ALL ADAPTED TOPOLOGIES ARE EQUAL

JULIO BACKHOFF-VERAGUAS, DANIEL BARTL, MATHIAS BEIGLBÖCK, AND MANU EDER

Abstract. A number of researchers have introduced topological structures on the set of laws of stochastic processes. A unifying goal of these authors is to strengthen the usual weak topology in order to adequately capture the temporal structure of stochastic processes.

Aldous defines an extended weak topology based on the weak convergence of prediction processes. In the economic literature, Hellwig introduced the information topology to study the stability of equilibrium problems. Bion-Nadal and Talay introduce a version of the Wasserstein distance between the laws of diffusion processes. Pflug and Pichler consider the nested distance (and the weak nested topology) to obtain continuity of stochastic multistage programming problems. These distances can be seen as a symmetrization of Lassalle's causal transport problem, but there are also further natural ways to derive a topology from causal transport.

Our main result is that all of these seemingly independent approaches define *the same topology* in finite discrete time. Moreover we show that this 'weak adapted topology' is characterized as the coarsest topology that guarantees continuity of optimal stopping problems for continuous bounded reward functions.

Keywords: Aldous' extended weak topology, Hellwig's information topology, nested distance, causal optimal transport, stability of optimal stopping, Vershik's iterated Kantorovich distance

1. Introduction

1.1. **Outline.** If some type of natural phenomenon is modelled through a stochastic process, one might expect that the model does not describe reality in an entirely accurate way. To be able to study the impact of such inaccuracies on the problems one is trying to solve, it makes sense to equip the set of laws of stochastic processes with a suitable notion of distance or topology.

Denoting by $\Omega := \mathcal{X}^N$ the path space (where X is some Polish space and $N \in \mathbb{N}$), the set of laws of stochastic processes is $\mathcal{P}(\Omega)$, i.e. the set of probability measures on Ω.

Clearly, $\mathcal{P}(\Omega)$ carries the usual weak topology. However, this topology does not respect the time evolution of stochastic processes which has a number of potentially inconvenient consequences: e.g., problems of optimal stopping / utility maximization / stochastic programming are not continuous, arbitrary processes can be approximated by processes which are deterministic after the first period, etc. In the following we describe a number of approaches which have been developed by different authors to deal with these (and related) problems. Our main result (Theorem 1.1) is that all of these approaches actually define the same topology in the present discrete time setup. Moreover, this topology is the weakest topology which allows for continuity of optimal stopping problems.

1.2. **Adapted Wasserstein distances, nested distance.** A number of authors have independently introduced variants of the Wasserstein distance which take the temporal structure of processes into account: the definition of 'iterated Kantorovich

distance' by Vershik [58, 59] might be seen as a first construction in this direction. The topic is also considered by Rüschendorf [56]. Independently, Pflug and Pflug–Pichler [50, 54, 51, 52, 53, 28] introduce the nested distance and describe the concept's rich potential for the approximation of stochastic multi-period optimization problems. Lassalle [44] considers the 'causal transport problem' that leads to a corresponding notion of distance. Once again independently of these developments, Bion-Nadal and Talay [15] define an adapted version of the Wasserstein distance between laws of solutions to SDEs.

To set the stage for describing these 'adapted' variants let us fix $p \geq 1$ and recall the definition of the usual p-Wassterstein distance.

$(\mathcal{X}, \rho_{\mathcal{X}})$ is now a Polish metric space. On $\Omega = \mathcal{X}^N$ we use the Polish metric $\rho_\Omega((x_t)_t, (y_t)_t) := (\sum_t \rho_{\mathcal{X}}(x_t, y_t)^p)^{1/p}$. Typically, when clear from the context we will omit the subscript for the metric. We use $(X_t)_t$ to denote the canonical process on Ω, i.e. X_t is the projection onto the t-th factor of $\Omega = \mathcal{X}^N$. On $\Omega \times \Omega$ call $X = (X_t)_t$ the projection on the first factor and call $Y = (Y_t)_t$ the projection on the second factor. For $\mu, \nu \in \mathcal{P}(\Omega)$ we denote by $\mathrm{Cpl}\,(\mu, \nu)$ the set of probability measures π on $\Omega \times \Omega$ for which $X \sim \mu$ and $Y \sim \nu$ under π, i.e. for which the distribution of X under π is μ and that of Y under π is ν. In applications, a particular role is played by Monge couplings. A Monge coupling from μ to ν is a coupling π for which $Y = T(X)$ π-a.s. for some Borel mapping $T : \Omega \to \Omega$ that transports μ to ν, i.e. satisfies $T_\#(\mu) = \nu$.

For $\mu, \nu \in \mathcal{P}_p(\Omega)$, i.e. for probability measures on Ω with finite p-th moment their p-Wasserstein distance is

$$\mathcal{W}_p(\mu, \nu) := \inf \left\{ \mathbb{E}^\pi \left(\rho(X, Y)^p\right)^{1/p} : \pi \in \mathrm{Cpl}(\mu, \nu) \right\} . \tag{1}$$

Following, [55] the infimum in (1) remains unchanged if one minimizes only over Monge couplings in many situations.

To motivate the formal definition of the adapted cousins in (5) and (6) below, we start with an informal discussion in terms of Monge mappings: In probabilistic terms, the preservation of mass assumption $T_\#(\mu) = \nu$ asserts

$$\big(T_1(X_1, \dots, X_N), \dots, T_N(X_1, \dots, X_N)\big) \sim \nu, \tag{2}$$

which ignores the evolution of μ and ν (resp.) in time. Rather it would appear more natural to restrict to mappings $(T_k)_{k=1}^N$ which are adapted in the sense that T_k depends only on $X_1, \dots, X_k$. Adapted Wasserstein distances can be defined following precisely this intuition, relying on a suitable version of adaptedness on the level of couplings:

The set $\mathrm{Cpl}_c(\mu, \nu)$ of *causal couplings*[1] consists of all $\pi \in \mathrm{Cpl}(\mu, \nu)$ such that

$$\pi\big((Y_1, \dots, Y_t) \in A | X\big) = \pi\big((Y_1, \dots, Y_t) \in A | X_1, \dots X_t\big). \tag{3}$$

for all $t \leq N$ and $A \subseteq \mathcal{X}^t$ measurable, cf. [44]. The set of all *bi-causal* couplings $\mathrm{Cpl}_{bc}(\mu, \nu)$ consists of all $\pi \in \mathrm{Cpl}_c(\mu, \nu)$ such that the distribution of (Y, X) under π is also in $\mathrm{Cpl}_c(\nu, \mu)$, i.e. that (3) also holds with the roles of X and Y reversed.

The term *causal* was introduced by Lassalle [44], who considers a causal transport problem in which the usual set of couplings is replaced by the set of causal couplings. The resulting concept is not actually a metric as it lacks symmetry, but as suggested by Soumik Pal, this is easily mended and we formally define the *causal* - and *symmetrized-causal p-Wasserstein distance*, resp. as follows:

[1] Intuitively, at time t, given the past $(X_1, \dots, X_t)$ of X, the distribution of Y_t does not depend on the future $(X_{t+1}, \dots, X_N)$ of X. For absolutely continuous measures μ, the weak closure of the set of adapted Monge couplings, i.e. of those $\pi \in \mathrm{Cpl}\,(\mu, \nu)$ for which $Y = T(X)$ π-a.s. for T adapted, is precisely the set of all causal couplings, see [42].

For $\mu, \nu \in \mathcal{P}_p(\Omega)$ set

$$\mathcal{CW}_p(\mu,\nu) := \inf\left\{\mathbb{E}^\pi\left(\rho(X,Y)^p\right)^{1/p} : \pi \in \mathrm{Cpl}_c(\mu,\nu)\right\} \tag{4}$$

$$\mathcal{SCW}_p(\mu,\nu) := \max\left(\mathcal{CW}_p(\mu,\nu),\, \mathcal{CW}_p(\nu,\mu)\right) . \tag{5}$$

We use the term *adapted Wasserstein distance* for

$$\mathcal{AW}_p(\mu,\nu) := \inf\left\{\mathbb{E}^\pi\left(\rho(X,Y)^p\right)^{1/p} : \pi \in \mathrm{Cpl}_{bc}(\mu,\nu)\right\} . \tag{6}$$

Rüschendorf [56] refers to $\mathcal{AW}_p$ as 'modified Wasserstein distance'. Pflug-Pichler [50, Definition 1] use the names multi-stage distance of order p and nested distance. It can also be considered as a discrete time version of the 'Wasserstein-type distance' of Bion-Nadal and Talay [15]. In [4] we use a slightly modified definition of $\mathcal{AW}_p$ which scales better with the number of time-periods N but leads to an equivalent metric (for fixed p and N). We shall discuss further properties of $\mathcal{AW}_p$ (and in particular the connection with Vershik's iterated Kantorovich distance) in Section 1.8 below.

1.3. **Hellwig's information topology.** The information topology introduced by Hellwig in [29] (as well as Aldous' extended weak topology which we discuss next) is based on the idea that an essential part of the structure of a process is the information that we may deduce about the future behaviour of the process given its behaviour up to current time t. For a process whose law is μ, this information is captured by the conditional law $\mathcal{L}^\mu(X_{t+1},\dots,X_N|X_1 = x_1,\dots,X_t = x_t)$ of $X_{t+1},\dots,X_N$ given $X_1 = x_1,\dots,X_t = x_t$ under μ.

$\mathcal{L}^\mu(X_{t+1},\dots,X_N|X_1 = x_1,\dots,X_t = x_t)$ is also the disintegration $\mu_{x_1,\dots,x_t}$ of $\mu \in \mathcal{P}(\Omega)$ w.r.t. the first t coordinates.

Hellwig's *information topology* is the initial topology w.r.t. a family of maps $(\mathcal{I}_t)_{t=1}^{N-1}$ which are defined based on these disintegrations:

$$\begin{aligned}\mathcal{I}_t : \mathcal{P}(\Omega) &\to \mathcal{P}\left(\mathcal{X}^t \times \mathcal{P}\left(\mathcal{X}^{N-t}\right)\right)\\ \mathcal{I}_t(\mu) &:= k^t_\#(\mu)\\ k^t(x_1,\dots,x_N) &:= (x_1,\dots,x_t,\mu_{x_1,\dots,x_t})\end{aligned}$$

Equivalently, $\mathcal{I}_t(\mu)$ is the joint law of

$$X_1,\dots,X_t,\mathcal{L}^\mu(X_{t+1},\dots,X_N|X_1,\dots,X_t)$$

under μ, and Hellwig's information topology is therefore the coarsest topology which makes continuous for all t the maps which send a probability μ to the joint law describing the evolution of the coordinate process up to time t and the prediction about the future behaviour of the coordinate process after t.

The work of Hellwig [29] was motivated by questions of stability in dynamic economic models/games; see the related articles [38, 57, 30, 10].

1.4. **Aldous' extended weak topology.** Aldous [3] introduces a type of convergence for pairs of filtrations and continuous time stochastic processes on them that he calls *extended weak convergence* [3, Definition 15.2]. Restricted to our current setting, his definition can be paraphrased in a similar manner as that of the information topology. Aldous' idea is to represent a stochastic process with law μ through the associated *prediction process*[2], that is, the process given by

$$Z^\mu_0 := \mathcal{L}(X) = \mu, Z^\mu_1 := \mathcal{L}^\mu(X|X_1),\dots,Z^\mu_N := \mathcal{L}^\mu(X|X_1,\dots,X_N).$$

That is, $(Z^\mu_t)_{t=0}^N$ is a measure-valued martingale that makes increasingly accurate predictions about the full trajectory of the process X.

[2]The definition of the prediction process goes back at least to Knight [39].

Rather then comparing the laws of processes directly, the *extended weak topology* is derived from the weak topology on the corresponding prediction processes (plus the original processes). I.e. formally, the extended weak topology on $\mathcal{P}(\Omega)$ is the initial topology w.r.t. the map

$$\mathcal{E} : \mathcal{P}(\Omega) \to \mathcal{P}\Big(\Omega \times \mathcal{P}(\Omega)^{N+1}\Big)$$

which sends μ to the joint distribution of

$$(X, Z^\mu) = (X_1, \dots, X_N, \mu, \mathcal{L}^\mu(X|X_1), \mathcal{L}^\mu(X|X_1, X_2), \dots, \mathcal{L}^\mu(X|X_1, \dots, X_N))$$

under μ.

Note that, to stay faithful to Aldous' original definition, we defined $\mathcal{E}$ to map μ not just to the law of the prediction process but to the joint law of the original process and its prediction process. One easily checks that the original process may be omitted in our setting without changing the resulting topology.

1.5. **The optimal stopping topology.** The usual weak topology on $\mathcal{P}(\Omega)$ is the coarsest topology which makes continuous all the functions

$$\mu \mapsto \int f \, \mathrm{d}\mu$$

for $f : \Omega \to \mathbb{R}$ continuous and bounded.

One may follow a similar pattern and look at the coarsest topology which makes continuous the outcomes of all sequential decision procedures. Perhaps the easiest way to formalize this is to look at optimal stopping problems. In detail, write $AC(\Omega)$ for the set of all processes $(L_t)_{t=1}^N$ which are adapted, bounded and satisfy that $x \mapsto L_t(x)$ is continuous for each $t \leq N$. Write $v^L(\mu)$ for the corresponding value function, given that the process X follows the law μ, i.e.

$$v^L(\mu) := \inf\{\mathbb{E}^\mu\left(L_\tau\right) : \tau \leq N \text{ is a stopping time}\}.$$

The *optimal stopping topology* on $\mathcal{P}(\Omega)$ is the coarsest topology which makes the functions

$$\mu \mapsto v^L(\mu)$$

continuous for all $(L_t)_{t=1}^N \in AC(\Omega)$.

1.6. **Main result.** We can now state our main result:

Theorem 1.1. *Let $(\mathcal{X}, \rho_\mathcal{X})$ be a Polish metric space, where $\rho_\mathcal{X}$ is a bounded metric and set $\Omega := \mathcal{X}^N$. Then the following topologies on $\mathcal{P}(\Omega)$ are equal*

(1) the topology induced by $\mathcal{AW}_p$
(2) the topology induced by $\mathcal{SCW}_p$
(3) Hellwig's information topology
(4) Aldous' extended weak topology
(5) the optimal stopping topology.

The assumption that $\rho_\mathcal{X}$ is bounded serves only to simplify the statement of the theorem, because in this case the topology induced by $\mathcal{W}_p$ coincides with the weak topology. For every Polish space there is a bounded complete metric which induces the topology (given any complete metric $\rho_\mathcal{X}$, replace it by e.g. $\min(1, \rho_\mathcal{X})$).

1.6.1. *p-Wasserstein and unbounded metrics.* There is an analogous statement, Theorem 1.2 below, which drops the assumption that $\rho_{\mathcal{X}}$ is bounded. To be able to state it, we introduce slight variations of Hellwig's information topology, of Aldous' extended weak topology and of the optimal stopping topology:

In [29] Hellwig equips the target spaces of $\mathcal{I}_t$ with the weak topology – or more precisely he equips $\mathcal{P}\big(\mathcal{X}^{N-t}\big)$ with the weak topology, $\mathcal{X}^t \times \mathcal{P}\big(\mathcal{X}^{N-t}\big)$ with the product topology and finally $\mathcal{P}\big(\mathcal{X}^t \times \mathcal{P}\big(\mathcal{X}^{N-t}\big)\big)$ with the weak topology based on this topology. One may easily define a p-Wasserstein version of Hellwigs information topology by using the recipe 'replace the weak topology by the p-Wasserstein metric everywhere'. Concretely, if we restrict $\mathcal{I}_t$ to $\mathcal{P}_p(\Omega)$, we may view it as a map into $\mathcal{P}_p\big(\mathcal{X}^t \times \mathcal{P}_p\big(\mathcal{X}^{N-t}\big)\big)$, where the last space carries the metric

$$\rho_{\mathcal{P}_p(\mathcal{X}^t \times \mathcal{P}_p(\mathcal{X}^{N-t}))}(\mu,\nu) := \inf_{\gamma \in \mathrm{Cpl}(\mu,\nu)} \Big(\int \rho((x_i)_{i\le t},(y_i)_{i\le t})^p + \mathcal{W}_p(\hat\mu,\hat\nu)^p \, \mathrm{d}\gamma((x_i)_{i\le t},\hat\mu,(y_i)_{i\le t},\hat\nu) \Big)^{1/p}.$$

We will call the resulting variant of Hellwigs information topology on $\mathcal{P}_p(\Omega)$ the *$\mathcal{W}_p$-information topology.*

Similarly, one may systematically replace every occurrence of the weak topology in the definition of the extended weak topology by the p-Wasserstein metric. We call the resulting topology on $\mathcal{P}_p(\Omega)$ the *extended $\mathcal{W}_p$-topology.*

Just like the weak topology is the coarsest topology which makes integration of continuous bounded functions continuous, the p-Wasserstein topology is the coarsest topology which makes integration of continuous functions bounded by $c \cdot (1 + \rho(x_0, x)^p)$ continuous. Following this analogy, we define $AC_p(\Omega)$ as the set of all processes $(L_t)_{t=1}^N$ which are adapted, bounded by $x \mapsto c \cdot (1 + \rho(x_0, x)^p)$ for some $c \in \mathbb{R}_+$ and satisfy that $x \mapsto L_t(x)$ is continuous for each $t \le N$.

The *$\mathcal{W}_p$-optimal stopping topology* on $\mathcal{P}_p(\Omega)$ is the coarsest topology which makes the functions

$$\mu \mapsto v^L(\mu)$$

continuous for all $(L_t)_{t=1}^N \in AC_p(\Omega)$.

With these we may state the following generalization of Theorem 1.1:

Theorem 1.2. *Let $(\mathcal{X}, \rho_{\mathcal{X}})$ be a Polish metric space and set $\Omega := \mathcal{X}^N$. Then the following topologies on $\mathcal{P}_p(\Omega)$ are equal*

(1) the topology induced by $\mathcal{AW}_p$
(2) the topology induced by $\mathcal{SCW}_p$
(3) the $\mathcal{W}_p$-information topology
(4) the extended $\mathcal{W}_p$-topology
(5) the $\mathcal{W}_p$-optimal stopping topology.

Clearly, one recovers Theorem 1.1 from Theorem 1.2 by choosing a bounded metric on $\mathcal{X}$, because the $\mathcal{W}_p$-information topology for bounded $\rho_{\mathcal{X}}$ is just the information topology, the extended $\mathcal{W}_p$-topology for bounded $\rho_{\mathcal{X}}$ is just the extended weak topology and the $\mathcal{W}_p$-optimal stopping topology for bounded $\rho_{\mathcal{X}}$ is just the optimal stopping topology.

The relationship between the topologies listed in Theorem 1.1 and those listed in Theorem 1.2 is similar to the non-adapted case where we know that usual p-Wasserstein convergence is equivalent to usual weak convergence plus convergence of the p-th moments.

Lemma 1.3. *Convergence in any of the topologies of Theorem 1.2 is equivalent to convergence in any of the topologies of Theorem 1.1 (where for building $\mathcal{SCW}_p$ and*

$\mathcal{AW}_p$, $\rho_{\mathcal{X}}$ is replaced by a bounded compatible complete metric e.g. $\min(1, \rho_{\mathcal{X}})$*) plus convergence of p-th moments on Ω w.r.t. (the original) ρ_Ω.*

We prove Lemma 1.3 in Section 6, making use of (parts of) Theorem 1.1 and Theorem 1.2.

1.7. **Further remarks on related work.**

1.7.1. *Some further articles of successors of Aldous.* One of the original applications of Aldous' weak extended topology concerned the stability of optimal stopping [3]. This corresponds to one half of (4)=(5) in Theorem 1.1, but in a much more general setting. This line of work has been continued by Lamberton and Pagès [43], Coquet and Toldo [19], among others.

Aldous' extended weak topology was also inspiring and instrumental for the development of the theory of convergence of filtrations, and the associated questions of stability of the martingale representation property and Doob-Meyer decompositions. In this regard, see the works by Hoover et al [35, 33] and by Mémin et al [18, 46]. The related question of stability of stochastic differential equations (as well as their backwards version) with respect to the *driving noise* has particularly seen a burst of activity in the last two decades. For brevity's sake we only refer to the recent article by Papapantoleon, Posamaï, and Saplaouras [48] for an overview of the many available works in this direction.

1.7.2. *Previous applications of adapted Wasserstein distances.* Pflug, Pichler and co-authors [50, 54, 51, 52, 53, 28] have extensively developed and applied the notion of nested distaces for the purpose of scenario generation, stability, sensitivity bounds, and distributionally robust stochastic optimization, in the context of operations research.

Acciaio, Zalashko, and one of the present authors consider in [2] the adapted Wasserstein distance in continuous time in connection with utility maximization, enlargement of filtrations and optimal stopping.

Causal couplings have appeared in the work by Yamada and Watanabe [60], Jacod and Mémin [36] as well as Kurtz [40, 41], concerning weak solutions of stochastic differential equations, and by Rüschendof [56] concerning approximation theorems in probability theory. The term 'causal' is first used by Lassalle [44], who uses it in an additional constraint for the transport problem and gives an alternative derivation of the Talagrand inequality for the Wiener measure. Causal couplings are also present in the numerical scheme suggested in [1] for (extended mean-field) stochastic control.

The article [6] connects adapted Wasserstein distance (in continuous time) to martingale optimal transport (cf. [32, 12, 26, 22, 16, 31, 17, 11, 13] among many others). Several familiar objects appear as solutions to variational problems in this context. E.g. geometric Brownian motion is the martingale which is closest in $\mathcal{AW}_2$ to usual Brownian motion subject having a log normal distribution at the terminal time-point, the local vol model is closest to Brownian motion subject to matching 1-d marginals.

Bion-Nadal and Talay [15] introduce an adapted Wasserstein-type distance on the set of diffusion SDEs and show that this distance corresponds to the computation of a tractable stochastic control problem. They also apply their results to the problem of fitting diffusion models to given marginals.

In [4] the present authors consider adapted Wasserstein distances in relation to stability in finance: Lipschitz continuity of utility maximization/hedging are established w.r.t. to the underlying models in discrete and continuous time.

1.8. **Another formulation of the adapted Wasserstein distance and of Hellwigs information topology.** Here we give an alternative formulation of the adapted Wasserstein distance / nested distance due to Pflug and Pichler.

Again, $\mathcal{X}$ is a Polish space and $\rho = \rho_{\mathcal{X}}$ is a compatible metric on $\mathcal{X}$. Starting with $V_N^p := 0$ we define

$$V_t^p(x_1,\dots,x_t,y_1,\dots,y_t) := \tag{7}$$
$$\inf_{\gamma^{t+1}\in \mathrm{Cpl}(\mu_{x_1,\dots,x_t},\nu_{y_1,\dots,y_t})} \iint \begin{pmatrix} V_{t+1}^p(x_1,\dots,x_{t+1},y_1,\dots,y_{t+1}) \\ +\ \rho(x_{t+1},y_{t+1})^p \end{pmatrix} \mathrm{d}\gamma^{t+1}(x_{t+1},y_{t+1}).$$

The nested distance is finally obtained in a backwards recursive way by

$$\mathcal{ND}_p(\mu,\nu)^p = \inf_{\gamma^1\in \mathrm{Cpl}(\mathrm{proj}_{1\#}(\mu),\mathrm{proj}_{1\#}(\nu))} \iint (V_1^p(x_1,y_1) +\ \rho(x_1,y_1)^p)\ \mathrm{d}\gamma^1(x_1,y_1). \tag{8}$$

Then $\mathcal{AW}_p = \mathcal{ND}_p$. We refer to [7] for the (straightforward) justification.

