Adapted Wasserstein Distances and Applications to Distributionally Robust Optimization

Yifan Jiang
University College
University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Trinity 2025

This thesis is dedicated to my parents

林联 and 蒋雪明

for their unwavering support and unconditional love.

Acknowledgements

靡不有初，鲜克有终。*To begin is common; to preserve and succeed is exceptional.* — 诗经 *(Classic of Poetry)*

I would like to express my heartfelt gratitude to my supervisor Jan Obłój for his patient supervision over the past four years. His sincere guidance and thoughtful suggestions have pushed and shaped me into a better researcher. I always enjoyed our discussions, and what I have learned from him is invaluable, and goes far beyond mathematics.

I am deeply grateful to my co-supervisor, Gui-Qiang Chen, whose unwavering positivity has been inspiring. My thanks also extend to his family, Anna and Richard for their warm hospitality during the annual gatherings—it made me feel at home, even from thousands of miles away.

I am honored to have Rama Cont and Nizar Touzi serve as the examiners for this thesis. I would also like to thank my early-stage examiners José Carrillo, Samuel Cohen, and Zhongmin Qian for their valuable comments and suggestions.

My sincere thanks go to the MCF and PDE research groups for providing a friendly and stimulating research environment. Special thanks to WIP members Benjamin Joseph, Fang Rui Lim, Vlad Tuchilus for bearing with my many picky questions. Thank you to my CDT cohort, especially Karolina Bassa, Vladislav Cherepanov, Rivka Mitchell, Shyam Popat, Milena Vuletić, who have been a great company during my DPhil journey.

I am indebted to many others who have encouraged, inspired, accompanied me through the past years.

To my dear friend, Maths, for making my life so challenging yet so rewarding.

To my parents, for encouraging me to chase my dream and supporting that dream as if it were theirs.

Finally, to my fiancée, Yuxin, for standing firm beside me when we had nothing but each other. Not everyone is fortunate enough to meet their special one at the age of thirteen. My heart is full of gratitude for the countless moments, big and small, that we have shared. Through the highest peaks and the deepest valleys of our journey, we were, are, and will always be the optimal coupling.

Statement of Authorship

I, Yifan Jiang, hereby declare that this thesis and the work presented in it are my own, except where otherwise stated. Where this thesis includes work to which others have contributed, the nature and extent of my contribution and the contributions of others are clearly indicated.

Chapters 3, 4, and 5 are based on the following works, which have been submitted for publication or are in preparation:

1. Y. Jiang and F. R. Lim. *A transfer principle for computing the adapted Wasserstein distance between stochastic processes*, June 2025. arXiv:2505.21337.
2. Y. Jiang. *Duality of causal distributionally robust optimization: The discrete-time case*, January 2024. arXiv:2401.16556.
3. Y. Jiang and J. Obłój. *Sensitivity of causal distributionally robust optimization*, August 2024. arXiv:2408.17109.

In all cases, my contribution was primary. I conceived the research problems, developed the theoretical framework, carried out the numerical experiments, and wrote the first drafts of the manuscripts.

Abstract

This thesis studies adapted Wasserstein distances and their applications to distributionally robust optimization (DRO) problems in a dynamic context. In Chapter 3, we propose a transfer principle to study the adapted 2-Wasserstein distance between stochastic processes. We obtain an explicit formula for the distance between real-valued mean-square continuous Gaussian processes by introducing causal factorization, an infinite-dimensional analogue of the Cholesky decomposition for operators on Hilbert spaces. We discuss the existence and uniqueness of this causal factorization and link it to the canonical representation of Gaussian processes. As a byproduct, we characterize mean-square continuous Gaussian Volterra processes in terms of their natural filtrations. Moreover, for real-valued fractional stochastic differential equations, we show that the synchronous coupling between the driving fractional noises attains the adapted Wasserstein distance under some monotonicity conditions. Our results cover a wide class of stochastic processes which are neither Markov processes nor semi-martingales, including fractional Brownian motions and fractional Ornstein–Uhlenbeck processes.

Subsequently, we contribute to the adapted Wasserstein distributionally robust optimization (AW-DRO) problem in both discrete- and continuous- time settings. This framework addresses decision-making under model uncertainty by optimizing for the worst-case scenario, where uncertainty is captured by penalizing potential models in function of their adapted Wasserstein distance to a given reference model. In Chapter 4, we derive a dynamic duality formula that reformulates the worst-case expectation as a tractable minimax problem. The inner maximum can be computed recursively in discrete time, or solved by a path-dependent Hamilton–Jacobi–Bellman equation in continuous time. We further extend these duality results from the worst-case expectation to the worst-case expected shortfall, a non-linear expectation. Finally, we apply the AW-DRO framework to optimal stopping problems in discrete time. We recast the original problem as a classical Wasserstein DRO on a nested space by introducing a novel relaxation that considers stopping times with respect to general filtrations.

In Chapter 5, we study the AW-DRO via sensitivity analysis. We introduce a real-valued parameter into the penalty function to reflect the strength of model uncertainty. Our main results derive the first-order sensitivity of the worst-case expectation with respect to the penalization parameter. Moreover, we investigate the case where a martingale constraint is imposed on the underlying model, as is common for pricing measures in mathematical finance. By introducing different scaling regimes, we obtain the continuous-time sensitivities as nontrivial limits of their discrete-time counterparts. Of independent interest, we also establish a novel stochastic Fubini theorem for a two-fold forward and Itô integral.

Contents

List of Figures

List of Symbols

$\mathbb{R}^*$	Extended real line $\mathbb{R} \cup \{+\infty\}$
$C(X;\mathbb{R})$	Continuous functions on a topological space X
$L^p(X,\mu;\mathbb{R})$	L^p integrable functions on a measure space (X,μ)
$W^{k,p}$	Sobolev space with indices (k,p)
$\mathcal{W}_p$	p-Wasserstein distance
$\mathcal{CW}_p$	Causal p-Wasserstein distance
$\mathcal{AW}_p$	Adapted p-Wasserstein distance
$\Pi(\mu,\nu)$	Couplings between probability measures μ and ν
$\Pi_{\mathfrak{c}}(\mu,\nu)$	Causal couplings between probability measures μ and ν
$\Pi_{\mathfrak{bc}}(\mu,\nu)$	Bi-causal couplings between probability measures μ and ν
$\Pi(\mu,*)$	Couplings with a fixed first marginal μ
$\Pi_{\mathfrak{c}}(\mu,*)$	Causal couplings with a fixed first marginal μ
$\Pi_{\mathfrak{bc}}(\mu,*)$	Bi-causal couplings with a fixed first marginal μ
$\mathcal{T}_{\mathfrak{c}}(\mu,\nu)$	Causal optimal transport cost between μ and ν
$\mathcal{T}_{\mathfrak{bc}}(\mu,\nu)$	Bi-causal optimal transport cost between μ and ν
φ^*	Convex conjugate of a function φ
φ_*	Concave conjugate of a function φ
${}^{\mathrm{o}}A$	Optional projection of a stochastic process A
${}^{\mathrm{p}}A$	Predictable projection of a stochastic process A

$\mathbb{D}$	Pathwise Malliavin derivative
$\mathbf{D}$	Classical Malliavin derivative
$\mathcal{D}$	Horizontal derivative
∇_x	Vertical derivative
∂_ω	Fréchet derivative
I	Index set: $\{0,\dots,N\}$ in discrete time and $[0,T]$ in continuous time
$\mathcal{X}$	Canonical path space
$\mathcal{B}(\mathcal{X})$	Borel σ-algebra of $\mathcal{X}$
$\mathscr{P}(\mathcal{X})$	Borel probability measures on $\mathcal{X}$
$\mathscr{P}_p(\mathcal{X})$	Borel probability measures on $\mathcal{X}$ with a finite p-th moment
$\mathscr{M}(\mathcal{X})$	Martingale measures on $\mathcal{X}$
δ_x	Dirac measure at $x \in \mathcal{X}$
$\mu \otimes \nu$	Product of two measures μ and ν
${}^{\mu}\mathcal{F}$	Completion of a σ-algebra $\mathcal{F}$ under μ
$\omega \otimes_s \eta$	Concatenation of two paths ω and η at time s
B_H	Fractional Brownian motion with Hurst parameter H
$\mathbb{X}$	Adapted stochastic process $\mathbb{X} = (\Omega, \mathcal{F}, \mathbf{F}, P, X)$
AP	Space of adapted stochastic processes
NP	Space of naturally filtered stochastic processes
$\lVert\cdot\rVert_\infty$	Uniform norm
$\lVert\cdot\rVert_{H_0^1}$	Cameron–Martin norm
$\lVert\cdot\rVert_{\mathrm{tr}}$	Trace norm
$\lVert\cdot\rVert_{\mathbf{F}}$	Frobenious norm
$\lVert\cdot\rVert_{\mathrm{HS}}$	Hilbert–Schmidt norm

H_μ	Hilbert space $L^2([0,T],\mu;\mathbb{R})$ where μ a positive measure on $[0,T]$
$\mathfrak{B}(H_{\mu_1}, H_{\mu_2})$	Bounded linear operators from H_{μ_1} to H_{μ_2}
$\mathfrak{B}_2(H_{\mu_1}, H_{\mu_2})$	Hilbert–Schmidt operators from H_{μ_1} to H_{μ_2}
$\mathfrak{N}(H_\mu)$	Nest algebra on H_μ
$\mathfrak{D}(H_\mu)$	Diagonal algebra on H_μ
$\mathrm{span}(S)$	Linear span of S a subset of a linear space
$\overline{S}$	Closure of S a subset of a topological space

Chapter 1

Introduction

Stochastic modeling in a dynamic setting is a staple of applied mathematics, with ubiquitous applications in fields ranging from mathematical finance to machine learning. Traditionally, models were derived from theoretical considerations, combined with calibration to data, and often had nice analytic representations. More recently, models are often obtained through a purely data-driven approach, where they are constructed from the empirical distributions of observed samples. In both scenarios, however, the model can easily be misspecified, leading to a discrepancy between the postulated model and the ground-truth distribution. It is therefore of fundamental importance to develop methods that are robust to such model uncertainty.

A distributionally robust approach is particularly appealing in this context which offers a connecting bridge between model-based and model-free methods. Distributionally robust optimization (DRO) is formulated as a minimax problem where a decision-maker seeks an optimal strategy under the worst-case stochastic model drawn from an 'ambiguity set'—a collection of plausible models centered around a given reference model. The performance and tractability of a DRO problem critically depend on how this ambiguity set is chosen. In recent years, ambiguity sets constructed using the Wasserstein distance from optimal transport have gained significant attention in operations research (Mohajerin Esfahani and Kuhn, 2018, Blanchet and Murthy, 2019, Gao and Kleywegt, 2022, Gao, 2023), mathematical finance (Obłój and Wiesel, 2021, Blanchet et al., 2022, Nendel and Sgarabottolo, 2024, Wu and Jaimungal, 2023), and machine learning (Blanchet et al., 2019, Bai et al., 2023, 2025, Nietert et al., 2023) due to their statistical guarantees and interpretability. We refer to Rahimian and Mehrotra (2022), Kuhn et al. (2025) for a comprehensive exposition.

However, for dynamic problems involving stochastic processes, the classical Wasserstein distance reveals a fundamental limitation. It views stochastic processes as path-valued random variables and metrizes the weak topology of probability measures on the path space, a topology that proves too coarse for many applications in mathematical finance. This 'static' viewpoint fails to properly capture the 'dynamic' of information carried by a process's filtration. Consequently, crucial quantities, such as the value of an optimal stopping problem, are not continuous with respect to the weak topology. Different notions of adapted topologies have been proposed to refine the weak topology, such as Aldous's extended weak topology (Aldous, 1981), Hellwig's information topology (Hellwig, 1996), Hoover–Keisler topology (Hoover and Keisler, 1984, Hoover, 1987), nested distance (Pflug and Pichler, 2012), and the adapted Wasserstein distance (Lassalle, 2018, Bion–Nadal and Talay, 2019). In the seminal paper Backhoff-Veraguas et al. (2020b), these notions are unified and proven to be all equivalent to the initial topology of the optimal stopping problems in a discrete-time setting. The essence of all aforementioned adapted topologies is to consider not only the law but also the conditional law of the stochastic process with respect to its natural filtration. Or, in other words, to incorporate the information flow carried by the underlying process.

In this thesis, we focus on the adapted Wasserstein distance and its applications to DRO problems. The adapted Wasserstein distance was first introduced in Lassalle (2018) as a dynamic counterpart of the Wasserstein distance. For $p \geq 1$ and two stochastic processes X_1 and X_2, their adapted p-Wasserstein distance is given by

$$\mathcal{AW}_p(X_1, X_2) := \inf_{\pi \in \Pi_{\mathfrak{bc}}(X_1, X_2)} E_\pi[d(X_1, X_2)^p]^{1/p} \tag{1.1}$$

where d is a distance on the path space, and $\Pi_{\mathfrak{bc}}(X_1, X_2)$ is a subset of couplings with an additional *(bi-)causality* constraint. Heuristically speaking, a coupling is *causal* if

given the past of X_1, the past of X_2 and the future of X_1 are independent,

as stated in Backhoff-Veraguas et al. (2017). The term *bi-causal* means that this condition holds symmetrically for both processes. We refer to Chapter 2 for a formal definition. A related notion is the p-causal Wasserstein distance $\mathcal{CW}_p$ obtained by replacing the set of bi-causal couplings with the larger set of causal couplings in (1.1).

This leads to the adapted Wasserstein distributionally robust optimization (AW-DRO) problem, a variant of DRO tailored for dynamic contexts. The aim of AW-DRO is to optimize the worst-case expectation given by

$$\sup_{\nu \in B(\mu)} E_\nu[f(X)], \tag{1.2}$$

where the ambiguity set $B(\mu)$ is an adapted Wasserstein ball centered at the reference model μ.

However, both computing the adapted Wasserstein distance itself and solving the AW-DRO problem are difficult, due to the additional causality constraint imposed on the couplings. This thesis aims to bridge this gap by establishing new theoretical results that enhance the tractability and applicability of the adapted Wasserstein distance and the AW-DRO framework.

1.1 Outline and contributions

In this section we will summarize the main results presented in the different chapters of this thesis without going into the technical details. A more detailed introduction and literature review will be provided at the beginning of each individual chapter.

Chapter 2 serves as a warm-up, where we introduce basic notations and concepts. From there, we formally introduce the discrete- and continuous- time adapted Wasserstein distances between probability measures on path spaces, and discuss the properties of the induced adapted Wasserstein topology. We then introduce the extension of the adapted Wasserstein distance to the space of adapted stochastic processes with more general filtrations. This extension, which highlights the crucial role of the filtration, is essential for the duality results developed in Chapter 4. The material in this chapter is based on several pioneering works Lassalle (2018), Acciaio et al. (2020), Bartl et al. (2024).

In Chapter 3, we propose a transfer principle to study the adapted 2-Wasserstein distance between continuous-time stochastic processes. Heuristically, it says if for stochastic process X_i there exists a representation $X_i = T_i(Y_i)$ such that X_i and Y_i generate the same natural filtration for $i = 1, 2$, then the adapted Wasserstein distance between X_1 and X_2 can be transferred to a bi-causal optimal transport problem between Y_1 and Y_2. In essence, this principle encourages us to choose an appropriate representation of a stochastic process leading to a nicer parameterization of the bi-causal coupling, and hence simplifies the computation of the adapted Wasserstein distance. In Section 3.4, we first apply the transfer principle and obtain an explicit formula for the distance between real-valued mean-square continuous Gaussian processes by introducing the causal factorization as an infinite-dimensional analogue of the Cholesky decomposition for operators on Hilbert spaces. In Section 3.3, we discuss the existence and uniqueness of this causal factorization and link it to the canonical representation of Gaussian processes (Hida, 1960, Hida and Hitsuda, 1993). As a byproduct, we characterize mean-square continuous Gaussian Volterra processes in terms of their natural filtrations. Moreover, for real-valued fractional stochastic differential equations, in Section 3.5 we show that the synchronous coupling between the driving fractional noises attains the adapted Wasserstein distance under some monotonicity conditions. We apply the transfer principle and reformulate the adapted Wasserstein distance as a stochastic control problem. By leveraging the functional Itô formula of Viens and Zhang (2019), we verify the synchronous coupling induces a classical solution to the associated path-dependent Hamilton–Jacobi–Bellman equation, and thus we show the optimality. Our results cover a wide class of stochastic processes which are neither Markov processes nor semi-martingales, including fractional Brownian motions and fractional Ornstein–Uhlenbeck processes. To the best of our knowledge, unlike existing literature Lassalle (2018), Bion–Nadal and Talay (2019), Backhoff-Veraguas et al. (2022b), these are the very first results considering processes beyond semi-martingales. This chapter corresponds to the article Jiang and Lim (2025).

Subsequently, we contribute to the adapted Wasserstein distributionally robust optimization (AW-DRO) problem in both discrete- and continuous- time settings. This framework addresses decision-making under model uncertainty by optimizing for the worst-case expectation (1.2). Our contribution here is to provide tractable theoretical results via duality theory and sensitivity analysis.

In Chapter 4, we derive a dynamic duality formula that reformulates the worst-case expectation as a tractable minimax problem. We first extend (1.2) to a general penalized form

$$V = \sup_{\nu \in \mathscr{P}(\mathcal{X})} \{E_\nu[f(X)] - L(\mathcal{T}_{\mathfrak{bc}}(\mu, \nu))\},$$

where L is a convex penalty function, and $\mathcal{T}_{\mathfrak{bc}}(\mu, \nu) := \inf_{\pi \in \Pi_{\mathfrak{bc}}(\mu,\nu)} E[c(X, Y)]$ is the optimal bi-causal transport cost for a given cost function c. Here, we take infimum over $\mathscr{P}(\mathcal{X})$, the set of probability measures on the path space $\mathcal{X}$. In this setup, model uncertainty is captured by penalizing potential models in function of their bi-causal optimal transport cost to a given reference model μ. By taking L as an indicator function and an appropriate c, one can easily recover (1.2). The main duality result states the following:

$$V = \inf_{\lambda \geq 0} \{L^*(\lambda) + U(\lambda)\},$$

where L^* is the convex conjugate of L, and U is a convex function which can be computed recursively in discrete time, or solved by a path-dependent Hamilton–Jacobi–Bellman equation in continuous time. We remark that our discrete-time result retrieves existing Wasserstein DRO duality results (Gao and Kleywegt, 2022, Blanchet and Murthy, 2019) as a special case. For the continuous-time case, our approach adapts the non-Markovian stochastic control framework from Bally et al. (2016, Chapter 8.3). In Section 4.6, we further extend these duality results from the worst-case expectation to the worst-case expected shortfall, a non-linear expectation. We demonstrate the extra risk from model uncertainty for an exotic option. Finally, in Section 4.7 we apply the AW-DRO framework to optimal stopping problems in discrete time. We recast the original problem as a classical Wasserstein DRO on a nested space by introducing a novel relaxation that considers stopping times with respect to general filtrations. A preliminary version of this chapter appeared in Jiang (2024).

In Chapter 5, we study AW-DRO via a sensitivity analysis, an approach with deep roots in mathematical finance. Just as the 'Greeks' measure the sensitivity of derivative prices to model parameters, we aim to compute the sensitivity of our robust price (1.2) to model uncertainty itself. To this end, we fix $p, q > 1$ with $1/p + 1/q = 1$ and introduce a parameterized variant of (1.2) as

$$V(\delta) = \sup_{\nu \in \mathscr{P}(\mathcal{X})} \{E_\nu[f(X)] - L_\delta(\mathcal{CW}_p(\mu, \nu))\},$$

where $L_\delta(\cdot) = \delta L(\cdot/\delta)$ and $\mathcal{CW}_p$ is the p-causal Wasserstein distance. The penalty strength is controlled through the real-valued parameter δ which, in the special case of an indicator penalty, is simply the radius of the uncertainty ball. Instead of computing the exact value of the worst-case expectation, our main results derive the first-order sensitivity of $V(\delta)$ with respect to the penalization parameter, i.e., a non-parametric 'Greek' for model uncertainty. In the discrete-time setting, Theorem 5.15 provides an explicit formula for the sensitivity:

$$V(\delta) = V(0) + \delta\Upsilon + o(\delta) \text{ with } \Upsilon = L^*(\|{}^{\mathrm{o}}\mathbb{D}f\|_{L^q(\mu)}),$$

where L^* is the convex conjugate of L, $\mathbb{D}$ is a gradient induced by the metric chosen on the underlying path space, and ${}^{\mathrm{o}}$ denotes the optional projection. Motivated by the pricing measures in financial applications, we analyze a variant with martingale constraints imposed on the underlying model:

$$V_{\text{Mart}}(\delta) = \sup_{\nu \in \mathscr{M}(\mathcal{X})} \{E_\nu[f(X)] - L_\delta(\mathcal{AW}_p(\mu, \nu))\}.$$

For $p = q = 2$, as a special case of Theorem 5.18, we derive

$$V_{\text{Mart}}(\delta) = V_{\text{Mart}}(0) + \delta\Upsilon_{\text{Mart}} + o(\delta) \text{ with } \Upsilon_{\text{Mart}} = L^*(\|{}^{\mathrm{o}}\mathbb{D}f - {}^{\mathrm{p}}\mathbb{D}f\|_{L^2(\mu)}),$$

which notably involves both the optional projection ^{o} and the predictable projection ^{p}. We then consider different scaling regimes, which allow us to obtain the continuous-time sensitivities as nontrivial limits of their discrete-time counterparts. To establish our results we obtain several novel results which are of independent interest. In particular, we introduce a pathwise Malliavin derivative as the natural limit of the gradient $\mathbb{D}$ in continuous time. On the common domain, it coincides with the classical Malliavin derivative almost surely under the Wiener measure. We also establish a novel stochastic Fubini theorem for a two-fold forward and Itô integral in Theorem 5.12. Our results open new avenues to hedge model uncertainty, with the sensitivity Υ serving as a new tool for risk management. The contents in this chapter appeared in Jiang and Obłój (2024).

Chapter 2

Adapted Wasserstein distances

2.1 Basic definitions

We begin with a gentle introduction to the classical Wasserstein distance and basic notations. Let $(\mathcal{X}, d_{\mathcal{X}})$ be a Polish metric space and $\mathscr{P}(\mathcal{X})$ be the set of Borel probability measures on $\mathcal{X}$. We take $p \geq 1$, and by $\mathscr{P}_p(\mathcal{X})$ we denote the set of probability measures with finite p-th moment, i.e.,

$$\mathscr{P}_p(\mathcal{X}) := \left\{ \mu \in \mathscr{P}(\mathcal{X}) : \int_{\mathcal{X}} d_{\mathcal{X}}(x, x_0)^p \mu(\mathrm{d}x) < \infty \right\}, \text{ for some } x_0 \in \mathcal{X}.$$

It is clear that $\mathscr{P}_p(\mathcal{X})$ is independent of the choice of x_0 as $d_{\mathcal{X}}$ is a metric. The p-Wasserstein distance between $\mu, \nu \in \mathscr{P}_p(\mathcal{X})$ is defined as

$$\mathcal{W}_p(\mu, \nu) := \inf_{\pi \in \Pi(\mu, \nu)} \left(\int_{\mathcal{X} \times \mathcal{X}} d_{\mathcal{X}}(x_1, x_2)^p \pi(\mathrm{d}x_1, \mathrm{d}x_2) \right)^{1/p},$$

where $\Pi(\mu, \nu)$ is the set of couplings between μ and ν given by

$$\Pi(\mu, \nu) = \{\pi \in \mathscr{P}(\mathcal{X} \times \mathcal{X}) : \pi(\cdot \times \mathcal{X}) = \mu(\cdot) \text{ and } \pi(\mathcal{X} \times \cdot) = \nu(\cdot)\}.$$

The Wasserstein space $(\mathscr{P}_p(\mathcal{X}), \mathcal{W}_p)$ can be viewed as a lift of $(\mathcal{X}, d_{\mathcal{X}})$ to the space of probability measures, and it inherits many topological properties from the base space. For example, the map $x \mapsto \delta_x$ gives an isometric embedding of $(\mathcal{X}, d_{\mathcal{X}})$ into $(\mathscr{P}_p(\mathcal{X}), \mathcal{W}_p)$, where δ_x is the Dirac measure at $x \in \mathcal{X}$.

We say a sequence of probability measures μ_n converges weakly to $\mu \in \mathscr{P}(\mathcal{X})$ if for any bounded continuous function $f : \mathcal{X} \to \mathbb{R}$ we have the convergence of $\int f \, \mathrm{d}\mu_n$ to $\int f \, \mathrm{d}\mu$. The Wasserstein distance metrizes the weak convergence on $\mathscr{P}_p(\mathcal{X})$.

Proposition 2.1 (Villani (2009), Remark 6.9)**.** *Let $\mu_n, \mu \in \mathscr{P}_p(\mathcal{X})$. Then μ_n converges to μ in $\mathcal{W}_p$ if and only if μ_n converges to μ in the weak topology and the p-th moment $\int_{\mathcal{X}} d_{\mathcal{X}}(x, x_0)^p \mu_n(\mathrm{d}x)$ converges to $\int_{\mathcal{X}} d_{\mathcal{X}}(x, x_0)^p \mu(\mathrm{d}x)$.*

In particular, if $d_{\mathcal{X}}$ is bounded, then all p-Wasserstein topologies are equivalent to the weak topology. The following theorem shows that the Wasserstein space inherits the Polish property from the underlying metric space.

Theorem 2.2 (Villani (2009), Theorem 6.18)**.** *$(\mathscr{P}_p(\mathcal{X}), \mathcal{W}_p)$ is a Polish metric space, and in particular, it is complete.*

In this thesis we are interested in the probability measures on path spaces which are interpreted as laws of stochastic processes. We will consider both discrete- and continuous- time settings. In discrete time, we take time index set I as $\{0, 1, \dots, N\}$ and the canonical path space $\mathcal{X} = \mathcal{X}_0 \times \mathcal{X}_1 \times \cdots \mathcal{X}_N$ the product of Polish spaces; in continuous time, we consider $I = [0, T]$ and $\mathcal{X} = C([0, T]; \mathbb{R}^N)$ the continuous path space. We equip $\mathcal{X}$ with the uniform metric in both cases. Let $\mathrm{Id}_{\mathcal{X}} : \mathcal{X} \to \mathcal{X}$ be the identity map on $\mathcal{X}$. It is interpreted as the canonical process and induces a natural filtration $\mathbf{F} = (\mathcal{F}_t)_{t \in I}$ given by $\mathcal{F}_t = \sigma(\mathrm{Id}_{\mathcal{X}}(s) : 0 \leq s \leq t)$ which denotes the information available at time t.

To incorporate the information structure of the underlying space, the Wasserstein distance is modified using the key concept of a *causal* coupling. Heuristically, this is a transport plan that is constrained to move mass without using any information from the future.

Definition 2.3 (Causal coupling)**.** Let $\mu, \nu \in \mathscr{P}(\mathcal{X})$. We say a coupling $\pi \in \Pi(\mu, \nu)$ is *causal* if

$$x_1 \mapsto \pi_{x_1}(V_t) \text{ is } {}^{\mu}\mathcal{F}_t\text{-measurable}$$

for any $V_t \in \mathcal{F}_t$ and $t \in I$, where ${}^{\mu}\mathcal{F}_t$ is the completion of $\mathcal{F}_t$ under μ and π_{x_1} is the disintegration kernel given by $\pi(\mathrm{d}x_1, \mathrm{d}x_2) = \mu(\mathrm{d}x_1)\pi_{x_1}(\mathrm{d}x_2)$. A causal coupling π is *bi-causal* if further $[(x_1, x_2) \mapsto (x_2, x_1)]_{\#}\pi$ is causal. We write the set of causal (bi-causal) couplings between μ and ν as $\Pi_{\mathfrak{c}}(\mu, \nu)$ $(\Pi_{\mathfrak{bc}}(\mu, \nu))$.

We give a few examples of (bi-)causal couplings.

1. The product measure $\pi = \mu \otimes \nu$ is always a bi-causal coupling.

2. The Knothe–Rosenblatt coupling between discrete-time scalar processes is a bi-causal coupling (Backhoff-Veraguas et al., 2017, Proposition 5.8).

3. Let $T : \mathcal{X} \to \mathcal{X}$ be a transport map such that $T_{\#}\mu = \nu$. Then $\pi = (\mathrm{Id}_{\mathcal{X}}, T)_{\#}\mu$ is a causal coupling if T is non-anticipative, i.e., $T_t(x) = T_t(x(\cdot \wedge t))$. If further T is invertible and T^{-1} is non-anticipative, then π is bi-causal.

4. Let $X(t) = \int_0^t \{3 - 12(s/t) + 10(s/t)^2\}\,\mathrm{d}B(s)$ be Lévy's non-canonical representation of the Brownian motion (Lévy, 1956). The joint law $\pi = \mathrm{Law}(B, X)$ gives a causal coupling between Wiener measure, but it is not bi-causal (Jiang and Obłój, 2024, Remark 2.3).

5. Assume that SDE $X(t) = x + \int_0^t b(X(s))\,\mathrm{d}s + \int_0^t \sigma(X(s))\,\mathrm{d}B(s)$ has a unique strong solution. The joint law $\pi = \mathrm{Law}(B, X)$ yields a causal coupling as the solution is given by a non-anticipative Itô map $X = F(B)$. Moreover, this coupling is actually bi-causal as shown in Cont and Lim (2024, Theorem 3.2).

For here and what follows, we should interpret 'distance' in a generalized sense; it is only a non-negative function $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_{\geq 0} \cup \{+\infty\}$ vanishing on the diagonal and satisfying a 'directional' triangle inequality

$$d(x, y) + d(y, z) \geq d(x, z) \text{ for any } x, y, z \in \mathcal{X}.$$

Definition 2.4 (Adapted Wasserstein distance)**.** Let $\mu, \nu \in \mathscr{P}(\mathcal{X})$ and d a (generalized) distance on $\mathcal{X}$. The adapted p-Wasserstein distance between μ and ν is defined as

$$\mathcal{AW}_p(\mu, \nu) := \inf_{\pi \in \Pi_{\mathfrak{bc}}(\mu,\nu)} \left(\int_{\mathcal{X}\times\mathcal{X}} d(x_1, x_2)^p \pi(\mathrm{d}x_1, \mathrm{d}x_2) \right)^{1/p}.$$

Remark 2.5. A related notion appeared in the literature is the causal p-Wasserstein 'distance' defined as

$$\mathcal{CW}_p(\mu, \nu) := \inf_{\pi \in \Pi_{\mathfrak{c}}(\mu,\nu)} \left(\int_{\mathcal{X}\times\mathcal{X}} d(x_1, x_2)^p \pi(\mathrm{d}x_1, \mathrm{d}x_2) \right)^{1/p},$$

Unlike the adapted Wasserstein distance, $\mathcal{AW}_p$, which is symmetric (assuming the underlying metric d is), the causal Wasserstein distance $\mathcal{CW}_p$ is generally not. This asymmetry comes from the directional nature of the causality constraint. We follow the convention in the literature and interpret it as a generalized distance.

From the trivial inclusion $\Pi_{\mathfrak{bc}}(\mu, \nu) \subseteq \Pi_{\mathfrak{c}}(\mu, \nu)$, it is clear that the causal Wasserstein distance is dominated by the adapted one. Nevertheless, a crucial result in discrete time shows that the convergence of $\lim_{n\to\infty} \mathcal{CW}_p(\mu, \mu_n) = 0$ implies the convergence of $\lim_{n\to\infty} \mathcal{AW}_p(\mu, \mu_n) = 0$, see for example Backhoff-Veraguas et al. (2020b), Pammer (2024).

Remark 2.6. In discrete time, d is usually a distance compatible with $d_{\mathcal{X}}$. In continuous time, the choice of d is more flexible and is often tailored to the specific application. For example, d is taken as the uniform norm in Bartl et al. (2025a); the

L^2 norm in Bion–Nadal and Talay (2019), Backhoff-Veraguas et al. (2022b) and Chapter 3; Cameron–Martin norm in Lassalle (2018), Acciaio et al. (2020) and Chapter 4; $\mathcal{H}^2$ martingale norm[1] in Backhoff-Veraguas et al. (2020a) and Chapter 5, etc.

At this point, it is unclear whether the causal/adapted Wasserstein distance satisfies the triangle inequality. We continue with the discussion on the adapted Wasserstein topology and postpone the proof of triangle inequality to the end of this chapter where we accommodate the adapted Wasserstein distance in a more general framework of stochastic processes.

2.2 Adapted Wasserstein space

We start with a simple example in discrete time.

Example 2.7. In Figure 2.1, we consider a two-step case with the path space $\mathcal{X} = \{0\} \times \mathbb{R} \times \mathbb{R}$, and take d as the Euclidean distance. Let $\mu_n = \frac{1}{2}\delta_{(0,1/n,1)} + \frac{1}{2}\delta_{(0,-1/n,-1)}$, and $\mu = \frac{1}{2}\delta_{(0,0,1)} + \frac{1}{2}\delta_{(0,0,-1)}$, where δ_x denotes the Dirac measure concentrated at the path x. An optimal transport map between μ and μ_n for the classical Wasserstein distance $\mathcal{W}_p$ is given by

$$T_n(x) = \begin{cases} (0, 1/n, 1), x = (0, 0, 1), \\ (0, -1/n, 1), x = (0, 0, -1). \end{cases}$$

However, T_n is not a non-anticipative map as at time $t = 1$ it splits the mass to $1/n$ and $-1/n$ based on the future trajectory at time $t = 2$. Indeed, one can verify that μ^n does not converge to μ in $\mathcal{AW}_p$.

The example above also demonstrates that Wasserstein distance is too coarse for dynamic problems. If we consider the optimal stopping problem

$$\mathrm{OS}(\mu) := \sup_{\tau \in \mathscr{T}} E_\mu[X_\tau]$$

over the set of $\mathbf{F}$-stopping time $\mathscr{T}$, we observe a stark discontinuity: $\mathrm{OS}(\mu) = 0$ while $\lim_{n\to\infty} \mathrm{OS}(\mu_n) = 1/2$. The reason for this discontinuity becomes clear from the Snell envelope theorem, which gives

$$\mathrm{OS}(\mu) = E_\mu[\max\{X_0, E_\mu[\max\{X_1, E_\mu[X_2|X_1, X_2]\}|X_1]\}].$$

[1]Strictly speaking, in Jiang and Obłój (2024), Backhoff-Veraguas et al. (2020a) the distance d is chosen such as $d(x_1, x_2) = \sqrt{[x_1 - x_2]_T}$ π-a.s. holds for any bi-causal coupling π between two square-integrable martingale measures.

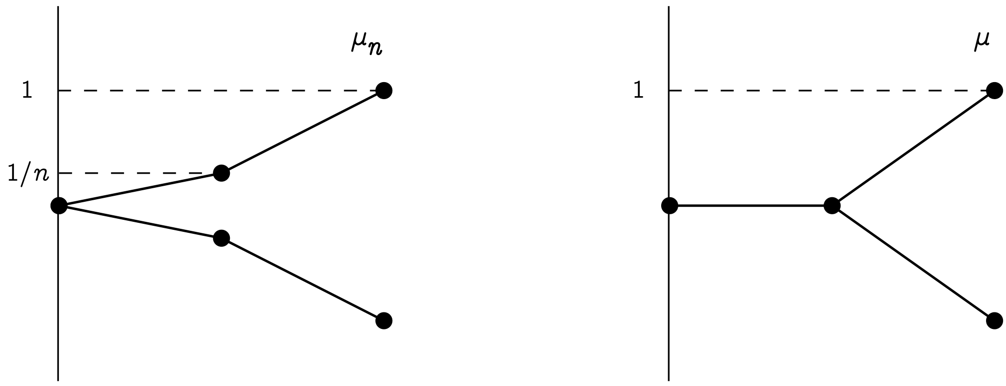

Figure 2.1: An example of convergence in the Wasserstein distance but not in the adapted Wasserstein distance.

The value of the optimal stopping problem depends not just on the law of the process, but on its nested conditional laws.

The challenge of finding a topology under which such stochastic optimization problems are continuous is not new. A key approach, which we will introduce shortly, involves encoding the information carried by a process as a distribution on a nested path space. This leads to the nested distribution and the induced nested weak topology studied in Pflug (2010), Pflug and Pichler (2012), Bartl et al. (2024).

Definition 2.8 (Nested space)**.** Let $\mathcal{X} = \mathcal{X}_0 \times \mathcal{X}_1 \times \cdots \times \mathcal{X}_N$. We recursively define $\widehat{X}_N = \mathcal{X}_N$ and

$$\widehat{\mathcal{X}}_n = \widehat{\mathcal{X}}_n^- \times \widehat{\mathcal{X}}_n^+ := \mathcal{X}_n \times \mathscr{P}(\widehat{\mathcal{X}}_{n+1}) \quad \text{for } 0 \le n \le N-1.$$

For any $\hat{x}_n \in \widehat{\mathcal{X}}_n$, we write it as $\hat{x}_n = (\hat{x}_n^-, \hat{x}_n^+)$ with $\hat{x}_n^- \in \mathcal{X}_n$ and $\hat{x}_n^+ \in \mathscr{P}(\widehat{\mathcal{X}}_{n+1})$. We say $\widehat{\mathcal{X}} = \widehat{\mathcal{X}}_0$ is the nested space associated to $\mathcal{X}$.

Definition 2.9 (Nested distribution)**.** Let $\mu \in \mathscr{P}(\mathcal{X})$, and $\mathrm{Id}_{\mathcal{X}} : \mathcal{X} \to \mathcal{X}$ the canonical process on $\mathcal{X}$. We recursively define $\widehat{\mathrm{Id}}_N = \mathrm{Id}_{\mathcal{X}_N}$ and $\widehat{\mathrm{Id}}_n : \mathcal{X} \to \widehat{\mathcal{X}}_n$ as

$$\widehat{\mathrm{Id}}_n = (\widehat{\mathrm{Id}}_n^-, \widehat{\mathrm{Id}}_n^+) := (\mathrm{Id}_n, \mathrm{Law}(\widehat{\mathrm{Id}}_{n+1} | \mathcal{F}_n)) \quad \text{for } 0 \le n \le N,$$

where $\mathbf{F} = (\mathcal{F}_n)$ is the natural filtration on $\mathcal{X}$. We say $\hat{\mu} = \mathrm{Law}(\widehat{\mathrm{Id}}_0) \in \mathscr{P}(\widehat{\mathcal{X}})$ is the nested distribution associated to μ. We say $\mu_n \in \mathscr{P}(\mathcal{X})$ converges to μ in the nested weak topology if the corresponding nested distribution $\hat{\mu}_n$ converges to $\hat{\mu}_n$ in weak topology as probability measures on $\widehat{\mathcal{X}}$.

The following result is adapted from Bartl et al. (2024, Theorem 1.3).

Proposition 2.10. *In discrete time, the adapted Wasserstein distance $\mathcal{AW}_p$ is not a complete metric on $\mathscr{P}_p(\mathcal{X})$. There exists a Polish metric $d_{\widehat{\mathcal{X}}}$ on $\widehat{\mathcal{X}}$ such that for any $\mu, \nu \in \mathscr{P}(\mathcal{X})$, we have $\mathcal{AW}_p(\mu, \nu) = \mathcal{W}_p(\hat{\mu}, \hat{\nu})$, where $\hat{\mu}$ and $\hat{\nu}$ are the nested distributions associated to μ and ν respectively. Moreover, $(\mathscr{P}_p(\widehat{\mathcal{X}}), \mathcal{W}_p)$ is isometric to the completion of $(\mathscr{P}_p(\mathcal{X}), \mathcal{AW}_p)$.*

An immediate implication of the above result is an 'adapted' version of Proposition 2.1 which characterizes the convergence in the adapted Wasserstein topology.

Proposition 2.11. *Let $\mu_n, \mu \in \mathscr{P}_p(\mu)$. Then μ_n converges to μ in $\mathcal{AW}_p$ if and only if μ_n converges to μ in the nested weak topology and $\int_{\mathcal{X}} d(x, x_0)^p \mu_n(\mathrm{d}x)$ converges to $\int_{\mathcal{X}} d(x, x_0)^p \mu(\mathrm{d}x)$.*

Proof. From Propositions 2.1 and 2.10, it suffices to show that the convergence of $\int_{\mathcal{X}} d(x, x_0)^p \mu_n(\mathrm{d}x)$ to $\int_{\mathcal{X}} d(x, x_0)^p \mu(\mathrm{d}x)$ is equivalent to the convergence of their nested counterparts $\int_{\widehat{\mathcal{X}}} d_{\widehat{\mathcal{X}}}(\hat{x}, \hat{x}_0)^p \hat{\mu}_n(\mathrm{d}\hat{x})$ to $\int_{\widehat{\mathcal{X}}} d_{\widehat{\mathcal{X}}}(\hat{x}, \hat{x}_0)^p \hat{\mu}(\mathrm{d}\hat{x})$. Notice the choice of x_0 and $\hat{x}_0$ is arbitrary. We fix an $x_0 \in \mathcal{X}$ and let $\mu_0 = \delta_{x_0} \in \mathscr{P}(\mathcal{X})$. We observe that the nested distribution $\hat{\mu}_0$ associated to μ_0 is again a Dirac measure, and we write it as $\delta_{\hat{x}_0} \in \mathscr{P}(\widehat{\mathcal{X}})$. It follows from the isometry in Proposition 2.10 that

$$\int_{\mathcal{X}} d(x, x_0)^p \mu_n(\mathrm{d}x) = \mathcal{AW}_p(\mu_n, \mu_0)^p = \mathcal{W}_p(\hat{\mu}_n, \hat{\mu}_0)^p = \int_{\widehat{\mathcal{X}}} d_{\widehat{\mathcal{X}}}(\hat{x}, \hat{x}_0)^p \hat{\mu}_n(\mathrm{d}\hat{x}).$$

Therefore, we derive the desired equivalence. □

A natural question arises: does the incompleteness of $\mathcal{AW}_p$ stem merely from the inadequacy of the natural filtration on the path space $\mathcal{X}$? For instance, the sequence μ_n constructed in Example 2.7 can be made to converge if we equip the Wasserstein limit μ with an enlarged filtration $\mathbf{G} = \{\mathcal{G}_0, \mathcal{G}_1, \mathcal{G}_2\}$ where $\mathcal{G}_1$ already contains the information from $\mathcal{F}_2$. In this case, the filtration at time 1 'knows' the future at time 2. And the optimal stopping problem with respect to this enlarged filtration $\mathbf{G}$ matches the limit of $\mathrm{OS}(\mu_n)$.

The above arguments suggest a potential path to retrieve the completeness: we only need to consider all possible filtrations on the path space $\mathcal{X}$. However, this is not the case. In Figure 2.2, we consider $\nu_n = \frac{1}{4}(\delta_{(0,1/n,1)} + \delta_{(0,1/n,-1)} + \delta_{(0,0,1)} + \delta_{(0,0,-1)})$ and $\nu = \frac{1}{2}(\delta_{(0,0,1)} + \delta_{(0,0,-1)})$. A direct calculate gives the limit of the corresponding optimal stopping values is $\lim_{n\to\infty} \mathrm{OS}(\nu_n) = 1/4$.

On the other hand, for *any* random time $\tau : \mathcal{X} \to \{0, 1, 2\}$, $E_\nu[X_\tau]$ can only take value in $\{0, 1/2, -1/2\}$. Since $1/4$ is not in this set, it is impossible to find a stopping time for ν on $\mathcal{X}$ such that recovers the limit of the optimal stopping values. This implies that the space $\mathcal{X}$ is not sufficient to describe the true adapted Wasserstein limit of ν_n.

Heuristically, the formal $\mathcal{AW}_p$ limit of ν_n would require external randomness—a setting akin to a 'parallel universe' where in one branch the future is known, and in the other it is not. One has to enlarge the base space $\mathcal{X}$ to accommodate such a realization.

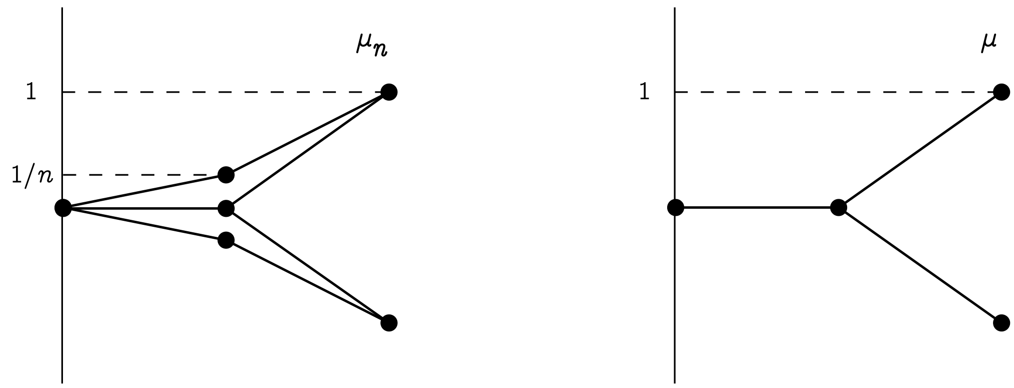

Figure 2.2: Adapted stochastic processes on the canonical path space are incomplete under the $\mathcal{AW}_p$.

Motivated by the above example, we introduce the extension of the adapted Wasserstein distance to general adapted stochastic processes. From now on, the time index set I can be either discrete or continuous.

Definition 2.12. We say an adapted stochastic process $\mathbb{X}$ is given by a 5-tuple $(\Omega^{\mathbb{X}}, \mathcal{F}^{\mathbb{X}}, P^{\mathbb{X}}, \mathbf{F}^{\mathbb{X}}, X)$ where $(\Omega^{\mathbb{X}}, \mathcal{F}^{\mathbb{X}}, \mathbf{F}^{\mathbb{X}}, P^{\mathbb{X}})$ is a filtered Polish probability space, and $X : \Omega^{\mathbb{X}} \to \mathcal{X}$ is a stochastic process adapted to the filtration $\mathbf{F}^{\mathbb{X}} = (\mathcal{F}_t^{\mathbb{X}})_{t\in I}$.

For an adapted process $\mathbb{X} = (\Omega^{\mathbb{X}}, \mathcal{F}^{\mathbb{X}}, P^{\mathbb{X}}, \mathbf{F}^{\mathbb{X}}, X)$, if $\mathbf{F}^{\mathbb{X}}$ coincides with $\mathbf{F}^X$ the natural filtration generated by X, then we say $\mathbb{X}$ is a *naturally filtered* process. With a slight abuse of notation, we may also use X to refer the 5-tuple of a naturally filtered process since itself determines the natural filtration. We denote the space of adapted processes and naturally filtered processes by AP and NP respectively.

The notion of the causality can be naturally carried over to adapted stochastic processes as follows.

Definition 2.13 (Causal coupling)**.** Let $\mathbb{X}_i = (\Omega^{\mathbb{X}_i}, \mathcal{F}^{\mathbb{X}_i}, \mathbf{F}^{\mathbb{X}_i}, P^{\mathbb{X}_i}, X_i)$ be two adapted stochastic processes for $i = 1, 2$. We say a coupling $\pi \in \Pi(P^{\mathbb{X}_1}, P^{\mathbb{X}_2})$ is *causal* if

$$\omega_1 \mapsto \pi_{\omega_1}(V_t) \text{ is } {}^{P^{\mathbb{X}_1}}\mathcal{F}_t^{\mathbb{X}_1}\text{-measurable}$$

for any $V_t \in \mathcal{F}_t^{\mathbb{X}_2}$ and $t \in I$, where π_{ω_1} is the disintegration kernel of π with respect to $\Omega^{\mathbb{X}_1}$. A causal coupling π is *bi-causal* if further $[(\omega_1, \omega_2) \mapsto (\omega_2, \omega_1)]_{\#}\pi$ is causal. With a slight abuse of notation, we write the set of causal (bi-causal) couplings between $\mathbb{X}_1$ and $\mathbb{X}_2$ as $\Pi_{\mathfrak{c}}(\mathbb{X}_1, \mathbb{X}_2)$ ($\Pi_{\mathfrak{bc}}(\mathbb{X}_1, \mathbb{X}_2)$).

Definition 2.14 (Adapted Wasserstein distance)**.** Let $\mathbb{X}_i = (\Omega^{\mathbb{X}_i}, \mathcal{F}^{\mathbb{X}_i}, \mathbf{F}^{\mathbb{X}_i}, P^{\mathbb{X}_i}, X_i)$ be two adapted stochastic processes for $i = 1, 2$. The adapted Wasserstein distance between $\mathbb{X}_1$ and $\mathbb{X}_2$ is defined as

$$\mathcal{AW}_p(\mathbb{X}_1, \mathbb{X}_2) := \inf_{\pi \in \Pi_{\mathfrak{bc}}(\mathbb{X}_1, \mathbb{X}_2)} E_\pi[d(X_1, X_2)^p]^{1/p}. \tag{2.1}$$

In the same fashion, we define the causal Wasserstein distance $\mathcal{CW}_p(\mathbb{X}_1, \mathbb{X}_2)$ by replacing the bi-causal coupling $\Pi_{\mathfrak{bc}}(\mathbb{X}_1, \mathbb{X}_2)$ with the causal coupling $\Pi_{\mathfrak{c}}(\mathbb{X}_1, \mathbb{X}_2)$.

Remark 2.15. The recent work of Bartl et al. (2025a) pointed out that the above definition is too strong in a continuous-time context. Indeed, the induced topology is strictly stronger than the initial topology of optimal stopping problems, and Donsker's theorem does not hold in the adapted Wasserstein topology. A relaxed version of (2.1) is proposed such that all adapted topologies are equivalent in a continuous-time setting. Nevertheless, the current definition enjoys better analytic properties and provide a uniform framework for both discrete time and continuous time.

Remark 2.16. The above definitions are consistent with the definitions introduced in Section 2.1 by viewing the 5-tuple $(\mathcal{X}, \mathcal{F}, \mathbf{F}, \mu, \mathrm{Id}_{\mathcal{X}})$ as an adapted stochastic process on the canonical path space.

2.3 Gluing and shadowing

We first present an equivalent characterization of the causality.

Proposition 2.17. *Let $\mathbb{X}_i = (\Omega^{\mathbb{X}_i}, \mathcal{F}^{\mathbb{X}_i}, \mathbf{F}^{\mathbb{X}_i}, P^{\mathbb{X}_i}, X_i)$ be two adapted stochastic processes for $i = 1, 2$. For a coupling $\pi \in \Pi(P^{\mathbb{X}_1}, P^{\mathbb{X}_2})$ the following statements are equivalent:*

(i) $\pi \in \Pi_{\mathfrak{c}}(\mathbb{X}_1, \mathbb{X}_2)$.

(ii) $\{\emptyset, \Omega^{\mathbb{X}_1}\} \otimes \mathcal{F}_t^{\mathbb{X}_2}$ *is conditionally independent of* $\mathcal{F}_T^{\mathbb{X}_1} \otimes \{\emptyset, \Omega^{\mathbb{X}_1}\}$ *under* π *given* $\mathcal{F}_t^{\mathbb{X}_1} \otimes \{\emptyset, \Omega^{\mathbb{X}_1}\}$ *for any* $t \in I$.

Proof. We notice that $\pi \in \Pi_{\mathfrak{c}}(\mathbb{X}_1, \mathbb{X}_2)$ is equivalent to the ${}^{P^{\mathbb{X}_1}}\mathcal{F}_t^{\mathbb{X}_1}$-measurability of the map

$$\omega_1 \mapsto \pi_{\omega_1}(B_t) = E_\pi[\mathbb{1}_{B_t}(\omega_2)|\mathcal{F}_T^{\mathbb{X}_1} \otimes \{\emptyset, \Omega^{\mathbb{X}_2}\}](\omega_1)$$

for any $B_t \in \mathcal{F}_t^{\mathbb{X}_2}$ and $t \in I$. This is further equivalent to the identity

$$E_\pi[\mathbb{1}_{B_t}(\omega_2)|\mathcal{F}_T^{\mathbb{X}_1} \otimes \{\emptyset, \Omega^{\mathbb{X}_2}\}](\omega_1) = E_\pi[\mathbb{1}_{B_t}(\omega_2)|\mathcal{F}_t^{\mathbb{X}_1} \otimes \{\emptyset, \Omega^{\mathbb{X}_2}\}](\omega_1)$$

for any $B_t \in \mathcal{F}_t^{\mathbb{X}_2}$ and $t \in I$. By the tower property of the condition expectation, we derive the equivalence between (i) and (ii). □

We introduce the gluing between two coupling measures.

Definition 2.18 (Gluing)**.** Let μ, ν, η be three probability measures on Polish spaces $\mathcal{X}$, $\mathcal{Y}$, and $\mathcal{Z}$ respectively. For any $\pi_1 \in \Pi(\mu, \nu)$ and $\pi_2 \in \Pi(\nu, \eta)$, their gluing is defined as $\pi_3(\mathrm{d}x, \mathrm{d}z) := \varpi(\mathrm{d}x, \mathcal{Y}, \mathrm{d}z)$ where $\varpi \in \mathscr{P}(\mathcal{X} \times \mathcal{Y} \times \mathcal{Z})$ is given by

$$\varpi(\mathrm{d}x, \mathrm{d}y, \mathrm{d}z) := \pi_{1,y}(\mathrm{d}x)\pi_{2,y}(\mathrm{d}z)\nu(\mathrm{d}y).$$

In particular, we have $\pi_3 := \varpi|_{(\mathcal{X} \times \mathcal{Z})} \in \Pi(\mu, \eta)$.

Proposition 2.19 (Pammer (2024), Corollary 2.12)**.** *Let* $\mathbb{X}_i = (\Omega^{\mathbb{X}_i}, \mathcal{F}^{\mathbb{X}_i}, \mathbf{F}^{\mathbb{X}_i}, P^{\mathbb{X}_i}, X_i)$ *be three adapted stochastic processes for* $i = 1, 2, 3$. *If* $\pi_1 \in \Pi_{\mathfrak{c}}(\mathbb{X}_1, \mathbb{X}_2)$ *and* $\pi_2 \in \Pi_{\mathfrak{c}}(\mathbb{X}_2, \mathbb{X}_3)$*, then their gluing* $\pi_3 \in \Pi_{\mathfrak{c}}(\mathbb{X}_1, \mathbb{X}_3)$*.*

Proof. Notice that we have

$$\varpi(\mathrm{d}\omega_1, \mathrm{d}\omega_2, \mathrm{d}\omega_3) = \pi_{2,\omega_2}(\mathrm{d}\omega_3)\pi_1(\mathrm{d}\omega_1, \mathrm{d}\omega_2) = \pi_{2,\omega_2}(\mathrm{d}\omega_3)\pi_{1,\omega_1}(\mathrm{d}\omega_2)\mu(\mathrm{d}\omega_1).$$

This yields the disintegration kernel of π_3 with respect to $\Omega^{\mathbb{X}_1}$ as

$$\pi_{3,\omega_1}(\mathrm{d}\omega_3) = \int_{\Omega^{\mathbb{X}_2}} \pi_{2,\omega_2}(\mathrm{d}\omega_3)\pi_{1,\omega_1}(\mathrm{d}\omega_2).$$

As $\pi_1 \in \Pi_{\mathfrak{c}}(\mathbb{X}_1, \mathbb{X}_2)$ and $\pi_2 \in \Pi_{\mathfrak{c}}(\mathbb{X}_2, \mathbb{X}_3)$, we obtain for any $W_t \in \mathcal{F}_t^{\mathbb{X}_3}$

$$\omega_1 \mapsto \pi_{3,\omega_1}(W_t) \text{ is } {}^{P^{\mathbb{X}_1}}\mathcal{F}_t^{\mathbb{X}_1}\text{-measurable.}$$

Therefore, we derive $\pi_3 \in \Pi_{\mathfrak{c}}(\mathbb{X}_1, \mathbb{X}_3)$.

