{ip}_t(r_N(\mathbb{X})) = J_t^N,$$

which would yield the assertion of the lemma.

We proceed to show this claim by backward induction. Indeed, when $t = N$, we have that $J_N^N = X_N = \mathrm{ip}_N(r_N(\mathbb{X}))$. Next, assume the claim to be true for $t + 1$. Then we have $\mathbb{P}$-a.s.

$$\mathrm{ip}_t^-(r_N(\mathbb{X})) = J_t^{N,-} \quad \text{and} \quad \mathrm{ip}_t^+(r_N(\mathbb{X})) = \mathcal{L}_{\mathbb{P}}(J_{t+1}^N | \mathcal{F}_t).$$

So, it remains to show that $\mathcal{L}_{\mathbb{P}}(J_{t+1}^N | \mathcal{F}_t)$ coincides with $J_t^{N,+}$. To this end, define the σ-algebras

$$\widetilde{\mathcal{F}}_t^N := \sigma(I_s^{k,N} : 1 \le k \le N, 1 \le s \le k \wedge t).$$

By Lemma 3.3, for $t + 1 \le k \le N$, we have $\mathbb{P}^{\mathbb{X}^{(N)}}$-a.s.

$$I_t^{k,N,+} = \mathcal{L}_{\mathbb{P}^{\mathbb{X}^{(N)}}}(I_{t+1}^{k,N} | \mathcal{F}_t^{\mathbb{X}^{(N)}}) = \mathcal{L}_{\mathbb{P}^{\mathbb{X}^{(N)}}}(I_{t+1}^{k,N} | \widetilde{\mathcal{F}}_t^N). \tag{B.3}$$

Setting $\mathcal{F}_t^N = \sigma(J_s^k : 1 \le k \le N, 1 \le s \le k \wedge t)$, we obtain from (B.2) that

$$\mathcal{L}(J_t^{k,+}, \mathcal{L}_{\mathbb{P}}(J_{t+1}^k | \mathcal{F}_t^N)) = \mathcal{L}(I_t^{k,N,+}, \mathcal{L}_{\mathbb{P}^{\mathbb{X}^{(N)}}}(I_{t+1}^{k,N} | \widetilde{\mathcal{F}}_t^N)). \tag{B.4}$$

Let $M \geq N$ and set $k = N$. Then (B.3) and (B.4) yield $\mathbb{P}$-a.s.

$$\mathcal{L}_{\mathbb{P}}(J_{t+1}^N | \mathcal{F}_t^N) = J_t^{N,+} = \mathcal{L}(J_{t+1}^N | \mathcal{F}_t^M). \tag{B.5}$$

As $M \geq N$ is arbitrary, $(\mathcal{F}_t^M)_{M=1}^\infty$ is increasing and $\lim_{M \to \infty} \mathcal{F}_t^M$ generates $\mathcal{F}_t$, we find that (B.5) also holds true when replacing $\mathcal{F}_t^M$ by $\mathcal{F}_t$. This concludes the proof. ∎

Proof of Theorem B.1. (i) This follows from Lemma B.2 and the fact that $\mathcal{A}\mathcal{W}_p^{(N)}$ is a metric for each $N \in \mathbb{N}$.

(ii) To verify completeness, consider a Cauchy sequence $\mathbb{X}^m \in \mathrm{FP}_p^{(\infty)}$, $m \in \mathbb{N}$. Passing to subsequences and using a diagonalization argument, there exist an increasing sequence $(m_k)_{k \geq 1}$ and filtered processes $\mathbb{Y}^{(N)} \in \mathrm{FP}_p^{(N)}$ such that for each $N \in \mathbb{N}$, $r_N(\mathbb{X}^{m_k}) \to \mathbb{Y}^{(N)}$. By the consistency result (i.e., Lemma B.3), there is a process $\mathbb{Y} \in \mathrm{FP}_p^{(\infty)}$ such that $\mathbb{Y}^{(N)} = r_N(\mathbb{Y})$, $N \in \mathbb{N}$, and then $\mathbb{X}^{m_k} \to \mathbb{Y}$.