For $N > 1$ the adapted Wasserstein distance is not complete. As was established in [5], a natural complete space into which $(\mathcal{P}_p(\Omega), \mathcal{AW}_p)$ embeds is given by the space of *nested distributions*:

Consider the sequence of metric spaces

$$\begin{aligned} \mathcal{X}_{N:N} &:= (\mathcal{X}, \rho_{N:N}), & \rho_{N:N} &:= \rho = (\rho^p)^{1/p}, \\ \mathcal{X}_{N-1:N} &:= \left(\mathcal{X} \times \mathcal{P}_p(\mathcal{X}_{N:N}), \rho_{N-1:N}\right), & \rho_{N-1:N} &:= \left(\rho^p + \mathcal{W}^p_{\rho_{N:N},p}\right)^{1/p}, \\ &\vdots & &\vdots \\ \mathcal{X}_{1:N} &:= \left(\mathcal{X} \times \mathcal{P}_p(\mathcal{X}_{2:N}), \rho_{1:N}\right), & \rho_{1:N} &:= \left(\rho^p + \mathcal{W}^p_{\rho_{2:N},p}\right)^{1/p}, \end{aligned}$$

where at each stage t, the space $\mathcal{P}_p(\mathcal{X}_{t:N})$ is endowed with the p-Wasserstein distance with respect to the metric $\rho_{t:N}$ on $\mathcal{X}_{t:N}$, which we denote by $\mathcal{W}_{\rho_{t:N},p}$. The space of *nested distributions* (of depth N) is defined as $\mathcal{P}_p(\mathcal{X}_{1:N})$. We endow $\mathcal{P}_p(\mathcal{X}_{1:N})$ with the complete metric $\mathcal{W}_{\rho_{1:N},p}$.

The space of nested distributions was defined by Pflug [49]. Notably the idea to iterate the formation of Wasserstein spaces and metrics goes back to Vershik [58, 59] who uses the name 'iterated Kantorovich distance'. The main interest of Vershik (and his successors) lies in the classification of filtrations (in the language of ergodic theory). We refer to the work of Emery and Schachermayer [24] for a survey from a probabilistic perspective and to Janvresse, Laurent and de la Rue [37] for a contemporary article (again from a probabilistic viewpoint).

$\mathcal{P}_p(\Omega)$ is naturally embedded in the set of nested distributions of depth N through the map $\mathcal{N}$ given by

$$\mathcal{N}(\mu) := \mathcal{L}\left(X_1, \mathcal{L}\left(X_2, \cdots \mathcal{L}\left(X_{N-1}, \mathcal{L}\left(X_N \middle| \bar{X}_1^{N-1}\right) \middle| \bar{X}_1^{N-2}\right) \cdots \middle| X_1\right)\right) \tag{9}$$

where $(X_1,\dots,X_N)$ is a vector with law μ, $\mathcal{L}$ again denotes (conditional) law and we use $\bar{X}_1^t$ as a shorthand for the vector $X_1,\dots,X_t$.

Following [5], we have:

Theorem 1.4. *The map $\mathcal{N}$ defined in* (9) *embeds the metric space $(\mathcal{P}_p(\Omega), \mathcal{AW}_p)$* isometrically *into the complete separable metric space $(\mathcal{P}_p(\mathcal{X}_{1:N}), \mathcal{W}_{\rho_{1:N},p})$.*

Remark 1.5. When $\mathcal{X}$ has no isolated points, $\mathcal{P}_p(\mathcal{X}_{1:N})$ is actually the completion of $\mathcal{P}_p(\Omega)$, i.e. $\mathcal{P}_p(\Omega)$ considered as a subset of $\mathcal{P}_p(\mathcal{X}_{1:N})$ is dense.

1.8.1. *Hellwig's information topology in terms of adapted Wasserstein distances.* We note that Hellwig's definition of the information topology can also be rephrased

using the concept of adapted Wasserstein distance: Assume that $\rho_{\mathcal{X}}$ is a bounded metric and for $t \leq N$, set

$$\Omega = \mathcal{X}^N = \underbrace{\mathcal{X}^t}_{=:X_1^{(t)}} \times \underbrace{\mathcal{X}^{N-t}}_{=:X_2^{(t)}} = X_1^{(t)} \times X_2^{(t)}.$$

I.e. for each t, we consider Ω as the product of two Polish spaces (which one might consider as 'history' and 'future'). Extending the defintion of $\mathcal{AW}_p$ in the obvious way to products of not necessarily equal Polish spaces, we can then equip $\mathcal{P}_p\Big(X_1^{(t)} \times X_2^{(t)}\Big)$ with a *one period* adapted Wasserstein distance $\mathcal{AW}_p^{(t)}, p \geq 1$. Setting for $\mu, \nu \in \mathcal{P}(\Omega)$

$$\mathcal{IW}_p(\mu, \nu) := \sum_{t=1}^{N} \mathcal{AW}_p^{(t)}(\mu, \nu), \quad p \geq 1, \tag{10}$$

we obtain a compatible metric for the information topology. This is relatively straightforward (whereas the full version of Theorem 1.1 is not straightforward as far as we are concerned).

1.9. **Preservation of Compactness.** We close this section with a result about the preservation of relative compactness which we shall use in Sections 4 and 6, but which also might be of independent interest. Specifically, in [8, 9] the two-step version of Lemma 1.6 is used as a crucial tool in the investigation of the weak transport problem.

A more detailed investigation of compactness in $\mathcal{P}(\Omega)$ with the weak adapted topology is the topic of the companion paper to this one, [23].

Assume for simplicity that $\rho_{\mathcal{X}}$ is a bounded metric. Then we have

Lemma 1.6 (Compactness lemma)**.** *$A \subseteq \mathcal{P}(\Omega)$ is relatively compact w.r.t. the usual weak topology iff $\mathcal{N}[A] \subseteq \mathcal{P}(\mathcal{X}_{1:N})$ is relatively compact.*

We note that Lemma 1.6 is essentially a consequence of the characterization of compact subsets in $\mathcal{P}(\mathcal{P}(X))$; in a somewhat different framework it was first proved in [34]. The version stated here follows by repeated application of [23, Lemma 3.3]/[8, Lemma 2.6].

The implication that $\mathcal{N}[A]$ relatively compact implies A relatively compact is rather easy to see, but the other direction that A relatively compact implies $\mathcal{N}[A]$ relatively compact is nontrivial since the mapping $\mathcal{N} : \mathcal{P}(\Omega) \to \mathcal{P}(\mathcal{X}_{1:N})$ is not continuous when $\mathcal{P}(\Omega)$ is endowed with the usual weak topology (except for trivial cases). Lemma 1.6 would *not* be true if we were to replace relative compactness by compactness.

The assumption that $\rho_{\mathcal{X}}$ is bounded is inessential. A version of Lemma 1.6 holds if we replace $\mathcal{P}(\Omega)$ by $\mathcal{P}_p(\Omega)$ and the weak topology by the one induced by the p-Wasserstein metric.

A similar result based on Hellwig's information toplogy, relating relative compactness in $\mathcal{P}(\Omega)$ to relative compactness in $\prod_{t=1}^{N-1} \mathcal{P}\big(\mathcal{X}^t \times \mathcal{P}\big(\mathcal{X}^{N-t}\big)\big)$, is also true.

2. Preparations

The rest of the paper will essentially be devoted to proving Theorem 1.1, or really its generalization Theorem 1.2.

In Section 3 we prove that Hellwig's information topology equals the topology induced by $\mathcal{AW}_p$, i.e. (3) = (1) in Theorem 1.2. In a sense, of all the topologies listed in Theorem 1.2, Hellwig's information toplogy 'looks' the coarsest – or at least like one of the coarser ones, while the topology induced by $\mathcal{AW}_p$ 'looks' the finest.

In Section 4 we sandwich the topology induced by $\mathcal{SCW}_p$ between Hellwig's information topology and the toplogy induced by $\mathcal{AW}_p$, i.e. we show (3) $\leq$ (2) $\leq$ (1) in Theorem 1.2.

In Section 5 we show that Aldous' extended weak topology is equal to Hellwig's information topology, i.e. (4) = (3) in Theorem 1.2.

In Section 6 we prove Lemma 1.3.

In Section 7 we prove that the optimal stopping topology is coarser than the topology induced by $\mathcal{AW}_p$ and finer than Hellwig's ($\mathcal{W}_p$-)information topology, i.e. (3) $\leq$ (5) $\leq$ (1) in Theorem 1.2.

2.1. **Notation.** The nested structure of spaces like for example $\mathcal{P}_p(\mathcal{X}_{1:N})$ introduced in Section 1.8 is (at least for the authors) not so easy to gain an intuition for. It seems rather challenging to picture probability measures on probability measures on probability measures... etc.

Therefore, much of the proofs in the following two sections will be about bookkeeping and not getting lost in these nested structures. In most other contexts we would regard such bookkeeping as abstract nonsense better swept under the rug, but in the context of the present paper we believe that it really constitutes an important and nontrivial ingredient in successfully carrying out the proofs.

To aid in this endeavour we make some notational preparations and introduce a few conventions.

2.1.1. *Operations on Spaces.* In the introduction we described the topologies listed in Theorems 1.1 and 1.2 as initial topologies w.r.t. maps into more complex spaces. These spaces are built up from just a few basic operations, and in most cases the maps can also be constructed using a few relatively simple ingredients.

For spaces, the operations in question are

- product formation, i.e. for spaces $\mathcal{X}$ and $\mathcal{Y}$ we may form their product space $\mathcal{X} \times \mathcal{Y}$,
- and passing from a space $\mathcal{X}$ to the space $\mathcal{P}(\mathcal{X})$ of probability measures on $\mathcal{X}$.

Here we run into some tension between the various existing definitions in the literature. While Hellwig and Aldous originally defined their topologies based on equipping the space $\mathcal{P}(\mathcal{X})$ of probability measures on some space $\mathcal{X}$ with the weak topology, without any mention of metrics, $\mathcal{AW}_p$ is a metric built on the p-Wasserstein metric, and Theorem 1.4 exhibits this metric as the 'initial metric' w.r.t. an embedding of $\mathcal{P}_p(\Omega)$ (not $\mathcal{P}(\Omega)$) into $(\mathcal{P}_p(\mathcal{X}_{1:N}), \mathcal{W}_{\rho_{1:N},p})$.

Luckily, when the base metric $\rho_{\mathcal{X}}$ on $\mathcal{X}$ is bounded and we decide that we only care about topologies and not the metrics that induce them, all of these distinctions vanish, and one may hope for these fine distinctions to not be so important in the end.

To give as uniform and as streamlined a treatment as possible of all the various ways in which these metric and topological spaces can be related to each other we employ the following strategy: A lot of our arguments are agnostic to the distinction between $\mathcal{P}$ and $\mathcal{P}_p$, and to whether we are talking about metric or topological spaces etc. They only rely on properties of the operations of product formation and formation of spaces of probability measures and on properties of maps between various spaces built using these operations which hold in either case. For the rest of the paper we will therefore drop the p in $\mathcal{P}_p$ and other explicit mentions of these distinctions. The reader may decide to read the paper using either of the following two sets of conventions, which are to be applied recursively:

Convention 1 (weak topologies)

- $\mathcal{X}$, $\mathcal{Y}$, $\mathcal{Z}$, $\mathcal{A}$, $\mathcal{B}$, $\mathcal{C}$, etc. are Polish spaces.

- $\mathcal{X} \times \mathcal{Y}$ is a topological space with the product topology (again Polish).
- $\mathcal{P}(\mathcal{X})$ is a topological space with the weak topology (also Polish).
- 'space' will mean Polish space.

Convention 2 ($\mathcal{W}_p$)

- $p \geq 1$ is fixed throughout the paper
- $\mathcal{X}$, $\mathcal{Y}$, $\mathcal{Z}$, $\mathcal{A}$, $\mathcal{B}$, $\mathcal{C}$, etc. are Polish (i.e. complete separable) metric spaces with metrics $\rho_{\mathcal{X}}$, $\rho_{\mathcal{Y}}$, $\rho_{\mathcal{Z}}$, $\rho_{\mathcal{A}}$, $\rho_{\mathcal{B}}$, $\rho_{\mathcal{C}}$, etc. respectively.
- $\mathcal{X} \times \mathcal{Y}$ is a Polish metric space with the metric

$$\rho_{\mathcal{X}\times\mathcal{Y}}((x_1,y_1),(x_2,y_2)) := \left(\rho_{\mathcal{X}}(x_1,x_2)^p + \rho_{\mathcal{Y}}(y_1,y_2)^p\right)^{1/p} .$$

- $\mathcal{P}(\mathcal{X})$ is a Polish metric space with the p-Wasserstein metric

$$\rho_{\mathcal{P}(\mathcal{X})}(\mu,\nu) := \inf_{\gamma\in \mathrm{Cpl}(\mu,\nu)} \left(\int \rho(x_1,x_2)^p \,\mathrm{d}\gamma(x_1,x_2)\right)^{1/p} .$$

- The subscript on the metric ρ may be dropped when clear from the context.
- 'space' will mean Polish metric space.

Unless specified otherwise everything said from here on will be true for either way of reading. Convention 1 will lead to a direct proof of Theorem 1.1, while Convention 2 will give a proof of the more general version, Theorem 1.2. Occasionally an argument will require us to talk directly about metrics to establish continuity of some map. When one only cares about Theorem 1.1 and not Theorem 1.2 these sections can be read while assuming that $p = 1$ and that all metrics mentioned are bounded.

Another space we will need is

Definition 2.1. $\mathcal{F}(\mathcal{A} \rightsquigarrow \mathcal{B}) \subseteq \mathcal{P}(\mathcal{A} \times \mathcal{B})$ is the space of probability measures on $\mathcal{A} \times \mathcal{B}$ which are concentrated on the graph of a measuruable function, i.e.:

$$\mathcal{F}(\mathcal{A} \rightsquigarrow \mathcal{B}) := \left\{ \mu \in \mathcal{P}(\mathcal{A}\times\mathcal{B}) \,\middle|\, \exists f : \mathcal{A} \to \mathcal{B} \text{ measurable s.t. } \mu(\mathrm{graph}(f)) = 1 \right\} .$$

The space $\mathcal{F}(\mathcal{A} \rightsquigarrow \mathcal{B})$ carries the subspace topology / the restriction of the metric on $\mathcal{P}(\mathcal{A} \times \mathcal{B})$.

2.1.2. *Maps between spaces.* Assuming Convention 1, when $f : \mathcal{X} \to \mathcal{Y}$ is a continuous map, the pushforward under f, i.e. the map which sends $\mu \in \mathcal{P}(\mathcal{X})$ to the measure $\nu \in \mathcal{P}(\mathcal{Y})$ with $\nu(A) = \mu(f^{-1}[A])$ is also continuous.

Similarly, assuming Convention 2, when $f : \mathcal{X} \to \mathcal{Y}$ is a Lipschitz-continuous map between metric spaces the pushforward under f is also Lipschitz-continous from $\mathcal{P}(\mathcal{X})$ to $\mathcal{P}(\mathcal{Y})$.

We will use $\mathcal{P}(f) : \mathcal{P}(\mathcal{X}) \to \mathcal{P}(\mathcal{Y})$ to denote the pushforward under f, to emphasize the fact that $\mathcal{P}$ is a *functor*, i.e. that it sends a diagram with a 'nice' (read continuous/Lipschitz) map

$$\mathcal{X} \xrightarrow{f} \mathcal{Y}$$

to a similar diagram

$$\mathcal{P}(\mathcal{X}) \xrightarrow{\mathcal{P}(f)} \mathcal{P}(\mathcal{Y})$$

where the map is also 'nice', and that $\mathcal{P}(f \circ g) = \mathcal{P}(f) \circ \mathcal{P}(g)$ and $\mathcal{P}(1_{\mathcal{X}}) = 1_{\mathcal{P}(\mathcal{X})}$ (where $1_{\mathcal{X}}$ is the identity function on $\mathcal{X}$).

For a product of spaces $\mathcal{X} \times \mathcal{Y}$, the projection onto $\mathcal{X}$ will alternatively be denoted by either $\mathrm{proj}_{\mathcal{X}}$ or by the same letter that is used for the space, but in a non-calligrapic font, i.e. $X : \mathcal{X} \times \mathcal{Y} \to \mathcal{X}$.

If μ is defined on some product $\prod_i \mathcal{X}_i$ of spaces, we also introduce a shorthand notation for marginals of μ, i.e. for the pushforward of μ under projection onto the product of some subset of the original factors:

$$\mu_{\restriction (\mathcal{X}_{i_j})_j} = \mathcal{P}\big((X_{i_j})_j\big)\,(\mu)\ .$$

If $f : \mathcal{A} \to \mathcal{B}$ and $g : \mathcal{A} \to \mathcal{C}$ are functions we write (f, g) for the function

$$\begin{aligned}(f, g) &: \mathcal{A} \to \mathcal{B} \times \mathcal{C}\\ (f, g)(a) &:= (f(a), g(a))\ .\end{aligned}$$

If we want to specify a map from, say $\mathcal{A}\times\mathcal{B}\times\mathcal{C}$ to $\mathcal{X}$ but we only really care about one of the variables we will use an underscore '_' instead of naming the unused variables, as in $(a, \_, \_) \mapsto f(a)$. Similarly, when integrating we may also use _ to denote unused variables, i.e. for $\mu \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ we might write $\int f(y)\,\mathrm{d}\mu(\_, y)$.

Two important maps will be the disintegration map $\mathrm{dis}_{\mathcal{A}}^{\mathcal{B}}$ and its left inverse $\mathrm{int}_{\mathcal{A}}^{\mathcal{B}}$.

The disintegration map

$$\mathrm{dis}_{\mathcal{A}}^{\mathcal{B}} : \mathcal{P}(\mathcal{A} \times \mathcal{B}) \to \mathcal{F}\,(\mathcal{A} \rightsquigarrow \mathcal{P}(\mathcal{B}))$$

sends a probability μ on $\mathcal{A} \times \mathcal{B}$ to the measure

$$\mathcal{P}\big((a, \_) \mapsto (a, \mu_a)\big)(\mu)$$

where $a \mapsto \mu_a$ is a classical disintegration of μ, i.e. if $\bar{\mu} = \mathrm{dis}_{\mathcal{A}}^{\mathcal{B}}(\mu)$ then

$$\int f(a, b)\,\mathrm{d}\nu(b)\,\mathrm{d}\bar{\mu}(a, \nu) = \int f(a, b)\,\mathrm{d}\mu_a(b)\,\mathrm{d}\mu(a, \_) = \int f(a, b)\,\mathrm{d}\mu(a, b)\ .$$

The disintegration map is measurable (see for example [14, Proposition 7.27]) and injective. It is not continuous w.r.t. the weak topologies or the Wasserstein metrics.

When writing $\mathrm{dis}_{\mathcal{A}}^{\mathcal{B}}$ we will not insist that $\mathcal{A}$ has to be the first factor in the domain of $\mathrm{dis}_{\mathcal{A}}^{\mathcal{B}}$ – $\mathcal{A}$ and $\mathcal{B}$ may even be products themselves, whose factors are intermingled in the product that makes up the domain of $\mathrm{dis}_{\mathcal{A}}^{\mathcal{B}}$. Also, we may sometimes omit $\mathcal{B}$, only specifying the variable(s) w.r.t. which we are disintegrating, not the ones which are left over, as in $\mathrm{dis}_{\mathcal{A}}$.

The map

$$\begin{aligned}\mathrm{int}_{\mathcal{A}}^{\mathcal{B}} &: \mathcal{P}(\mathcal{A} \times \mathcal{P}(\mathcal{B})) \to \mathcal{P}(\mathcal{A} \times \mathcal{B})\\ \mathrm{int}_{\mathcal{A}}^{\mathcal{B}}(\mu) &:= f \mapsto \int f(a, b)\,\mathrm{d}\nu(b)\,\mathrm{d}\mu(a, \nu)\end{aligned}$$

is (Lipschitz-)continuous.

The pair $\mathrm{dis}_{\mathcal{A}}^{\mathcal{B}}$, $\mathrm{int}_{\mathcal{A}}^{\mathcal{B}}$ enjoy the following properties:

(1) $\mathrm{int}_{\mathcal{A}}^{\mathcal{B}}$ is the left inverse of the disintegration map, i.e.

$$\mathrm{int}_{\mathcal{A}}^{\mathcal{B}} \circ \mathrm{dis}_{\mathcal{A}}^{\mathcal{B}} = 1_{\mathcal{P}(\mathcal{A}\times\mathcal{B})}\ .$$

This is a direct consequence of the definition of the disintegration.

(2) $\mathrm{int}_{\mathcal{A}\restriction\mathcal{F}(\mathcal{A}\rightsquigarrow\mathcal{P}(\mathcal{B}))}^{\mathcal{B}}$ is injective. Therefore,

(3) $\mathrm{dis}_{\mathcal{A}}^{\mathcal{B}} \circ \mathrm{int}_{\mathcal{A}\restriction\mathcal{F}(\mathcal{A}\rightsquigarrow\mathcal{P}(\mathcal{B}))}^{\mathcal{B}} = 1_{\mathcal{F}(\mathcal{A}\rightsquigarrow\mathcal{P}(\mathcal{B}))}$, i.e. $\mathrm{dis}_{\mathcal{A}}^{\mathcal{B}}$ and $\mathrm{int}_{\mathcal{A}}^{\mathcal{B}}$ are inverse bijections between $\mathcal{P}(\mathcal{A} \times \mathcal{B})$ and $\mathcal{F}\,(\mathcal{A} \rightsquigarrow \mathcal{P}(\mathcal{B}))$.

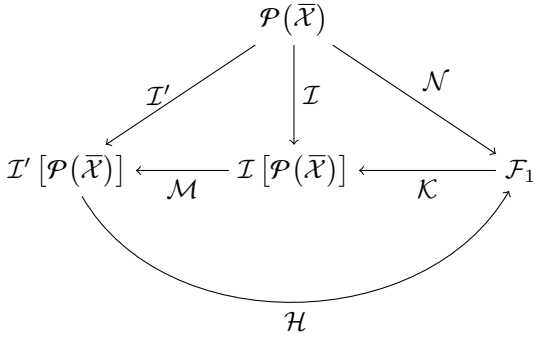
The last two properties are just a reformulation of the known fact that the disintegration of a measure is almost-surely uniquely defined.

2.1.3. *Processes which take values in different spaces at different times.* Already in the introduction, in Section 1.8.1, we found it convenient to extend the definition of $\mathcal{AW}_p$ to products of not necessarily equal Polish spaces 'in the obvious way'. To accommodate for reapplication of concepts in a similar style as seen there we make the minor generalization of letting all the processes we talk about take values in different spaces at different times – typically at time t they will take values in a space $\mathcal{X}_t$.

Denote by $\overline{\mathcal{X}}_j^k := \prod_{i=j}^k \mathcal{X}_i$ and define $\overline{\mathcal{X}} := \overline{\mathcal{X}}_1^N$, $\overline{\mathcal{X}}^k := \overline{\mathcal{X}}_1^k$, $\overline{\mathcal{X}}_j := \overline{\mathcal{X}}_j^N$.

3. Hellwig's $\mathcal{W}_p$-information topology is equal to the topology induced by $\mathcal{AW}_p$

In this section we show (3) = (1) in Theorem 1.2. We will do so by identifying both topologies as initial topologies w.r.t. a single map each, i.e. finding a space which is homeomorphic to $\mathcal{P}(\overline{\mathcal{X}})$ with Hellwig's ($\mathcal{W}_p$-)information topology and one which is homeomorphic to $\mathcal{P}(\overline{\mathcal{X}})$ with the topology induced by $\mathcal{AW}_p$ and then showing that these spaces are homeomorphic in the right way. As an auxilliary tool we will introduce another topology on $\mathcal{P}(\overline{\mathcal{X}})$ which wasn't mentioned in the introduction, but which is very similar to Hellwig's. The proof strategy can be summarized by saying that we want to show that the following diagram is commutative.



(11)

Here $\mathcal{N}$ is the map which induces the same topology as $\mathcal{AW}_p$, $\mathcal{I}$ induces Hellwig's topology and $\mathcal{I}'$ induces what we will call the *reduced* information topology. We shortly restate their definitions below. The maps $\mathcal{K}$, $\mathcal{M}$, $\mathcal{H}$ are still to be found.

