Mathematical Programming (2020) 181:377–403
https://doi.org/10.1007/s10107-019-01391-2

FULL LENGTH PAPER

Series B

Martingale characterizations of risk-averse stochastic optimization problems

Alois Pichler[1] · **Ruben Schlotter**[1]

Received: 12 February 2018 / Accepted: 19 March 2019 / Published online: 27 March 2019

Abstract
This paper addresses risk awareness of stochastic optimization problems. Nested risk measures appear naturally in this context, as they allow beneficial reformulations for algorithmic treatments. The reformulations presented extend usual dynamic equations by involving risk awareness in the problem formulation. Nested risk measures are built on risk measures, which originate by conditioning on the history of a stochastic process. We derive martingale properties of these risk measures and use them to prove continuity. It is demonstrated that stochastic optimization problems, which incorporate risk awareness via nesting risk measures, are continuous with respect to the natural distance governing these optimization problems, the nested distance.

Keywords Risk measures · Stochastic optimization · Stochastic processes

Mathematics Subject Classification 90C15 · 60B05 · 62P05

1 Introduction

Risk measures have been found useful in various disciplines of applied mathematics, particularly in mathematical finance and in stochastic optimization. Many applications involve them in various places to account for risk. It is hence natural to investigate risk measures in a multistage or dynamic optimization framework as well. One of the first occurrences of dynamic risk measures in the literature is Riedel [27], Ruszczyński and Shapiro [31] (consider also the references therein) discuss conditional risk measures.

It seems that there is no general consensus on how to incorporate risk measures in a more general framework which involves time. One of the conceptual difficulties

Special Issue Math. Prog. on "The interface between optimization and probability".

✉ Alois Pichler
alois.pichler@math.tu-chemnitz.de

[1] Technische Universität Chemnitz, 09126 Chemnitz, Germany

arising in a problem setting involving time is time consistency. In short, the decisions considered optimal at some stage of time should not be rejected from a later perspective.

Risk-averse multistage stochastic programs incorporate risk awareness in multistage decision making. These problems have been considered in Ruszczyński [30] and Dentcheva and Ruszczyński [6], while applications can be found in Philpott and de Matos [22], Philpott et al. [23] or Maggioni et al. [16], e.g., where stochastic dual dynamic programming methods are addressed, cf. also Römisch and Guigues [29], Girardeau et al. [10]. In economics, the spread between *risk-averse* and *risk-neutral* preferences is associated with a risk or insurance premium. For this, the prevailing idea of risk in these papers is the interpretation as insurance on a rolling horizon basis.

This paper introduces conditional risk functionals based on the history of the governing stochastic process. These functionals are nested to obtain risk functionals accounting for the risk at each stage of the stochastic process. We elaborate their continuity properties and for important cases we compare them with simple risk measures spanning the entire horizon as a whole.

Building on the idea in Pflug [19] (cf. also Dentcheva and Ruszczyński [7]) we introduce the nested distance via conditional probabilities. We relate these concepts by verifying that nested risk functionals are continuous with respect to the nested distance and provide an explicit expression of the modulus of continuity.

Martingales are present in stochastic optimization since its very beginning, cf. Rockafellar and Wets [28]. The approach taken here to verify the results is based on generalized martingales. They reflect the evolution of risk over time, as risk measures replace risk-neutral expectations. It is demonstrated that the nested distance, as well as nested risk measures, follow martingale characteristics in this generalized sense.

It is a consequence that risk-averse multistage stochastic programs are continuous with respect to the nested distance. The optimal solutions constitute a stochastic process, which again follows a martingale-like pattern. We finally give a verification theorem. This is a risk-averse generalization of dynamic programming equations, which are well-known from dynamic optimization. Pichler and Shapiro [26] observed recently that dynamic equations and martingales are available in exactly the situations presented.

Outline of the paper Section 3 introduces nested risk functionals after an introductory discussion (Sect. 2). Section 4 addresses the main features of the nested distance which are important and relevant to cover the discussion on continuity of the multistage stochastic programs in Sect. 5. Risk martingales are introduced in Sect. 6. We conclude with the main result in Sect. 7.