To see denseness of $\mathcal{P}_p(\mathbb{R}^\infty)$, we can (for instance) note that

$$\bigcup_{N \in \mathbb{N}} \left\{ (\Omega, \mathcal{F}, \mathbb{P}, (\mathcal{F}_t)_{t=1}^\infty, (X_t)_{t=1}^\infty) : (\Omega, \mathcal{F}, \mathbb{P}, (\mathcal{F}_t)_{t=1}^N, (X_t)_{t=1}^N) \in \mathrm{FP}_p^{(N)}, \right.$$
$$\left. X_s = X_N, \ \mathcal{F}_s = \mathcal{F}_N, \ s \geq N \right\} \tag{B.6}$$

is dense; hence the claim follows from Theorem 5.4.

(iii) This follows again from the denseness of the set in (B.6) together with Theorem 5.4.

(iv) and (v) First note that a function $\Psi : \mathrm{FP}_p^{(\infty)} \to \mathrm{FP}_p^{(\infty)}$ is continuous if and only if for every N, $\Psi \circ r_N$ is continuous. Hence, (iv) follows from Proposition 6.1 and (v) follows exactly as in the proof of Proposition 6.8.

(vi) This is a straightforward consequence of Lemma B.2 and Theorem 1.6.

(vii) A set $K \subseteq \mathrm{FP}_p^{(\infty)}$ is precompact if and only if all the sets $r_N[K]$, $N \geq 1$, are precompact; and a set $\tilde{K} \subseteq \mathcal{P}_p(\mathbb{R}^\infty)$ is precompact if and only if all the sets $r_N[\tilde{K}]$, $N \geq 1$, are precompact (where we interpret r_N as a function on $\mathcal{P}_p(\mathbb{R}^\infty)$ by identifying again laws and processes). The result thus follows from Theorem 1.7. ∎

Appendix C. Comments on the continuous-time case

Depending on the context and intended application, different authors have considered different 'adapted' notions of equivalence and similarity for stochastic processes. As we discuss below, these notions can be rephrased using ideas from adapted transport as considered above. Therefore, we believe that the framework and results developed in the present paper provide a blueprint for the respective theories in the continuous-time case. We describe some natural 'adapted' topologies/distances from the perspective of the present paper.

On the one end of the spectrum, the direct continuous-time extension of the distance considered in the current paper is

$$\mathcal{A}\mathcal{W}_p(\mathbb{X}, \mathbb{Y}) = \inf_{\pi \in \mathrm{Cpl}_{\mathrm{bc}}(\mathbb{X}, \mathbb{Y})} \mathbb{E}_\pi[d^p(X, Y)]^{1/p},$$

where d is a distance on the paths. This or very closely related notions are considered (on path spaces), for instance, in [2, 13, 16] and 'almost all' probabilistic operations (such as stochastic integration) are continuous with respect to this topology, see, e.g., [13]. We also note that results of the present paper concerning, e.g., the representation of the completion based on filtrations as well as geodesic properties of the distance will carry over to this setting. A disadvantage is that this topology is not separable. Moreover, there are not too many relatively compact sets, and scaled random walks do not converge to Brownian motion (no matter which distance d one chooses on the paths), cf. the argument in [61, p. 240]. This means that $\mathcal{A}\mathcal{W}_p$ is well suited, e.g., for analyzing the sensitivity of stochastic optimization problems with respect to the input process, but less suited to analyze the transition from discrete to continuous time, or to establish the existence of certain 'extremal' processes.

Bion-Nadal and Talay [35] single out a specific instance of $\mathcal{A}\mathcal{W}_2$, where d is induced by the L^2-distance and provide a Hamilton–Jacobi–Bellman equation for the calculation of $\mathcal{A}\mathcal{W}_2$. In particular, they obtain a numerically tractable version of an adapted Wasserstein distance. This line of research is continued by Backhoff–Källblad–Robinson [19].