As introduced in Section 1.3 Hellwig's ($\mathcal{W}_p$-)information topology is induced by a family of maps $\mathcal{I}_t$, given by:

$$\mathcal{I}_t : \mathcal{P}(\overline{\mathcal{X}}) \to \mathcal{F}\left(\overline{\mathcal{X}}^t \rightsquigarrow \mathcal{P}(\overline{\mathcal{X}}_{t+1})\right)$$
$$\mathcal{I}_t := \mathrm{dis}_{\overline{\mathcal{X}}^t}^{\overline{\mathcal{X}}_{t+1}} \ .$$

Equivalently, the information topology is the initial topology w.r.t. the map

$$\mathcal{I} : \mathcal{P}(\overline{\mathcal{X}}) \to \prod_{t=1}^{N-1} \mathcal{F}\left(\overline{\mathcal{X}}^t \rightsquigarrow \mathcal{P}(\overline{\mathcal{X}}_{t+1})\right)$$
$$\mathcal{I}(\mu) := (\mathcal{I}_t(\mu))_t \ .$$

We saw in Section 1.8 that $\mathcal{AW}_p$ is induced by an embedding $\mathcal{N} : \mathcal{P}(\overline{\mathcal{X}}) \to \mathcal{P}(\mathcal{X}_{1:N})$. Rephrasing the definition there, $\mathcal{N}$ is obtained by defining recursively from $t = N-1$ to $t = 1$:

$$\mathcal{N}^N := 1_{\mathcal{P}(\overline{\mathcal{X}})}$$
$$\mathcal{N}^t := \mathrm{dis}_{\overline{\mathcal{X}}^t}^{\mathcal{X}_{t+1:N}} \circ \mathcal{N}^{t+1}$$

and setting

$$\mathcal{N} := \mathcal{N}^1 \ .$$

In fact, because dis maps into the space of measures concentrated on the graph of a function, $\mathcal{N}$ also maps into a smaller space, which we call $\mathcal{F}_1$, and which is again defined by recursion down from $N-1$ to 1:

$$\begin{aligned} \mathcal{F}_N &:= \mathcal{P}(\mathcal{X}_N) \\ \mathcal{F}_t &:= \mathcal{F}\left(\mathcal{X}_t \rightsquigarrow \mathcal{F}_{t+1}\right) \ . \end{aligned}$$

I.e. $\mathcal{F}_1$ is $\mathcal{P}(\mathcal{X}_{1:N})$ with all occurences of $\mathcal{P}(\cdot \times \cdot)$ replaced by $\mathcal{F}\left(\cdot \rightsquigarrow \cdot\right)$. Remember that we had

$$\begin{aligned} \mathcal{X}_{N:N} &:= \mathcal{X}_N \\ \mathcal{X}_{t:N} &:= \mathcal{X}_t \times \mathcal{P}(\mathcal{X}_{t+1:N}) \ . \end{aligned}$$

For convenience, let us also define

$$\mathcal{P}_t := \mathcal{P}(\mathcal{X}_{t:N}) \ .$$

The fact that

$$\mathcal{N}^t : \mathcal{P}\big(\overline{\mathcal{X}}\big) \to \mathcal{F}\left(\overline{\mathcal{X}}^t \rightsquigarrow \mathcal{F}_{t+1}\right)$$

and that therefore $\mathcal{N}$ maps into $\mathcal{F}_1$ is a consequence of Lemma 3.1 below.

Finally, $\mathcal{I}'$ is defined as follows

$$\begin{aligned} \mathcal{I}' : \mathcal{P}\big(\overline{\mathcal{X}}\big) &\to \prod_{t=1}^{N-1} \mathcal{F}\left(\overline{\mathcal{X}}^t \rightsquigarrow \mathcal{P}(\mathcal{X}_{t+1})\right) \\ \mathcal{I}'(\mu) &:= (\mathcal{I}'_t(\mu))_t \\ \mathcal{I}'_t : \mathcal{P}\big(\overline{\mathcal{X}}\big) &\to \mathcal{F}\left(\overline{\mathcal{X}}^t \rightsquigarrow \mathcal{P}(\mathcal{X}_{t+1})\right) \\ \mathcal{I}'_t &:= \mathrm{dis}^{\mathcal{X}_{t+1}}_{\overline{\mathcal{X}}^t} \circ \mathcal{P}\big(\mathrm{proj}_{\overline{\mathcal{X}}^{t+1}}\big) \ . \end{aligned}$$

I.e. the reduced information topology, like the information topology, makes continuous predictions about the behaviour of the process after time t given information about its behaviour up to time t, only now we are just predicting what the process will do in the next step, not for the rest of time.

$\mathcal{I}$, $\mathcal{I}'$ and $\mathcal{N}$ are injective and therefore bijections onto their codomains. This means that the values of the maps $\mathcal{K}$, $\mathcal{M}$, $\mathcal{H}$ in diagram (11) as functions between sets are really already prescribed. The task consists in finding a representation for them which makes it clear that they are continuous.

Lemma 3.1. $\mathrm{dis}^{\mathcal{B}\times\mathcal{Y}}_{\mathcal{A}}$ *restricted to* $\mathcal{F}\left(\mathcal{A}\times\mathcal{B}\rightsquigarrow\mathcal{Y}\right)$ *maps onto* $\mathcal{F}\big(\mathcal{A}\rightsquigarrow\mathcal{F}\left(\mathcal{B}\rightsquigarrow\mathcal{Y}\right)\big)$.

Proof. We first show that it maps into $\mathcal{F}\big(\mathcal{A}\rightsquigarrow\mathcal{F}\left(\mathcal{B}\rightsquigarrow\mathcal{Y}\right)\big)$. Let $\nu\in\mathcal{F}\big(\mathcal{A}\times\mathcal{B}\rightsquigarrow\mathcal{Y}\big)$ and let $g:\mathcal{A}\times\mathcal{B}\to\mathcal{Y}$ be a function witnessing this fact, i.e. $\nu(f) = \int f(a,b,g(a,b))\,\mathrm{d}\nu(a,b,\_)$.

Let $\alpha := \mathrm{dis}^{\mathcal{B}\times\mathcal{Y}}_{\mathcal{A}}(\nu)$. Then

$$\int\int 1_{g(a,b)\neq y}\,\mathrm{d}\beta(b,y)\,\mathrm{d}\alpha(a,\beta) = \int 1_{g(a,b)\neq y}\,\mathrm{d}\nu(a,b,y) = 0 \ .$$

This means that for α-a.a. (a,β) we have $\int 1_{g(a,b)\neq y}\,\mathrm{d}\beta(b,y)=0$, i.e. β is concentrated on the graph of the function $b\mapsto g(a,b)$.

To see that any $\alpha\in\mathcal{F}\big(\mathcal{A}\rightsquigarrow\mathcal{F}\left(\mathcal{B}\rightsquigarrow\mathcal{Y}\right)\big)$ can be obtained as the image of some $\nu\in\mathcal{F}\left(\mathcal{A}\times\mathcal{B}\rightsquigarrow\mathcal{Y}\right)$ under $\mathrm{dis}^{\mathcal{B}\times\mathcal{Y}}_{\mathcal{A}}$, note that for such α, by the existence of measurably dependent (classical) disintegrations (see for example [14, Proposition 7.27]), $\nu := \mathrm{int}^{\mathcal{B}\times\mathcal{Y}}_{\mathcal{A}}(\alpha)\in\mathcal{F}\left(\mathcal{A}\times\mathcal{B}\rightsquigarrow\mathcal{Y}\right)$, and $\mathrm{dis}^{\mathcal{B}\times\mathcal{Y}}_{\mathcal{A}}(\nu)=\alpha$. □

3.1. **Homeomorphisms.** We give a plain language description of what follows in this section:

The continuity of $\mathcal{M}$ will be quite trivial, because we are just discarding information.

The components $\mathcal{K}_k : \mathcal{F}_1 \to \mathcal{F}\left(\overline{\mathcal{X}}^k \rightsquigarrow \mathcal{P}\big(\overline{\mathcal{X}}_{k+1}\big)\right)$ of the map $\mathcal{K}$ are obtained by 'folding' both the 'head' and the 'tail' of $\mathcal{F}_1$ using iterated application of the map int.

$$\overbrace{\mathcal{F}\Bigg(\mathcal{X}_1 \rightsquigarrow \mathcal{F}\Bigg(\cdots \rightsquigarrow \mathcal{F}\Big(\mathcal{X}_k \rightsquigarrow \mathcal{F}\big(\mathcal{X}_{k+1} \rightsquigarrow \mathcal{F}(\cdots \rightsquigarrow \mathcal{P}(\mathcal{X}_N)\dots)\big)\Big)\dots\Bigg)\Bigg)}^{\text{head}} \overbrace{\qquad\qquad}^{\text{tail}}$$

By continuity of int, it's easy to see that $\mathcal{K}_k$ is continuous. To show that the map $\mathcal{K}$ with the components $\mathcal{K}_k$ is the map we are looking for, we basically show that

$$\mathcal{I}^{-1} \circ \mathcal{K}_k = \mathcal{N}^{-1} \ . \tag{12}$$

$\mathcal{N}^{-1}$ is again another way of 'folding' all of $\mathcal{F}_1$ using int to arrive at $\mathcal{P}\big(\overline{\mathcal{X}}\big)$. As $\mathcal{I}^{-1}$ is also int, showing (12) amounts to showing that these two different ways of 'folding' – first the head and tail and then in a last step the junction between k and $k+1$ on the one hand, and from front to back on the other hand – do the same thing. This may be intuitively clear to the reader. The proof works by repeated application of Lemma 3.5, which represents one step of 'folding order doesn't matter'. Using Lemma 3.5 the proof is completely analogous to the proof that for an operation $\star$ satisfying $(a \star b) \star c = a \star (b \star c)$, i.e. for an associative operation, one has

$$\begin{aligned}\big(\dots((x_1 \star x_2) \star x_3) \star \dots) \star x_k\big) \star \big(\dots((x_{k+1} \star x_{k+2}) \star x_{k+3}) \star \dots) \star x_N\big) \\ = \big(\dots((x_1 \star x_2) \star x_3) \star \dots) \star x_N\big) \ .\end{aligned}$$

As we know, for such an operation any way of parenthesizing the multiplication of N elements gives the same result. An analogous statement holds for int, though we do not formally state or prove this.

Finally, in Lemma 3.9, using Lemma 3.8 as the main ingredient we prove the 'hard direction', i.e. that $\mathcal{H}$ is continuous. If the continuity of $\mathcal{M}$ and $\mathcal{K}$ as informally described here seem obvious to the reader they may wish to skip ahead to Lemma 3.8 and Lemma 3.9.

Remark 3.2. The reader interested in working out the details and analogies between 'folding' using int and associative binary operations might be interested in reading about *monads* in the context of Category Theory first. (See for example Chapter VI in [45].) In fact, $(\mathcal{P}, \boldsymbol{\eta}, \boldsymbol{\mu})$ forms a monad, where

$$\boldsymbol{\eta}_{\mathcal{X}} : \mathcal{X} \to \mathcal{P}(\mathcal{X})$$

sends an element x of $\mathcal{X}$ to the dirac measure at x and

$$\begin{aligned}\boldsymbol{\mu}_{\mathcal{X}} &: \mathcal{P}(\mathcal{P}(\mathcal{X})) \to \mathcal{P}(\mathcal{X}) \\ \boldsymbol{\mu}_{\mathcal{X}}(\nu) &:= f \mapsto \iint f(x)\, \mathrm{d}\nu'(x)\, \mathrm{d}\nu(\nu') \ .\end{aligned}$$

This monad is studied in a little more detail in [27]. int can be obtained from $\boldsymbol{\mu}$ and a *tensorial strength* $t_{\mathcal{A},\mathcal{B}} : \mathcal{A} \times \mathcal{P}(\mathcal{B}) \to \mathcal{P}(\mathcal{A} \times \mathcal{B})$ in the sense described for example in [47].

To show that $\mathcal{M}$ is continuous we will need the following lemma.

Lemma 3.3. $\mathrm{dis}_{\mathcal{A}}^{\mathcal{B}}$ *is natural in* $\mathcal{B}$*, i.e. for* $f : \mathcal{B} \to \mathcal{B}'$ *the following diagram commutes.*

$$\begin{array}{ccc} \mathcal{P}(\mathcal{A} \times \mathcal{B}') & \xleftarrow{\quad \mathcal{P}(1_{\mathcal{A}} \times f) \quad} & \mathcal{P}(\mathcal{A} \times \mathcal{B}) \\ \Big\downarrow{\scriptstyle \mathrm{dis}_{\mathcal{A}}^{\mathcal{B}'}} & & \Big\downarrow{\scriptstyle \mathrm{dis}_{\mathcal{A}}^{\mathcal{B}}} \\ \mathcal{F}\left(\mathcal{A} \rightsquigarrow \mathcal{P}(\mathcal{B}')\right) & \xleftarrow[\quad \mathcal{P}(1_{\mathcal{A}} \times \mathcal{P}(f)) \quad]{} & \mathcal{F}\left(\mathcal{A} \rightsquigarrow \mathcal{P}(\mathcal{B})\right) \end{array}$$

Proof. This is just straigtforward calculation using the definitions. □

Applying Lemma 3.3 with $\mathcal{A} = \overline{\mathcal{X}}^k$, $\mathcal{B} = \overline{\mathcal{X}}_{k+1}$, $\mathcal{B}' = \mathcal{X}_{k+1}$ and $f = \mathrm{proj}_{\mathcal{X}_{k+1}} : \overline{\mathcal{X}}_{k+1} \to \mathcal{X}_{k+1}$ we get that

$$\mathcal{I}'_k = \mathrm{dis}_{\overline{\mathcal{X}}^k}^{\mathcal{X}_{k+1}} \circ \mathcal{P}\Big(1_{\overline{\mathcal{X}}^k} \times \mathrm{proj}_{\mathcal{X}_{k+1}}\Big) = \mathcal{P}\Big(1_{\overline{\mathcal{X}}^k} \times \mathcal{P}\Big(\mathrm{proj}_{\mathcal{X}_{k+1}}\Big)\Big) \circ \mathrm{dis}_{\overline{\mathcal{X}}^k}^{\overline{\mathcal{X}}_{k+1}}$$

Setting $\mathcal{M}_k := \mathcal{P}\Big(1_{\overline{\mathcal{X}}^k} \times \mathcal{P}\Big(\mathrm{proj}_{\mathcal{X}_{k+1}}\Big)\Big)$ we get $\mathcal{I}'_k = \mathcal{M}_k \circ \mathcal{I}_k$ and then setting $\mathcal{M}((\mu_k)_k) := (\mathcal{M}_k(\mu_k))_k$ gives $\mathcal{I}' = \mathcal{M} \circ \mathcal{I}$.

There is an analogue of Lemma 3.3 which we list here for completeness.

Lemma 3.4. $\mathrm{int}_{\mathcal{A}}^{\mathcal{B}} : \mathcal{P}(\mathcal{A} \times \mathcal{P}(\mathcal{B})) \to \mathcal{P}(\mathcal{A} \times \mathcal{B})$ *is natural in* $\mathcal{B}$*, i.e. for* $f : \mathcal{B} \to \mathcal{B}'$ *the following diagram commutes:*

$$\begin{array}{ccc} \mathcal{P}(\mathcal{A} \times \mathcal{P}(\mathcal{B}')) & \xleftarrow{\quad \mathcal{P}(1_{\mathcal{A}} \times \mathcal{P}(f)) \quad} & \mathcal{P}(\mathcal{A} \times \mathcal{P}(\mathcal{B})) \\ \Big\downarrow{\scriptstyle \mathrm{int}_{\mathcal{A}}^{\mathcal{B}'}} & & \Big\downarrow{\scriptstyle \mathrm{int}_{\mathcal{A}}^{\mathcal{B}}} \\ \mathcal{P}(\mathcal{A} \times \mathcal{B}') & \xleftarrow[\quad \mathcal{P}(1_{\mathcal{A}} \times f) \quad]{} & \mathcal{P}(\mathcal{A} \times \mathcal{B}) \end{array}$$

In particular, if $\mathcal{B} \subseteq \mathcal{B}'$ *then*

$$\mathrm{int}_{\mathcal{A}\ \restriction \mathcal{P}(\mathcal{A} \times \mathcal{P}(\mathcal{B}))}^{\mathcal{B}'} = \mathrm{int}_{\mathcal{A}}^{\mathcal{B}}$$

if we regard $\mathcal{P}(\mathcal{A} \times \mathcal{P}(\mathcal{B}))$ *as a subset of* $\mathcal{P}(\mathcal{A} \times \mathcal{P}(\mathcal{B}'))$ *by recursively using the recipe: 'if* $\mathcal{B}$ *is a subset of* $\mathcal{B}'$*, then we can view* $\mathcal{P}(\mathcal{B})$ *as the subset of those* $\mu \in \mathcal{P}(\mathcal{B}')$ *which are concentrated on* $\mathcal{B}$*'.*

Proof. Again this is just calculation. □

We already implicity used the 'in particular'-part of Lemma 3.4 when we said that $\mathcal{N}$ can be regarded both as a map into $\mathcal{P}(\mathcal{X}_{1:N})$ and into $\mathcal{F}_1$ but the use there seemed too trivial to warrant much mention. There will be more such tacit uses.

Now we show that $\mathcal{K}$ is continuous. We claim that it can be written as

$$\mathcal{K}(\mu) = (\mathcal{K}_k(\mu))_k$$

where

$$\mathcal{K}_k = \mathcal{P}\Big(1_{\overline{\mathcal{X}}^k} \times \big(\mathrm{int}_{\overline{\mathcal{X}}_{k+1}^{N-1}}^{\mathcal{X}_{N:N}} \circ \cdots \circ \mathrm{int}_{\overline{\mathcal{X}}_{k+1}^{k+2}}^{\mathcal{X}_{k+3:N}} \circ \mathrm{int}_{\mathcal{X}_{k+1}}^{\mathcal{X}_{k+2:N}}\big)\Big) \circ \mathrm{int}_{\overline{\mathcal{X}}^{k-1}}^{\mathcal{X}_{k:N}} \circ \cdots \circ \mathrm{int}_{\overline{\mathcal{X}}^2}^{\mathcal{X}_{3:N}} \circ \mathrm{int}_{\overline{\mathcal{X}}^1}^{\mathcal{X}_{2:N}} \quad ,$$

or without the dots, letting $\fatbslash$ denote concatenation of functions, e.g. $\fatbslash_{i=3}^{1} f_i = f_3 \circ f_2 \circ f_1$:

$$\mathcal{K}_k = \mathcal{P}\Big(1_{\overline{\mathcal{X}}^k} \times \big(\fatbslash_{i=N-1}^{k+1} \operatorname{int}^{\mathcal{X}_{i+1:N}}_{\overline{\mathcal{X}}^i_{k+1}}\big)\Big) \circ \fatbslash_{i=k-1}^{1} \operatorname{int}^{\mathcal{X}_{i+1:N}}_{\overline{\mathcal{X}}^i} \ .$$

To prove this we will repeatedly apply the following lemma.

Lemma 3.5 (int is 'associative')**.** int *satisfies the following relation:*

$$\operatorname{int}^{\mathcal{C}}_{\mathcal{A}\times\mathcal{B}} \circ \operatorname{int}^{\mathcal{B}\times\mathcal{P}(\mathcal{C})}_{\mathcal{A}} = \operatorname{int}^{\mathcal{B}\times\mathcal{C}}_{\mathcal{A}} \circ \mathcal{P}\big(1_{\mathcal{A}} \times \operatorname{int}^{\mathcal{C}}_{\mathcal{B}}\big)$$

These maps can be seen in the following commutative diagram.

$$\begin{array}{ccc} \mathcal{P}(\mathcal{A}\times\mathcal{B}\times\mathcal{P}(\mathcal{C})) & \xleftarrow{\operatorname{int}^{\mathcal{B}\times\mathcal{P}(\mathcal{C})}_{\mathcal{A}}} & \mathcal{P}(\mathcal{A}\times\mathcal{P}(\mathcal{B}\times\mathcal{P}(\mathcal{C}))) \\ \Big\downarrow{\scriptstyle \operatorname{int}^{\mathcal{C}}_{\mathcal{A}\times\mathcal{B}}} & & \Big\downarrow{\scriptstyle \mathcal{P}(1_{\mathcal{A}}\times\operatorname{int}^{\mathcal{C}}_{\mathcal{B}})} \\ \mathcal{P}(\mathcal{A}\times\mathcal{B}\times\mathcal{C}) & \xleftarrow[\operatorname{int}^{\mathcal{B}\times\mathcal{C}}_{\mathcal{A}}]{} & \mathcal{P}(\mathcal{A}\times\mathcal{P}(\mathcal{B}\times\mathcal{C})) \end{array}$$

Proof. This is just expanding the definition. Both maps send a measure $\alpha \in \mathcal{P}(\mathcal{A}\times\mathcal{P}(\mathcal{B}\times\mathcal{P}(\mathcal{C})))$ to the measure μ with

$$\int f \,\mathrm{d}\mu = \int f(a,b,c)\,\mathrm{d}\gamma(c)\,\mathrm{d}\beta(b,\gamma)\,\mathrm{d}\alpha(a,\beta) \ .$$

□

Lemma 3.6. *The following relation holds.*

$$\operatorname{int}^{\overline{\mathcal{X}}_{k+1}}_{\overline{\mathcal{X}}^k} \circ \mathcal{K}_k = \fatbslash_{i=N-1}^{1} \operatorname{int}^{\mathcal{X}_{i+1:N}}_{\overline{\mathcal{X}}^i} \tag{13}$$

Proof. Again, this is just repeated application of Lemma 3.5. Below we define $\mathcal{T}_l$ for $N \geq l \geq k$ and show that

$$\operatorname{int}^{\overline{\mathcal{X}}_{k+1}}_{\overline{\mathcal{X}}^k} \circ \fatbslash_{i=N-1}^{k+1} \mathcal{P}\left(1_{\overline{\mathcal{X}}^k} \times \operatorname{int}^{\mathcal{X}_{i+1:N}}_{\overline{\mathcal{X}}^i_{k+1}}\right) = \mathcal{T}_l \tag{14}$$

for all $N \geq l \geq k$ by showing $\mathcal{T}_l = \mathcal{T}_{l-1}$ for all $N \geq l > k$. The left hand side of (14) is the left hand side of (13) with the common tail $\fatbslash_{i=k-1}^{1} \operatorname{int}^{\mathcal{X}_{i+1:N}}_{\overline{\mathcal{X}}^i}$ of the left and right side in (13) dropped. $\mathcal{T}_k$ will be the right hand side of (13) with the common part dropped.

$$\mathcal{T}_l := \fatbslash_{i=N-1}^{l} \operatorname{int}^{\mathcal{X}_{i+1:N}}_{\overline{\mathcal{X}}^i} \quad \circ \operatorname{int}^{\overline{\mathcal{X}}^l_{k+1}\times\mathcal{P}_{l+1}}_{\overline{\mathcal{X}}^k} \circ \quad \fatbslash_{i=l-1}^{k+1} \mathcal{P}\left(1_{\overline{\mathcal{X}}^k} \times \operatorname{int}^{\mathcal{X}_{i+1:N}}_{\overline{\mathcal{X}}^i_{k+1}}\right)$$

Here we regard $\fatbslash_r^s \ldots$ with $r < s$ (an empty product in our context) as the identity function. For $l = N$ the first factor is an empty product and therefore clearly (14) is true for $l = N$. To get from $\mathcal{T}_l$ to $\mathcal{T}_{l-1}$ we leave the first factor alone and apply Lemma 3.5 with $\mathcal{A} = \overline{\mathcal{X}}^k$, $\mathcal{B} = \overline{\mathcal{X}}^{l-1}_{k+1}$ and $\mathcal{C} = \mathcal{X}_{l:N}$. This transforms

$$\operatorname{int}^{\overline{\mathcal{X}}^l_{k+1}\times\mathcal{P}_{l+1}}_{\overline{\mathcal{X}}^k} \circ \mathcal{P}\left(1_{\overline{\mathcal{X}}^k} \times \operatorname{int}^{\mathcal{X}_{l:N}}_{\overline{\mathcal{X}}^{l-1}_{k+1}}\right)$$

into

$$\operatorname{int}^{\mathcal{X}_{l:N}}_{\overline{\mathcal{X}}^{l-1}} \circ \operatorname{int}^{\overline{\mathcal{X}}^{l-1}_{k+1}\times\mathcal{P}_l}_{\overline{\mathcal{X}}^k}$$

and therefore $\mathcal{T}_l$ into $\mathcal{T}_{l-1}$. □

Lemma 3.7. *The right hand triangle in* (11) *commutes, i.e.*

$$\mathcal{K}_k \circ \mathcal{N} = \mathcal{I}_k \ .$$

Proof. Prepending $\mathcal{N}$ to (13) gives

$$\mathrm{int}^{\overline{\mathcal{X}}_{k+1}}_{\overline{\mathcal{X}}^k{}_{\restriction \mathcal{F}\left(\overline{\mathcal{X}}^k \rightsquigarrow \mathcal{P}\left(\overline{\mathcal{X}}_{k+1}\right)\right)}} \circ \mathcal{K}_k \circ \mathcal{N} = 1_{\mathcal{P}\left(\overline{\mathcal{X}}\right)}$$

and appending $\mathcal{I}_k$ gives

$$\mathcal{K}_k \circ \mathcal{N} = \mathcal{I}_k \ .$$

□

Now we will show that $\mathcal{H}$ is continuous. We will postpone the proof of Lemma 3.8 below, which is the crucial non-bookkeeping ingredient in the proof of Lemma 3.9 below, until the end of this section. The methods used in the proof of Lemma 3.8 differ significantly from the rest in this section and make use of the concept of the *modulus of continuity* for measures, and results relating to it, introduced in the companion paper [23] to this one.