2 Notation and preliminaries

We consider the Polish spaces $\left(\Xi_t, d_t\right)$, $t \in \{1, \dots, T\}$. We shall associate $t \in \{1, \dots, T\}$ with *stage* or *time* advancing in discrete steps from 1 to T, where $T \in \{1, 2, \dots\}$ is the time horizon (terminal time) or final stage. Each space Ξ_t, $t = 1, \dots T$, contains the information revealed at time t. In what follows it will often be sufficient to consider the spaces $\Xi_t = \mathbb{R}^{m_t}$.

The product $\Xi := \Xi_{1:T} := \Xi_1 \times \cdots \times \Xi_T$ is endowed with the metric d and

$$(\Xi,\ d) \tag{1}$$

is Polish as well (for example, choose $d(x, y) := \ell_p(x, y) := \left(\sum_{t=1}^{T} d_t(x_t, y_t)^p\right)^{1/p}$). We denote elements $x \in \Xi_{1:T}$ by $x_{1:T} := x = (x_1, \ldots, x_T)$ and by pr_t the canonical (i.e., coordinate) projection $\mathsf{pr}_t(x_{1:T}) := x_{1:t}$ onto the subspace $\Xi_{1:t} := \Xi_1 \times \cdots \times \Xi_t$. To allow a compact notation we also introduce the empty tuple $x_{1:0} = ()$.

On the Borel sets $\mathcal{F}_T := \mathscr{B}(\Xi_{1:T})$ we consider the probability measure

$$P\colon \mathcal{F}_T \to [0, 1].$$

The probability measures restricted to the sub-sigma algebra $\mathcal{F}_t := \sigma(\mathsf{pr}_t)$ are the image measures defined by

$$P_t(A) := P^{\mathsf{pr}_t}(A) = P\left(A \times \Xi_{t+1} \times \cdots \times \Xi_T\right),$$

where $A \in \mathscr{B}(\Xi_{1:t})$, the Borel sigma algebra on $\Xi_{1:t}$. The sequence $\mathcal{F} := \mathcal{F}_{0:T} := (\mathcal{F}_t)_{t=0}^{T}$ is the canonical (i.e., coordinate) filtration and $(\Xi_{1:T}, \mathcal{F}_{0:T}, P)$ is a filtered probability space (a.k.a. stochastic basis), where we include the trivial sigma algebra $\mathcal{F}_0 := \{\emptyset,\ \Xi_{1:T}\}$ for completeness and convenience.

The disintegration theorem (cf. Dellacherie and Meyer [5, III-70] or Ambrosio et al. [1, Sect. 5.3]) allows 'disintegrating' the probability measure with respect to the coordinates.

Theorem 1 (Disintegration theorem) *There is a regular kernel, i.e., a P_t-a.s. uniquely defined family of measures $P(\cdot \mid x_{1:t})$ so that*

(i) *$x_{1:t} \mapsto P(B \mid x_{1:t})$ is measurable for every $B \in \mathscr{B}(\Xi_{t+1} \times \cdots \times \Xi_T)$ and*
(ii) *$P(A \times B) = \int_A P(B \mid x_{1:t})\, P_t(\mathrm{d}x_{1:t})$, where $A \in \mathscr{B}(\Xi_1 \times \cdots \times \Xi_t)$ and $B \in \mathscr{B}(\Xi_{t+1} \times \cdots \times \Xi_T)$.*

The conditional probability measures

$$P_{t+1}(\cdot \mid x_{1:t}) \ \ \textit{on} \ \ \mathscr{B}(\Xi_{t+1}) \tag{2}$$

are called (*regular*) kernels *and the substring $x_{1:t}$ is also called a* fiber.

By disintegrating the measures P_t and composing their kernels at subsequent stages we obtain the nested expressions

$$P_t(A_1 \times \cdots \times A_t) = \int_{A_1} \int_{A_2} \ldots \int_{A_t} P_t\left(\mathrm{d}x_t \mid x_{1:t-1}\right) \ldots P_2\left(\mathrm{d}x_2 \mid x_{1:1}\right) P_1(\mathrm{d}x_1) \tag{3}$$

and the conditional probability measures

$$P(A_{t+1} \times \cdots \times A_T \mid x_{1:t}) = \int_{A_{t+1}} \cdots \int_{A_T} P_T\left(\mathrm{d}x_T \mid x_{1:T-1}\right) \ldots P_{t+1}\left(\mathrm{d}x_{t+1} \mid x_{1:t}\right). \tag{4}$$

Both expressions reveal the initial probability measure P, which can be seen by substituting $t = T$ in (3) or $t = 0$ in (4).