A further variation of the distance $\mathcal{A}\mathcal{W}_p$ for semi-martingales absolutely continuous with respect to Wiener measure and with Cameron–Martin cost is considered by Lasalle [72] who uses it to provide new interpretations of transport-information inequalities on the Wiener space. Föllmer [49, 50] investigates yet another adapted Wasserstein distance for semi-martingales with drift dominated by quadratic variation and extends Talagrand's inequality to measures which are not absolutely continuous with respect to Wiener measure.

On another end of the spectrum of weak adapted topologies lie the contributions initiated by Aldous [9] and Hoover–Keisler [61, 62]. Historically, the first extension of the weak topology to take information into account was the extended weak topology as defined by Aldous [9]. The idea is to identify the law of a process with the law of its prediction process and then to measure the distance of stochastic processes through the distance of the respective prediction processes. Hoover and Keisler [61, 62] build on this idea and construct an infinitely iterated prediction process which captures further properties of the underlying filtration of the stochastic process. In view of Theorem 7.1, this infinite iteration is already necessary if the goal is to capture the information required for optimal stopping problems. It would be possible to metrize the Hoover–Keisler topology through a Wasserstein-type distance if one modifies the bicausality condition (in spirit similar to the J_1 metric on càdlàg paths):

$$\inf_{\pi \in \mathrm{Cpl}(\mathbb{X}, \mathbb{Y})} (\mathbb{E}_\pi [d^p(X, Y)]^{1/p} + \text{penalization how far } \pi \text{ is from being bicausal}).$$

Although the continuity of probabilistic operations with respect to this distance becomes more subtle, this distance has important advantages. First, as established by Hoover [61], many discrete-time objects such as scaled random walks converge in this topology to their continuous-time limit. Further, just as in our paper, tightness of the laws of the processes guarantees relative compactness and thus there are many relatively compact sets.

We believe that the theory on adapted optimal transport developed in this paper forms the foundation for analyzing the continuous-time theories from the (geometric and metric) perspective of optimal transport. Although we expect it can provide significant new insights, exploring the details is beyond the scope of this paper, and we defer this aspect to future research.

References

[1] Acciaio, B., Backhoff-Veraguas, J., Pammer, G.: Quantitative fundamental theorem of asset pricing. arXiv:2209.15037 (2024)

[2] Acciaio, B., Backhoff-Veraguas, J., Zalashko, A.: Causal optimal transport and its links to enlargement of filtrations and continuous-time stochastic optimization. Stochastic Process. Appl. **130**, 2918–2953 (2020) Zbl 1435.90012 MR 4080734

[3] Acciaio, B., Beiglböck, M., Pammer, G.: Weak transport for non-convex costs and model-independence in a fixed-income market. Math. Finance **31**, 1423–1453 (2021) Zbl 1522.91255 MR 4332852

[4] Acciaio, B., Hou, S.: Convergence of adapted empirical measures on $\mathbb{R}^d$. Ann. Appl. Probab. **34**, 4799–4835 (2024) Zbl 07927502 MR 4801582

[5] Acciaio, B., Kratsios, A., Pammer, G.: Designing universal causal deep learning models: The geometric (hyper)transformer. Math. Finance **34**, 671–735 (2024) Zbl 1533.91509 MR 4718938

[6] Agueh, M., Carlier, G.: Barycenters in the Wasserstein space. SIAM J. Math. Anal. **43**, 904–924 (2011) Zbl 1223.49045 MR 2801182

[7] Akbari, S., Ganassali, L., Kiyavash, N.: Learning causal graphs via monotone triangular transport maps. arXiv:2305.18210 (2023)

[8] Aksamit, A., Deng, S., Obłój, J., Tan, X.: The robust pricing-hedging duality for American options in discrete time financial markets. Math. Finance **29**, 861–897 (2019) Zbl 1432.91116 MR 3974286