Lemma 3.8. *Let*

$$\mathrm{dom}\Big(\mathcal{J}^{\mathcal{Y}}_{\mathcal{A},\mathcal{B}}\Big) \subseteq \mathcal{F}\left(\mathcal{A} \rightsquigarrow \mathcal{P}(\mathcal{B})\right) \times \mathcal{F}\left(\mathcal{A} \times \mathcal{B} \rightsquigarrow \mathcal{Y}\right)$$

be the set of all (μ', μ) *s.t.*

$$\mathrm{int}^{\mathcal{B}}_{\mathcal{A}}(\mu') = \mu_{\restriction \mathcal{A} \times \mathcal{B}} \ . \tag{15}$$

The function

$$\mathcal{J}^{\mathcal{Y}}_{\mathcal{A},\mathcal{B}} : \mathrm{dom}\Big(\mathcal{J}^{\mathcal{Y}}_{\mathcal{A},\mathcal{B}}\Big) \to \mathcal{F}\left(\mathcal{A} \rightsquigarrow \mathcal{F}\left(\mathcal{B} \rightsquigarrow \mathcal{Y}\right)\right)$$
$$\mathcal{J}^{\mathcal{Y}}_{\mathcal{A},\mathcal{B}}(\mu', \mu) := \mathrm{dis}^{\mathcal{B} \times \mathcal{Y}}_{\mathcal{A}}(\mu)$$

is continuous.

Clearly, as a function between sets, $\mathcal{J}^{\mathcal{Y}}_{\mathcal{A},\mathcal{B}}(\mu', \mu)$ only depends on μ. But, as we know, $\mathrm{dis}^{\mathcal{B} \times \mathcal{Y}}_{\mathcal{A}}$ is *not* continuous. Only when we refine the topology on the source space, which we encode by regarding $\mathcal{J}^{\mathcal{Y}}_{\mathcal{A},\mathcal{B}}$ as a map from the above subset of a product space, does it become continuous.

Lemma 3.9. $\mathcal{H}$ *is continuous.*

Proof. We will inductively define

$$\mathcal{H}^k : \mathcal{I}'\left[\mathcal{P}\left(\overline{\mathcal{X}}\right)\right] \to \mathcal{P}\left(\overline{\mathcal{X}}^k \times \mathcal{P}_{k+1}\right)$$

(again down from $N-1$ to 1) so that they will be continuous by construction (and by virtue of Lemma 3.8). Also by construction, we will have $\mathcal{H}^k \circ \mathcal{I}' = \mathcal{N}^k$. $\mathcal{H}$ will be $\mathcal{H}^1$ so that $\mathcal{H} \circ \mathcal{I}' = \mathcal{N}$.

Set $\mathcal{H}^{N-1} := \mathrm{proj}_{N-1}$, the projection from $\prod_{k=1}^{N-1} \mathcal{F}\left(\overline{\mathcal{X}}^k \rightsquigarrow \mathcal{P}(\mathcal{X}_{k+1})\right)$ onto the last factor. $\mathcal{H}^{N-1} \circ \mathcal{I}' = \mathcal{I}'_{N-1} = \mathrm{dis}^{\mathcal{X}_N}_{\overline{\mathcal{X}}^{N-1}} = \mathcal{N}^{N-1}$ by definition. Given $\mathcal{H}^{k+1}$ define

$$\mathcal{H}^k(\mu) := \mathcal{J}^{\mathcal{F}_{k+2}}_{\overline{\mathcal{X}}^k, \mathcal{X}_{k+1}}\left(\mathrm{proj}_k(\mu), \mathcal{H}^{k+1}(\mu)\right) \ ,$$

where proj_k is the projection from $\prod_{k=1}^{N-1} \mathcal{F}\left(\overline{\mathcal{X}}^k \rightsquigarrow \mathcal{P}(\mathcal{X}_{k+1})\right)$ onto the k-th factor.

For this to be well-defined we need to check that for $\mu \in \mathcal{I}'\left[\mathcal{P}\left(\overline{\mathcal{X}}\right)\right]$ we have

$$\mathrm{int}^{\mathcal{X}_{k+1}}_{\overline{\mathcal{X}}^k}(\mathrm{proj}_k(\mu)) = \mathcal{P}\big(\mathrm{proj}_{\overline{\mathcal{X}}^{k+1}}\big)\left(\mathcal{H}^{k+1}(\mu)\right) \ .$$

I.e. for $\nu \in \mathcal{P}\left(\overline{\mathcal{X}}\right)$ we want

$$\mathrm{int}^{\mathcal{X}_{k+1}}_{\overline{\mathcal{X}}^k}(\mathrm{proj}_k(\mathcal{I}'(\nu))) = \mathcal{P}\big(\mathrm{proj}_{\overline{\mathcal{X}}^{k+1}}\big)\left(\mathcal{H}^{k+1}(\mathcal{I}'(\nu))\right)$$

The composite of the maps on the left-hand side is equal to

$$\mathrm{int}^{\mathcal{X}_{k+1}}_{\overline{\mathcal{X}}^k} \circ \mathcal{I}'_k = \mathrm{int}^{\mathcal{X}_{k+1}}_{\overline{\mathcal{X}}^k} \circ \mathrm{dis}^{\mathcal{X}_{k+1}}_{\overline{\mathcal{X}}^k} \circ \mathcal{P}\big(\mathrm{proj}_{\overline{\mathcal{X}}^{k+1}}\big) = \mathcal{P}\big(\mathrm{proj}_{\overline{\mathcal{X}}^{k+1}}\big) \ .$$

On the right-hand side we get by induction hypothesis

$$\mathcal{P}\big(\mathrm{proj}_{\overline{\mathcal{X}}^{k+1}}\big) \circ \mathcal{N}^{k+1} \ . \tag{16}$$

Using that $\mathcal{P}(\mathrm{proj}_{\mathcal{A}}) \circ \mathrm{dis}_{\mathcal{A}}^{\mathcal{B}} = \mathcal{P}(\mathrm{proj}_{\mathcal{A}})$ we see for $l \geq k+1$

$$\begin{aligned}\mathcal{P}\big(\mathrm{proj}_{\overline{\mathcal{X}}^{k+1}}\big) \circ \mathcal{P}\big(\mathrm{proj}_{\overline{\mathcal{X}}^{l}}\big) \circ \mathcal{N}^{l} = \\ \mathcal{P}\big(\mathrm{proj}_{\overline{\mathcal{X}}^{k+1}}\big) \circ \mathcal{P}\big(\mathrm{proj}_{\overline{\mathcal{X}}^{l}}\big) \circ \mathrm{dis}_{\overline{\mathcal{X}}^{l}}^{\mathcal{X}_{l+1:N}} \circ \mathcal{N}^{l+1} = \\ \mathcal{P}\big(\mathrm{proj}_{\overline{\mathcal{X}}^{k+1}}\big) \circ \mathcal{P}\big(\mathrm{proj}_{\overline{\mathcal{X}}^{l}}\big) \circ \mathcal{N}^{l+1} = \\ \mathcal{P}\big(\mathrm{proj}_{\overline{\mathcal{X}}^{k+1}}\big) \circ \mathcal{P}\big(\mathrm{proj}_{\overline{\mathcal{X}}^{l+1}}\big) \circ \mathcal{N}^{l+1} \ ,\end{aligned}$$

i.e. by induction (16) is also equal to $\mathcal{P}\big(\mathrm{proj}_{\overline{\mathcal{X}}^{k+1}}\big)$.

As a composite of continuous maps $\mathcal{H}^k$ is clearly continuous. (This is where we use Lemma 3.8.) As a map between sets $\mathcal{H}^k$ is just

$$\mathrm{dis}_{\overline{\mathcal{X}}^{k}}^{\mathcal{X}_{k+1:N}} \circ \mathcal{H}^{k+1} = \mathrm{dis}_{\overline{\mathcal{X}}^{k}}^{\mathcal{X}_{k+1:N}} \circ \mathcal{N}^{k+1} = \mathcal{N}^{k}$$

by induction hypothesis and definition of $\mathcal{N}^k$. □

3.2. **Proof of Lemma 3.8.** In this part we prove Lemma 3.8. Here we use several of the ideas developed in the companion paper [23]. In particular we will need [23, Lemma 4.2] which we reproduce below.

Lemma 3.10 ([23, Lemma 4.2])**.** *Let $\mu \in \mathcal{F}(\mathcal{X} \rightsquigarrow \mathcal{Y})$. For any $\varepsilon > 0$ there is a $\delta > 0$ s.t. if*

$$\begin{aligned}&\nu \in \mathcal{P}(\mathcal{X} \times \mathcal{Y}) \text{ with } \mathcal{W}_p(\mu, \nu) < \delta \text{ and} \\ &\gamma \in \mathrm{Cpl}(\mu, \nu) \text{ with } \textstyle\int \rho(x_1, x_2)^p \, \mathrm{d}\gamma(x_1, y_1, x_2, y_2) < \delta^p\end{aligned}$$

then

$$\textstyle\int \rho(y_1, y_2)^p \, \mathrm{d}\gamma(x_1, y_1, x_2, y_2) < \varepsilon^p \ .$$

For easy reference we also restate Lemma 3.8.

Lemma 3.8. *Let*

$$\mathrm{dom}\Big(\mathcal{J}_{\mathcal{A},\mathcal{B}}^{\mathcal{Y}}\Big) \subseteq \mathcal{F}\big(\mathcal{A} \rightsquigarrow \mathcal{P}(\mathcal{B})\big) \times \mathcal{F}\big(\mathcal{A} \times \mathcal{B} \rightsquigarrow \mathcal{Y}\big)$$

be the set of all (μ', μ) s.t.

$$\mathrm{int}_{\mathcal{A}}^{\mathcal{B}}(\mu') = \mu_{\restriction \mathcal{A} \times \mathcal{B}} \ . \tag{15}$$

The function

$$\begin{aligned}&\mathcal{J}_{\mathcal{A},\mathcal{B}}^{\mathcal{Y}} : \mathrm{dom}\Big(\mathcal{J}_{\mathcal{A},\mathcal{B}}^{\mathcal{Y}}\Big) \to \mathcal{F}\big(\mathcal{A} \rightsquigarrow \mathcal{F}(\mathcal{B} \rightsquigarrow \mathcal{Y})\big) \\ &\mathcal{J}_{\mathcal{A},\mathcal{B}}^{\mathcal{Y}}(\mu', \mu) := \mathrm{dis}_{\mathcal{A}}^{\mathcal{B} \times \mathcal{Y}}(\mu)\end{aligned}$$

is continuous.

Proof of Lemma 3.8. Let $(\mu', \mu) \in \mathrm{dom}(\mathcal{J}_{\mathcal{A},\mathcal{B}}^{\mathcal{Y}})$. Let $\varepsilon > 0$.

Choose $\delta > 0$ according to Lemma 3.10 with $\mathcal{X} = \mathcal{A} \times \mathcal{B}$, i.e. s.t. for any $\nu \in \mathcal{P}(\mathcal{A} \times \mathcal{B} \times \mathcal{Y})$ with $\mathcal{W}_p(\mu, \nu) < \delta$ and any $\gamma \in \mathrm{Cpl}(\mu, \nu)$ with $\int \rho(a_1, a_2)^p + \rho(b_1, b_2)^p \, \mathrm{d}\gamma(a_1, b_1, \_, a_2, b_2, \_) < \delta^p$ we have $\int \rho(y_1, y_2)^p \, \mathrm{d}\gamma(\_, y_1, \_, y_2) < \varepsilon^p$.

Let $(\nu', \nu) \in \mathrm{dom}(\mathcal{J}_{\mathcal{A},\mathcal{B}}^{\mathcal{Y}})$ with $\max(\rho(\mu, \nu), \rho(\mu', \nu')) < \min(\delta, \varepsilon)$.

This means we can find $\gamma' \in \mathrm{Cpl}(\mu', \nu')$ with

$$\textstyle\int \rho(a_1, a_2)^p + \mathcal{W}_p(\hat{b}_1, \hat{b}_2)^p \, \mathrm{d}\gamma'(a_1, \hat{b}_1, a_2, \hat{b}_2) < \min(\delta^p, \varepsilon^p) \ . \tag{17}$$

Let $(a, b) \mapsto f_a(b)$ and $(a, b) \mapsto g_a(b) : \mathcal{A} \times \mathcal{B} \to \mathcal{Y}$ be measurable functions on whose graph μ and ν, respectively, are concentrated. Let $\bar{\mu} := \mathcal{J}_{\mathcal{A},\mathcal{B}}^{\mathcal{Y}}(\mu', \mu)$, $\bar{\nu} := \mathcal{J}_{\mathcal{A},\mathcal{B}}^{\mathcal{Y}}(\nu', \nu)$.

As noted in the proof of Lemma 3.1 we know that for $\bar{\mu}$-a.a. $(a, \dot{\mu})$ the measure $\dot{\mu}$ is concentrated on the graph of the function f_a (and similarly for $\bar{\nu}$). This together with $\mathcal{P}(1_{\mathcal{A}} \times \mathcal{P}(\mathrm{proj}_{\mathcal{B}}))(\bar{\mu}) = \mu'$ (which is a consequence of (15)) implies that

$$\int h \,\mathrm{d}\bar{\mu} = \int h \left(a, \mathcal{P}(1_{\mathcal{B}}, f_a)(\hat{b})\right) \mathrm{d}\mu'(a, \hat{b})$$

(again similarly for $\bar{\nu}$).

From this we see that the measure $\bar{\gamma} \in \mathcal{P}(\mathcal{A} \times \mathcal{F}(\mathcal{B} \rightsquigarrow \mathcal{Y}) \times \mathcal{A} \times \mathcal{F}(\mathcal{B} \rightsquigarrow \mathcal{Y}))$ defined as

$$\int h \,\mathrm{d}\bar{\gamma} := \int h \left(a_1, \mathcal{P}(1_{\mathcal{B}}, f_{a_1})(\hat{b}_1), a_2, \mathcal{P}(1_{\mathcal{B}}, g_{a_2})(\hat{b}_2)\right) \mathrm{d}\gamma'(a_1, \hat{b}_1, a_2, \hat{b}_2)$$

is in $\mathrm{Cpl}(\bar{\mu}, \bar{\nu})$.

We may measurably select almost-witnesses $\hat{\gamma}_{\hat{b}_1, \hat{b}_2} \in \mathrm{Cpl}(\hat{b}_1, \hat{b}_2)$ for the distances $\mathcal{W}_p(\hat{b}_1, \hat{b}_2)$ s.t. building on (17) we get

$$\int \rho(a_1, a_2)^p + \int \rho(b_1, b_2)^p \,\mathrm{d}\hat{\gamma}_{\hat{b}_1, \hat{b}_2}(b_1, b_2) \,\mathrm{d}\gamma'(a_1, \hat{b}_1, a_2, \hat{b}_2) < \min(\delta^p, \varepsilon^p) \,. \tag{18}$$

Now

$$\begin{aligned}
&\rho(\bar{\mu}, \bar{\nu})^p \leq \int \rho^p_{\mathcal{P}(\mathcal{A} \times \mathcal{P}(\mathcal{B} \times \mathcal{Y}))} \,\mathrm{d}\bar{\gamma} \\
&= \int \rho(a_1, a_2)^p + \mathcal{W}_p\left(\mathcal{P}(1_{\mathcal{B}}, f_{a_1})(\hat{b}_1), \mathcal{P}(1_{\mathcal{B}}, g_{a_2})(\hat{b}_2)\right)^p \,\mathrm{d}\gamma'(a_1, \hat{b}_1, a_1, \hat{b}_2) \\
&\leq \int \rho(a_1, a_2)^p + \int \rho(b_1, b_2)^p + \rho\left(f_{a_1}(b_1), g_{a_2}(b_2)\right)^p \,\mathrm{d}\hat{\gamma}_{\hat{b}_1, \hat{b}_2}(b_1, b_2) \,\mathrm{d}\gamma'(a_1, \hat{b}_1, a_2, \hat{b}_2) \\
&= \int \rho(a_1, a_2)^p + \rho(b_1, b_2)^p + \rho(y_1, y_2)^p \,\mathrm{d}\gamma(a_1, b_1, y_1, a_2, b_2, y_2)
\end{aligned} \tag{19}$$

where $\gamma \in \mathrm{Cpl}(\mu, \nu)$ is defined as

$$\int h \,\mathrm{d}\gamma = \iint h\left(a_1, b_1, f_{a_1}(b_1), a_2, b_2, g_{a_2}(b_2)\right) \,\mathrm{d}\hat{\gamma}_{\hat{b}_1, \hat{b}_2}(b_1, b_2) \,\mathrm{d}\gamma'(a_1, \hat{b}_1, a_2, \hat{b}_2) \,.$$

The integral over the first two summands in (19) is less than $\min(\delta^p, \varepsilon^p)$ by (18). By our choice of δ in the beginning this implies that the integral over the last summand is also less than ε^p, so that overall

$$\rho(\bar{\mu}, \bar{\nu})^p < 2\varepsilon^p \,.$$

Es ε was arbitrary this concludes the proof. □

4. The symmetrized causal Wasserstein distance $\mathcal{SCW}_p$

In this section we prove that the topology induced by $\mathcal{SCW}_p$ is sandwiched between Hellwig's $\mathcal{W}_p$-information topology and the topology induced by $\mathcal{AW}_p$, and therefore by what we have already seen in the previous section equal to both of them. Our arguments in this section make explicit use of metrics. The reader who is only interested in the simpler version of our main theorem, Theorem 1.1 may assume that $p = 1$ and that all metrics are bounded.

Remember that for $\mu, \nu \in \mathcal{P}(\overline{\mathcal{X}})$ we have

$$\mathcal{CW}_p(\mu, \nu)^p = \inf_{\substack{\gamma \in \mathrm{Cpl}(\mu, \nu) \\ \gamma \text{ causal}}} \int \sum_{t=1}^{N} \rho(x_t, y_t)^p \,\mathrm{d}\gamma((x_t)_t, (y_t)_t) \tag{20}$$

$$\mathcal{SCW}_p(\mu, \nu) = \max\left(\mathcal{CW}_p(\mu, \nu), \mathcal{CW}_p(\nu, \mu)\right) \tag{21}$$

$$\mathcal{AW}_p(\mu, \nu)^p = \inf_{\substack{\gamma \in \mathrm{Cpl}(\mu, \nu) \\ \gamma \text{ bicausal}}} \int \sum_{t=1}^{N} \rho(x_t, y_t)^p \,\mathrm{d}\gamma((x_t)_t, (y_t)_t) \,. \tag{22}$$

In proving this we will take a slightly roundabout route. First we will focus on the case where $\overline{\mathcal{X}} = \mathcal{X}_1 \times \mathcal{X}_2$ is the product of just two spaces, i.e. where we have only two time points. Moreover, for expositional purposes, let us for the moment

assume that $\mathcal{X}_1$ and $\mathcal{X}_2$ are both compact. Generalizing from this setting will not be very hard.

In the compact, two-time-point case we will show equality of the two topologies in question by extending both to a larger (compact) space and showing equality of the topologies on that larger space.

In more detail:

When there are only two timepoints Hellwig's $\mathcal{W}_p$-information topology and the topology induced by $\mathcal{AW}_p$ trivially coincide. Both are induced by emedding $\mathcal{P}(\mathcal{X}_1 \times \mathcal{X}_2)$ into $\mathcal{P}(\mathcal{X}_1 \times \mathcal{P}(\mathcal{X}_2))$ via $\mathrm{dis}_{\mathcal{X}_1}^{\mathcal{X}_2}$. The latter space carries its standard metric $\rho_{\mathcal{P}(\mathcal{X}_1 \times \mathcal{P}(\mathcal{X}_2))}$, which – as was already established in Theorem 1.4 in Section 1.8 of the introduction – is an extension of $\mathcal{AW}_p$. To highlight this connection, in this section we will also refer to that metric as $\overline{\mathcal{AW}}_p$. As a reminder,

$$\overline{\mathcal{AW}}_p(\mu,\nu)^p = \inf_{\gamma \in \mathrm{Cpl}(\mu,\nu)} \int \rho(x_1,y_1)^p + \mathcal{W}_p(\xi_2,\eta_2)^p \, \mathrm{d}\gamma(x_1,\xi_2,y_1,\eta_2)$$

where $\mathcal{W}_p$ is the normal Wasserstein distance (on $\mathcal{P}(\mathcal{X}_2)$ in this case). We will find an extension $\overline{\mathcal{CW}}_p$ of $\mathcal{CW}_p$ to $\mathcal{P}(\mathcal{X}_1 \times \mathcal{P}(\mathcal{X}_2))$, which still satisfies all properties of a metric except for symmetry and which is dominated by $\overline{\mathcal{AW}}_p$. Symmetrizing this extension gives a metric (which we will call $\overline{\mathcal{SCW}}_p$). The identity function from $\mathcal{P}(\mathcal{X}_1 \times \mathcal{P}(\mathcal{X}_2))$ topologized with $\overline{\mathcal{AW}}_p$ to $\mathcal{P}(\mathcal{X}_1 \times \mathcal{P}(\mathcal{X}_2))$ topologized with $\overline{\mathcal{SCW}}_p$ will then be a continuous bijection from a compact space (this is where we use compactness of $\mathcal{X}_1$, $\mathcal{X}_2$) to a *Hausdorff* space, i.e. a homeomorphism.

The next subsection will be devoted to finding an expression for the extension of $\mathcal{CW}_p$ to $\mathcal{P}(\mathcal{X}_1 \times \mathcal{P}(\mathcal{X}_2))$ and proving that it satisfies all the properties mentioned above.

Remark 4.1. When $\mathcal{X}_1$ contains no isolated points, because $\mathcal{P}(\mathcal{X}_1 \times \mathcal{P}(\mathcal{X}_2))$ is the metric completion of $\mathcal{P}(\mathcal{X}_1 \times \mathcal{X}_2)$ w.r.t. $\mathcal{AW}_p$ and because the above properties imply that $\mathcal{CW}_p$ is (uniformly) continuous w.r.t. $\mathcal{AW}_p$, we have already uniquely identified $\overline{\mathcal{CW}}_p$. Still, we want to find an expression that allows us to work with $\overline{\mathcal{CW}}_p$ and in particular that allows us to prove that $\overline{\mathcal{SCW}}_p$ is a metric and not just a pseudometric, i.e. that the induced topology is in fact Hausdorff. This is exactly what we gain from assuming compact base spaces and passing to the completion: instead of having to find a lower bound for $\mathcal{SCW}_p(\mu,\nu)$ in terms of $\mathcal{AW}_p(\mu,\nu)$ (and possibly μ) we now just have to prove that if $\mu \neq \nu$ then $\overline{\mathcal{SCW}}_p(\mu,\nu) > 0$.

4.1. **Extending the causal 'distance'.** So now we are working with two Polish metric spaces $\mathcal{X}_1$, $\mathcal{X}_2$. Remember that we denote the 'canonical process' on $\overline{\mathcal{X}} := \mathcal{X}_1 \times \mathcal{X}_2$ by $(X_i)_{i=1,2}$, i.e. $X_i : \overline{\mathcal{X}} \to \mathcal{X}_i$ is the projection onto the i-th coordinate.