□

Corollary 2.20. *Both the causal and the adapted* p*-Wasserstein distance satisfy the 'directional' triangle inequality. Moreover, if* d *is symmetric than* $\mathcal{AW}_p$ *is also symmetric.*

Proof. It follows directly from Proposition 2.19 and the triangle inequality of d. $\square$

Definition 2.21 (Shadowing)**.** Let μ, ν, μ', ν' be probability measures on Polish spaces $\mathcal{X}$, $\mathcal{Y}$, $\mathcal{X}'$, and $\mathcal{Y}'$ respectively. For any $\pi \in \Pi(\mu, \nu)$, $\gamma_1 \in \Pi(\mu', \mu)$, $\gamma_2 \in \Pi(\nu, \nu')$, we say the shadowing of π under (γ_1, γ_2) is given by

$$\pi'(\mathrm{d}x', \mathrm{d}y') := \int_{\mathcal{X}\times\mathcal{Y}} \gamma_{1,x}(\mathrm{d}x')\gamma_{2,y}(\mathrm{d}y')\pi(\mathrm{d}x, \mathrm{d}y).$$

Proposition 2.22 (Eckstein and Pammer (2024), Lemma 3.4)**.** *Let $\mathbb{X}_1, \mathbb{X}_2, \mathbb{X}_1', \mathbb{X}_2'$ be four adapted processes. Assume $\pi \in \Pi_{\mathfrak{c}}(\mathbb{X}_1, \mathbb{X}_2)$, $\gamma_1 \in \Pi_{\mathfrak{c}}(\mathbb{X}_1', \mathbb{X}_1)$, and $\gamma_2 \in \Pi_{\mathfrak{c}}(\mathbb{X}_2, \mathbb{X}_2')$. Then the shadowing of π under (γ_1, γ_2) is a causal coupling between $\mathbb{X}_1'$ and $\mathbb{X}_2'$.*

We summarize Propositions 2.19 and 2.22 in the commutative diagram 2.3 below.

$$\begin{tikzcd}
 & \mathbb{X}_2 \arrow[dr, bend left, "\pi_2"] & \\
\mathbb{X}_1 \arrow[ur, bend left, "\pi_1"] \arrow[rr, "\pi_3"'] & & \mathbb{X}_3
\end{tikzcd}$$

(a) Gluing.

$$\begin{tikzcd}
\mathbb{X}_1' \arrow[r, "\pi'"] \arrow[d, "\gamma_1"'] & \mathbb{X}_2' \\
\mathbb{X}_1 \arrow[r, "\pi"'] & \mathbb{X}_2 \arrow[u, "\gamma_2"']
\end{tikzcd}$$

(b) Shadowing.

Figure 2.3: Gluing and shadowing of causal couplings. The arrows represent the direction of the causality.

The following proposition establishes the equivalence between two classical viewpoints in stochastic analysis for naturally filtered processes. The first perspective considers various stochastic processes on an abstract probability space, while the second, or canonical, perspective considers various probability measures on the fixed space of all possible sample paths.

Proposition 2.23. *Let $\mu, \nu \in \mathscr{P}_p(\mathcal{X})$ and $\mathbb{X}_i = (\Omega^{\mathbb{X}_i}, \mathcal{F}^{\mathbb{X}_i}, \mathbf{F}^{\mathbb{X}_i}, P^{\mathbb{X}_i}, X_i)$ be two naturally filtered processes with $(X_1)_\# P^{\mathbb{X}_1} = \mu$ and $(X_2)_\# P^{\mathbb{X}_2} = \nu$. Then we have*

$$\mathcal{AW}_p(\mu, \nu) = \mathcal{AW}_p(\mathbb{X}_1, \mathbb{X}_2).$$

In particular, if X_1 and X_2 have the same law, then $\mathcal{AW}_p(\mathbb{X}_1, \mathbb{X}_2) = 0$.

Proof. Let $\gamma_1 = (\mathrm{Id}_{\Omega^{\mathbb{X}_1}}, X_1)_\# P^{\mathbb{X}_1}$ and $\gamma_2 = (X_2, \mathrm{Id}_{\Omega^{\mathbb{X}_2}})_\# P^{\mathbb{X}_2}$. Notice that we have $X_1 = \mathrm{Id}_{\mathcal{X}}$ γ_1-a.s., where we naturally lift both processes to the product probability space $\Omega^{\mathbb{X}_1} \times \mathcal{X}$. This implies that X_1 and $\mathrm{Id}_{\mathcal{X}}$ generate the same γ_1-completed natural filtration, i.e., ${}^{\gamma_1}(\{\emptyset, \Omega^{\mathbb{X}_1}\} \otimes \mathcal{F}_t) = {}^{\gamma_1}(\mathcal{F}_t^{\mathbb{X}_1} \otimes \{\emptyset, \mathcal{X}\})$. By Proposition 2.17 (ii), we deduce that γ_1 is a bi-causal coupling. Similarly, we have γ_2 is a bi-causal coupling.

By Proposition 2.22 for any bi-causal coupling $\pi \in \Pi_{\mathfrak{bc}}(\mathbb{X}_1, \mathbb{X}_2)$ its shadowing under (γ_1, γ_2) gives a bi-causal coupling $\pi' \in \Pi_{\mathfrak{bc}}(\mu, \nu)$, and vice versa. Moreover, it satisfies that $(X_1, X_2)_{\#}\pi = \pi'$. Therefore, we deduce $\mathcal{AW}_p(\mu, \nu) = \mathcal{AW}_p(\mathbb{X}_1, \mathbb{X}_2)$.

□

Chapter 3

A transfer principle for computing adapted Wasserstein distances

3.1 Introduction

Stochastic processes, the building block of stochastic analysis, can be viewed as path-valued random variables. From this perspective, the convergence of stochastic processes can naturally be induced by the weak convergence of their laws as probability measures on the path space. However, this 'static' viewpoint turns out to be insufficient for 'dynamic' problems, especially for many key applications in mathematical finance and beyond. In particular, the value of a stochastic optimal stopping problem is not continuous with respect to this weak topology (Backhoff-Veraguas et al., 2020a, 2022b). Different notions of adapted topologies have been proposed to refine the weak topology. We focus on the adapted Wasserstein distance which was first introduced in Lassalle (2018) as a dynamic counterpart of the Wasserstein distance for stochastic processes. We fix a Polish probability space $(\Omega, \mathcal{F}, P)$ throughout this chapter. Recall Definition 2.14 and the discussion above, for two naturally filtered stochastic processes X_1 and X_2 on $(\Omega, \mathcal{F}, P)$, their adapted 2-Wasserstein distance is given by

$$\mathcal{AW}_2(X_1, X_2) := \inf_{\pi \in \Pi_{\mathfrak{bc}}(X_1, X_2)} E_\pi[d(X_1, X_2)^2]^{1/2}, \tag{3.1}$$

where we will take d as the L^2 distance on the path space throughout this chapter, and $\Pi_{\mathfrak{bc}}(X_1, X_2)$ is the set of bi-causal couplings between X_1 and X_2. It has been applied to the analysis of various aspects of robust finance such as stability (Backhoff-Veraguas et al., 2020a), sensitivity (Bartl and Wiesel, 2023, Mirmominov and Wiesel, 2024), and model risk (Han, 2025, Sauldubois and Touzi, 2024). However, computing the adapted Wasserstein distance analytically, or even numerically, is difficult, due to the additional causality constraint. Even in discrete time, few explicit formulas have been obtained for the adapted Wasserstein distance, see Gunasingam and Wong (2025), Acciaio et al. (2024), Backhoff-Veraguas et al. (2017), etc. In continuous time, to the best of our knowledge, there has been little to no results beyond the semi-martingale framework, see Lassalle (2018), Bion–Nadal and Talay (2019), Backhoff-Veraguas et al. (2022b), etc.

In this chapter, we leverage a simple yet effective transfer principle to compute the explicit adapted Wasserstein distance between Gaussian processes and identify the optimal coupling between fractional stochastic differential equations. Let X_i, Y_i be natural filtered stochastic processes for $i = 1, 2$. Given a transport map T_i such that $X_i = T_i(Y_i)$ and X_i, Y_i generate the same natural filtration, it immediately follows from Definition 2.13 that $\Pi_{\mathfrak{bc}}(X_1, X_2) = \Pi_{\mathfrak{bc}}(Y_1, Y_2)$ and

$$\mathcal{AW}_2(X_1, X_2) = \inf_{\Pi_{\mathfrak{bc}}(Y_1,Y_2)} E_\pi[d(T_1(Y_1), T_2(Y_2))^2]^{1/2}. \tag{3.2}$$

This principle transfers the original transport problem from X_1 and X_2 to Y_1 and Y_2 which could have a much simpler structure. In particular, if Y_1 and Y_2 have independent marginals, then under any bi-causal coupling, one can verify

$$Y_1(t) \text{ is independent of } Y_2(s) \text{ for distinct } s, t \in I. \tag{3.3}$$

In Backhoff-Veraguas et al. (2022b), this principle has been already applied to transfer bi-causal couplings between SDEs to bi-causal couplings between Brownian motions. To illustrate the idea, we consider a simpler example of discrete-time Gaussian processes from Gunasingam and Wong (2025). Let $X_i \sim \mathcal{N}(0, \Sigma_i)$ be an N-step 1D non-degenerate Gaussian process. We construct $X_i = K_i Y_i$ where K_i is the Cholesky decomposition of Σ_i and $Y_i \sim \mathcal{N}(0, \mathrm{Id}_N)$ is a standard Gaussian. Indeed, X_i and Y_i generate the same natural filtration as K_i is lower triangular and invertible. By applying the transfer principle and (3.3), we can calculate $\mathcal{AW}_2(X_1, X_2)$ as

$$\begin{aligned}
\mathcal{AW}_2(X_1, X_2)^2 &= \mathrm{tr}(\Sigma_1 + \Sigma_2) - 2 \sup_{\pi \in \Pi_{\mathfrak{bc}}(X_1,X_2)} E_\pi[\langle X_1, X_2 \rangle] \\
&= \mathrm{tr}(\Sigma_1 + \Sigma_2) - 2 \sup_{\pi \in \Pi_{\mathfrak{bc}}(Y_1,Y_2)} E_\pi[\langle K_1 Y_1, K_2 Y_2 \rangle] \\
&= \mathrm{tr}(\Sigma_1 + \Sigma_2) - 2 \sup_{\pi \in \Pi_{\mathfrak{bc}}(Y_1,Y_2)} \sum_{n=1}^{N} (K_1^* K_2)_{n,n} E_\pi[Y_1(n) Y_2(n)].
\end{aligned}$$

This gives $\mathcal{AW}_2(X_1, X_2)^2 = \mathrm{tr}(\Sigma_1 + \Sigma_2) - 2 \sum_{n=1}^{N} |(K_1^* K_2)_{n,n}|$ as choosing $E_\pi[Y_1(n) Y_2(n)]$ to match the sign of the diagonal element $(K_1^* K_2)_{n,n}$ attains the supremum. Heuristically, we can view Y_i as a 'nicer' coordinate system which leads to a 'nicer' parameterization of the set of bi-causal couplings, and hence simplifies the computation.

Our first main result extends the above example to a continuous-time setting and computes the adapted Wasserstein distance between mean-square continuous Gaussian processes. To apply the transfer principle, in Section 3.3, we introduce a notion of 'canonical causal factorization' as an infinite-dimensional analogue of the Cholesky decomposition for operators on Hilbert spaces. This notion naturally bridges an algebraic object 'nest algebra' (Davidson, 1988) and a probabilistic object 'canonical representation' (Hida, 1960) of Gaussian processes. Our results give an explicit formula of the adapted Wasserstein distance in terms of the canonical causal factorization of the covariance operator, or equivalently, of the canonical representation of the Gaussian process, see Theorem 3.18. For example, any fractional Brownian motion B_H has a Molchan–Golosov representation given by $B_H(t) = \int_0^t k_H(t,s)\,\mathrm{d}B(s)$, where $H \in (0,1)$ is the Hurst parameter, k_H is the Molchan–Golosov kernel (Molchan and Golosov, 1969, Decreusefond and Üstünel, 1999)

$$k_H(t,s) = \Gamma(H+1/2)^{-1}(t-s)^{H-1/2}F(H-1/2, 1/2-H, H+1/2, 1-t/s)\mathbb{1}_{\{s\le t\}},$$

$F(a,b,c,z)$ is the Gaussian hypergeometric function $F(a,b,c,z) = \sum_{n=0}^{\infty} \frac{(a)_n(b)_n}{(c)_n}\frac{z^n}{n!}$, and $(x)_n = \Gamma(x+n)/\Gamma(x)$ is the Pochhammer symbol. We have the following result as a direct application of Theorem 3.18.

Theorem 3.1. *Let B_{H_i} be the fractional Brownian motion with Hurst parameter $H_i \in (0,1)$. Then the adapted 2-Wasserstein distance between B_{H_1} and B_{H_2} is given by*

$$\mathcal{AW}_2(B_{H_1}, B_{H_2})^2 = \int_0^T \int_0^T (k_{H_1}(t,s) - k_{H_2}(t,s))^2\,\mathrm{d}t\,\mathrm{d}s.$$

Moreover, the optimal coupling is given by the synchronous coupling between B_{H_1} and B_{H_2}, i.e., they are driven by the same Brownian motion in their Molchan–Golosov representations.

Our second result considers the adapted Wasserstein distances between fractional stochastic differential equations. By applying the transfer principle, we reformulate the adapted Wasserstein distance as a stochastic optimal control problem of fractional SDEs. The control only appears as the correlation between the driving noises. We show the optimality of the synchronous coupling by adapting the path-dependent HJB equation framework from Viens and Zhang (2019), see Theorem 3.34. In particular, for SDEs driven by fractional Brownian motions, we have the following result.

Theorem 3.2. *Let X_i be the solution of the following fractional SDE*

$$X_i(t) = x_i + \int_0^t b_i(X_i(s))\,\mathrm{d}s + \int_0^t \sigma_i(X_i(s))\,\mathrm{d}B_{H_i}(s),$$

where B_{H_i} is the fractional Brownian motion with Hurst parameter $H_i \in (1/2, 1)$. We assume that $b_i, \sigma_i \in C^2$ with bounded first and second derivatives, and b_i'', σ_i'' are uniformly continuous. Moreover, σ_i is positive, bounded, and bounded away from zero. Then, $\mathcal{AW}_2(X_1, X_2)$ is attained by the synchronous coupling between B_{H_1} and B_{H_2}.

Admittedly, the regularity constraint in the above result is not optimal, as is often the case in classical stochastic control theory, where strong assumptions are imposed to ensure the verification theorem. In a forthcoming work, we aim to relax the regularity constraint through a time-discretization approximation in the spirit of Backhoff-Veraguas et al. (2022b) and extend results to stochastic Volterra equations with monotone kernels.

To the best of our knowledge, this is the first work to investigate the adapted Wasserstein distance between fractional processes. We stress that these processes are neither semi-martingales nor Markovian, which precludes a direct application of techniques from the existing literature. Their ability to capture long-range dependence and rough path behavior has led to impactful applications, notably in finance (Baillie, 1996, Rogers, 1997, Cont, 2005), in physics (Metzler and Klafter, 2000), in engineering (Lévy-Véhel et al., 2005), and filtering theory (Decreusefond and Üstünel, 1998).

3.1.1 Related literature

We review the existing literature on the computation of adapted Wasserstein distances. For broader literature related causal optimal transport problems, we refer readers to Backhoff-Veraguas et al. (2020b), Bartl et al. (2024, 2025a) and references therein. In discrete time, Gunasingam and Wong (2025) computed explicitly the adapted Wasserstein distance between two 1D Gaussian processes. More recently, Acciaio et al. (2024) extended the previous result to multi-dimensional Gaussian processes and also considered an entropic regularization. Both of these results leveraged a dynamic programming principle from Backhoff-Veraguas et al. (2017), which is distinct from the transfer principle considered in this work. Instead of computing the explicit formula of the adapted Wasserstein distance, the Knothe–Rosenblatt coupling is identified as the optimal coupling between co-monotone distributions in discrete time (Rüschendorf, 1985, Backhoff-Veraguas et al., 2017). In continuous time, it is shown in Lassalle (2018), for a Cameron–Martin cost, the adapted Wasserstein distance between an arbitrary probability measure and the Wiener measure is equal to the square-root of its relative entropy with respect to the Wiener measure. For L^2 cost, it is shown in Bion–Nadal and Talay (2019) and later in Backhoff-Veraguas et al. (2022b), Robinson and Szölgyenyi (2024) that the synchronous coupling is the optimal coupling between two 1D SDEs.

Another line of research is to numerically compute the adapted Wasserstein distance by approximation or regularization. These results are mainly in a discrete-time setting. For instance, Eckstein and Pammer (2024) proposed numerical algorithms to compute the entropic regularized adapted Wasserstein distance. In Pflug and Pichler (2016), Backhoff-Veraguas et al. (2022a), Acciaio and Hou (2024), the authors studied various smoothed adapted empirical measures and derived the convergence rate to their limit under the adapted Wasserstein distance.

The notion of causality underpinning the adapted Wasserstein distance, when placed in the context of linear transformations between (finite-dimensional) vector spaces, naturally corresponds to the triangularity of these transformations. A suitable generalization of these triangular forms to Hilbert spaces is the nest algebra, which originates from the work of Ringrose (1965). The nest algebra is a prime example of non-selfadjoint algebras and reflexive algebras in the sense of Arveson (1974). Early research focused on the structure of compact operators in nest algebras, see Ringrose (1962), Erdos (1968), etc. Further developments include the characterization of the radical (Ringrose, 1965), unitary invariants (Erdos, 1967), and similarity invariants (Larson, 1985). The causal factorization introduced in Section 3.3 is motivated by several pioneering works (Pitts, 1988, Anoussis and Katsoulis, 1997, 1998). We refer interested readers to Davidson (1988) for a more complete reference.

In order to study the prediction theory of Gaussian processes, Lévy (1956) introduced the canonical representation of a Gaussian process, which provides a full description of its natural filtration. This canonical representation and the related notion of multiplicity was systematically investigated in Hida (1960), Hida and Hitsuda (1993), Hitsuda (1968) and extended by Cramér (1971) to general stochastic processes. In the sequel, we clarify the connection between the canonical representation of Gaussian processes and the (canonical) causal factorization of their covariance operators.

3.1.2 Outline

The rest of the chapter is organized as follows. In Section 3.2, we recall basic definitions and properties of the adapted Wasserstein distance and the canonical representation of Gaussian processes. In Section 3.3, we introduce the (canonical) causal factorization and discuss its existence and uniqueness. A characterization of the Gaussian Volterra processes is given in Corollary 3.10 which we believe is of independent interest. In Section 3.4, we apply the transfer principle to compute the adapted 2-Wasserstein distance between Gaussian processes. An explicit formula for the distance is given in Theorems 3.18 and 3.25 for the unit multiplicity case and the higher multiplicity case respectively. An optimal coupling is identified in both cases. In Theorem 3.29, we consider the best martingale approximation to a fractional Brownian motion with respect to the adapted 2-Wasserstein distance. In Section 3.5, we study the adapted Wasserstein distance between fractional SDEs via a stochastic control reformulation. We establish a verification theorem for additive fractional SDEs and reduce the multiplicative case into the additive case via a Lamperti transform (Lamperti, 1964). Some technical estimates are postponed to Section 3.6.

3.2 Preliminaries

3.2.1 Notations

Let μ, μ_1, μ_2 be positive measures on $[0,T]$. We write $H_\mu = L^2([0,T], \mu; \mathbb{R})$ and $\langle \cdot, \cdot \rangle_\mu$ as the inner product on H_μ with the induced norm $\| \cdot \|_\mu$. We write $H_{\mu,t} = \{f \in H_\mu : \operatorname{supp}(f) \subseteq [0,t]\}$ as a closed subspace of H_μ. We equip H_μ with the Borel σ-algebra $\mathcal{B}(H_\mu)$ and its natural filtration $\mathbf{H}_\mu = (\mathcal{H}_{\mu,t})_{t\in[0,T]}$, where $\mathcal{H}_{\mu,t} := \sigma(f \in H_{\mu,t})$. Here, we identify $H_{\mu,t}$ with its dual $H^*_{\mu,t}$. By $\mathfrak{B}(H_{\mu_1}, H_{\mu_2})$ we denote the set of bounded linear operator $A : H_{\mu_1} \to H_{\mu_2}$, and we write $\mathfrak{B}(H_\mu) = \mathfrak{B}(H_\mu, H_\mu)$. Given a closed subspace $N \subseteq H_\mu$, we denote the orthogonal projection onto N by P_N. Let S be a subset of a normed linear space. By $\overline{\operatorname{span}(S)}$, we denote the closure of the linear span of S.

We say an operator $A \in \mathfrak{B}(H_\mu)$ is positive if $\langle Af, f\rangle_\mu \geq 0$ for any $f \in H_\mu$. We say an operator $A \in \mathfrak{B}(H_\mu)$ is trace-class, if $\|A\|_{\mathrm{tr}} := \sum_{k\geq 1} \langle |A| e_k, e_k \rangle_\mu$ is finite for an orthonormal basis $(e_k)_{k\geq 1}$ of H_μ. An operator $K : H_{\mu_1} \to H_{\mu_2}$ is Hilbert–Schmidt, if KK^* is trace-class where K^* is the dual operator of K. Its Hilbert–Schmidt norm is defined as $\|K\|_{\mathrm{HS}} = \sqrt{\operatorname{tr}(KK^*)}$. We denote the set of Hilbert–Schmidt operators from H_{μ_1} to H_{μ_2} by $\mathfrak{B}_2(H_{\mu_1}, H_{\mu_2})$. There exists an isometry from $\mathfrak{B}_2(H_{\mu_1}, H_{\mu_2})$

to $L^2([0,T]^2, \mu_1 \otimes \mu_2; \mathbb{R})$. In fact, every Hilbert–Schmidt operator K has a kernel $k \in L^2([0,T]^2, \mu_1 \otimes \mu_2; \mathbb{R})$ such that $Kf(t) = \int_0^T k(t,s) f(s) \mu_1(\mathrm{d}s) \in H_{\mu_2}$. We omit the subscript if $\mu = \lambda$ the Lebesgue measure on $[0,T]$.

By $\mu_1 \gg \mu_2$ we denote μ_2 is absolutely continuous with respect to μ_1 and write their Radon–Nikodym derivative as $\frac{\mathrm{d}\mu_2}{\mathrm{d}\mu_1}$. We denote the geometric mean of μ_1 and μ_2 by

$$\sqrt{\mu_1 \mu_2}(\mathrm{d}t) := \sqrt{\frac{\mathrm{d}\mu_1}{\mathrm{d}(\mu_1+\mu_2)} \frac{\mathrm{d}\mu_2}{\mathrm{d}(\mu_1+\mu_2)}}(t)(\mu_1+\mu_2)(\mathrm{d}t).$$

Let $C([0,T];\mathbb{R})$ be the continuous path space. For a functional f on $C([0,T];\mathbb{R})$, we say f is Fréchet differentiable at $\omega \in C([0,T];\mathbb{R})$ if there exists a linear functional $\partial_\omega f(\omega) \in C([0,T];\mathbb{R})^*$ such that for any $\eta \in C([0,T];\mathbb{R})$ it holds

$$f(\omega + \eta) - f(\omega) = \langle \eta, \partial_\omega f(\omega) \rangle + o(\|\eta\|).$$

We call $\partial_\omega f$ the Fréchet derivative of f. Similarly, we define the second Fréchet derivative of f and denote it as $\partial_\omega^2 f$. Given two linear functionals $f, g \in C([0,T];\mathbb{R})^*$, we denote their tensor product as a bilinear functional given by

$$\langle (\eta_1, \eta_2), f \otimes g \rangle = \langle \eta_1, f \rangle \langle \eta_2, g \rangle,$$

for any $\eta_1, \eta_2 \in C([0,T];\mathbb{R})$.

3.2.2 Canonical representation of Gaussian processes

We say $X : \Omega \times [0,T] \to \mathbb{R}$ is a 1D Gaussian process if for any $t_1, \ldots, t_n \in [0,T]$, the random vector $(X(t_1), \ldots, X(t_n))$ is Gaussian. In this chapter, we focus on centered and mean-square continuous Gaussian processes, i.e.,

$$E[X(t)] = 0 \text{ for any } t \in [0,T], \text{ and } t \mapsto X(t) \in L^2(\Omega, P) \text{ is continuous.}$$

Notice that mean-square continuity of X implies X has path in $H = L^2([0,T], \lambda; \mathbb{R})$ almost surely. Hence, $X_{\#}P$ yields a Gaussian measure on H whose covariance operator $\Sigma : H \to H$ is given by $\langle \Sigma f, g \rangle := E[\langle f, X \rangle \langle g, X \rangle]$. Moreover, Σ has a unique continuous kernel $R(t,s) = E[X(t)X(s)]$ such that

$$\Sigma f(t) = \int_0^T f(s) R(t,s)\, \mathrm{d}s.$$

We say a Gaussian process X is *deterministic* if the behavior of X is completely determined by its behavior in an infinitesimal time, i.e.,

$$\bigcap_{t>0} \overline{\operatorname{span}\{X(s) : s \in [0,t]\}} = \overline{\operatorname{span}\{X(s) : s \in [0,T]\}} \subseteq L^2(\Omega, P);$$

and X is *purely nondeterministic* if the information of X must have entered as a new impulse at some definite time in the past, i.e.,

$$\bigcap_{t>0} \overline{\operatorname{span}\{X(s) : s \in [0,t]\}} = \{0\}. \tag{3.4}$$

We shall not confuse a deterministic Gaussian process with a deterministic path-valued random variable which is supported on a single path. For example, $X(t) = t\xi$ where $\xi \sim \mathcal{N}(0,1)$ is a deterministic Gaussian process but not a deterministic random variable. We remark that in Corollary 3.10, we show that (3.4) is equivalent to the condition that $\mathcal{F}_{0+}^X = \bigcap_{t>0} \mathcal{F}_t^X$ is trivial.

In what follows, we introduce the canonical representation of a Gaussian process which was initiated by Lévy (1956), and systematically studied by Hida (1960) and Cramér (1971). It states that a centered, mean-square continuous, and purely non-deterministic Gaussian process is essentially driven by a countable number of 'noises'. Such a representation is canonical in the sense that the 'noises' precisely generate the same natural filtration as the one of the Gaussian process. We adapt Hida and Hitsuda (1993, Theorem 4.1) to our setting.

Theorem 3.3. *Let X be a centered, mean-square continuous, and purely nondeterministic Gaussian process. Then there exists $N \in \mathbb{N} \cup \{\infty\}$ uniquely determined by X called the multiplicity of X, such that X has a canonical representation in the form of*

$$X(t) = \sum_{n=1}^{N} \int_0^t k^n(t,s)\, \mathrm{d}M^n(s), \tag{3.5}$$

satisfying the following conditions:

(i) $\{M^n\}_{n=1}^N$ *are independent Gaussian martingales with independent increments,*

(ii) $\mu^n(t) := [M^n](t)$ *is continuous, non-decreasing, and* $\mu^1(\mathrm{d}t) \gg \mu^2(\mathrm{d}t) \gg \cdots$,

(iii) $t \mapsto k^n(t,\cdot) \in H_{\mu^n}$ *is continuous and* $\operatorname{supp}(k^n(t,\cdot)) \subseteq [0,t]$,

(iv) $\mathbf{F}^X = \mathbf{F}^M$ *with* $M = (M_1, \dots, M_N)$.

In general, it is not easy to find the canonical representation of a Gaussian process. The following result from Hida and Hitsuda (1993, Theorem 4.4) gives a characterization of the canonical representation.

Theorem 3.4. *Let X be a Gaussian process with a representation of the form of (3.5). Then it is a canonical representation if and only if for any $T' \in [0,T]$ and $f^n \in H_{\mu^n}$,*

$$g(t) = \sum_{n=1}^{N} \int_0^t k^n(t,s) f^n(s) \mu^n(\mathrm{d}s) = 0 \textit{ for all } t \in [0,T']$$

implies $f^n = 0$ on $[0,T']$ for all n.

3.3 Causal factorization

In this section, we introduce the *causal* factorization as an analogue of Cholesky decomposition for positive operators on infinite dimensional Hilbert space. We first recall some basic properties of Cholesky decomposition. For any positive definite matrix $A \in \mathbb{R}^{N \times N}$, there exists a lower triangular matrix L such that $A = LL^*$. If A is nondegenerate, such a decomposition L is unique up to a multiplication by a diagonal matrix D with diagonal entries being $\{1, -1\}$.

It is clear that lower triangularity is not an intrinsic property, but depends on the choice of the basis. From a geometric viewpoint, let $V = \mathbb{R}^N$ and $V_n = \text{span}\{e_1, \ldots, e_n\}$, where $\mathbf{e} = \{e_n\}_{n=1}^N$ is an orthonormal basis of V. A map $A : V \to V$ is lower triangular (with respect to $\mathbf{e}$) if and only if $V_n^\perp$ is invariant under A for any $1 \leq n \leq N$.

We focus on the covariance operator Σ associated to a centered, mean-square continuous, and purely nondeterministic Gaussian process X, and denote the set of such operators as $\mathfrak{C}(H)$. Notice $\mathfrak{C}(H)$ is a proper subset of positive trace operators on H with continuous kernel.

Definition 3.5. Let $\Sigma \in \mathfrak{C}(H)$. For a positive continuous measure μ on $[0,T]$ and Hilbert–Schmidt operator $K : H_\mu \to H$, we say (K, μ) is a *causal* factorization of Σ if $\Sigma = KK^*$ and K is *causal* in the sense that $K : (H_\mu, \mathcal{H}_{\mu,t}) \to (H, \mathcal{H}_t)$ is measurable for any $t \in [0,T]$. We say (K, μ) is a canonical causal factorization if further K_t is injective for any $t \in [0,T]$, where $K_t := P_{H_t} K|_{H_{\mu,t}}$.

The following property justifies the causality condition as a natural extension to the lower traingularity in continuous time context.

Proposition 3.6. *Let $K : (H_\mu, \mathbf{H}_\mu) \to (H, \mathbf{H})$ be a bounded linear operator. Then K is causal if and only if K maps $H_{\mu,t}^\perp$ into $H_t^\perp$ for any $t \in [0,T]$.*

Proof. Let us first assume that K is causal. For $h \in H_t$, we write $K^*(h) = f + g$, where $f \in H_{\mu,t}^{\perp}$ and $g \in H_{\mu,t}$. We notice that from the causality of K

$$
\begin{aligned}
K^{-1}(\{x \in H : \langle h, x\rangle \leq 0\}) &= \{x \in H_\mu : \langle h, K(x)\rangle \leq 0\} \\
&= \{x \in H_\mu : \langle f + g, x\rangle_\mu \leq 0\} \in \mathcal{H}_{\mu,t}.
\end{aligned}
$$

Since $\mathcal{H}_{\mu,t} = \sigma(\{h \in H_\mu : \operatorname{supp}(h) \subseteq [0,t]\})$, we derive $H_{\mu,t}^{\perp} \subseteq U$ for any $U \in \mathcal{H}_{\mu,t}$. In particular, $H_{\mu,t}^{\perp} \subseteq \{x \in H_\mu : \langle f+g, x\rangle_\mu \leq 0\}$ and hence $f = 0$. Therefore, $P_{H_{\mu,t}^{\perp}} K^* P_{H_t} = 0$. Taking the adjoint on both sides, we deduce $P_{H_t} K P_{H_{\mu,t}^{\perp}} = 0$, i.e., K maps $H_{\mu,t}^{\perp}$ into $H_t^{\perp}$.

On the other hand, if $H_{\mu,t}^{\perp}$ is mapped into H_t under K for any $t \in [0,T]$, then $H_t^{\perp}$ is mapped into $H_{\mu,t}$ under K^*. For any $h \in H_t^{\perp}$ and $r \in \mathbb{R}$, we have

$$
K^{-1}(\{x \in H : \langle h, x\rangle \leq r\}) = \{x \in H_\mu : \langle h, K(x)\rangle_\mu \leq r\} = \{x \in H_\mu : \langle K^*(h), x\rangle_\mu \leq r\}.
$$

The causality follows directly from the fact that $K^*(h) \in H_{\mu,t}$. $\square$

Remark 3.7. When $\mu = \lambda$, the set of operators $K : H_\mu \to H_\mu$ which leaves $H_{\mu,t}^{\perp}$ invariant forms a non-selfadjoint algebra. This algebra is called the nest algebra first introduced in Ringrose (1965), and we denote it as $\mathfrak{N}(H_\mu)$. The diagonal algebra $\mathfrak{D}(H_\mu)$ is a subalgebra of $\mathfrak{N}(H_\mu)$ consisting of operators K such that both K and K^* are in $\mathfrak{N}(H_\mu)$. See Davidson (1988) for a detailed reference.

3.3.1 Existence

We investigate the existence of causal factorization. Similar to Cholesky decomposition, it does exist for any $\Sigma \in \mathfrak{C}(H)$. The proof is based on a factorization result (Anoussis and Katsoulis, 1998, Theorem 13) in nest algebra.

Proposition 3.8. *Let μ be a positive continuous measure, and $\mathcal{R}_\mu = \{\operatorname{range}(A) : A \in \mathfrak{N}(H_\mu)\}$. For any $A \in \mathfrak{B}(H_\mu)$, there exists $B \in \mathfrak{N}(H_\mu)$ such that $AA^* = BB^*$ if and only if* $\operatorname{range}(A) \in \mathcal{R}_\mu$.

Theorem 3.9. *Let $\Sigma \in \mathfrak{C}(H)$. Then there exists a causal factorization (K, μ) of Σ.*

Proof. Let X be a centered, mean-square continuous, and purely nondeterministic Gaussian process associated to Σ. By Theorem 3.3, we have a canonical representation of X given by

$$
X(t) = \sum_{n=1}^{N} \int_0^t k^n(t,s)\, \mathrm{d}M^n(s). \tag{3.6}
$$

Recall we write $\mu^n(\mathrm{d}t) = [M^n](\mathrm{d}t)$, and $\mu^1 \gg \mu^2 \gg \cdots$. Let $\mu = \lambda + \mu^1$. Since $\Sigma \in \mathfrak{C}(H)$, we notice Σ uniquely determines a continuous kernel given by $R(t,s) = E[X(t)X(s)]$. Hence, it uniquely induces an operator $\Sigma_\mu : H_\mu \to H_\mu$ given by

$$\Sigma_\mu f(t) = \int_0^T f(s)R(t,s)\mu(\mathrm{d}s).$$

The representation (3.6) yields a representation of Σ_μ as $\Sigma_\mu = \sum_{n=1}^N K^n(K^n)^*$, where $K^n : H_\mu \to H_\mu$ is given by

$$K^n f(t) = \int_0^t k^n(t,s)\sqrt{\frac{\mathrm{d}\mu^n}{\mathrm{d}\mu}}(s)f(s)\mu(\mathrm{d}s).$$

In particular, $K^n \in \mathfrak{N}(H_\mu)$. If the multiplicity N was finite, then we could apply Anoussis and Katsoulis (1998, Proposition 27) which states the sum of two factorizable operators can still be factored in the nest algebra. This would give us a $K_\mu \in \mathfrak{N}(H_\mu)$ such that $\Sigma_\mu = K_\mu(K_\mu)^*$. We could construct $K : H_\mu \to H$ as

$$Kf(t) = \int_0^t k_\mu(t,s)\sqrt{\frac{\mathrm{d}\lambda}{\mathrm{d}\mu}}(s)f(s)\mu(\mathrm{d}s),$$

where k_μ is the kernel of K_μ. Then it is direct to verify (K,μ) would be a causal factorization of Σ.

Now, we proceed with the case $N = \infty$. The spirit of the proof aligns with Anoussis and Katsoulis (1998, Proposition 27), but we extend it to a countable sum of operators. By Proposition 3.8, it suffices to show $\mathrm{range}(\Sigma_\mu^{1/2}) \in \mathcal{R}_\mu$. We define $T : \bigoplus_{n=1}^\infty H_\mu \to \bigoplus_{n=1}^\infty H_\mu$ as

$$T(f_1, f_2, \dots) = \left(\sum_{n=1}^\infty K^n f_n, 0, \dots\right).$$

Notice that T is a bounded linear operator and $\mathrm{range}(T) = \mathrm{range}((TT^*)^{1/2})$ by Douglas (1966, Theorem 1). This yields $\mathrm{range}(\Sigma_\mu^{1/2}) = \sum_{n=1}^\infty \mathrm{range}(K^n)$. We construct a sequence of partial isometries $\{U_n\}_{n=1}^\infty$ in $\mathfrak{N}(H_\mu)$ with full range and mutually orthogonal initial spaces. Let $e_n \in H_\mu$ with $\mathrm{supp}(e_n) \in [\frac{1}{n+1}, \frac{1}{n}]$. We take $\{I_{n,m}\}_{n,m=1}^\infty$ where $I_{n,m}$ are infinite and mutually disjoint subsets of $\mathbb{Z}_+$. We define closed subspaces of H_μ by $E_{n,m} := \overline{\mathrm{span}\{e_k : k \in I_{n,m}, k > m\}}$ and $F_m := \{f \in H_\mu : \mathrm{supp}(f) \subseteq [\frac{1}{m+1}, \frac{1}{m}]\}$. In particular, $E_{n,m}$ are mutually orthogonal. Since $E_{n,m}$ is infinite dimensional, we can find a partial isometry $P_{n,m} \in \mathfrak{N}(H_\mu)$ with initial space $E_{n,m}$ and range F_m. By taking $U_n = \sum_{m=1}^\infty P_{n,m}$, we have $U_n \in \mathfrak{N}(H_\mu)$ with full range and mutually orthogonal initial spaces. Therefore, $\sum_{n=1}^\infty \mathrm{range}(K^n) = \mathrm{range}(\sum_{n=1}^\infty K^n U_n) \in \mathcal{R}_\mu$. We conclude the proof by applying Proposition 3.8 and notice $\mathrm{range}(\Sigma_\mu^{1/2}) \in \mathcal{R}_\mu$. $\square$

We say a process X is Gaussian Volterra if there exists a Volterra representation $X(t) = \int_0^t k(t,s)\,\mathrm{d}M(s)$ with M a continuous Gaussian martingale with independent increments. The above result gives a characterization of mean-square continuous Gaussian Volterra processes.

Corollary 3.10. *Let X be a centered, mean-square continuous Gaussian process. The following statements are equivalent:*

(i) $\mathcal{F}_{0+}^X = \bigcap_{t>0} \mathcal{F}_t^X$ *is trivial.*

(ii) X *is purely nondeterministic.*

(iii) There exists a Gaussian Volterra process $\widetilde{X}$ *such that* X *and* $\widetilde{X}$ *share the same law.*

Proof. (i) $\Rightarrow$ (ii). Notice that $\overline{\operatorname{span}\{X(s) : s \le t\}} \subseteq \mathcal{F}_t^X$. Therefore, $\mathcal{F}_{0+}^X$ is trivial implies that $\bigcap_{t>0} \overline{\operatorname{span}\{X(s) : s \le t\}} = \{0\}$, and hence X is purely nondeterministic.

(ii) $\Rightarrow$ (iii). If X is purely nondeterministic, by Theorem 3.9 there exists a causal factorization (K,μ) of Σ, the covariance operator of X. We can take a Gaussian Volterra process $\widetilde{X}(t) = \int_0^t k(t,s)\,\mathrm{d}M(s)$, where k is the kernel of K, and M is a Gaussian martingale with independent increments and $\mu(\mathrm{d}t) = [M](\mathrm{d}t)$. It is clear that $\widetilde{X}$ has the same covariance operator as X, and hence they share the same law.

(iii) $\Rightarrow$ (i). $\widetilde{X}(t) = \int_0^t k(t,s)\,\mathrm{d}M(s)$ is a Gaussian Volterra process and shares the same law as X. In particular, M is continuous Gaussian martingale with independent increments, and it is a deterministic continuous time change of the standard Brownian motion. Therefore, $\mathcal{F}_{0+}^{\widetilde{X}} \subseteq \mathcal{F}_{0+}^M$ is trivial, and so as $\mathcal{F}_{0+}^X$.

□

On the other hand, a canonical causal factorization does not always exist. In particular, the following result links the canonical causal factorization to Gaussian processes with unit multiplicity.

Theorem 3.11. *Let $\Sigma \in \mathfrak{C}(H)$ and be associated to a Gaussian process X. The following statements are equivalent:*

(i) X *is of unit multiplicity and has a canonical representation* $X(t) = \int_0^t k(t,s)\,\mathrm{d}M(s)$.

(ii) Σ *has a canonical causal factorization* (K,μ).

(iii) Σ *has a causal factorization* (K,μ) *such that the span of* $\{k(r,\cdot) : r \in [0,t]\}$ *is dense in* $H_{\mu,t}$ *for any* $t \in [0,T]$*, where* k *is the kernel of* K.

Proof. (*i*) $\Leftrightarrow$ (*ii*). We notice by Theorem 3.4, $X(t) = \int_0^t k(t,s)\,\mathrm{d}M(s)$ is a canonical representation, if and only if for any $T' \in [0,T]$ and $f \in H_\mu$,

$$g(t) = \int_0^t k(t,s)f(s)\mu(\mathrm{d}s) = 0 \text{ for all } t \in [0,T']$$

implies $f = 0$ on $[0,T']$. This is equivalent to the injectivity of $K_{T'} = P_{H_{T'}}K|_{H_{\mu_{T'}}}$ for any $T' \in [0,T]$ where K is given by $Kf(t) = \int_0^t k(t,s)f(s)\mu(\mathrm{d}s)$. Since $K \in \mathfrak{N}(H_\mu)$, this is further equivalent to (K,μ) is a canonical causal factorization of Σ.

(*ii*) $\Leftrightarrow$ (*iii*). Notice $K_t = P_{H_t}K|_{H_{\mu,t}}$ is injective if and only if the range of K_t^* is dense in $H_{\mu,t}$. Since $\Sigma \in \mathfrak{C}(H)$, $R(t,s) = E[X(t)X(s)] = \int_0^{t\wedge s} k(t,r)k(s,r)\mu(\mathrm{d}r)$ is continuous. This implies $r \mapsto k(r,\cdot) \in H_\mu$ is continuous. Therefore, $\mathrm{range}(K_t^*) = \{f(s) = \int_0^t k(r,s)g(r)\,\mathrm{d}r : g \in H_t\}$ is dense if and only if the span of $\{k(r,\cdot) : r \in [0,t]\}$ is dense in $H_{\mu,t}$ ☐

3.3.2 Uniqueness

In the finite dimensional case, Cholesky decomposition is unique up to a diagonal matrix. This is saying for nondegenerate, lower-triangular matrices K_1, K_2 satisfying $K_1K_1^* = K_2K_2^*$, there exists a diagonal matrix D such that $K_1 = K_2D$. However, it is not the case for the causal factorization.

Proposition 3.12. *Let $\Sigma \in \mathfrak{C}(H)$ and (K,μ) be a causal factorization of Σ. For any partial isometry $U \in \mathfrak{N}(H_\mu)$ with range dense in H_μ, (KU,μ) is again a causal factorization of Σ. Moreover, U is not necessarily in the diagonal algebra $\mathfrak{D}(H_\mu)$, i.e., U^* is not necessarily in $\mathfrak{N}(H_\mu)$.*

Proof. Since U is a partial isometry with a dense range, we have $UU^* = \mathrm{Id}$ on H_μ. Therefore, (KU,μ) is a causal factorization of Σ. For example, we can take U as in the proof of Theorem 3.9. And in particular, U is not diagonal. ☐

Such non-uniqueness generates non-canonical representations of the same Gaussian process. In the following example, we include Levy's non-canonical representation of Brownian motion (Lévy, 1957).

Example 3.13. We define a partial isometry U on H given by $U^*\mathbb{1}_{[0,t]}(s) = [3 - 12(s/t)+10(s/t)^2]\mathbb{1}_{[0,t]}(s)$. It is direct to verify $X(t) = \int_0^t\{3-12(s/t)+10(s/t)^2\}\,\mathrm{d}B(s)$ is a Brownian motion. However, this representation is not canonical. Notice $X(t)$ is independent of $\int_0^T s\,\mathrm{d}B(s)$, which implies $\mathcal{F}_T^X \subsetneq \mathcal{F}_T^B$.

If we restrict ourselves to the canonical causal factorization, we retrieve a uniqueness result analogous to the one for Cholesky decomposition.

Proposition 3.14. *Let* $\Sigma \in \mathfrak{C}(H)$*. Assume* (K_1, μ) *and* (K_2, μ) *are two canonical causal factorizations of* Σ*. Then, there exists a diagonal operator* $D \in \mathfrak{D}(H_\mu)$ *such that* $K_1 = K_2 D$*. Moreover,* D *is a multiplication operator given by* $Df(t) = (\mathbb{1}_S(t) - \mathbb{1}_{[0,T]\setminus S}(t))f(t)$ *for a measurable set* $S \subseteq [0, T]$*.*

Proof. Since (K_2, μ) is canonical, we have K_2 is injective and hence $\overline{\mathrm{range}(K_2^*)} = H_\mu$. Since $K_1 K_1^* = K_2 K_2^*$, we deduce K_1^* and K_2 share the same null space. We can define an operator $\widetilde{D}$ from range K_2^* to range(K_1^*) such that $\widetilde{D}(K_2^* f) = K_1^* f$. Moreover, $\widetilde{D}$ can be uniquely extended to an operator on $H_\mu = \overline{\mathrm{range}(K_2^*)}$. Therefore, by taking $D = \widetilde{D}^*$, we derive $K_1 = K_2 D$. Noticing $K_2 D K_1^* = K_1 K_1^* = K_2 K_2^*$ and K_2 is injective, we deduce $DK_1^* = K_2^*$ and $D^* = K_1^{-1} K_2$. This yields that D is an orthogonal operator on H_μ.

Now, we consider two canonical representations induced by K_1 and K_2

$$X(t) = \int_0^T k_1(t, s)\,\mathrm{d}M_1(s) = \int_0^T k_2(t, s)\,\mathrm{d}M_2(s),$$

where k_i is the kernel of K_i. Since M_1 and M_2 generate the same filtration as X, M_1 is a $\mathbf{F}^{M_2}$–martingale. Moreover, M_2 is a continuous Ocone martingale with deterministic quadratic variation. Therefore, by martingale representation theorem Vostrikova and Yor (2007, Proposition), we have $M_1(t) = \int_0^t \rho(s)\,\mathrm{d}M_2(s)$ for some predictable process $\rho(s)$ taking value in $\{-1, 1\}$. Together with the fact that $K_2 = K_1 D^*$, we deduce $k_2(t, \cdot) = Dk_1(t, \cdot) = \rho(\omega, \cdot)k_1(t, \cdot) \in H_\mu$, $P(\mathrm{d}\omega)$-a.s $\lambda(\mathrm{d}t)$-a.e. This implies $\rho(\omega, \cdot) \in H_\mu$ is deterministic and has the form of $\rho(s) = \mathbb{1}_S(s) - \mathbb{1}_{[0,T]-S}(s)$. Otherwise, $k_1(t, \cdot) = 0$ on a positive measure set which contradicts the injectivity of K_1. Moreover, we notice the span of $\{k_1(t, \cdot) : t \in [0, T]\}$ is dense in H_μ as K_1 is injective, and we conclude $Df(t) = f(t)(\mathbb{1}_S(t) - \mathbb{1}_{[0,T]-S}(t))$. □

Remark 3.15. Following the same lines of arguments, we can show that all orthogonal operators O in the nest algebra $\mathfrak{N}(H_\mu)$ with $P_{H_{\mu,t}} O|_{H_{\mu,t}}$ surjective for all $t \in [0, T]$ are diagonal. This is of sharp contrast to the result of Davidson (1998) which shows the abundance of the unitary operators in a nest algebra on a complex Hilbert space. Indeed, under their setting, any contraction in $\mathfrak{N}(H_\mu)$ can be represented as a finite convex combination of unitary operators, and hence there are non-diagonal unitary operators in $\mathfrak{N}(H_\mu)$.

3.4 Gaussian processes

Before we present our main theorem, we show that we can always decompose a Gaussian process into a deterministic part and a purely nondeterministic part. These two parts are 'orthogonal', and we can calculate the adapted Wasserstein distance separately.

Lemma 3.16. *For any mean-square continuous Gaussian process X, there exists a decomposition $X = Y + Z$ where Y is purely nondeterministic and Z is deterministic. Moreover, Y and Z are independent mean-square continuous Gaussian processes. The adapted Wasserstein distance between X_1 and X_2 can be decomposed as*

$$\mathcal{AW}_2(X_1, X_2)^2 = \mathcal{AW}_2(Y_1, Y_2)^2 + \mathcal{W}_2(Z_1, Z_2)^2.$$

Remark 3.17. The Wasserstein distance between two Gaussian processes Z_1 and Z_2 is well studied (see Dowson and Landau (1982), Gelbrich (1990)), and can be calculated explicitly given the covariance operators of Z_1 and Z_2.

Proof. The first statement is a generalization of Wold decomposition to general second order stochastic processes, see Cramér (1971). The deterministic process Z is given by $Z(t) = P_{0+}X(t)$ where P_t is the orthogonal projection from $L^2(\Omega, P)$ to the closed subspace $\overline{\operatorname{span}\{X_s : 0 \leq s \leq t\}}$ and $P_{0+} = \lim_{t\to 0+} P_t$. It is direct to verify (X, Z) is jointly Gaussian and so is (Y, Z). Therefore, the independence of Y and Z follows from the orthogonality of the projection. Since mean-square continuity can be preserved by the orthogonal projection, we have Y and Z are mean-square continuous.

We proceed to show the decomposition of the adapted Wasserstein distance between X_1 and X_2. Noticing under any bi-causal coupling $\pi \in \Pi_{\mathfrak{bc}}(X_1, X_2)$, $\mathcal{F}_t^{X_1}$ is conditionally independent of $\mathcal{F}_T^{X_2}$ given $\mathcal{F}_t^{X_2}$. This implies $\mathcal{F}_T^{Z_1} = \mathcal{F}_t^{Z_1}$ is conditionally independent of $\mathcal{F}_T^{Y_2}$ given $\mathcal{F}_t^{X_2}$. Hence, we deduce

$$E_\pi[\langle Z_1, Y_2\rangle] = E_\pi[E_\pi[\langle Z_1, Y_2\rangle|\mathcal{F}_t^{X_2}]] = E_\pi[\langle E_\pi[Z_1|\mathcal{F}_t^{X_2}], E_\pi[Y_2|\mathcal{F}_t^{X_2}]\rangle].$$

Notice that $E_\pi[Y_2(\cdot)|\mathcal{F}_t^{X_2}] = E_P[X_2(\cdot) - Z_2(\cdot)|\mathcal{F}_t^{X_2}] = (P_t - P_{0+})X_2(\cdot)$. By Lebesgue dominated convergence theorem, we derive $E_\pi[\langle Z_1, Y_2\rangle] = 0$ by taking t to 0. Therefore, we have $\mathcal{AW}_2(X_1, X_2)^2 \geq \mathcal{AW}_2(Y_1, Y_2)^2 + \mathcal{AW}_2(Z_1, Z_2)^2$. Finally, noticing for deterministic process Z_i, it holds $\mathcal{F}_{0+}^{Z_i} = \mathcal{F}_T^{Z_i}$, and hence $\mathcal{AW}_2(Z_1, Z_2) = \mathcal{W}_2(Z_1, Z_2)$.

For the reverse direction, we consider the optimal bi-causal coupling $\pi_Y (\pi_Z)$ attaining the adapted Wasserstein distance between Y_1 and Y_2 (Z_1 and Z_2). Then we construct a bi-causal coupling from the independent product $\pi_Y \otimes \pi_Z$. Let

$\hat{\pi} = (Y_1 + Z_1, Y_2 + Z_2)_{\#}(\pi_Y \otimes \pi_Z)$. By Shadowing Proposition 2.22, there exists $\pi \in \Pi_{\mathfrak{bc}}(X_1, X_2)$ such that $(X_1, X_2)_{\#}\pi = \hat{\pi}$. Hence, this yields $\mathcal{AW}(X_1, X_2)^2 \leq E_\pi[\|X_1 - X_2\|^2] = \mathcal{AW}_2(Y_1, Y_2)^2 + \mathcal{W}_2(Z_1, Z_2)^2$.

□

3.4.1 Unit multiplicity

We present an explicit adapted Wasserstein distance formula for Gaussian processes of unit multiplicity.

Theorem 3.18. *Let X_i be a centered, mean-square continuous, and purely non-deterministic Gaussian process of unit multiplicity, with canonical representation $X_i(t) = \int_0^t k_i(t,s)\,\mathrm{d}M_i(s)$ for $i = 1, 2$. Then, the adapted Wasserstein distance between X_1 and X_2 is given by*

$$\begin{aligned}\mathcal{AW}_2(X_1, X_2)^2 = \int_0^T \|k_1(\cdot, s)\|^2 \mu_1(\mathrm{d}s) + \int_0^T \|k_2(\cdot, s)\|^2 \mu_2(\mathrm{d}s) \\ - 2\int_0^T |\langle k_1(\cdot, s), k_2(\cdot, s)\rangle| \sqrt{\mu_1\mu_2}(\mathrm{d}s),\end{aligned} \tag{3.7}$$

where $\mu_i(\mathrm{d}s) = [M_i](\mathrm{d}s)$.

Equivalently, let Σ_i be the covariance operator of X_i, (K_i, μ_i) be a canonical causal factorization of Σ_i. We have the adapted Wasserstein distance

$$\mathcal{AW}_2(X_1, X_2)^2 = \operatorname{tr}(\Sigma_1 + \Sigma_2) - 2\int_0^T \|\,\mathrm{d}H_{\mu_1} K_1^* K_2 \,\mathrm{d}H_{\mu_2}\|_{\mathrm{HS}}, \tag{3.8}$$

where

$$\int_0^T \|\,\mathrm{d}H_{\mu_1} K_1^* K_2 \,\mathrm{d}H_{\mu_2}\|_{\mathrm{HS}} := \lim_{\|\mathbf{P}\|\to 0} \sum_{(s,t)\in\mathbf{P}} \|(P_{\mu_1,t} - P_{\mu_1,s}) K_1^* K_2 (P_{\mu_2,t} - P_{\mu_2,s})\|_{\mathrm{HS}},$$

and $P_{\mu_i,t}$ denotes the projection of H_{μ_i} to the subspace $H_{\mu_i,t} = \{f \in H_{\mu_i} : \operatorname{supp}(f) \subseteq [0,t]\}$. Here, the limit is taken over all partitions $\mathbf{P}$ of $[0,T]$ with mesh size $\|\mathbf{P}\|$ converging to 0.