[9] Aldous, D. J.: Weak convergence and general theory of processes. https://www.stat.berkeley.edu/~aldous/Papers/weak-gtp.pdf, visited on 14 December 2024 (1981)

[10] Ambrosio, L., Gigli, N., Savaré, G.: Gradient flows in metric spaces and in the space of probability measures. 2nd ed., Lectures Math. ETH Zürich, Birkhäuser, Basel (2008) Zbl 1145.35001 MR 2401600

[11] Backhoff, J., Bartl, D., Beiglböck, M., Wiesel, J.: Estimating processes in adapted Wasserstein distance. Ann. Appl. Probab. **32**, 529–550 (2022) Zbl 1492.60104 MR 4386535

[12] Backhoff, J., Beiglböck, M., Lin, Y., Zalashko, A.: Causal transport in discrete time and applications. SIAM J. Optim. **27**, 2528–2562 (2017) Zbl 1387.90168 MR 3738324

[13] Backhoff-Veraguas, J., Bartl, D., Beiglböck, M., Eder, M.: Adapted Wasserstein distances and stability in mathematical finance. Finance Stoch. **24**, 601–632 (2020) Zbl 1440.91036 MR 4118990

[14] Backhoff-Veraguas, J., Bartl, D., Beiglböck, M., Eder, M.: All adapted topologies are equal. Probab. Theory Related Fields **178**, 1125–1172 (2020) Zbl 1476.60079 MR 4168395

[15] Backhoff Veraguas, J., Beiglböck, M., Eder, M., Pichler, A.: Fundamental properties of process distances. Stochastic Process. Appl. **130**, 5575–5591 (2020) Zbl 1454.60048 MR 4127339

[16] Backhoff-Veraguas, J., Beiglböck, M., Huesmann, M., Källblad, S.: Martingale Benamou–Brenier: A probabilistic perspective. Ann. Probab. **48**, 2258–2289 (2020) Zbl 1473.60071 MR 4152642

[17] Backhoff-Veraguas, J., Beiglböck, M., Pammer, G.: Weak monotone rearrangement on the line. Electron. Commun. Probab. **25**, article no. 18 (2020) Zbl 1447.60047 MR 4069738

[18] Backhoff-Veraguas, J., Fontbona, J., Rios, G., Tobar, F.: Bayesian learning with Wasserstein barycenters. ESAIM Probab. Stat. **26**, 436–472 (2022) Zbl 07730417 MR 4519227

[19] Backhoff-Veraguas, J., Källblad, S., Robinson, B. A.: Adapted Wasserstein distance between the laws of SDEs. arXiv:2209.03243 (2024)

[20] Backhoff-Veraguas, J., Pammer, G.: Applications of weak transport theory. Bernoulli **28**, 370–394 (2022) Zbl 1516.49042 MR 4337709

[21] Bartl, D., Wiesel, J.: Sensitivity of multiperiod optimization problems with respect to the adapted Wasserstein distance. SIAM J. Financial Math. **14**, 704–720 (2023) Zbl 1520.91364 MR 4599319

[22] Bayraktar, E., Dolinsky, Y., Guo, J.: Continuity of utility maximization under weak convergence. Math. Financ. Econ. **14**, 725–757 (2020) Zbl 1461.91288 MR 4135703

[23] Bayraktar, E., Dolinskyi, L., Dolinsky, Y.: Extended weak convergence and utility maximisation with proportional transaction costs. Finance Stoch. **24**, 1013–1034 (2020) Zbl 1448.91271 MR 4152280

[24] Bayraktar, E., Han, B.: Fitted value iteration methods for bicausal optimal transport. arXiv:2306.12658 (2023)

[25] Bayraktar, E., Han, B.: Equilibrium transport with time-inconsistent costs: An application to matching problems in the job market. arXiv:2302.01498 (2024)