To differentiate between the different roles that $\overline{\mathcal{X}}$ may play - i.e. is it the space for the left measure μ or the right measure ν when measuring the 'distance' $\mathcal{CW}_p(\mu,\nu)$ - we will also refer to $\overline{\mathcal{X}}$, $\mathcal{X}_i$ by the aliases $\overline{\mathcal{Y}}$, $\mathcal{Y}_i$ respectively. (And later $\overline{\mathcal{Z}}$, $\mathcal{Z}_i$ as well.) Analogously, we have $Y_i : \overline{\mathcal{Y}} \to \mathcal{Y}_i$. (And $Z_i : \overline{\mathcal{Z}} \to \mathcal{Z}_i$.)

In this section we will repeatedly make use of the following construction:

Definition 4.2. Let $\mathcal{A}$, $\mathcal{B}$, $\mathcal{C}$ be Polish metric spaces. Let $\mu \in \mathcal{P}(\mathcal{A} \times \mathcal{B})$ and $\nu \in \mathcal{P}(\mathcal{B} \times \mathcal{C})$ with $\mu_{\restriction \mathcal{B}} = \nu_{\restriction \mathcal{B}}$. We define

$$\mu \underset{\mathcal{B}}{\otimes} \nu \in \mathcal{P}(\mathcal{A} \times \mathcal{B} \times \mathcal{C})$$

as the measure given by

$$\begin{aligned} \int h \, \mathrm{d}(\mu \underset{\mathcal{B}}{\otimes} \nu) &:= \int h(a,b,c) \, \mathrm{d}\nu_b(c) \, \mathrm{d}\mu(a,b) \\ &= \int h(a,b,c) \, \mathrm{d}\mu_b(a) \, \mathrm{d}\nu(b,c) \end{aligned} \tag{23}$$

where $b \mapsto \nu_b$ is a disintegration of ν w.r.t. $\mathcal{B}$ and similarly for μ.

We further define

$$\mu \fatsemi_{\mathcal{B}} \nu := \left(\mu \underset{\mathcal{B}}{\otimes} \nu\right)_{\restriction \mathcal{A}\times\mathcal{C}} \in \mathcal{P}(\mathcal{A}\times\mathcal{C}) \ .$$

Remark 4.3. If μ is a probability on $\mathcal{A}\times\mathcal{B}$ and ν is a probability on $\mathcal{B}\times\mathcal{C}$, another way of saying what $\mu \underset{\mathcal{B}}{\otimes} \nu$ is, is to state that it is a probability on $\mathcal{A}\times\mathcal{B}\times\mathcal{C}$ s.t. the law of (A,B) is equal to μ, the law of (B,C) is equal to ν (where per our convention A is the projectio onto $\mathcal{A}$, etc.), and A is conditionally independent from C given B. (For the notion of conditional independence see for example [21, Definition II.43].)

Another helpful intuition comes from looking at the case where $\mu \in \mathcal{F}(\mathcal{A} \rightsquigarrow \mathcal{B})$ is concentrated on the graph of some measurable function $f : \mathcal{A} \to \mathcal{B}$ and $\nu \in \mathcal{F}(\mathcal{B} \rightsquigarrow \mathcal{C})$ is concentrated on the graph of a measurable function $g : \mathcal{B} \to \mathcal{C}$. $\mu \fatsemi_{\mathcal{B}} \nu$ is then concentrated on the graph of $g \circ f : \mathcal{A} \to \mathcal{C}$. In some contexts $g \circ f$ is also written as $f \fatsemi g$, which is where we borrowed the symbol from.

Remark 4.4. We will often encounter the situation that one of the factors $\mathcal{A}$, $\mathcal{B}$ or $\mathcal{C}$ in Definition 4.2 is itself a product of spaces and the individual factors may not always be so nicely sorted. We will rely on naming in the subscript the space(s) along which to join the measures μ and ν. For example if $\mu \in \mathcal{P}(\mathcal{A}_1 \times \mathcal{B}_1 \times \mathcal{A}_1 \times \mathcal{B}_2)$ and $\nu \in \mathcal{P}(\mathcal{B}_2 \times \mathcal{C}_1 \times \mathcal{B}_1 \times \mathcal{C}_2)$ we might write

$$\mu \underset{\mathcal{B}_1,\mathcal{B}_2}{\otimes} \nu \in \mathcal{P}(\mathcal{A}_1 \times \mathcal{B}_1 \times \mathcal{A}_2 \times \mathcal{B}_2 \times \mathcal{C}_1 \times \mathcal{C}_2)$$

to refer to the measure that we get when in (23) we use $(b_1, b_2) \in \mathcal{B}_1 \times \mathcal{B}_2$ as the middle variable b. We will not be systematic about the order of the factors in the resulting product space on which e.g. $\mu \underset{\mathcal{B}_1,\mathcal{B}_2}{\otimes} \nu$ is a measure, again relying on naming our spaces for disambiguation.

For future reference we paraphrase the definition of a causal transport plan given in (3) in the introduction.

Lemma 4.5. *Let μ be a measure on $\overline{\mathcal{X}} = \mathcal{X}_1 \times \mathcal{X}_2$ and ν be a measure on $\overline{\mathcal{Y}} = \mathcal{Y}_1 \times \mathcal{Y}_2$. $\gamma \in \mathrm{Cpl}(\mu,\nu)$ is a causal transference plan from μ to ν iff under γ*

$$X_2 \text{ and } Y_1 \text{ are conditionally independent given } X_1.$$

Proof. One way of formulating conditional independence is as in (3), see for example [21, Definition II.43, Theorem II.45]. □

In other words, $\gamma \in \mathrm{Cpl}(\mu,\nu)$ is a causal transference plan iff $\gamma_{\restriction \mathcal{X}_1,\mathcal{X}_2,\mathcal{Y}_1} = \mu \underset{\mathcal{X}_1}{\otimes} \gamma_{\restriction \mathcal{X}_1,\mathcal{Y}_1}$.

We start by reexpressing $\mathcal{CW}_p$ in different ways until we find one which also makes sense in $\mathcal{P}(\mathcal{X}_1 \times \mathcal{P}(\mathcal{X}_2))$.

Let $\mu \in \mathcal{P}(\overline{\mathcal{X}})$ and $\nu \in \mathcal{P}(\overline{\mathcal{Y}})$. Then

$$\begin{aligned}\mathcal{CW}_p(\mu,\nu)^p &= \inf_{\substack{\gamma\in\mathrm{Cpl}(\mu,\nu)\\ \gamma \text{ causal}}} \int \rho(x_1,y_1)^p + \rho(x_2,y_2)^p \,\mathrm{d}\gamma(x_1,x_2,y_1,y_2)\\ &= \inf_{\gamma\in C_1} \int \rho(x_1,y_1)^p + \rho(x_2,y_2)^p \,\mathrm{d}\gamma(x_1,x_2,y_1,y_2)\end{aligned}$$

where

$$C_1 = \left\{\gamma \in \mathrm{Cpl}(\mu,\nu) \,\middle|\, \gamma = \left(\mu \underset{\mathcal{X}_1}{\otimes} \gamma_{\restriction \mathcal{X}_1,\mathcal{Y}_1}\right) \underset{\mathcal{X}_2,\mathcal{Y}_1}{\otimes} \gamma_{\restriction \mathcal{X}_2,\mathcal{Y}_1,\mathcal{Y}_2}\right\} \ .$$

This is true because, on the one hand clearly a $\gamma \in C_1$ is causal by Lemma 4.5 and the alternative characterization of $\underset{\mathcal{X}_1}{\otimes}$. On the other hand, given any causal $\gamma \in \mathrm{Cpl}(\mu,\nu)$, again by Lemma 4.5, $\gamma_{\restriction \mathcal{X}_1,\mathcal{X}_2,\mathcal{Y}_1} = \mu \underset{\mathcal{X}_1}{\otimes} \gamma_{\restriction \mathcal{X}_1,\mathcal{Y}_1}$, and we may define

$\gamma' := \left(\mu \underset{\mathcal{X}_1}{\otimes} \gamma_{\restriction \mathcal{X}_1,\mathcal{Y}_1}\right) \underset{\mathcal{X}_2,\mathcal{Y}_1}{\otimes} \gamma_{\restriction \mathcal{X}_2,\mathcal{Y}_1,\mathcal{Y}_2} \in \mathrm{Cpl}\,(\mu,\nu)$. Now $\gamma_{\restriction \mathcal{X}_1,\mathcal{X}_2,\mathcal{Y}_1} = \gamma'_{\restriction \mathcal{X}_1,\mathcal{X}_2,\mathcal{Y}_1}$ and $\gamma_{\restriction \mathcal{X}_2,\mathcal{Y}_1,\mathcal{Y}_2} = \gamma'_{\restriction \mathcal{X}_2,\mathcal{Y}_1,\mathcal{Y}_2}$, so in particular

$$\int \rho(x_1,y_1)^p + \rho(x_2,y_2)^p \,\mathrm{d}\gamma(x_1,x_2,y_1,y_2) = \int \rho(x_1,y_1)^p + \rho(x_2,y_2)^p \,\mathrm{d}\gamma'(x_1,x_2,y_1,y_2)\ .$$

We may name the different building blocks of $\gamma \in C_1$ to get

$$\mathcal{CW}_p\,(\mu,\nu)^p = \inf_{(\gamma,\beta)\in C_2} \int \rho(x_1,y_1)^p \,\mathrm{d}\gamma(x_1,y_1) + \int \rho(x_2,y_2)^p \,\mathrm{d}\beta(y_1,x_2,y_2)$$

with

$$C_2 = \Big\{(\gamma,\beta) \in \mathrm{Cpl}\,(\mu_{\restriction \mathcal{X}_1},\nu_{\restriction \mathcal{Y}_1}) \times \mathcal{P}(\mathcal{Y}_1\times\mathcal{X}_2\times\mathcal{Y}_2) \;\Big|\; \beta_{\restriction \mathcal{X}_2,\mathcal{Y}_1} = \mu \,{}^{g}_{g}{}_{\mathcal{X}_1}\, \gamma \text{ and } \beta_{\restriction \mathcal{Y}_1,\mathcal{Y}_2} = \nu\Big\}\ ,$$

i.e. there is a bijection between C_1 and C_2 given by sending $\gamma' \in C_1$ to $(\gamma,\beta) \in C_2$ where $\gamma := \gamma'_{\restriction \mathcal{X}_1,\mathcal{Y}_1}$, $\beta := \gamma'_{\restriction \mathcal{X}_2,\mathcal{Y}_1,\mathcal{Y}_2}$, and, in the other direction, by sending $(\gamma,\beta) \in C_2$ to $\gamma' := \left(\mu \underset{\mathcal{X}_1}{\otimes} \gamma\right) \underset{\mathcal{X}_2,\mathcal{Y}_1}{\otimes} \beta$.

We can apply the bijection $\mathrm{dis}_{\mathcal{Y}_1} : \mathcal{P}(\mathcal{Y}_1\times\mathcal{X}_2\times\mathcal{Y}_2) \to \mathcal{F}\,(\mathcal{Y}_1 \rightsquigarrow \mathcal{P}(\mathcal{X}_2\times\mathcal{Y}_2))$ to β. Translating the conditions on $(\gamma,\beta) \in C_2$ to conditions on $(\gamma, \mathrm{dis}_{\mathcal{Y}_1}(\beta))$ we arrive at

$$\mathcal{CW}_p\,(\mu,\nu)^p = \inf_{(\gamma,\beta)\in C_3} \int \rho(x_1,y_1)^p \,\mathrm{d}\gamma + \int\int \rho(x_2,y_2)^p \,\mathrm{d}\beta'(x_2,y_2)\,\mathrm{d}\beta(y_1,\beta')$$

where

$$C_3 = \Big\{(\gamma,\beta) \in \mathrm{Cpl}\,(\mu_{\restriction \mathcal{X}_1},\nu_{\restriction \mathcal{Y}_1}) \times \mathcal{F}\,(\mathcal{Y}_1 \rightsquigarrow \mathcal{P}(\mathcal{X}_2\times\mathcal{Y}_2)) \;\Big|\; \mathcal{P}(1_{\mathcal{Y}_1}\times\mathcal{P}(Y_2))\,(\beta) = \mathrm{dis}_{\mathcal{Y}_1}(\nu) \text{ and } \mathcal{P}(1_{\mathcal{Y}_1}\times\mathcal{P}(X_2))\,(\beta) = \mathrm{dis}_{\mathcal{Y}_1}\left(\gamma \,{}^{g}_{g}{}_{\mathcal{X}_1}\, \mu\right)\Big\}\ .$$

Let $(\gamma,\beta) \in C_3$ and let $(y_1,\beta') \mapsto \tilde\beta'_{y_1,\beta'}$ be a measurable mapping with $\tilde\beta'_{y_1,\beta'} \in \mathrm{Cpl}\left(\beta'_{\restriction \mathcal{X}_2}, \beta'_{\restriction \mathcal{Y}_2}\right)$ for β-a.a. (y_1,β'). Then we have that also $(\gamma,\tilde\beta) \in C_3$, where $\tilde\beta \in \mathcal{F}\,(\mathcal{Y}_1 \rightsquigarrow \mathcal{P}(\mathcal{X}_2\times\mathcal{Y}_2))$ is defined by

$$\tilde\beta := f \mapsto \int f(y_1, \tilde\beta'_{y_1,\beta'})\,\mathrm{d}\beta(y_1,\beta')\ .$$

By employing a β-a.e. measurable selector this implies that

$$\begin{aligned}\mathcal{CW}_p\,(\mu,\nu)^p &= \inf_{(\gamma,\beta)\in C_3} \textstyle\int \rho(x_1,y_1)^p \,\mathrm{d}\gamma + \int \inf_{\tilde\beta' \in \mathrm{Cpl}\left(\beta'_{\restriction \mathcal{X}_2}, \beta'_{\restriction \mathcal{Y}_2}\right)} \int \rho(x_2,y_2)^p \,\mathrm{d}\tilde\beta'(x_2,y_2)\,\mathrm{d}\beta(y_1,\beta') \\ &= \inf_{(\gamma,\beta)\in C_3} \textstyle\int \rho(x_1,y_1)^p \,\mathrm{d}\gamma + \int \mathcal{W}_p\left(\beta'_{\restriction \mathcal{X}_2}, \beta'_{\restriction \mathcal{Y}_2}\right)^p \,\mathrm{d}\beta(y_1,\beta')\ .\end{aligned}$$

We need

Lemma 4.6. *If $\kappa \in \mathcal{P}(\mathcal{A}\times\mathcal{B})$ and $\lambda \in \mathcal{F}\,(\mathcal{B} \rightsquigarrow \mathcal{C})$ then the only measure $\eta \in \mathcal{P}(\mathcal{A}\times\mathcal{B}\times\mathcal{C})$ with $\eta_{\restriction \mathcal{A}\times\mathcal{B}} = \kappa$ and $\eta_{\restriction \mathcal{B}\times\mathcal{C}} = \lambda$ is $\kappa \underset{\mathcal{B}}{\otimes} \lambda$.*

Proof. If η satisfies the properties above and $b \mapsto \kappa_b$, $b \mapsto \lambda_b$ are (classical) disintegrations of κ, λ w.r.t. $\mathcal{B}$, then a (classical) disintegration $b \mapsto \eta_b$ of η w.r.t. $\mathcal{B}$ has to satisfy $\eta_{b\restriction \mathcal{A}} = \kappa_b$ and $\eta_{b\restriction \mathcal{C}} = \lambda_b$ a.s. As λ_b is a Dirac measure a.s. this forces η_b to be $\kappa_b \otimes \lambda_b$ almost surely. □

This implies that for $(\gamma, \beta) \in C_3$ the distribution of

$$(y_1, \beta') \mapsto (y_1, \beta'_{\restriction \mathcal{X}_2}, \beta'_{\restriction \mathcal{Y}_2}) \tag{24}$$

under β is already determined by γ, i.e. because the distribution of $(y_1, \beta') \mapsto (y_1, \beta'_{\restriction \mathcal{X}_2})$ is $\mathrm{dis}_{\mathcal{Y}_1}\big(\gamma \,\mathbin{⨟}_{\mathcal{X}_1}\, \mu\big)$ and the distribution of $(y_1, \beta') \mapsto (y_1, \beta'_{\restriction \mathcal{Y}_2})$ is $\mathrm{dis}_{\mathcal{Y}_1}(\nu)$, the distribution of (24) under β must be equal to

$$\mathrm{dis}_{\mathcal{Y}_1}\big(\gamma \,\mathbin{⨟}_{\mathcal{X}_1}\, \mu\big) \underset{\mathcal{Y}_1}{\otimes} \mathrm{dis}_{\mathcal{Y}_1}(\nu) \ .$$

This means that we may get rid of β:

$$\begin{aligned}\mathcal{CW}_p(\mu,\nu)^p = \inf_{\gamma \in \mathrm{Cpl}\left(\mu_{\restriction \mathcal{X}_1}, \nu_{\restriction \mathcal{Y}_1}\right)} & \int \rho(x_1,y_1)^p \,\mathrm{d}\gamma \\ & + \int \mathcal{W}_p(\mu',\nu')^p \,\mathrm{d}\left(\mathrm{dis}_{\mathcal{Y}_1}\big(\gamma \,\mathbin{⨟}_{\mathcal{X}_1}\, \mu\big) \underset{\mathcal{Y}_1}{\otimes} \mathrm{dis}_{\mathcal{Y}_1}(\nu)\right)(y_1,\mu',\nu')\end{aligned}$$

For the final step we need another lemma:

Lemma 4.7. *Let $\lambda \in \mathcal{P}(\mathcal{A} \times \mathcal{B})$ and $\beta \in \mathcal{P}(\mathcal{B} \times \mathcal{C})$. Let $\hat{C}$ denote the projection onto $\mathcal{P}(\mathcal{C})$. Then*

$$\mathrm{dis}_{\mathcal{A}}(\lambda \,\mathbin{⨟}_{\mathcal{B}}\, \beta)$$

is equal to the distribution of

$$(A, \mathbb{E}^{\eta}(\hat{C}|A)) \text{ under } \eta := \lambda \underset{\mathcal{B}}{\otimes} \mathrm{dis}_{\mathcal{B}}(\beta) \ .$$

Proof. Let $a \mapsto \lambda_a$ be a version of the (classical) disintegration of λ w.r.t. $\mathcal{A}$ and let $b \mapsto \beta_b$ be a disintegration of β w.r.t. $\mathcal{B}$.

As one easily checks, a version of the (classical) disintegration of $\lambda \,\mathbin{⨟}_{\mathcal{B}}\, \beta$ w.r.t. $\mathcal{A}$ is given by $a \mapsto \int \beta_b \,\mathrm{d}\lambda_a(b)$, so that $\mathrm{dis}_{\mathcal{A}}(\lambda \,\mathbin{⨟}_{\mathcal{B}}\, \beta)$ is equal to

$$\mathcal{P}\Big(a \mapsto (a, \textstyle\int \beta_b \,\mathrm{d}\lambda_a(b))\Big)(\lambda_{\restriction \mathcal{A}}) \ .$$

By the same argument a version of the disintegration of $\lambda \,\mathbin{⨟}_{\mathcal{B}}\, \mathrm{dis}_{\mathcal{B}}(\beta)$ w.r.t. $\mathcal{A}$ is given by $h := a \mapsto \int \mathrm{dis}_{\mathcal{B}}(\beta)_b \,\mathrm{d}\lambda_a(b)$, where $b \mapsto \mathrm{dis}_{\mathcal{B}}(\beta)_b$ is a disintegration of $\mathrm{dis}_{\mathcal{B}}(\beta)$ w.r.t. $\mathcal{B}$. But such a disintegration is given by $b \mapsto \delta_{\beta_b}$, (where δ_{β_b} is the dirac measure at β_b). So $h = a \mapsto \int \delta_{\beta_b} \,\mathrm{d}\lambda_a(b)$. This means (a version of) $\mathbb{E}^{\eta}(\hat{C}|A)$ is given by

$$\mathbb{E}^{\eta}(\hat{C}|A)(a,\_,\_) = \textstyle\int \hat{c} \,\mathrm{d}(\int \delta_{\beta_b} \,\mathrm{d}\lambda_a(b))(\hat{c}) = \iint \hat{c} \,\mathrm{d}\delta_{\beta_b}(\hat{c}) \,\mathrm{d}\lambda_a(b) = \int \beta_b \,\mathrm{d}\lambda_a(b) \ ,$$

so that the distribution of $(A, \mathbb{E}^{\eta}(\hat{C}|A))$ under η is also given by

$$\mathcal{P}\Big(a \mapsto (a, \textstyle\int \beta_b \,\mathrm{d}\lambda_a(b))\Big)(\lambda_{\restriction \mathcal{A}}) \ .$$

□

Using this lemma with $\mathcal{A} = \mathcal{Y}_1$, $\mathcal{B} = \mathcal{X}_1$, $\mathcal{C} = \mathcal{X}_2$, $\lambda = \gamma$, $\beta = \mu$ and writing $\hat{X}_2$, $\hat{Y}_2$ for the projections onto $\mathcal{P}(\mathcal{X}_2)$, $\mathcal{P}(\mathcal{Y}_2)$ respectively, we find:

$$\mathcal{CW}_p(\mu,\nu)^p = \inf_{\gamma \in \mathrm{Cpl}\left(\mu_{\restriction \mathcal{X}_1}, \nu_{\restriction \mathcal{Y}_1}\right)} \mathbb{E}^{\gamma}(\rho(X_1,Y_1)^p) + \mathbb{E}^{\eta(\gamma)}\left(\mathcal{W}_p\left(\mathbb{E}^{\eta(\gamma)}(\hat{X}_2|Y_1), \hat{Y}_2\right)^p\right)$$

where $\eta(\gamma) := \mathrm{dis}_{\mathcal{X}_1}(\mu) \underset{\mathcal{X}_1}{\otimes} \gamma \underset{\mathcal{Y}_1}{\otimes} \mathrm{dis}_{\mathcal{Y}_1}(\nu)$.