Remark 2 The kernels derived from the projected measures (2) are conditioned on the history $x_{1:t}$ and they do depend explicitly on the entire history up to t. In the Markovian case this dependence reduces (simplifies) to

$$P_{t+1}(\cdot \mid x_{1:t}) = P_{t+1}(\cdot \mid x_t).$$

An important algorithm in stochastic optimization is Stochastic Dual Dynamic Programming (SDDP). In this context the probabilities are typically assumed to be *stagewise independent,* i.e.,

$$P_{t+1}(\cdot \mid x_{1:t}) = P_{t+1}(\cdot)$$

(cf. Goulart and da Costa [11]).

3 Conditional and nested risk measures

To define conditional risk functionals we recall the definition of *law invariant, coherent risk functionals* $\mathcal{R}\colon L \to \mathbb{R}$ defined on some vector space L of $\mathbb{R}$-valued random variables first. They satisfy the following axioms introduced by Artzner et al. [2].

- A1 Monotonicity: $\mathcal{R}(Y_0) \leq \mathcal{R}(Y_1)$, provided that $Y_0 \leq Y_1$ almost surely;
- A2 Translation equivariance: $\mathcal{R}(Y + c) = \mathcal{R}(Y) + c$ for $c \in \mathbb{R}$;
- A3 Convexity: $\mathcal{R}\big((1-\lambda)Y_0 + \lambda Y_1\big) \leq (1-\lambda)\mathcal{R}(Y_0) + \lambda\mathcal{R}(Y_1)$ for $\lambda \in (0,1)$;
- A4 Positively homogeneity: $\mathcal{R}(\lambda Y) = \lambda\mathcal{R}(Y)$ for $\lambda \geq 0$;
- A5 Law invariance: $\mathcal{R}(Y) = \mathcal{R}(Y')$, whenever Y and Y' have the same law, i.e., $P(Y \leq y) = P(Y' \leq y)$ for all $y \in \mathbb{R}$.

We shall make frequently use of the following proposition, which is an immediate consequence of the monotonicity Axiom A1. For a mathematical rigorous and formal definition of the essential infimum we refer to Karatzas and Shreve [14, Appendix A].

Proposition 3 *Consider a family $Y_\iota \in L$, $\iota \in I$, which is uniformly bounded from below, $c \leq Y_\iota$ for some $c \in \mathbb{R}$. The essential infimum of the family of random variables apparently satisfies $c \leq \operatorname{ess\,inf}_{\iota' \in I} Y_{\iota'} \leq Y_\iota$ for every $\iota \in I$. Hence, $\mathcal{R}(\operatorname{ess\,inf}_{\iota' \in I} Y_{\iota'})$ is finite and well-defined and by the monotonicity Axiom A1, $\mathcal{R}(\operatorname{ess\,inf}_{\iota' \in I} Y_{\iota'}) \leq \mathcal{R}(Y_\iota)$. Consequently we have that*

$$\mathcal{R}\left(\operatorname*{ess\,inf}_{\iota' \in I} Y_{\iota'}\right) \leq \inf_{\iota \in I} \mathcal{R}(Y_\iota).$$

The Average Value-at-Risk at level $\alpha \in [0, 1)$ defined on L^1 by

$$\mathsf{AV@R}_\alpha(Y) := \inf_{q\in\mathbb{R}} \left\{ q + \frac{1}{1-\alpha}\,\mathbb{E}(Y-q)_+ \right\} \tag{5}$$

is the most prominent coherent risk functional satisfying the axioms A1–A5 above. The Average Value-at-Risk at risk level $\alpha = 0$ is the expectation,

$$\mathsf{AV@R}_0(Y) = \mathbb{E}\, Y$$

and, for $Y \in L^\infty$, the convenient setting

$$\mathsf{AV@R}_1(Y) := \lim_{\alpha \nearrow 1} \mathsf{AV@R}_\alpha(Y) = \operatorname{ess\,sup} Y$$

continuously extends the Average Value-at-Risk to $\alpha = 1$.

The Average Value-at-Risk turns out to be of central importance, it can be interpreted as an extreme point in the set of risk functionals and, similarly to Choquet's representation, every law invariant, coherent risk functional is a convex combination of AV@Rs. The statement is a consequence of the Fenchel–Moreau theorem in convex analysis (cf. Föllmer and Schied [9, Lemma 4.55] or Shapiro et al. [35], Shapiro [32], Pichler and Shapiro [25]). In addition, note that Jouini et al. [12] insure that every law invariant risk functional on L^∞ is automatically lower semicontinuous. The following general representation (Kusuoka's representation, cf. Kusuoka [15]) highlights this relation.