Remark 3.19. The distance does not depend on the choice of the canonical representation. Indeed, if k_1 and $\tilde{k}_1$ are kernels of two canonical representations of X_1, by Proposition 3.14 we have $\tilde{k}_1(\cdot, s) = k_1(\cdot, s)(\mathbb{1}_S(s) - \mathbb{1}_{[0,T]\setminus S}(s))$. Hence, plugging $\tilde{k}_1$ into (3.7) does not change its value.

Remark 3.20. One shall not expect to relax the condition of the canonical representation. We consider the non-canonical representation of Brownian motion given in Example 3.13. Naively plugging in the formula, we would obtain a positive quantity for the adapted Wasserstein distance between two standard Brownian motions.

Remark 3.21. Although we focus on mean-square continuous Gaussian processes, the proof can be easily adapted to the discrete-time case. Moreover, (3.8) is consistent with the discrete-time result given in Gunasingam and Wong (2025). In discrete-time case, the triangular integral $\int_0^T \| \mathrm{d}H_{\mu_1} K_1^* K_2 \, \mathrm{d}H_{\mu_2} \|_{\mathrm{HS}}$ can be interpreted as the sum of the diagonal elements of $K_1^* K_2$. Here, the notation of triangular integral is adapted from the literature of nest algebra, e.g., Davidson (1988).

Proof of Theorem 3.18. Since $\mathbf{F}^{X_i} = \mathbf{F}^{M_i}$, by definition we obtain $\Pi_{\mathfrak{bc}}(X_1, X_2) = \Pi_{\mathfrak{bc}}(M_1, M_2)$. We apply the transfer principle and derive that

$$\begin{aligned}\sup_{\pi \in \Pi_{\mathfrak{bc}}(X_1,X_2)} E_\pi[\langle X_1, X_2 \rangle] &= \sup_{\pi \in \Pi_{\mathfrak{bc}}(M_1,M_2)} E_\pi[\langle X_1, X_2 \rangle] \\ &= \sup_{\pi \in \Pi_{\mathfrak{bc}}(M_1,M_2)} E_\pi\left[\int_0^T \int_0^T k_1(t,s) k_2(t,s) [M_1, M_2](\mathrm{d}s) \, \mathrm{d}t\right].\end{aligned}$$

The second equality follows from the fact that M_1 and M_2 remain martingales with respect to the product filtration under any bi-causal coupling, see Acciaio et al. (2020, Remark 2.3). By Fubini theorem and Kunita–Watanabe inequality, we derive

$$\begin{aligned}\sup_{\pi \in \Pi_{\mathfrak{bc}}(X_1,X_2)} E_\pi[\langle X_1, X_2 \rangle] &= \sup_{\pi \in \Pi_{\mathfrak{bc}}(M_1,M_2)} E_\pi\left[\int_0^T \int_0^T k_1(t,s) k_2(t,s) \, \mathrm{d}t [M_1, M_2](\mathrm{d}s)\right] \\ &= \sup_{\rho(\cdot) \in [-1,1]} E_\pi\left[\int_0^T \left(\int_0^T k_1(t,s) k_2(t,s) \, \mathrm{d}t\right) \rho(s) \sqrt{\mu_1 \mu_2}(\mathrm{d}s)\right] \\ &= \int_0^T |\langle k_1(\cdot, s), k_2(\cdot, s) \rangle| \sqrt{\mu_1 \mu_2}(\mathrm{d}s).\end{aligned}$$

The second equality follows from the fact that $\sqrt{\mu_1 \mu_2} \gg [M_1, M_2]$ and the Radon–Nikodym density ρ takes values in $[-1, 1]$. Moreover, the optimal bi-causal coupling is induced by a Gaussian coupling

$$M_1(t) = \int_0^t \sqrt{\frac{\mathrm{d}\mu_1}{\mathrm{d}(\mu_1 + \mu_2)}}(s) \, \mathrm{d}\widetilde{M}(s) \quad \text{and } M_2(t) = \int_0^t \sqrt{\frac{\mathrm{d}\mu_2}{\mathrm{d}(\mu_1 + \mu_2)}}(s) \rho(s) \, \mathrm{d}\widetilde{M}(s),$$

where ρ attains the supremum in the above estimate and $\widetilde{M}$ is a Gaussian martingale with independent increments, $[\widetilde{M}](\mathrm{d}s) = (\mu_1 + \mu_2)(\mathrm{d}s)$.

For (3.8), we fix $s, t \in [0, T]$. We notice for any $f \in H_{\mu_2}$

$$(P_{\mu_1,t} - P_{\mu_1,s}) K_1^* K_2 (P_{\mu_2,t} - P_{\mu_2,s}) f(r_1) = \int_s^t \langle k_1(\cdot, r_1), k_2(\cdot, r_2)\rangle f(r_2) \mu_2(\mathrm{d}r_2).$$

This gives

$$\|(P_{\mu_1,t} - P_{\mu_1,s}) K_1^* K_2 (P_{\mu_2,t} - P_{\mu_2,s})\|_{\mathrm{HS}} = \left[\int_s^t \int_s^t |\langle k_1(\cdot, r_1), k_2(\cdot, r_2)\rangle|^2 \mu_1(\mathrm{d}r_1)\mu_2(\mathrm{d}r_2)\right]^{1/2}.$$

Since X_i is mean-square continuous, we have $(r_1, r_2) \mapsto \langle k_1(\cdot, r_1), k_2(\cdot, r_2)\rangle$ is uniformly continuous on $[0, T] \times [0, T]$. This allows us to conclude

$$\int_0^T \| \,\mathrm{d}H_{\mu_1} K_1^* K_2 \,\mathrm{d}H_{\mu_2}\|_{\mathrm{HS}} = \int_0^T |\langle k_1(\cdot, s), k_2(\cdot, s)\rangle| \sqrt{\mu_1 \mu_2}(\mathrm{d}s).$$

□

We give several examples.

Example 3.22. We consider the adapted Wasserstein distance between a standard Brownian motion B and a Cantor Gaussian martingale C. The covariance operator of the Cantor Gaussian martingale C is given by $E[C(t)C(s)] = F(t \wedge s)$, where F is the Cantor function, also known as the Devil's staircase. In particular, $F(\mathrm{d}t)$ is mutually singular to the Lebesgue measure. This implies that under any bi-causal coupling $B(t)$ and $C(t)$ are uncorrelated which gives

$$\mathcal{AW}_2(B, C)^2 = \int_0^T (t + F(t)) \,\mathrm{d}t.$$

In fact, every bi-causal coupling attains the adapted Wasserstein distance. On the other hand, one can easily construct a non-bi-causal coupling by the time change of Brownian motion under which B and C are not independent anymore and have a transport cost strictly less than $\int_0^T (t + F(t)) \,\mathrm{d}t$.

Example 3.23. We consider the adapted Wasserstein distance between two fractional Brownian motions. For a fractional Brownian motion B_H with Hurst parameter H, it has a stochastic representation given by

$$B_H(t) = \int_0^t k_H(t, s) \,\mathrm{d}B(s),$$

where k_H is the Molchan–Golosov kernel, see Molchan and Golosov (1969), Decreusefond and Üstünel (1999) for example. In particular, this gives a canonical representation of B_H, see Jost (2006, Theorem 5.1). Therefore, plugging this canonical representation into Theorem 3.18, we obtain Theorem 3.1 and have

$$\mathcal{AW}_2(B_{H_1}, B_{H_2})^2 = \int_0^T \int_0^T (k_{H_1}(t, s) - k_{H_2}(t, s))^2 \,\mathrm{d}t \,\mathrm{d}s = \|K_{H_1} - K_{H_2}\|_{\mathrm{HS}}^2.$$

We remark that the synchronous coupling is the unique optimal bi-causal coupling.

Example 3.24. We consider the adapted Wasserstein distance between fractional Ornstein–Uhlenbeck processes given by

$$X_i(0) = x_i - \lambda_i \int_0^t X_i(s)\,\mathrm{d}s + B_{H_i}(t),$$

whose solution is given by

$$X_i(t) = \mathrm{e}^{-\lambda_i t} x_i + \int_0^t \mathrm{e}^{\lambda_i (s-t)} \mathrm{d}B_{H_i}(s).$$

Let $\widetilde{X}_i(t) = X_i(t) - \mathrm{e}^{-\lambda_i t} x_i$. Then, $\widetilde{X}_i$ is a centered Gaussian process, and

$$\mathcal{AW}_2(X_1, X_2)^2 = \mathcal{AW}_2(\widetilde{X}_1, \widetilde{X}_2)^2 + \int_0^T |\, \mathrm{e}^{-\lambda_1 t} x_1 - \mathrm{e}^{-\lambda_2 t} x_2|^2\,\mathrm{d}t.$$

By Cheridito et al. (2003, Proposition A.1), we can show $\widetilde{X}_i$ is of unit multiplicity and with a canonical representation given by

$$\begin{aligned}\widetilde{X}_i(t) = \int_0^t \mathrm{e}^{\lambda_i (s-t)} \mathrm{d}B_{H_i}(s) &= \int_0^t \left(k_{H_i}(t,s) + \int_s^t \mathrm{e}^{\lambda_i (t-r)} k_{H_i}(r,s)\,\mathrm{d}r \right) \mathrm{d}B(s) \\ &:= \int_0^t k_{OU_i}(t,s)\,\mathrm{d}B(s).\end{aligned}$$

By Theorem 3.18, we derive $\mathcal{AW}_2(\widetilde{X}_1, \widetilde{X}_2)^2 = \int_0^T \int_0^T (k_{OU_1}(t,s) - k_{OU_2}(t,s))^2\,\mathrm{d}t\,\mathrm{d}s$ as $k_{OU_i} \geq 0$.

3.4.2 Higher multiplicity

We can also extend the result to the case of higher multiplicity.

Theorem 3.25. *Let X_1 and X_2 be two centered, mean-square continuous, and purely-nondeterministic Gaussian processes with canonical representations*

$$X_1(t) = \sum_{i=1}^m \int_0^t k_1^i(t,s)\,\mathrm{d}M_1^i(s) \quad \textit{and} \quad X_2(t) = \sum_{j=1}^n \int_0^t k_2^j(t,s)\,\mathrm{d}M_2^j(s).$$

Then, the adapted Wasserstein distance between X_1 and X_2 is given by

$$\begin{aligned}\mathcal{AW}_2(X_1, X_2)^2 = \sum_{i=1}^m \int_0^T \|k_1^i(\cdot,s)\|^2 \mu_1^i(\mathrm{d}s) + \sum_{j=1}^n \int_0^T \|k_2^j(\cdot,s)\|^2 \mu_2^j(\mathrm{d}s) \\ - 2\int_0^T \left\| \langle \tilde{k}_1^i(\cdot,s), \tilde{k}_2^j(\cdot,s)\rangle_{i,j} \right\|_{\mathrm{tr}} \sqrt{\mu_1^1 \mu_2^1}(\mathrm{d}s),\end{aligned}$$

where $\tilde{k}_1^i(\cdot,s) = \sqrt{\frac{\mathrm{d}\mu_1^i}{\mathrm{d}\mu_1^1}}(s) k_1^i(\cdot,s)$ *and* $\tilde{k}_2^j(\cdot,s) = \sqrt{\frac{\mathrm{d}\mu_2^j}{\mathrm{d}\mu_2^1}}(s) k_2^j(\cdot,s)$.

Remark 3.26. We point out that even though the Gaussian process X_1 is one-dimensional, its natural filtration is 'multi-dimensional'. Indeed, we can use X_1 to reconstruct a multi-dimensional Gaussian martingale $M_1 = (M_1^1, \dots, M_1^m)$ with independent components, sharing the same natural filtration as X_1. Hence, the adapted Wasserstein distance between higher multiplicity Gaussian processes is similar to the discrete-time multi-dimensional case (Acciaio et al., 2024) where a trace norm is present. In the same fashion, one can derive the adapted Wasserstein distance between multi-dimensional Gaussian processes with arbitrary multiplicity. For brevity, we only present the one-dimensional case.

Remark 3.27. Gaussian processes with higher multiplicity do exist in theory, although they are mostly pathological and not common in practice. For example, the independent sum of a standard Brownian motion and a fractional Brownian motion with $H > 3/4$ is equivalent to a standard Brownian motion (Cheridito, 2001), and hence the mixture is still a Gaussian process of unit multiplicity Hida and Hitsuda (1993, Theorem 6.3). In Hida and Hitsuda (1993, Chapter 4), a Gaussian process with multiplicity 2 is constructed explicitly by taking $X(t) = B_1(t) + F(t)B_2(t)$, where B_1, B_2 are independent standard Brownian motions, and F' is integrable but F is nowhere locally square integrable.

The following is an elementary algebraic lemma which we require for the proof of Theorem 3.25.

Lemma 3.28. *Let $A \in \mathbb{R}^{m\times m}$, $B \in \mathbb{R}^{n\times n}$, and $C \in \mathbb{R}^{m\times n}$. Assume A and B are semi-positive definite. Then, for any $\Gamma \in \mathbb{R}^{m\times n}$ such that $\left(\begin{smallmatrix} A & \Gamma \\ \Gamma^* & B \end{smallmatrix}\right) \geq 0$ we have*

$$\operatorname{tr}(C\Gamma^*) \leq \|A^{1/2}CB^{1/2}\|_{\operatorname{tr}}.$$

Moreover, the equality can be attained by $\Gamma = A^{1/2}UVB^{1/2}$ where U and V are given by the singular value decomposition $A^{1/2}CB^{1/2} = U\Sigma V$.

Proof. We first show the results for nondegenerate A and B. We notice $\left(\begin{smallmatrix} A & \Gamma \\ \Gamma^* & B \end{smallmatrix}\right) \geq 0$ is equivalent to $I \geq (A^{-1/2}\Gamma B^{-1/2})^*(A^{-1/2}\Gamma B^{-1/2})$. Moreover, the singular value decomposition gives

$$\operatorname{tr}(C\Gamma^*) = \operatorname{tr}(A^{1/2}CB^{1/2}(A^{-1/2}\Gamma B^{-1/2})^*) \leq \operatorname{tr}(\Sigma) = \|A^{1/2}CB^{1/2}\|_{\operatorname{tr}}.$$

Now we consider the general case. Since $\left(\begin{smallmatrix} A & \Gamma \\ \Gamma^* & B \end{smallmatrix}\right) \geq 0$ is equivalent to $\left(\begin{smallmatrix} A_\varepsilon & \Gamma \\ \Gamma^* & B_\varepsilon \end{smallmatrix}\right) \geq 0$ for any $\varepsilon > 0$ where $A_\varepsilon = A + \varepsilon I$ and $B_\varepsilon = B + \varepsilon I$. We derive $\operatorname{tr}(C\Gamma^*) \leq \|A_\varepsilon^{1/2}CB_\varepsilon^{1/2}\|_{\operatorname{tr}}$ for any $\varepsilon > 0$. Therefore, we conclude the proof by taking the limit $\varepsilon \to 0$ and noticing the equality can be attained by $\Gamma = A^{1/2}UVB^{1/2}$. $\square$

Proof of Theorem 3.25. We put emphasis on the difference between the unit multiplicity case and the higher multiplicity case and only sketch the similar part. We write $M_1 = (M_1^1, \dots, M_1^m)$, $M_2 = (M_2^1, \dots, M_2^n)$. Similar to the unit multiplicity case we notice

$$\begin{aligned}
\sup_{\pi\in\Pi_{\mathfrak{bc}}(X_1,X_2)} E_\pi[\langle X_1, X_2\rangle] &= \sup_{\pi\in\Pi_{\mathfrak{bc}}(M_1,M_2)} E_\pi[\langle X_1, X_2\rangle] \\
&= \sup_{\pi\in\Pi_{\mathfrak{bc}}(M_1,M_2)} \sum_{i,j} E_\pi\left[\int_0^T \int_0^T k_1^i(t,s)k_2^j(t,s)[M_1^i, M_2^j](\mathrm{d}s)\,\mathrm{d}t\right] \\
&= \sup_{\pi\in\Pi_{\mathfrak{bc}}(M_1,M_2)} \sum_{i,j} E_\pi\left[\int_0^T \langle k_1^i(\cdot,s), k_2^j(\cdot,s)\rangle \Gamma^{i,j}(s)\sqrt{\mu_1^1\mu_2^1}(\mathrm{d}s)\right],
\end{aligned}$$

where $\Gamma^{i,j}$ is the density of $[M_1^i, M_2^j]$ with respect to $\sqrt{\mu_1^1\mu_2^1}$. By Kunita–Watanabe inequality, we derive

$$\begin{pmatrix} \operatorname{diag}\left(\frac{\mathrm{d}\mu_1^1}{\mathrm{d}\mu_1^1}, \dots, \frac{\mathrm{d}\mu_1^m}{\mathrm{d}\mu_1^1}\right) & \Gamma \\ \Gamma^* & \operatorname{diag}\left(\frac{\mathrm{d}\mu_2^1}{\mathrm{d}\mu_2^1}, \dots, \frac{\mathrm{d}\mu_2^n}{\mathrm{d}\mu_2^1}\right) \end{pmatrix}(s) \geq 0.$$

By Lemma 3.28, we conclude the proof. In particular, the supremum is induced by the Gaussian coupling

$$\begin{cases} M_1(t) = \displaystyle\int_0^t \sqrt{\operatorname{diag}\left(\frac{\mathrm{d}\mu_1^1}{\mathrm{d}\nu}, \dots, \frac{\mathrm{d}\mu_1^m}{\mathrm{d}\nu}\right)}(s)\,\mathrm{d}\widetilde{M}(s), \\ M_2(t) = \displaystyle\int_0^t \sqrt{\operatorname{diag}\left(\frac{\mathrm{d}\mu_2^1}{\mathrm{d}\nu}, \dots, \frac{\mathrm{d}\mu_2^n}{\mathrm{d}\nu}\right)}(s)\Gamma^*(s)\,\mathrm{d}\widetilde{M}(s), \end{cases}$$

where $\nu = \mu_1^1 + \mu_2^1$, Γ a deterministic process attains the supremum in the above estimate, $\widetilde{M}$ is a Gaussian martingale with independent increments and $[\widetilde{M}](\mathrm{d}s) = \mathrm{Id}\,\nu(\mathrm{d}s)$.

□

3.4.3 A martingale approximation to the fractional BMs

It is well-known that, except in the case $H = 1/2$, the fractional Brownian motion is neither a martingale nor a Markov process. Hence, models based on fractional Brownian motions in practice are often less tractable and lead to difficulty in numerical simulation. To this end, we use the transfer principle to derive the best martingale approximation of a fractional Brownian motion in terms of their adapted Wasserstein distance, i.e.,

$$\inf_M \mathcal{AW}_2(B_H, M)^2, \text{ where } M \text{ is a } \mathbf{F}^{B_H}\text{-martingale.} \tag{3.9}$$

Theorem 3.29. *Let k_H be the Molchan–Golosov kernel of the fractional Brownian motion B_H. Then, the solution to (3.9) is given by*

$$M_H(t) = \int_0^t \frac{1}{T-r} \int_r^T k_H(s,r)\,\mathrm{d}s\,\mathrm{d}B(r).$$

Proof. Since $B_H(t) = \int_0^t k_H(t,s)\,\mathrm{d}B(s)$ is a canonical representation, we have $\mathbf{F}^{B_H} = \mathbf{F}^B$. Without loss of generality, we may restrict (3.9) to the set of centered and square integrable martingales. Under any bi-causal coupling π, M is still a $\mathbf{F}^B$-martingale. By martingale representation theorem, we deduce

$$M(t) = \int_0^t \rho(r)\,\mathrm{d}B(r), \text{ where } \rho \text{ is a } \mathbf{F}^B\text{-predictable process.}$$

Therefore, we have

$$\begin{aligned}
\inf_M \mathcal{AW}_2(B_H, M)^2 &= \inf_\rho E\left[\int_0^T \left|B_H(s) - \int_0^s \rho(r)\,\mathrm{d}B(r)\right|^2 \mathrm{d}s\right] \\
&= \inf_\rho E\left[\int_0^T \left|\int_0^s (k_H(s,r) - \rho(r))\,\mathrm{d}B(r)\right|^2 \mathrm{d}s\right] \\
&= \inf_\rho E\left[\int_0^T \int_r^T (k_H(s,r) - \rho(r))^2\,\mathrm{d}s\,\mathrm{d}r\right].
\end{aligned}$$

It is clear that the optimal ρ is given by $\rho_H(r) = \frac{1}{T-r}\int_r^T k_H(s,r)\,\mathrm{d}s$. □

We can interpret M_H as the martingale whose volatility is given by the average volatility of the prediction process of B_H. To be more precise, we introduce the prediction process Θ_H of B_H as the double-indexed process given by

$$\Theta_H(s;t) := E[B_H(t)|\mathcal{F}_s^{B_H}] = \int_0^s k_H(t,r)\,\mathrm{d}B(r) \text{ for } 0 \le s \le t.$$

In particular, for any fixed $t \in [0,T]$, $\Theta_H(\cdot\,;t)$ is a martingale with volatility given by $k_H(t,\cdot)$. Therefore, the volatility of the martingale M_H at the current time r, $\rho_H(r)$, is given by the current volatility of the prediction process $\Theta_H(\cdot\,;t)$ averaged over the future period $[r,T]$.

3.5 Fractional SDEs

In this section, we investigate the adapted Wasserstein distance between 1D fractional SDEs. Let X_i be the solution to

$$X_i(t) = x_i + \int_0^t b_i(X_i(s))\,\mathrm{d}s + \int_0^t \sigma_i(X_i(s))\,\mathrm{d}Z_i(s), \tag{3.10}$$

where $Z_i(t) = \int_0^t k_i(t,s)\,\mathrm{d}B_i(s)$ and B_i is a standard Brownian motion.

Assumption 3.30. We assume

- $b_i, \sigma_i \in C^2$ with bounded first and second derivatives.
- b_i'' and σ_i'' are uniformly continuous with a modulus of continuity ϱ_i.
- σ_i is positive, bounded, and bounded away from 0.

Assumption 3.31. We assume $Z_i(t) = \int_0^t k_i(t,s)\,\mathrm{d}B_i(s)$ is a canonical representation. Moreover, k_i satisfies

- $k_i(t,s) \geq 0$ for any $t, s \in [0,T]$.
- $k_i(\cdot, s) \in C^1([0,T];\mathbb{R})$ for any $s \in (0,T]$.
- $|k_i(t,s)| \leq Cs^{1/2-H}|t-s|^{H-1/2}$ and $|\partial_t k_i(t,s)| \leq Cs^{1/2-H}|t-s|^{H-3/2}$ for some $H \in (1/2, 1)$.

Assumption 3.32. We assume either of the following conditions holds:

(i) (b_i/σ_i) is non-decreasing.

(ii) $k_1(\cdot, s)$ and $k_2(\cdot, s)$ are both non-decreasing for any $s \in (0,T]$.

The following well-posedness result is standard, and for example, can be found in Friz and Hairer (2020, Section 8.3), Viens and Zhang (2019, Theorem A.1).

Lemma 3.33. *Under Assumptions 3.30 and 3.31, fractional SDE (3.10) is well-posed with a unique α-Hölder continuous strong solution for any $\alpha < H$. The stochastic integral $\int_0^t \sigma_i(X_i(s))\,\mathrm{d}Z_i(s)$ can be interpreted as a Young integral. Moreover, $E[\sup_{t\in\in[0,T]} |X_i(t)|^p] < \infty$ for any $p \geq 1$.*

Theorem 3.34. *Under Assumptions 3.30, 3.31 and 3.32, the adapted Wasserstein distance between X_1 and X_2 is attained by the synchronous coupling between B_1 and B_2, i.e., the noises Z_1 and Z_2 are driven by the same Brownian motion. In particular the synchronous coupling is a bi-causal coupling between X_1 and X_2.*

Remark 3.35. Assumptions 3.31 and 3.32 includes the Riemann–Liouville fractional kernel $RL_H(t,s) = \Gamma(H+1/2)^{-1}(t-s)^{H-1/2}\mathbb{1}_{\{t\geq s\}}$, as well as the Molchan–Golosov kernel $k_H(t,s)$ for $H \in (1/2,1)$.

We split the proof of Theorem 3.34 into two steps. The first step is to show, by a stochastic control reformulation, the results hold for the additive noise, i.e., $\sigma_i \equiv 1$. In the second step, we apply Lamperti transform to reduce the general case to the additive noise case.

3.5.1 Additive noise

By strong well-posedness Lemma 3.33, we reduce the problem to a minimization over the bi-causal coupling between the driving Brownian motions.

Lemma 3.36. *Let $\sigma_i \equiv 1$ for $i = 1, 2$. Under Assumptions 3.30 and 3.31, we have* $\Pi_{\mathfrak{bc}}(X_1, X_2) = \Pi_{\mathfrak{bc}}(B_1, B_2)$.

Proof. It suffices to show $\mathbf{F}^{X_i} = \mathbf{F}^{B_i}$. From the strong well-posedness, we have $\mathcal{F}_t^{X_i} \subseteq \mathcal{F}_t^{B_i}$ for any $t \in [0, T]$. Moreover, we notice

$$Z_i(t) = \int_0^t k_i(t, s)\, \mathrm{d}B_i(s) = X_i(t) - x_i - \int_0^t b_i(s, X_i(s))\, \mathrm{d}s \in \mathcal{F}_t^{X_i}.$$

This implies $\mathcal{F}_t^{B_i} = \mathcal{F}_t^{Z_i} \subseteq \mathcal{F}_t^{X_i}$ from the canonical representation of Z_i. Therefore, $\mathcal{F}_t^{X_i} = \mathcal{F}_t^{B_i}$ and we conclude the proof. $\square$

Now, similar to Bion–Nadal and Talay (2019), we address the bi-causal optimal transport problem as a stochastic control problem with the control of the correlation of the driving Brownian motions. We consider a controlled system

$$\begin{cases} X_1(t) = x_1 + \displaystyle\int_0^t b_1(X_1(s))\, \mathrm{d}s + \int_0^t k_1(t, s)\, \mathrm{d}B_1(s), \\ X_2^u(t) = x_2 + \displaystyle\int_0^t b_2(X_2^u(s))\, \mathrm{d}s + \int_0^t k_2(t, s)\, \mathrm{d}B_2^u(s), \end{cases}$$

where $\mathrm{d}B_2^u(t) = \sin(u(t))\, \mathrm{d}B_1(t) + \cos(u(t))\, \mathrm{d}\widetilde{B}_1(t)$ and $\widetilde{B}_1$ is a Brownian motion independent to B_1. We notice the control only enters the system through the correlation of the driving Brownian motions. Our aim is to minimize

$$\inf_{u \in \mathbb{U}([0,T])} E\left[\int_0^T |X_1(t) - X_2^u(t)|^2\, \mathrm{d}t\right],$$

over $\mathbb{U}([0, T])$ the set of $(\mathbf{F}^{B_1} \vee \mathbf{F}^{\widetilde{B}_1})$-progressively measurable processes. We immediately see that X_1 no longer enjoys the flow property in the sense that

$$X_1(t) \neq \widetilde{X}_1^{s,X_1}(t), \quad \text{where} \quad \widetilde{X}_1^{s,X_1}(t) := X_1(s) + \int_s^t b(X_1(r))\, \mathrm{d}r + \int_s^t k_1(t, s)\, \mathrm{d}B_1(r).$$

Therefore, the classical approach of dynamic programming does not apply directly. To go around this issue an auxiliary system Θ is introduced in Viens and Zhang (2019) to retrieve the flow property. We adapt their framework to our setting as

$$\begin{cases} \Theta_1(s; t) = x_1 + \displaystyle\int_0^s b_1(\Theta_1(r; r))\, \mathrm{d}r + \int_0^s k_1(t, r)\, \mathrm{d}B_1(r), \\ \Theta_2^u(s; t) = x_2 + \displaystyle\int_0^s b_2(\Theta_2^u(r; r))\, \mathrm{d}r + \int_0^s k_2(t, r)\, \mathrm{d}B_2^u(r). \end{cases}$$

In particular, $(X_1(t), X_2^u(t)) = (\Theta_1(t;t), \Theta_2^u(t;t))$ and

$$\mathcal{AW}_2(X_1, X_2)^2 = \inf_{\pi \in \Pi_{\mathfrak{bc}}(B_1,B_2)} E_\pi[\|X_1 - X_2\|^2] = \inf_{u \in \mathbb{U}([0,T])} E\left[\int_0^T |\Theta_1(t;t) - \Theta_2^u(t;t)|^2 \,\mathrm{d}t\right].$$

We can view $\{(\Theta_1(t;\cdot), \Theta_2^u(t;\cdot)) : t \in [0,T]\}$ as an infinite dimensional flow taking values in $C([0,T];\mathbb{R}^2)$. Naturally, we define the value function $v : [0,T] \times C([0,T];\mathbb{R}^2) \to \mathbb{R}$ as

$$v(r, \omega_1, \omega_2) := \inf_{u \in \mathbb{U}([r,T])} E\left[\int_r^T |\Theta_1^{r,\omega_1}(t;t) - \Theta_2^{r,\omega_2,u}(t;t)|^2 \,\mathrm{d}t\right],$$

where

$$\begin{cases} \Theta_1^{r,\omega_1}(\cdot\,;t) = \omega_1(t) + \int_r^\cdot b_1(\Theta_1^{r,\omega_1}(s;s)) \,\mathrm{d}s + \int_r^\cdot k_1(t,s) \,\mathrm{d}B_1(s), \\ \Theta_2^{r,\omega_2,u}(\cdot\,;t) = \omega_2(t) + \int_r^\cdot b_2(\Theta_2^{r,\omega_2,u}(s;s)) \,\mathrm{d}s + \int_r^\cdot k_2(t,s) \,\mathrm{d}B_2^u(s). \end{cases}$$

We denote the time derivative by ∂_t and the first and second Fréchet derivatives by ∂_{ω_i} and $\partial^2_{\omega_i\omega_j}$, respectively. The corresponding HJB equation is given by

$$(\partial_t + \mathcal{L}_1 + \mathcal{L}_2 + \mathcal{H})V(r, \omega_1, \omega_2) = -|\omega_1(r) - \omega_2(r)|^2 \quad \text{with} \quad V(T, \omega_1, \omega_2) = 0. \tag{3.11}$$

Here, $\mathcal{L}_i$ and $\mathcal{H}$ are given by

$$\begin{aligned} \mathcal{L}_i V(r, \omega_1, \omega_2) &= \langle b_i(\omega_i(r))\mathbb{1}_{[0,T]}(\cdot), \partial_{\omega_i} V(r, \omega_1, \omega_2)(\cdot)\rangle \\ &\quad + \frac{1}{2}\langle (k_1(\cdot, r), k_1(\cdot, r)), \partial^2_{\omega_i\omega_i} V(r, \omega_1, \omega_2)(\cdot)\rangle \end{aligned}$$

and

$$\mathcal{H}V(r, \omega_1, \omega_2) = \inf_{a \in [-1,1]} a\langle (k_1(\cdot, r), k_2(\cdot, r)), \partial^2_{\omega_1\omega_2} V(r, \omega_1, \omega_2)(\cdot)\rangle.$$

We denote the expected cost under the synchronous coupling by V_*, which is given by

$$V_*(r, \omega_1, \omega_2) := E\left[\int_r^T \left|\Theta_{1,*}^{r,\omega_1}(t;t) - \Theta_{2,*}^{r,\omega_2}(t;t)\right|^2 \mathrm{d}t\right],$$

where

$$\begin{cases} \Theta_{1,*}^{r,\omega_1}(\cdot\,;t) = \omega_1(t) + \int_r^\cdot b_1(\Theta_{1,*}^{r,\omega_1}(s;s)) \,\mathrm{d}s + \int_r^\cdot k_1(t,s) \,\mathrm{d}B_1(s), \\ \Theta_{2,*}^{r,\omega_2}(\cdot\,;t) = \omega_2(t) + \int_r^\cdot b_2(\Theta_{2,*}^{r,\omega_2}(s;s)) \,\mathrm{d}s + \int_r^\cdot k_2(t,s) \,\mathrm{d}B_1(s). \end{cases}$$

Our plan is to verify V_* coincides with the value function v. To achieve this, we adapt a functional Itô formula from Viens and Zhang (2019, Theorem 3.10) to our setting. We note that the kernel k_i has the same singularity as the fractional Brownian motion kernel $k_H(t,s)$ for $H \in (1/2, 1)$, where the singularity can only occur as s approaches 0. Hence, for any $r > 0$ Viens and Zhang (2019, Theorem 3.10) is directly applicable. The derivatives involed here are the time and Fréchet derivatives, in contrast to the horizontal and vertical derivatives studied in Dupire (2009), Cont and Fournié (2010). We also remark that a *pathwise* Itô formula for non-anticipative functionals is derived in Cont and Fournié (2010).

Lemma 3.37 (Functional Itô formula)**.** *Let $u : [0,T] \times C([0,T];\mathbb{R}^2) \to \mathbb{R}$ be a purely anticipative functional, i.e., $u(t,\omega_1,\omega_2) = u(t,\omega_1(\cdot \vee t),\omega_2(\cdot \vee t))$ for any $t \in [0,T]$ and $\omega_i \in C([0,T];\mathbb{R})$. Assume $u \in C^{1,2}$, and there exists a modulus of continuity ρ such that for any $\eta, \tilde{\eta} \in C([0,T];\mathbb{R})$, u satisfies the following conditions:*

(i) for any $\omega_1, \omega_2 \in C([0,T];\mathbb{R})$,

$$|\langle \eta, \partial_{\omega_i} u(r,\omega_1,\omega_2)\rangle| \le C(1 + \|\omega_1\|_\infty + \|\omega_2\|_\infty)\|\eta\|_\infty,$$
$$|\langle (\eta,\tilde{\eta}), \partial^2_{\omega_i\omega_j} u(r,\omega_1,\omega_2)\rangle| \le C(1 + \|\omega_1\|_\infty + \|\omega_2\|_\infty)\|\eta\|_\infty\|\tilde{\eta}\|_\infty;$$

(ii) for any other $\omega_1', \omega_2' \in C([0,T];\mathbb{R})$,

$$\begin{aligned}&|\langle (\eta,\tilde{\eta}), \partial^2_{\omega_i\omega_j} u(r,\omega_1,\omega_2) - \partial^2_{\omega_i\omega_j} u(r,\omega_1',\omega_2')\rangle| \\ &\qquad \le C(1 + \|\omega_1\|_\infty + \|\omega_2\|_\infty)\|\eta\|_\infty\|\tilde{\eta}\|_\infty \rho(\|\omega_1-\omega_1'\|_\infty + \|\omega_2-\omega_2'\|_\infty).\end{aligned}$$

Then under Assumptions 3.30 and 3.31, we have

$$\begin{aligned}&u(t,\Theta^{r,\omega_1}_{1,*}(t;\cdot),\Theta^{r,\omega_2}_{2,*}(t;\cdot)) \\ &= u(r,\omega_1,\omega_2) + \int_r^t (\partial_t + \mathcal{L}_1 + \mathcal{L}_2) u(s,\Theta^{r,\omega_1}_{1,*}(s;\cdot),\Theta^{r,\omega_2}_{2,*}(s;\cdot))\,\mathrm{d}s \\ &\qquad + \int_r^t \langle (k_1(\cdot,s),k_2(\cdot,s)), \partial^2_{\omega_1\omega_2} u(s,\Theta^{r,\omega_1}_{1,*}(s;\cdot),\Theta^{r,\omega_2}_{2,*}(s;\cdot))\rangle\,\mathrm{d}s \\ &\qquad + \int_r^t \langle k_1(\cdot,s), \partial_{\omega_1} u(s,\Theta^{r,\omega_1}_{1,*}(s;\cdot),\Theta^{r,\omega_2}_{2,*}(s;\cdot))\rangle\,\mathrm{d}B_1(s) \\ &\qquad + \int_r^t \langle k_2(\cdot,s), \partial_{\omega_2} u(s,\Theta^{r,\omega_1}_{1,*}(s;\cdot),\Theta^{r,\omega_2}_{2,*}(s;\cdot))\rangle\,\mathrm{d}B_1(s).\end{aligned}$$

The following technical lemma states that V_* is sufficiently regular to apply the functional Itô formula of Viens and Zhang (2019).

Lemma 3.38. *Under Assumptions 3.30, 3.31, and 3.32, V_* satisfies conditions in Lemma 3.37, and is a classical solution to*

$$(\partial_t+\mathcal{L}_1+\mathcal{L}_2)V_*(r,\omega_1,\omega_2)+\langle (k_1(\cdot,r),k_2(\cdot,r)), \partial^2_{\omega_1\omega_2} V_*(r,\omega_1,\omega_2)(\cdot)\rangle = -|\omega_1(r)-\omega_2(r)|^2. \tag{3.12}$$

Moreover, there is a probabilistic representation of $\partial^2_{\omega_1\omega_2} V_*$ given by

$$\langle (\eta_1, \eta_2), \partial^2_{\omega_1\omega_2} V_*(r, \omega_1, \omega_2)\rangle = -2E\bigg[\int_r^T \langle \eta_1, \Gamma^{r,\omega_1}_{1,*}(t)\rangle\langle \eta_2, \Gamma^{r,\omega_2}_{2,*}(t)\rangle \, \mathrm{d}t\bigg], \tag{3.13}$$

where $\Gamma^{r,\omega_i}_{i,*}$ is the unique solution to

$$\Gamma^{r,\omega_i}_{i,*}(t) = \delta(t) + \int_r^t b_i'(\Theta^{r,\omega_i}_{i,*}(s;s))\Gamma^{r,\omega_i}_{i,*}(s)\, \mathrm{d}s. \tag{3.14}$$

To not distract the readers, we postpone the proof of this technical result to Section 3.6 and continue with the main line of our results.

Theorem 3.39. *Under Assumptions 3.30, 3.31, and 3.32, V_* is a classical solution to the path-dependent HJB equation (3.11). Moreover, V_* coincides with the value function v, and in particular, the adapted Wasserstein distance is given by $\mathcal{AW}_2(X_1, X_2) = V_*(0, x_1\mathbb{1}_{[0,T]}, x_2\mathbb{1}_{[0,T]})^{1/2}$.*

Remark 3.40. We point out that a similar stochastic control approach was taken in Bion–Nadal and Talay (2019) where they rely on the regularity and well-posedness of nonlinear parabolic equations. However, to the best of knowledge, there is no well-posedness result for nonlinear *functional* parabolic equations on Banach space which can be directly applied to our setting. Our estimates are based on probabilistic methods. We manage to show the existence of the classical solution to the path-dependent HJB equation by a direct construction. It is interesting and challenging to build a viscosity solution theory of this type of path-dependent HJB equations. We leave this as a future research direction.

Remark 3.41. Following the same line of proof, we can show that for any non-decreasing f_i with bounded first, second, and third derivatives, synchronous coupling is still an optimal coupling for the bi-causal optimal transport problem

$$\inf_{\pi\in\Pi_{\mathfrak{bc}}(X_1,X_2)} E_\pi\bigg[\int_0^T |f_1(X_1(t)) - f_2(X_2(t))|^2 \, \mathrm{d}t\bigg].$$

Also, see Remark 3.48 for more details.

Proof. We prove V_* is a classical solution to the HJB equation (3.11). By Lemma 3.38, it suffices to verify that $\langle (k_1(\cdot, r), k_2(\cdot, r)), \partial^2_{\omega_1\omega_2} V_*(r, \omega_1, \omega_2)\rangle \le 0$. Recall we define $\Gamma^{r,\omega_i}_{i,*}$ in (3.14), and it admits a unique solution

$$\Gamma^{r,\omega_i}_{i,*}(t) = \delta(t) + \int_r^t \exp\bigg(\int_s^t b_i'(\Theta^{r,\omega_i}_{i,*}(\tau;\tau))\, \mathrm{d}\tau\bigg) b_i'(\Theta^{r,\omega_i}_{i,*}(s;s))\delta(s)\, \mathrm{d}s. \tag{3.15}$$

We discuss two cases in Assumption 3.32 separately. If b_i is non-decreasing, from (3.15), we derive $\langle \eta_i, \Gamma_i^{r,\omega_i}\rangle \geq 0$ for any $\eta_i \geq 0$. Plugging it into (3.13), we conclude V_* is a classical solution to HJB equation (3.11) as $k_i(\cdot, r) \geq 0$. If $k_1(\cdot, r)$ and $k_2(\cdot, r)$ are both non-decreasing, by applying integration by part to (3.15), we derive

$$\langle k_i(\cdot, r), \Gamma_{i,*}^{r,\omega_i}(t)\rangle = \int_r^t \exp\left(\int_s^t b_i'(\Theta_{i,*}^{r,\omega_i}(\tau;\tau))\,\mathrm{d}\tau\right) k_i(\mathrm{d}s, r)$$

have the same sign for $i = 1, 2$. Therefore, V_* is a classical solution to (3.11).

We show that V_* coincides with the value function v. We fix a control $u \in \mathbb{U}([r, T])$ and, by Lemma 3.38, we apply functional Itô formula to $V_*(t, \Theta_1^{r,\omega_1}(t;\cdot), \Theta_2^{r,\omega_2,u}(t;\cdot))$. We obtain

$$\begin{aligned}
&V_*(r, \omega_1, \omega_2)\\
&= -E\left[\int_r^T (\partial_t + \mathcal{L}_1 + \mathcal{L}_2) V_*(t, \Theta_1^{r,\omega_1}(t;\cdot), \Theta_2^{r,\omega_2,u}(t;\cdot))\,\mathrm{d}t\right]\\
&\quad - E\left[\int_r^T \sin(u(t))\langle (k_1(t,\cdot), k_2(t,\cdot)), \partial_{\omega_1\omega_2} V_*(t, \Theta_1^{r,\omega_1}(t;\cdot), \Theta_2^{r,\omega_2,u}(t;\cdot))\rangle\,\mathrm{d}t\right]\\
&\leq -E\left[\int_r^T (\partial_t + \mathcal{L}_1 + \mathcal{L}_2 + \mathcal{H}) V_*(t, \Theta_1^{r,\omega_1}(t;\cdot), \Theta_2^{r,\omega_2,u}(t;\cdot))\,\mathrm{d}t\right]\\
&= E\left[\int_r^T |\Theta_1^{r,\omega_1}(t;t) - \Theta_2^{r,\omega_2,u}(t;t)|^2\,\mathrm{d}t\right].
\end{aligned}$$

The above inequality follows from the fact that V_* satisfies HJB equation (3.11). Therefore, taking infimum over $\mathbb{U}([r, T])$ we deduce

$$V_*(r, \omega_1, \omega_2) \leq \inf_{u\in\mathbb{U}([r,T])} E\left[\int_r^T \left|\Theta_1^{s,\omega_1(t)}(t;t) - \Theta_2^{s,\omega_2(t),u}(t;t)\right|^2 \mathrm{d}t\right] = v(r, \omega_1, \omega_2).$$

On the other hand, we notice $u(r) \equiv \pi/2$ gives an optimal control, and hence $V_* = v$. $\square$

3.5.2 Multiplicative noise

Now we return to (3.10) with diffusion coefficient σ_i satisfying Assumption 3.30. We write

$$g_i(x) = \int_{x_i}^x \frac{1}{\sigma_i(\xi)}\,\mathrm{d}\xi \quad \text{and} \quad Y_i(t) = g_i(X_i(t)).$$

Notice that under Assumptions 3.30 and 3.31, X_i and Z_i are α–Hölder with $\alpha > 1/2$. This yields, Y_i, the Lamperti transform of X_i, satisfies

$$Y_i(t) = \int_0^t \frac{b_i(g_i^{-1}(Y_i(s)))}{\sigma_i(g_i^{-1}(Y_i(s)))}\,\mathrm{d}s + Z_i(t).$$

Lemma 3.42. *Under Assumptions 3.30 and 3.31, we have* $\mathbf{F}^{X_i} = \mathbf{F}^{B_i}$.

Proof. By Lemma 3.33, X_i is a strong solution and hence $\mathcal{F}_t^{X_i} \subseteq \mathcal{F}_t^{B_i}$ for any $t \in [0,T]$. On the other hand, we notice $Z_i(t) = Y_i(t) - \int_0^t \frac{b_i(g_i^{-1}(Y_i(s)))}{\sigma_i(g_i^{-1}(Y_i(s)))}\,\mathrm{d}s$, which implies $\mathcal{F}_t^{Z_i} \subseteq \mathcal{F}_t^{Y_i}$. Therefore, we deduce

$$\mathcal{F}_t^{B_i} = \mathcal{F}_t^{Z_i} \subseteq \mathcal{F}_t^{Y_i} \subseteq \mathcal{F}_t^{X_i} \subseteq \mathcal{F}_t^{B_i}.$$

□

The above lemma allows us to reduce the adapted Wasserstein distance between X_1 and X_2 to a bi-causal optimal transport problem between Y_1 and Y_2 as

$$\mathcal{AW}_2(X_1, X_2)^2 = \inf_{\pi \in \Pi_{\mathfrak{bc}}(Y_1,Y_2)} E\left[\int_0^T |g_1^{-1}(Y_1(t)) - g_2^{-1}(Y_2(t))|^2\,\mathrm{d}t\right].$$

We construct $(\tilde{b}_i, \tilde{\sigma}_i) = \left(\frac{b_i \circ g_i^{-1}}{\sigma_i \circ g_i^{-1}}, 1\right)$. A direct calculation gives $\tilde{b}_i' = b_i' \circ g_i^{-1} - \frac{(b_i \circ g_i^{-1})(\sigma_i' \circ g_i^{-1})}{\sigma_i \circ g_i^{-1}}$ and

$$\tilde{b}_i'' = (b_i'' \circ g_i^{-1})(\sigma \circ g_i^{-1}) - (b_i' \circ g_i^{-1})(\sigma_i' \circ g_i^{-1}) - (b_i \circ g_i^{-1})(\sigma_i'' \circ g_i^{-1}) + \frac{(b_i \circ g_i^{-1})(\sigma_i' \circ g_i^{-1})^2}{\sigma_i \circ g_i^{-1}}.$$

If b_i were bounded, we could verify $(\tilde{b}_i, \tilde{\sigma}_i)$ satisfies Assumptions 3.30, and $(\tilde{b}_i/\tilde{\sigma}_i)$ is non-decreasing if (b_i/σ_i) is. Applying Remark 3.41 we could conclude the proof of Theorem 3.34. For unbounded b_i, we take a sequence of functions $b_i^n \in C_b^2$ satisfying Assumption 3.30 and converging to b_i pointwise. In particular, we can assume $b_i^n = b_i$ on $[-n, n]$, and $|(b_i^n)'| \le |b_i'| \le L$. We define

$$X_i^n(t) = x_i + \int_0^t b_i^n(X_i^n(s))\,\mathrm{d}s + \int_0^t \sigma_i(X_i^n(s))\,\mathrm{d}Z_i(s).$$

By the triangle inequality, we obtain

$$\mathcal{AW}_2(X_1^n, X_2^n) \le \mathcal{AW}_2(X_1, X_2) + \mathcal{AW}_2(X_1^n, X_1) + \mathcal{AW}_2(X_2^n, X_2).$$

In order to show the synchronous coupling is optimal, we only need to show $\mathcal{AW}_2(X_i^n, X_i)$ goes to 0 since the synchronous coupling is already optimal between X_1^n and X_2^n by previous arguments.

Lemma 3.43. *Under Assumptions 3.30 and 3.31, we have* $\lim_{n\to\infty} \mathcal{AW}_2(X_i^n, X_i) = 0$.

Proof. By Lemma 3.42, we have $\mathbf{F}^{X_i^n} = \mathbf{F}^{Z_i^n}$, and hence the synchronous coupling π_{sync} between Z_i^n and Z_i is a bi-causal coupling between X_i^n and X_i. We write $Y_i^n = g_i(X_i^n)$ and $\tilde{b}_i^n = \frac{b_i^n \circ g_i^{-1}}{\sigma_i \circ g_i^{-1}}$. By our construction of b_i^n, we have $\tilde{b}_i^n = \tilde{b}_i$ on $[-n, n]$ and $|(b_i^n)'| \leq |b_i'| \leq L$. Without loss of generality, we may assume $|b_i^n(x)| + |b_i(x)| \leq L(1 + |x|)$ for possibly larger L.

Since σ_i is bounded and bounded away from 0, we derive that

$$\mathcal{AW}_2(X_i^n, X_i)^2 \leq E_{\pi_{\text{sync}}}[\|X_i^n - X_i\|^2 = E_{\pi_{\text{sync}}}[\|g_i^{-1}(Y_i^n) - g_i^{-1}(Y_i)\|^2] \leq C E_{\pi_{\text{sync}}}[\|Y_i^n - Y_i\|^2].$$

Therefore, it suffices to show Y_i^n converges to Y_i in H in L^2. Notice under π_{sync}, we have

$$\begin{aligned}
|Y_i^n(t) - Y_i(t)|^2 &\leq 2\left(\int_0^t |\tilde{b}_i^n(Y_i^n(s)) - \tilde{b}_i^n(Y_i(s))| \,\mathrm{d}s\right)^2 + 2\left(\int_0^t |\tilde{b}_i^n(Y_i(s)) - \tilde{b}_i(Y_i(s))| \,\mathrm{d}s\right)^2 \\
&\leq 2TL^2 \int_0^t |Y_i^n(s) - Y_i(s)|^2 \,\mathrm{d}s + 2T \int_0^t |\tilde{b}_i^n(Y_i(s)) - \tilde{b}_i(Y_i(s))|^2 \mathbb{1}_{\{|Y_i(s)| \geq n\}} \,\mathrm{d}s \\
&\leq 2TL^2 \int_0^t |Y_i^n(s) - Y_i(s)| \,\mathrm{d}s + 2TL^2 \int_0^t (1 + |Y_i(s)|)^2 \mathbb{1}_{\{|Y_i(s)| \geq n\}} \,\mathrm{d}s.
\end{aligned}$$

By Gronwall inequality, we obtain

$$E_{\pi_{\text{sync}}}[\|Y_i^n - Y_i\|^2] \leq C \int_0^T E_{\pi_{\text{sync}}}\left[|Y_i(s)|^2 \mathbb{1}_{\{|Y_i(s)| \geq n\}}\right] \mathrm{d}s.$$

By Lemma 3.33, Y_i is in L^2, and hence we derive the L^2 convergence of Y_i^n. $\square$

3.6 Some additional estimates

Recall

$$\Theta_{i,*}^{r,\omega_i}(\cdot\,; t) = \omega_i(t) + \int_r^{\cdot} b_i(\Theta_{i,*}^{r,\omega_i}(s; s)) \,\mathrm{d}s + \int_r^{\cdot} k_i(t, s) \,\mathrm{d}B_1(s), \tag{3.16}$$

and

$$\Gamma_{i,*}^{r,\omega_i}(t) = \delta(t) + \int_r^t b_i'(\Theta_{i,*}^{r,\omega_i}(s; s)) \Gamma_{i,*}^{r,\omega_i}(s) \,\mathrm{d}s. \tag{3.17}$$

Proposition 3.44. *Let $s \in [r, T]$ and $\eta \in C([0, T]; \mathbb{R})$. Under Assumptions 3.30 and 3.31, the following estimates hold with a deterministic constant C independent of ω_i and η*

$$\sup_{t \in [r,T]} E[|\Theta_{i,*}^{r,\omega_i}(t; t)|] \leq C(1 + \|\omega_i\|_\infty) \text{ and } \sup_{t \in [r,T]} |\Theta_{i,*}^{r,\omega_i+\eta}(t; t) - \Theta_{i,*}^{r,\omega_i}(t; t)| \leq C\|\eta\|_\infty.$$

Proof. It follows directly from the Gronwall inequality and the boundedness of b'. $\square$

Proposition 3.45. *Let $s \in [r,T]$ and $\tilde{\eta}, \eta \in C([0,T];\mathbb{R})$. Under Assumptions 3.30 and 3.31, the following estimates hold with a deterministic constant C independent of ω_i, η, and $\tilde{\eta}$*

$$\sup_{t\in[r,T]} |\langle \tilde{\eta}, \Gamma^{r,\omega_i}_{i,*}(t)\rangle| \le C\|\tilde{\eta}\|_\infty \text{ and } \sup_{t\in[r,T]} |\langle \tilde{\eta}, \Gamma^{r,\omega_i+\eta}_{i,*}(t) - \Gamma^{r,\omega_i}_{i,*}(t)\rangle| \le C\|\tilde{\eta}\|_\infty \|\eta\|_\infty.$$

Proof. It follows directly from the Gronwall inequality and the boundedness of b' and b''. □

The following result shows that $\Gamma^{r,\omega_i}_{i,*}$ is the first variation process of $\Theta^{r,\omega_i}_{i,*}$.

Proposition 3.46. *Let $\eta \in C([0,T];\mathbb{R})$. Under Assumptions 3.30 and 3.31, there exists a deterministic constant C independent of ω_i and η such that*

$$\sup_{t\in[r,T]} |\Theta^{r,\omega_i+\eta}_{i,*}(t;t) - \Theta^{r,\omega_i}_{i,*}(t;t) - \langle \eta, \Gamma^{r,\omega_i}_{i,*}(t)\rangle| \le C\|\eta\|^2_\infty.$$

Proof. Write $\Delta\Theta(t) = \Theta^{r,\omega_i+\eta}_{i,*}(t;t) - \Theta^{r,\omega_i}_{i,*}(t;t)$ and $R_1(t) = \Delta\Theta(t) - \langle \eta, \Gamma^{r,\omega_i}_{i,*}(t)\rangle$. Plugging (3.16) and (3.17), we notice that

$$\begin{aligned} R_1(t) &= \int_r^t b_i'(\Theta^{r,\omega_i}_{i,*}(s;s)) R_1(s)\,\mathrm{d}s \\ &\quad + \int_r^t \left(\int_0^1 \left[b_i'(\Theta^{r,\omega_i}_{i,*}(s;s) + \lambda\Delta\Theta(s)) - b_i'(\Theta^{r,\omega_i}_{i,*}(s;s)) \right] \mathrm{d}\lambda \right) \Delta\Theta(s)\,\mathrm{d}s. \end{aligned}$$

By Gronwall inequality and Proposition 3.44, we deduce

$$\sup_{t\in[r,T]} |R_1(t)| \le C\|\eta\|_\infty \int_r^T \int_0^1 \left| b_i'(\Theta^{r,\omega_i}_{i,*}(s;s) + \lambda\Delta\Theta(s)) - b_i'(\Theta^{r,\omega_i}_{i,*}(s;s)) \right| \mathrm{d}\lambda\,\mathrm{d}s.$$

Since b_i'' is bounded and $\sup_{t\in[r,T]} |\Delta\Theta(t)| \le C\|\eta\|_\infty$, we derive $\sup_{t\in[r,T]} |R_1(t)| \le C\|\eta\|^2_\infty$. □

We define

$$\Xi^{r,\omega_i}_{i,*}(t) = \int_r^t \exp\left(\int_s^t b_i'(\Theta^{r,\omega_i}_{i,*}(\tau;\tau))\,\mathrm{d}\tau \right) b_i''(\Theta^{r,\omega_i}_{i,*}(s;s)) \Gamma^{r,\omega_i}_{i,*}(s) \otimes \Gamma^{r,\omega_i}_{i,*}(s)\,\mathrm{d}s, \quad (3.18)$$

which is the unique solution to

$$\Xi^{r,\omega_i}_{i,*}(t) = \int_r^t b_i'(\Theta^{r,\omega_i}_{i,*}(s;s)) \Xi^{r,\omega_i}_{i,*}(s)\,\mathrm{d}s + \int_r^t b_i''(\Theta^{r,\omega_i}_{i,*}(s;s)) \Gamma^{r,\omega_i}_{i,*}(s) \otimes \Gamma^{r,\omega_i}_{i,*}(s)\,\mathrm{d}s. \quad (3.19)$$

Proposition 3.47. *Let $\eta, \tilde{\eta} \in C([0,T];\mathbb{R})$. Under Assumptions 3.30 and 3.31, there exists a deterministic constant C independent of ω_i, η, and $\tilde{\eta}$ such that*

$$\sup_{t\in[r,T]} |\langle \tilde{\eta}, \Gamma^{r,\omega_i+\eta}_{i,*}(t) - \Gamma^{r,\omega_i}_{i,*}(t)\rangle - \langle (\eta,\tilde{\eta}), \Xi^{r,\omega_i}_{i,*}(t)\rangle| \le C\|\tilde{\eta}\|_\infty \|\eta\|_\infty \varrho_i(\|\eta\|_\infty),$$

where ϱ_i is the modulus of continuity of b_i''.