[26] Bayraktar, E., Zhou, Z.: No-arbitrage and hedging with liquid American options. Math. Oper. Res. **44**, 468–486 (2019) Zbl 1433.91169 MR 3959081

[27] Beiglböck, M., Cox, A. M. G., Huesmann, M.: Optimal transport and Skorokhod embedding. Invent. Math. **208**, 327–400 (2017) Zbl 1371.60072 MR 3639595

[28] Beiglböck, M., Cox, A. M. G., Huesmann, M.: The geometry of multi-marginal Skorokhod embedding. Probab. Theory Related Fields **176**, 1045–1096 (2020) Zbl 1469.60126 MR 4087489

[29] Beiglböck, M., Henry-Labordère, P., Penkner, F.: Model-independent bounds for option prices—a mass transport approach. Finance Stoch. **17**, 477–501 (2013) Zbl 1277.91162 MR 3066985

[30] Beiglböck, M., Jourdain, B., Margheriti, W., Pammer, G.: Approximation of martingale couplings on the line in the adapted weak topology. Probab. Theory Related Fields **183**, 359–413 (2022) Zbl 1504.60067 MR 4421177

[31] Beiglböck, M., Jourdain, B., Margheriti, W., Pammer, G.: Stability of the weak martingale optimal transport problem. Ann. Appl. Probab. **33**, 5382–5412 (2023) Zbl 1532.49041 MR 4677736

[32] Beiglböck, M., Lacker, D.: Denseness of adapted processes among causal couplings. arXiv:1805.03185 (2020)

[33] Beiglböck, M., Nutz, M., Touzi, N.: Complete duality for martingale optimal transport on the line. Ann. Probab. **45**, 3038–3074 (2017) Zbl 1417.60032 MR 3706738

[34] Beiglböck, M., Pammer, G., Schrott, S.: Denseness of biadapted Monge mappings. arXiv:2210.15554 (2023)

[35] Bion-Nadal, J., Talay, D.: On a Wasserstein-type distance between solutions to stochastic differential equations. Ann. Appl. Probab. **29**, 1609–1639 (2019) Zbl 1412.60115 MR 3914552

[36] Bonnier, P., Liu, C., Oberhauser, H.: Adapted topologies and higher rank signatures. Ann. Appl. Probab. **33**, 2136–2175 (2023) Zbl 1537.60038 MR 4583667

[37] Breeden, D. T., Litzenberger, R. H.: Prices of state-contingent claims implicit in option prices. J. Bus. **51**, 621–651 (1978)

[38] Chen, S., Lim, S., Mémoli, F., Wan, Z., Wang, Y.: The Weisfeiler–Lehman distance: Reinterpretation and connection with GNNs. In: Proceedings of 2nd Annual Workshop on Topology, Algebra, and Geometry in Machine Learning (TAG-ML), Proc. Mach. Learn. Res. (PMLR) 221, Proceedings of Machine Learning Research, New York, 404–425 (2023)

[39] Cheridito, P., Eckstein, S.: Optimal transport and Wasserstein distances for causal models. arXiv:2303.14085 (2024)

[40] Cheridito, P., Kiiski, M., Prömel, D. J., Soner, H. M.: Martingale optimal transport duality. Math. Ann. **379**, 1685–1712 (2021) Zbl 1462.49072 MR 4238277

[41] Cox, A. M. G., Obłój, J., Touzi, N.: The Root solution to the multi-marginal embedding problem: an optimal stopping and time-reversal approach. Probab. Theory Related Fields **173**, 211–259 (2019) Zbl 1478.60133 MR 3916107

[42] Cuturi, M., Doucet, A.: Fast computation of wasserstein barycenters. In: Proceedings of the 31st International Conference on Machine Learning, Proc. Mach. Learn. Res. (PMLR) 32, Proceedings of Machine Learning Research, Bejing, 685–693 (2014)

[43] Dolinsky, Y.: Hedging of game options under model uncertainty in discrete time. Electron. Commun. Probab. **19**, article no. 19 (2014) Zbl 1304.91216 MR 3183572