Definition 4 A function $\sigma\colon [0, 1) \to \mathbb{R}$ is a *distortion function*, if $\sigma(\cdot)$ is non-decreasing, $\sigma(\cdot) \geq 0$ and $\int_0^1 \sigma(u)\mathrm{d}u = 1$.

Proposition 5 (Kusuoka's representation, cf. Pflug and Pichler [20]) *Every law invariant, coherent risk functional $\mathcal{R}\colon L^\infty \to \mathbb{R}$ has the representation*

$$\mathcal{R}(Y) = \sup_{\sigma\in\mathcal{S}} \mathcal{R}_\sigma(Y), \tag{6}$$

where $\mathcal{S}$ is an appropriate collection of distortion functions and

$$\mathcal{R}_\sigma(Y) := \sup\left\{ \mathbb{E}\, Y\zeta \;\middle|\; \begin{array}{l} \zeta \geq 0,\ \mathbb{E}\,\zeta = 1 \text{and} \\ \mathit{AV@R}_\alpha(\zeta) \leq \frac{1}{1-\alpha}\int_\alpha^1 \sigma(u)\mathrm{d}u \text{ for all } \alpha \in (0,1) \end{array} \right\}. \tag{7}$$

In applications, as well as in what follows we restrict the domain of the risk functional to $L = L^\infty(\Xi)$, although a larger domain is occasionally possible (cf. Pichler [24]).

Example 6 The Kusuoka representation of the Average Value-at-Risk according to Proposition 5 is given by $\mathcal{S} = \{\sigma_\alpha(\cdot)\}$, where the distortion function is $\sigma_\alpha(u) := \begin{cases} \frac{1}{1-\alpha} & \text{if } u \geq \alpha, \\ 0 & \text{else.} \end{cases}$

The representation of the distortion risk functional (6) implicitly involves the probability measure P via the expectation $\mathbb{E}$ and the Average Value-at-Risk in (7). We want to make the probability measure P explicit by rewriting (6) as

$$\mathcal{R}(Y) = \mathcal{R}_{\mathcal{S};P}(Y) := \sup\left\{ \mathbb{E}_P\, Y\zeta \;\middle|\; \begin{array}{l} \zeta \geq 0,\ \mathbb{E}_P\,\zeta = 1 \text{ and} \\ \mathsf{AV@R}_{\alpha;P}(\zeta) \leq \frac{1}{1-\alpha}\int_\alpha^1 \sigma(u)\mathrm{d}u,\ \alpha \in (0,1) \\ \quad \text{for some } \sigma(\cdot) \in \mathcal{S} \end{array} \right\}, \tag{8}$$

where the expectation in $\mathsf{AV@R}_{\alpha;P}$ is with respect to the probability measure P as well, cf. (5). The probability measures considered are on the Borel σ-algebra of the metric space (Ξ, d).

3.1 Conditional risk measures

To define conditional versions of risk measures on product spaces we employ the conditional measures available by the disintegration theorem, Theorem 1.

Definition 7 Let $\mathcal{S}_{t+1}$ be a collection of distortion functions and $Y \in L^\infty$. The *conditional risk measure* or risk *measure conditioned on the fiber* $x_{1:t}$ of the regular kernels of the probability measure P is

$$\mathcal{R}_{\mathcal{S}_{t+1}}(Y \mid x_{1:t}) := \sup_{\sigma \in \mathcal{S}_{t+1}} \mathcal{R}_{\sigma;P(\cdot|x_{1:t})}(Y). \tag{9}$$

Notice that $Y\colon \Xi_{1:T} \to \mathbb{R}$ is a random variable. As a consequence of Theorem 1(i) and the representations (6) and (7), the mapping

$$\begin{aligned} \mathcal{R}_{\mathcal{S}_{t+1}}(Y \mid \cdot)\colon \Xi_{1:t} &\to \mathbb{R} \\ x_{1:t} &\mapsto \mathcal{R}_{\mathcal{S}_{t+1}}\big(Y \mid x_{1:t}\big) \end{aligned} \tag{10}$$

is a random variable on $\Xi_{1:t}$, which is P_t a.s. well-defined and measurable with respect to $\mathcal{F}_t$ if $\mathcal{S}_{t+1}$ is finite. For $t = 0$, the conditional risk functional (9) is

$$\mathcal{R}_{\mathcal{S}_1}(Y \mid x_{1:0}) = \mathcal{R}_{\mathcal{S}_1}(Y) = \sup_{\sigma \in \mathcal{S}_1} \mathcal{R}_{\sigma;P}(Y) = \mathcal{R}_{\mathcal{S}_1}(Y),$$

a deterministic number.