Proof. Write $\Delta\Gamma(t) = \langle \tilde{\eta}, \Gamma^{r,\omega_i+\eta}_{i,*}(t) - \Gamma^{r,\omega_i}_{i,*}(t)\rangle$ and $R_2(t) = \Delta\Gamma(t) - \langle(\eta,\tilde{\eta}), \Xi^{r,\omega_i}_{i,*}(t)\rangle$. Plugging (3.17) and (3.19), we notice that

$$\begin{aligned}
R_2(t) = &\int_r^t b_i'(\Theta^{r,\omega_i}_{i,*}(s;s))R_2(s)\,\mathrm{d}s\\
&+ \int_r^t (b_i'(\Theta^{r,\omega_i+\eta}_{i,*}(s;s)) - b_i'(\Theta^{r,\omega_i}_{i,*}(s;s)))\Delta\Gamma(s)\,\mathrm{d}s\\
&+ \int_r^t \big[b_i'(\Theta^{r,\omega_i+\eta}_{i,*}(s;s)) - b_i'(\Theta^{r,\omega_i}_{i,*}(s;s)) - b_i''(\Theta^{r,\omega_i}_{i,*}(s;s))\langle\eta,\Gamma^{r,\omega_i}_{i,*}(s)\rangle\big]\langle\tilde{\eta},\Gamma^{r,\omega_i}_{i,*}(s)\rangle\,\mathrm{d}s.
\end{aligned}$$

By Gronwall inequality, we deduce

$$\begin{aligned}
&\sup_{t\in[r,T]}|R_2(t)|\\
\lesssim &\int_r^T |b_i'(\Theta^{r,\omega_i+\eta}_{i,*}(s;s)) - b_i'(\Theta^{r,\omega_i}_{i,*}(s;s))||\Delta\Gamma(s)|\,\mathrm{d}s\\
&+ \int_r^T |b_i'(\Theta^{r,\omega_i+\eta}_{i,*}(s;s)) - b_i'(\Theta^{r,\omega_i}_{i,*}(s;s)) - b_i''(\Theta^{r,\omega_i}_{i,*}(s;s))\langle\eta,\Gamma^{r,\omega_i}_{i,*}(s)\rangle||\langle\tilde{\eta},\Gamma^{r,\omega_i}_{i,*}(s)\rangle|\,\mathrm{d}s\\
:= &I_1 + I_2.
\end{aligned}$$

By Proposition 3.44 and Proposition 3.45, we notice $I_1 \lesssim \|\tilde{\eta}\|_\infty\|\eta\|^2_\infty$. For I_2, we plug in the estimates from Proposition 3.46 and obtain

$$\begin{aligned}
I_2 \lesssim &\|\tilde{\eta}\|_\infty \sup_{t\in[r,T]}|R_1(t)|\\
&+ \|\tilde{\eta}\|_\infty \int_r^t |b_i'(\Theta^{r,\omega_i+\eta}_{i,*}(s;s)) - b_i'(\Theta^{r,\omega_i}_{i,*}(s;s)) - b_i''(\Theta^{r,\omega_i}_{i,*}(s;s))\Delta\Theta(s)|\,\mathrm{d}s\\
\lesssim &\|\tilde{\eta}\|_\infty\|\eta\|^2_\infty + \|\tilde{\eta}\|_\infty \int_r^T\int_0^1 |b_i''(\Theta^{r,\omega_i}_{i,*}(s;s) + \lambda\Delta\Theta(s)) - b_i''(\Theta^{r,\omega_i}_{i,*})||\Delta\Theta(s)|\,\mathrm{d}\lambda\,\mathrm{d}s.
\end{aligned}$$

Notice that b_i'' is bounded with a module of continuity ϱ_i,and $\sup_{t\in[r,T]}|\Delta\Theta(t)| \le C\|\eta\|_\infty$. By Lebesgue dominated convergence theorem, we show that $\sup_{t\in[r,T]} R_2(t) \le C\|\tilde{\eta}\|_\infty\|\eta\|_\infty\varrho_i(\|\eta\|_\infty)$. ☐

Let $c \in C^3(\mathbb{R}^2;\mathbb{R})$ be a general cost with derivatives growing at most linearly. We consider

$$u(r,\omega_1,\omega_2) := E\bigg[\int_r^T c(\Theta^{r,\omega_1}_{1,*}(t;t), \Theta^{r,\omega_2}_{2,*}(t;t))\,\mathrm{d}t\bigg].$$

Remark 3.48. For example, we can take $c(x,y) = |f_1(x) - f_2(y)|^2$, where f_i has bounded first, second, and third derivatives.

Proposition 3.49. *Under Assumptions 3.30 and 3.31, we have u is twice Fréchet differentiable and weakly continuous. In particular, for $i, j = 1, 2$,*

$$\partial_{\omega_i} u(r, \omega_1, \omega_2) = E\left[\int_r^T \partial_i c(\Theta_{1,*}^{r,\omega_1}(t;t), \Theta_{2,*}^{r,\omega_2}(t;t))\Gamma_{i,*}^{r,\omega_i}(t)\,\mathrm{d}t\right], \tag{3.20}$$

$$\begin{aligned}\partial^2_{\omega_i\omega_j} u(r, \omega_1, \omega_2) = {}& E\left[\int_r^T \partial^2_{ij} c(\Theta_{1,*}^{r,\omega_1}(t;t), \Theta_{2,*}^{r,\omega_2}(t;t))\Gamma_{i,*}^{r,\omega_i}(t)\otimes\Gamma_{j,*}^{r,\omega_j}(t)\,\mathrm{d}t\right] \\ &+ \delta_{i,j} E\left[\int_r^T \partial_i c(\Theta_{1,*}^{r,\omega_1}(t;t), \Theta_{2,*}^{r,\omega_2}(t;t))\Xi_{i,*}^{r,\omega_i}(t)\,\mathrm{d}t\right],\end{aligned} \tag{3.21}$$

where $\delta_{i,j}$ is the Kronecker symbol.

Proof. The linear growth of $\partial^2_{ij} c$ yields

$$\left| c(\tilde{\theta}_1, \tilde{\theta}_2) - c(\theta_1, \theta_2) - \sum_{i=1,2} \partial_i c(\theta_1, \theta_2)(\tilde{\theta}_i - \theta_i) \right| \le C(1 + \sum_{i=1,2}(|\tilde{\theta}_i| + |\theta_i|)) \sum_{i=1,2}(\tilde{\theta}_i - \theta_i)^2.$$

Plugging $\tilde{\theta}_i = \Theta_{i,*}^{r,\omega_i+\eta_i}(t;t)$ and $\theta_i = \Theta_{i,*}^{r,\omega_i}(t;t)$ into the above estimates, and by Proposition 3.46, we deduce

$$\begin{aligned}u(r, \omega_1 + \eta_1, \omega_2 + \eta_2) - u(r, \omega_1, \omega_2) = {}& \sum_{i=1,2} E\left[\int_r^T \partial_i c(\Theta_{1,*}^{r,\omega_1}(t;t), \Theta_{2,*}^{r,\omega_2}(t;t))\langle \eta_i, \Gamma_{i,*}^{r,\omega_i}(t)\rangle\,\mathrm{d}t\right] \\ &+ o(\|\eta_1\|_\infty + \|\eta_2\|_\infty).\end{aligned}$$

Therefore, u is Fréchet differentiable, and (3.20) is verified. To show (3.21), we only need to notice that $\partial_i c, \partial^3_{ijk} c$ has a linear growth and $\sup_{t\in[r,T]}\langle \tilde{\eta}_i, \Gamma_{i,*}^{r,\omega_i}(t)\rangle \le C\|\tilde{\eta}_i\|_\infty$. By Proposition 3.47 and similar arguments as above, we deduce

$$\begin{aligned}&\langle \tilde{\eta}, \partial_{\omega_1} u(r, \omega_1 + \eta_1, \omega_2 + \eta_2) - \partial_{\omega_1} u(r, \omega_1, \omega_2)\rangle \\ &= E\left[\int_r^T \partial^2_{12} c(\Theta_{1,*}^{r,\omega_1}(t;t), \Theta_{2,*}^{r,\omega_2}(t;t))\langle \tilde{\eta}, \Gamma_{1,*}^{r,\omega_1}(t)\rangle\langle \eta_2, \Gamma_{2,*}^{r,\omega_2}(t)\rangle\,\mathrm{d}t\right] \\ &\quad + E\left[\int_r^T \partial^2_{11} c(\Theta_{1,*}^{r,\omega_1}(t;t), \Theta_{2,*}^{r,\omega_2}(t;t))\langle \tilde{\eta}, \Gamma_{1,*}^{r,\omega_i}(t)\rangle\langle \eta_1, \Gamma_{1,*}^{r,\omega_i}(t)\rangle\,\mathrm{d}t\right] \\ &\quad + E\left[\int_r^T \partial_1 c(\Theta_{1,*}^{r,\omega_1}(t;t), \Theta_{2,*}^{r,\omega_2}(t;t))\langle (\tilde{\eta}, \eta_1), \Xi_{1,*}^{r,\omega_1}(t)\rangle\,\mathrm{d}t\right] + o(\|\eta_1\|_\infty + \|\eta_2\|_\infty).\end{aligned}$$

Therefore, u is twice Fréchet differentiable and weakly continuous with derivatives given in (3.20) and (3.21). ∎

Proof of Lemma 3.38. We first show that u satisfies all conditions in Lemma 3.37. We recall the regularity condition here. For any $\eta, \tilde{\eta} \in C([0, T]; \mathbb{R})$, it holds that

(i) for any $\omega_1, \omega_2 \in C([0,T];\mathbb{R})$,

$$
\begin{aligned}
|\langle \eta, \partial_{\omega_i} u(r,\omega_1,\omega_2)\rangle| &\le C(1+\|\omega_1\|_\infty + \|\omega_2\|_\infty)\|\eta\|_\infty \\
|\langle (\eta,\tilde\eta), \partial^2_{\omega_i\omega_j} u(r,\omega_1,\omega_2)\rangle| &\le C(1+\|\omega_1\|_\infty + \|\omega_2\|_\infty)\|\eta\|_\infty\|\tilde\eta\|_\infty;
\end{aligned}
\tag{3.22}
$$

(ii) for any other $\omega_1', \omega_2' \in C([0,T];\mathbb{R})$, there exists a modulus of continuity ρ such that

$$
\begin{aligned}
&|\langle (\eta,\tilde\eta), \partial^2_{\omega_i\omega_j} u(r,\omega_1,\omega_2) - \partial^2_{\omega_i\omega_j} u(r,\omega_1',\omega_2')\rangle| \\
&\qquad \le C(1+\|\omega_1\|_\infty + \|\omega_2\|_\infty)\|\eta\|_\infty\|\tilde\eta\|_\infty \rho(\|\omega_1-\omega_1'\|_\infty + \|\omega_2-\omega_2'\|_\infty).
\end{aligned}
\tag{3.23}
$$

We first verify (3.22). By Propositions 3.44 and 3.45, we have

$$
\sup_{t\in[r,T]} E[|\Theta^{r,\omega_i}_{i,*}(t;t)|] \le C(1+\|\omega_i\|_\infty) \text{ and } |\langle \eta, \Gamma^{r,\omega_i}_{i,*}(t)\rangle| \le C\|\eta\|_\infty.
$$

Plugging the above into (3.20), we derive

$$
\begin{aligned}
|\langle \eta, \partial_{\omega_i} u(r,\omega_1,\omega_2)\rangle| &\le C\|\eta\|_\infty E\big[1+|\Theta^{r,\omega_1}_{1,*}(t;t)| + |\Theta^{r,\omega_2}_{2,*}(t;t)|\big] \\
&\le C(1+\|\omega_1\|_\infty + \|\omega_2\|_\infty)\|\eta\|_\infty.
\end{aligned}
$$

For the second derivative, we notice

$$
|\langle (\eta,\tilde\eta), \Gamma^{r,\omega_i}_{i,*}(t)\otimes\Gamma^{r,\omega_j}_{j,*}(t)\rangle| = |\langle \eta, \Gamma^{r,\omega_i}_{i,*}(t)\rangle\langle \tilde\eta, \Gamma^{r,\omega_j}_{j,*}(t)\rangle| \le C\|\eta\|_\infty\|\tilde\eta\|_\infty.
$$

Moreover, from (3.18) and the boundedness of b_i', b_i'', we deduce

$$
|\langle (\eta,\tilde\eta), \Xi^{r,\omega_i}_{i,*}(t)\rangle| \le C\int_r^t |\langle (\eta,\tilde\eta), \Gamma^{r,\omega_i}_{i,*}(s)\otimes\Gamma^{r,\omega_i}_{i,*}(s)\rangle|\,\mathrm{d}s \le C\|\eta\|_\infty\|\tilde\eta\|_\infty.
$$

Therefore, by Proposition 3.49 and the linear growth of $\partial_i c, \partial^2_{ij} c$, we derive

$$
|\langle (\eta,\tilde\eta), \partial^2_{\omega_i\omega_j} u(r,\omega_1,\omega_2)\rangle| \le C(1+\|\omega_1\|_\infty + \|\omega_2\|_\infty)\|\eta\|_\infty\|\tilde\eta\|_\infty.
$$

Now, we start to verify (3.23). Since $\partial^2_{ij} c$ has a linear growth, we have

$$
\begin{aligned}
&\Big|\partial_i c(\Theta^{r,\omega_1}_{1,*}(t;t), \Theta^{r,\omega_2}_{2,*}(t;t)) - \partial_i c(\Theta^{r,\omega_1'}_{1,*}(t;t), \Theta^{r,\omega_2'}_{2,*}(t;t))\Big| \\
&\le C(1+|\Theta^{r,\omega_1}_{1,*}(t;t)| + |\Theta^{r,\omega_2}_{2,*}(t;t)| + |\Theta^{r,\omega_1'}_{1,*}(t;t)| + |\Theta^{r,\omega_2'}_{2,*}(t;t)|) \\
&\qquad \times (|\Theta^{r,\omega_1}_{1,*}(t;t) - \Theta^{r,\omega_1'}_{1,*}(t;t)| + |\Theta^{r,\omega_2}_{2,*}(t;t) - \Theta^{r,\omega_2'}_{2,*}(t;t)|) \\
&\le C(1+|\Theta^{r,\omega_1}_{1,*}(t;t)| + |\Theta^{r,\omega_2}_{2,*}(t;t)| + \|\omega_1-\omega_1'\|_\infty + \|\omega_2-\omega_2'\|_\infty) \\
&\qquad \times (\|\omega_1-\omega_1'\|_\infty + \|\omega_2-\omega_2'\|_\infty).
\end{aligned}
$$

Similarly, as $\partial^3_{ijk}c$ has a linear growth, we have

$$\begin{aligned}&\left|\partial^2_{ij}c(\Theta^{r,\omega_1}_{1,*}(t;t),\Theta^{r,\omega_2}_{2,*}(t;t)) - \partial^2_{ij}c(\Theta^{r,\omega_1'}_{1,*}(t;t),\Theta^{r,\omega_2'}_{2,*}(t;t))\right|\\ &\leq C(1+|\Theta^{r,\omega_1}_{1,*}(t;t)|+|\Theta^{r,\omega_2}_{2,*}(t;t)|+\|\omega_1-\omega_1'\|_\infty+\|\omega_2-\omega_2'\|_\infty)\\ &\qquad\times(\|\omega_1-\omega_1'\|_\infty+\|\omega_2-\omega_2'\|_\infty).\end{aligned}$$

By Proposition 3.45, we have

$$\begin{aligned}&|\langle(\eta,\tilde\eta),\Gamma^{r,\omega_i}_{i,*}(t)\otimes\Gamma^{r,\omega_j}_{j,*}(t)-\Gamma^{r,\omega_i'}_{i,*}(t)\otimes\Gamma^{r,\omega_j'}_{j,*}(t)\rangle|\\ &\leq C|\langle\eta,\Gamma^{r,\omega_i}_{i,*}(t)-\Gamma^{r,\omega_i'}_{i,*}(t)\rangle\langle\tilde\eta,\Gamma^{r,\omega_j}_{j,*}(t)\rangle|+C|\langle\eta,\Gamma^{r,\omega_i'}_{i,*}(t)\rangle\langle\tilde\eta,\Gamma^{r,\omega_j}_{j,*}(t)-\Gamma^{r,\omega_j'}_{j,*}(t)\rangle|\\ &\leq C\|\eta\|_\infty\|\tilde\eta\|_\infty(\|\omega_i-\omega_i'\|_\infty+\|\omega_j-\omega_j'\|_\infty).\end{aligned}$$

Plugging the above estimates into (3.18), we derive

$$|\langle(\eta,\tilde\eta),\Xi^{r,\omega_i}_{i,*}(t)-\Xi^{r,\omega_i'}_{i,*}(t)\rangle|\leq C\|\eta\|_\infty\|\tilde\eta\|_\infty\varrho_i(\|\omega_i-\omega_i'\|_\infty),$$

where ϱ_i is the modulus of continuity of b_i''. Combining the above estimates, we conclude (3.23).

Now, we show that $\partial_t u$ exists and is continuous. By the Markov property of $(\Theta^{r,\omega_1}_{1,*},\Theta^{r,\omega_2}_{2,*})$, we have

$$u(r,\omega_1,\omega_2)=E\left[\int_r^{r+\delta}c\big(\Theta^{r,\omega_1}_{1,*}(t;t),\Theta^{r,\omega_2}_{2,*}(t;t)\big)\,\mathrm{d}t+u(r+\delta,\Theta^{r,\omega_1}_{1,*}(r+\delta;\cdot),\Theta^{r,\omega_2}_{2,*}(r+\delta;\cdot))\right].\tag{3.24}$$

Since we have verified (3.22) and (3.23), applying Itô formula we obtain

$$\begin{aligned}&u(r+\delta,\Theta^{r,\omega_1}_{1,*}(r+\delta;\cdot),\Theta^{r,\omega_2}_{2,*}(r+\delta;\cdot))-u(r+\delta,\omega_1,\omega_2)\\ &=\int_r^{r+\delta}(\mathcal{L}_1+\mathcal{L}_2)u(s,\Theta^{r,\omega_1}_{1,*}(s;\cdot),\Theta^{r,\omega_2}_{2,*}(s;\cdot))\,\mathrm{d}s\\ &\quad+\int_r^{r+\delta}\langle(k_1(\cdot,s),k_2(\cdot,s)),\partial^2_{\omega_1\omega_2}u(s,\Theta^{r,\omega_1}_{1,*}(s;\cdot),\Theta^{r,\omega_2}_{2,*}(s;\cdot))\rangle\,\mathrm{d}s\\ &\quad+\int_r^{r+\delta}\langle k_1(\cdot,s),\partial_{\omega_1}u(s,\Theta^{r,\omega_1}_{1,*}(s;\cdot),\Theta^{r,\omega_2}_{2,*}(s;\cdot))\rangle\,\mathrm{d}B_1(s)\\ &\quad+\int_r^{r+\delta}\langle k_2(\cdot,s),\partial_{\omega_2}u(s,\Theta^{r,\omega_1}_{1,*}(s;\cdot),\Theta^{r,\omega_2}_{2,*}(s;\cdot))\rangle\,\mathrm{d}B_1(s).\end{aligned}$$

Plug the above identity into (3.24) and divide both sides by δ. Let δ go to 0, and we deduce u satisfies

$$(\partial_t+\mathcal{L}_1+\mathcal{L}_2)u(r,\omega_1,\omega_2)+\langle(k_1(\cdot,r),k_2(\cdot,r)),\partial^2_{\omega_1\omega_2}u(r,\omega_1,\omega_2)(\cdot)\rangle=-c(\omega_1(r),\omega_2(r)).\tag{3.25}$$

This gives the continuity of $\partial_t u$. We conclude the proof by noticing $u=V_*$ if we take $c(x,y)=|x-y|^2$. □

Chapter 4

Adapted Wasserstein DRO: Duality

4.1 Introduction

As Box famously said (Box, 1976), *all models are wrong, but some are useful.* In real-world applications, practitioners often need to trade off between a model's fidelity (its capability to capture system features) and a model's tractability (its capability to provide interpretable solutions). A postulated model may capture some important aspects of reality, while inevitably ignoring some other aspects. This is especially true in mathematical finance, where the model is often derived from theoretical considerations, possibly combined with some calibration to market data. This inherent Knightian uncertainty (Knight, 1921) regarding the model is of fundamental importance and a subject of intense studies in mathematics and economics alike. Mathematical frameworks such as risk measures (Föllmer and Schied, 2008) and sublinear expectations (Peng, 2019) have been developed to take into account such uncertainty. More recently, Wasserstein distributionally robust optimization (W-DRO) has emerged as a powerful tool to counter model uncertainty (Mohajerin Esfahani and Kuhn, 2018, Blanchet and Murthy, 2019, Gao and Kleywegt, 2022). It is formulated as minimax problem that optimizes a worst-case objective, evaluated over a collection of models. This collection is often referred to as the ambiguity set and is taken as a Wasserstein ball centered at a reference model.

In this chapter, we extend Wasserstein distributionally robust optimization (W-DRO) to a dynamic setting, where model uncertainty is quantified by the causal/adapted Wasserstein distance. This distance not only captures the spatial differences between two models but also the information flow they generate. The choice of ambiguity set here is natural in a dynamic setting and, to some extent, becomes essential when considering optimal stopping problems. Crucially, it allows for models with potentially different support from the reference model, similar to the classical W-DRO, while excluding models that can only be obtained from an *anticipative* perturbation of the reference model.

Despite its appealing theoretical framework, adapted Wasserstein distributionally robust optimization (AW-DRO) often suffers from the computational challenges inherited from causal optimal transport. Our contribution here is to fill this gap by providing a tractable dynamic duality formula for the AW-DRO problem in both discrete and continuous-time settings. We focus on the distributional model risk:

$$\sup_{\nu \in B_\delta(\mu)} E_\nu[f(X)],$$

where $B_\delta(\mu)$ is an ambiguity set given by a causal Wasserstein ball or an adapted Wasserstein ball. We will extend this ambiguity set constraint to a more general penalized form

$$V_\star = \sup_{\nu \in \mathscr{P}(\mathcal{X})} \{E_\nu[f(X)] - L(\mathcal{T}_\star(\mu,\nu))\}, \tag{4.1}$$

where L is a penalty function, and $\mathcal{T}_\star$ represents the optimal causal/bi-causal transport cost given by

$$\mathcal{T}_\star(\mu,\nu) = \inf_{\pi \in \Pi_\star(\mu,\nu)} E_\pi[c(X,Y)],$$

with $\star = \mathfrak{c}/\mathfrak{bc}$ respectively. By choosing an indicator penalization L and an appropriate cost c, one can easily recover the corresponding ambiguity set $B_\delta(\mu)$.

While $V_\mathfrak{c}$ and $V_\mathfrak{bc}$ are not equal a priori, we will show that under mild regularity conditions, these two penalizations are equivalent. This allows us to omit the subscript when there is no ambiguity. Our main results, detailed in Sections 4.4 and 4.5, show that

$$V = \inf_{\lambda \geq 0}\{L^*(\lambda) + U(\lambda)\},$$

where L^* is the convex conjugate of L, and U is a convex function given by a dynamic programming principle. In particular, U can be computed recursively in discrete time, or solved by a path-dependent Hamilton–Jacobi–Bellman (HJB) equation in continuous time. Moreover, in Sections 4.6 and 4.7 we replace the linear expectation $\nu \mapsto E_\nu[f(X)]$ in (4.1) with a nonlinear functional $\nu \mapsto F(\nu)$ and derive duality formulas in discrete time. In Section 4.6, we take F as the expected shortfall, a concave functional of the model; in Section 4.7, F is taken as the value of an optimal stopping problem, which is in general not concave.

To the best of our knowledge, this is the first result that tackles the duality of AW-DRO in continuous time and considers optimal stopping problems. In continuous time, we reformulate the dual problem as a non-Markovian stochastic control problem and utilize the tools from functional Itô calculus (Cont and Fournié, 2010, Bally et al., 2016) to establish a path-dependent HJB equation. For the AW-DR optimal stopping problems, we leverage a novel concave relaxation which lifts the optimal stopping problem to the space of adapted stochastic processes. The original problem is then recast as a Wasserstein DRO problem on the 'nested space'.

4.1.1 Related literature

Distributionally robust optimization (DRO) provides a framework for decision-making when the underlying stochastic model is uncertain. Unlike classical stochastic optimization, which assumes a single reference distribution, DRO considers a collection of plausible distributions often defined via statistical distances such as the Wasserstein distance or ϕ-divergences and optimizes for the worst-case expectation over this ambiguity set. This review is focused on the transport-type DRO; for a broader survey of the field, we refer interested readers to Rahimian and Mehrotra (2022), Kuhn et al. (2025).

Wasserstein DRO has found broad applications in operations research, mathematical finance (Bartl et al., 2020, Blanchet et al., 2022), and machine learning (Bai et al., 2023, 2025, Blanchet and Murthy, 2019), where robustness to model misspecification is crucial. A key development in Wasserstein DRO is its duality theory, which reveals a close connection to regularized optimization. This theory has been progressively generalized, starting from the data-driven case where the reference measure is an empirical measure (Mohajerin Esfahani and Kuhn, 2018), extending to Borel measures on Euclidean spaces (Gao and Kleywegt, 2022), general Polish spaces (Blanchet and Murthy, 2019), and to spaces with the interchangeability property, such as Suslin spaces (Zhang et al., 2024). More recently, variants of the classical Wasserstein DRO also have been introduced. These include the robust optimized certainty equivalents in a penalized form (Bartl et al., 2020), a weak optimal transport-type DRO (Kupper et al., 2023), and problems with marginal uncertainty in both source and target distributions (Fan et al., 2023).

Few results are available for the adapted Wasserstein DRO in a dynamic context, with all existing work focusing solely on a discrete-time setting. Addressing the causality constraint is a key challenge. In a recent work Han (2025), the author reformulated the causality constraint as an infinite-dimensional linear constraint, which leads to a dual optimization problem over an infinite-dimensional test function space. Our proposed dynamic duality provides a more tractable solution. It not only accommodates a more general penalized setting (4.1), but critically, it leverages the temporal structure of causal couplings. We remark that our the discrete-time duality Theorem 4.24 was also independently obtained in Gao et al. (2022), Yang et al. (2022).

Our continuous-time results rely on tools from functional Itô calculus, which was first proposed by Dupire (2009), and systematically studied in Cont and Fournié (2010, 2013). It was then applied to study non-Markovian stochastic control problems and their associated path-dependent HJB equations. A verification theorem for the classical solution was established in Bally et al. (2016, Chapter 8.3). The viscosity solution theory, however, has proven more intricate, leading to several proposed notions, for example in Tang and Zhang (2015), Ekren et al. (2014, 2016a,b), etc.

4.1.2 Outline

The rest of the chapter is organized as follows. In Section 4.2, we introduce the basic notations and tools. In Section 4.3, as an intermediate step, we derive a general duality formula for AW-DRO problems where the penalty is given as a function of causal optimal transport problem. The dual problem involves a causal transport problem with a fixed source distribution only. A generalized Fenchel–Moreau duality theorem is proved in Lemma 4.21. In Section 4.4, we focus on the discrete-time setting and derive a dynamic duality formula in Theorem 4.24. Under a mild continuity condition, the equivalence between the causal penalization and bi-causal penalization is established in Theorem 4.27. In Section 4.5, we consider a continuous-time setting with a penalty given by the Cameron–Martin adapted Wasserstein distance. We reformulate the dual problem as a stochastic control problem and identify the worst-case distribution via a path-dependent HJB equation in Theorem 4.30. In Sections 4.6 and 4.7, we extend Theorem 4.24 in two directions. We replace the linear expectation with the expected shortfall in Section 4.6, and study a numerical example of a two-step exotic option. In Section 4.7, we consider optimal stopping problems in the AW-DRO framework. A duality formula is derived in Theorem 4.42 by lifting the original problem to the space of adapted processes.

4.2 Preliminaries

Let $N \in \mathbb{Z}_+$ be the number of steps. For any $1 \le n \le m \le N$ and an N-tuple $(\theta_1, \dots, \theta_N)$, we denote the truncation $(\theta_n, \dots, \theta_m)$ by $\theta_{n:m}$.

By $\Pi(\mu, *)$ we denote the set of couplings with a fixed first marginal μ. Accordingly, $\Pi_{\mathfrak{c}}(\mu, *)$ and $\Pi_{\mathfrak{bc}}(\mu, *)$ represent the respective subsets of causal and bi-causal couplings.

4.2.1 Analytic sets and universal measurability

The analytic sets (also known as Suslin sets) are widely applied to reconcile the measurability issue in dynamic programming principle. We give a minimal introduction here to serve our purpose.

Definition 4.1. Let $\mathcal{X}$ be a Polish space. We say S a subset of $\mathcal{X}$ is analytic if S is a continuous image of a Polish space, and a function $\varphi : \mathcal{X} \to \mathbb{R}$ is upper semi-analytic if the level set $\{\varphi \ge r\}$ is analytic for any $r \in \mathbb{R}$. The universal σ-algebra of $\mathcal{X}$ is defined as $\bigcap_{\mu \in \mathscr{P}(\mathcal{X})} {}^{\mu}\mathcal{B}(X)$.

To ease the notation, we will not distinguish any Borel measure μ and its completion. So, for any universally measurable function φ, $\mu(\varphi)$ is understood as the integration with respect to the completion of μ.

Proposition 4.2 (Bertsekas and Shreve (1996), Corollary 7.42.1)**.** *Let $S \subseteq \mathcal{X}$ be an analytic set. Then S is universally measurable, and therefore any upper semi-analytic function is universally measurable.*

Proposition 4.3 (Bertsekas and Shreve (1996), Proposition 7.39)**.** *Let $\mathcal{X}$ and $\mathcal{Y}$ be Polish spaces and $D \subseteq \mathcal{X} \times \mathcal{Y}$ an analytic set. Then the projection $\mathrm{pj}_{\mathcal{X}}(D) := \{x \in \mathcal{X} : (x, y) \in D$ for some $y\}$ is an analytic subset of $\mathcal{X}$.*

We also recall the following proposition from Bertsekas and Shreve (1996, Proposition 7.50).

Proposition 4.4 (Analytic selection theorem)**.** *Let $\mathcal{X}$ and $\mathcal{Y}$ be Polish spaces and $D \subseteq \mathcal{X} \times \mathcal{Y}$ an analytic set, and $\varphi : D \to \mathbb{R}$ an upper semi-analytic function. Define $\tilde{\varphi} : \mathrm{pj}_{\mathcal{X}}(D) \to \mathbb{R}^*$ by*

$$\tilde{\varphi}(x) = \sup_{y \in D_x} \varphi(x, y),$$

where $D_x = \{y \in \mathcal{Y} : (x, y) \in D\}$*. Then for any* $\varepsilon > 0$*, there exists an analytically measurable function* $s : \mathrm{pj}_{\mathcal{X}}(D) \to \mathcal{X}$ *such that* $(x, s(x)) \in D$ *for any* $x \in \mathrm{pj}_{\mathcal{X}}(D)$*, and*

$$\varphi(x, s(x)) \geq \begin{cases} \tilde{\varphi}(x) - \varepsilon & \text{if} \quad \tilde{\varphi} < \infty, \\ 1/\varepsilon & \text{if} \quad \tilde{\varphi} = \infty. \end{cases}$$

The following result is adapted from Zhang et al. (2024).

Lemma 4.5. *Let* $(\mathcal{X}, \mathcal{B})$ *be a Polish space equipped with its Borel* σ*-algebra,* $\varphi : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ *upper semi-analytic. Define* $\tilde{\varphi} : \mathcal{X} \to \mathbb{R}^*$ *as*

$$\tilde{\varphi}(x) = \sup_{y \in \mathcal{X}} \varphi(x, y).$$

Then $\tilde{\varphi}$ *is universally measurable; moreover, it holds that*

$$E_\mu[\sup_{y \in \mathcal{X}} \varphi(X, y)] = \sup_{\pi \in \Pi(\mu, *)} E_\pi[\varphi(X, Y)].$$

Proof. Notice that for any $r \in \mathbb{R}$ we have

$$\big\{x : \sup_{y \in \mathcal{X}} \varphi(x, y) > r\big\} = \mathrm{pj}_{\mathcal{X}}(\{(x, y) : \varphi(x, y) > r\}).$$

Since φ is upper semi-analytic, we have $\tilde{\varphi}$ is also upper semi-analytic by Proposition 4.3. Moreover, by Proposition 4.2, we know $\tilde{\varphi}$ is universally measurable. It follows from the definition that

$$E_\mu[\sup_{y \in \mathcal{X}} \varphi(X, y)] \geq \sup_{\pi \in \Pi(\mu, *)} E_\pi[\varphi(X, Y)].$$

Now, by Proposition 4.4 there exists an analytically measurable function $s_n : \mathcal{X} \to \mathcal{X}$ such that

$$\varphi(x, s_n(x)) \geq \begin{cases} \tilde{\varphi}(x) - \frac{1}{n} & \text{if} \quad \tilde{\varphi}(x) < \infty, \\ n & \text{if} \quad \tilde{\varphi}(x) = \infty. \end{cases}$$

Moreover, there exists a Borel measurable function t_n such that $t_n = s_n$ holds μ-a.s. Then we take $\pi_n = (\mathrm{Id}, t_n)_\# \mu \in \Pi(\mu, *)$. We derive

$$E_{\pi_n}[\varphi(X, Y)] = E_\mu[\varphi(X, t_n(X))] \geq E_\mu[\tilde{\varphi} \mathbb{1}_{\{\tilde{\varphi} < \infty\}}] + n E_\mu[\mathbb{1}_{\{\tilde{\varphi} = \infty\}}] - \frac{1}{n}.$$

As $n \to \infty$, we conclude

$$\sup_{\pi \in \Pi(\mu, *)} E_\pi[\varphi(X, Y)] \geq E_\mu[\sup_{y \in \mathcal{X}} \varphi(X, y)].$$

$\square$

4.2.2 Convex analysis

We recall several basic concepts and results in convex analysis.

Definition 4.6. Let $\varphi : \mathbb{R} \to \mathbb{R}^*$ be a convex function. We define the *domain* of φ as

$$\mathrm{dom}(\varphi) = \{x \in \mathbb{R} : \varphi(x) < \infty\}.$$

We say φ is *proper* if $\mathrm{dom}(\varphi) \neq \emptyset$, and φ is *closed* if $\mathrm{dom}(\varphi)$ is closed. A function ψ is proper closed concave if and only if $-\psi$ is proper closed convex.

Definition 4.7. Let $\varphi : \mathbb{R} \to \mathbb{R}^*$ be a convex function. The convex conjugate of $\varphi : \mathbb{R} \to \mathbb{R}^*$ is defined as $\varphi^*(y) = \sup_{x\in\mathbb{R}}\{xy - \varphi(x)\}$. Similarly, we define the concave conjugate of ψ as $\psi_*(y) = \inf_{x\in\mathbb{R}}\{xy - \psi(x)\}$.

The following result is the celebrated convex duality theorem from Rockafellar (1997, Corollary 12.2.1).

Theorem 4.8 (Fenchel–Moreau Theorem)**.** *Let $\varphi : \mathbb{R} \to \mathbb{R}^*$ be a closed proper convex (concave) function. Then, we have $\varphi^{**} = \varphi$ ($\varphi_{**} = \varphi$).*

We introduce the subdifferential of a convex function as an extension of the classical derivative to the convex functions which are not necessarily differentiable.

Definition 4.9 (Subdifferential)**.** Let $\varphi : \mathbb{R} \to \mathbb{R}^*$ be a convex function. For any $x \in \mathbb{R}$, we define the *subdifferential* of φ at x as

$$\partial\varphi(x) = \{y \in \mathbb{R} : \varphi(x') \geq \varphi(x) + y(x' - x) \quad \forall x' \in \mathbb{R}\}.$$

Proposition 4.10 (Rockafellar (1997), Theorem 23.5)**.** *Let $\varphi : \mathbb{R} \to \mathbb{R}^*$ be a convex function. If φ is proper and closed, then $y \in \partial\varphi(x)$ if and only if $x \in \partial\varphi^*(y)$. And we have the equality $\varphi(x) + \varphi^*(y) = xy$.*

4.2.3 Horizontal and vertical derivatives

Following Dupire (2009), Cont and Fournié (2010), Bally et al. (2016), we introduce the horizontal and vertical derivatives of a non-anticipative functional. We start with the definition of a non-anticipative functional on the càdlàg path space $D([0, T]; \mathbb{R}^n)$.

Definition 4.11. We say a functional $F : [0, T] \times D([0, T]; \mathbb{R}^n) \to \mathbb{R}$ is non-anticipative if $F(t, x) = F(t, X(\cdot \wedge t))$ for any $t \in [0, T]$ and $x \in D([0, T]; \mathbb{R}^n)$.

Definition 4.12. A non-anticipative functional F is said to be *horizontally* differentiable at $(t,x) \in [0,T] \times D([0,T];\mathbb{R}^n)$ if the limit

$$\mathcal{D}F(t,x) = \lim_{h \to 0^+} \frac{F(t+h, x(\cdot \wedge t)) - F(t, x(\cdot \wedge t))}{h} \text{ exists.}$$

We call $\mathcal{D}F(t,x)$ the *horizontal* derivative of F at (t,x).

Definition 4.13. A non-anticipative functional F is said to be *vertically* differentiable at $(t,x) \in [0,T] \times D([0,T];\mathbb{R}^n)$ if the map

$$\mathbb{R}^n \ni e \mapsto F(t, x(\cdot \wedge t) + e\mathbb{1}_{[t,T]}) \in \mathbb{R}$$

is differentiable at 0. We call its gradient at 0 the *vertical* derivative of F at (t,x) and denote it by $\nabla_x F(t,x)$.

Before we proceed to the non-anticipative functional on the continuous path space, we introduce some regularity conditions. We equip the product space $[0,T] \times D([0,T];\mathbb{R}^n)$ with the metric

$$d_\infty((t,x),(t',x')) := |t-t'| + \sup_{s \in [0,T]} \|x(\cdot \wedge t) - x'(\cdot \wedge t')\|_\infty.$$

Definition 4.14 (Left continuity)**.** We say a non-anticipative functional F is left continuous if for any $(t,x) \in [0,T] \times D([0,T];\mathbb{R}^n)$ and any sequence $(t_n, x_n) \overset{d_\infty}{\to} (t,x)$ in $[0,T] \times D([0,T];\mathbb{R}^n)$ with t_n increasing, we have

$$F(t_n, x_n) \to F(t,x).$$

By C_l we denote the space of left continuous non-anticipative functionals which are also continuous at any fixed time $t \in [0,T]$.

Definition 4.15 (Boundedness preserving)**.** We say a non-anticipative functional F is boundedness preserving if for any $M > 0$ and $t_0 < T$ there exists C_{M,t_0} such that

$$|F(t,x)| \le C_{M,t_0} \quad \text{if} \quad \|x(\cdot \wedge t_0)\|_\infty \le M.$$

By C_b we denote the space of bounded preserving non-anticipative functionals.

The following class of regular non-anticipative functionals plays a special role in functional Itô calculus.

Definition 4.16. We say a non-anticipative functional F is in the class $C_b^{1,2}$ if the following conditions hold:

(i) F admits horizontal and the first and second vertical derivatives for any $(t, x) \in [0, T] \times D([0, T]; \mathbb{R}^n)$.

(ii) $F, \mathcal{D}F, \nabla_x F, \nabla_x^2 F \in C_l$.

(iii) $\mathcal{D}F, \nabla_x F, \nabla_x^2 F \in C_b$.

Theorem 4.17 (Bally et al. (2016), Theorem 5.27, Theorem 5.28)**.** *Let F_1, F_2 be two non-anticipative functionals in $C_b^{1,2}$. If F_1 and F_2 are equal on all continuous paths, then their first and second horizontal derivatives coincide on all continuous paths.*

We say a non-anticipative functional $F : [0, T] \times C([0, T]; \mathbb{R}^n)$ is of class $C_b^{1,2}$ if there exists a non-anticipative functional $G : [0, T] \times D([0, T]; \mathbb{R}^n) \to \mathbb{R}$ such that $F = G$ on $[0, T] \times C([0, T]; \mathbb{R}^n)$ and $G \in C_b^{1,2}$. The vertical derivative of F is defined as the one of G. The above theorem ensures such definition is independent of the choice of G.

4.3 A convex duality

In this section, we present a convex duality in Proposition 4.19 which works in both discrete and continuous-time settings. It decouples the loss function L and acts as an intermediate step to derive the duality for adapted Wasserstein DRO problems.

Assumption 4.18. We assume the following conditions:

(i) $L : \mathbb{R}^* \to \mathbb{R}^*$ is a non-decreasing closed proper convex function with $L(0) = 0$ and $L(+\infty) = +\infty$.

(ii) $c : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^*$ is lower semi-analytic and non-negative with $c(x, y) = 0$ if and only if $x = y$ for any $x, y \in \mathcal{X}$.

(iii) $f : \mathcal{X} \to \mathbb{R}$ is upper semi-analytic and μ-integrable.

Proposition 4.19. *Under Assumption 4.18, we have the duality*

$$V_{\mathfrak{c}} = \sup_{\nu \in \mathscr{P}(\mathcal{X})} \{E_\nu[f(X)] - L(\mathcal{T}_{\mathfrak{c}}(\mu, \nu))\} = \inf_{\lambda \geq 0} \{L^*(\lambda) + U(\lambda)\},$$

where L^* is the convex conjugate of L, and $U(\lambda) = \sup_{\pi \in \Pi_{\mathfrak{c}}(\mu, *)} E_\pi[f(Y) - \lambda c(X, Y)]$. Moreover, there exists a dual optimizer λ_* attaining the infimum. If there exist $\nu_* \in \mathcal{X}$ and $\pi_* \in \Pi_{\mathfrak{c}}(\mu, \nu_*)$ such that

$$L^*(\lambda_*) = \lambda_* E_{\pi_*}[c(X, Y)] - L(E_{\pi_*}[c(X, Y)]) \text{ and } U(\lambda_*) = E_{\pi_*}[f(Y) - \lambda_* c(X, Y)], \tag{4.2}$$

then it holds

$$V_{\mathfrak{c}} = E_{\nu_*}[f(X)] - L(\mathcal{T}_{\mathfrak{c}}(\mu, \nu_*)) = L^*(\lambda_*) + U(\lambda_*).$$

Remark 4.20. Let $\mathcal{P} \ni \mu$ be a convex subset of $\mathscr{P}(\mathcal{X})$ and $\Pi_{\mathfrak{c}}(\mu, \mathcal{P}) = \bigcup_{\nu \in \mathcal{P}} \Pi_{\mathfrak{c}}(\mu, \nu)$. One can extend the above duality to a closed proper concave functional $F : \mathcal{P} \to \mathbb{R}$. Under Assumption 4.18 (i) and (ii), it holds

$$\sup_{\nu \in \mathcal{P}} \{F(\nu) - L(\mathcal{T}_{\mathfrak{c}}(\mu, \nu))\} = \inf_{\lambda \geq 0} \{L^*(\lambda) + \sup_{\nu \in \mathcal{P}} \{F(\nu) - \mathcal{T}_{\mathfrak{c}}(\mu, \nu)\}\}.$$

Proof. Step 1. We first show the case $L = +\infty \mathbb{1}_{(\delta, +\infty]}$ for some $\delta \geq 0$. We write

$$V(\delta) = \sup_{\mathcal{T}_{\mathfrak{c}}(\mu, \nu) \leq \delta} E_\nu[f(X)].$$

We claim that $V(\delta)$ is proper and concave by naturally setting V as $-\infty$ on the negative real line. The concavity of V follows directly from the convexity of the set of causal couplings $\Pi_{\mathfrak{c}}(\mu, *)$. By Assumption 4.18 (ii) and (iii), we obtain V is proper as $V(0) = E_\mu[f(X)] < \infty$. If $V(\delta) = +\infty$ for some $\delta > 0$, then $V(\delta) = +\infty$ for any $\delta > 0$. In this case it is direct to verify the desired duality

$$V(\delta) = \inf_{\lambda \geq 0} \{\lambda \delta + \sup_{\pi \in \Pi_{\mathfrak{c}}(\mu, *)} \{E_\pi[f(Y) - \lambda c(X, Y)]\}\} = +\infty.$$

Now, we focus on the case $V(\delta) < +\infty$ for any $\delta \geq 0$. In particular, V is closed, proper, and concave. We calculate the concave conjugate of V as

$$\begin{aligned}
V_*(\lambda) = \inf_{\delta \geq 0} \{\lambda \delta - \sup_{\mathcal{T}_{\mathfrak{c}}(\mu, \nu) \leq \delta} E_\nu[f(X)]\} &= \inf_{\delta \geq 0} \inf_{\mathcal{T}_{\mathfrak{c}}(\mu, \nu) \leq \delta} \{\lambda \delta - E_\nu[f(X)]\} \\
&= \inf_{\nu \in \mathscr{P}(\mathcal{X})} \inf_{\delta \geq \mathcal{T}_{\mathfrak{c}}(\mu, \nu)} \{\lambda \delta - E_\nu[f(X)]\} \\
&= \inf_{\nu \in \mathscr{P}(\mathcal{X})} \{\lambda \mathcal{T}_{\mathfrak{c}}(\mu, \nu) - E_\nu[f(X)]\} \\
&= \inf_{\pi \in \Pi_{\mathfrak{c}}(\mu, *)} E_\pi[\lambda c(X, Y) - f(Y)].
\end{aligned}$$

By Fenchel–Moreau Theorem 4.8, we have $V = (V_*)_*$ which yields

$$\begin{aligned}
V(\delta) &= \inf_{\lambda \geq 0} \{\lambda \delta - V_*(\lambda)\} \\
&= \inf_{\lambda \geq 0} \{\lambda \delta - \inf_{\pi \in \Pi_{\mathfrak{c}}(\mu, *)} \{E_\pi[\lambda c(X, Y) - f(Y)]\}\} \\
&= \inf_{\lambda \geq 0} \{\lambda \delta + \sup_{\pi \in \Pi_{\mathfrak{c}}(\mu, *)} \{E_\pi[f(Y) - \lambda c(X, Y)]\}\}.
\end{aligned}$$

The desired duality is shown by noticing $L^*(\lambda) = \delta \lambda$ when $L = +\infty \mathbb{1}_{(\delta, +\infty]}$.

Step 2. We consider a general penalization L satisfying Assumption 4.18 (i). We notice that

$$\sup_{\nu\in\mathscr{P}(\mathcal{X})}\{E_\nu[f(X)]-L(\mathcal{T}_\mathfrak{c}(\mu,\nu))\}\le\sup_{\nu\in\mathscr{P}(\mathcal{X})}\{V(\mathcal{T}_\mathfrak{c}(\mu,\nu))-L(\mathcal{T}_\mathfrak{c}(\mu,\nu))\}$$
$$\le\sup_{\delta\ge0}\{V(\delta)-L(\delta)\}.$$

On the other hand, since L is non-decreasing, we have for any $\delta\ge0$

$$V(\delta)-L(\delta)=\sup_{\mathcal{T}_\mathfrak{c}(\mu,\nu)\le\delta}\{E_\nu[f(X)]-L(\delta)\}\le\sup_{\mathcal{T}_\mathfrak{c}(\mu,\nu)\le\delta}\{E_\nu[f(X)]-L(\mathcal{T}_\mathfrak{c}(\mu,\nu))\}.$$

Hence, we derive $\sup_{\nu\in\mathscr{P}(\mathcal{X})}\{E_\nu[f(X)]-L(\mathcal{T}_\mathfrak{c}(\mu,\nu))\}=\sup_{\delta\ge0}\{V(\delta)-L(\delta)\}$. Plugging the result from *Step 1*, we obtain

$$\begin{aligned}\sup_{\nu\in\mathscr{P}(\mathcal{X})}\{E_\nu[f(X)]-L(\mathcal{T}_\mathfrak{c}(\mu,\nu))\}&=\sup_{\delta\ge0}\{V(\delta)-L(\delta)\}\\&=\sup_{\delta\ge0}\inf_{\lambda\ge0}\{\lambda\delta-L(\delta)+\sup_{\pi\in\Pi_\mathfrak{c}(\mu,*)}E_\pi[f(Y)-\lambda c(X,Y)]\}\\&:=\sup_{\delta\ge0}\inf_{\lambda\ge0}\{\lambda\delta-L(\delta)+U(\lambda)\}.\end{aligned}$$

Here, $U(\lambda)=\sup_{\pi\in\Pi_\mathfrak{c}(\mu,*)}E_\pi[f(Y)-\lambda c(X,Y)]$ either is equal to $+\infty$ for any $\lambda\ge0$ or is a closed proper convex function of λ. In the former case, it is direct to verify

$$\sup_{\nu\in\mathscr{P}(\mathcal{X})}\{E_\nu[f(X)]-L(\mathcal{T}_\mathfrak{c}(\mu,\nu))\}=\inf_{\lambda\ge0}\{L^*(\lambda)+U(\lambda)\}=+\infty.$$

In the latter case, the duality will follow from Lemma 4.21 below.

Step 3. We notice by definition (L^*+U) is lower semi-continuous and goes to infinity at infinity. Therefore, there exists a minimizer λ_*. Given a pair (ν_*,π_*) satisfying the slackness condition (4.2), we derive

$$\begin{aligned}E_{\nu_*}[f(X)]-L(\mathcal{T}_c(\mu,\nu_*))&\ge\lambda_*E_{\pi_*}[c(X,Y)]-L(E_{\pi_*}[c(X,Y)])+E_{\pi_*}[f(Y)-\lambda_*c(X,Y)]\\&=L^*(\lambda_*)+U(\lambda)=V_\mathfrak{c}.\end{aligned}$$

The reverse direction is trivial and we complete the proof. □

We did not find a direct reference for the following lemma, so we include a proof for the completeness.

Lemma 4.21. *Let $\varphi,\psi:\mathbb{R}\to\mathbb{R}^*$ be two closed proper convex functions. Then, we have the following minimax theorem*

$$\sup_{x\in\mathbb{R}}\inf_{y\in\mathbb{R}}\{xy-\varphi(x)+\psi(y)\}=\inf_{y\in\mathbb{R}}\sup_{x\in\mathbb{R}}\{xy-\varphi(x)+\psi(y)\}.$$

Remark 4.22. This lemma generalizes Fenchel–Moreau theorem. By taking ψ as a linear function, we retrieve the classical Fenchel–Moreau theorem $\varphi^{**} = \varphi$.

Proof. By definition, we have

$$\sup_{x\in\mathbb{R}} \inf_{y\in\mathbb{R}} \{xy - \varphi(x) + \psi(y)\} \leq \inf_{y\in\mathbb{R}} \sup_{x\in\mathbb{R}} \{xy - \varphi(x) + \psi(y)\}.$$

For the other direction, we know

$$\inf_{y\in\mathbb{R}} \sup_{x\in\mathbb{R}} \{xy - \varphi(x) + \psi(y)\} = \inf_{y\in\mathbb{R}} \{\varphi^*(y) + \psi(y)\}.$$

Without loss of generality, we take y_0 satisfying $0 \in \partial\varphi^*(y_0) + \partial\psi(y_0)$ and $x_0 \in \partial\varphi^*(y_0) \cap -\partial\psi(y_0)$. Otherwise, we have φ^* and ψ are monotone and bounded from below, and thus

$$\begin{aligned}\inf_{y\in\mathbb{R}} \{\varphi^*(y) + \psi(y)\} &= \inf_{y\in\mathbb{R}} \varphi^*(y) + \inf_{y\in\mathbb{R}} \psi(y) \\ &= -\varphi(0) + \inf_{y\in\mathbb{R}} \psi(y) \leq \sup_{x\in\mathbb{R}} \inf_{y\in\mathbb{R}} \{xy - \varphi(x) + \psi(y)\}.\end{aligned}$$

By Proposition 4.10, we have

$$\varphi(x_0) + \varphi^*(y_0) = x_0 y_0,$$

and

$$\psi(y) - \psi(y_0) + x_0(y - y_0) \geq 0.$$

Therefore, we derive

$$\begin{aligned}\sup_{x\in\mathbb{R}} \inf_{y\in\mathbb{R}} \{xy - \varphi(x_0) + \psi(y)\} &\geq \inf_{y\in\mathbb{R}} \{x_0 y - \varphi(x) + \psi(y)\} \\ &= \inf_{y\in\mathbb{R}} \{x_0 y_0 - \varphi(x_0) + \psi(y_0) + [\psi(y) - \psi(y_0) + x_0(y - y_0)]\} \\ &\geq \varphi^*(y_0) + \psi(y_0) \geq \inf_{y\in\mathbb{R}} \sup_{x\in\mathbb{R}} \{xy - \varphi(x) + \psi(y)\}.\end{aligned}$$

□

4.4 Discrete-time results

In this section, we focus on the discrete-time setting. We take $I = \{0, 1, \ldots, N\}$ and $\mathcal{X} = \mathcal{X}_0 \times \mathcal{X}_1 \times \cdots \times \mathcal{X}_N$. In light of Proposition 4.19, it suffices to compute $\sup_{\pi\in\Pi_c(\mu,*)} E_\pi[f(Y) - \lambda c(X, Y)]$, which can be viewed as causal optimal transport problem with constraint of the source measure only. We will exploit a key dynamic temporal structure of causal couplings from the following proposition.

Proposition 4.23 (Backhoff-Veraguas et al. (2017)). *The following statements are equivalent:*

- *$\pi \in \Pi(\mu, *)$ is a causal coupling with the first marginal μ.*
- *Decomposing π in terms of successive regular kernels*

$$\pi(\mathrm{d}x, \mathrm{d}y) = \pi_0(\mathrm{d}x_0, \mathrm{d}y_0)\pi_{x_0,y_0}(\mathrm{d}x_1, \mathrm{d}y_1) \cdots \pi_{x_{0:N-1},y_{0:N-1}}(\mathrm{d}x_N, \mathrm{d}y_N),$$

  for any $1 \le n \le N$ and π-almost surely $x_{0:n-1}$, $y_{0:n-1}$ we have

$$\pi_{x_{0:n-1},y_{0:n-1}}(\mathrm{d}x_n, \mathrm{d}y_n) \in \Pi(\mu_{x_{0:n-1}}(\mathrm{d}x_n), *),$$

  and $\pi_0(\mathrm{d}x_0, \mathrm{d}y_0) \in \Pi(\mu_0(\mathrm{d}x_0), *)$.