[44] Eckstein, S., Pammer, G.: Computational methods for adapted optimal transport. Ann. Appl. Probab. **34**, 675–713 (2024) Zbl 07829153 MR 4696288

[45] Eder, M.: Compactness in adapted weak topologies. arXiv:1905.00856 (2019)

[46] El Hamri, M., Bennani, Y., Falih, I.: Hierarchical optimal transport for unsupervised domain adaptation. Mach. Learn. **111**, 4159–4182 (2022) Zbl 07694461 MR 4514167

[47] Émery, M., Schachermayer, W.: On Vershik's standardness criterion and Tsirelson's notion of cosiness. In: Séminaire de Probabilités, XXXV, Lecture Notes in Math. 1755, Springer, Berlin, 265–305 (2001) Zbl 1001.60039 MR 1837293

[48] Ethier, S. N., Kurtz, T. G.: Markov processes: Characterization and convergence. Wiley Ser. Probab. Math. Statist. Probab. Math. Statist., John Wiley & Sons, New York (1986) Zbl 1089.60005 MR 838085

[49] Föllmer, H.: Doob decomposition, Dirichlet processes, and entropies on Wiener space. In: Dirichlet forms and related topics, Springer Proc. Math. Stat. 394, Springer, Singapore, 119–141 (2022) Zbl 1497.60065 MR 4508843

[50] Föllmer, H.: Optimal couplings on Wiener space and an extension of Talagrand's transport inequality. In: Stochastic analysis, filtering, and stochastic optimization, Springer, Cham, 147–175 (2022) Zbl 1504.60088 MR 4433813

[51] Gigli, N.: On the geometry of the space of probability measures in $\mathbb{R}^n$ endowed with the quadratic optimal transport distance. Ph.D. thesis, Scuola Normale Superiore di Pisa (2004)

[52] Glanzer, M., Pflug, G. C., Pichler, A.: Incorporating statistical model error into the calculation of acceptability prices of contingent claims. Math. Program. **174**, 499–524 (2019) Zbl 1421.90095 MR 3935089

[53] Gozlan, N., Roberto, C., Samson, P.-M., Tetali, P.: Kantorovich duality for general transport costs and applications. J. Funct. Anal. **273**, 3327–3405 (2017) Zbl 1406.60032 MR 3706606

[54] Guo, G., Obłój, J.: Computational methods for martingale optimal transport problems. Ann. Appl. Probab. **29**, 3311–3347 (2019) Zbl 1433.49043 MR 4047982

[55] Guyon, J., Menegaux, R., Nutz, M.: Bounds for VIX futures given S&P 500 smiles. Finance Stoch. **21**, 593–630 (2017) Zbl 1422.91698 MR 3663638

[56] Han, B.: Distributionally robust risk evaluation with a causality constraint and structural information. arXiv:2203.10571 (2024)

[57] Hellwig, M. F.: Sequential decisions under uncertainty and the maximum theorem. J. Math. Econom. **25**, 443–464 (1996) Zbl 0870.90006 MR 1394333

[58] Hellwig, M. F., Schmidt, K. M.: Discrete-time approximations of the Holmstrom–Milgrom Brownian-motion model of intertemporal incentive provision. Econometrica **70**, 2225–2264 (2002) Zbl 1141.60364 MR 1939896

[59] Hobson, D., Neuberger, A.: Model uncertainty and the pricing of American options. Finance Stoch. **21**, 285–329 (2017) Zbl 1380.91131 MR 3590709

[60] Hoover, D. N.: A characterization of adapted distribution. Ann. Probab. **15**, 1600–1611 (1987) Zbl 0634.60033 MR 905352

[61] Hoover, D. N.: Convergence in distribution and Skorokhod convergence for the general theory of processes. Probab. Theory Related Fields **89**, 239–259 (1991) Zbl 0725.60005 MR 1113218