By Lemma 4.6 the function $\eta : \mathrm{Cpl}(\mu_{\restriction \mathcal{X}_1}, \nu_{\restriction \mathcal{Y}_1}) \to \mathrm{Cpl}(\mathrm{dis}_{\mathcal{X}_1}(\mu), \mathrm{dis}_{\mathcal{Y}_1}(\nu))$ is a bijection, so we may as well write

$$\mathcal{CW}_p(\mu,\nu)^p = \inf_{\gamma \in \mathrm{Cpl}\left(\mathrm{dis}_{\mathcal{X}_1}(\mu), \mathrm{dis}_{\mathcal{Y}_1}(\nu)\right)} \mathbb{E}^{\gamma}(\rho(X_1,Y_1)^p) + \mathbb{E}^{\gamma}\left(\mathcal{W}_p\left(\mathbb{E}^{\gamma}(\hat{X}_2|Y_1), \hat{Y}_2\right)^p\right) \ .$$

Finally, under any $\gamma \in \mathrm{Cpl}(\mathrm{dis}_{\mathcal{X}_1}(\mu), \mathrm{dis}_{\mathcal{Y}_1}(\nu))$ we know that $\hat{Y}_2$ is almost surely equal to a function of Y_1, so that the completions of the sigma-algebras generated

by Y_1 and $\vec{Y} := (Y_1, \hat{Y}_2)$ respectively are equal. This means that $\mathbb{E}^\gamma(\hat{X}_2|Y_1) = \mathbb{E}^\gamma(\hat{X}_2|\vec{Y})$ a.s. and we arrive at our final expression for $\mathcal{CW}_p(\mu,\nu)$:

$$\mathcal{CW}_p(\mu,\nu) = \inf_{\gamma\in \mathrm{Cpl}\left(\mathrm{dis}_{\mathcal{X}_1}(\mu),\mathrm{dis}_{\mathcal{Y}_1}(\nu)\right)} \left(\mathbb{E}^\gamma\left(\rho(X_1,Y_1)^p + \mathcal{W}_p\left(\mathbb{E}^\gamma\left(\hat{X}_2|\vec{Y}\right),\hat{Y}_2\right)^p\right)\right)^{1/p}.$$

Now this expression is trivial to generalize to $\mu \in \mathcal{P}(\mathcal{X}_1 \times \mathcal{P}(\mathcal{X}_2))$ and $\nu \in \mathcal{P}(\mathcal{Y}_1 \times \mathcal{P}(\mathcal{Y}_2))$, i.e. for such μ, ν we set

$$\overline{\mathcal{CW}}_p(\mu,\nu) := \inf_{\gamma\in\mathrm{Cpl}(\mu,\nu)} \left(\mathbb{E}^\gamma\left(\rho(X_1,Y_1)^p + \mathcal{W}_p\big(\mathbb{E}^\gamma\left(\hat{X}_2|\vec{Y}\right),\hat{Y}_2\big)^p\right)\right)^{1/p}. \tag{25}$$

To summarize our discussion up to this point:

Lemma 4.8. *The function*

$$\overline{\mathcal{CW}}_p : \mathcal{P}(\mathcal{X}_1 \times \mathcal{P}(\mathcal{X}_2))^2 \to \mathbb{R}_+$$

as defined in (25) *is really an extension of*

$$\mathcal{CW}_p : \mathcal{P}(\mathcal{X}_1 \times \mathcal{X}_2)^2 \to \mathbb{R}_+$$

as defined in (20) *(when $\mathcal{P}(\mathcal{X}_1 \times \mathcal{X}_2)$ is embedded into $\mathcal{P}(\mathcal{X}_1 \times \mathcal{P}(\mathcal{X}_2))$ via $\mathrm{dis}_{\mathcal{X}_1}$).*

Next we promised to show

Lemma 4.9. *$\overline{\mathcal{CW}}_p$ is bounded by $\overline{\mathcal{AW}}_p$, i.e.*

$$\overline{\mathcal{CW}}_p(\mu,\nu) \le \inf_{\gamma\in\mathrm{Cpl}(\mu,\nu)} \left(\mathbb{E}^\gamma\left(\rho(X_1,Y_1)^p + \mathcal{W}_p(\hat{X}_2,\hat{Y}_2)^p\right)\right)^{1/p} = \overline{\mathcal{AW}}_p(\mu,\nu).$$

Proof. By the conditional version of Jensen's inequality applied to the convex function $(\hat{x},\hat{y}) \mapsto \mathcal{W}_p(\hat{x},\hat{y})^p$ we have

$$\mathcal{W}_p\left(\mathbb{E}^\gamma\left(\hat{X}_2|\vec{Y}\right),\hat{Y}_2\right)^p = \mathcal{W}_p\left(\mathbb{E}^\gamma\left((\hat{X}_2,\hat{Y}_2)|\vec{Y}\right)\right)^p \le \mathbb{E}^\gamma\left(\mathcal{W}_p\left(\hat{X}_2,\hat{Y}_2\right)^p|\vec{Y}\right).$$

□

Remark 4.10. For the reader who may be sceptical of whether Jensen's inequality holds in this rather unusual setting, where we have a convex function

$$\mathcal{W}_p : \mathcal{P}(\mathcal{X}_2) \times \mathcal{P}(\mathcal{Y}_2) \to \mathbb{R}_+$$

and conditional expectations on spaces of measures we remark that for the Wasserstein distance in particular this is very easy to check. The proof is just integrating transport plans between $\hat{X}_2$ and $\hat{Y}_2$ w.r.t. the distribution of these conditioned on $\vec{Y}$ (in this case) to get transport plans between $\mathbb{E}^\gamma(\hat{X}_2|\vec{Y})$ and $\mathbb{E}^\gamma(\hat{Y}_2|\vec{Y})$.

Lemma 4.11. *Let $\mu,\nu,\lambda \in \mathcal{P}(\mathcal{X}_1 \times \mathcal{P}(\mathcal{X}_2))$. Then*

$$\overline{\mathcal{CW}}_p(\mu,\lambda) \le \overline{\mathcal{CW}}_p(\mu,\nu) + \overline{\mathcal{CW}}_p(\nu,\lambda).$$

Proof. Using our naming convention we have

$$\mu \in \mathcal{P}(\mathcal{X}_1 \times \mathcal{P}(\mathcal{X}_2)), \quad \nu \in \mathcal{P}(\mathcal{Y}_1 \times \mathcal{P}(\mathcal{Y}_2)), \quad \lambda \in \mathcal{P}(\mathcal{Z}_1 \times \mathcal{P}(\mathcal{Z}_2)).$$

We denote the projections onto $\mathcal{P}(\mathcal{X}_2)$, $\mathcal{P}(\mathcal{Y}_2)$, $\mathcal{P}(\mathcal{Z}_2)$ by $\hat{X}_2$, $\hat{Y}_2$, $\hat{Z}_2$ respectively. $\vec{Y} = (Y_1,\hat{Y}_2)$, $\vec{Z} := (Z_1,\hat{Z}_2)$.

Let $\gamma \in \mathrm{Cpl}(\mu,\nu)$ and $\eta \in \mathrm{Cpl}(\nu,\lambda)$. In the following let $\mathbb{E}$ refer to (conditional) expectation w.r.t. $\kappa := \gamma \underset{\mathcal{Y}_1,\mathcal{P}(\mathcal{Y}_2)}{\otimes} \eta$, and let $\|\cdot\|_{L_p}$ refer to the L_p-norm w.r.t. κ.

Combining the triangle inequalities for ρ, $\mathcal{W}_p$ and the $\|\cdot\|_{L_p}$ we get

$$\|\rho(X_1,Z_1)\|_{L_p} \le \|\rho(X_1,Y_1)\|_{L_p} + \|\rho(Y_1,Z_1)\|_{L_p} \tag{26}$$

$$\|\mathcal{W}_p\left(\mathbb{E}\left(\hat{X}_2|\vec{Z}\right),\hat{Z}_2\right)\|_{L_p} \le \|\mathcal{W}_p\left(\mathbb{E}\left((\hat{X}_2,\hat{Y}_2)|\vec{Z}\right)\right)\|_{L_p} + \|\mathcal{W}_p\left(\mathbb{E}\left(\hat{Y}_2|\vec{Z}\right),\hat{Z}_2\right)\|_{L_p} \tag{27}$$

By the conditional Jensen inequality

$$\begin{aligned}\mathcal{W}_p\left(\mathbb{E}\left((\hat{X}_2,\hat{Y}_2)|\vec{Z}\right)\right)^p &= \mathcal{W}_p\left(\mathbb{E}\Big(\mathbb{E}\big((\hat{X}_2,\hat{Y}_2)\big|\vec{Y},\vec{Z}\big)\Big|\vec{Z}\Big)\right)^p \\ &\leq \mathbb{E}\left(\mathcal{W}_p\Big(\mathbb{E}\big((\hat{X}_2,\hat{Y}_2)\big|\vec{Y},\vec{Z}\big)\Big)^p\middle|\vec{Z}\right)\end{aligned}$$

and therefore

$$\left\|\mathcal{W}_p\left(\mathbb{E}\left((\hat{X}_2,\hat{Y}_2)|\vec{Z}\right)\right)\right\|_{L_p}^p \leq \left\|\mathcal{W}_p\left(\mathbb{E}\left((\hat{X}_2,\hat{Y}_2)|\vec{Y},\vec{Z}\right)\right)^p\right\|_{L_p} .$$

By construction, $(\hat{X}_2,\hat{Y}_2)$ is conditionally independent from $\vec{Z}$ given $\vec{Y}$, so that $\mathbb{E}\left((\hat{X}_2,\hat{Y}_2)|\vec{Y},\vec{Z}\right) = \mathbb{E}\left((\hat{X}_2,\hat{Y}_2)|\vec{Y}\right)$ (this basic fact about conditional independence can be found for example as Theorem 45 in [21]). Combining this with (27) gives

$$\left\|\mathcal{W}_p\left(\mathbb{E}\left(\hat{X}_2|\vec{Z}\right),\hat{Z}_2\right)\right\|_{L_p} \leq \left\|\mathcal{W}_p\left(\mathbb{E}\left(\hat{X}_2|\vec{Y}\right),\hat{Y}_2\right)\right\|_{L_p} + \left\|\mathcal{W}_p\left(\mathbb{E}\left(\hat{Y}_2|\vec{Z}\right),\hat{Z}_2\right)\right\|_{L_p} . \tag{28}$$

Putting together (26) and (28) with the triangle inequality for ℓ_p we get

$$\begin{aligned}\mathcal{CW}_p(\mu,\lambda) &= \left(\|\rho(X_1,Z_1)\|_{L_p}^p + \left\|\mathcal{W}_p\left(\mathbb{E}\left(\hat{X}_2|\vec{Z}\right),\hat{Z}_2\right)\right\|_{L_p}^p\right)^{1/p} \\ &\leq \left(\|\rho(X_1,Y_1)\|_{L_p}^p + \left\|\mathcal{W}_p\left(\mathbb{E}\left(\hat{X}_2|\vec{Y}\right),\hat{Y}_2\right)\right\|_{L_p}^p\right)^{1/p} \\ &\quad + \left(\|\rho(Y_1,Z_1)\|_{L_p}^p + \left\|\mathcal{W}_p\left(\mathbb{E}\left(\hat{Y}_2|\vec{Z}\right),\hat{Z}_2\right)\right\|_{L_p}^p\right)^{1/p} \\ &= \mathcal{CW}_p(\mu,\nu) + \mathcal{CW}_p(\nu,\lambda) .\end{aligned}$$

□

Lemma 4.12. *$\overline{\mathcal{CW}}_p$ is uniformly continuous w.r.t. $\overline{\mathcal{AW}}_p$ on $\mathcal{P}(\mathcal{X}_1\times\mathcal{P}(\mathcal{X}_2))^2$.*

Proof. Let $\mu,\nu,\mu',\nu' \in \mathcal{P}(\mathcal{X}_1\times\mathcal{P}(\mathcal{X}_2))$. We repeatedly use Lemma 4.11:

$$\overline{\mathcal{CW}}_p(\mu,\nu) \leq \overline{\mathcal{CW}}_p(\mu,\nu') + \overline{\mathcal{CW}}_p(\nu',\nu) \leq \overline{\mathcal{CW}}_p(\mu,\mu') + \overline{\mathcal{CW}}_p(\mu',\nu') + \overline{\mathcal{CW}}_p(\nu',\nu)$$

therefore

$$\overline{\mathcal{CW}}_p(\mu,\nu) - \overline{\mathcal{CW}}_p(\mu',\nu') \leq \overline{\mathcal{CW}}_p(\mu,\mu') + \overline{\mathcal{CW}}_p(\nu',\nu) .$$

Switching the roles of (μ,ν) and (μ',ν') implies

$$\begin{aligned}&|\overline{\mathcal{CW}}_p(\mu,\nu) - \overline{\mathcal{CW}}_p(\mu',\nu')| \\ &\quad\leq \max\left(\overline{\mathcal{CW}}_p(\mu,\mu'),\overline{\mathcal{CW}}_p(\mu',\mu)\right) + \max\left(\overline{\mathcal{CW}}_p(\nu,\nu'),\overline{\mathcal{CW}}_p(\nu',\nu)\right) \\ &\quad\leq \overline{\mathcal{AW}}_p(\mu,\mu') + \overline{\mathcal{AW}}_p(\nu,\nu') .\end{aligned}$$

□

Lemma 4.13. *The infimum in* (25) *is attained.*

Proof. This is an application of [8, Theorem 1.2].

For self-containedness and because it's a nice application of the nested distance, we also sketch the argument here. We know that $\mathrm{Cpl}(\mu,\nu)$ is compact. The problem is that $\gamma \mapsto \mathbb{E}^\gamma\left(\mathcal{W}_p\left(\mathbb{E}^\gamma\left(\hat{X}_2|\vec{Y}\right),\vec{Y}\right)^p\right)$ is not (lower semi-) continuous. But we may switch to a topology which is better *adapted* to the problem at hand. Namely the two-timepoint $\mathcal{AW}_p$-topology. In this case the space for the first timepoint is $\mathcal{Y}_1\times\mathcal{P}(\mathcal{Y}_2)$ and that for the second is $\mathcal{X}_1\times\mathcal{P}(\mathcal{X}_2)$. In effect that means that instead of $\gamma \in \mathrm{Cpl}(\mu,\nu)$ we are now looking at $\gamma' \in \mathcal{F}\left(\mathcal{Y}_1\times\mathcal{P}(\mathcal{Y}_2) \rightsquigarrow \mathcal{P}(\mathcal{X}_1\times\mathcal{P}(\mathcal{X}_2))\right)$. The function that we are optimizing over can be written as

$$\hat{C} := \gamma' \mapsto \mathbb{E}^{\gamma'}\left(C(Y_1,\hat{Y}_2,\hat{\vec{X}})\right)$$

where

$$C(y_1, \hat{y}_2, \xi) = \int \rho(x_1, y_1)\, \mathrm{d}\xi(x_1, \_) + \mathcal{W}_p\left(\mathrm{bary}(\xi_{\restriction \mathcal{P}(\mathcal{X}_2)}), \hat{y}_2\right)$$
$$\mathrm{bary}(\lambda) = \int x\, \mathrm{d}\lambda(x)$$

C is a continuous function and so is $\hat{C}$. Now $\mathrm{dis}_{\mathcal{Y}_1 \times \mathcal{P}(\mathcal{Y}_2)}(\mathrm{Cpl}(\mu, \nu))$ is not compact, but

$$\Big\{\gamma' \in \mathcal{P}(\mathcal{Y}_1 \times \mathcal{P}(\mathcal{Y}_2) \times \mathcal{P}(\mathcal{X}_1 \times \mathcal{P}(\mathcal{X}_2))) \;\Big|\; \gamma'_{\restriction \mathcal{Y}_1 \times \mathcal{P}(\mathcal{Y}_2)} = \nu\,, \quad \mathrm{int}_{\mathcal{Y}_1 \times \mathcal{P}(\mathcal{Y}_2)}(\gamma')_{\restriction \mathcal{P}(\mathcal{X}_1 \times \mathcal{P}(\mathcal{X}_2))} = \mu\Big\}$$

is. So we can find a minimizer γ' of $\hat{C}$ in this set. To return to $\mathrm{Cpl}(\mu, \nu)$, or more precisely $\mathrm{dis}_{\mathcal{Y}_1 \times \mathcal{P}(\mathcal{Y}_2)}(\mathrm{Cpl}(\mu, \nu))$, we can send γ' to the distribution γ'' of $(Y_1, \hat{Y}_2, \mathbb{E}^{\gamma'}(\vec{\hat{X}}|\vec{Y}))$. Because C is continuous and convex in its last argument and by (the conditional version of) Jensens inequality (which could again be proved 'by hand' here) $\hat{C}(\gamma'') \leq \hat{C}(\gamma')$. $\mathrm{int}_{\mathcal{Y}_1 \times \mathcal{P}(\mathcal{Y}_2)}(\gamma'')$ is the sought after minimizer of (25). □

Lemma 4.14. *Let $\mu, \nu \in \mathcal{P}(\mathcal{X}_1 \times \mathcal{P}(\mathcal{X}_2))$. Then $\overline{\mathcal{CW}}_p(\mu, \nu) = \overline{\mathcal{CW}}_p(\nu, \mu) = 0$ implies $\mu = \nu$.*

Proof. Call

$$\vec{\mathcal{X}} := \mathcal{X}_1 \times \mathcal{P}(\mathcal{X}_2) \qquad \vec{\mathcal{Y}} := \mathcal{Y}_1 \times \mathcal{P}(\mathcal{Y}_2) \qquad \vec{\mathcal{Z}} := \mathcal{Z}_1 \times \mathcal{P}(\mathcal{Z}_2)\ .$$

To have labels for our spaces, see μ, ν as

$$\mu \in \mathcal{P}(\vec{\mathcal{X}})\ , \quad \nu \in \mathcal{P}(\vec{\mathcal{Y}})\ , \quad \mu \in \mathcal{P}(\vec{\mathcal{Z}})\ .$$

Let $\gamma \in \mathrm{Cpl}(\mu, \nu) \subseteq \mathcal{P}(\vec{\mathcal{X}} \times \vec{\mathcal{Y}})$ s.t. $\mathbb{E}^{\gamma}(\rho(X_1, Y_1)^p) + \mathbb{E}^{\gamma}(\mathcal{W}_p(\mathbb{E}^{\gamma}(\hat{X}_2|\vec{Y}), \hat{Y}_2)^p) = 0$.

Let $\eta \in \mathrm{Cpl}(\nu, \mu) \subseteq \mathcal{P}(\vec{\mathcal{Y}} \times \vec{\mathcal{Z}})$ s.t. $\mathbb{E}^{\eta}(\rho(Y_1, Z_1)^p) + \mathbb{E}^{\eta}(\mathcal{W}_p(\mathbb{E}^{\eta}(\hat{Y}_2|\vec{Z}), \hat{Z}_2)^p) = 0$.

All the following considerations happen under $\gamma \underset{\vec{\mathcal{Y}}}{\otimes} \eta$. Clearly, $Z_1 = Y_1 = X_1$ a.s.

Moreover, because $\mathbb{E}(\hat{X}_2|\vec{Y}, \vec{Z}) = \mathbb{E}(\hat{X}_2|\vec{Y})$, the random variables $\hat{Z}_2, \hat{Y}_2, \hat{X}_2$ form a martingale w.r.t. the filtration generated by $\vec{Z}, \vec{Y}, \vec{X}$. The distribution of $\hat{Z}_2$ is equal to the distribution of $\hat{X}_2$. Both of these statements are also true if we integrate some bounded measurable function w.r.t. our random variables, i.e. for any bounded measurable $f : \mathcal{X}_2 \to \mathbb{R}$ we have that $\int f\, \mathrm{d}\hat{Z}_2, \int f\, \mathrm{d}\hat{Y}_2, \int f\, \mathrm{d}\hat{X}_2$ is a martingale and that the distribution of $\int f\, \mathrm{d}\hat{Z}_2$ is equal to the distribution of $\int f\, \mathrm{d}\hat{X}_2$. But this means that we must have $\int f\, \mathrm{d}\hat{Z}_2 = \int f\, \mathrm{d}\hat{Y}_2 = \int f\, \mathrm{d}\hat{X}_2$ a.s. (Lemma 4.15 below). As this is true for all f from a countable generator of the sigma-algebra on $\mathcal{X}_2$, we have $\hat{Z}_2 = \hat{Y}_2 = \hat{X}_2$ a.s. □

Lemma 4.15. *Let X_1, X_2, X_3 be a bounded martingale over $\mathbb{R}$. If the distribution of X_1 is equal to the distribution of X_3 then $X_1 = X_2 = X_3$ a.s.*

Proof. This is a consequence of the strict version of Jensen's inequality applied to any everywhere strictly convex function. (Take for example $x \mapsto x^2$.) □

Remark 4.16. The reason we took the detour of turning our probability-measure-valued martingale into a family of martingales on $\mathbb{R}$ and arguing on these is because this way we avoid having to exhibit a continuous, everywhere strictly convex function on $\mathcal{P}(\mathcal{X}_2)$.

As a reminder:

Definition 4.17. For $\mu, \nu \in \mathcal{P}(\mathcal{X}_1 \times \mathcal{P}(\mathcal{X}_2))$,

$$\overline{\mathcal{SCW}}_p(\mu,\nu) := \max(\overline{\mathcal{CW}}_p(\mu,\nu), \overline{\mathcal{CW}}_p(\nu,\mu)) \ .$$

Theorem 4.18. *$\overline{\mathcal{SCW}}_p$ is a metric on $\mathcal{P}(\mathcal{X}_1 \times \mathcal{P}(\mathcal{X}_2))$ satisfying*

$$\overline{\mathcal{SCW}}_p(\mu,\nu) \leq \overline{\mathcal{AW}}_p(\mu,\nu) \ .$$

Proof. This follows from Lemma 4.11, Lemma 4.14 and Lemma 4.9. □

Remark 4.19. As outlined at the beginning of this section we now know enough to conclude that the topology induced by $\mathcal{SCW}_p$ is equal to the topology induced by $\mathcal{AW}_p$ in the case where $\mathcal{X}_1$ and $\mathcal{X}_2$ are both compact. The non-compact case is not much harder. We need the following lemma.

Lemma 4.20. *The map*

$$\mathrm{int}_{\mathcal{X}_1} : \mathcal{P}(\mathcal{X}_1 \times \mathcal{P}(\mathcal{X}_2)) \to \mathcal{P}(\mathcal{X}_1 \times \mathcal{X}_2)$$

is a contraction when we equip the source space with $\overline{\mathcal{SCW}}_p$ and the target space with $\mathcal{W}_p$. More specifically for $\mu, \nu \in \mathcal{P}(\mathcal{X}_1 \times \mathcal{P}(\mathcal{X}_2))$

$$\mathcal{W}_p(\mathrm{int}_{\mathcal{X}_1}(\mu), \mathrm{int}_{\mathcal{X}_1}(\nu)) \leq \overline{\mathcal{CW}}_p(\mu,\nu) \ . \tag{29}$$

Proof. We prove the second statement. Let $\mu \in \mathcal{P}(\vec{\mathcal{X}})$, $\nu \in \mathcal{P}(\vec{\mathcal{Y}})$. Given $\gamma \in \mathrm{Cpl}(\mu,\nu)$ and $\varepsilon > 0$ the task is to find $\gamma' \in \mathrm{Cpl}(\mathrm{int}_{\mathcal{X}_1}\mu, \mathrm{int}_{\mathcal{Y}_1}\nu)$ s.t.

$$\mathbb{E}^{\gamma'}(\rho(X_1,Y_1)^p + \rho(X_2,Y_2)^p) \leq \mathbb{E}^{\gamma}(\rho(X_1,Y_1)^p) + \mathbb{E}^{\gamma}(\mathcal{W}_p(\mathbb{E}^{\gamma}(\hat{X}_2|\vec{Y}), \hat{Y}_2)^p) + \varepsilon \ . \tag{30}$$

We take inspiration from the discussion at the beginning of this section. Let $\Xi : \vec{\mathcal{X}} \times \vec{\mathcal{Y}} \to \mathcal{P}(\mathcal{X}_2 \times \mathcal{Y}_2)$ be a measurable selector satisfying

$$\begin{aligned} &\Xi \in \mathrm{Cpl}(\mathbb{E}^{\gamma}(\hat{X}_2|\vec{Y}), \hat{Y}_2) \quad \gamma\text{-a.s. and} \\ &\mathbb{E}^{\Xi}(\rho(X_2,Y_2)^p) \leq \mathcal{W}_p(\mathbb{E}^{\gamma}(\hat{X}_2|\vec{Y}), \hat{Y}_2)^p + \varepsilon \quad \gamma\text{-a.s.} \end{aligned}$$

The obvious choice for γ', namely $f \mapsto \mathbb{E}^{\gamma}\left(\mathbb{E}^{\Xi}(f(X_1,X_2,Y_1,Y_2))\right)$ will not work because in general it gets the relationship between X_1 and X_2 wrong, i.e. its first marginal may not be $\mathrm{int}_{\mathcal{X}_1}(\mu)$. Instead we again define $\gamma_L \in \mathcal{P}(\mathcal{X}_1 \times \mathcal{X}_2 \times \mathcal{Y}_1)$ and $\gamma_R \in \mathcal{P}(\mathcal{X}_2 \times \mathcal{Y}_1 \times \mathcal{Y}_2)$ and set $\gamma' := \gamma_L \overset{\otimes}{}_{\mathcal{X}_2,\mathcal{Y}_1} \gamma_R$.

$$\begin{aligned} \gamma_L &:= f \mapsto \mathbb{E}^{\gamma}\left(\mathbb{E}^{\hat{X}_2}(f(X_1,X_2,Y_1))\right) \\ \gamma_R &:= f \mapsto \mathbb{E}^{\gamma}\left(\mathbb{E}^{\Xi}(f(X_2,Y_1,Y_2))\right) \end{aligned}$$

Clearly, if we can actually define γ' as announced, then (30) will hold, because then

$$\begin{aligned} \mathbb{E}^{\gamma'}(\rho(X_1,Y_1)) &= \mathbb{E}^{\gamma}\left(\mathbb{E}^{\hat{X}_2}(\rho(X_1,Y_1))\right) = \mathbb{E}^{\gamma}(\rho(X_1,Y_1)) \\ \mathbb{E}^{\gamma'}(\rho(X_2,Y_2)) &= \mathbb{E}^{\gamma}\left(\mathbb{E}^{\Xi}(\rho(X_2,Y_2))\right) \leq \mathbb{E}^{\gamma}(\mathcal{W}_p(\mathbb{E}^{\gamma}(\hat{X}_2|\vec{Y}), \hat{Y}_2)^p) + \varepsilon \ . \end{aligned}$$

It remains to check that γ_L and γ_R can actually be composed, i.e. that (X_2, Y_1) has the same distribution under γ_L and γ_R.

$$\begin{aligned} \mathbb{E}^{\gamma_R}(h(X_2,Y_1)) &= \mathbb{E}^{\gamma}\left(\mathbb{E}^{\Xi}(h(X_2,Y_1))\right) = \mathbb{E}^{\gamma}\left(\mathbb{E}^{\mathbb{E}^{\gamma}(\hat{X}_2|\vec{Y})}(h(X_2,Y_1))\right) = \\ &\mathbb{E}^{\gamma}\left(\mathbb{E}^{\gamma}\left(\mathbb{E}^{\hat{X}_2}(h(X_2,Y_1))\,\middle|\,\vec{Y}\right)\right) = \mathbb{E}^{\gamma}(\mathbb{E}^{\hat{X}_2}(h(X_2,Y_1))) = \mathbb{E}^{\gamma_L}(h(X_2,Y_1)) \end{aligned}$$

The step in the middle has its own Lemma 4.21 below. □

Lemma 4.21. *Let $\mathbb{P}$ be a probability on $\mathcal{P}(\mathcal{X}) \times \mathcal{Y}$, for Polish spaces $\mathcal{X}, \mathcal{Y}$. Let $h : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ be a measurable function. Then*

$$\mathbb{E}^{\mathbb{E}(\hat{X}|Y)}\left(h(X,Y)\right) = \mathbb{E}\left(\mathbb{E}^{\hat{X}}\left(h(X,Y)\right)\Big|Y\right) \quad \mathbb{P}\text{-a.s.},$$

where $\mathbb{E}$ without superscript is the (conditional) expectiation w.r.t. $\mathbb{P}$ and $\hat{X}$ is the projection onto $\mathcal{P}(\mathcal{X})$.