Theorem 4.24. *Under Assumption 4.18, we have a dynamic duality formula*

$$V_{\mathfrak{c}} = \sup_{\nu \in \mathscr{P}(\mathcal{X})} \{E_\nu[f(X)] - L(\mathcal{T}_{\mathfrak{c}}(\mu, \nu))\} = \inf_{\lambda \ge 0}\{L^*(\lambda) + U(\lambda)\}, \tag{4.3}$$

where U is given by

$$U(\lambda) = \mu_0(\sup_{y_0 \in \mathcal{X}_0} \{\cdots \mu_{x_{0:N-1}}(\sup_{y_N \in \mathcal{X}_N} \{f(y) - \lambda c(x, y)\}) \cdots \}). \tag{4.4}$$

Remark 4.25. If $N = 0$, we retrieve the classical Wasserstein DRO duality results (Blanchet and Murthy, 2019, Bartl et al., 2020, Zhang et al., 2024) in this static setting.

Proof. We first show that the following quantities are well-defined:

$$U_N(x_{0:N-1}, y_{0:N-1}) := \sup_{\pi_N \in \Pi(\mu_{x_{0:N-1}}, *)} E_{\pi_N}[f(y_{0:N-1}, Y_N) - \lambda c(x_{0:N-1}, X_N, y_{0:N-1}, Y_N)]$$

and for $1 \le n \le N-1$

$$U_n(x_{0:n-1}, y_{0:n-1}) := \sup_{\pi_n \in \Pi(\mu_{x_{0:n-1}}, *)} E_{\pi_n}[U_{n+1}(x_{0:n-1}, X_n, y_{0:n-1}, Y_n)].$$

We claim U_N is upper semi-analytic. It follows from Bertsekas and Shreve (1996, Proposition 7.48) that

$$D = \{(x_{0:N-1}, y_{0:N-1}, \pi) : (x, y) \in \mathcal{X} \times \mathcal{X},\ \pi \in \Pi(\mu_{x_{0:N-1}}, *)\}$$

is Borel, and

$$(x_{0:N-1}, y_{0:N-1}, \pi) \mapsto E_\pi[f(Y) - \lambda c(X, Y)]$$

is upper semi-analytic. By Bertsekas and Shreve (1996, Proposition 7.47), we obtain U_N is upper semi-analytic. Therefore, U_N is universally measurable, and this implies U_{N-1} is well-defined. Recursively, we can show U_n is again upper semi-analytic for any $1 \le n \le N$. By Proposition 4.23, we decompose the optimization problem $\sup_{\pi \in \Pi_{\mathfrak{c}}(\mu,*)} E_\pi[f(Y) - \lambda c(X,Y)]$ into single step problems as

$$\begin{aligned} U(\lambda) &= \sup_{\pi \in \Pi_{\mathfrak{c}}(\mu,*)} E_\pi[f(Y) - \lambda c(X,Y)] \\ &= \sup_{\pi_0 \in \Pi(\mu_0,*)} E_{\pi_0}\Big[\cdots \sup_{\pi_N \in \Pi(\mu_{x_{1:N-1}},*)} E_{\pi_N}[f(Y) - \lambda c(X,Y)] \cdots\Big]. \end{aligned}$$

As we have shown U_n is upper semi-analytic for any $1 \le n \le N$, applying Lemma 4.5 we derive (4.4). $\square$

Our next result answers the question when the bi-causal and the causal ambiguity sets are equivalent in terms of their corresponding distributional risks.

Assumption 4.26. We assume there exists $p \ge 1$ such that the following conditions hold:

(i) $L : \mathbb{R}^* \to \mathbb{R}^*$ is non-decreasing and continuous on its domain $\{L < \infty\}$.

(ii) $c : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is continuous and has a polynomial growth

$$|c(x,y)| \le C(1 + d_{\mathcal{X}}(\tilde{x},x)^p + d_{\mathcal{X}}(\tilde{x},y)^p) \quad \text{for some } \tilde{x} \in \mathcal{X}.$$

(iii) $f : \mathcal{X} \to \mathbb{R}$ is continuous and has a polynomial growth

$$|f(x)| \le C(1 + d_{\mathcal{X}}(\tilde{x},x)^p) \quad \text{for some } \tilde{x} \in \mathcal{X}.$$

Theorem 4.27. *We assume that $\mathcal{X}$ has no isolated points. Under Assumption 4.26, we have $V_{\mathfrak{c}} = V_{\mathfrak{bc}}$, i.e.,*

$$\sup_{\nu \in \mathscr{P}(\mathcal{X})} \{E_\nu[f(X)] - L(\mathcal{T}_{\mathfrak{c}}(\mu,\nu))\} = \sup_{\nu \in \mathscr{P}(\mathcal{X})} \{E_\nu[f(X)] - L(\mathcal{T}_{\mathfrak{bc}}(\mu,\nu))\}.$$

Proof. We lift the problem to the space of adapted stochastic processes introduced in Chapter 2. Recall by AP we denote the space of adapted stochastic processes with paths in $\mathcal{X}$, and by NP we denote the space of naturally filtered stochastic processes. It is clear that we can reformulate the desired identity as

$$\sup_{\mathbb{Y} \in \mathsf{NP}} \{E_Q[f(Y)] - L(\mathcal{T}_{\mathfrak{c}}(\mathbb{X},\mathbb{Y}))\} = \sup_{\mathbb{Y} \in \mathsf{NP}} \{E_Q[f(Y)] - L(\mathcal{T}_{\mathfrak{bc}}(\mathbb{X},\mathbb{Y}))\},$$

where $\mathbb{X} = (\mathcal{X}, \mathcal{F}, \mathbf{F}, \mu, X)$ and $\mathbb{Y} = (\Omega^{\mathbb{Y}}, \mathcal{F}^{\mathbb{Y}}, \mathbf{F}^{\mathbb{Y}}, Q, Y)$. Since we assume $\mathcal{X}$ has no isolated points, by Bartl et al. (2024, Theorem 5.4) NP is a dense subset of AP in $\mathcal{AW}_p$. By Assumption 4.26 (ii) and Eckstein and Pammer (2024, Theorem 3.6), both $\mathbb{Y} \mapsto \mathcal{T}_{\mathfrak{c}}(\mathbb{X}, \mathbb{Y})$ and $\mathbb{Y} \mapsto \mathcal{T}_{\mathfrak{bc}}(\mathbb{X}, \mathbb{Y})$ are continuous with respect to $\mathcal{AW}_p$. Combining these facts with Assumption 4.26 (iii), we derive

$$\sup_{\mathbb{Y} \in \mathsf{NP}} \{E_Q[f(Y)] - L(\mathcal{T}_{\mathfrak{c}}(\mathbb{X}, \mathbb{Y}))\} = \sup_{\mathbb{Y} \in \mathsf{AP}} \{E_Q[f(Y)] - L(\mathcal{T}_{\mathfrak{c}}(\mathbb{X}, \mathbb{Y}))\},$$

and

$$\sup_{\mathbb{Y} \in \mathsf{NP}} \{E_Q[f(Y)] - L(\mathcal{T}_{\mathfrak{bc}}(\mathbb{X}, \mathbb{Y}))\} = \sup_{\mathbb{Y} \in \mathsf{AP}} \{E_Q[f(Y)] - L(\mathcal{T}_{\mathfrak{bc}}(\mathbb{X}, \mathbb{Y}))\}.$$

For any $\mathbb{Y} \in \mathsf{AP}$, we construct $\widetilde{\mathbb{Y}} = (\Omega^{\mathbb{Y}}, \mathcal{F}^{\mathbb{Y}}, \widetilde{\mathbf{F}}^{\mathbb{Y}}, Q, Y) \in \mathsf{AP}$ where $\widetilde{\mathbf{F}}^{\mathbb{Y}} = \{\mathcal{F}^{\mathbb{Y}}\}_{t \in I}$ is the constant filtration with the richest σ-algebra. This yields the inclusion $\Pi_{\mathfrak{c}}(\mathbb{X}, \mathbb{Y}) \subseteq \Pi_{\mathfrak{bc}}(\mathbb{X}, \widetilde{\mathbb{Y}})$ as any coupling is causal form $\widetilde{\mathbb{Y}}$ to $\mathbb{X}$. Hence, we deduce $\mathcal{T}_{\mathfrak{c}}(\mathbb{X}, \mathbb{Y}) \geq \mathcal{T}_{\mathfrak{bc}}(\mathbb{X}, \widetilde{\mathbb{Y}})$, and together with Assumption 4.26 (i) we show

$$\begin{aligned}
\sup_{\mathbb{Y} \in \mathsf{AP}} \{E_Q[f(Y)] - L(\mathcal{T}_{\mathfrak{c}}(\mathbb{X}, \mathbb{Y}))\} &= \sup_{\mathbb{Y} \in \mathsf{NP}} \{E_Q[f(Y)] - L(\mathcal{T}_{\mathfrak{c}}(\mathbb{X}, \mathbb{Y}))\} \\
&\leq \sup_{\mathbb{Y} \in \mathsf{NP}} \{E_Q[f(Y)] - L(\mathcal{T}_{\mathfrak{bc}}(\mathbb{X}, \widetilde{\mathbb{Y}}))\} \\
&\leq \sup_{\mathbb{Y} \in \mathsf{AP}} \{E_Q[f(Y)] - L(\mathcal{T}_{\mathfrak{bc}}(\mathbb{X}, \mathbb{Y}))\}.
\end{aligned}$$

The reverse direction is trivial, and we conclude the proof. □

We remark that this result generalizes the arguments in Bartl and Wiesel (2023, Lemma 3.1) to abstract Polish spaces.

4.5 Continuous-time results

We set up the continuous time problem as follows. Let $(\Omega, \mathcal{F}, P)$ be a Polish probability space supporting a standard n-dimensional Brownian motion B. We consider a path-dependent SDE given by

$$\alpha(t) = \int_0^t b(s, \alpha(\cdot \wedge s))\,\mathrm{d}s + \int_0^t \sigma(s, \alpha(\cdot \wedge s))\,\mathrm{d}B(s). \tag{4.5}$$

Assumption 4.28. We assume SDE (4.5) has a unique strong solution α, and the law of $\alpha(t)$ is non-atomic for any $0 < t \leq T$.

The strong well-posedness holds, for example in Bally et al. (2016, Theorem 8.1), if the coefficients b and σ are Lipschitz continuous and have linear growth. In a

Markovian setting, the celebrated Hörmander theorem gives a sufficient condition when the law of $\alpha(t)$ is non-atomic. More recently, it has been extended to a path-dependent SDE setting in Ohashi et al. (2021).

We take the reference model as $\mu = \alpha_{\#}P$ and are interested in the case where $c(x, y) = \|x - y\|^2_{H_0^1}$ as the square of the Cameron–Martin norm on the path space $\mathcal{X} = C_0([0, T]; \mathbb{R}^n)$.

Assumption 4.29. We assume the following conditions:

(i) $L : \mathbb{R}^* \to \mathbb{R}^*$ is a non-decreasing closed proper convex function with $L(0) = 0$ and $L(+\infty) = +\infty$.

(ii) $f : \mathcal{X} \to \mathbb{R}$ is continuous and has a quadratic growth $|f(x)| \leq C(1 + \|x\|^2_\infty)$.

Theorem 4.30. *Assume the semi-linear path-dependent PDE*

$$\begin{aligned}&\mathcal{D}U(t, x, y; \lambda) + b(t, x)^\intercal(\nabla_x + \nabla_y)U(t, x, y; \lambda)\\&\quad + \frac{1}{2}\operatorname{tr}(\sigma(t, x)^\intercal\sigma(t, x)(\nabla^2_{xx} + \nabla^2_{yy} + 2\nabla^2_{xy})U(t, x, y; \lambda)) + \frac{1}{4\lambda}\|\nabla_y U(t, x, y; \lambda)\|^2 = 0\end{aligned} \tag{4.6}$$

with boundary condition $U(T, x, y; \lambda) = f(y)$ has a non-anticipative $C_b^{1,2}$ solution U such that $\nabla_y U(t, x, y)$ is Lipschitz with a linear growth. Then under Assumptions 4.28 and 4.29 we have

$$V_{\mathfrak{c}} = \sup_{\nu \in \mathscr{P}(\mathcal{X})} \{E_\nu[f(Y)] - L(\mathcal{T}_{\mathfrak{c}}(\mu, \nu))\} = \inf_{\lambda \geq 0}\{L^*(\lambda) + U(\lambda)\},$$

where

$$U(\lambda) := U(0, 0, 0; \lambda) = \sup_{\pi \in \Pi_{\mathfrak{c}}(\mu, *)} E_\pi[f(Y) - \lambda c(X, Y)].$$

Remark 4.31. When the reference model μ is the Wiener measure γ, the above path-dependent PDE (4.6) reads as

$$\mathcal{D}U(t, x, y; \lambda) + \frac{1}{2}\operatorname{tr}((\nabla^2_{xx} + \nabla^2_{yy} + 2\nabla_{xy})U(t, x, y; \lambda)) + \frac{1}{4\lambda}\|\nabla_y U(t, x, y; \lambda)\|^2 = 0.$$

The change of the variable $U = 2\lambda\log(W)$ turns the above equation to a linear PDE and yields a unique classical solution given explicitly by

$$U(t, x, y; \lambda) = 2\lambda\log\left(\int_{C([0,T];\mathbb{R}^n)} \exp\left(\frac{f(z)}{2\lambda}\right)\gamma_{y(\cdot\wedge t)}(\mathrm{d}z)\right).$$

The following result is adapted from Beiglböck and Lacker (2020, Theorem 1.2) which shows that the set of Monge causal couplings with a first marginal μ are dense in $\Pi_{\mathfrak{c}}(\mu, *)$. We postpone the proof to the end of the current section.

Lemma 4.32. *Let $\Gamma_{\mathfrak{c}}(\mu,*)$ denote the set of Monge causal couplings with a fixed first marginal μ. Under Assumption 4.28, we have*

$$\sup_{\pi\in\Pi_{\mathfrak{c}}(\mu,*)} E_\pi[f(Y)-\lambda c(X,Y)] = \sup_{\pi\in\Gamma_{\mathfrak{c}}(\mu,*)} E_\pi[f(Y)-\lambda c(X,Y)].$$

Proof of Theorem 4.30. Step 1. We first show the solution to (4.6) coincides with the value function of a non-Markovian stochastic optimal control problem. For brevity of the notation, we fix $\lambda>0$ and omit the λ argument in U. For two paths $\omega,\eta\in C_0([0,T];\mathbb{R}^n)$ with $\omega(s)=\eta(s)$ we introduce their concatenation at time s as

$$\omega\otimes_s\eta(t) := \begin{cases}\omega(t), & \text{if } t\le s,\\ \eta(t), & \text{if } t>s.\end{cases}$$

Let $(\alpha^{s,x},\beta^{s,x,y,u})$ be a controlled system given by

$$\begin{cases}\alpha^{s,x}(\cdot) = x(s) + \displaystyle\int_s^\cdot b(t,x\otimes_s\alpha^{s,x})\,\mathrm{d}t + \int_s^\cdot \sigma(t,x\otimes_s\alpha^{s,x})\,\mathrm{d}B(t),\\ \beta^{s,x,y,u}(\cdot) = y(s) + \displaystyle\int_s^\cdot [b(t,x\otimes_s\alpha^{s,x})+u(t)]\,\mathrm{d}t + \int_s^\cdot \sigma(t,x\otimes_s\alpha^{s,x})\,\mathrm{d}B(t),\end{cases}$$

with the aim of maximizing the objective

$$J(s,x,y,u) = E_P\left[f(y\otimes_s\beta^{s,x,y,u}) - \lambda\int_s^T \|u(t)\|^2\,\mathrm{d}t\right]$$

over $\mathbb{U}([s,T])$ the set of $\mathbf{F}^B$-progressively measurable processes. The *Hamiltonian* associated to the control problem $\mathcal{H}:[0,T]\times C([0,T];\mathbb{R}^{2n})\times(\mathbb{R}^n)^2\times(\mathbb{R}^{n\times n})^3$ is given by

$$\begin{aligned}&\mathcal{H}(t,x,y,\rho_x,\rho_y,A_{xx},A_{yy},A_{xy})\\ &=\sup_{u\in\mathbb{R}^n}\left\{b(t,x)^\intercal(\rho_x+\rho_y)+\frac{1}{2}\operatorname{tr}(\sigma(t,x)^\intercal\sigma(t,x)(A_{xx}+A_{yy}+2A_{xy}))+u^\intercal\rho_y-\lambda\|u\|^2\right\}\\ &=b(t,x)^\intercal(\rho_x+\rho_y)+\frac{1}{2}\operatorname{tr}(\sigma(t,x)^\intercal\sigma(t,x)(A_{xx}+A_{yy}+2A_{xy}))+\frac{\|\rho_y\|^2}{4\lambda}.\end{aligned}$$

In particular, the above supremum is attained when $u=\frac{1}{2\lambda}\rho_y$. Since we assume U is a $C_b^{1,2}$ solution, by Bally et al. (2016, Theorem 8.16), it is clear that

$$U(s,x,y)\ge\sup_{u\in\mathbb{U}([s,T])} J(s,x,y,u).$$

On the other hand, as we assume $\nabla_y U$ is Lipschitz, there exists a unique solution to

$$\begin{cases}\alpha^{s,x}(\cdot) = x(s)+\displaystyle\int_s^\cdot b(t,x\otimes_s\alpha^{s,x})\,\mathrm{d}t+\int_s^\cdot\sigma(t,x\otimes_s\alpha^{s,x})\,\mathrm{d}B(t),\\ \beta_*^{s,x,y}(\cdot) = y(s)+\displaystyle\int_s^\cdot\left[b(t,x\otimes_s\alpha^{s,x})+\frac{1}{2\lambda}\nabla_y U(t,x\otimes_s\alpha^{s,x},y\otimes_s\beta_*^{s,x,y})\right]\mathrm{d}t\\ \qquad\qquad\qquad\qquad+\displaystyle\int_s^\cdot\sigma(t,x\otimes_s\alpha^{s,x})\,\mathrm{d}B(t).\end{cases}\tag{4.7}$$

This yields that $u_*(t) = \frac{1}{2\lambda}\nabla_y U(t, x \otimes_s \alpha^{s,x}, y \otimes_s \beta_*^{s,x,y}) \in \mathbb{U}([s,T])$ is an optimal control since $u = \frac{1}{2\lambda}\rho_y$ attained the supremum in the Hamiltonian $\mathcal{H}$. Therefore, we derive

$$U(s,x,y) = J(s,x,y,u_*) = \sup_{u \in \mathbb{U}([s,T])} J(s,x,y,u).$$

Step 2. We show $U(0,0,0;\lambda) = \sup_{\pi \in \Pi_{\mathfrak{c}}(\mu,*)} E_\pi[f(Y) - \lambda c(X,Y)]$. We notice that $(B,\alpha)_\# P$ is a bi-causal coupling by Cont and Lim (2024, Theorem 3.2), and in particular, $(\alpha, B)_\# P$ is causal. Moreover, $(B,u)_\# P$ is a causal coupling for any $u \in \mathbb{U}([0,T])$ since u is $\mathbf{F}^B$-progressively measurable. Hence, by the gluing lemma 2.19 we derive that $(\alpha, u)_\# P$ is a causal coupling which further implies $\pi^u := (\alpha, \beta^u)_\# P$ is a causal coupling as $\mathrm{d}\beta^u = \mathrm{d}\alpha + \mathrm{d}u$. This gives us

$$\begin{aligned} U(0,0,0;\lambda) = E_P\left[f(\beta^{u_*}) - \lambda \int_0^T \|u^*(t)\|^2 \,\mathrm{d}t\right] &= E_{\pi^{u_*}}[f(Y) - \lambda c(X,Y)] \\ &\leq \sup_{\pi \in \Pi_{\mathfrak{c}}(\mu,*)} E_\pi[f(Y) - \lambda c(X,Y)]. \end{aligned}$$

For the reverse direction, let $\widehat{\mathbb{U}}([0,T];\mathbb{R}^n)$ be the set of $\mathbf{F}^\alpha$-progressively measurable processes. Recall $\Gamma_{\mathfrak{c}}(\mu,*)$ is the set of causal Monge couplings with a fixed first marginal μ. For any $\pi \in \Gamma_{\mathfrak{c}}(\mu,*)$ with $E_\pi[c(X,Y)] < \infty$, there exists a non-anticipative map $\Phi : [0,T] \times C([0,T];\mathbb{R}^n) \to \mathbb{R}^n$ such that $Y(t) = X(t) + \int_0^t \Phi(s,X)\,\mathrm{d}s$ holds π-a.s. This induces a control $u^\pi(t) := \Phi(t,\alpha) \in \widehat{\mathbb{U}}([0,T])$. Therefore, together with Lemma 4.32 we have

$$\begin{aligned} U(0,0,0;\lambda) \geq \sup_{u \in \widehat{\mathbb{U}}([0,T];\mathbb{R}^n)} J(0,0,0,u) &\geq \sup_{\pi \in \Gamma_{\mathfrak{c}}(\mu,*)} E_\pi[f(Y) - \lambda c(X,Y)] \\ &= \sup_{\pi \in \Pi_{\mathfrak{c}}(\mu,*)} E_\pi[f(Y) - \lambda c(X,Y)]. \end{aligned}$$

Notice that Assumption 4.29 implies Assumption 4.18. Applying Proposition 4.19, we conclude the proof.

□

Proposition 4.33. *Under the conditions of Theorem 4.30, if further $U(\lambda)$ is differentiable on its domain, then we have $V_{\mathfrak{c}} = V_{\mathfrak{bc}}$, i.e.,*

$$\sup_{\nu \in \mathscr{P}(\mathcal{X})} \{E_\nu[f(Y)] - L(\mathcal{T}_{\mathfrak{c}}(\mu,\nu))\} = \sup_{\nu \in \mathscr{P}(\mathcal{X})} \{E_\nu[f(Y)] - L(\mathcal{T}_{\mathfrak{bc}}(\mu,\nu))\}.$$

Proof. Step 1. We recall that $\pi_* = (\alpha, \beta_*)_\# P$ attains the supremum in

$$U(\lambda) = \sup_{\pi \in \Pi_{\mathfrak{c}}(\mu,*)} E_\pi[f(Y) - \lambda c(X,Y)]$$

where (α, β_*) is given in (4.7). We claim that π_* is actually a bi-causal coupling. To see this, we can reformulate (4.7) as

$$\alpha(t) = -\int_0^t \frac{1}{2\lambda} \nabla_y U(s, \alpha, \beta_*)\, \mathrm{d}s + \beta_*(t).$$

As we assume $\nabla_y U$ is Lipschitz, there exists a non-anticipative functional F such that $\alpha(t) = F(t, \beta_*)$ holds π_*-a.s. This implies that π_* is a causal from β_* to α, and hence it is a bi-causal coupling.

Step 2. Following Proposition 4.19, there exists a minimizer λ_* of $\inf_{\lambda \geq 0}\{L^*(\lambda) + U(\lambda)\}$. This implies $0 \in \partial L^*(\lambda_*) + \partial U(\lambda_*)$. We notice that

$$\begin{aligned} U(\lambda) - U(\lambda_*) &\geq E_{\pi_*}[f(Y) - \lambda c(X,Y)] - E_{\pi_*}[f(Y) - \lambda_* c(X,Y)] \\ &= -E_{\pi_*}[c(X,Y)](\lambda - \lambda_*). \end{aligned}$$

Together with the assumption that U is differentiable, we obtain $\{-E_{\pi_*}[c(X,Y)]\} = \partial U(\lambda_*)$. Hence, we deduce that $E_{\pi_*}[c(X,Y)] \in \partial L^*(\lambda_*)$. Therefore, by Proposition 4.10 we obtain

$$L^*(\lambda_*) = \lambda_* E_{\pi_*}[c(X,Y)] - L(E_{\pi_*}[c(X,Y)]).$$

It follows from the bi-causality of π_* and Proposition 4.19 that

$$V_{\mathfrak{c}} = E_{\pi_*}[f(Y)] - L(E_{\pi_*}[c(X,Y)]) \leq \sup_{\nu \in \mathscr{P}(\mathcal{X})} \{E_\nu[f(Y)] - L(\mathcal{T}_{\mathfrak{bc}}(\mu, \nu))\} = V_{\mathfrak{bc}}.$$

The reverse direction is trivial as $\Pi_{\mathfrak{bc}}(\mu, *) \subseteq \Pi_{\mathfrak{c}}(\mu, *)$.

□

Proof of Lemma 4.32. When $\lambda = 0$, it is clear that the identity holds with both sides equal to $\sup_{x \in \mathcal{X}} f(x)$. For $\lambda > 0$, we rely on the proof of Beiglböck and Lacker (2020, Theorem 5.2). Let $Z = Y - X$. We fix a $\pi_0 \in \Pi_{\mathfrak{c}}(\mu, *)$ such that $E_{\pi_0}[\|Z\|^2_{H_0^1}] < \infty$. We notice $\tilde{\pi} = (X, Z)_\# \pi_0$ is also in $\Pi_{\mathfrak{c}}(\mu, *)$. By Beiglböck and Lacker (2020, Theorem 1.2), there exist piecewise linear $\mathbf{F}^X$-adapted processes $\{Z_n\}_{n \geq 1}$ such that $(X, Z_n)_\# \pi_0$ converges weakly to $\tilde{\pi}$. In this case, Z_n can be further chosen such as $\lim_{n \to \infty} E_{\pi_0}[\|Z_n\|^2_{H_0^1}] = E_{\pi_0}[\|Z\|^2_{H_0^1}]$. Therefore, $\pi_n = (X, X + Z_n)_\# \pi_0 \in \Gamma_{\mathfrak{c}}(\mu, *)$ and satisfies

$$\sup_{\pi \in \Gamma_{\mathfrak{c}}(\mu, *)} E_\pi[f(Y) - \lambda c(X,Y)] \geq \lim_{n \to \infty} E_{\pi_n}[f(Y) - \lambda c(X,Y)] = E_{\pi_0}[f(Y) - \lambda c(X,Y)].$$

The arbitrary choice of π_0 allows us to derive

$$\sup_{\pi \in \Gamma_{\mathfrak{c}}(\mu, *)} E_\pi[f(Y) - \lambda c(X,Y)] \geq \sup_{\pi \in \Pi_{\mathfrak{c}}(\mu, *)} E_\pi[f(Y) - \lambda c(X,Y)].$$

The reverse direction is trivial as $\Gamma_{\mathfrak{c}}(\mu, *) \subseteq \Pi_{\mathfrak{c}}(\mu, *)$. □

4.6 Distributionally robust expected shortfall

In this section, we consider the extension of our duality formula from a linear expectation to a convex risk measure in a discrete-time setting. Let $f : \mathcal{X} \to \mathbb{R}$ be a path-dependent payoff of a contingent claim and $\alpha \in (0,1)$ a risk aversion parameter. The expected shortfall ES_α is a law-invariant convex risk measure given by

$$\mathrm{ES}_\alpha(f;\mu) := \sup_{\eta \ll \mu,\, \frac{\mathrm{d}\eta}{\mathrm{d}\mu} \le \alpha^{-1}} E_\eta[f(X)].$$

We are interested in its performance under a model misspecification, and write its adapted Wasserstein distributionally robust counterpart $\mathbf{ES}_\alpha$ as

$$\mathbf{ES}_\alpha(f;\mu) := \sup_{\nu \in \mathscr{P}(\mathcal{X})} \{\mathrm{ES}_\alpha(f;\nu) - L(\mathcal{T}_{\mathfrak{c}}(\mu,\nu))\}.$$

Assumption 4.34. We assume the following conditions:

(i) $L : \mathbb{R}^* \to \mathbb{R}^*$ is a non-decreasing closed proper convex function with $L(0) = 0$ and $L(+\infty) = +\infty$.

(ii) $c : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^*$ is lower semi-continuous and non-negative with $c(x,y) = 0$ if and only if $x = y$ for any $x, y \in \mathcal{X}$.

(iii) $f : \mathcal{X} \to \mathbb{R}$ is non-negative and upper semi-continuous.

We derive the following duality representation of $\mathbf{ES}_\alpha(f)$.

Theorem 4.35. *Under Assumption 4.34, we have*

$$\mathbf{ES}_\alpha(f;\mu) = \inf_{\lambda,\gamma \ge 0} \{L^*(\lambda) + \gamma + U(\lambda,\gamma)\},$$

where $U(\lambda,\gamma) = \sup_{\pi \in \Pi_{\mathfrak{c}}(\mu,*)} E_\pi[\alpha^{-1}(f(Y) - \gamma)^+ - \lambda c(X,Y)]$ *can be computed by*

$$U(\lambda,\gamma) = \mu_0\Big(\sup_{y_0 \in \mathcal{X}_0} \{\cdots \mu_{x_{0:N-1}}\big(\sup_{y_N \in \mathcal{X}_N} \{\alpha^{-1}(f(y) - \gamma)^+ - \lambda c(x,y)\}\big) \cdots\}\Big).$$

Proof. It is shown in Föllmer and Schied (2008, Theorem 4.39) that the expected shortfall has a dual representation given by

$$\mathrm{ES}_\alpha(f;\mu) = \inf_{\gamma \ge 0} \{\gamma + \alpha^{-1} E_\mu[(f(X) - \gamma)^+]\}.$$

In particular, this implies $\mu \mapsto \mathrm{ES}_\alpha(f;\mu)$ is a concave functional. By Remark 4.20, we derive

$$\begin{aligned}\mathbf{ES}_\alpha(f;\mu) &= \inf_{\lambda\geq 0}\{L^*\lambda + \sup_{\nu\in\mathscr{P}(\mathcal{X})}\{\mathrm{ES}_\alpha(f;\mu) - \mathcal{T}_{\mathfrak{c}}(\mu,\nu)\}\} \\ &= \inf_{\lambda\geq 0}\{L^*(\lambda) + \sup_{\pi\in\Pi_{\mathfrak{c}}(\mu,*)}\inf_{\gamma\geq 0}\{\gamma + E_\pi[\alpha^{-1}(f(Y)-\gamma)^+ - \lambda c(X,Y)]\}\}.\end{aligned}$$

Now it suffices to show that we can swap the inner sup and inf and apply the same arguments as Theorem 4.24. We fix $\lambda \geq 0$ and consider the first case when

$$\sup_{\pi\in\Pi_{\mathfrak{c}}(\mu,*)} E_\pi[\alpha^{-1}f(Y) - \lambda c(X,Y)] = +\infty. \tag{4.8}$$

The inequality

$$\gamma + E_\pi[\alpha^{-1}(f(Y)-\gamma)^+ - \lambda c(X,Y)] \geq \gamma - \alpha^{-1}\gamma + E_\pi[\alpha^{-1}f(Y) - \lambda c(X,Y)]$$

yields

$$\inf_{\gamma\geq 0}\sup_{\pi\in\Pi_{\mathfrak{c}}(\mu,*)}\{\gamma + E_\pi[\alpha^{-1}(f(Y)-\gamma)^+ - \lambda c(X,Y)]\} = +\infty.$$

On the other hand, we take a sequence of $\pi_n \in \Pi_{\mathfrak{c}}(\mu,*)$ such that $E_{\pi_n}[(\alpha^{-1}f(Y) - \lambda c(X,Y))]$ goes to infinity. We write $\pi_n(\mathrm{d}x,\mathrm{d}y) = \mu(\mathrm{d}x)\eta_n(x,\mathrm{d}y)$ and define

$$\tilde{\pi}_n(\mathrm{d}x,\mathrm{d}y) = \mu(\mathrm{d}x)[\alpha\eta_n(x,\mathrm{d}y) + (1-\alpha)\delta_x(\mathrm{d}y)].$$

It is direct to verify that $\tilde{\pi}_n$ is again causal as it is a linear combination of two causal couplings. We further notice

$$\begin{aligned}&\inf_{\gamma\geq 0}\{\gamma + E_{\tilde{\pi}_n}[\alpha^{-1}(f(Y)-\gamma)^+ - \lambda c(X,Y)]\} \\ &= \inf_{\gamma\geq 0}\{\gamma + \alpha^{-1}E_{\tilde{\pi}_n}[(f(Y)-\gamma)^+]\} - \lambda\alpha E_{\pi_n}[c(X,Y)] \\ &\geq \sup_{0\leq\xi\leq\alpha^{-1}, E_{\tilde{\pi}_n}[\xi]\leq 1} E_{\tilde{\pi}_n}[f(Y)\xi] - \lambda\alpha E_{\pi_n}[c(X,Y)] \\ &\geq \alpha E_{\pi_n}[\alpha^{-1}f(Y) - \lambda c(X,Y)].\end{aligned}$$

The second equality follows from the duality representation of ES_α; the last inequality follows by taking $\xi = \frac{\mathrm{d}\pi_n}{\mathrm{d}\tilde{\pi}_n} \in [0,\alpha^{-1}]$. Therefore, we deduce the duality holds with

$$\sup_{\pi\in\Pi_{\mathfrak{c}}(\mu,*)}\inf_{\gamma\geq 0}\{\gamma + E_\pi[\alpha^{-1}(f(Y)-\gamma)^+ - \lambda c(X,Y)]\} = +\infty.$$

The second case is to consider

$$\sup_{\pi\in\Pi_{\mathfrak{c}}(\mu,*)} E_\pi[\alpha^{-1}f(Y) - \lambda c(X,Y)] < \infty. \tag{4.9}$$

We write

$$I := \inf_{\gamma \geq 0} \sup_{\pi \in \Pi_{\mathfrak{c}}(\mu,*)} \{\gamma + E_\pi[\alpha^{-1}(f(Y) - \gamma)^+ - \lambda c(X,Y)]\}.$$

Since $\gamma_{\max} := \sup_{\pi \in \Pi_{\mathfrak{c}}(\mu,*)} E_\pi[\alpha^{-1} f(Y) - \lambda c(X,Y)] < \infty$, it follows from Fan (1953, Theorem 2) that

$$\begin{aligned} I &= \inf_{\gamma \in [0,\gamma_{\max}]} \sup_{\pi \in \Pi_{\mathfrak{c}}(\mu,*)} \{\gamma + E_\pi[\alpha^{-1}(f(Y) - \gamma)^+ - \lambda c(X,Y)]\} \\ &= \sup_{\pi \in \Pi_{\mathfrak{c}}(\mu,*)} \inf_{\gamma \in [0,\gamma_{\max}]} \{\gamma + E_\pi[\alpha^{-1}(f(Y) - \gamma)^+] - \lambda c(X,Y)\}. \end{aligned}$$

Let $\pi_\delta \in \Pi_{\mathfrak{c}}(\mu,*)$ such that

$$\inf_{\gamma \in [0,\gamma_{\max}]} \{\gamma + E_{\pi_\delta}[\alpha^{-1}(f(Y) - \gamma)^+ - \lambda c(X,Y)]\} \geq I - \delta.$$

Since the support of any probability measure on a Polish space is σ-compact, we may assume $\mathcal{X}_n$ is σ-compact for simplicity. In particular, we can write $\mathcal{X}_n = \lim_{m\to\infty} \mathcal{K}_n^m$, where $\{\mathcal{K}_n^m\}_{m\geq 1}$ is ascending and compact. We write $\mathcal{K}^m = \mathcal{K}_0^m \times \mathcal{K}_1^m \times \cdots \times \mathcal{K}_N^m$ and $\Pi^m := \{\pi \in \Pi_{\mathfrak{c}}(\mu,*) : \mathrm{supp}(\pi) \subseteq \mathcal{X} \times \mathcal{K}^m\}$. Without loss of generality, we assume

$$E_{\pi_\delta}[\mathbb{1}_{\mathcal{X}\times(\mathcal{K}^m)^c}] + E_{\pi_\delta}[\mathbb{1}_{\mathcal{X}\times(\mathcal{K}^m)^c}(\alpha^{-1} f(Y) - \lambda c(X,Y))] \leq \frac{1}{m}.$$

We fix $\tilde{x} \in \mathcal{K}^1$ and define $\tilde{\pi} = [x \mapsto (x,\tilde{x})]_{\#}\mu$. We construct a causal coupling π^m as

$$\pi^m(\cdot) = \pi_\delta(\cdot \cap (\mathcal{X} \times \mathcal{K}^m)) + (1 - \pi_\delta(\mathcal{X} \times \mathcal{K}^m)\tilde{\pi}(\cdot) \in \Pi^m$$

— it follows from the fact that $\pi_\delta(\cdot \cap (\mathcal{X} \times \mathcal{K}^m))/\pi_\delta(\mathcal{X} \times \mathcal{K}^m)$ is causal. Therefore, we derive

$$\begin{aligned} &\sup_{\pi \in \Pi^m} \inf_{\gamma \in [0,\gamma_{\max}]} \{\gamma + E_\pi[\alpha^{-1}(f(Y) - \gamma)^+ - \lambda c(X,Y)]\} \\ \geq &\inf_{\gamma \in [0,\gamma_{\max}]} \{\gamma + E_{\pi^m}[\alpha^{-1}(f(Y) - \gamma)^+ - \lambda c(X,Y)]\} \\ \geq &\inf_{\gamma \in [0,\gamma_{\max}]} \{\gamma + E_{\pi_\delta}[\alpha^{-1}(f(Y) - \gamma)^+ - \lambda c(X,Y)] - \frac{1}{m} - \frac{1}{m} E_\mu[\lambda c(X,x_0)]\} \\ \geq &\, I - \delta - \frac{1}{m}(1 + E_\mu[\lambda c(X,x_0)]). \end{aligned}$$

Here, the second last inequality follows from the estimate (4.6). On the other hand, Π^m is compact under the weak topology, see Lassalle (2018, Theorem 3 (i)). Therefore, we apply the minimax theorem from Fan (1953) and derive

$$
\begin{aligned}
&\sup_{\pi\in\Pi_{\mathfrak{c}}(\mu,*)}\inf_{\gamma\geq 0}\{\gamma+E_\pi[\alpha^{-1}(f(Y)-\gamma)^+-\lambda c(X,Y)]\}\\
\geq &\lim_{m\to\infty}\sup_{\pi\in\Pi^m}\inf_{\gamma\geq 0}\{\gamma+E_\pi[\alpha^{-1}(f(Y)-\gamma)^+-\lambda c(X,Y)]\}\\
= &\lim_{m\to\infty}\inf_{\gamma\geq 0}\sup_{\pi\in\Pi^m}\{\gamma+E_\pi[\alpha^{-1}(f(Y)-\gamma)^+-\lambda c(X,Y)]\}\\
= &\lim_{m\to\infty}\inf_{\gamma\in[0,\gamma_{\max}]}\sup_{\pi\in\Pi^m}\{\gamma+E_\pi[\alpha^{-1}(f(Y)-\gamma)^+-\lambda c(X,Y)]\}\\
= &\lim_{m\to\infty}\sup_{\pi\in\Pi^m}\inf_{\gamma\in[0,\gamma_{\max}]}\{\gamma+E_\pi[\alpha^{-1}(f(Y)-\gamma)^+-\lambda c(X,Y)]\}\\
\geq &I-\delta.
\end{aligned}
$$

The arbitrary choice of δ allows us to conclude the proof. □

Example 4.36. We consider a two-step exotic option. Let $X=(X_0,X_1,X_2)$ be the underlying asset, $f(x)=(x_2-x_1+1-K)^+$ the payoff of the option. We set interest rates and dividends to zero for simplicity. Let μ be a reference pricing measure of the underlying. In this example we assume μ is the finite marginal of a geometric Brownian motion S given by

$$\mathrm{d}S_t=\sigma S_t\,\mathrm{d}W_t,\quad S_0=1.$$

We set $\alpha=0.95$, $\sigma=0.2$, $(X_0,X_1,X_2)\sim(S_0,S_{0.5},S_1)$, $c(x,y)=|x-y|^2$, and $L=+\infty\mathbb{1}_{(0.3^2,+\infty]}$. In Figure 4.1, we plot the expected shortfall of the exotic option under different strikes. The classical expected shortfall ES_α, the AW-DR expected shortfall $\mathbf{ES}_\alpha$, the W-DR expected shortfall are in solid, dashed, and dotted respectively. The gap between the solid and the dashed lines corresponds to the extra risk coming from the model uncertainty.

We emphasize that, in certain cases, restricting to non-anticipative model uncertainty does not lead to a reduction in risk. For example, consider the calendar spread payoff $f(x)=(x_2-K)^+-(x_1-K)^+$ with any separable cost c.

Remark 4.37. We can extend the risk-indifference pricing (Xu, 2006) to a model misspecification context. For simplicity, we assume zero initial capital and liability. The risk-indifference (sell) price is the minimum price a trader will charge so that the total risk of their portfolio will not increase. Here, we take the risk measure as the distributionally robust expected shortfall to counter the model uncertainty. The corresponding distributionally robust risk-indifference price is given by

$$\inf_{H\in\mathbb{H}}\mathbf{ES}_\alpha(f+(H\circ X)_N;\mu)-\inf_{H\in\mathbb{H}}\mathbf{ES}_\alpha((H\circ X)_N;\mu),$$

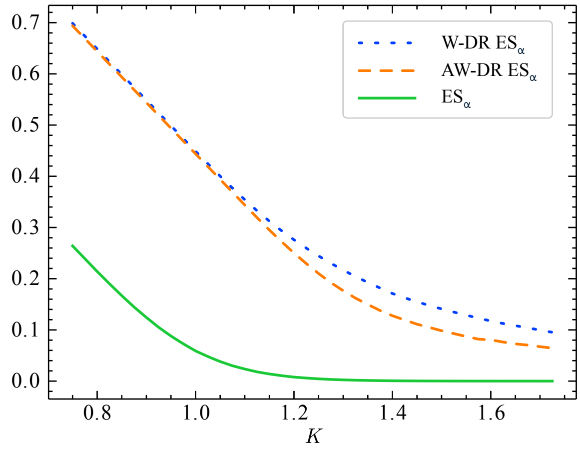

Figure 4.1: Comparison of W-DR expected shortfall (dotted), AW-DR expected shortfall (dashed), and standard expected shortfall (solid) for an exotic option.

where $\mathbb{H}$ is the set of all predictable hedging strategies and $(H \circ X)_N = \sum_{n=1}^N H_n(X_n - X_{n-1})$ is the discrete stochastic integral. By Theorem 4.35, we write

$$\inf_{H \in \mathbb{H}} \mathbf{ES}_\alpha(f + (H \circ X)_N; \mu) = \inf_{\lambda, \gamma \geq 0} \{L^*(\lambda) + \gamma + \inf_{H \in \mathbb{H}} U(\lambda, \gamma, H)\},$$

where

$$U(\lambda, \gamma) = \mu_0(\sup_{y_0 \in \mathcal{X}_0} \{ \cdots \mu_{x_{0:N-1}}(\sup_{y_N \in \mathcal{X}_N} \{\alpha^{-1}(f(y) + \sum_{n=1}^N H_n(y_n - y_{n-1}) - \gamma)^+ - \lambda c(x, y)\}) \cdots \}).$$

As α goes to 0, it is known that the risk-indifference price under ES_α converges to the superhedging price. Under the current context, the robust superhedging price is given by

$$\rho(f) := \inf\{x : x + (H \circ X)_N \geq f - L(\mathcal{T}_\mathfrak{c}(\mu, \nu)) \quad \text{for some } H \in \mathbb{H} \quad \nu\text{-a.s.}\}.$$

We stress that this is a nondominated framework and is different from the setting in Bouchard and Nutz (2015). Indeed, we consider all possible measures $\nu \in \mathscr{P}(\mathcal{X})$ with a penalization $L(\mathcal{T}_\mathfrak{c}(\mu, \nu))$, whereas in Bouchard and Nutz (2015) the authors considered a collection of measures which is stable under the concatenation of kernels. If $\mathcal{X}$ is compact and f is continuous, one can write down a pricing–hedging duality as

$$\rho(f) = \sup_{\nu \in \mathscr{P}(\mathcal{X})} \sup_{\eta \ll \nu, \eta \in \mathscr{M}(\mathcal{X})} \{E_\eta[f(X)] - L(\mathcal{T}_\mathfrak{c}(\mu, \nu))\},$$

where $\mathscr{M}(\mathcal{X})$ denotes the set of martingale measures on $\mathcal{X}$. However, it is unclear if this holds under a more general setting, and we leave it for future work.

4.7 Extension to optimal stopping problems

In this section, our aim is to derive a duality formula for AW-DR optimal stopping problems. We stick to a discrete-time setup, and let $f_n : \mathcal{X}_{0:n} \to \mathbb{R}$ for $n \in I$ denote the payoff if the process stops at time n. We introduce the optimal stopping problem as

$$\mathrm{OS}(f; \mu) := \sup_{\tau \in \mathscr{T}} E_\mu[f_\tau(X_{0:\tau})],$$

where $\mathscr{T}$ is the set of $\mathbf{F}$-stopping times. The corresponding adapted Wasserstein distributionally robust counterpart is given by

$$\mathbf{OS}(f; \mu) := \sup_{\nu \in \mathscr{P}(\mathcal{X})} \{\sup_{\tau \in \mathscr{T}} E_\nu[f_\tau(X_{0:\tau})] - L(\mathcal{T}_{\mathfrak{bc}}(\mu, \nu))\}. \tag{4.10}$$

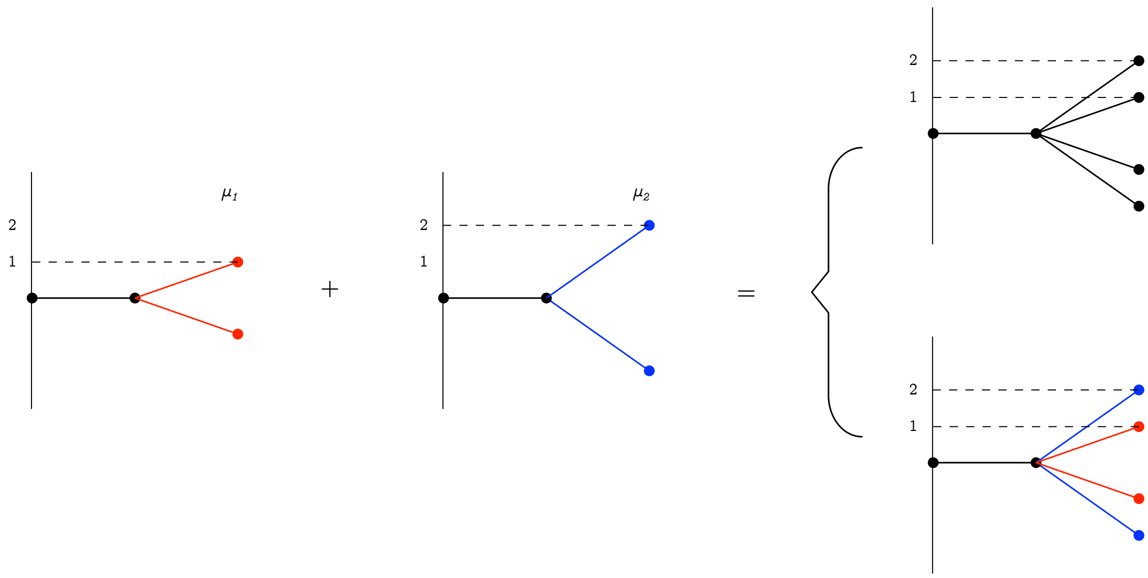

Figure 4.2: Different averages of adapted stochastic processes $\mathbb{X}_1 = (\mathcal{X}, \mathcal{F}, \mathbf{F}, \mu_1, X)$ and $\mathbb{X}_2 = (\mathcal{X}, \mathcal{F}, \mathbf{F}, \mu_2, X)$. The upper process is a naturally filtered stochastic process, while the lower one is the process carrying extra information from the coin toss.

Before we proceed, we first observe that $\mu \mapsto \mathrm{OS}(f; \mu)$ fails to be concave in general. This prohibits us to apply Remark 4.20 directly as in Section 4.6. In Figure 4.2, we illustrate a concrete example by taking $\mu_1 = \frac{1}{2}(\delta_{(0,0,1)} + \delta_{(0,0,-1)})$, $\mu_2 = \frac{1}{2}(\delta_{(0,0,2)} + \delta_{(0,0,-2)})$, and $f_n(x_{0:n}) = |x_n^2 - 1|$. Under this setting, it is direct to compute

$$\frac{1}{2}(\mathrm{OS}(f; \mu_1) + \mathrm{OS}(f; \mu_2)) = \frac{1}{2}(1 + 3) > \frac{3}{2} = \mathrm{OS}(f; (\mu_1 + \mu_2)/2),$$

and hence $\mathrm{OS}(f; u)$ is not concave in μ. In order to retrieve the concavity, we consider a relaxation of the control set $\mathscr{T}$. Heuristically, we can realize a process with law

$\frac{1}{2}(\mu_1 + \mu_2)$ by tossing a fair coin at the initial step and then follow the process with law μ_1 or μ_2 depending on the outcome of the coin. If the stopping time was with respect to the natural filtration augmented by the coin toss, we would be able to stop the process depending on the outcome of the coin toss. This would allow us to achieve an optimal stopping value exactly equal to $\frac{1}{2}(\mathrm{OS}(f;\mu_1) + \mathrm{OS}(f;\mu_2))$.

In light of the above arguments, we lift the AW-DR optimal stopping problem to the space of adapted processes AP to allow richer filtrations. We define

$$\mathbb{OS}(f;\mathbb{X}) := \sup_{\mathbb{Y}\in\mathsf{AP}} \{ \sup_{\tau\in\mathscr{T}^{\mathbb{Y}}} E_Q[f_\tau(Y_{0:\tau})] - L(\mathcal{T}_{\mathfrak{bc}}(\mathbb{X},\mathbb{Y}))\},$$

where $\mathbb{X} = (\mathcal{X}, \mathcal{F}, \mathbf{F}, \mu, X)$, $\mathbb{Y} = (\Omega^{\mathbb{Y}}, \mathcal{F}^{\mathbb{Y}}, \mathbf{F}^{\mathbb{Y}}, Q, Y)$, and $\mathscr{T}^{\mathbb{Y}}$ the set of $\mathbf{F}^{\mathbb{Y}}$-stopping time. It is direct to verify $\mathbf{OS}(f;\mu)$ is equivalent to the above formulation if we replace the optimizing set AP by NP. One may argue that on AP there is no natural linear structure as different adapted stochastic processes may live on different filtered probability spaces. To resolve this issue, we recall the nested space and the nested distribution introduced in Chapter 2.

Definition 4.38 (Nested space)**.** We recursively define $\widehat{X}_N = \mathcal{X}_N$ and

$$\widehat{\mathcal{X}}_n = \widehat{\mathcal{X}}_n^- \times \widehat{\mathcal{X}}_n^+ := \mathcal{X}_n \times \mathscr{P}(\widehat{\mathcal{X}}_{n+1}) \quad \text{for } 0 \le n \le N-1.$$

For any $\hat{x}_n \in \widehat{\mathcal{X}}_n$, we write it as $\hat{x}_n = (\hat{x}_n^-, \hat{x}_n^+)$ with $\hat{x}_n^- \in \mathcal{X}_n$ and $\hat{x}_n^+ \in \mathscr{P}(\widehat{\mathcal{X}}_{n+1})$. We say $\widehat{\mathcal{X}} = \widehat{\mathcal{X}}_0$ is the nested space associated to $\mathcal{X}$.

We naturally extend the nested distribution to AP.

Definition 4.39 (Nested distribution)**.** For a given adapted stochastic process $\mathbb{X} = (\Omega^{\mathbb{X}}, \mathcal{F}^{\mathbb{X}}, \mathbf{F}^{\mathbb{X}}, X, P^{\mathbb{X}}) \in \mathsf{AP}$, the associated information process $\widehat{X}$ is given recursively by $\widehat{X}_N := X_N$ and

$$\widehat{X}_n = (\widehat{X}_n^-, \widehat{X}_n^+) := (X_n, \mathrm{Law}(\widehat{X}_{n+1}^{\mathbb{X}} | \mathcal{F}_n^{\mathbb{X}})).$$

We say $\widehat{P}^{\mathbb{X}} = \mathrm{Law}(\widehat{X}_0) \in \mathscr{P}(\widehat{\mathcal{X}})$ is the nested distribution associated to $\mathbb{X}$.

Assumption 4.40. We assume there exists $p \ge 1$ such that the following conditions hold:

(i) $L : \mathbb{R}^* \to \mathbb{R}^*$ is non-decreasing and continuous on its domain $\{L < \infty\}$.

(ii) $c : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is continuous and has a polynomial growth

$$|c(x,y)| \le C(1 + d_{\mathcal{X}}(\tilde{x}, x)^p + d_{\mathcal{X}}(\tilde{x}, y)^p) \quad \text{for some } \tilde{x} \in \mathcal{X}.$$

(iii) $f_n : \mathcal{X}_{0:n} \to \mathbb{R}$ is continuous and has a polynomial growth

$$|f_n(x_{0:n})| \leq C(1 + d_{\mathcal{X}}(\tilde{x}_{0:n}, x_{0:n})^p) \quad \text{for some } \tilde{x}_{0:n} \in \mathcal{X}_{0:n}.$$

The following result shows that the bi-causal optimal transport problem between two adapted stochastic processes can be solved by a dynamic programming principle, and this is equivalent to a classical optimal transport problem on the nested space. The proof is similar to Bartl et al. (2024, Theorem 3.10), and we postpone it to the end of this section.