[62] Hoover, D. N., Keisler, H. J.: Adapted probability distributions. Trans. Amer. Math. Soc. **286**, 159–201 (1984) Zbl 0548.60019 MR 756035

[63] Horvath, B., Lemercier, M., Liu, C., Lyons, T., Salvi, C.: Optimal stopping via distribution regression: a higher rank signature approach. arXiv:2304.01479 (2023)

[64] Jourdain, B., Margheriti, W.: One dimensional martingale rearrangement couplings. ESAIM Probab. Stat. **26**, 495–527 (2022) Zbl 1528.60034 MR 4525181

[65] Kadkhodaie, Z., Guth, F., Mallat, S., Simoncelli, E. P.: Learning multi-scale local conditional probability models of images. arXiv:2303.02984 (2023)

[66] Kallenberg, O.: Foundations of modern probability. Probab. Appl., Springer, New York (1997) Zbl 0892.60001 MR 1464694

[67] Kechris, A. S.: Classical descriptive set theory. Grad. Texts in Math. 156, Springer, New York (1995) Zbl 0819.04002 MR 1321597

[68] Kim, Y.-H., Pass, B.: Wasserstein barycenters over Riemannian manifolds. Adv. Math. **307**, 640–683 (2017) Zbl 1373.60006 MR 3590527

[69] Kirui, K. B., Pflug, G. C., Pichler, A.: New algorithms and fast implementations to approximate stochastic processes. arXiv:2012.01185 (2020)

[70] Knight, F. B.: A predictive view of continuous time processes. Ann. Probab. **3**, 573–596 (1975) Zbl 0317.60018 MR 383513

[71] Kurtz, T.: The Yamada–Watanabe–Engelbert theorem for general stochastic equations and inequalities. Electron. J. Probab. **12**, 951–965 (2007) Zbl 1157.60068 MR 2336594

[72] Lassalle, R.: Causal transport plans and their Monge–Kantorovich problems. Stoch. Anal. Appl. **36**, 452–484 (2018) Zbl 1385.93087 MR 3784142

[73] Le Gouic, T., Loubes, J.-M.: Existence and consistency of Wasserstein barycenters. Probab. Theory Related Fields **168**, 901–917 (2017) Zbl 1406.60019 MR 3663634

[74] Lisini, S.: Characterization of absolutely continuous curves in Wasserstein spaces. Calc. Var. Partial Differential Equations **28**, 85–120 (2007) Zbl 1132.60004 MR 2267755

[75] McCann, R. J.: A convexity theory for interacting gases and equilibrium crystals. Ph.D. thesis, Princeton University, https://www.math.toronto.edu/~mccann/papers/short.pdf, visited on 14 December 2024 (1994)

[76] Moulos, V.: Bicausal optimal transport for Markov chains via dynamic programming. In: IEEE International Symposium on Information Theory (ISIT), IEEE, 1688–1693 (2021)

[77] Nielsen, F., Sun, K.: Chain rule optimal transport. In: Progress in information geometry—theory and applications, Signals Commun. Technol., Springer, Cham, 191–217 (2021) Zbl 1473.49052 MR 4238624

[78] Panaretos, V. M., Zemel, Y.: Statistical aspects of Wasserstein distances. Annu. Rev. Stat. Appl. **6**, 405–431 (2019) MR 3939527

[79] Pflug, G. C.: Version-independence and nested distributions in multistage stochastic optimization. SIAM J. Optim. **20**, 1406–1420 (2009) Zbl 1198.90307 MR 2587733

[80] Pflug, G. C., Pichler, A.: A distance for multistage stochastic optimization models. SIAM J. Optim. **22**, 1–23 (2012) Zbl 1262.90118 MR 2902682

[81] Pflug, G. C., Pichler, A.: Multistage stochastic optimization. Springer Ser. Oper. Res. Financ. Eng., Springer, Cham (2014) Zbl 1317.90220 MR 3288310