Note that X is on both sides introduced by the expectation operator which carries a superscript, while Y may on both sides be interpreted as coming from the outermost context. On the right hand side Y may also be seen as having been introduced by the outermost conditional expectation operator. (As this operator conditions on Y this is the same thing.)

Proof. Both sides are clearly Y-measurable. We prove that for $h(x,y) = f(x)g_1(y)$, multiplying by $g_2(Y)$ and taking expectation gives the same result. By definition of the conditional expectation

$$\mathbb{E}\left(\mathbb{E}\left(\hat{X}|Y\right)g(Y)\right) = \mathbb{E}\left(\hat{X}g(Y)\right) \ .$$

Applying the continuous linear function $\gamma \mapsto \mathbb{E}^{\gamma}\left(f(X)\right)$ this gives

$$\mathbb{E}\left(\mathbb{E}^{\mathbb{E}(\hat{X}|Y)}\left(f(X)\right)g(Y)\right) = \mathbb{E}\left(\mathbb{E}^{\hat{X}}\left(f(X)\right)g(Y)\right) \ .$$

Again by the definition of the conditional expectation:

$$\mathbb{E}\left(\mathbb{E}\left(\mathbb{E}^{\hat{X}}\left(f(X)g_1(Y)\right)\Big|Y\right)g_2(Y)\right) = \mathbb{E}\left(\mathbb{E}^{\hat{X}}\left(f(X)g_1(Y)\right)g_2(Y)\right) =$$
$$\mathbb{E}\left(\mathbb{E}^{\hat{X}}\left(f(X)\right)g_1(Y)g_2(Y)\right) = \mathbb{E}\left(\mathbb{E}^{\mathbb{E}(\hat{X}|Y)}\left(f(X)\right)g_1(Y)g_2(Y)\right) =$$
$$\mathbb{E}\left(\mathbb{E}^{\mathbb{E}(\hat{X}|Y)}\left(f(X)g_1(Y)\right)g_2(Y)\right)$$

where for the third equality we plugged in the previous equation. □

Alternative proof of Lemma 4.20 when $\mathcal{X}_1$ has no isolated points. When the space $\mathcal{X}_1$ has no isolated points one can show that the space $\mathcal{F}\left(\mathcal{X}_1 \rightsquigarrow \mathcal{P}(\mathcal{X}_2)\right)$ is dense in $\mathcal{P}(\mathcal{X}_1 \times \mathcal{P}(\mathcal{X}_2))$. This allows for a shorter proof of Lemma 4.20:

By the original definition (20) of $\mathcal{CW}_p$ on the space $\mathcal{P}(\mathcal{X}_1 \times \mathcal{X}_2)$ the inequality (29) holds on $\mathcal{F}\left(\mathcal{X}_1 \rightsquigarrow \mathcal{P}(\mathcal{X}_2)\right) \times \mathcal{F}\left(\mathcal{X}_1 \rightsquigarrow \mathcal{P}(\mathcal{X}_2)\right)$. Both $\mathcal{CW}_p$ and $(\mu,\nu) \mapsto \mathcal{W}_p\left(\mathrm{int}_{\mathcal{X}_1}(\mu), \mathrm{int}_{\mathcal{X}_1}(\nu)\right)$ are uniformly continuous on $\mathcal{P}(\vec{\mathcal{X}}) \times \mathcal{P}(\vec{\mathcal{X}})$ w.r.t. some product metric of $\overline{\mathcal{AW}}_p$ with itself. $\mathcal{F}\left(\mathcal{X}_1 \rightsquigarrow \mathcal{P}(\mathcal{X}_2)\right)$ is dense in $\mathcal{P}(\vec{\mathcal{X}})$, and therefore $\mathcal{F}\left(\mathcal{X}_1 \rightsquigarrow \mathcal{P}(\mathcal{X}_2)\right) \times \mathcal{F}\left(\mathcal{X}_1 \rightsquigarrow \mathcal{P}(\mathcal{X}_2)\right)$ is dense in $\mathcal{P}(\vec{\mathcal{X}}) \times \mathcal{P}(\vec{\mathcal{X}})$. This implies that (29) holds on all of $\mathcal{P}(\vec{\mathcal{X}}) \times \mathcal{P}(\vec{\mathcal{X}})$. □

Theorem 4.22. *The topology induced by $\overline{\mathcal{SCW}}_p$ on $\mathcal{P}(\mathcal{X}_1 \times \mathcal{P}(\mathcal{X}_2))$ is equal to the toplogy induced by $\overline{\mathcal{AW}}_p$ on that space.*

Proof. As both topologies are metric and therefore first-countable we may argue on sequences. Let $(\mu_n)_n$ be a sequence in $\mathcal{P}(\mathcal{X}_1 \times \mathcal{P}(\mathcal{X}_2))$. As $\overline{\mathcal{SCW}}_p(\mu_n, \mu) \leq \overline{\mathcal{AW}}_p(\mu_n, \mu)$, if $(\mu_n)_n$ converges to μ w.r.t. $\overline{\mathcal{AW}}_p$ it also converges to μ w.r.t. $\overline{\mathcal{SCW}}_p$.

Now assume that a sequence $(\mu_n)_n$ in $\mathcal{P}(\mathcal{X}_1 \times \mathcal{P}(\mathcal{X}_2))$ converges to μ w.r.t. $\overline{\mathcal{SCW}}_p$. We will show that every subsequence of $(\mu_n)_n$ has a subsequence which converges to μ w.r.t. $\overline{\mathcal{AW}}_p$. Our assumption implies that the set $K := \{\mu_n \,|\, n \in \mathbb{N}\}$ is relatively compact. As $\mathrm{int}_{\mathcal{X}_1}$ is continuous as a map from $\mathcal{P}(\mathcal{X}_1 \times \mathcal{P}(\mathcal{X}_2))$ with the topology induced by $\overline{\mathcal{SCW}}_p$ to $\mathcal{P}(\mathcal{X}_1 \times \mathcal{X}_2)$ with the toplogy induced by $\mathcal{W}_p$ (Lemma 4.20), we have that $\mathrm{int}_{\mathcal{X}_1}[K] = \{\mathrm{int}_{\mathcal{X}_1}(\mu_n) \,|\, n \in \mathbb{N}\}$ is also relatively compact. By Lemma 1.6/[23, Lemma 3.3] this implies that K is relatively compact in $\mathcal{P}(\mathcal{X}_1 \times \mathcal{P}(\mathcal{X}_2))$ with the topology induced by $\overline{\mathcal{AW}}_p$. Now let $(\mu_{n_k})_k$ be some

subsequence of $(\mu_n)_n$. As K is relatively compact we can find a subsequence $(\mu_{n_{k_j}})_j$ of $(\mu_{n_k})_k$, which converges w.r.t. $\overline{\mathcal{AW}}_p$ to some $\mu' \in \mathcal{P}(\mathcal{X}_1 \times \mathcal{P}(\mathcal{X}_2))$. As $\overline{\mathcal{SCW}}_p\left(\mu_{n_{k_j}}, \mu'\right) \leq \overline{\mathcal{AW}}_p\left(\mu_{n_{k_j}}, \mu'\right)$ this sequence also converges to μ' w.r.t. $\overline{\mathcal{SCW}}_p$. But $(\mu_{n_{k_j}})_j$ also converges to μ w.r.t. $\overline{\mathcal{SCW}}_p$. Because the topology induced by $\overline{\mathcal{SCW}}_p$ is Hausdorff (Lemma 4.14), we must have $\mu' = \mu$, i.e. $(\mu_{n_{k_j}})_j$ converges to μ w.r.t. $\overline{\mathcal{AW}}_p$. □

Now we return to the general case of N time-points.

Theorem 4.23. *The topology induced by $\mathcal{SCW}_p$ on $\mathcal{P}(\overline{\mathcal{X}})$ is equal to Hellwig's $\mathcal{W}_p$-information topology and to the topology induced by $\mathcal{AW}_p$.*

Proof. As every bicausal transport plan between μ and ν can be interpreted as a causal transport plan from μ to ν and also as a causal transport plan from ν to μ we have that $\mathcal{SCW}_p(\mu, \nu) \leq \mathcal{AW}_p(\mu, \nu)$. This means that the identity from $\mathcal{P}(\overline{\mathcal{X}})$ with the topology induced by $\mathcal{AW}_p$ to $\mathcal{P}(\overline{\mathcal{X}})$ with the topology induced by $\mathcal{SCW}_p$ is continuous. For the other direction we show that the identity from $\mathcal{P}(\overline{\mathcal{X}})$ with the topology induced by $\mathcal{SCW}_p$ to $\mathcal{P}(\overline{\mathcal{X}})$ with the $\mathcal{W}_p$-information topology is continuous, i.e. we show that each of the maps

$$\mathrm{dis}_{\overline{\mathcal{X}}^t}^{\overline{\mathcal{X}}_{t+1}} = \mathcal{I}_t : \mathcal{P}(\overline{\mathcal{X}}) \to \mathcal{F}\left(\overline{\mathcal{X}}^t \rightsquigarrow \mathcal{P}(\overline{\mathcal{X}}_{t+1})\right)$$

is continuous when $\mathcal{P}(\overline{\mathcal{X}})$ gets the topology induced by $\mathcal{SCW}_p$.

If $\mu, \nu \in \mathcal{P}(\overline{\mathcal{X}})$ and $\gamma \in \mathrm{Cpl}(\mu, \nu)$ is causal, then, in particular, γ is 'causal at the timestep from t to $t+1$', i.e. γ is causal when regarded as a coupling between $\mu, \nu \in \mathcal{P}(\overline{\mathcal{X}}^t \times \overline{\mathcal{X}}_{t+1})$. This means that if we define $\mathcal{SCW}'_p$ like $\mathcal{SCW}_p$, but only require causality based on the decomposition of $\overline{\mathcal{X}}$ as $\overline{\mathcal{X}}^t \times \overline{\mathcal{X}}_{t+1}$, then $\mathcal{SCW}'_p(\mu, \nu) \leq \mathcal{SCW}_p(\mu, \nu)$, i.e. the identity from $\mathcal{P}(\overline{\mathcal{X}})$ with the topology induced by $\mathcal{SCW}_p$ to $\mathcal{P}(\overline{\mathcal{X}})$ with the topology induced by $\mathcal{SCW}'_p$ is continuous. By Theorem 4.22 the map

$$\mathrm{dis}_{\overline{\mathcal{X}}^t}^{\overline{\mathcal{X}}_{t+1}} : \mathcal{P}(\overline{\mathcal{X}}^t \times \overline{\mathcal{X}}_{t+1}) \to \mathcal{F}\left(\overline{\mathcal{X}}^t \rightsquigarrow \mathcal{P}(\overline{\mathcal{X}}_{t+1})\right)$$

is continuous when we equip $\mathcal{P}(\overline{\mathcal{X}}^t \times \overline{\mathcal{X}}_{t+1})$ with the topology induced by $\mathcal{SCW}'_p$. Now $\mathcal{I}_t$ is continuous as a composite of continuous maps. □

5. Aldous' extended weak convergence

In this section we show that Aldous extended $\mathcal{W}_p$-/weak topology is equal to Hellwig's ($\mathcal{W}_p$-)information topology.

We recall and paraphrase here the definition, already given in the introduction, of Aldous' topology.

Definition 5.1. Given $\mu \in \mathcal{P}(\overline{\mathcal{X}})$ let $\mu_{(x_i)_{i=1}^j}$ be the value of a (classical) disintegration of μ w.r.t. the first j coordinates at $(x_i)_{i=1}^j$. (By convention $\mu_{(x_i)_{i=1}^0} = \mu$). Define

$$\mathcal{E} : \mathcal{P}(\overline{\mathcal{X}}) \to \mathcal{P}\left(\overline{\mathcal{X}} \times \prod_{j=0}^{N} \mathcal{P}(\overline{\mathcal{X}})\right)$$

$$\mathcal{E}(\mu) := \mathcal{P}\left((x_i)_{i=1}^N \mapsto \left((x_i)_{i=1}^N, \left(\delta_{(x_i)_{i=1}^j} \otimes \mu_{(x_i)_{i=1}^j}\right)_{j=0}^N\right)\right)(\mu)\ .$$

The *extended $\mathcal{W}_p$−/weak topology* on $\mathcal{P}(\overline{\mathcal{X}})$ is the initial topology w.r.t. $\mathcal{E}$.

Remark 5.2. Reasonable people may disagree about whether the most faithful / useful transcription of Aldous' definition should include the factors $j = 0$ and $j = N$ in the above product of spaces. When including $j = N$, as we did, one has to interpret $\delta_{(x_i)_{i=1}^N} \otimes \mu_{(x_i)_{i=1}^N}$ simply as $\delta_{(x_i)_{i=1}^N}$. We leave it as an exercise to the reader to check that either or both may be dropped in the definition of $\mathcal{E}$ without affecting the resulting topology on $\mathcal{P}(X)$.

Theorem 5.3. *The ($\mathcal{W}_p$-)extended weak topology is equal to the ($\mathcal{W}_p$-)information topology.*

Proof. We construct continuous maps

$$\mathcal{A}'_k : \mathcal{P}\left(\overline{\mathcal{X}} \times \prod_{j=0}^{N} \mathcal{P}\big(\overline{\mathcal{X}}\big)\right) \to \mathcal{P}\big(\overline{\mathcal{X}}^k \times \mathcal{P}\big(\overline{\mathcal{X}}_{k+1}\big)\big)$$
$$\mathcal{A} : \prod_{k=1}^{N-1} \mathcal{F}\left(\overline{\mathcal{X}}^k \rightsquigarrow \mathcal{P}\big(\overline{\mathcal{X}}_{k+1}\big)\right) \to \mathcal{P}\left(\overline{\mathcal{X}} \times \prod_{j=0}^{N} \mathcal{P}\big(\overline{\mathcal{X}}\big)\right)$$

such that

$$\mathcal{A}'_k \circ \mathcal{E} = \mathcal{I}_k$$
$$\mathcal{A} \circ \mathcal{I} = \mathcal{E} \ .$$

The first equality above implies that the identity on $\mathcal{P}\big(\overline{\mathcal{X}}\big)$ is continuous from the extended weak topology to the information topology, the second implies that it is continuous in the other direction.

$\mathcal{A}'_k$ is very simple. We just need to select the right factors and then discard the unnecessary $\delta_{(x_i)_{i=1}^k}$ part of the measure component. Formally

$$\mathcal{A}'_k := \mathcal{P}\Big(\big((x_i)_{i=1}^N, (\nu_j)_{j=0}^N\big) \mapsto \big((x_i)_{i=1}^k, \nu_{k\,|\,\overline{\mathcal{X}}_{k+1}}\big)\Big) \ ,$$

which is cleary continuous.

We construct $\mathcal{A}$ recursively, by constructing as a composite of continuous maps

$$\mathcal{A}^m : \prod_{k=1}^{N-1} \mathcal{F}\left(\overline{\mathcal{X}}^k \rightsquigarrow \mathcal{P}\big(\overline{\mathcal{X}}_{k+1}\big)\right) \to \mathcal{P}\left(\overline{\mathcal{X}}^m \times \prod_{j=0}^{m} \mathcal{P}\big(\overline{\mathcal{X}}\big)\right)$$

satisfying

$$\mathcal{A}^m(\mathcal{I}(\mu)) = \mathcal{P}\Big((x_i)_{i=1}^N \mapsto \big((x_i)_{i=1}^m, (\delta_{(x_i)_{i=1}^k} \otimes \mu_{(x_i)_{i=1}^k})_{k=0}^m\big)\Big)(\mu) \ . \tag{31}$$

$\mathcal{A}^0\left((\nu_k)_{k=1}^{N-1}\right) := \delta_{\mathrm{int}_{\mathcal{X}_1}(\nu_1)}$. We need the helper functions

$$h_m : \mathcal{F}\left(\overline{\mathcal{X}}^m \rightsquigarrow \mathcal{P}\big(\overline{\mathcal{X}}_{m+1}\big)\right) \to \mathcal{F}\left(\overline{\mathcal{X}}^m \rightsquigarrow \mathcal{P}\big(\overline{\mathcal{X}}\big)\right)$$
$$h_m := \mathcal{P}\Big(\big((x_i)_{i=1}^m, \rho\big) \mapsto \big((x_i)_{i=1}^m, \delta_{(x_i)_{i=1}^m} \otimes \rho\big)\Big) \ .$$

Given $\mathcal{A}^m$ satisfying the induction hypothesis we set

$$\mathcal{A}^{m+1}\left((\nu_k)_{k=1}^{N-1}\right) := \mathcal{P}(s_{m+1})\left(\mathcal{A}^m\left((\nu_k)_{k=1}^{N-1}\right) \underset{\overline{\mathcal{X}}^m}{\otimes} h_{m+1}(\nu_{m+1})\right)$$

where s_{m+1} is the obvious permutation of the coordinates to get the factors into the right order. $\mathcal{A}^{m+1}$ is continuous because by [23, Theorem 4.1] $\underset{\overline{\mathcal{X}}^m}{\otimes}$ is continuous when one of the arguments is an element of some $\mathcal{F}(\mathcal{B} \rightsquigarrow \mathcal{C})$. That (31) still holds for $m+1$ is a straightforward calculation. This way we get to $\mathcal{A}^{N-1}$. Finally, set

$$\mathcal{A}\left((\nu_k)_{k=1}^{N-1}\right) := \mathcal{P}(s_N)\left(\mathcal{A}^{N-1}\left((\nu_k)_{k=1}^{N-1}\right) \underset{\overline{\mathcal{X}}^{N-1}}{\otimes} \mathrm{dis}_{\mathcal{X}_1}(\nu_1)\right)$$

where

$$s_N\left(((x_i)_{i=1}^{N-1},(\rho_j)_{j=1}^{N-1},x_N)\right) := \left((x_i)_{i=1}^{N},(\rho_j)_{j=1}^{N-1},\delta_{((x_i)_{i=1}^{N})}\right) .$$

□

6. Bounded vs unbounded metrics

Because we will need it in the next section we interject here a proof of Lemma 1.3, which we restate below.

Lemma 1.3. *Convergence in any of the topologies of Theorem 1.2 is equivalent to convergence in any of the topologies of Theorem 1.1 (where for building $\mathcal{SCW}_p$ and $\mathcal{AW}_p$, $\rho_{\mathcal{X}}$ is replaced by a bounded compatible complete metric e.g. $\min(1,\rho_{\mathcal{X}})$) plus convergence of p-th moments on $\overline{\mathcal{X}}$ w.r.t. (the original) $\rho_{\overline{\mathcal{X}}}$.*

Proof of Lemma 1.3. We provide the proof only for Hellwig's topology, i.e. (3) of Theorem 1.2 and Theorem 1.1, respectively. As we have already seen in the previous sections, the topologies (2)–(4) are equivalent topologies, and the result therefore carries over to them. The ($\mathcal{W}_p$-)optimal stopping topology, (5), is treated below. It is clear that convergence w.r.t. $\mathcal{W}_p$-information topology implies convergence in Hellwig'g information topology plus convergence of p-th moments. For the reverse implication, let $1 \leq t \leq N-1$, and denote by $\mathcal{A} := \overline{\mathcal{X}}^t$ the first t and by $\mathcal{B} := \overline{\mathcal{X}}_{t+1}$ the last $N-t$ coordinates. Now assume that $(\mu_n)_n$ converges to μ in Hellwig's information topology and that the p-th moments converge. The classical (not adapted) version of the very lemma we prove here implies that $\mu_n \to \mu$ in $\mathcal{W}_p$; in particular $K := \{\mu_n : n\} \subset \mathcal{P}_p(\mathcal{A}\times\mathcal{B})$ is relatively compact. Lemma 1.6 (or really [23, Lemma 3.3]/[8, Lemma 2.6]) therefore guarantees that $\mathrm{dis}_{\mathcal{A}}^{\mathcal{B}}[K] \subset \mathcal{P}_p(\mathcal{A}\times\mathcal{P}_p(\mathcal{B}))$ is relatively compact.

Every subsequence of $(\mathrm{dis}_{\mathcal{A}}^{\mathcal{B}}(\mu_n))_n$ therefore has a subsequence $(\mathrm{dis}_{\mathcal{A}}^{\mathcal{B}}(\mu_{n_k}))_k$ which converges w.r.t. the topology on $\mathcal{P}_p(\mathcal{A}\times\mathcal{P}_p(\mathcal{B}))$ (i.e. the one coming from nested Wasserstein metrics) to some $\mu' \in \mathcal{P}_p(\mathcal{A}\times\mathcal{P}_p(\mathcal{B}))$. Because convergence in $\mathcal{P}_p(\mathcal{A}\times\mathcal{P}_p(\mathcal{B}))$ is stronger than convergence in $\mathcal{P}(\mathcal{A}\times\mathcal{P}(\mathcal{B}))$ (i.e. in the nested *weak* sense) we must also have $\mathrm{dis}_{\mathcal{A}}^{\mathcal{B}}(\mu_{n_k}) \overset{k}{\to} \mu'$ in $\mathcal{P}(\mathcal{A}\times\mathcal{P}(\mathcal{B}))$. But also, by assumption, $\mathrm{dis}_{\mathcal{A}}^{\mathcal{B}}(\mu_{n_k}) \overset{k}{\to} \mathrm{dis}_{\mathcal{A}}^{\mathcal{B}}(\mu)$ in $\mathcal{P}(\mathcal{A}\times\mathcal{P}(\mathcal{B}))$ and therefore $\mu' = \mathrm{dis}_{\mathcal{A}}^{\mathcal{B}}(\mu)$. □

7. Optimal Stopping

In this section we investigate the relation between the ($\mathcal{W}_p$-)optimal stopping topology and the adapted Wasserstein topology. Lemma 7.1 states that the topology induced by $\mathcal{AW}_p$ ((1) of Theorem 1.2) is finer than the $\mathcal{W}_p$-optimal stopping topology. Lemma 7.5 states that the $\mathcal{W}_p$-optimal stopping topology is finer than the $\mathcal{W}_p$-information topology ((3) of Theorem 1.2). This will finish the proof of Theorem 1.2.