Proposition 4.41. *Let* $\mathbb{X} = (\Omega^{\mathbb{X}}, \mathcal{F}^{\mathbb{X}}, \mathbf{F}^{\mathbb{X}}, P, X)$ *and* $\mathbb{Y} = (\Omega^{\mathbb{Y}}, \mathcal{F}^{\mathbb{Y}}, \mathbf{F}^{\mathbb{Y}}, Q, Y)$ *be two adapted stochastic processes. Under Assumption 4.40, there exists a continuous function* $\hat{c} : \widehat{\mathcal{X}} \times \widehat{\mathcal{X}} \to \mathbb{R}$ *such that*

$$\mathcal{T}_{\mathfrak{bc}}(\mathbb{X}, \mathbb{Y}) = \inf_{\pi \in \Pi(\widehat{P}, \widehat{Q})} E_\pi[\hat{c}(\widehat{X}, \widehat{Y})] := \widehat{\mathcal{T}}(\widehat{P}, \widehat{Q}).$$

Theorem 4.42. *We recursively define* $\hat{f}_n : \widehat{\mathcal{X}}_{0:n} \to \mathbb{R}$ *for* $n \in I$ *as follows. Let* $f_N(\hat{x}_{0:N}) := f(x^-_{0:N})$ *and for* $n = N-1, \ldots, 0$

$$\hat{f}_n(\hat{x}_{0:n}) := \max\{f_n(\hat{x}^-_{0:n}), \hat{x}^+_n(\hat{f}_{n+1}(\hat{x}_{0:n}, \cdot))\}.$$

We assume $\mathcal{X}$ *has no isolated points. Under Assumption 4.40, we have*

$$\mathbf{OS}(f;\mu) = \inf_{\lambda \geq 0}\{L^*(\lambda) + \hat{\mu}(\sup_{\hat{y}}\{\hat{f}(\hat{y}) - \hat{c}(\hat{x}, \hat{y})\})\},$$

where $\hat{\mu} \in \mathscr{P}(\widehat{\mathcal{X}})$ *is the nested distribution associated to* $\mathbb{X} = (\mathcal{X}, \mathcal{F}, \mathbf{F}, \mu, X)$ *and* $\hat{c}$ *is the cost function given in Proposition 4.41.*

Proof. Step 1. We first show $\mathbf{OS}(f;\mu) = \mathbb{OS}(f;\mathbb{X})$. Since we assume $\mathcal{X}$ has no isolated points, by Bartl et al. (2024, Theorem 5.4) NP is a dense subset of AP in $\mathcal{AW}_p$. By Assumption 4.40 (ii) and Eckstein and Pammer (2024, Theorem 3.6), $\mathbb{Y} \mapsto \mathcal{T}_{\mathfrak{c}}(\mathbb{X}, \mathbb{Y})$ are continuous with respect to $\mathcal{AW}_p$. Moreover, by Assumption 4.40 (iii) and Bartl et al. (2024, Proposition 6.1 (ii)), the value of optimal stopping problem is continuous in $\mathcal{AW}_p$. These properties yield

$$\begin{aligned}\mathbb{OS}(f;\mathbb{X}) &= \sup_{\mathbb{Y} \in \mathsf{AP}}\{\sup_{\tau \in \mathscr{T}^{\mathbb{Y}}} E_Q[f_\tau(Y_{0:\tau})] - L(\mathcal{T}_{\mathfrak{bc}}(\mathbb{X}, \mathbb{Y}))\} \\ &= \sup_{\mathbb{Y} \in \mathsf{NP}}\{\sup_{\tau \in \mathscr{T}^{\mathbb{Y}}} E_Q[f_\tau(Y_{0:\tau})] - L(\mathcal{T}_{\mathfrak{bc}}(\mathbb{X}, \mathbb{Y}))\} = \mathbf{OS}(f;\mu).\end{aligned}$$

Step 2. We now derive the duality formula for $\mathbb{OS}(f;\mathbb{X})$. Notice by Snell envelope theorem, we have

$$\sup_{\tau\in\mathscr{T}^{\mathbb{Y}}} E_Q[f_\tau(Y_{0:\tau})] = E_{\widehat{Q}}[\hat{f}(\widehat{X})],$$

where $\widehat{Q}$ is the nested distribution associated to $\mathbb{Y}$. Combining this with Proposition 4.41 yields

$$\begin{aligned}\mathbb{OS}(f;\mathbb{X}) &= \sup_{\mathbb{Y}\in\mathsf{AP}}\{\sup_{\tau\in\mathscr{T}^{\mathbb{Y}}} E_Q[f_\tau(Y_{0:\tau})] - L(\mathcal{T}_{\mathfrak{bc}}(\mathbb{X},\mathbb{Y}))\}\\ &\leq \sup_{\widehat{Q}\in\mathscr{P}(\widehat{\mathcal{X}})}\{E_{\widehat{Q}}[\hat{f}(\widehat{X})] - L(\widehat{\mathcal{T}}(\hat{\mu},\widehat{Q})\}.\end{aligned}$$

On the other hand, for any $\widehat{Q}\in\mathscr{P}(\widehat{\mathcal{X}})$ we can construct an adapted stochastic process $\mathbb{Y}$ such that $\widehat{Q}$ is the nested distribution associated to $\mathbb{Y}$. Indeed, we can take $\mathbb{Y} = (\widehat{\mathcal{X}}_{0:N},\mathcal{F},\mathbf{F},Q,\widehat{X}^-)$, where $\mathcal{F}=\mathcal{B}(\widehat{\mathcal{X}}_{0:N})$, $\mathbf{F}=(\mathcal{F}_n)_{n=0}^N$ with $\mathcal{F}_n=\sigma(\widehat{X}_{0:n})$, $\widehat{X}^-$ is the projection from $\widehat{\mathcal{X}}_{0:N}$ to $\widehat{\mathcal{X}}_{0:n}$, and

$$Q(\mathrm{d}\hat{x}_0,\mathrm{d}\hat{x}_1,\dots,\mathrm{d}\hat{x}_N) = \widehat{Q}(\mathrm{d}\hat{x}_0)\hat{x}_0^+(\mathrm{d}\hat{x}_1)\dots\hat{x}_{N-1}^+(\mathrm{d}\hat{x}_N).$$

Therefore, we derive

$$\mathbb{OS}(f;\mathbb{X}) = \sup_{\widehat{Q}\in\mathscr{P}(\widehat{\mathcal{X}})}\{E_{\widehat{Q}}[\hat{f}(\widehat{X})] - L(\widehat{\mathcal{T}}(\hat{\mu},\widehat{Q})\}.$$

The right-hand side is a static Wasserstein DRO problem. Applying Theorem 4.24 with $N=0$, we derive the desired duality. $\square$

Proof of Proposition 4.41. We recursively define $\hat{c}_n:\widehat{\mathcal{X}}_{0:n}\times\widehat{\mathcal{X}}_{0:n}\to\mathbb{R}$ for $n\in I$. Let $\hat{c}_N(\hat{x}_{0:N},\hat{y}_{0:N}) = c(x^-_{0:n},y^-_{0:n})$ and

$$\hat{c}_n(\hat{x}_{0:n},\hat{y}_{0:n}) = \sup_{\pi\in\Pi(\hat{x}_n^+,\hat{y}_n^+)} E_\pi[\hat{c}_{n+1}(\hat{x}_{0:n},\widehat{X}_{n+1},\hat{y}_{0:n},\widehat{Y}_{n+1})].$$

Notice that each $\hat{c}_n$ is a classic optimal transport problem, and hence it is continuous by Assumption 4.40 (ii) and Bogachev and Popova (2021, Corollary 3.6). We recall the definition of $\widehat{X}_n$ in Definition 4.39

$$\widehat{X}_n = (\widehat{X}_n^-,\widehat{X}_n^+) := (X_n,\mathrm{Law}(\widehat{X}^{\mathbb{X}}_{n+1}|\mathcal{F}^{\mathbb{X}}_n)).$$

Now we notice

$$\begin{aligned}\inf_{\pi\in\Pi_{\mathfrak{bc}}(\mathbb{X},\mathbb{Y})} E_\pi[c(X,Y)] &= \inf_{\pi\in\Pi_{\mathfrak{bc}}(\mathbb{X},\mathbb{Y})} E_\pi[\hat{c}_N(\widehat{X},\widehat{Y})]\\ &= \inf_{\pi\in\Pi_{\mathfrak{bc}}(\mathbb{X},\mathbb{Y})} E_\pi[E_\pi[\hat{c}_N(\widehat{X},\widehat{Y})|\mathcal{F}^{\mathbb{X}}_{n-1}\otimes\mathcal{F}^{\mathbb{Y}}_{n-1}]\\ &= \inf_{\pi\in\Pi_{\mathfrak{bc}}(\mathbb{X},\mathbb{Y})} E_\pi[\sup_{\gamma\in\Pi(\widehat{X}_n^+,\widehat{Y}_n^+)} E_{(\xi,\eta)\sim\gamma}[\hat{c}_N(\widehat{X}_{0:N-1},\xi,\widehat{Y}_{0:N-1},\eta)]]\\ &= \inf_{\pi\in\Pi_{\mathfrak{bc}}(\mathbb{X},\mathbb{Y})} E_\pi[\hat{c}_{N-1}(\widehat{X}_{0:N-1},\widehat{Y}_{0:N-1})].\end{aligned}$$

The last equality follows from an extension of Backhoff-Veraguas et al. (2017, Proposition 2.4). We apply the above argument repeatedly, and obtain

$$\mathcal{T}_{\mathfrak{bc}}(\mathbb{X}, \mathbb{Y}) = \inf_{\pi \in \Pi(\hat{P}, \hat{Q})} E_\pi[\hat{c}(\widehat{X}, \widehat{Y})],$$

where we set $\hat{c} = \hat{c}_0$. $\square$

Chapter 5

Adapted Wasserstein DRO: Sensitivity

5.1 Introduction

Naturally, the issue of model uncertainty and its impact on agents' course of action is a topic with a long history, including in decision theory and mathematical finance. Any lecture in mathematical finance covering derivatives' pricing, will typically also cover the 'Greeks', the sensitivities of the prices to key model parameters, used throughout the financial industry. Parametric, or similarly specific, sensitivities of optimal investment problems have been considered in a number of works, see Larsen and Žitković (2007), Mostovyi and Sîrbu (2019) and the references therein. Specifically, the framework of robust utility maximization, which benefits from a strong axiomatic justification (Gilboa and Schmeidler, 1989, Maccheroni et al., 2006), has been studied in depth. Its analytic properties such as the existence of optimal strategies (Tevzadze et al., 2013, Neufeld and Nutz, 2018), for formulation of a dynamic programming principle (Schied, 2007, Zitković, 2009, Källblad et al., 2018), and relation to 2BSDEs (Matoussi et al., 2015), are well-understood. However, except special cases, these works rarely allow for any explicit computations.

In this chapter, we study adapted Wasserstein DRO by computing the explicit sensitivity to model uncertainty itself. To this end, we fix $p, q > 1$ with $1/p + 1/q = 1$ and introduce a parameterized variant of AW-DRO as

$$V(\delta) = \sup_{\nu \in \mathscr{P}(\mathcal{X})} \{E_\nu[f(X)] - L_\delta(\mathcal{CW}_p(\mu, \nu))\}, \tag{5.1}$$

where L is a penalty function, $L_\delta(\cdot) = \delta L(\cdot/\delta)$, and $\mathcal{CW}_p$ is the p-causal Wasserstein distance. The penalty strength is controlled through the real-valued parameter δ

which, in the special case of an indicator penalty, is simply the radius of the uncertainty ball. Furthermore, we focus on a case of central interest in robust finance: pricing financial derivatives in a misspecified non-arbitrage market. This requires imposing a martingale constraint on the set of models, leading to the problem:

$$V_{\mathrm{Mart}}(\delta) = \sup_{\nu \in \mathscr{M}(\mathcal{X})} \{E_\nu[f(X)] - L_\delta(\mathcal{AW}_p(\mu, \nu))\}, \tag{5.2}$$

where $\mathscr{M}(\mathcal{X})$ denotes the set of martingale measures. We note that we penalize the two problems using different, though related, distances. As established in Chapter 4, these penalizations are equivalent in the discrete-time settings we consider here.

We set up the discrete-time framework as follows. Let $I = \{0, 1, \dots, N\}$ and $\mathcal{X} = \{0\} \times (\mathbb{R}^n)^N$. By $|\cdot|$ we denote the l_p-norm on $\mathbb{R}^n$ and by $|\cdot|_*$ we denote the l_q-norm on $\mathbb{R}^n$. Let $\Delta : \mathcal{X} \to \mathcal{X}$ be the increment map given by $(0, x_1, \dots, x_N) \mapsto (0, x_1, \dots, x_N - x_{N-1})$. We equip $\mathcal{X}$ with the metric given by

$$d_N(x, y) = \left(\sum_{n=1}^{N} |\Delta x_n - \Delta y_n|^p \right)^{1/p}. \tag{5.3}$$

We set $\mathcal{CW}_p$ in (5.1) and $\mathcal{AW}_p$ in (5.2) the distances induced by d_N. In Theorem 5.15 we show that under mild regularity conditions the sensitivity of (5.1) is given by

$$\Upsilon := \lim_{\delta \to 0} \frac{1}{\delta}(V(\delta) - V(0)) = L^* \left(E_\mu \left[\sum_{n=1}^{N} |E_\mu[\mathbb{D}_n f(X) | \mathcal{F}_n]|_*^q \right]^{1/q} \right), \tag{5.4}$$

where L^* is the convex conjugate of L, and $\mathbb{D} = (\mathbb{D}_1, \dots, \mathbb{D}_N)$ is the pullback of $\nabla = (\partial_1, \dots, \partial_N)$ under Δ. For the martingale constraint problem (5.2) with $p = 2$, Theorem 5.18 shows that

$$\begin{aligned} \Upsilon_{\mathrm{Mart}} :&= \lim_{\delta \to 0} \frac{1}{\delta}(V_{\mathrm{Mart}}(\delta) - V_{\mathrm{Mart}}(0)) \\ &= L^* \left(E_\mu \left[\sum_{n=1}^{N} |E_\mu[\mathbb{D}_n f(X) | \mathcal{F}_n] - E_\mu[\mathbb{D}_n f(X) | \mathcal{F}_{n-1}]|_*^2 \right]^{1/2} \right). \end{aligned} \tag{5.5}$$

Our discrete-time results (5.4) and (5.5) thus offer a multi-step extension to the single-period sensitivities studied in Bartl et al. (2021).

With the discrete-time results in hand, we investigate the corresponding continuous-time limits in different scaling regimes. We focus on stochastic processes with paths in $\mathcal{X} = C_0([0, T]; \mathbb{R}^n)$, i.e., the space of continuous paths starting at 0. For any path $\omega \in \mathcal{X}$, we denote by $\omega^N = (0, \omega(T/N), \ldots, \omega(T))$ its discretization. We first introduce the *hyperbolic* scaling:

$$d(\omega, \eta) := \lim_{N\to\infty} \left(\frac{N}{T}\right)^{1-1/p} d_N(\omega^N, \eta^N) = \|\omega - \eta\|_{W_0^{1,p}},$$

where the Sobolev norm is given by $\|\omega\|_{W_0^{1,p}} = \|\dot{\omega}\|_{L^p}$. This scaling regime is natural for capturing uncertainty in the drift of a process. Under such scaling, we show in Theorem 5.22 that the sensitivity converges to a natural continuous-time analogue of (5.4) given by

$$\Upsilon = \lim_{\delta\to 0} \frac{1}{\delta}(V(\delta) - V(0)) = L^*\left(E_\mu\left[\int_0^T |E_\mu[\mathbb{D}_t f(X)|\mathcal{F}_t]|_*^q \, \mathrm{d}t\right]^{1/q}\right), \tag{5.6}$$

where $\mathbb{D}_t$ is a novel *pathwise* Malliavin derivative (intuitively) given by

$$\langle \mathbb{D}_t f(\omega), e\rangle := \lim_{\varepsilon\to 0} \frac{f(\omega + \varepsilon e \mathbb{1}_{[0,t]}) - f(\omega)}{\varepsilon}.$$

We defer the formal definition of $\mathbb{D}$ and the relation to the classical Malliavin derivative to Section 5.2. However, a similar limiting argument under the martingale constraint would feature the difference of the optional and predictable projections, which is known to be a thin process and would thus give $\Upsilon_{\text{Mart}} = 0$. Alternatively, this is also clear from Doob's decomposition: processes at a finite bi-causal distance under $d(\omega, \eta) = \|\omega - \eta\|_{W_0^{1,p}}$ will only differ by their finite variation part. In fact, there is no other martingale measure in any causal Wasserstein neighborhood of a given martingale measure μ in the current scaling.

Put differently, the *hyperbolic* scaling is not critical for the martingale constraint problem. To allow ambiguity in the volatility, we have to 'zoom out' by using a *parabolic* scaling. For $p = 2$, this corresponds to a limiting distance given by

$$d(\omega, \eta) = \lim_{N\to\infty} d_N(\omega^N, \eta^N) = \sqrt{[\omega - \eta]_T},$$

where $[\cdot]_T$ denotes the scalar quadratic variation at terminal time T. We take μ as the Wiener measure, denote the classical Malliavin derivative by $\mathbf{D}$ and focus on the objective of the form of

$$f(X) = U(H), \quad \text{for } H = \int_0^T \langle h(t, X(\cdot \wedge t)), \mathrm{d}X(t)\rangle,$$

for reasons explored in Remark 5.27. In Theorem 5.30, we derive the sensitivity

$$\Upsilon_{\text{Mart}} = L^*\left(E_\mu\left[\int_0^T \|\varphi(t)\|_{\mathbf{F}}^2 \, \mathrm{d}t\right]^{1/2}\right),$$

where $\|\cdot\|_{\mathbf{F}}$ is the Frobenius norm, $^\intercal$ is the linear transpose, and

$$\varphi(t)=E_\mu[\mathbf{D}_{t+}\mathbf{D}_tU(H)-U'(H)\mathbf{D}_t^\intercal h(t,X(\cdot\wedge t))|\mathcal{F}_t].$$

The terms *hyperbolic* and *parabolic* scaling are borrowed from the literature on hydrodynamic limits of interacting particle systems, see Kipnis and Landim (1999). Roughly speaking, if the microparticle system has a non-zero mean, its distribution profile converges to a macroscopic continuity equation under hyperbolic scaling, characterizing a drift process. Conversely, if the particle system is centered, its distribution profile converges to a macroscopic Fokker–Planck equation under parabolic scaling, characterizing a diffusion process. In our context, hyperbolic scaling corresponds to the model's drift uncertainty, while parabolic scaling corresponds to its volatility uncertainty. In particular, the adverse distribution is approximately a perturbation of the reference model by a drift process and a diffusion process, respectively.

For the continuous-time results, both scaling regimes are novel since we do not require an inner product structure on $C_0([0,T];\mathbb{R}^n)$. Under hyperbolic scaling, essentially we derive a *pathwise* first order expansion along any path h in the Sobolev space $W_0^{1,p}([0,T];\mathbb{R}^n)$. We introduce a *pathwise* Malliavin derivative such that the directional derivative of f along η can be represented as $\langle \mathbb{D}f,\dot\eta\rangle$. For parabolic scaling, we study the property of the forward integral against a martingale $\int_0^T\langle\Phi(t),\mathrm{d}^-M(t)\rangle$. A new stochastic Fubini theorem is established in Theorem 5.12 which states that we can swap the order of the forward integral and the Itô integral with an extra correction term. The sensitivity Υ_{Mart} then follows from a first order expansion in Lemma 5.36.

5.1.1 Related literature

To study (5.1) and (5.2), there are two lines of research: duality and sensitivity. We refer to Chapter 4 for a review on the duality approach and focus here on the literature concerning sensitivity. In the static setting, the classical Wasserstein DRO sensitivity was first derived in Bartl et al. (2021) with an indicator penalty and later extended to a general penalization in Nendel and Sgarabottolo (2024). These foundational results have recently been generalized to multi-step dynamic contexts, with a focus on settings that have a specific dynamic Markovian structure (Fuhrmann et al., 2023, Langner et al., 2024, Neufeld and Sester, 2024, Mirmominov and Wiesel, 2024).

Under an adapted Wasserstein setting, partial discrete-time sensitivity results were derived independently of this chapter in Bartl and Wiesel (2023) for the unconstrained case, and in Sauldubois and Touzi (2024) for the martingale-constrained case. For both works, the authors took an indicator penalty $L = +\infty \mathbb{1}_{(1,\infty)}$ and considered a standard l_p metric $\tilde{d}_N(x,y) = (\sum_{n=1}^N |x_n - y_n|^p)^{1/p}$ rather than d_N given in (5.3). We emphasize the choice is critical: while equivalent in discrete time, our choice allows obtaining continuous time results as limits of the discrete time results. For the martingale constraint problem in Theorem 5.18, we can formally rewrite it as an unconstrained problem with a Lagrange multiplier:

$$V_{\mathrm{Mart}}(\delta) = \sup_{\nu \in \mathscr{P}(\mathcal{X})} \inf_{h \in \mathbb{H}_b} \{E_\nu[f(X) - h \circ X] - L_\delta(\mathcal{AW}_p(\mu,\nu))\},$$

where $\mathbb{H}_b$ is the set of bounded predictable processes and $h \circ X$ denotes the discrete stochastic integral. This formulation aligns with Bartl and Wiesel (2023, Theorem 2.4). However, Bartl and Wiesel (2023, Assumption 2.4) essentially imposes the uniqueness of the optimizer h^* for

$$\inf_{h \in \mathbb{H}_b} E_\mu[f(X) - h \circ X],$$

which is of sharp contrast to the martingale constraint problem where any $h \in \mathbb{H}_b$ is an optimizer.

To the best of our knowledge, this is the first work that studies the continuous-time AW-DRO sensitivity as a limit of its discrete-time counterpart. The closest setting in continuous time is the recent work by Bartl et al. (2025b), which considers L^p-balls around the drift and the volatility of the reference model under a strong formulation. We discuss the link between our works in detail in Section 5.4.3 below. Other related works, such as Herrmann and Muhle-Karbe (2017), Herrmann et al. (2017) consider related sensitivities albeit in a more specific setup with the penalty in (5.1) depending on the payoff in a way which makes the sensitivity universal and independent of agent's risk aversion.

5.1.2 Outline

The rest of the chapter is organized as follows. Section 5.2 introduces the necessary notations and concepts. In Section 5.3, we develop the key technical tools for our analysis, including a novel pathwise Malliavin derivative and properties of the forward integral. We then present our main sensitivity results in Section 5.4, complete with examples and a discussion of possible extensions. The proofs are detailed in the subsequent sections: Section 5.4.4 outlines a unified framework, which is then applied to derive the discrete-time sensitivity (Section 5.5) and the continuous-time sensitivities under hyperbolic scaling (Section 5.6.1) and parabolic scaling (Section 5.6.2). Finally, proofs of auxiliary technical results are postponed to Section 5.7.

5.2 Notations

Throughout the chapter, we fix $N, d \geq 1$, $T > 0$, $p > 1$ and let $q = p/(p-1)$. We use the generic notation $\langle \cdot, \cdot \rangle$ for the scalar product, where the space is clear from the context. We consider stochastic processes taking value in $\mathbb{R}^n$ and recall that we use the notation $|\cdot|$ for the l_p-norm on $\mathbb{R}^n$, and $|\cdot|_*$ for the l_q-norm on $\mathbb{R}^n$. In discrete-time, we take time index $I = \{0, 1, \ldots, N\}$ and refer to the canonical path space as $\mathcal{X} = \{0\} \times (\mathbb{R}^n)^N$ equipped with its natural filtration; in continuous-time, we take time index $I = [0, T]$ and refer to the canonical path space as $\mathcal{X} = C_0([0, T]; \mathbb{R}^n)$. On the product space $\mathcal{X} \times \mathcal{X}$, we denote the first and the second coordinate processes by X and Y. By $\mathbf{F} = (\mathcal{F}_t)_{t \in I}$ and $\mathbf{G} = (\mathcal{G}_t)_{t \in I}$ we denote the natural filtrations generated by X and Y respectively.

Let $W_0^{1,p}$ be the Sobolev subspace of C_0 equipped with the norm

$$\|x\|_{W_0^{1,p}} = \|\dot{x}\|_{L^p}.$$

Let $A = (A(t))_{t \in I}$ be a stochastic process. In discrete time we write

$$\|A\|_{L^q(\mu)} = \left(\sum_{n=1}^{N} E_\mu[|A_n|_*^q] \right)^{1/q};$$

and in continuous time we write

$$\|A\|_{L^q(\mu)} = \left(\int_0^T E_\mu[|A(t)|_*^q] \, \mathrm{d}t \right)^{1/q}.$$

For A with $E_P[\|A\|_\infty] < \infty$, we denote the optional projection of A by ${}^{\mathrm{o}}A$ which is the unique optional process (up to a P-null set) such that for any bounded optional stopping time τ

$${}^{\mathrm{o}}A_\tau = E_P[A_\tau | \mathcal{F}_\tau].$$

We denote the predictable projection of A by ${}^{\mathrm{p}}A$ which is the unique process (up to a P-null set) such that for any bounded predictable stopping time τ

$${}^{\mathrm{p}}A_\tau = E_P[A_\tau | \mathcal{F}_{\tau-}].$$

Note that in discrete time there are no issues with the pathwise regularity, and the projections are simply given by

$${}^{\mathrm{o}}X_n = E[X_n | \mathcal{F}_n] \quad \text{and} \quad {}^{\mathrm{p}}X_n = E[X_n | \mathcal{F}_{n-1}].$$

For a multidimensional martingale M, we denote $[\![M]\!]$ its matrix-valued quadratic variation process and $[M]$ its trace, which we refer to as the scalar quadratic variation process. In the sequel, we use *regular martingale* to refer to a continuous-time square-integrable continuous martingale M with absolutely continuous quadratic variation. By $\mathscr{M}(\mathcal{X})$ we denote the space of regular martingale measures on $\mathcal{X}$. By $\langle\cdot,\cdot\rangle_{\mathbf{F}}$ and $\|\cdot\|_{\mathbf{F}}$ we denote the Frobenius inner product and norm respectively.

We adopt Landau symbols o and O. By $o(r)$ we denote a quantity bounded by $l(r)r$ with $\lim_{r\to 0} l(r) = 0$, and by $O(r)$ we denote a quantity bounded by $l(r)r$ with $\lim_{r\to 0} l(r) < \infty$. We stress that $l(r)$ is deterministic and independent of the underlying probability measure.

By $\Pi(\mu,*)$ we denote the set of couplings with a fixed first marginal μ. Accordingly, $\Pi_{\mathfrak{c}}(\mu,*)$ and $\Pi_{\mathfrak{bc}}(\mu,*)$ represent the respective subsets of causal and bi-causal couplings.

5.3 Some tools from stochastic calculus

5.3.1 Malliavin derivative

We follow Nualart (2006) to give a brief introduction to the (classical) Malliavin calculus. We then introduce the notion of *pathwise* Malliavin derivative. It arises naturally as the limit of discrete objects, and we show it coincides with the classical version on the intersection of their domains. We believe such a pathwise approach to Malliavin calculus can unlock many interesting results. It links with the functional Itô calculus of Dupire (2009), Cont and Fournié (2013) and will be key to extending our sensitivity results in continuous time to reference measures beyond the Brownian case. We plan to pursue this direction of research in a future chapter, see also Section 5.4.3.

Let X be a d-dim Brownian motion on a filtered probability space $(\Omega,\mathcal{F},(\mathcal{F}_t)_{t\in I},P)$. For a smooth cylindrical random variable $F = f(\int_0^T\langle h(t),\mathrm{d}X(t)\rangle)$ where $f\in C_b^1$ and $h\in L^2([0,T],\mathbb{R}^n)$, its Malliavin derivative $\mathbf{D}F$ is given by

$$\mathbf{D}_tF := f'\bigg(\int_0^T\langle h(s),\mathrm{d}X(s)\rangle\bigg)h(t).$$

It is well-known that $\mathbf{D}$ is closable on $L^2(P,\mathcal{F}_T^X)$. Hence, we do not distinguish $\mathbf{D}$ from its closure and denote its domain by $\mathbf{D}^{1,2}$. The predictable projection of the Malliavin derivative solves the martingale representation problem.

Theorem 5.1 (Clark–Ocone formula)**.** *Assume $Z\in\mathbf{D}^{1,2}$. Then we have*

$$Z = E_P[Z] + \int_0^T\langle {}^{\mathrm{p}}\mathbf{D}_tZ,\mathrm{d}X(t)\rangle.$$

Malliavin derivative can be interpreted as a 'gradient' operator on the 'tangent space', the Cameron–Martin space $W_0^{1,2}$.

Proposition 5.2. *Assume $f : \mathcal{X} \to \mathbb{R}$ and $f(X) \in \mathbf{D}^{1,2}$. Then for any $\eta \in W_0^{1,2}$, we have*

$$\lim_{\varepsilon\to 0} \frac{f(X+\varepsilon\eta) - f(X)}{\varepsilon} = \int_0^T \langle \mathbf{D}_t f(X), \dot{\eta}\rangle \, \mathrm{d}t \quad P\text{-a.s.}. \tag{5.7}$$

5.3.2 Pathwise Malliavin derivative

To introduce pathwise Malliavin derivative, we start with the discrete-time setup. Recall that $\Delta : \mathcal{X} \to \mathcal{X}$ is the increment map given by

$$\Delta x = (0, \Delta x_1, \ldots, \Delta x_N) := (0, x_1, x_2 - x_1, \ldots, x_N - x_{N-1}).$$

Then the *pathwise* Malliavin derivative $\mathbb{D} = (\mathbb{D}_1, \ldots, \mathbb{D}_N)$ is defined as the pullback of $\nabla = (\partial_1, \ldots, \partial_N)$ under Δ, i.e., for any smooth f

$$\langle \mathbb{D} f(x), y\rangle := \langle \nabla f(x), \Delta^{-1} y\rangle = \langle (\Delta^{-1})^* \nabla f(x), y\rangle, \tag{5.8}$$

where $(\Delta^{-1})^*$ is the adjoint operator of Δ^{-1}. Expressed explicitly, this gives

$$\mathbb{D}_n = \sum_{k=n}^{N} \partial_k, \quad 1 \le n \le N. \tag{5.9}$$

Here, $\mathbb{D} f$ is an analogue of Malliavin derivative in the sense that for any deterministic path $x, y \in \mathcal{X}$

$$\lim_{\varepsilon\to 0} \frac{f(x+\varepsilon y) - f(x)}{\varepsilon} = \langle \mathbb{D} f(x), \Delta y\rangle.$$

In continuous-time setting, while we focus on stochastic processes with continuous path, it is natural to first define *pathwise* Malliavin derivative for functionals on the càdlàg path space $D([0,T];\mathbb{R}^n)$. This approach closely aligns with the functional Itô calculus from Cont and Fournié (2013). Proofs of results in this section are deferred to Section 5.7.

Definition 5.3 (Pathwise Malliavin derivative)**.** Let $f : D([0,T];\mathbb{R}^n) \to \mathbb{R}$ be a functional on the càdlàg path space. We say that f is *pathwise Malliavin differentiable* if there exists $\mathbb{D} f : D([0,T];\mathbb{R}^n) \times [0,T] \to \mathbb{R}^d$ such that

$$\langle \mathbb{D}_t f(\omega), e\rangle = \lim_{\varepsilon\to 0} \frac{f(\omega + \varepsilon e \mathbb{1}_{[t,T]}) - f(\omega)}{\varepsilon}, \quad \forall \omega \in D([0,T];\mathbb{R}^n),\ e \in \mathbb{R}^n.$$

We call $\mathbb{D} f = (\mathbb{D}_t f)_{t\in[0,T]}$ the *pathwise Malliavin derivative* of f.

Definition 5.4 (Left limit)**.** For a functional $F : [0,T] \times D([0,T];\mathbb{R}^n) \to \mathbb{R}$, we say F has a left limit at $t \in [0,T]$ if for any ω^n converging uniformly to ω and t^n increasing to t the following limit exists

$$\lim_{n\to\infty} F(t^n, \omega^n) = F(t-, \omega).$$

Definition 5.5 (Boundedness preservation)**.** We say a functional $F : [0,T]\times D([0,T];\mathbb{R}^n) \to \mathbb{R}$ is boundedness preserving if for any K

$$\sup_{t\in[0,T]} \sup_{\|\omega\|_\infty < K} |F(t,\omega)| < \infty.$$

Proposition 5.6. *Assume f and $\mathbb{D}_t f$, $t \in I$, are all continuous with respect to the uniform topology. Then, for any simple step function $\eta = \sum_{k=1}^n e_k \mathbb{1}_{[t_k,1]}$, we have*

$$\lim_{\varepsilon\to 0} \frac{f(\omega+\varepsilon\eta) - f(\omega)}{\varepsilon} = \sum_{k=1}^n \langle \mathbb{D}_{t_k} f(\omega), e_k\rangle.$$

If further $\mathbb{D}f$ has a left limit for all $t \in [0,T]$ and is boundedness preserving then, for any path $\eta \in AC_0([0,T];\mathbb{R}^n)$, we have

$$\lim_{\varepsilon\to 0} \frac{f(\omega+\varepsilon\eta) - f(\omega)}{\varepsilon} = \int_0^T \langle \mathbb{D}_t f(\omega), \dot\eta(t)\rangle \,\mathrm{d}t. \tag{5.10}$$

We note that (5.10) fully characterizes (up to a Lebesgue null set) the left limit of the Malliavin derivative. This, in particular, offers us a natural way to define the pathwise Malliavin derivative for a functional defined on the continuous path space $f : C_0([0,T];\mathbb{R}^n) \to \mathbb{R}$.

Definition 5.7. We denote $\mathbb{D}_b^1$ the space of functionals $f : C_0([0,T];\mathbb{R}^n) \to \mathbb{R}$ which admit an extension $\tilde f$ to $D([0,T];\mathbb{R}^n)$ satisfying assumptions in Proposition 5.6, and for such f we set

$$\mathbb{D}_t f(\omega) := \mathbb{D}_t \tilde f(\omega), \quad t \in [0,T],\ \omega \in C_0([0,T];\mathbb{R}^n).$$

We remark that $\mathbb{D}f$ is measurable and does not depend on the choice of the extension $\tilde f$ as the right-hand side only depends on the values of $\tilde f$ on the continuous path space.

Example 5.8. We give a few examples of functionals $f \in \mathbb{D}_b^1$.

1. $f(\omega) = g(\omega(t_1), \dots, \omega(t_n))$, where $g \in C_b^1$.
   This gives $\mathbb{D}_t f(\omega) = \sum_{k=1}^n \partial_k g(\omega(t_1), \dots, \omega(t_n))\mathbb{1}_{[0,t_k]}$.

2. $f(\omega) = \int_0^T g(\omega(t))\theta(\mathrm{d}t)$ where $g \in C_b^1$ and θ is a finite measure. This gives $\mathbb{D}_t f(\omega) = \int_t^T \nabla g(\omega(s))\theta(\mathrm{d}s)$.

The following property shows that the classical and the *pathwise* Malliavin derivatives agree on their common domain.

Proposition 5.9. *Let $\mathcal{X} = C_0([0,T];\mathbb{R}^n)$, μ the Wiener measure, and X be the canonical process. Assume $f \in \mathbb{D}_b^1$ and $f(X) \in \mathbf{D}^{1,2}$. Then, we have*

$$(\mathbb{D}_t f)(X) = \mathbf{D}_t f(X) \quad \mathrm{d}\mu \otimes \mathrm{d}t\text{–}a.e..$$

5.3.3 Forward integral against regular martingales

Let X be a d-dim Brownian motion on a filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\in I}, P)$. The forward integration was introduced as an anticipative extension to the Itô integral, and it plays a crucial role in our sensitivity analysis. We first recall its definition.

Definition 5.10 (Russo and Vallois (1993))**.** Let A, B be two measurable stochastic processes (not necessary adapted). The forward integral is given by

$$\int_0^T \langle A(t), \mathrm{d}^- B(t)\rangle = \lim_{\varepsilon\to 0} \int_0^T \frac{1}{\varepsilon}\langle A(t), B((t+\varepsilon)\wedge T) - B(t)\rangle\,\mathrm{d}t,$$

if the right hand side limit exists in probability. We say A is B forward–γ integrable if the above limit exists in L^γ. A family $\{A_\lambda\}_{\lambda\in\Lambda}$ of B forward–γ integrable processes is said to be B uniformly forward–γ integrable if

$$\sup_{\lambda\in\Lambda}\left|\int_0^T \langle A_\lambda(t), \mathrm{d}^- B(t)\rangle - \int_0^T \frac{1}{\varepsilon}\langle A_\lambda(t), B((t+\varepsilon)\wedge T) - B(t)\rangle\,\mathrm{d}t\right|$$

converges to 0 in L^γ as ε goes to 0.

We focus on the case where the integrator of the forward integral is a regular martingale. Unlike Itô integral, the expectation of a forward integral can be nonzero as a consequence of the anticipative nature of the integrand.

Proposition 5.11. *Let M be a regular martingale. We assume $\Phi(t) \in \mathcal{F}_T^X$ for any $t \in I$, and Φ is M forward–1 integrable. If the right limit $\lim_{s\to t+} \mathbf{D}_s\Phi(t) = \mathbf{D}_{t+}\Phi(t)$ converges in L^2 and $\sup_{s,t\in[0,T]} \|\mathbf{D}_s\Phi(t)\|_{\mathbf{F}} \in L^2$, then we have*

$$E_P\left[\int_0^T \langle \Phi(t), \mathrm{d}^- M(t)\rangle\right] = E_P\left[\int_0^T \langle \mathbf{D}_{t+}\Phi(t), \mathrm{d}[\![X, M]\!](t)\rangle_{\mathbf{F}}\right],$$

Proof. By definition of the forward-1 integrability and L^1 convergence, we have

$$E_P\left[\int_0^T \langle \Phi(t), \mathrm{d}^- M(t)\rangle\right] = \lim_{\varepsilon\to 0} E_P\left[\int_0^T \frac{1}{\varepsilon}\langle \Phi(t), M((t+\varepsilon)\wedge T) - M(t)\rangle \,\mathrm{d}t\right].$$

By applying Clark–Ocone formula to $\Phi(t)$, we derive

$$\begin{aligned}
E_P\left[\int_0^T \langle \Phi(t), \mathrm{d}^- M(t)\rangle\right] &= \lim_{\varepsilon\to 0} E_P\left[\int_0^T \frac{1}{\varepsilon}\left\langle \int_0^T {}^{\mathrm{p}}\mathbf{D}_s^{\intercal}\Phi(t)\,\mathrm{d}X(s), \int_t^{(t+\varepsilon)\wedge T} \mathrm{d}M(s)\right\rangle \mathrm{d}t\right] \\
&= \lim_{\varepsilon\to 0} E_P\left[\int_0^T \frac{1}{\varepsilon}\int_t^{(t+\varepsilon)\wedge T} \langle {}^{\mathrm{p}}\mathbf{D}_s\Phi(t), \mathrm{d}[\![X, M]\!](s)\rangle_{\mathbf{F}}\,\mathrm{d}t\right] \\
&= E_P\left[\int_0^T \langle {}^{\mathrm{p}}\mathbf{D}_{t+}\Phi(t), \mathrm{d}[\![X, M]\!](t)\rangle_{\mathbf{F}}\right],
\end{aligned}$$

where the second line follows from the Itô isometry and Fubini theorem, and the last line follows from the dominated convergence theorem and $\sup_{s,t\in[0,T]} \|\mathbf{D}_s\Phi(t)\|_{\mathbf{F}} \in L^2$. We conclude the proof by noticing $[\![X, M]\!]$ is a predictable process. □

We give now a stochastic Fubini theorem which details the correction term arising from an interchange between a forward and an Itô integral. To the best of our knowledge, this is a novel result which we believe is of independent interest. Its proof is deferred to Section 5.7.

Theorem 5.12 (Stochastic Fubini theorem)**.** *Let $\gamma \in [1, 2)$. Let $\Psi : I \times I \times \Omega \to \mathbb{R}^{d\times d}$, M be a regular martingale. We assume that $\Psi(s,\cdot)$ is predictable for any $s \in I$ and satisfies the following conditions:*

1. *$\{\Psi(\cdot, t)\}_{t\in[0,T]}$ is M. uniformly forward–γ integrable,*
2. *$\sup_{s,t\in[0,T]} \|\Psi(s, t)\|_{\mathbf{F}} \in L^{2\gamma/(2-\gamma)}$,*
3. *$\lim_{t\to s+} E_P[\|\Psi(s, t) - \Psi(t, t)\|^2_{\mathbf{F}}] = 0$ for any $s \in [0, T]$.*

Then, we have $\int_\cdot^T \Psi(\cdot, t)^{\intercal}\,\mathrm{d}X(t)$ is M. forward–γ integrable. Moreover,

$$\begin{aligned}
&\int_0^T \left\langle \int_s^T \Psi(s, t)^{\intercal}\,\mathrm{d}X(t), \mathrm{d}^- M(s)\right\rangle \\
&= \int_0^T \left\langle \int_0^t \Psi(s, t)\,\mathrm{d}^- M(s), \mathrm{d}X(t)\right\rangle + \int_0^T \langle \Psi(t, t), \mathrm{d}[\![X, M]\!](t)\rangle_{\mathbf{F}}.
\end{aligned}$$

5.4 Sensitivity of AW-DRO

With the tools developed so far, we can now give rigorous statements of our main results on the sensitivity of AW-DRO problems to model uncertainty. We recall that $p > 1$ is fixed, $q = \frac{p}{p-1}$, and the parameterized penalty is given by $L_\delta(\cdot) = \delta L(\cdot/\delta)$. We impose the following assumption on L. The growth condition ensures that when δ goes to 0, the adversarial distribution will converge to the reference model μ. We stress that the proofs offer direct characterizations of the first-order optimal adversarial model perturbations. We make this explicit only for the first theorem, see Remark 5.16. As highlighted in (Bai et al., 2023) this can be as important as the sensitivity computation itself.

Assumption 5.13. We assume that $L : [0, +\infty) \to [0, +\infty]$ is continuous, non-decreasing, and satisfies

$$L(0) = 0 \quad \text{and} \quad \liminf_{u\to\infty} \frac{L(u)}{u^p} = +\infty.$$

We write the convex conjugate of L as L^* given by

$$L^*(v) = \sup_{u\geq 0}\{uv - L(u)\}.$$

5.4.1 Discrete-time results

We start with the discrete-time setting. We state and discuss the results, with the proofs deferred to Section 5.5. Let $\mathcal{CW}_p$ and $\mathcal{AW}_p$ be the distances induced by the metric

$$d_N(x, y) = \left(\sum_{n=1}^{N} |\Delta x_n - \Delta y_n|^p\right)^{1/p},$$

and recall $\mathbb{D}$ is the discrete Malliavin derivative in (5.9). We consider

$$V(\delta) = \sup_{\nu\in\mathscr{P}(\mathcal{X})} \{E_\nu[f(X)] - L_\delta(\mathcal{CW}_p(\mu, \nu))\}. \tag{5.11}$$

Assumption 5.14. We assume that $f : \mathcal{X} \to \mathbb{R}$ is continuously differentiable. Moreover, $\mathbb{D}f$ satisfies for any $x \in \mathcal{X}$

$$|\mathbb{D}f(x)| \leq C(1 + |x|^{p-1}). \tag{5.12}$$

Theorem 5.15. *Under Assumptions 5.13 and 5.14, we have*

$$V(\delta) = V(0) + \Upsilon\delta + o(\delta),$$

where

$$\Upsilon := \lim_{\delta\to 0}\frac{1}{\delta}(V(\delta)-V(0)) = L^*\left(E_\mu\left[\sum_{n=1}^N \big|E_\mu[\mathbb{D}_n f(X)|\mathcal{F}_n]\big|_*^q\right]^{1/q}\right) = L^*\big(\|{}^{\mathrm{o}}\mathbb{D}f\|_{L^q(\mu)}\big).$$

Remark 5.16. The first-order optimal adversarial model ν_δ which approximates (5.11) for small δ is given explicitly in the proof as $\nu_\delta = (\mathrm{Id} + u\delta\Delta^{-1}\circ\Phi)_\#\mu$, where Φ is given in (5.27) and u is such that $\Upsilon = u\|{}^{\mathrm{o}}\mathbb{D}f\|_{L^q(\mu)} - L(u)$.

We now turn to the problem under martingale constraint. Let $\mu \in \mathscr{M}(\mathcal{X})$ and consider

$$V_{\mathrm{Mart}}(\delta) = \sup_{\nu\in\mathscr{M}(\mathcal{X})}\{E_\nu[f(X)] - L_\delta(\mathcal{AW}_p(\mu,\nu))\}.$$

Remark 5.17. We note that in the discrete-time setting the bi-causal and the causal penalizations are interchangeable by Proposition 5.35. We use different penalizations for $V(\delta)$ and $V_{\mathrm{Mart}}(\delta)$ to be consistent with their continuous-time counterparts.

Theorem 5.18. *Let $\mathbb{H}^q$ denote the set of predictable processes h with $\|h\|_{L^q(\mu)} < \infty$. Under Assumptions 5.13 and 5.14, we have*

$$V_{\mathrm{Mart}}(\delta) = V_{\mathrm{Mart}}(0) + \Upsilon_{\mathrm{Mart}}\delta + o(\delta),$$

where

$$\Upsilon_{\mathrm{Mart}} = L^*\left(\inf_{h\in\mathbb{H}^q} E_\mu\left[\sum_{n=1}^N \big|E_\mu[\mathbb{D}_n f(X)|\mathcal{F}_n] - h_n\big|_*^q\right]^{1/q}\right) = L^*\left(\inf_{h\in\mathbb{H}^q}\|{}^{\mathrm{o}}\mathbb{D}f - h\|_{L^q(\mu)}\right).$$

In particular, if $p=2$, we obtain

$$\Upsilon_{\mathrm{Mart}} = L^*\big(\|{}^{\mathrm{o}}\mathbb{D}f - {}^{\mathrm{p}}\mathbb{D}f\|_{L^2(\mu)}\big).$$

Example 5.19. In the context of derivatives pricing in mathematical finance, the AW-DRO sensitivity under martingale constraint can be viewed as a *nonparametric* Greek. It captures the sensitivity of the option price to model uncertainty. This was first observed in Bartl et al. (2021) in the context of perturbations of the distribution of the underlying stock price process at a given time (the maturity). Having derived sensitivities in a dynamic context, we can consider perturbations of the actual model for the price process.

We consider $d=1$ and a discrete-monitored Asian option whose payoff is given by

$$f(X) = \max\big\{0, \bar{X} - K\big\} \quad \text{with} \quad \bar{X} = \frac{1}{N+1}\sum_{n=0}^N X_n. \tag{5.13}$$

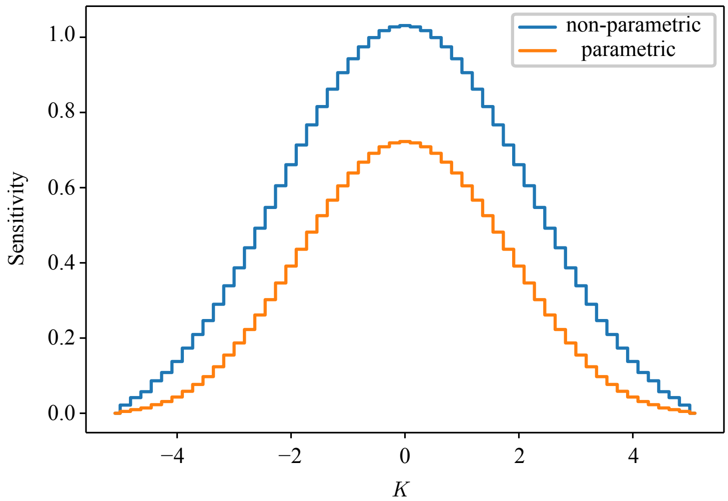

Figure 5.1: Comparison of the parametric sensitivity and the nonparametric sensitivity of the Asian option price under different strikes K.

Let $\mu \in \mathscr{M}(\mathcal{X})$ be the reference risk-neutral (pricing) measure. Without loss of generality, we assume μ is centered, otherwise we can always shift the market by a constant and absorb it into K. Notice that

$$\mathbb{D}_n f(X) = \frac{N+1-n}{N+1}\mathbb{1}_{\{\bar{X}\geq K\}}.$$

For simplicity, we take $p = 2$ and $L = +\infty\mathbb{1}_{(1,\infty)}$. By Theorem 5.18, we derive the nonparametric 'Greek' of the Asian option as

$$\begin{aligned}
\Upsilon_{\mathrm{Mart}} &= \left(E_\mu\left[\sum_{n=1}^{N}\left|E_\mu[\mathbb{D}_n f(X)|\mathcal{F}_n] - E_\mu[\mathbb{D}_n f(X)|\mathcal{F}_{n-1}]\right|^2\right]\right)^{1/2} \\
&= \left(E_\mu\left[\sum_{n=1}^{N}\frac{(N+1-n)^2}{(N+1)^2}\left|\mu(\bar{X}\geq K|\mathcal{F}_n) - \mu(\bar{X}\geq K|\mathcal{F}_{n-1})\right|^2\right]\right)^{1/2}. \qquad (5.14)
\end{aligned}$$

To compare this result with a parametric sensitivity, consider $\mu(\lambda)$ to be the distribution of a symmetric random walk with jump size $\pm\lambda$, and set the reference model as $\mu = \mu(1)$, the distribution of the simple symmetric random walk. In Figure 5.1, we compare the nonparametric sensitivity (5.14) with the parametric sensitivity to the jump size λ. Whilst the parametric sensitivity captures the main risk, the nonparametric one dominates it, as expected. Inspecting the first-order optimal adversarial model perturbation, see (5.27), reveals that it involves both the jump size and their symmetry being broken simultaneously.

5.4.2 Continuous-time results

We switch now to the continuous-time setting. As before, we state and discuss the main results, with proofs deferred to 5.6. We start with the hyperbolic scaling where the metric d is given by

$$d(\omega,\eta)=\limsup_{N\to\infty}\left(\frac{N}{T}\right)^{1-1/p}d_N(\omega,\eta)=\begin{cases}\|\omega-\eta\|_{W_0^{1,p}} & \text{if } \omega-\eta\in W_0^{1,p},\\ +\infty & \text{elsewhere.}\end{cases}$$

Assumption 5.20. Function $f:\mathcal{X}\to\mathbb{R}$ is in $\mathbb{D}_b^1$ and satisfies

- $|f(\omega)|\leq C(1+\|\omega\|_\infty^p)$, $\omega\in\mathcal{X}$;
- $\mathbb{D}_t f$ is continuous, and $|\mathbb{D}_t f(\omega)|\leq C(1+\|\omega\|_\infty^{p-1})$, $t\in[0,T]$.

Remark 5.21. We recall that $\mathbb{D}_b^1$ was introduced in Definition 5.7. Note that we do not expect the growth of $\mathbb{D}_t f$ to imply a control on the growth of f. This is because $\mathbb{D}_t f$ is only a directional derivative along a proper subspace of the tangent space.

Theorem 5.22. *Let $p>1$ and suppose Assumptions 5.13 and 5.20 hold, and that $\mu\in\mathscr{P}(\mathcal{X})$ satisfies $E_\mu[\sup_{t\in[0,T]}|X(t)|^p]<\infty$. Then, with $V(\delta)$ given in (5.11), we have*

$$V(\delta)=V(0)+\Upsilon\delta+o(\delta),$$

where

$$\Upsilon=\lim_{\delta\to0}\frac{1}{\delta}(V(\delta)-V(0))=L^*\big(\|{}^{o}\mathbb{D}f\|_{L^q(\mu)}\big).$$

Remark 5.23. We point out when the reference model is the Wiener measure γ and $f(X)\in\mathbf{D}^{1,2}$, the intertwining formula (Bally et al., 2016, Theorem 7.9) gives

$$E_\gamma[\mathbf{D}f(X)|\mathcal{F}_t]=\nabla_x E_\gamma[f(X)|\mathcal{F}_t]\quad \gamma\text{-a.s.},$$

where ∇_x denotes the vertical derivative. Moreover, from Proposition 5.9 the Malliavin derivative coincides with the *pointwise* Malliavin derivative on their common domain, and this leads to

$$E_\gamma[\mathbf{D}f(X)|\mathcal{F}_t]=E_\gamma[\mathbb{D}f(X)|\mathcal{F}_t]\quad \gamma\text{-a.s.}$$

Therefore, we can represent the sensitivity using the vertical derivative as $\Upsilon=L^*(\|\nabla_x E_\gamma[f|\mathcal{F}_t]\|_{L^q(\gamma)})$. One can extend the above arguments to more general martingale reference models by taking an appropriate metric d.

Example 5.24. We consider Merton's (Merton, 1969) classical setting of a log investor maximizing the expected utility of their terminal wealth. The stock price process follows the standard Black–Scholes model, with S solving

$$\mathrm{d}S(t) = \alpha S(t)\,\mathrm{d}t + \sigma S(t)\,\mathrm{d}X(t),$$

where X is a Brownian motion. The agent invests their wealth into the stock and a riskless asset which grows at a constant interest rate r. Suppose their initial wealth is κ and let $\theta(t)$ denote the *proportion* of their wealth invested in the risky asset at time t, which is assumed to be predictable. Their wealth process $(K^{\theta}(t))_{t\in I}$ evolves according to

$$\mathrm{d}K^{\theta}(t) = (r + \lambda\theta(t)\sigma)K^{\theta}(t)\,\mathrm{d}t + \sigma\theta(t)K^{\theta}(t)\,\mathrm{d}X(t),$$

where $\lambda = (\alpha - r)/\sigma$, known as the market price of risk, is the key market parameter the investor has to estimate. Merton's problem of maximizing $\mathbb{E}[\log(K^{\theta}(T))]$ over the choice of θ is solved taking $\theta(t) = \lambda/\sigma$, and the resulting wealth satisfies $K^*(T) = \kappa\exp((r + \lambda^2/2)T + \lambda X(T))$. Agent's expected utility is given by $V(0) = \mathbb{E}[\log(K^*(T))] = \log(\kappa)+(r+\lambda^2/2)T$ and its parametric sensitivity to λ, which captures the agent welfare's sensitivity to the estimated parameter, is $\frac{\partial}{\partial\lambda}V(0) = \lambda T$. The general sensitivity to model uncertainty, around μ the Wiener measure, can be computed using Theorem 5.22 for

$$f(X) = \log(K^*(T)) = \log(\kappa) + (r + \lambda^2/2)T + \lambda X_T.$$

Taking $p = 2$ and $L = +\infty\mathbb{1}_{(1,\infty)}$, we obtain $\Upsilon = \lambda\sqrt{T}$. We see that Υ recovers the parametric sensitivity but with a different scaling in time. Indeed, $\sqrt{T}$ is the natural Brownian scaling in time, and we kept the uncertainty penalty L independent of time. If instead, we set $L = +\infty\mathbb{1}_{(\sqrt{T},\infty)}$, introducing the natural Brownian scaling into the size of the uncertainty ball considered, then we obtain $\Upsilon = \lambda T$, as before. Naturally, the parametric sensitivity only makes sense in the specific context of Black–Scholes price dynamics, and Υ offers its natural nonparametric extension to general investment settings.

Example 5.25. We stress that our continuous-time sensitivity results can also go beyond the semi-martingale framework. Consider an objective given by a pathwise rough integral introduced in Cont and Perkowski (2019). We fix a positive integer l

and a sequence of partitions $\mathfrak{p} = \{p_1, p_2, \dots\}$ where $p_n = \{0 = t_0 < \cdots < t_{i_n} = T\}$. By $\mathbb{V}_{2l}(\mathfrak{p})$, we denote the set of paths with finite $2l$-variation along $\mathfrak{p}$ in the sense that

$$\lim_{n\to\infty} \sum_{[t_i,t_{i+1}]\in p_n, t_i\le t} |\omega(t_{i+1}) - \omega(t_i)|^{2l} = [\omega](t)^{2l}$$

for some continuous function $[\omega]^{2l}$. For any path $\omega \in \mathbb{V}_{2l}(\mathfrak{p})$ and $g \in C^{2l+1}$, we define the rough integral

$$\int_0^T g'(\omega(t)) \bullet \mathrm{d}\omega(t) := \lim_{n\to\infty} \sum_{[t_i,t_{i+1}]\in p_n} \sum_{k=1}^{2l-1} \frac{g^{(k)}(\omega(t_i))}{k!} (\omega(t_{i+1}) - \omega(t_i))^k.$$

Then we take $f(\omega) = \int_0^T g'(\omega(t)) \bullet \mathrm{d}\omega(t)$ and the reference measure as the law of the fractional Brownian motion with Hurst parameter $H = 1/2l$. By Cont and Perkowski (2019, Theorem 1.5), we notice

$$\begin{aligned} E_\mu[f(X)] &= E_\mu\left[g(X(T)) - g(X(0)) - \frac{1}{(2l)!}\int_0^T g^{(2l)}(X(s))\,\mathrm{d}[X]^{2l}(s)\right] \\ &= E_\mu\left[g(X(T)) - g(X(0)) - \frac{E_\mu[|X(1)|^{2l}]}{(2l)!}\int_0^T g^{(2l)}(X(s))\,\mathrm{d}s\right]. \end{aligned}$$

For simplicity, we take $d = 1$, $p = 2$, and $L = +\infty\mathbb{1}_{(1,\infty)}$. Therefore, by Theorem 5.22, we derive

$$\Upsilon = E_\mu\left[\int_0^T \left| E_\mu\left[g'(X(T)) - \frac{E_\mu[|X(1)|^{2l}]}{(2l)!}\int_t^T g^{(2l+1)}(X(s))\,\mathrm{d}s \,\middle|\, \mathcal{F}_t\right]\right|^2 \mathrm{d}t\right]^{1/2}.$$

As mentioned in the introduction, the hyperbolic scaling is not critical for the martingale constraint problem since, by Doob's decomposition theorem, the difference between two martingales is of infinite variation. This gives an infinite transport cost under hyperbolic scaling, i.e., martingale measures are a 'totally disconnected' set:

Remark 5.26. Let $p > 1$ and $\mu \in \mathscr{M}(\Omega)$ satisfy $E_\mu[\sup_{t\in[0,T]} |X(t)|^p] < \infty$. If a martingale measure $\nu \in \mathscr{M}(\Omega)$ satisfies $\mathcal{CW}_p(\mu, \nu) < \infty$, then $\mu = \nu$.