[82] Pflug, G. C., Pichler, A.: Dynamic generation of scenario trees. Comput. Optim. Appl. **62**, 641–668 (2015) Zbl 1337.90047 MR 3426120

[83] Pflug, G. C., Pichler, A.: From empirical observations to tree models for stochastic optimization: Convergence properties. SIAM J. Optim. **26**, 1715–1740 (2016) Zbl 1346.90645 MR 3543169

[84] Pham, H., Wei, X.: Discrete time McKean–Vlasov control problem: A dynamic programming approach. Appl. Math. Optim. **74**, 487–506 (2016) Zbl 1360.49018 MR 3575613

[85] Pichler, A., Schlotter, R.: Martingale characterizations of risk-averse stochastic optimization problems. Math. Program. **181**, 377–403 (2020) Zbl 1445.90068 MR 4109507

[86] Pichler, A., Shapiro, A.: Mathematical foundations of distributionally robust multistage optimization. SIAM J. Optim. **31**, 3044–3067 (2021) Zbl 1481.90242 MR 4344889

[87] Pichler, A., Weinhardt, M.: The nested Sinkhorn divergence to learn the nested distance. Comput. Manag. Sci. **19**, 269–293 (2022) Zbl 07557211 MR 4423922

[88] Rabin, J., Peyré, G., Delon, J., Bernot, M.: Wasserstein barycenter and its application to texture mixing. In: Scale space and variational methods in computer vision, Lecture Notes in Comput. Sci. 6667, Springer, Berlin, 435–446 (2012)

[89] Rüschendorf, L.: The Wasserstein distance and approximation theorems. Z. Wahrsch. Verw. Gebiete **70**, 117–129 (1985) Zbl 0554.60024 MR 795791

[90] Salvi, C., Lemercier, M., Liu, C., Horvath, B., Damoulas, T., Lyons, T.: Higher order kernel mean embeddings to capture filtrations of stochastic processes. In: Advances in neural information processing systems, 34, Curran Associates, Red Hook, NY, 16635–16647 (2021)

[91] Vershik, A. M.: Decreasing sequences of measurable partitions and their applications. Sov. Math. Dokl. **11**, 1007–1011 (1970) Zbl 0238.28011

[92] Vershik, A. M.: The theory of decreasing sequences of measurable partitions. St. Petersburg Math. J. **6**, 705–761 (1995) Zbl 0853.28009 MR 1304093

[93] Vershik, A. M.: The theory of filtrations of subalgebras, standardness, and independence. Russian Math. Surveys **72**, 257–333 (2017) Zbl 1370.05220 MR 3635438

[94] Villani, C.: Optimal transport. Old and new. Grundlehren Math. Wiss. 338, Springer, Berlin (2009) Zbl 1156.53003 MR 2459454

[95] Wiesel, J.: Continuity of the martingale optimal transport problem on the real line. Ann. Appl. Probab. **33**, 4645–4692 (2023) Zbl 07789644 MR 4674061

[96] Wiesel, J. C. W.: Measuring association with Wasserstein distances. Bernoulli **28**, 2816–2832 (2022) Zbl 07594079 MR 4474563

[97] Xu, T., Acciaio, B.: Conditional COT-GAN for video prediction with kernel smoothing. arXiv:2106.05658 (2022)

[98] Xu, T., Wenliang, L. K., Munn, M., Acciaio, B.: COT-GAN: Generating sequential data via causal optimal transport. In: Advances in neural information processing systems, 33, Curran Associates, Red Hook, NY, 8798–8809 (2020)

[99] Yamada, T., Watanabe, S.: On the uniqueness of solutions of stochastic differential equations. J. Math. Kyoto Univ. **11**, 155–167 (1971) Zbl 0236.60037 MR 278420

[100] Yurochkin, M., Claici, S., Chien, E., Mirzazadeh, F., Solomon, J. M.: Hierarchical optimal transport for document representation. In: Advances in neural information processing systems, 32, Curran Associates, Red Hook, NY, 1601–1611 (2019)