Recall that

$$v^L(\mu) := \inf\left\{\mathbb{E}^\mu\left(L_\tau(X)\right)) : 0 \leq \tau \leq N \text{ is a stopping time}\right\}$$

for $L = (L_t)_{t=0}^N \in AC_p(\Omega)$.

Lemma 7.1. *Let $L \in AC_p(\Omega)$. Then $\mu \mapsto v^L(\mu)$ is continuous w.r.t. $\mathcal{AW}_p$. In fact, one has*

$$|v^L(\mu) - v^L(\nu)| \leq \inf\left\{\mathbb{E}^\pi\left(\max_{0\leq t\leq N}|L_t(X)-L_t(Y)|\right) : \pi \in \mathrm{Cpl}_{bc}(\mu,\nu)\right\}. \quad (32)$$

for every $\mu,\nu \in \mathcal{P}_p(\Omega)$.

Proof. Let $\mu, \nu \in \mathcal{P}_p(\Omega)$ and assume that $v^L(\mu) \leq v^L(\nu)$. Moreover, let $\pi \in \mathrm{Cpl}_{bc}(\mu,\nu)$ and $\varepsilon > 0$ be arbitrary, and fix a stopping time τ satisfying $\mathbb{E}^\nu\left(L_\tau(Y)\right) \leq v^L(\nu) + \varepsilon$. For $u \in [0,1]$ define

$$\begin{aligned}\sigma(X,u) &:= \inf\{t \in \{0,\cdots,T\} : \pi(\tau(Y) \leq t | X) \geq u\} \\ &= \inf\{t \in \{0,\cdots,T\} : \pi(\tau(Y) \leq t | X_1,\dots,X_t) \geq u\},\end{aligned}$$

where the equality holds by the properties of stopping times and since π is causal. We then have that

$$\begin{aligned}\int_{[0,1]} \mathbb{E}^\pi\left(L_{\sigma(X,u)}(X)\right) \mathrm{d}u &= \sum_{t=0}^{T} \int_{[0,1]} \mathbb{E}^\pi\left(L_t(X) 1_{\pi(\tau(Y)\leq t|X) \geq u > \pi(\tau(Y)\leq t-1|X)}\right) \mathrm{d}u \\ &= \sum_{t=0}^{T} \mathbb{E}^\pi\left(L_t(X) 1_{\tau(Y)=t}\right) = \mathbb{E}^\pi\left(L_{\tau(Y)}(X)\right).\end{aligned}$$

As further $\sigma(\cdot,u)$ is a stopping time for every fixed $u \in [0,1]$ one has $v^L(\mu) \leq \int_{[0,1]} \mathbb{E}^\pi\left(L_{\sigma(X,u)}(X)\right) \mathrm{d}u$ and therefore

$$\begin{aligned}v^L(\mu) - v^L(\nu) &\leq \mathbb{E}^\pi\left(L_{\tau(Y)}(X) - L_{\tau(Y)}(Y)\right) + \varepsilon \\ &\leq \mathbb{E}^\pi\left(\max_{0\leq t\leq N} |L_t(X) - L_t(Y)|\right) + \varepsilon.\end{aligned}$$

Changing the role of μ and ν and using that $\varepsilon > 0$ and $\pi \in \mathrm{Cpl}_{bc}(\mu,\nu)$ was arbitrary yields (32).

Now assume that $\mathcal{AW}_p(\mu_n,\mu) \to 0$ and that $\pi_n \in \mathrm{Cpl}\,(\mu_n,\mu)$ is less than $1/n$ away from attaining the infimum $\mathcal{AW}_p(\mu_n,\mu)$. Then $\mathcal{W}_p(\pi_n,\pi) \to 0$, where $\pi \in \mathrm{Cpl}\,(\mu,\mu)$ is the identity coupling $\mathcal{P}(1_\Omega, 1_\Omega)(\mu)$ of μ. (A coupling between π_n and π is given by $\mathcal{P}((x,y) \mapsto (x,y,y,y))(\pi_n)$.) Because $(x,y) \mapsto \max_{0\leq t\leq N} |L_t(x) - L_t(y)|$ is a continuous function of growth of at most order p, we get that

$$\mathbb{E}^{\pi_n}\left(\max_{0\leq t\leq N} |L_t(X) - L_t(Y)|\right) \to \mathbb{E}^{\pi}\left(\max_{0\leq t\leq N} |L_t(X) - L_t(Y)|\right) = 0\,.$$

Together with (32) this implies that v^L is continuous w.r.t. $\mathcal{AW}_p$. □

Remark 7.2. The above proof reveals that if L_t is Lipschitz with constant $c > 0$ for every t, then $|v^L(\mu) - v^L(\nu)| \leq c\,\mathcal{SCW}_1(\mu,\nu)$.

In order to show that the optimal stopping topology is finer than the $\mathcal{W}_p$-information topology, we need to make a few preparations.

Lemma 7.3. *Let $\mathcal{A}$ be a Polish space. Then the family*

$$\left\{\mathcal{P}(\mathcal{A}) \ni \mu \mapsto G\left(\int_{\mathcal{A}} h_1 \,\mathrm{d}\mu, \dots, \int_{\mathcal{A}} h_L \,\mathrm{d}\mu\right) : \begin{array}{l} L \in \mathbb{N}, G \in C_b(\mathbb{R}^L) \\ (h_i)_{i\leq L} \subset C_b(\mathcal{A}) \end{array}\right\} \tag{33}$$

is convergence determining for the weak topology on $\mathcal{P}(\mathcal{P}(\mathcal{A}))$, that is, a sequence of probability measures $(\mu_n)_n$ in $\mathcal{P}(\mathcal{P}(\mathcal{A}))$ converges weakly to a probability measure $\mu \in \mathcal{P}(\mathcal{P}(\mathcal{A}))$ if and only if $\int F \,\mathrm{d}\mu_n \to \int F \,\mathrm{d}\mu$ for all F in (33).

This follows from the Stone-Weierstrass theorem in case of compact $\mathcal{A}$ and readily extends to general Polish spaces e.g. via Stone-Čech compactification.

Lemma 7.4. *Let $\mathcal{A}$ be a Polish space. The family of functions*

$$\left\{\mu \mapsto G\left(\int_{\mathcal{A}} h \,\mathrm{d}\mu\right) : h \in C_b(\mathcal{A}), G \in C_b(\mathbb{R})\right\} \tag{34}$$

is convergence determining for the weak topology on $\mathcal{P}(\mathcal{P}(\mathcal{A}))$.

Proof. Let L, G, and $(h_i)_{i\leq L}$ as in (33). Moreover, let $m \in \mathbb{R}$ such that $|h_i| \leq m$ for all $1 \leq i \leq L$ and define $I := [-m, m]^L$. Then $I \subset \mathbb{R}^L$ is compact and satisfies

$$\left(\int h_1 d\mu, \ldots, \int h_L \, \mathrm{d}\mu \right) \in I \quad \text{for all } \mu \in \mathcal{P}(\mathcal{A}) \,.$$

Let $\sigma \colon \mathbb{R} \to \mathbb{R}$ be some fixed bounded continuous sigmoid function such as $\sigma(r) = (1 + e^{-r})^{-1}$ or $\sigma(r) = \max(0, \min(r, 1))$.

By the universal approximation result of Cybenko [20, Theorem 2], the set

$$\left\{ x \mapsto \sum_{i=1}^{m} u_i \sigma(v_i \cdot x + w_i) : \begin{array}{l} m \in \mathbb{N}, (u_i)_{i\leq m} \subset \mathbb{R}, \\ (v_i)_{i\leq m} \subset \mathbb{R}^L, (w_i)_{i\leq m} \subset \mathbb{R} \end{array} \right\}$$

is dense in $C(I, \mathbb{R})$ w.r.t. the supremum norm. As a result, it is enough to replace G in (33) by functions of the form $x \mapsto \sum_{i=1}^{m} u_i \sigma(v_i \cdot x + w_i)$. Evaluating the latter function on the vector $x = (\int h_1 \, \mathrm{d}\mu, \ldots, \int h_L \, \mathrm{d}\mu)$ yields

$$\begin{aligned} \sum_{i=1}^{m} u_i \sigma \left(\sum_{k=1}^{L} v_i^k \int h_k \, \mathrm{d}\mu + w_i \right) &= \sum_{i=1}^{m} u_i \sigma \left(\int \left(\sum_{k=1}^{L+1} v_i^k h_k \right) \mathrm{d}\mu \right) \\ &= \sum_{i=1}^{m} u_i \sigma \left(\int \bar{h}_i \, \mathrm{d}\mu \right), \end{aligned}$$

upon defining $v_i^{L+1} := b_i$, $w_{L+1} := 1$, and finally $\bar{h}_i := \sum_{k=1}^{L+1} v_i^k h_k$ for every i. The result follows from Lemma 7.3. □

Lemma 7.5. *The $\mathcal{W}_p$-optimal stopping topology is finer than the $\mathcal{W}_p$-information topology.*

Proof. The choice $L_T := -\rho(x, x_0)^p - 1$ and $L_t := 0$ for $t \neq T$ shows that convergence in the $\mathcal{W}_p$-optimal stopping topology implies convergence of the p-th moments. Thus, we are left to show that convergence in the optimal stopping topology implies convergence in Hellwig's information topology. Then, by the part of Lemma 1.3 which has already been established, we obtain convergence in the $\mathcal{W}_p$-information topology.

Fix $1 \leq t \leq N-1$ and denote by $\mathcal{A} := \overline{\mathcal{X}}^t$ the first t and by $\mathcal{B} := \overline{\mathcal{X}}_{t+1}$ the last $N - t$ coordinates. As $C_b(\mathcal{A})$ is convergence determining for $\mathcal{P}(\mathcal{A})$, and $\{\nu \mapsto G(\int_{\mathcal{B}} h \, \mathrm{d}\nu) : h \in C_b(\mathcal{B}), G \in C_b(\mathbb{R})\}$ is, by Lemma 7.4, convergence determining for $\mathcal{P}(\mathcal{P}(\mathcal{B}))$, it follows e.g. from [25, Proposition 4.6 (p.115)] that

$$\left\{ (a, \nu) \mapsto f(a) g \left(\int_{\mathcal{B}} h(b) \, \mathrm{d}\nu(b) \right) : f \in C_b(\mathcal{A}),\, g \in C_b(\mathbb{R}),\, h \in C_b(\mathcal{B}) \right\}, \tag{35}$$

is convergence determining for the weak topology on $\mathcal{P}(\mathcal{A} \times \mathcal{P}(\mathcal{B}))$. Since h in (35) is bounded, one can actually take g in (35) to be compactly supported. But a continuous compactly supported function can be approximated uniformly by piecewise linear functions. The latter are linear combinations of functions of the form $z \mapsto \min(c, dz)$ where $c, d \in \mathbb{R}$. It therefore follows that

$$\left\{ (a, \nu) \mapsto \min \left(f(a) \,, \int_{\mathcal{B}} f(a) h(b) \, \mathrm{d}\nu(b) \right) : f \in C_b(\mathcal{A}), h \in C_b(\mathcal{B}) \right\}, \tag{36}$$

is also convergence determining for the weak topology on $\mathcal{P}(\mathcal{A} \times \mathcal{P}(\mathcal{B}))$. Let F be a function in (36), defined via $f \in C_b(\mathcal{A})$ and $h \in C_b(\mathcal{B})$, and let $m \in \mathbb{R}$ be a bound for $|f|$ and $|h|$. Define $L \in AC_p(\Omega)$ via

$$L_t := f \circ \overline{X}^t \quad L_T := (f \circ \overline{X}^t) \cdot (h \circ \overline{X}_{t+1}) \quad \text{and } L_s := m + 1 \text{ for } s \neq t, T.$$

(Where $\overline{X}^t$ is the projection onto the first t coordinates and $\overline{X}_{t+1}$ is the projection onto the remaining $N - t$ coordinates.)

By dynamic programming (the Snell-envelope theorem) one has

$$\begin{aligned} v^L(\mu) &= \mathbb{E}^\mu \left(\min \left(f(\overline{X}^t), \mathbb{E}^\mu \left(f(\overline{X}^t) h(\overline{X}_{t+1}) | \overline{X}^t \right) \right) \right) \\ &= \int_{\mathcal{A} \times \mathcal{P}(\mathcal{B})} F \, \mathrm{d}(\mathrm{dis}_{\mathcal{A}}^{\mathcal{B}}(\mu)) \end{aligned}$$

for every $\mu \in \mathcal{P}(\mathcal{A} \times \mathcal{B})$. This implies that the optimal stopping topology is finer than the initial topology of $\mu \mapsto \int F \, \mathrm{d}(\mathrm{dis}_{\mathcal{A}}^{\mathcal{B}}(\mu))$ over F in (36). As (36) is convergence determining for the weak topology on $\mathcal{P}(\mathcal{A} \times \mathcal{P}(\mathcal{B}))$, the optimal stopping topology is indeed finer than the information topology, and as observed at the beginning of this proof therefore the $\mathcal{W}_p$-optimal stopping topology is finer than the $\mathcal{W}_p$-information topology. □

Acknowledgements

D. Bartl has been funded by the Vienna Science and Technology Fund (WWTF) through project VRG17-005. M. Beiglboeck and M. Eder gratefully acknowledge financial support by the FWF through grant Y782. J. Backhoff gratefully acknowledges financial support from the Austrian Science Fund (FWF) under grant P30750, as well as the Vienna University of Technology.

References

[1] B. Acciaio, J. Backhoff-Veraguas, and R. Carmona. Extended mean field control problems: stochastic maximum principle and transport perspective. *arXiv preprint arXiv:1802.05754*, 2018.

[2] B. Acciaio, J. Backhoff-Veraguas, and A. Zalashko. Causal optimal transport and its links to enlargement of filtrations and continuous-time stochastic optimization. *ArXiv e-prints*, 2016.

[3] D. J. Aldous. Weak convergence and general theory of processes. Unpublished incomplete draft of monograph; Department of Statistics, University of California, Berkeley, CA 94720, July 1981.

[4] J. Backhoff-Veraguas, D. Bartl, M. Beiglböck, and M. Eder. Adapted Wasserstein Distances and Stability in Mathematical Finance. *arXiv e-prints*, page arXiv:1901.07450, Jan 2019.

[5] J. Backhoff-Veraguas, M. Beiglböck, M. Eder, and A. Pichler. Fundamental properties of process distances. *ArXiv e-prints*, 2017.

[6] J. Backhoff-Veraguas, M. Beiglböck, M. Huesmann, and S. Källblad. Martingale Benamou–Brenier: a probabilistic perspective. *ArXiv e-prints*, Aug. 2017.

[7] J. Backhoff-Veraguas, M. Beiglböck, Y. Lin, and A. Zalashko. Causal transport in discrete time and applications. *SIAM Journal on Optimization*, 27(4):2528–2562, 2017.

[8] J. Backhoff Veraguas, M. Beiglböck, and G. Pammer. Existence, Duality, and Cyclical monotonicity for weak transport costs. *arXiv e-prints*, page arXiv:1809.05893, Sep 2018.

[9] J. Backhoff-Veraguas and G. Pammer. Stability of martingale optimal transport and weak optimal transport. *arXiv e-prints*, page arXiv:1904.04171, Apr 2019.

[10] M. Barbie and A. Gupta. The topology of information on the space of probability measures over Polish spaces. *Journal of Mathematical Economics*, 52(C):98–111, 2014.

[11] M. Beiglböck, A. Cox, and M. Huesmann. Optimal transport and Skorokhod embedding. *Invent. Math.*, 208(2):327–400, 2017.

[12] M. Beiglböck, P. Henry-Labordère, and F. Penkner. Model-independent bounds for option prices: A mass transport approach. *Finance Stoch.*, 17(3):477–501, 2013.

[13] M. Beiglböck, M. Nutz, and N. Touzi. Complete Duality for Martingale Optimal Transport on the Line. *Ann. Probab., to appear*, 2016.

[14] D. P. Bertsekas and S. E. Shreve. *Stochastic optimal control*, volume 139 of *Mathematics in Science and Engineering*. Academic Press, Inc. [Harcourt Brace Jovanovich, Publishers], New York-London, 1978. The discrete time case.

[15] J. Bion-Nadal and D. Talay. On a Wasserstein-type distance between solutions to stochastic differential equations. *Ann. Appl. Probab.*, 29(3):1609–1639, 2019.

[16] B. Bouchard and M. Nutz. Arbitrage and duality in nondominated discrete-time models. *The Annals of Applied Probability*, 25(2):823–859, 2015.

[17] L. Campi, I. Laachir, and C. Martini. Change of numeraire in the two-marginals martingale transport problem. *Finance Stoch.*, 21(2):471–486, June 2017.

[18] F. Coquet, J. Mémin, and L. Słominski. On weak convergence of filtrations. In *Séminaire de probabilités XXXV*, pages 306–328. Springer, 2001.
[19] F. Coquet and S. Toldo. Convergence of values in optimal stopping and convergence of optimal stopping times. *Electron. J. Probab.*, 12:no. 8, 207–228, 2007.
[20] G. Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, 2(4):303–314, 1989.
[21] C. Dellacherie and P.-A. Meyer. *Probabilities and Potential, A*, volume 29 of *North-Holland Mathematics Studies*. North-Holland Publishing Co., Amsterdam, 1978.
[22] Y. Dolinsky and H. M. Soner. Martingale optimal transport and robust hedging in continuous time. *Probab. Theory Relat. Fields*, 160(1-2):391–427, 2014.
[23] M. Eder. Compactness in adapted weak topologies. Preprint available at `https://www.mat.univie.ac.at/~eder/AdaptedCompactness`, 2019.
[24] M. Émery and W. Schachermayer. On Vershik's standardness criterion and Tsirelson's notion of cosiness. In *Séminaire de Probabilités, XXXV*, volume 1755 of *Lecture Notes in Math.*, pages 265–305. Springer, Berlin, 2001.
[25] S. N. Ethier and T. G. Kurtz. *Markov processes: characterization and convergence*, volume 282. John Wiley & Sons, 2009.
[26] A. Galichon, P. Henry-Labordère, and N. Touzi. A stochastic control approach to no-arbitrage bounds given marginals, with an application to lookback options. *Ann. Appl. Probab.*, 24(1):312–336, 2014.
[27] M. Giry. A categorical approach to probability theory. In B. Banaschewski, editor, *Categorical Aspects of Topology and Analysis*, pages 68–85. Springer, Berlin, Heidelberg, 1982. Lecture Notes in Mathematics, vol 915.
[28] M. Glanzer, G. Pflug, and A. Pichler. Incorporating statistical model error into the calculation of acceptability prices of contingent claims. *ArXiv e-prints*, 2017.
[29] M. F. Hellwig. Sequential decisions under uncertainty and the maximum theorem. *J. Math. Econom.*, 25(4):443–464, 1996.
[30] M. F. Hellwig and K. M. Schmidt. Discrete–time approximations of the holmström–milgrom brownian–motion model of intertemporal incentive provision. *Econometrica*, 70(6):2225–2264, 2002.
[31] D. Hobson and M. Klimmek. Robust price bounds for the forward starting straddle. *Finance Stoch.*, 9(1):189–214, Apr. 2015.
[32] D. Hobson and A. Neuberger. Robust bounds for forward start options. *Math. Finance*, 22(1):31–56, 2012.
[33] D. Hoover. Convergence in distribution and skorokhod convergence for the general theory of processes. *Probability theory and related fields*, 89(3):239–259, 1991.
[34] D. N. Hoover. Extending probability spaces and adapted distribution. In *Séminaire de Probabilités, XXVI*, volume 1526 of *Lecture Notes in Math.*, pages 560–574. Springer, Berlin, 1992.
[35] D. N. Hoover and H. J. Keisler. Adapted probability distributions. *Transactions of the American Mathematical Society*, 286(1):159–201, 1984.
[36] J. Jacod and J. Mémin. Weak and strong solutions of stochastic differential equations: existence and stability. In *Stochastic integrals*, pages 169–212. Springer, 1981.
[37] E. Janvresse, S. Laurent, and T. de la Rue. Standardness of monotonic Markov filtrations. *Markov Process. Related Fields*, 22(4):697–736, 2016.
[38] J. S. Jordan. The continuity of optimal dynamic decision rules. *Econometrica: Journal of the Econometric Society*, pages 1365–1376, 1977.
[39] F. B. Knight et al. A predictive view of continuous time processes. *The annals of Probability*, 3(4):573–596, 1975.
[40] T. Kurtz. The Yamada-Watanabe-Engelbert theorem for general stochastic equations and inequalities. *Electron. J. Probab*, 12:951–965, 2007.
[41] T. Kurtz et al. Weak and strong solutions of general stochastic models. *Electronic Communications in Probability*, 19, 2014.
[42] D. Lacker. Dense sets of joint distributions appearing in filtration enlargements, stochastic control, and causal optimal transport. *ArXiv e-prints*, 2018.
[43] D. Lamberton and G. Pagès. Sur l'approximation des réduites. *Ann. Inst. H. Poincaré Probab. Statist.*, 26(2):331–355, 1990.
[44] R. Lassalle. Causal transference plans and their Monge-Kantorovich problems. *Stochastic Analysis and Applications*, 36(3):452–484, 2018.
[45] S. MacLane. *Categories for the Working Mathematician*. Springer-Verlag, New York, 1971. Graduate Texts in Mathematics, Vol. 5.
[46] J. Mémin. Stability of doob-meyer decomposition under extended convergence. *Acta Mathematicae Applicatae Sinica*, 19(2):177–190, 2003.

[47] E. Moggi. Computational lambda-calculus and monads. In *Proceedings of the Fourth Annual Symposium on Logic in Computer Science*, pages 14–23, Piscataway, NJ, USA, 1989. IEEE Press.
[48] A. Papapantoleon, D. Possamai, and A. Saplaouras. Stability results for martingale representations: the general case. *arXiv preprint arXiv:1806.01172*, 2018.
[49] G. C. Pflug. Version-independence and nested distributions in multistage stochastic optimization. *SIAM Journal on Optimization*, 20(3):1406–1420, 2009.
[50] G. C. Pflug and A. Pichler. A distance for multistage stochastic optimization models. *SIAM J. Optim.*, 22(1):1–23, 2012.
[51] G. C. Pflug and A. Pichler. *Multistage stochastic optimization*. Springer Series in Operations Research and Financial Engineering. Springer, Cham, 2014.
[52] G. C. Pflug and A. Pichler. Dynamic generation of scenario trees. *Comput. Optim. Appl.*, 62(3):641–668, 2015.
[53] G. C. Pflug and A. Pichler. From empirical observations to tree models for stochastic optimization: convergence properties. *SIAM J. Optim.*, 26(3):1715–1740, 2016.
[54] A. Pichler. Evaluations of risk measures for different probability measures. *SIAM J. Optim.*, 23(1):530–551, 2013.
[55] A. Pratelli. On the equality between Monge's infimum and Kantorovich's minimum in optimal mass transportation. *Ann. Inst. H. Poincaré Probab. Statist.*, 43(1):1–13, 2007.
[56] L. Rüschendorf. The Wasserstein distance and approximation theorems. *Z. Wahrsch. Verw. Gebiete*, 70(1):117–129, 1985.
[57] T. Van Zandt. Information, measurability, and continuous behavior. *Journal of Mathematical Economics*, 38(3):293–309, 2002.
[58] A. M. Vershik. Decreasing sequences of measurable partitions and their applications. *Sov. Mat. Dokl.*, 11(4):1007 – 1011, 1970.
[59] A. M. Vershik. Theory of decreasing sequences of measurable partitions. *Algebra i Analiz*, 6(4):1–68, 1994.
[60] T. Yamada and S. Watanabe. On the uniqueness of solutions of stochastic differential equations. *Journal of Mathematics of Kyoto University*, 11(1):155–167, 1971.

Department of Mathematics, University of Vienna, Austria
E-mail address, J. Backhoff-Veraguas: julio.backhoff@univie.ac.at
E-mail address, D. Bartl: daniel.bartl@univie.ac.at
E-mail address, M. Beiglböck: mathias.beiglboeck@univie.ac.at
E-mail address, M. Eder: manuel.eder@univie.ac.at