As a consequence, we have to zoom out the scaling. Put differently, we need a metric which allows us to alter the quadratic variation of the path. We focus on the case $p = 2$ and the reference measure μ being the Wiener measure, leaving the general case for future studies. Adopting a parabolic scaling, we formally set d as

$$d(\omega, \eta) = \limsup_{N\to\infty} d_N(\omega, \eta) = \sqrt{[\omega - \eta]_T}.$$

Remark 5.27. Consider formally taking continuous limit of Υ_{Mart} in Theorem 5.18 for $p = 2$:

$$\begin{aligned}
&\lim_{N\to\infty} L^*\left(E_\mu\left[\sum_{n=1}^{N}\left|E_\mu[\mathbf{D}_{nT/N}f(X)|\mathcal{F}_{nT/N}] - E_\mu[\mathbf{D}_{nT/N}f(X)|\mathcal{F}_{(n-1)T/N}]\right|_*^2\right]^{1/2}\right)\\
&\quad= \lim_{N\to\infty} L^*\left(E_\mu\left[\sum_{n=1}^{N}\int_{(n-1)T/N}^{nT/N} \|E_\mu[\mathbf{D}_t\mathbf{D}_{nT/N}f(X)|\mathcal{F}_t]\|_{\mathbf{F}}^2\,\mathrm{d}t\right]^{1/2}\right)\\
&\quad\approx L^*\left(E_\mu\left[\int_0^T \|E_\mu[\mathbf{D}_t\mathbf{D}_{t+}f(X)|\mathcal{F}_t]\|_{\mathbf{F}}^2\,\mathrm{d}t\right]^{1/2}\right),
\end{aligned}$$

where we apply the Clark–Ocone formula in the second line and denote the Frobenius norm by $\|\cdot\|_{\mathbf{F}}$. Surprisingly, this limit does not always coincide with Υ_{Mart} for some objectives of our interest such as $f(X) = \frac{1}{2}[X]_T$. If the above limit were true, we would have $\Upsilon_{\text{Mart}} = L^*(0) = 0$ since $\mathbf{D}_t f(X) = 0$ for any $t \in [0,T]$. However, a direct computation shows that $\Upsilon_{\text{Mart}} = L^*(\sqrt{T})$.

Motivated by the above remark, we consider objective functionals f of the form $f(\omega) = U\left(\int_0^T \langle h(t,\omega(\cdot\wedge t)), \mathrm{d}\omega(t)\rangle\right)$.

Assumption 5.28. We assume that $h : I \times \mathcal{X} \to \mathbb{R}^n$ is continuous, bounded, and $U \in C^2$ with a bounded second derivative. There exists $\gamma \in (1,2)$ such that under any $\pi \in \Pi_{\mathfrak{bc}}(\mu,*)$ with $\pi(\mathcal{X}\times\cdot) \in \mathscr{M}(\mathcal{X})$ the following holds:

- $E_\pi[\sup_{t\in[0,T]} |h(t,X(\cdot\wedge t)) - h(t,Y(\cdot\wedge t))|^2] = o(E_\pi[[X-Y]_T]^{1/2})$.
- $E_\pi\left[\sup_{t\in[0,T]} |h(t,Y(\cdot\wedge t)) - h(t,X(\cdot\wedge t)) - \int_0^t \mathbf{D}_s^{\intercal} h(t,X(\cdot\wedge t))\,\mathrm{d}^-(Y-X)(s)|^\gamma\right]$ $= o(E_\pi[[X-Y]_T]^{\gamma/2})$.
- $\{\mathbf{D}h(t,X(\cdot\wedge t))\}_{t\in[0,T]}$ is $(Y-X)$ uniformly forward–γ integrable.
- $\sup_{s,t\in[0,T]} \|\mathbf{D}_s h(t,X(\cdot\wedge t))\|_{\mathbf{F}} \in L^{2\gamma/(2-\gamma)}$.
- $\lim_{t\to s+} E_\pi[\|\mathbf{D}_s h(t,X(\cdot\wedge t)) - \mathbf{D}_t h(t,X(\cdot\wedge t))\|_{\mathbf{F}}^2] = 0$ for any $s \in [0,T]$.
- For $H = \int_0^T \langle h(t,X(\cdot\wedge t)), \mathrm{d}X(t)\rangle$, $\mathbf{D}_{s+}H = \lim_{t\to s+}\mathbf{D}_t H$ and $\mathbf{D}_{s+}\mathbf{D}_s H = \lim_{t\to s+}\mathbf{D}_t\mathbf{D}_s H$ both converge in L^2 for any $s\in[0,T]$.

Remark 5.29. We remark that d and f can be defined in a pathwise sense. By Karandikar (1995, Theorem 3) there exists a map $q : \mathcal{X} \to \mathbb{R}$ such that $q(\omega) = [\omega]_T$

holds almost surely under any martingale measure μ. This allows us to interpret $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ as $d(\omega, \eta) := q(\omega - \eta)$. In particular, under any bi-causal coupling π between martingale measures, we have the desired identity

$$d(X, Y) = q(X - Y) = [X - Y]_T \quad \pi\text{-a.s.}$$

Similarly, under Assumption 5.28, h is continuous and by Karandikar (1995, Theorem 3) there exists $f : \mathcal{X} \to \mathbb{R}$ such that for any martingale measure μ

$$f(X) = U\left(\int_0^T \langle h(t, X(\cdot \wedge t)), \mathrm{d}X(t)\rangle\right) \quad \mu\text{-a.s.}$$

Theorem 5.30. *Take $p = 2$ and let μ be the Wiener measure. Let Assumptions 5.13 and 5.28 hold, and denote $H = \int_0^T \langle h(t, X(\cdot \wedge t)), \mathrm{d}X(t)\rangle$. Recall*

$$V_{\mathrm{Mart}}(\delta) = \sup_{\nu \in \mathscr{M}(\mathcal{X})} \{E_\nu[U(H)] - L_\delta(\mathcal{AW}_2(\mu, \nu))\}. \tag{5.15}$$

Then, we have

$$V_{\mathrm{Mart}}(\delta) = V_{\mathrm{Mart}}(0) + \Upsilon_{\mathrm{Mart}}\delta + o(\delta),$$

where

$$\Upsilon_{\mathrm{Mart}} = L^*\left(E_\mu\left[\int_0^T \|\varphi(t)\|_{\mathrm{F}}^2 \,\mathrm{d}t\right]^{1/2}\right) \tag{5.16}$$

and

$$\varphi(s) = {}^{\mathrm{p}}\{\mathbf{D}_{s+}\mathbf{D}_s U(H) - U'(H)\mathbf{D}_s^{\intercal} h(s, X(\cdot \wedge s))\}. \tag{5.17}$$

Example 5.31. Let $\sigma \in C_b^{1,2}$. Then $h(t, \omega) = \sigma(t, \omega(t))$ satisfies Assumption 5.28. The first condition follows from BDG inequality and the boundedness of $\partial_x\sigma$. For the second one, we notice that $\mathbf{D}_s h(t, X(\cdot \wedge t)) = \partial_x\sigma(t, X(t))\mathbb{1}_{s \le t}$. The estimate follows from Taylor expansion and BDG inequality by noticing

$$\begin{aligned}&\left|h(t, Y(\cdot \wedge t)) - h(t, X(\cdot \wedge t)) - \int_0^t \mathbf{D}_s^{\intercal} h(t, X(\cdot \wedge t))\,\mathrm{d}^-(Y - X)(s)\right| \\ &\qquad = |\sigma(t, Y(t)) - \sigma(t, X(t)) - \langle \partial_x\sigma(t, X(t)), (Y(t) - X(t))\rangle| \le C|Y(t) - X(t)|^{1+\varepsilon},\end{aligned}$$

for some $\varepsilon \in (0, 1)$. Taking $\gamma = 2/(1 + \varepsilon)$, we obtain the required estimate. For the third condition, we notice $\int_0^T \mathbf{D}_s h(t, X(\cdot \wedge t))\,\mathrm{d}^-(Y - X)(s) = \partial_x\sigma(t, X(t))(Y(t) - X(t))$ and uniform forward integrablity follows from the boundedness of $\partial_x\sigma$. The fourth condition is again satisfied by the boundedness of $\partial_x\sigma$. The fifth condition is a consequence of the continuity and the boundedness of $\partial_x\sigma$. The last condition follows from the fact that $\mathbf{D}_t H = \sigma(t, X(t)) + \int_t^T \partial_x\sigma(r, X(r))\,\mathrm{d}X(r)$ and $\mathbf{D}_t\mathbf{D}_s H = \partial_x\sigma(t, X(t)) + \int_t^T \partial_x^2\sigma(r, X(r))\,\mathrm{d}X(r)$ for $t \ge s$.

Example 5.32. We continue the theme of Example 5.19 and explore a nonparametric Greek for option pricing in a continuous time setting. Suppose the price process follows

$$\mathrm{d}S(t) = S(t)\sigma(t, X(t))\,\mathrm{d}X(t),$$

where X is a Brownian motion under the reference measure μ and $\sigma \in C_b^{1,2}$. The payoff of a log contract is given by $-\log(S(T)/S(0))$ and Itô's formula gives its price as

$$E_\mu[-\log(S(T)/S(0))] = E_\mu\left[\frac{1}{2}\int_0^T \sigma(t, X(t))^2\,\mathrm{d}[X](t)\right] = E_\mu\left[\frac{1}{2}\left(\int_0^T \sigma(t, X(t))\,\mathrm{d}X(t)\right)^2\right].$$

We take $L = +\infty\mathbb{1}_{(1,\infty)}$ and applying Theorem 5.30, noting its assumptions are satisfied by Example 5.31, calculate the nonparametric 'Greek' of a log contract as

$$\Upsilon_{\mathrm{Mart}} = E_\mu\left[\int_0^T |\varphi(t)|^2\,\mathrm{d}t\right]^{1/2},$$

where

$$\begin{aligned}\varphi(s) &= {}^{\mathrm{p}}\left\{\mathbf{D}_{s+}\mathbf{D}_s\left(\frac{1}{2}\left(\int_0^T \sigma(t, X(t))\,\mathrm{d}X(t)\right)^2\right) - \int_0^T \sigma(t, X(t))\,\mathrm{d}X(t)\partial_x\sigma(s, X(s))\right\}\\ &= \sigma(s, X(s))^2 + \int_s^T E_\mu\left[(\partial_x\sigma(t, X(t)))^2 + \sigma(t, X(t))\partial_x^2\sigma(t, X(t))\middle|\mathcal{F}_s\right]\mathrm{d}t.\end{aligned}$$

If σ is constant then the price of the log contract is given by $p(\sigma) := \frac{1}{2}\sigma^2 T$ while the above formula gives $\Upsilon_{\mathrm{Mart}} = \sigma^2\sqrt{T}$. The difference in scaling in time comes from two factors: one is similar to what we saw in Example 5.24 above, the other is due σ being the annualized volatility, while Υ_{Mart} is not annualized. To wit, note that a δ perturbation of the underlying Brownian motion, X to $(1+\delta)X$ corresponds to $\sqrt{T}\delta$ causal perturbation whilst changing the log contract price from $p(\sigma)$ to $p(\sigma(1+\delta))$. Thus, the sensitivity to log contract price corresponding to Υ_{Mart} is

$$\lim_{\delta\to 0}\frac{1}{2}\frac{\sigma^2(1+\delta/\sqrt{T})^2 T - \sigma^2 T}{\delta} = \sigma^2\sqrt{T}.$$

In general, taking time scaling into the account, we would expect $p'(\sigma) \leq \frac{\sqrt{T}}{\sigma}\Upsilon_{\mathrm{Mart}}$ and the fact that we actually have equality shows that, up to the first order, the worst adversarial change to the dynamics comes from a constant shift to σ. Finally, we note that, as in Example 5.24, we could take $L = +\infty\mathbb{1}_{(\sqrt{T}/\sigma,\infty)}$ leading to $\Upsilon_{\mathrm{Mart}} = \sigma T = p'(\sigma)$.

5.4.3 Extensions and further results

Filtrations in discrete time

In Theorem 5.15, we work with the canonical filtration $\mathbf{F}$. This was taken for ease of notation and to mimic the continuous time results. We can extend our result to any enlarged filtration $\widetilde{\mathbf{F}}$, and following the same lines of arguments the sensitivity is again given by $\Upsilon = L^*\big(\| {}^{o}\mathbb{D}f\|_{L^q(\mu)}\big)$, where the optional projection is with respect to $\widetilde{\mathbf{F}}$. The notion of causality introduced in Section 5.2 is easily transferred to this setting, see Acciaio et al. (2020) for more details. In particular, if we equip the reference model with the largest filtration, i.e., $\tilde{\mathcal{F}}_t = \mathcal{F}$ for any $t \in I$, then any coupling is a causal coupling. Under such an extended setting, we retrieve the static Wasserstein sensitivity (Bartl et al., 2021) as a specific corollary of Theorem 5.15.

Weak OT objective in discrete time

Another extension is to consider an optimal stopping problem

$$V_{\mathrm{OS}}(0) = \sup_{\tau \in \mathscr{T}} E_\mu[g(X_\tau)],$$

and its distributionally robust counterpart

$$V_{\mathrm{OS}}(\delta) = \sup_{\nu \in \mathscr{P}(\mathcal{X})} \left\{ \sup_{\tau \in \mathscr{T}} E_\nu[g(X_\tau)] - L_\delta(\mathcal{AW}_p(\mu,\nu)) \right\},$$

where $\mathscr{T}$ is the set of $(\mathcal{F}_t)$–stopping time. The dual representation for $V_{\mathrm{OS}}(\delta)$ is discussed in Jiang (2024). For simplicity, we focus on the two-period case, and by Snell's envelope we notice

$$\sup_{\tau \in \mathscr{T}} E_\nu[g(X_\tau)] = E_\nu[\max\{g(X_1), E_\nu[g(X_2)|X_1]\}] = E_\nu[f(X_1, \mathrm{Law}(X_2|X_1))].$$

In this case, the objective $f : \mathcal{X}_1 \times \mathscr{P}(\mathcal{X}_2) \to \mathbb{R}$ is a functional not only of the state but also of the conditional law of the state. Such problems are referred to as Weak OT and were studied in Gozlan et al. (2017). For this two-period setting, it can be shown that

$$\Upsilon_{\mathrm{OS}} := \lim_{\delta \to 0} (V_{\mathrm{OS}}(\delta) - V_{\mathrm{OS}}(0))/\delta = L^*\big(\|(\partial_x f, \partial_\mu f)\|_{L^q(\mu)}\big),$$

where ∂_μ is Lion's derivative, suitably adjusted to the change of coordinates by Δ. However, to the best of our knowledge, the general N-period case remains open.

Semi-martingale ambiguity in continuous time

Our results can be extended to a framework where the ambiguity set is the set of semi-martingales around the reference model. In order to allow both the drift and the volatility ambiguity, we need to decompose the semi-martingale into its finite-variation and its martingale part. For a process X, we write $X = X^a + X^m$ for its Doob's decomposition. We adopt the objective in Theorem 5.18:

$$f(X) = U(H), \quad \text{where} \quad H = \int_0^T \langle h(t, X(\cdot \wedge t)), \mathrm{d}X(t)\rangle.$$

Let $d = 1$, $p = 2$, μ be the Wiener measure and consider the DRO problem given by

$$V_{\mathrm{S.Mart}}(\delta) = \sup_{\nu \in B_\delta} E_\nu[U(H)],$$

where $B_\delta = B_\delta^a \cap B_\delta^m$ is the intersection of two bi-causal balls given by

$$B_\delta^a = \left\{\nu \in \mathscr{S}(\mathcal{X}) : \inf_{\pi \in \Pi_{\mathfrak{bc}}(\mu,\nu)} E_\pi\left[\|X^a - Y^a\|_{W_0^{1,2}}\right] \leq \delta\right\}$$

and

$$B_\delta^m = \left\{\nu \in \mathscr{S}(\mathcal{X}) : \inf_{\pi \in \Pi_{\mathfrak{bc}}(\mu,\nu)} E_\pi[[X^m - Y^m]_T]^{1/2} \leq \delta\right\}.$$

Here $\mathscr{S}(\mathcal{X})$ denotes the set of semi-martingale measures. In order to apply Theorem 5.22, a pathwise definition of the Itô integral is needed. For instance, we can adapt the approach in Example 5.25, and consider the objective of the form of

$$f(\omega) = U\left(\int_0^T g'(\omega(t))\,\mathrm{d}\omega(t)\right) = U\left(g(\omega_T) - g(0) - \frac{1}{2}\int_0^T g''(\omega(s))\,\mathrm{d}[\omega](s)\right).$$

We claim that

$$\begin{aligned}
&E_\pi[f(Y) - f(X)] \\
&= E_\pi[f(X + (Y-X)^a + (Y-X)^m) - f(X)] \\
&= E_\pi[f(X + (Y-X)^a) - f(X)] + E_\pi[f(X + (Y-X)^m) - f(X)] + o(\delta) \\
&= E_\pi\left[\int_0^T \langle {}^{o}\mathbb{D}f(t)(X), \mathrm{d}(Y-X)(t)^a\rangle\right] + E_\pi\left[\int_0^T \langle \varphi(t), \mathrm{d}[\![X, (Y-X)^m]\!](t)\rangle\right] + o(\delta),
\end{aligned} \tag{5.18}$$

where φ is given in (5.17), and where the second equality follows by controlling the higher order terms combining the arguments in the proofs of Theorems 5.22 and 5.30, and we omit the details. It follows that

$$\Upsilon_{\mathrm{S.Mart}} = \Upsilon + \Upsilon_{\mathrm{Mart}}. \tag{5.19}$$

This result allows us to detail the relation between our results and the recent work of Bartl et al. (2025a). Therein, the authors consider L^p-balls around the drift and the volatility of the reference model under a strong formulation. We adapted their results to the semi-martingale framework as follows. Let $(\Omega, \mathcal{F}, (\mathcal{F}_t), P)$ be a probability space supporting a Brownian motion W. Consider a DRO problem given by

$$\widetilde{V}_{\mathrm{S.Mart}}(\delta) = \inf_{h \in \mathbb{H}} \sup_{(b,\sigma) \in \widetilde{B}_\delta} E_P\Big[U\Big(\int_0^T h(t)\,\mathrm{d}X^{b,\sigma}(t)\Big)\Big],$$

where $\mathbb{H}$ is a set of predictable open-loop controls, $X(t)^{b,\sigma} = \int_0^t b(s)\,\mathrm{d}s + \int_0^t \sigma(s)\,\mathrm{d}W(s)$, and

$$\widetilde{B}_\delta = \left\{(b,\sigma) : E_P\Big[\int_0^T |b(t) - \bar{b}(t)|^p\,\mathrm{d}t\Big]^{1/p} \leq \delta, E_P\Big[\Big(\int_0^T |\sigma(t) - \bar{\sigma}(t)|^2\,\mathrm{d}t\Big)^{p/2}\Big]^{1/p} \leq \delta\right\},$$

thus an intersection of L^p-balls around the drift and volatility coefficients. Bartl et al. (2025b) shows that

$$\widetilde{\Upsilon}_{\mathrm{S.Mart}} = E_P\Big[\Big(\int_0^T |Y(t)h^*(t)|_*^q\,\mathrm{d}t\Big)\Big]^{1/q} + E_P\Big[\Big(\int_0^T |Z(t)h^*(t)|^2\,\mathrm{d}t\Big)^{q/2}\Big]^{1/q}, \quad (5.20)$$

where h^* is the unique optimal control of the reference model and (Y, Z) is the solution to the BSDE:

$$Y(t) = U'\Big(\int_0^T h^*(t)\,\mathrm{d}X(t)^{\bar{b},\bar{\sigma}}\Big) - \int_t^T Z(s)\,\mathrm{d}W(s), \quad t \in [0,T].$$

Notice that in our framework we do not specify a probability space, and hence we only consider feedback (closed-loop) controls h. The intersection between the two settings is obtained taking $\bar{b} = 0$, $\bar{\sigma} = \mathrm{Id}$ and $\mathbb{H} = \{h\}$ for some deterministic control h. While (5.20) was obtained only for $p > 3$, we find that (5.20) coincides with (5.19) by plugging $p = 2$. Roughly speaking, this indicates that the volatility ball $\widetilde{B}_\delta$, while more rigid, is actually equivalent to the bi-causal ball B_δ up to the first order approximation.

Remark 5.33. Above, we took an intersection of a drift ball and a volatility ball to obtain a direct comparison with Bartl and Wiesel (2023). However, from the causal-OT point of view, it is more natural to allow a trade-off between the two types of perturbations and to combine them into one cost function. This leads us to consider the bi-causal discrepancy given by

$$\mathcal{AW}_2(\mu,\nu) := \inf_{\pi \in \Pi_{\mathfrak{bc}}(\mu,\nu)} E_\pi\Big[\|X^a - Y^a\|^2_{W_0^{1,2}} + [X^m - Y^m]_T\Big]^{1/2}.$$

The corresponding DRO problem is given by

$$V_{\mathrm{S.Mart}}(\delta) = \sup_{\nu \in \mathscr{S}(\mathcal{X})} \{E_\nu[U(H)] - L_\delta(\mathcal{AW}_2(\mu, \nu))\}.$$

For simplicity, consider $L = +\infty \mathbb{1}_{(1,\infty)}$ as above. Combing the first order approximation (5.18) with the Cauchy inequality, in this setting one can derive

$$\Upsilon_{\mathrm{S.Mart}} = \sqrt{\Upsilon^2 + \Upsilon^2_{\mathrm{Mart}}}.$$

5.4.4 A general strategy of the proof

We present a key elementary lemma which provides a unified framework for the proofs. An upper bound of the sensitivity is given by estimates (5.21) and (5.22). A lower bound of the sensitivity is a consequence of the estimate (5.23). In the following proofs, we will verify all three estimates respectively in each case. The growth estimate (5.21) will follow from the growth assumption of f. Estimate (5.22) is an asymptotic estimate when the cost is small and is will be derived from the Hölder inequality or the Kunita–Watanabe inequality. For estimate (5.23), we will construct a sequence of couplings which attain the inequality in (5.22) asymptotically.

Lemma 5.34. *Let $\mathcal{P} \subseteq \Pi(\mu, *)$ be a given set of couplings and d a metric such that the following conditions hold: there exists a constant C such that for any $\pi \in \mathcal{P}$*

$$E_\pi[f(Y)] \leq C(1 + E_\pi[d(X,Y)^p]), \tag{5.21}$$

there exists r such that for any $\pi \in \mathcal{P}$

$$E_\pi[f(Y) - f(X)] \leq r E_\pi[d(X,Y)^p]^{1/p} + o(E_\pi[d(X,Y)^p]^{1/p}), \tag{5.22}$$

and for all $\delta > 0$ small enough, there exists $\pi_\delta \in \mathcal{P}$ such that $E_{\pi_\delta}[d(X,Y)^p] \leq u^p \delta^p$ and

$$E_{\pi_\delta}[f(Y) - f(X)] \geq ru\delta + o(\delta), \tag{5.23}$$

where u is given by $ur - L(u) = L^*(r)$. Then under Assumption 5.13, we have

$$\sup_{\pi \in \mathcal{P}} \big\{E_\pi[f(Y) - f(X)] - L_\delta(E_\pi[d(X,Y)^p]^{1/p})\big\} = L^*(r)\delta + o(\delta).$$

Proof. Since L satisfies $\liminf_{u \to \infty} \frac{L(u)}{u^p} = +\infty$, there exists M_1 and $C_1 > C$ such that $L_\delta(u) > C_1 \delta^{1-p} u^p$ for any $\delta < 1$ and $u > M_1$. Combined with (5.21), we see that for any $\delta < 1$ we can restrict to measures with uniformly bounded costs

$$\begin{aligned} &\sup_{\pi \in \mathcal{P}} \big\{E_\pi[f(Y) - f(X)] - L_\delta(E_\pi[d(X,Y)^p]^{1/p})\big\} \\ &\quad = \sup_{\pi \in \mathcal{Q}} \big\{E_\pi[f(Y) - f(X)] - L_\delta(E_\pi[d(X,Y)^p]^{1/p})\big\}, \end{aligned} \tag{5.24}$$

where $\mathcal{Q} = \mathcal{P} \cap \{\pi : E_\pi[d(X,Y)^p] \le M_1\}$. By (5.22), for $\pi \in \mathcal{Q}$ we have

$$E_\pi[f(Y) - f(X)] \le C_2 E_\pi[d(X,Y)^p]^{1/p},$$

for some constant C_2. Again by the growth assumption of L, there exists $M_2 > 1$ such that $L_\delta(u) > C_2\delta^{1-p}u^p$ for any $\delta < 1$ and $u > M_2$. Then on the set of $\{\pi \in \mathcal{Q} : E_\pi[d(X,Y)^p] > M_2^p\delta^p\}$ it holds that

$$\begin{aligned} E_\pi[f(Y) - f(X)] - L_\delta(E_\pi[d(X,Y)^p]^{1/p}) &\le C_2 E_\pi[d(X,Y)^p]^{1/p} - C_2\delta^{1-p}E_\pi[d(X,Y)^p] \\ &\le C_2 E_\pi[d(X,Y)^p]^{1/p}[1 - (E_\pi[d(X,Y)^p]^{1/p}/\delta)^{p-1}] < 0. \end{aligned}$$

Therefore, by taking $\mathcal{P}_\delta = \{\pi \in \mathcal{Q} : E_\pi[d(X,Y)^p] \le M_2^p\delta^p\}$, we obtain the desired estimate from (5.22)

$$\begin{aligned} &\sup_{\pi\in\mathcal{P}}\big\{E_\pi[f(Y) - f(X)] - L_\delta(E_\pi[d(X,Y)^p]^{1/p})\big\} \\ &\quad = \sup_{\pi\in\mathcal{P}_\delta}\big\{E_\pi[f(Y) - f(X)] - L_\delta(E_\pi[d(X,Y)^p]^{1/p})\big\} \\ &\quad\quad \le \sup_{\pi\in\mathcal{P}_\delta}\big\{rE_\pi[d(X,Y)^p]^{1/p} - \delta L(E_\pi[d(X,Y)^p]^{1/p}/\delta) + o(\delta)\big\} \le L^*(r)\delta + o(\delta). \end{aligned}$$

On the other hand, taking u such that $ur - L(u) = L^*(r)$ and π_δ in (5.23), we obtain the other inequality, and hence the desired equality:

$$\sup_{\pi\in\mathcal{P}_\delta}\big\{E_\pi[f(Y) - f(X)] - L_\delta(E_\pi[d(X,Y)^p]^{1/p})\big\} \ge ru\delta - \delta L(u) + o(\delta) = \delta L^*(r) + o(\delta).$$

□

5.5 Proofs of discrete-time results

Proof of Theorem 5.15. It is clear by definition that

$$V(\delta) = \sup_{\pi\in\Pi_{\mathfrak{c}}(\mu,*)}\big\{E_\pi[f(Y)] - L_\delta(E_\pi[d_N(X,Y)^p]^{1/p})\big\}.$$

By Lemma 5.34 it suffices to verify (5.21), (5.22) and (5.23) with $\mathcal{P} = \Pi_{\mathfrak{c}}(\mu,*)$ and

$$r = E_\mu\left[\sum_{n=1}^N \big|E_\mu[\mathbb{D}_n f(X)|\mathcal{F}_n]\big|_*^q\right]^{1/q} = \|\,{}^o\mathbb{D}f\|_{L^q(\mu)}.$$

We derive (5.21) by noticing $f(y) \le C(1 + |x|^p + |x-y|^p)$. Let $\pi \in \Pi_{\mathfrak{c}}(\mu, *)$, and we notice by Hölder inequality

$$\begin{aligned}E_\pi[f(Y) - f(X)] &= \int_0^1 E_\pi[\langle \mathbb{D} f(X + \lambda(Y - X)), \Delta Y - \Delta X\rangle] \, \mathrm{d}\lambda \\ &= \int_0^1 E_\pi[\langle {}^{\mathrm{o}}\mathbb{D} f(X + \lambda(Y - X)), \Delta Y - \Delta X\rangle] \, \mathrm{d}\lambda \\ &\le E_\pi[d_N(X,Y)^p]^{1/p} \int_0^1 \|{}^{\mathrm{o}}\mathbb{D} f(X + \lambda(Y - X))\|_{L^q(\pi)} \, \mathrm{d}\lambda.\end{aligned}$$

In order to verify (5.22), we need to show that for any π_n with $\lim_{n\to\infty} E_{\pi_n}[d_N(X,Y)^p] = 0$, it holds that

$$\limsup_{n\to\infty} \int_0^1 \|{}^{\mathrm{o}}\mathbb{D} f(X + \lambda(Y - X))\|_{L^q(\pi_n)} \, \mathrm{d}\lambda \le \|{}^{\mathrm{o}}\mathbb{D} f(X)\|_{L^q(\mu)}. \tag{5.25}$$

Notice that $\lim_{n\to\infty} E_{\pi_n}[d_N(X,Y)^p] = 0$ implies the convergence of π_n to the identical coupling $\tilde{\pi} = (\mathrm{Id}, \mathrm{Id})_{\#}\mu$ in p-Wasserstein distance. Therefore, by Assumption 5.14 and Jensen inequality, for any $\lambda \in [0,1]$, we obtain

$$\begin{aligned}&\limsup_{n\to\infty} \|{}^{\mathrm{o}}\mathbb{D} f(X + \lambda(Y - X)) - {}^{\mathrm{o}}\mathbb{D} f(X)\|_{L^q(\pi_n)} \\ &\quad\le \limsup_{n\to\infty} \|\mathbb{D} f(X + \lambda(Y - X)) - \mathbb{D} f(X)\|_{L^q(\pi_n)} \\ &\quad= \|\mathbb{D} f(X + \lambda(Y - X)) - \mathbb{D} f(X)\|_{L^q(\tilde{\pi})} = 0.\end{aligned}$$

Assumption 5.14 and dominated convergence theorem now give the desired inequality:

$$\begin{aligned}\limsup_{n\to\infty} \int_0^1 \|{}^{\mathrm{o}}\mathbb{D} f(X + \lambda(Y - X))\|_{L^q(\pi_n)} \, \mathrm{d}\lambda &\le \limsup_{n\to\infty} \|{}^{\mathrm{o}}\mathbb{D} f(X)\|_{L^q(\pi_n)} \\ &= \|{}^{\mathrm{o}}\mathbb{D} f\|_{L^q(\mu)},\end{aligned}$$

where in the last equality, we used the fact that $E_{\pi_n}[\mathbb{D}_n f(X) | \mathcal{F}_n \otimes \mathcal{G}_n] = E_\mu[\mathbb{D}_n f(X)|\mathcal{F}_n]$ since $\pi_n \in \Pi_{\mathfrak{c}}(\mu, *)$.

It remains to verify (5.23). We introduce $v : \mathbb{R}^n \to \mathbb{R}^n$ given by

$$v(e) = (e_1|e_1|^{q-2}, \ldots, e_d|e_d|^{q-2}). \tag{5.26}$$

We fix $u > 0$ and let $\pi_\delta = (\mathrm{Id}, \mathrm{Id} + u\delta\Delta^{-1} \circ \Phi)_{\#}\mu$ where

$$\Phi_n(X) = v(E_\mu[\mathbb{D}_n f(X)|\mathcal{F}_n]), \text{ for } n = 1, \ldots, N. \tag{5.27}$$

By construction, $\pi_\delta \in \Pi_{\mathfrak{c}}(\mu, *)$. We compute, using $pq - p = q$,

$$E_{\pi_\delta}[d_N(X,Y)^p] = u^p \delta^p E_\mu\left[\sum_{n=1}^N |\Phi_n(X)|^p\right] = \left(u\|{}^{\mathrm{o}}\mathbb{D} f\|_{L^q(\mu)}^{q/p}\right)^p \delta^p.$$

On the other hand, by the fundamental theorem of calculus and the definition of $\mathbb{D}$ in (5.8), we have

$$\begin{aligned} E_{\pi_\delta}[f(Y) - f(X)] &= E_\mu[f(X + u\delta\Delta^{-1} \circ \Phi(X)) - f(X)] \\ &= u\delta \int_0^1 E_\mu[\langle \mathbb{D} f(X + \lambda u\delta\Delta^{-1} \circ \Phi(X)), \Phi(X)\rangle] \, \mathrm{d}\lambda. \end{aligned}$$

Using the assumed growth estimates we can apply dominated convergence theorem, and taking conditional expectations, we see that the term under the integral, for any fixed λ, converges to $\| {}^{\mathrm{o}}\mathbb{D} f\|^q_{L^q(\mu)}$ as $\delta \to 0$, and hence

$$E_{\pi_\delta}[f(Y) - f(X)] = \| {}^{\mathrm{o}}\mathbb{D} f\|_{L^q(\mu)} \left(u \| {}^{\mathrm{o}}\mathbb{D} f\|^{q/p}_{L^q(\mu)} \right) \delta + o(\delta).$$

If $r = \| {}^{\mathrm{o}}\mathbb{D} f\|_{L^q(\mu)} > 0$ then above estimate is equivalent to (5.23), where we take u such that $L^*(r) = ur - L(u)$. If $\| {}^{\mathrm{o}}\mathbb{D} f\|_{L^q(\mu)} = 0$, then the first part has already implied the sensitivity $\Upsilon = L^*(0) = 0$. $\square$

In discrete-time setting, (Bartl and Wiesel, 2023, Lemma 3.1) shows bi-causal couplings $\Pi_{\mathfrak{bc}}(\mu, *)$ are dense in the set of causal couplings $\Pi_{\mathfrak{c}}(\mu, *)$. The proof immediately adapts if instead of couplings we consider transport maps. Here, we observe that both also extend to the setting under a martingale constraint. We state the results for maps as this is the version we need for the proof of Theorem 5.18. The proof is deferred to Section 5.7.

Proposition 5.35. *Let $\mathcal{X} = \{0\} \times (\mathbb{R}^n)^N$, $\mu \in \mathscr{P}(\mathcal{X})$ and $\pi = (\mathrm{Id}, \Phi)_\# \mu \in \Pi_{\mathfrak{c}}(\mu, *)$. Then for any $\varepsilon > 0$, there exists Φ^ε such that $\|\Phi - \Phi^\varepsilon\|_\infty < \varepsilon$ and $(\mathrm{Id}, \Phi^\varepsilon)_\# \mu \in \Pi_{\mathfrak{bc}}(\mu, *)$. Moreover, if $\mu, \Phi_\# \mu \in \mathscr{M}(\mathcal{X})$ then Φ^ε can be taken such that $\Phi^\varepsilon_\# \mu \in \mathscr{M}(\mathcal{X})$.*

Proof of Theorem 5.18. Notice that

$$V_{\mathrm{Mart}}(\delta) = \sup_{\pi \in \Pi_{\mathfrak{bc}}(\mu,*), \pi(\mathcal{X},\cdot) \in \mathscr{M}(\mathcal{X})} \{E_\pi[f(Y)] - L_\delta(E_\pi[d_N(X, Y)])\}.$$

By Lemma 5.34, it suffices to verify (5.21), (5.22) and (5.23) with $\mathcal{P} = \{\pi \in \Pi_{\mathfrak{bc}}(\mu, *) : \pi(\mathcal{X} \times \cdot) \in \mathscr{M}(\mathcal{X})\}$ and

$$r = \inf_{h \in \mathbb{H}^q} E_\mu \left[\sum_{n=1}^N \left| E_\mu[\mathbb{D}_n f | \mathcal{F}_n] - h_n \right|_*^q \right]^{1/q} = \inf_{h \in \mathbb{H}^q} \| {}^{\mathrm{o}}\mathbb{D} f - h\|_{L^q(\mu)},$$

where $\mathbb{H}^q$ is the set of predictable functionals in L^q. Estimate (5.21) follows directly from the unconstrained case. Let $\pi \in \mathcal{P}$ and $h \in \mathbb{H}^q \cap C^2_b$. We notice by Hölder inequality

$$\begin{aligned}
E_\pi[f(Y)-f(X)] &= \int_0^1 E_\pi[\langle \mathbb{D}f(X+\lambda(Y-X)), \Delta Y - \Delta X\rangle]\,\mathrm{d}\lambda \\
&= \int_0^1 E_\pi[\langle {}^{\mathrm{o}}\mathbb{D}f(X+\lambda(Y-X)) - h(X), \Delta Y - \Delta X\rangle]\,\mathrm{d}\lambda \\
&\le E_\pi[d_N(X,Y)^p]^{1/p} \int_0^1 \|{}^{\mathrm{o}}\mathbb{D}f(X+\lambda(Y-X)) - h(X)\|_{L^q(\pi)}\,\mathrm{d}\lambda.
\end{aligned}$$

Here, the second equality follows from $\pi \in \Pi_{\mathfrak{bc}}(\mu,*)$ and Proposition 5.37. Following the arguments used in the proof of Theorem 5.15, noting h is bounded, we derive

$$E_\pi[f(Y)-f(X)] \le E_\pi[d_N(X,Y)^p]^{1/p}\|{}^{\mathrm{o}}\mathbb{D}f - h\|_{L^q(\mu)} + o(E_\pi[d_N(X,Y)^p]^{1/p}).$$

We verify (5.22) by noticing

$$\inf_{h\in\mathbb{H}^q\cap C_b^2}\|{}^{\mathrm{o}}\mathbb{D}f - h\|_{L^q(\mu)} = \inf_{h\in\mathbb{H}^q}\|{}^{\mathrm{o}}\mathbb{D}f - h\|_{L^q(\mu)}.$$

We turn to showing (5.23). For h^* we denote the L^q predictable projection of ${}^{\mathrm{o}}\mathbb{D}f$, i.e., the unique optimizer of

$$\inf_{h\in\mathbb{H}} J[h] = \inf_{h\in\mathbb{H}}\|{}^{\mathrm{o}}\mathbb{D}f - h\|^q_{L^q(\mu)}. \tag{5.28}$$

We use v as defined in (5.26) and define $\Phi = (0,\Phi_1,\dots,\Phi_N) : \mathcal{X}\to\mathcal{X}$ where

$$\Phi_n(X) = v(E_\mu[\mathbb{D}_n f(X)|\mathcal{F}_n] - h_n^*) \text{ for } n = 1,\dots,N.$$

The first variation of (5.28) yields for any $g\in\mathbb{H}^q$

$$\delta J[h^*](g) = \lim_{\varepsilon\to 0}\frac{1}{\varepsilon}(J[h^*+\varepsilon g] - J[h^*]) = qE_\mu[\langle\Phi(X), g\rangle] \ge 0,$$

which implies that

$$E_\mu[\Phi_n(X)|\mathcal{F}_{n-1}] = 0. \tag{5.29}$$

We fix $u>0$ and let $\pi_\delta = (\mathrm{Id}, \mathrm{Id} + u\delta\Delta^{-1}\circ\Phi(X))_\#\mu$. A direct computation yields

$$E_{\pi_\delta}[d_N(X,Y)^p] = \left(u\|{}^{\mathrm{o}}\mathbb{D}f - h^*\|^{q/p}_{L^q(\mu)}\right)^p\delta^p.$$

On the other hand, we have

$$\begin{aligned}
E_{\pi_\delta}[f(Y)-f(X)] &= E_\mu[f(X+u\delta\Delta^{-1}\circ\Phi(X)) - f(X)] \\
&= u\delta\int_0^1 E_\mu[\langle\mathbb{D}f(X+\lambda u\delta\Delta^{-1}\circ\Phi(X)), \Phi(X)\rangle]\,\mathrm{d}\lambda.
\end{aligned}$$

Since f satisfies Assumption 5.14, the dominated convergence theorem gives

$$\begin{aligned}&\lim_{\delta\to 0}\int_0^1 E_\mu[\langle \mathbb{D}f(X+\lambda u\delta\Delta^{-1}\circ\Phi(X)),\Phi(X)\rangle]\,\mathrm{d}\lambda\\ &=E_\mu[\langle \mathbb{D}f(X),\Phi(X)\rangle]\\ &=E_\mu[\langle {}^{\mathrm{o}}\mathbb{D}f-h^*,\Phi\rangle]=\|{}^{\mathrm{o}}\mathbb{D}f-h^*\|^q_{L^q(\mu)},\end{aligned}$$

where the second equality is a consequence of (5.29). This implies that

$$E_{\pi_\delta}[f(Y)-f(X)]=\|{}^{\mathrm{o}}\mathbb{D}f-h^*\|_{L^q(\mu)}\Big(u\|{}^{\mathrm{o}}\mathbb{D}f-h^*\|^{q/p}_{L^q(\mu)}\Big)\delta+o(\delta). \tag{5.30}$$

The proof would be complete if we were able to show that the constructed $\pi_\delta\in\mathcal{P}$. By (5.29), indeed the second marginal of π_δ is a martingale measure. But, in general π_δ may not be a bi-causal coupling. Instead, by Proposition 5.35, we can approximate π_δ by a bi-causal coupling $\tilde\pi_\delta\in\mathcal{P}$. Assumption 5.14 ensures that taking ε small enough, e.g., $\varepsilon=\delta^2$, the estimate (5.30) still holds if we replace π_δ with $\tilde\pi_\delta$. This concludes the proof.

□

5.6 Proofs of continuous-time results

5.6.1 Hyperbolic scaling

Proof of Theorem 5.22. Recall that $\mathcal{X}=C_0([0,T];\mathbb{R}^n)$. By Lemma 5.34, it suffices to verify (5.21), (5.22) and (5.23) with $\mathcal{P}=\Pi_{\mathfrak{c}}(\mu,*)$ and

$$r=\left(\int_0^T E_\mu[|{}^{\mathrm{o}}\mathbb{D}_tf(X)|^q_*]\,\mathrm{d}t\right)^{1/q}=\|{}^{\mathrm{o}}\mathbb{D}f\|_{L^q(\mu)}.$$

We notice that (5.21) follows from

$$|f(y)|\le C(1+\|y\|^p_\infty)\le C(1+\|x\|^p_{W^{1,p}_0}+\|x-y\|^p_{W^{1,p}_0}).$$

Without loss of generality, we may assume $E_\pi[d(X,Y)^p]<\infty$. As we assume $f\in\mathbb{D}^1_b$, by Proposition 5.6 it holds

$$E_\pi[f(Y)-f(X)]=E_\pi\left[\int_0^1\int_0^T\langle \mathbb{D}_tf(X+\lambda(Y-X)),(X-Y)'(t)\rangle\,\mathrm{d}t\,\mathrm{d}\lambda\right].$$

In particular, we can choose a version of $(X-Y)'$ such that it is adapted to $(\mathcal{F}_t\otimes\mathcal{G}_t)_{t\in I}$. Therefore, applying Hölder inequality we deduce

$$\begin{aligned}E_\pi[f(Y)-f(X)] &= E_\pi\left[\int_0^1\int_0^T \langle \mathbb{D}_t f(X+\lambda(Y-X)),(X-Y)'(t)\rangle\,\mathrm{d}t\,\mathrm{d}\lambda\right]\\ &= E_\pi\left[\int_0^1\int_0^T \langle {}^{\mathrm{o}}\mathbb{D}_t f(X+\lambda(Y-X)),(X-Y)'(t)\rangle\,\mathrm{d}t\,\mathrm{d}\lambda\right]\\ &\le E_\pi[d(X,Y)^p]^{1/p}\int_0^1 \|{}^{\mathrm{o}}\mathbb{D}f(X+\lambda(Y-X))\|_{L^q(\pi)}\,\mathrm{d}\lambda.\end{aligned}$$

In order to verify (5.22), it suffices to show that for any sequence $\pi_n\in\mathcal{P}$ with $\lim_{n\to\infty}E_{\pi_n}[d(X,Y)^p]=0$ it holds

$$\limsup_{n\to\infty}\int_0^1 \|{}^{\mathrm{o}}\mathbb{D}f(X+\lambda(Y-X))\|_{L^q(\pi_n)}\,\mathrm{d}\lambda \le \|{}^{\mathrm{o}}\mathbb{D}f\|_{L^q(\mu)}. \tag{5.31}$$

Since $\|\cdot\|_\infty$ is dominated by $\|\cdot\|_{W_0^{1,p}}$ and $\mathbb{D}f$ is continuous with respect to the uniform topology, following the same arguments as for (5.25) in the proof of Theorem 5.15 we deduce (5.31). It remains to establish (5.23). We define $\Phi:\mathcal{X}\to AC_0([0,T];\mathbb{R}^n)$ as

$$\Phi_t(X)=\int_0^t v({}^{\mathrm{o}}\mathbb{D}_s f(X))\,\mathrm{d}s,$$

where $v:\mathbb{R}^n\to\mathbb{R}^n$ is, as previously, given by (5.26). We fix $u>0$ such that $L^*(r)=ur-L(u)$. For any $\delta>0$, we construct $\pi_\delta=(\mathrm{Id},\mathrm{Id}+u\delta\Phi)_\#\mu$. Since Φ is adapted, we have $\pi_\delta\in\Pi_{\mathfrak{c}}(\mu,*)$. A direct computation gives

$$E_{\pi_\delta}[d(X,Y)^p]=u^p\delta^pE_\mu\left[\int_0^T |v({}^{\mathrm{o}}\mathbb{D}_t f(X))|^p\,\mathrm{d}t\right]=\left(u\|{}^{\mathrm{o}}\mathbb{D}f\|_{L^q(\mu)}^{q/p}\right)^p\delta^p.$$

By Proposition 5.6, with $\eta=\Phi(X)$ and noting $\dot\eta_t=v({}^{\mathrm{o}}\mathbb{D}_t f(X))$, we derive

$$\begin{aligned}E_{\pi_\delta}[f(Y)-f(X)] &= E_\mu[f(X+u\delta\Phi(X))-f(X)]\\ &= u\delta\int_0^1\int_0^T E_\mu[\langle \mathbb{D}_t f(X+\lambda u\delta\Phi(X)),v({}^{\mathrm{o}}\mathbb{D}_t f(X))\rangle]\,\mathrm{d}t\,\mathrm{d}\lambda.\end{aligned}$$

By Assumption 5.20 and dominated convergence theorem, as $\delta\to0$, we have

$$\lim_{\delta\to0}\int_0^1\int_0^T E_\mu[\langle \mathbb{D}_t f(X+\lambda u\delta\Phi(X)),v({}^{\mathrm{o}}\mathbb{D}_t f(X))\rangle]\,\mathrm{d}t\,\mathrm{d}\lambda=\int_0^T E_\mu[|{}^{\mathrm{o}}\mathbb{D}_t f(X)|^q]\,\mathrm{d}t.$$

This implies that

$$E_{\pi_\delta}[f(Y)-f(X)]=\|{}^{\mathrm{o}}\mathbb{D}f\|_{L^q(\mu)}\left(u\|{}^{\mathrm{o}}\mathbb{D}f\|_{L^q(\mu)}^{q/p}\right)\delta+o(\delta),$$

and hence (5.23) holds. □

5.6.2 Parabolic scaling

In the proofs that follow, in a chain of inequalities, C may denote a different constant from one line to another. Recall that

$$V_{\mathrm{Mart}}(\delta) = \sup_{\nu \in \mathscr{M}(\mathcal{X})} \{E_\nu[U(H)] - L_\delta(\mathcal{AW}_2(\mu, \nu))\},$$

where $H = \int_0^T \langle h(t, X(\cdot \wedge t)), \mathrm{d}X(t)\rangle$.

Lemma 5.36. *Let Assumption 5.28 hold. Then, for any $\pi \in \Pi_{\mathfrak{bc}}(\mu, *)$ with $\pi(\mathcal{X} \times \cdot) \in \mathscr{M}(\mathcal{X})$ we have*

$$\begin{aligned}
&E_\pi\left[U\left(\int_0^T \langle h(t, Y(\cdot \wedge t)), \mathrm{d}Y(t)\rangle\right) - U\left(\int_0^T \langle h(t, X(\cdot \wedge t)), \mathrm{d}X(t)\rangle\right)\right] \\
&= E_\pi\left[U'(H) \int_0^T \langle h(t, X(\cdot \wedge t)), \mathrm{d}(Y - X)(t)\rangle\right] \\
&\quad + E_\pi\left[\int_0^T \left\langle \int_0^t \mathbf{D}_s^\intercal h(t, X(\cdot \wedge t))\, \mathrm{d}^-(Y - X)(s), \mathrm{d}X(t)\right\rangle\right] + o\Big(E_\pi[[X - Y]_T]^{1/2}\Big).
\end{aligned}$$

Proof. Since π is bi-causal coupling between martingale measures, by Proposition 5.37 we have (X, Y) is a joint martingale under π. By Assumption 5.28 and Itô isometry, we notice

$$\begin{aligned}
&E_\pi\left[\left|\int_0^T \langle h(t, Y(\cdot \wedge t)), \mathrm{d}Y(t)\rangle - \int_0^T \langle h(t, X(\cdot \wedge t)), \mathrm{d}X(t)\rangle\right|^2\right] \\
&\le CE_\pi\left[\left|\int_0^T \langle h(t, Y(\cdot \wedge t)), \mathrm{d}(X - Y)(t)\rangle\right|^2\right] \\
&\quad + CE_\pi\left[\left|\int_0^T \langle h(t, X(\cdot \wedge t)) - h(t, Y(\cdot \wedge t)), \mathrm{d}X(t)\rangle\right|^2\right] \\
&\le CE_\pi\left[[X - Y]_T \sup_{t \in [0,T]} |h(t, Y(\cdot \wedge t))|^2\right] \\
&\quad + CE_\pi\left[[X]_T \sup_{t \in [0,T]} |h(t, X(\cdot \wedge t)) - h(t, Y(\cdot \wedge t))|^2\right] \\
&= o(E_\pi[[X - Y]_T]^{1/2}),
\end{aligned}$$

where in the last step, we use the fact that h is bounded and $[X]_T = T$. Since U has a bounded second derivative, we have

$$|U(y) - U(x) - U'(x)(y - x)| \le C|y - x|^2.$$

Together with the previous estimate, this implies that

$$\begin{aligned}
&E_\pi\Big[U\Big(\int_0^T \langle h(t, Y(\cdot\wedge t)), \mathrm{d}Y(t)\rangle\Big) - U\Big(\int_0^T \langle h(t, X(\cdot\wedge t)), \mathrm{d}X(t)\rangle\Big)\Big] \\
&= E_\pi\Big[U'(H)\Big(\int_0^T \langle h(t, Y(\cdot\wedge t)), \mathrm{d}Y(t)\rangle - \int_0^T \langle h(t, X(\cdot\wedge t)), \mathrm{d}X(t)\rangle\Big)\Big] \\
&\quad + o(E_\pi[[X-Y]_T]^{1/2}).
\end{aligned}$$

For simplicity, we write

$$I_1 = \int_0^T \langle h(t, X(\cdot\wedge t)), \mathrm{d}(Y-X)(t)\rangle$$

and

$$I_2 = \int_0^T \Big\langle \int_0^t \mathbf{D}_s^{\intercal} h(t, X(\cdot\wedge t))\, \mathrm{d}^-(Y-X)(s), \mathrm{d}X(t)\Big\rangle.$$

Notice that $U'(H)$ has a finite moment of any order. To conclude the proof, it then suffices to show that for γ in Assumption 5.28

$$\begin{aligned}
&E_\pi\Big[\Big|\int_0^T \langle h(t, Y(\cdot\wedge t)), \mathrm{d}Y(t)\rangle - \int_0^T \langle h(t, X(\cdot\wedge t)), \mathrm{d}X(t)\rangle - I_1 - I_2\Big|^\gamma\Big] \\
&= o(E_\pi[[X-Y]_T]^{\gamma/2}).
\end{aligned} \tag{5.32}$$

Plugging I_1 and I_2 into (5.32), we obtain

$$\begin{aligned}
&E_\pi\Big[\Big|\int_0^T \langle h(t, Y(\cdot\wedge t)), \mathrm{d}Y(t)\rangle - \int_0^T \langle h(t, X(\cdot\wedge t)), \mathrm{d}X(t)\rangle - I_1 - I_2\Big|^\gamma\Big] \\
&\le CE_\pi\Big[\Big|\int_0^T \langle h(t, Y(\cdot\wedge t)) - h(t, X(\cdot\wedge t)), \mathrm{d}(Y-X)(t)\rangle\Big|^\gamma\Big] \\
&+ CE_\pi\Big[\Big|\int_0^T \Big\langle h(t, Y(\cdot\wedge t)) - h(t, X(\cdot\wedge t)) - \int_0^t \mathbf{D}_s^{\intercal} h(t, X(\cdot\wedge t))\, \mathrm{d}^-(Y-X)(s), \mathrm{d}X(t)\Big\rangle\Big|^\gamma\Big] \\
&:= J_1 + J_2.
\end{aligned}$$

It follows from BDG inequality and Hölder inequality that

$$\begin{aligned}
J_1 &\le CE_\pi\Big[[X-Y]_T^{\gamma/2} \sup_{t\in[0,T]} |h(t, Y(\cdot\wedge t)) - h(t, X(\cdot\wedge t))|^\gamma\Big] \\
&\le CE_\pi[[X-Y]_T]^{\gamma/2} E_\pi\Big[\sup_{t\in[0,T]} |h(t, Y(\cdot\wedge t)) - h(t, X(\cdot\wedge t))|^{2\gamma/(2-\gamma)}\Big]^{1-\gamma/2} \\
&\le CE_\pi[[X-Y]_T]^{\gamma/2} E_\pi\Big[\sup_{t\in[0,T]} |h(t, Y(\cdot\wedge t)) - h(t, X(\cdot\wedge t))|^2\Big]^{1-\gamma/2} \\
&= o(E_\pi[[X-Y]_T]^{\gamma/2}).
\end{aligned}$$

Here, the third line follows from the boundedness of h and $\gamma \in (1,2)$. For J_2, similarly by BDG inequality and Assumption 5.28 we deduce

$$\begin{aligned}J_2 &\le CE_\pi\left[\sup_{t\in[0,T]}\left|h(t,Y(\cdot\wedge t)) - h(t,X(\cdot\wedge t)) - \int_0^t \mathbf{D}_s^\intercal h(t,X(\cdot\wedge t))\,\mathrm{d}^-(Y-X)(s)\right|^\gamma [X]_T^{\gamma/2}\right]\\ &\le o(E_\pi[[X-Y]_T]^{\gamma/2}).\end{aligned}$$

□

The proof of Theorem 5.30 relies on the following proposition, which we adapt from Acciaio et al. (2020, Remark 2.3 (4)).

Proposition 5.37. *Let $\mu,\nu \in \mathscr{M}(\mathcal{X})$. Then $\pi \in \Pi_{\mathfrak{bc}}(\mu,\nu)$ implies that (X,Y) is an $(\mathcal{F}_t\otimes\mathcal{G}_t)_{t\in I}$–martingale. Suppose additionally that X and Y have the martingale representation property under μ and ν respectively. Then (X,Y) is an $(\mathcal{F}_t\otimes\mathcal{G}_t)_{t\in I}$–martingale under $\pi\in\Pi(\mu,\nu)$ implies that π is bi-causal.*

Proof of Theorem 5.30. Recall that $H = \int_0^T \langle h(t,X(\cdot\wedge t)), \mathrm{d}X(t)\rangle$ and

$$\varphi(t) = {}^{\mathrm{p}}\{\mathbf{D}_{t+}\mathbf{D}_t U(H) - U'(H)\mathbf{D}_t^\intercal h(t,X(\cdot\wedge t))\}.$$

We remark $\mathbf{D}_+\mathbf{D}U(H)$ is well-defined from the regularity condition of h. By Lemma 5.34, it suffices to verify (5.21), (5.22), and (5.23) with $\mathcal{P} = \{\pi\in\Pi_{\mathfrak{bc}}(\mu,*) : \Pi(\mathcal{X}\times\cdot)\in \mathscr{M}(\mathcal{X})\}$ and

$$r = E_\mu\left[\int_0^T \|\varphi(t)\|_{\mathbf{F}}^2\,\mathrm{d}t\right]^{1/2}.$$

To show (5.21), we notice that h is bounded and U has a bounded second derivative, and derive

$$\begin{aligned}E_\pi\left[U\left(\int_0^T\langle h(t,Y(\cdot\wedge t)),\mathrm{d}Y(t)\rangle\right)\right] &\le C\left(1+E_\pi\left[\left(\int_0^T\langle h(t,Y(\cdot\wedge t)),\mathrm{d}Y(t)\rangle\right)^2\right]\right)\\ &\le C(1+E_\pi[[Y-X]_T]).\end{aligned}$$

Now, we verify (5.22). Since $\mathbf{D}_s h(t,X(\cdot\wedge t))$ satisfies conditions in Theorem 5.12, by Theorem 5.12 we calculate

$$\begin{aligned}&\int_0^T\langle h(s,X(\cdot\wedge s)),\mathrm{d}^-(Y-X)(s)\rangle + \int_0^T\left\langle\int_0^t \mathbf{D}^\intercal(s)h(t,X(\cdot\wedge t))\,\mathrm{d}^-(Y-X)(s),\mathrm{d}X(t)\right\rangle\\ &= \int_0^T\langle h(s,X(\cdot\wedge s)),\mathrm{d}^-(Y-X)(s)\rangle + \int_0^T\left\langle\int_s^T \mathbf{D}_s h(t,X(\cdot\wedge t))\,\mathrm{d}X(t),\mathrm{d}^-(Y-X)(s)\right\rangle\\ &\qquad - \int_0^T\langle\mathbf{D}_s^\intercal h(s,X(\cdot\wedge s)),\mathrm{d}[\![X,Y-X]\!](s)\rangle_{\mathbf{F}}\\ &= \int_0^T\langle\mathbf{D}_s H,\mathrm{d}^-(Y-X)(s)\rangle - \int_0^T\langle\mathbf{D}_s^\intercal h(s,X(\cdot\wedge s)),\mathrm{d}[\![X,Y-X]\!](s)\rangle_{\mathbf{F}}.\end{aligned}$$

Therefore, Lemma 5.36 yields

$$
\begin{aligned}
&E_\pi\left[U\left(\int_0^T \langle h(t, Y(\cdot\wedge t)), \mathrm{d}Y(t)\rangle\right) - U\left(\int_0^T \langle h(t, X(\cdot\wedge t)), \mathrm{d}X(t)\rangle\right)\right]\\
&= E_\pi\left[U'(H)\int_0^T \langle \mathbf{D}_s H, \mathrm{d}^-(Y-X)(s)\rangle - U'(H)\int_0^T \langle \mathbf{D}_s^\intercal h(s, X(\cdot\wedge s)), \mathrm{d}[\![X, Y-X]\!](s)\rangle_{\mathbf{F}}\right]\\
&\qquad + o\big(E_\pi[d(X,Y)^2]^{1/2}\big).
\end{aligned}\tag{5.33}
$$

Since $U'(H)$ has a finite moment of any order, a simple application of Hölder inequality gives $\mathbf{D}U(H) = U'(H)\mathbf{D}H$ is $(Y-X)$ forward–γ' integrable for any $\gamma' \in [1,\gamma)$. Furthermore, we notice that $\mathbf{D}U(H)$ satisfies the assumption in Proposition 5.11, and hence we deduce

$$
E_\pi\left[U'(H)\int_0^T \langle \mathbf{D}_s H, \mathrm{d}^-(Y-X)(s)\rangle\right] = E_\pi\left[\int_0^T \langle \mathbf{D}_{s+}\mathbf{D}_s U(H), \mathrm{d}[\![X, Y-X]\!](s)\rangle_{\mathbf{F}}\right].
$$

Plugging the above equality into estimate (5.33) yields

$$
\begin{aligned}
&E_\pi\left[U\left(\int_0^T \langle h(t, Y(\cdot\wedge t)), \mathrm{d}Y(t)\rangle\right) - U\left(\int_0^T \langle h(t, X(\cdot\wedge t)), \mathrm{d}X(t)\rangle\right)\right]\\
&= E_\pi\left[\int_0^T \langle \mathbf{D}_{s+}\mathbf{D}_s U(H) - U'(H)\mathbf{D}_s^\intercal h(s, X(\cdot\wedge s)), \mathrm{d}[\![X, Y-X]\!](s)\rangle_{\mathbf{F}}\right] + o\big(E_\pi[d(X,Y)^2]^{1/2}\big)\\
&= E_\pi\left[\int_0^T \langle \varphi(s), \mathrm{d}[\![X, Y-X]\!](s)\rangle_{\mathbf{F}}\right] + o\big(E_\pi[d(X,Y)^2]^{1/2}\big).
\end{aligned}\tag{5.34}
$$

Hence, by Kunita–Watanabe inequality, we establish (5.22) holds as follows:

$$
\begin{aligned}
&E_\pi\left[U\left(\int_0^T \langle h(t, Y(\cdot\wedge t)), \mathrm{d}Y(t)\rangle\right) - U\left(\int_0^T \langle h(t, X(\cdot\wedge t)), \mathrm{d}X(t)\rangle\right)\right]\\
&\le E_\pi\left[\int_0^T \langle \varphi(s)\varphi(s)^\intercal, \mathrm{d}[\![X]\!](s)\rangle_{\mathbf{F}}\right]^{1/2} E_\pi\left[\int_0^T \langle \mathrm{Id}, \mathrm{d}[\![Y-X]\!](s)\rangle_{\mathbf{F}}\right]^{1/2} + o\big(E_\pi[d(X,Y)^2]^{1/2}\big)\\
&\le E_\mu\left[\int_0^T \|\varphi(t)\|_{\mathbf{F}}^2\,\mathrm{d}t\right]^{1/2} E_\pi[d(X,Y)^2]^{1/2} + o\big(E_\pi[d(X,Y)^2]^{1/2}\big).
\end{aligned}
$$

We turn now to (5.23) and define $\Phi_\cdot = \int_0^\cdot \varphi(s)\,\mathrm{d}X(s)$. We fix $u > 0$ such that $L^*(r) = ur - L(u)$. For $\delta > 0$, set $\pi_\delta = (\mathrm{Id}, \mathrm{Id} + u\delta\Phi)_{\#}\mu$. By direct computation, we have

$$
E_{\pi_\delta}[d(X,Y)^p] = u^2\delta^2 E_\mu[[\Phi]_T] = u^2\delta^2 E_\mu\left[\int_0^T \|\varphi(t)\|_{\mathbf{F}}^2\,\mathrm{d}t\right].
$$

On the other hand, we notice

$$
\begin{aligned}
&E_{\pi_\delta}\left[U'(H)\int_0^T \langle \mathbf{D}_s H, \mathrm{d}^-(Y-X)(s)\rangle - U'(H)\int_0^T \langle \mathbf{D}_s^\intercal h(s, X(\cdot\wedge s)), \mathrm{d}[\![X, Y-X]\!](s)\rangle_{\mathbf{F}}\right] \\
&= E_{\pi_\delta}\left[\int_0^T \langle \varphi(s), \mathrm{d}[\![X, Y-X]\!](s)\rangle_{\mathbf{F}}\right] \\
&= u\delta E_\mu\left[\int_0^T \langle \varphi(s), \varphi(s)\,\mathrm{d}[\![X]\!](s)\rangle_{\mathbf{F}}\right] = u\delta E_\mu\left[\int_0^T \|\varphi(t)\|_{\mathbf{F}}^2\,\mathrm{d}t\right].
\end{aligned}
$$

Hence, it follows from estimate (5.34) that

$$
\begin{aligned}
&E_{\pi_\delta}\left[U\left(\int_0^T \langle h(t, Y(\cdot\wedge t)), \mathrm{d}Y(t)\rangle\right) - U\left(\int_0^T \langle h(t, X(\cdot\wedge t)), \mathrm{d}X(t)\rangle\right)\right] \\
&= u\delta E_\mu\left[\int_0^T \|\varphi(t)\|_{\mathbf{F}}^2\,\mathrm{d}t\right] + o(\delta).
\end{aligned}
\tag{5.35}
$$

The proof would be complete if it was the case that $\pi_\delta \in \mathcal{P}$. However, in general π_δ is not a bi-causal coupling. To remedy this, we consider the following approximation to π_δ. Let φ^n be a sequence of bounded predictable processes such that

$$
\lim_{n\to\infty} E_\mu\left[\int_0^T \|\varphi(t) - \varphi^n(t)\|_{\mathbf{F}}^2\,\mathrm{d}t\right] = 0.
$$

We construct $\pi_\delta^n = (\mathrm{Id}, \mathrm{Id} + u\delta\Phi^n)_{\#}\mu$, where $\Phi^n_\cdot = \int_0^\cdot \varphi^n(s)\,\mathrm{d}X(s)$. Following the same argument as above, we have

$$
\begin{aligned}
&E_{\pi_\delta^n}\left[U\left(\int_0^T \langle h(t, Y(\cdot\wedge t)), \mathrm{d}Y(t)\rangle\right) - U\left(\int_0^T \langle h(t, X(\cdot\wedge t)), \mathrm{d}X(t)\rangle\right)\right] \\
&= u\delta E_\mu\left[\int_0^T \langle \varphi, \varphi^n\rangle_{\mathbf{F}}\,\mathrm{d}t\right] + o(\delta).
\end{aligned}
$$

We notice that indeed $X + u\delta\Phi^n$ is a regular martingale. Moreover, for δ sufficiently small, $X + u\delta\Phi^n$ is a non-degenerate martingale, and hence $X + u\delta\Phi^n$ has the martingale representation property. This implies for sufficiently small δ, π_δ^n is a bi-causal coupling by Proposition 5.37. Therefore, there exists a sequence of bi-causal couplings $\tilde{\pi}_\delta$ such that estimate (5.35) still holds if we replace π_δ by $\tilde{\pi}_\delta$. This concludes the proof.

□

5.7 Auxiliary proofs

In the section, we present the remaining proofs, in particular for the results in Section 5.3.

Proof of Proposition 5.35. The first part follows directly from (Bartl and Wiesel, 2023, Lemma 3.1). We prove the second part by induction. The base case $N = 1$ is trivial, and we assume the statement holds for $N-1$. Then there exists $(0, \Phi_1^\varepsilon, \ldots, \Phi_{N-1}^\varepsilon)$ with $|\Phi_n^\varepsilon - \Phi_n| < \varepsilon$ for $1 \le n \le N-1$ such that the projection of $\Phi_\#^\varepsilon \mu$ on the first $N-1$ marginals is a martingale. We may further assume $(0, \Phi_1^\varepsilon, \ldots, \Phi_{n-1}^\varepsilon)$ is an injection for any $1 \le n \le N-1$.

Now, it suffices to construct Φ_N^ε with $|\Phi_N^\varepsilon - \Phi_N|_\infty < \varepsilon$ such that $\Phi_\#^\varepsilon \mu \in \mathscr{M}(\mathcal{X})$ and $\Phi^\varepsilon : \mathcal{X} \to \mathcal{X}$ is injective. We write $\mathrm{pj}^\varepsilon(x) = \lfloor \frac{4x}{\varepsilon} \rfloor \frac{\varepsilon}{4}$ and construct

$$\Phi_N^\varepsilon(0, x_1, \ldots, x_N) = \mathrm{pj}^\varepsilon \circ \Phi_N(0, x_1, \ldots, x_N) + \varphi^\varepsilon(x_N) - r(0, x_1, \ldots, x_{N-1}),$$

where pj^ε is the projection to the $\varepsilon/4$ grid, $\varphi^\varepsilon : \mathbb{R}^n \to (0, \varepsilon/4)^d$ is a measurable bijection, and r is the residual given by

$$r(0, x_1, \ldots, x_{N-1}) = E_{\mu_{N-1}}[\mathrm{pj}^\varepsilon \circ \Phi_N(0, x_1, \ldots, x_{N-1}, X_N) + \varphi^\varepsilon(X_N)] - \Phi_{N-1}^\varepsilon(0, x_1, \ldots, x_{N-1}).$$

Here, μ_{N-1} is the disintegration kernel of μ given by

$$\mu(\mathrm{d}x_1, \ldots, \mathrm{d}x_N) = \mu(\mathrm{d}x_1, \ldots, \mathrm{d}x_{N-1}) \mu_{N-1}(x_1, \ldots, x_{N-1}, \mathrm{d}x_N).$$

It is clear that $|\Phi_N^\varepsilon - \Phi_N|_\infty < \varepsilon$ and $\Phi_\#^\varepsilon \mu \in \mathscr{M}(\mathcal{X})$. We now verify that Φ^ε is injective. Assume $\Phi^\varepsilon(0, x_1, \ldots, x_N) = \Phi^\varepsilon(0, x_1', \ldots, x_N')$. By induction assumption, we derive $x_n = x_n'$ for $1 \le n \le N-1$. Therefore, this implies

$$\mathrm{pj}^\varepsilon \circ \Phi_N(0, x_1, \ldots, x_N) + \varphi^\varepsilon(x_N) = \mathrm{pj}^\varepsilon \circ \Phi_N(0, x_1', \ldots, x_N') + \varphi^\varepsilon(x_N')$$

which further implies $x_N = x_N'$. Therefore, Φ^ε is injective and $(\mathrm{Id}, \Phi^\varepsilon)_\# \mu \in \Pi_{\mathfrak{bc}}(\mu, *)$. □

Proof of Propositions 5.6 and 5.9. We start by proving Proposition 5.6. If η is a simple step function, then the result is immediate from the continuity of $\mathbb{D}_t f$. This implies

$$f(\omega + \eta) - f(\omega) = \int_0^1 \sum_{k=1}^n \langle \mathbb{D}_{t_k} f(\omega + \lambda\eta), e_k \rangle \, \mathrm{d}\lambda.$$

Now, we assume $\eta \in AC_0([0,T]; \mathbb{R}^n)$ and take a sequence of simple step functions $\eta^n(t) = \sum_{k=1}^n (t_{k+1}^n - t_k^n) e_k^n \mathbb{1}_{[t_k^n, T]}(t)$ such that

$$\sup_{1 \le k \le n} (t_{k+1}^n - t_k^n)(|e_k^n| \vee 1) \to 0 \qquad \text{and} \qquad \sum_{k=1}^n e_k^n \mathbb{1}_{[t_k^n, t_{k+1}^n)} \to \dot{\eta} \text{ in } L^1.$$

We can always achieve the above by refining partitions. Therefore, we deduce that η^n converges to η in the uniform topology. This gives

$$\begin{aligned} f(\omega+\eta)-f(\omega) &= \lim_{n\to\infty} f(\omega+\eta^n)-f(\omega) \\ &= \lim_{n\to\infty}\int_0^1 \sum_{k=1}^{n}(t^n_{k+1}-t^n_k)\langle \mathbb{D}_{t^n_k} f(\omega+\lambda\eta^n), e^n_k\rangle \,\mathrm{d}\lambda \\ &= \lim_{n\to\infty}\int_0^1\int_0^T \sum_{k=1}^{n}\langle \mathbb{D}_{t^n_k} f(\omega+\lambda\eta^n), e^n_k \mathbb{1}_{[t^n_k,t^n_{k+1})}(t)\rangle \,\mathrm{d}t\,\mathrm{d}\lambda. \end{aligned}$$

Since $\mathbb{D}_t f$ has left limits, the integrand converges to $\langle \mathbb{D}_{t-} f(\omega+\lambda\eta), \dot\eta(t)\rangle$, which is equal to $\langle \mathbb{D}_t f(\omega+\lambda\eta), \dot\eta(t)\rangle$ $\mathrm{d}t\otimes\mathrm{d}\lambda$–a.e. Furthermore, as $\mathbb{D}f$ is boundedness preserving, we derive by the dominated convergence theorem that

$$f(\omega+\eta)-f(\omega) = \int_0^1\int_0^T \langle \mathbb{D}_t f(\omega+\lambda\eta), \dot\eta(t)\rangle \,\mathrm{d}t\,\mathrm{d}\lambda,$$

from which (5.10) follows.

To conclude that the pathwise and the classical Malliavin derivatives coincide, it suffices to observe that $W_0^{1,2}\subseteq AC_0$ and that, by (5.7) and (5.10), we have

$$\int_0^T \langle \mathbb{D}_t f(X), \dot\eta(t)\rangle \,\mathrm{d}t = \int_0^T \langle \mathbf{D}_t f(X), \dot\eta(t)\rangle \,\mathrm{d}t, \quad \forall \eta\in W_0^{1,2}.$$

□

Before we prove the stochastic Fubini theorem (Theorem 5.12), we first show that the forward integral indeed agrees with the Itô integral if the integrand is predictable. The below can be viewed as an L^γ extension of Russo and Vallois (1993, Proposition 1.1).

Proposition 5.38. *Let $\gamma\in[1,2)$, A be a predictable process, and B be a regular martingale with $\mathrm{d}[B](t)=\xi(t)\,\mathrm{d}t$. We assume either*

$$E_P\left[\sup_{t\in[0,T]}|A(t)|^{2\gamma/(2-\gamma)}\right]<\infty \quad \textit{and} \quad E_P\left[\int_0^T |\xi(t)|\,\mathrm{d}t\right]<\infty, \tag{5.36}$$

or

$$E_P\left[\int_0^T |A(t)|^2\,\mathrm{d}t\right]<\infty \quad \textit{and} \quad E_P\left[\sup_{t\in[0,T]}|\xi(t)|^{\gamma/(2-\gamma)}\right]<\infty. \tag{5.37}$$

Then A is B forward–γ integrable, and the forward integral coincides with the Itô integral, i.e.,

$$\int_0^T \langle A(t), \mathrm{d}^- B(t)\rangle = \int_0^T \langle A(t), \mathrm{d}B(t)\rangle.$$

Proof. We write the Hardy–Littlewood maximal process as

$$A^*(t) = \sup_{\varepsilon\in[0,T]} \frac{1}{\varepsilon}\int_{(t-\varepsilon)\vee 0}^{t} |A(s)|\,\mathrm{d}s \quad \text{for } t\in[0,T].$$

By Hardy–Littlewood maximal inequality, we have $\int_0^T |A^*(t)|^2\,\mathrm{d}t \le C\int_0^T |A(t)|^2\,\mathrm{d}t$, where C is a deterministic constant. Combining this with the assumption and Hölder inequality, we derive either

$$E_P\left[\left(\int_0^T |A^*(t)|^2\,\mathrm{d}[B](t)\right)^{\gamma/2}\right] \le E_P\left[\sup_{t\in[0,T]}|A(t)|^\gamma\left(\int_0^T |\xi(t)|\,\mathrm{d}t\right)^{\gamma/2}\right] < \infty,$$

or

$$E_P\left[\left(\int_0^T |A^*(t)|^2\,\mathrm{d}[B](t)\right)^{\gamma/2}\right] \le E_P\left[\left(\int_0^T |A^*(t)|^2\,\mathrm{d}t\right)^{\gamma/2}\sup_{t\in[0,T]}|\xi(t)|^{\gamma/2}\right] < \infty.$$

Then by Lebesgue dominated convergence theorem, we obtain

$$\lim_{\varepsilon\to 0} E_P\left[\left(\int_0^T \left|\frac{1}{\varepsilon}\int_{(t-\varepsilon)\vee 0}^{t} A(s)\,\mathrm{d}s - A(t)\right|^2 \mathrm{d}[B](t)\right)^{\gamma/2}\right] = 0,$$

as the integrand converges to 0 from Lebesgue differentiation theorem and is dominated by $C\Big(\int_0^T |A^*(t)|^2\,\mathrm{d}[B](t)\Big)^{\gamma/2}$. Therefore, by BDG inequality, we have the L^γ convergence

$$\lim_{\varepsilon\to 0}\int_0^T \left\langle \frac{1}{\varepsilon}\int_{(t-\varepsilon)\vee 0}^{t} A(s)\,\mathrm{d}s, \mathrm{d}B(t)\right\rangle = \int_0^T \langle A(t), \mathrm{d}B(t)\rangle.$$

On the other hand, by the Stochastic Fubini theorem (Veraar, 2012, Theorem 2.2), the above limit is equal to

$$\begin{aligned}
\lim_{\varepsilon\to 0}\int_0^T \left\langle \frac{1}{\varepsilon}\int_{(t-\varepsilon)\vee 0}^{t} A(s)\,\mathrm{d}s, \mathrm{d}B(t)\right\rangle &= \lim_{\varepsilon\to 0}\frac{1}{\varepsilon}\int_0^T \left\langle \int_0^T A(s)\mathbb{1}_{[s,(s+\varepsilon)\wedge T]}(t)\,\mathrm{d}s, \mathrm{d}B(t)\right\rangle \\
&= \lim_{\varepsilon\to 0}\frac{1}{\varepsilon}\int_0^T\int_0^T \langle A(s)\mathbb{1}_{[s,(s+\varepsilon)\wedge T]}(t), \mathrm{d}B(t)\rangle\,\mathrm{d}s \\
&= \lim_{\varepsilon\to 0}\frac{1}{\varepsilon}\int_0^T \langle A(t), B((t+\varepsilon)\wedge T) - B(t)\rangle\,\mathrm{d}t \\
&= \int_0^T \langle A(t), \mathrm{d}^- B(t)\rangle.
\end{aligned}$$

Therefore, A is B forward–γ integrable, and the forward integral coincides with the Itô integral.

□

Proof of Theorem 5.12. By the definition of the forward integral, we write

$$
\begin{aligned}
&\int_0^T \left\langle \int_s^T \Psi(s,t)^\intercal \, \mathrm{d}X(t), \mathrm{d}^- M(s) \right\rangle \\
&= \lim_{\varepsilon\to 0} \frac{1}{\varepsilon} \int_0^T \left\langle \int_s^T \Psi(s,t)^\intercal \, \mathrm{d}X(t), (M((s+\varepsilon)\wedge T) - M(s)) \right\rangle \mathrm{d}s \\
&= \lim_{\varepsilon\to 0} \frac{1}{\varepsilon} \int_0^T \left\langle \int_s^{(s+\varepsilon)\wedge T} (\Psi(s,t)^\intercal - \Psi(t,t)^\intercal) \, \mathrm{d}X(t), (M((s+\varepsilon)\wedge T) - M(s)) \right\rangle \mathrm{d}s \\
&\quad + \lim_{\varepsilon\to 0} \frac{1}{\varepsilon} \int_0^T \left\langle \int_s^{(s+\varepsilon)\wedge T} \Psi(t,t)^\intercal \, \mathrm{d}X(t), (M((s+\varepsilon)\wedge T) - M(s)) \right\rangle \mathrm{d}s \\
&\quad + \lim_{\varepsilon\to 0} \frac{1}{\varepsilon} \int_0^T \left\langle \int_s^T \Psi(s,t)(M((s+\varepsilon)\wedge t) - M(s)), \mathrm{d}X(t) \right\rangle \mathrm{d}s \\
&\quad - \lim_{\varepsilon\to 0} \frac{1}{\varepsilon} \int_0^T \left\langle \int_s^{(s+\varepsilon)\wedge T} \Psi(s,t)(M_t - M(s)), \mathrm{d}X(t) \right\rangle \mathrm{d}s \\
&:= J_1 + J_2 + J_3 - J_4.
\end{aligned}
$$

It suffices to show the L^γ convergence of J_1, J_2, J_3, and J_4 and compute their limits.

For J_1, by Hölder inequality, BDG inequality, and Fubini theorem, we obtain

$$
\begin{aligned}
&\lim_{\varepsilon\to 0} E_P\left[\left|\frac{1}{\varepsilon}\int_0^T \left\langle \int_s^{(s+\varepsilon)\wedge T} (\Psi(s,t)^\intercal - \Psi(t,t)^\intercal)\,\mathrm{d}X(t), M((s+\varepsilon)\wedge T) - M(s)\right\rangle \mathrm{d}s\right|^\gamma\right] \\
&\le \lim_{\varepsilon\to 0} \frac{C}{\varepsilon^\gamma}\int_0^T E_P\left[\left|\left\langle \int_s^{(s+\varepsilon)\wedge T} (\Psi(s,t)^\intercal - \Psi(t,t)^\intercal)\,\mathrm{d}X(t), M((s+\varepsilon)\wedge T) - M(s)\right\rangle\right|^\gamma\right] \mathrm{d}s \\
&\le \lim_{\varepsilon\to 0} \frac{C}{\varepsilon^\gamma}\int_0^T E_P\left[\left|\int_s^{(s+\varepsilon)\wedge T} (\Psi(s,t)^\intercal - \Psi(t,t)^\intercal)\,\mathrm{d}X(t)\right|^{2\gamma/(2-\gamma)}\right]^{1-\gamma/2} \\
&\qquad\qquad\qquad\qquad \times E_P\big[|M((s+\varepsilon)\wedge T) - M(s)|^2\big]^{\gamma/2}\,\mathrm{d}s \\
&\le \lim_{\varepsilon\to 0} \frac{C}{\varepsilon^\gamma}\int_0^T E_P\left[\left(\int_s^{(s+\varepsilon)\wedge T} \|\Psi(s,t) - \Psi(t,t)\|_{\mathbf{F}}^2\,\mathrm{d}t\right)^{\gamma/(2-\gamma)}\right]^{1-\gamma/2} \\
&\qquad\qquad\qquad\qquad E_P[[M]((s+\varepsilon)\wedge T) - [M](s)]^{\gamma/2}\,\mathrm{d}s \\
&\le \lim_{\varepsilon\to 0} C\left(\int_0^T \left(\frac{1}{\varepsilon}\int_s^{(s+\varepsilon)\wedge T} E_P\big[\|\Psi(s,t) - \Psi(t,t)\|_{\mathbf{F}}^2\big]\,\mathrm{d}t\right)^{\gamma/(2-\gamma)} \mathrm{d}s\right)^{1-\gamma/2} \\
&\qquad\qquad\qquad\qquad \times \left(\frac{1}{\varepsilon}\int_0^T E_P[[M]((s+\varepsilon)\wedge T) - [M](s)]\,\mathrm{d}s\right)^{\gamma/2} \\
&\le \lim_{\varepsilon\to 0} C\left(\int_0^T \left(\frac{1}{\varepsilon}\int_s^{(s+\varepsilon)\wedge T} E_P\big[\|\Psi(s,t) - \Psi(t,t)\|_{\mathbf{F}}^2\big]\,\mathrm{d}t\right)^{\gamma/(2-\gamma)} \mathrm{d}s\right)^{1-\gamma/2} E_P[[M]_T]^{\gamma/2}.
\end{aligned}
$$

Since we assume $\sup_{s,t\in[0,T]}\|\Psi(s,t)\| \in L^{2\gamma/(2-\gamma)}$, and $\lim_{t\to s+} E_P[\|\Psi(s,t)-\Psi(t,t)\|_{\mathbf{F}}^2] = 0$, by Lebesgue dominated convergence theorem we obtain the L^γ convergence of

$$
J_1 = \lim_{\varepsilon\to 0}\frac{1}{\varepsilon}\int_0^T \left\langle \int_s^{(s+\varepsilon)\wedge T} (\Psi(s,t)^\intercal - \Psi(s,t)^\intercal)\,\mathrm{d}X(t), (M((s+\varepsilon)\wedge T) - M(s))\right\rangle \mathrm{d}s = 0.
$$

We write $Y(s) = \int_0^s \Psi^\intercal(t,t)\,\mathrm{d}X(t)$ which is a regular martingale. Reorganizing the second term J_2, we obtain

$$
\begin{aligned}
J_2 &= \lim_{\varepsilon\to 0}\frac{1}{\varepsilon}\int_0^T\left\langle \int_s^{(s+\varepsilon)\wedge T}\Psi(t,t)^\intercal\,\mathrm{d}X(t), M((s+\varepsilon)\wedge T)-M(s)\right\rangle \mathrm{d}s\\
&= \lim_{\varepsilon\to 0}\frac{1}{\varepsilon}\int_0^T \langle Y((s+\varepsilon)\wedge T)-Y(s), M((s+\varepsilon)\wedge T)-M(s)\rangle\,\mathrm{d}s\\
&= \lim_{\varepsilon\to 0}\frac{1}{\varepsilon}\int_0^T (\langle Y((s+\varepsilon)\wedge T), M((s+\varepsilon)\wedge T)\rangle - \langle Y(s), M(s)\rangle)\,\mathrm{d}s\\
&\qquad - \lim_{\varepsilon\to 0}\frac{1}{\varepsilon}\int_0^T \langle M(s), Y((s+\varepsilon)\wedge T)-Y(s)\rangle\,\mathrm{d}s\\
&\qquad - \lim_{\varepsilon\to 0}\frac{1}{\varepsilon}\int_0^T \langle Y(s), M((s+\varepsilon)\wedge T)-M(s)\rangle\,\mathrm{d}s\\
&= \langle Y_T, M_T\rangle - \int_0^T\langle M(s),\mathrm{d}Y(s)\rangle - \int_0^T\langle Y(s),\mathrm{d}M(s)\rangle\\
&= \int_0^T \langle \Psi(s,s), \mathrm{d}[\![X,M]\!](s)\rangle_{\mathbf{F}}.
\end{aligned}
$$

The L^γ convergence of the second last line is justified from Proposition 5.38 by taking $(A,B)=(Y,M)$ and $(A,B)=(M,Y)$ respectively. For J_3, we interchange the Lebesgue integral and the Itô integral by (Veraar, 2012, Theorem 2.2) and derive

$$
\begin{aligned}
J_3 &= \lim_{\varepsilon\to 0}\frac{1}{\varepsilon}\int_0^T\left\langle\int_0^t \Psi(s,t)(M((s+\varepsilon)\wedge t)-M(s))\,\mathrm{d}s, \mathrm{d}X(t)\right\rangle\\
&= \int_0^T\left\langle\int_0^t \Psi(s,t)\,\mathrm{d}^- M(s), \mathrm{d}X(t)\right\rangle.
\end{aligned}
$$

The last equality does converge in L^γ by combining BDG inequality and the assumption that $\{\Psi(\cdot,t)\}_{t\in[0,T]}$ is M. uniformly forward–γ integrable. For the last term J_4, we apply (Veraar, 2012, Theorem 2.2) and BDG inequality. We derive

$$
\begin{aligned}
&\lim_{\varepsilon\to 0} E_P\left[\left|\frac{1}{\varepsilon}\left\langle\int_0^T\int_s^{(s+\varepsilon)\wedge T}\Psi(s,t)(M_t-M(s)), \mathrm{d}X(t)\right\rangle \mathrm{d}s\right|^\gamma\right]\\
&= \lim_{\varepsilon\to 0} E_P\left[\left|\frac{1}{\varepsilon}\int_0^T\left\langle\int_{(t-\varepsilon)\vee 0}^t \Psi(s,t)(M_t-M(s))\,\mathrm{d}s, \mathrm{d}X(t)\right\rangle\right|^\gamma\right]\\
&\le \lim_{\varepsilon\to 0} C E_P\left[\sup_{s,t\in[0,T]}\|\Psi(s,t)\|_{\mathbf{F}}^\gamma\left(\int_0^T\frac{1}{\varepsilon^2}\left(\int_{t-\varepsilon\vee 0}^t |M(t)-M(s)|\,\mathrm{d}s\right)^2\mathrm{d}t\right)^{\gamma/2}\right]\\
&\le \lim_{\varepsilon\to 0} C\left(\int_0^T E_P\left[\sup_{s\in[(t-\varepsilon)\vee 0,t]}|M(t)-M(s)|^2\right]\mathrm{d}t\right)^{\gamma/2} E_P\left[\sup_{s,t\in[0,T]}\|\Psi(s,t)\|_{\mathbf{F}}^{2\gamma/(2-\gamma)}\right]^{1-\gamma/2}\\
&\le \lim_{\varepsilon\to 0} C\left(\int_0^T E_P[[M](t)-[M]_{(t-\varepsilon)\vee 0}]\,\mathrm{d}t\right)^{\gamma/2} = 0.
\end{aligned}
$$

Summarizing above estimates, we conclude $\int_{\cdot}^{T} \Psi(\cdot, t)^{\intercal} \, \mathrm{d}X(t)$ is $M.$ forward–γ integrable and

$$
\begin{aligned}
&\int_0^T \left\langle \int_s^T \Psi(s,t)^{\intercal} \, \mathrm{d}X(t), \mathrm{d}^{-}M(s) \right\rangle \\
&= J_1 + J_2 + J_3 - J_4 \\
&= \int_0^T \left\langle \int_0^t \Psi(s,t) \, \mathrm{d}^{-}M(s), \mathrm{d}X(t) \right\rangle + \int_0^T \langle \Psi(t,t), \mathrm{d}[\![X, M]\!](t) \rangle_{\mathbf{F}}.
\end{aligned}
$$

$\square$

Bibliography

B. Acciaio and S. Hou. Convergence of adapted empirical measures on $\mathbb{R}^d$. *The Annals of Applied Probability*, 34(5):4799–4835, Oct. 2024.

B. Acciaio, J. Backhoff-Veraguas, and A. Zalashko. Causal optimal transport and its links to enlargement of filtrations and continuous-time stochastic optimization. *Stochastic Processes and their Applications*, 130(5):2918–2953, May 2020.

B. Acciaio, S. Hou, and G. Pammer. Entropic adapted Wasserstein distance on Gaussians, Dec. 2024. arXiv:2412.18794.

D. Aldous. Weak convergence and the general theory of processes. July 1981. Incomplete monograph.

M. Anoussis and E. G. Katsoulis. Factorisation in nest algebras. *Proceedings of the American Mathematical Society*, 125(1):87–92, 1997.

M. Anoussis and E. G. Katsoulis. Factorisation in nest algebras. II. *Transactions of the American Mathematical Society*, 350(1):165–183, 1998.

W. Arveson. Operator algebras and invariant subspaces. *Annals of Mathematics*, 100 (3):433–532, 1974.

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, Jan. 2017.

J. Backhoff-Veraguas, D. Bartl, M. Beiglböck, and M. Eder. Adapted Wasserstein distances and stability in mathematical finance. *Finance and Stochastics*, 24(3): 601–632, July 2020a.

J. Backhoff-Veraguas, D. Bartl, M. Beiglböck, and M. Eder. All adapted topologies are equal. *Probability Theory and Related Fields*, 178(3-4):1125–1172, Dec. 2020b.

J. Backhoff-Veraguas, D. Bartl, M. Beiglböck, and J. Wiesel. Estimating processes in adapted Wasserstein distance. *The Annals of Applied Probability*, 32(1):529–550, Feb. 2022a.

J. Backhoff-Veraguas, S. Källblad, and B. A. Robinson. Adapted Wasserstein distance between the laws of SDEs, Sept. 2022b. arXiv:2209.03243.

X. Bai, G. He, Y. Jiang, and J. Obłój. Wasserstein distributional robustness of neural networks. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.

X. Bai, G. He, Y. Jiang, and J. Obloj. Wasserstein distributional adversarial training for deep neural networks, Feb. 2025. arXiv:2502.09352.

R. T. Baillie. Long memory processes and fractional integration in econometrics. *Journal of Econometrics*, 73(1):5–59, July 1996.

V. Bally, L. Caramellino, and R. Cont. *Stochastic Integration by Parts and Functional Itô Calculus*. Advanced Courses in Mathematics - CRM Barcelona. Springer International Publishing, 2016.

D. Bartl and J. Wiesel. Sensitivity of multiperiod optimization problems with respect to the adapted Wasserstein distance. *SIAM Journal on Financial Mathematics*, 14 (2):704–720, June 2023.

D. Bartl, S. Drapeau, and L. Tangpi. Computational aspects of robust optimized certainty equivalents and option pricing. *Mathematical Finance*, 30(1):287–309, 2020.

D. Bartl, S. Drapeau, J. Obłój, and J. Wiesel. Sensitivity analysis of Wasserstein distributionally robust optimization problems. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 477(2256):20210176, Dec. 2021.

D. Bartl, M. Beiglböck, and G. Pammer. The Wasserstein space of stochastic processes. *Journal of the European Mathematical Society*, Dec. 2024.

D. Bartl, M. Beiglböck, G. Pammer, S. Schrott, and X. Zhang. The Wasserstein space of stochastic processes in continuous time, Jan. 2025a. arXiv:2501.14135.

D. Bartl, A. Neufeld, and K. Park. Sensitivity of robust optimization problems under drift and volatility uncertainty, Feb. 2025b. arXiv:2311.11248.

M. Beiglböck and D. Lacker. Denseness of adapted processes among causal couplings, May 2020. arXiv:1805.03185.

D. Bertsekas and S. E. Shreve. *Stochastic Optimal Control: The Discrete-Time Case*. Athena Scientific, Dec. 1996.

J. Bion–Nadal and D. Talay. On a Wasserstein-type distance between solutions to stochastic differential equations. *The Annals of Applied Probability*, 29(3), June 2019.

J. Blanchet and K. Murthy. Quantifying distributional model risk via optimal transport. *Mathematics of Operations Research*, 44(2):565–600, May 2019.

J. Blanchet, Y. Kang, and K. Murthy. Robust Wasserstein profile inference and applications to machine learning. *Journal of Applied Probability*, 56(3):830–857, Sept. 2019.

J. Blanchet, L. Chen, and X. Y. Zhou. Distributionally robust mean-variance portfolio selection with Wasserstein distances. *Management Science*, 68(9):6382–6410, Sept. 2022.

V. Bogachev and S. Popova. Optimal transportation of measures with a parameter, Nov. 2021. arXiv:2111.13014.

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

G. E. P. Box. Science and statistics. *Journal of the American Statistical Association*, 71(356):791–799, 1976.

P. Cheridito. Mixed fractional Brownian motion. *Bernoulli*, 7(6):913–934, 2001.

P. Cheridito, H. Kawaguchi, and M. Maejima. Fractional Ornstein–Uhlenbeck processes. *Electronic Journal of Probability*, 8:1–14, Jan. 2003.

R. Cont. Long range dependence in financial markets. In *Fractals in Engineering*, pages 159–179. Springer, 2005.

R. Cont and D.-A. Fournié. Change of variable formulas for non-anticipative functionals on path space. *Journal of Functional Analysis*, 259(4):1043–1072, Aug. 2010.

R. Cont and D.-A. Fournié. Functional Itô calculus and stochastic integral representation of martingales. *The Annals of Probability*, 41(1):109–133, Jan. 2013.

R. Cont and F. R. Lim. Causal transport on path space, Dec. 2024. arXiv:2412.02948.

R. Cont and N. Perkowski. Pathwise integration and change of variable formulas for continuous paths with arbitrary regularity. *Transactions of the American Mathematical Society, Series B*, 6(5):161–186, 2019.

H. Cramér. *Structural and Statistical Problems for a Class of Stochastic Processes: The First Samuel Stanley Wilks Lecture at Princeton University, March 7, 1970.* Princeton University Press, 1971.

K. R. Davidson. *Nest Algebras : Triangular Forms for Operator Algebras on Hilbert Space.* Pitman Research Notes in Mathematics Series. Longman Scientific & Technical and Wiley, 1988.

K. R. Davidson. The Russo-Dye theorem in nest algebras. *Proceedings of the American Mathematical Society*, 126(10):3055–3059, 1998.

L. Decreusefond and A. Üstünel. Stochastic analysis of the fractional Brownian motion. *Potential Analysis*, 10(2):177–214, Mar. 1999.

L. Decreusefond and A. S. Üstünel. Fractional Brownian motion: Theory and applications. *ESAIM: Proceedings*, 5:75–86, 1998.

R. G. Douglas. On majorization, factorization, and range inclusion of operators on Hilbert space. *Proceedings of the American Mathematical Society*, 17(2):413–415, 1966.

D. C. Dowson and B. V. Landau. The Fréchet distance between multivariate normal distributions. *Journal of Multivariate Analysis*, 12(3):450–455, Sept. 1982.

B. Dupire. Functional Itô calculus, July 2009. Bloomberg Portfolio Research Paper No. 2009-04-FRONTIERS.

S. Eckstein and G. Pammer. Computational methods for adapted optimal transport. *The Annals of Applied Probability*, 34(1A):675–713, Feb. 2024.

I. Ekren, C. Keller, N. Touzi, and J. Zhang. On viscosity solutions of path dependent PDEs. *The Annals of Probability*, 42(1):204–236, Jan. 2014.

I. Ekren, N. Touzi, and J. Zhang. Viscosity solutions of fully nonlinear parabolic path dependent PDEs: Part I. *The Annals of Probability*, 44(2):1212–1253, Mar. 2016a.

I. Ekren, N. Touzi, and J. Zhang. Viscosity solutions of fully nonlinear parabolic path dependent PDEs: Part II. *The Annals of Probability*, 44(4):2507–2553, July 2016b.

J. Erdos. Unitary invariants for nests. *Pacific Journal of Mathematics*, 23(2):229–256, Nov. 1967.

J. A. Erdos. Operators of finite rank in nest algebras. *Journal of the London Mathematical Society*, s1-43(1):391–397, 1968.

K. Fan. Minimax theorems. *Proceedings of the National Academy of Sciences*, 39(1): 42–47, Jan. 1953.

Y. Fan, H. Park, and G. Xu. Quantifying distributional model risk in marginal problems via optimal transport, July 2023. arXiv:2307.00779.

H. Föllmer and A. Schied. Stochastic finance: An introduction in discrete time. In *Stochastic Finance*. De Gruyter, Dec. 2008.

P. K. Friz and M. Hairer. *A Course on Rough Paths: With an Introduction to Regularity Structures*. Universitext. Springer International Publishing, 2020.

S. Fuhrmann, M. Kupper, and M. Nendel. Wasserstein perturbations of markovian transition semigroups. In *Annales de l'Institut Henri Poincare (B) Probabilites et statistiques*, volume 59, pages 904–932. Institut Henri Poincaré, 2023.

R. Gao. Finite-sample guarantees for Wasserstein distributionally robust optimization: Breaking the curse of dimensionality. *Operations Research*, 71(6):2291–2306, 2023.

R. Gao and A. Kleywegt. Distributionally robust stochastic optimization with Wasserstein distance. *Mathematics of Operations Research*, 48(2):603–655, 2022.

R. Gao, R. Arora, and Y. Huang. Data-driven multistage distributionally robust linear optimization with nested distance, Oct. 2022. Optimization Online.

M. Gelbrich. On a formula for the L^2 Wasserstein metric between measures on Euclidean and Hilbert spaces. *Mathematische Nachrichten*, 147(1):185–203, 1990.

I. Gilboa and D. Schmeidler. Maxmin expected utility with non-unique prior. *J. Math. Econom.*, 18(2):141–153, 1989.

N. Gozlan, C. Roberto, P.-M. Samson, and P. Tetali. Kantorovich duality for general transport costs and applications. *J. Funct. Anal.*, 273(11):3327–3405, 2017.

M. Gunasingam and T.-K. L. Wong. Adapted optimal transport between Gaussian processes in discrete time. *Electronic Communications in Probability*, 30:1–14, Jan. 2025.

B. Han. Distributionally robust risk evaluation with a causality constraint and structural information. *Mathematical Finance*, page mafi.12466, June 2025.

M. F. Hellwig. Sequential decisions under uncertainty and the maximum theorem. *Journal of Mathematical Economics*, 25(4):443–464, Jan. 1996.

S. Herrmann and J. Muhle-Karbe. Model uncertainty, recalibration, and the emergence of delta–vega hedging. *Finance Stoch.*, 21:873–930, 2017.

S. Herrmann, J. Muhle-Karbe, and F. T. Seifried. Hedging with small uncertainty aversion. *Finance Stoch.*, 21:1–64, 2017.

T. Hida. Canonical representations of Gaussian processes and their applications. *Memoirs of the College of Science, University of Kyoto. Series A: Mathematics*, 33 (1):109–155, Jan. 1960.

T. Hida and M. Hitsuda. *Gaussian Processes*, volume 120. American Mathematical Society, 1993.

M. Hitsuda. Representation of Gaussian processes equivalent to Wiener process. *Osaka Journal of Mathematics*, 5(2):299–312, Jan. 1968.

D. N. Hoover. A characterization of adapted distribution. *The Annals of Probability*, 15(4), Oct. 1987.

D. N. Hoover and H. J. Keisler. Adapted probability distributions. *Transactions of the American Mathematical Society*, 286(1):159–201, 1984.

Y. Jiang. Duality of causal distributionally robust optimization: The discrete-time case, Jan. 2024. arXiv:2401.16556.

Y. Jiang and F. R. Lim. A transfer principle for computing the adapted Wasserstein distance between stochastic processes, June 2025. arXiv:2505.21337.

Y. Jiang and J. Obłój. Sensitivity of causal distributionally robust optimization, Aug. 2024. arXiv:2408.17109.

C. Jost. Transformation formulas for fractional Brownian motion. *Stochastic Processes and their Applications*, 116(10):1341–1357, Oct. 2006.

S. Källblad, J. Obłój, and T. Zariphopoulou. Dynamically consistent investment under model uncertainty: The robust forward criteria. *Finance and Stochastics*, 22 (4):879–918, 2018.

R. L. Karandikar. On pathwise stochastic integration. *Stochastic Processes and their Applications*, 57(1):11–18, May 1995.

C. Kipnis and C. Landim. *Scaling Limits of Interacting Particle Systems*, volume 320 of *Grundlehren Der Mathematischen Wissenschaften.* Springer, 1999.

F. H. Knight. *Risk, Uncertainty and Profit.* Boston, New York, Houghton Mifflin Company, 1921.

D. Kuhn, S. Shafiee, and W. Wiesemann. Distributionally robust optimization, May 2025. arXiv:2411.02549.

M. Kupper, M. Nendel, and A. Sgarabottolo. Risk measures based on weak optimal transport, Dec. 2023. arXiv:2312.05973.

J. Lamperti. A simple construction of certain diffusion processes. *Journal of Mathematics of Kyoto University*, 4(1):161–170, Jan. 1964.

J. Langner, A. Neufeld, and K. Park. Markov–Nash equilibria in mean-field games under model uncertainty, 2024. arXiv:2410.11652.

K. Larsen and G. Žitković. Stability of utility-maximization in incomplete markets. *Stochastic Processes and Their Applications*, 117(11):1642–1662, 2007.

D. R. Larson. Nest algebras and similarity transformations. *Annals of Mathematics*, 121(2):409–427, 1985.

R. Lassalle. Causal transport plans and their Monge–Kantorovich problems. *Stochastic Analysis and Applications*, 36(3):452–484, May 2018.

P. Lévy. A special problem of Brownian motion, and a general theory of Gaussian random functions. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 2: Contributions to Probability Theory*, volume 3.2, pages 133–176. University of California Press, Jan. 1956.

P. Lévy. Fonctions aléatoires à corrélation linéaire. *Illinois Journal of Mathematics*, 1(2):217 – 258, 1957.

J. Lévy-Véhel, E. Lutton, and C. Tricot. *Fractals in Engineering*. Springer-Verlag, 2005.

F. Maccheroni, M. Marinacci, and A. Rustichini. Ambiguity aversion, robustness, and the variational representation of preferences. *Econometrica*, 74(6):1447–1498, 2006.

A. Matoussi, D. Possamaï, and C. Zhou. Robust utility maximization in nondominated models with 2BSDE: The uncertain volatility model. *Math. Finance*, 25(2): 258–287, 2015.

R. Merton. Lifetime portfolio selection under uncertainty: The continuous time case. *Rev. Econ. Stat.*, 51:247–257, 1969.

R. Metzler and J. Klafter. The random walk's guide to anomalous diffusion: A fractional dynamics approach. *Physics Reports*, 339(1):1–77, Dec. 2000.

R. Mirmominov and J. Wiesel. A dynamic programming principle for multiperiod control problems with bicausal constraints, 2024. arXiv:2410.23927.

P. Mohajerin Esfahani and D. Kuhn. Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations. *Mathematical Programming*, 171(1):115–166, Sept. 2018.

G. M. Molchan and Y. I. Golosov. Gaussian stationary processes with asymptotically a power spectrum. *Doklady Akademii Nauk SSSR*, 184(3):546–549, 1969.

O. Mostovyi and M. Sîrbu. Sensitivity analysis of the utility maximisation problem with respect to model perturbations. *Finance and Stochastics*, 23:595–640, 2019.

M. Nendel and A. Sgarabottolo. A parametric approach to the estimation of convex risk functionals based on Wasserstein distance, Aug. 2024. arXiv:2210.14340.

A. Neufeld and M. Nutz. Robust utility maximization with Lévy processes. *Mathematical Finance*, 28(1):82–105, 2018.

A. Neufeld and J. Sester. Robust Q-learning algorithm for markov decision processes under Wasserstein uncertainty. *Automatica*, 168:111825, 2024.

S. Nietert, Z. Goldfeld, and S. Shafiee. Outlier-robust Wasserstein DRO. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 62792–62820. Curran Associates, Inc., 2023.

D. Nualart. *The Malliavin calculus and related topics*, volume 1995. Springer, 2006.

J. Obłój and J. Wiesel. Distributionally robust portfolio maximization and marginal utility pricing in one period financial markets. *Mathematical Finance*, 31(4):1454–1493, 2021.

A. Ohashi, F. Russo, and E. Shamarova. Smoothness of densities for path-dependent SDEs under Hörmander's condition. *Journal of Functional Analysis*, 281(11): 109225, Dec. 2021.

G. Pammer. A note on the adapted weak topology in discrete time. *Electronic Communications in Probability*, 29(none):1–13, Jan. 2024.

S. Peng. *Nonlinear Expectations and Stochastic Calculus under Uncertainty: With Robust CLT and G-Brownian Motion*, volume 95 of *Probability Theory and Stochastic Modelling*. Springer, 2019.

G. C. Pflug. Version-independence and nested distributions in multistage stochastic optimization. *SIAM Journal on Optimization*, 20(3):1406–1420, Jan. 2010.

G. C. Pflug and A. Pichler. A distance for multistage stochastic optimization models. *SIAM Journal on Optimization*, 22(1):1–23, Jan. 2012.

G. C. Pflug and A. Pichler. From empirical observations to tree models for stochastic optimization: Convergence properties. *SIAM Journal on Optimization*, 26(3):1715–1740, Jan. 2016.

D. R. Pitts. Factorization problems for nests: Factorization methods and characterizations of the universal factorization property. *Journal of Functional Analysis*, 79 (1):57–90, July 1988.

H. Rahimian and S. Mehrotra. Frameworks and results in distributionally robust optimization. *Open Journal of Mathematical Optimization*, 3:1–85, July 2022.

J. R. Ringrose. Super-diagonal forms for compact linear operators. *Proceedings of the London Mathematical Society*, s3-12(1):367–384, 1962.

J. R. Ringrose. On some algebras of operators. *Proceedings of the London Mathematical Society*, s3-15(1):61–83, 1965.

B. A. Robinson and M. Szölgyenyi. Bicausal optimal transport for SDEs with irregular coefficients, Sept. 2024. arXiv:2403.09941.

R. T. Rockafellar. *Convex Analysis*. Princeton University Press, Jan. 1997.

L. C. G. Rogers. Arbitrage with fractional Brownian motion. *Mathematical Finance*, 7(1):95–105, 1997.

L. Rüschendorf. The Wasserstein distance and approximation theorems. *Probability Theory and Related Fields*, 70(1):117–129, Mar. 1985.

F. Russo and P. Vallois. Forward, backward and symmetric stochastic integration. *Probability Theory and Related Fields*, 97:403–421, 1993.

N. Sauldubois and N. Touzi. First order martingale model risk and semi-static hedging, Oct. 2024. arXiv:2410.06906.

A. Schied. Optimal investments for risk- and ambiguity-averse preferences: a duality approach. *Finance and Stochastics*, 11(1):107–129, 2007.

S. Tang and F. Zhang. Path-dependent optimal stochastic control and viscosity solution of associated Bellman equations. *Discrete and Continuous Dynamical Systems*, 35(11):5521–5553, Nov. 2015.

R. Tevzadze, T. Toronjadze, and T. Uzunashvili. Robust utility maximization for a diffusion market model with misspecified coefficients. *Finance and Stochastics*, 17: 535–563, 2013.

M. Veraar. The stochastic fubini theorem revisited. *Stochastics*, 84(4):543–551, 2012.

F. Viens and J. Zhang. A martingale approach for fractional Brownian motions and related path dependent PDEs. *The Annals of Applied Probability*, 29(6):3489–3540, Dec. 2019.

C. Villani. *Optimal Transport: Old and New.* Number 338 in Grundlehren Der Mathematischen Wissenschaften. Springer, 2009.

L. Vostrikova and M. Yor. Some invariance properties (of the laws) of Ocone’ s martingales. *Séminaire de Probabilités XXXIV*, pages 417–431, 2007.

D. Wu and S. Jaimungal. Robust risk-aware option hedging. *Applied Mathematical Finance*, 30(3):153–174, 2023.

M. Xu. Risk measure pricing and hedging in incomplete markets. *Annals of Finance*, 2(1):51–71, Jan. 2006.

J. Yang, L. Zhang, N. Chen, R. Gao, and M. Hu. Decision-making with side information: A causal transport robust approach, Oct. 2022. Optimization Online.

L. Zhang, J. Yang, and R. Gao. A short and general duality proof for Wasserstein distributionally robust optimization. *Operations Research*, page opre.2023.0135, July 2024.

G. Zitković. A dual characterization of self-generation and exponential forward performances. *Ann. Appl. Probab.*, 19(6):2176–2210, 2009.