3.2 Nested risk measures

The conditional risk measures (9) are well-defined on a fiber $x_{1:t}$. As each risk functional (10) is a random variable, they can be combined and considered in the following recursive, or nested way.

Definition 8 (*Nested risk functional*) Let $s, t \in \{1, \dots, T\}$ with $s < t$ and $Y \in L^\infty(\Xi_{1:T})$. The *nested risk functional* for a sequence $\mathcal{S}_{s+1:t} := \mathcal{S}_{s+1} \times \cdots \times \mathcal{S}_t$ of collections of distortion functionals is

$$\mathcal{R}_{\mathcal{S}_{s+1:t}}(Y \mid x_{1:s}) := \mathcal{R}_{\mathcal{S}_{s+1}}\Big(\dots \mathcal{R}_{\mathcal{S}_{t-1}}\big(\mathcal{R}_{\mathcal{S}_t}(Y \mid x_{1:t-1}) \mid x_{1:t-2}\big) \cdots \mid x_{1:s}\Big). \tag{11}$$

Remark 9 The nested risk functional $\mathcal{R}_{\mathcal{S}_{1:T}}(\cdot)$ maps real-valued random variables $Y\colon \Xi \to \mathbb{R}$ defined on Ξ to the real line. The nested risk functional satisfies generalizations of the axioms A1–A4, but it is *not* law invariant any longer, i.e., A5 is not necessarily satisfied.

The construction employed in Shapiro [33] to discuss rectangular sets is similar to nested risk measure given in Definition 8 above. Indeed, they can be recovered by choosing the feasible set as given in the general representation (8). A major difference is given by the fact that law invariant risk functionals have the Kusuoka representation (8), which is not the case for more general risk functionals.

Importantly, the nested risk measures are recursive as specified in the following proposition.

Proposition 10 *The nested risk functional $\mathcal{R}_{\mathcal{S}_{t+1:T}}$ is recursive, it holds that*

$$\mathcal{R}_{\mathcal{S}_{t+1:T}}(Y \mid x_{1:t}) = \mathcal{R}_{\mathcal{S}_{t+1:s}}\left(\mathcal{R}_{\mathcal{S}_{s+1:T}}(Y \mid x_{1:s}) \mid x_{1:t}\right) \tag{12}$$

whenever $0 \le t < s < T$.

Proof The assertion is an immediate consequence of the recursion (11) in Definition 8. □

Example 11 (*Conditional expectation*) The risk-neutral special case is given by choosing the simplest distortion functions $\mathcal{S}_{t+1} = \{\mathbb{1}\}$, i.e., the distortions consisting only of the constant function $\sigma(\cdot) = \mathbb{1}(\cdot) = 1$. In this case the risk functional (9) is

$$\mathcal{R}_{\mathcal{S}_{t+1};P}(Y \mid x_{1:t}) = \mathbb{E}(Y \mid x_{1:t}),$$

i.e.,

$$\mathcal{R}_{\mathcal{S}_{t+1};P}(Y \mid \cdot) = \mathbb{E}^{|\mathcal{F}_t}(Y)$$

(recall that $\mathbb{E}^{|\mathcal{F}_t}(Y)$ is indeed an $\mathcal{F}_t$ random variable). The recursion (12) reflects the tower property of the conditional expectation.

Definition 12 (*Nested Average Value-at-Risk, cf. Pflug and Römisch* [21]) The nested Average Value-at-Risk for $\alpha_{s+1:t} \in [0, 1]$ is a composition of AV@Rs at risk levels dependent on the state t where the α_{s+1} can be chosen as $\mathcal{F}_s$-measurable. More explicitly, we set

$$\begin{aligned}&\mathsf{nAV@R}_{\alpha_{s+1:t}}(Y \mid x_{1:s})\\ &\quad := \mathsf{AV@R}_{\alpha_{s+1};P(\cdot|x_{1:s})}\Big(\dots \mathsf{AV@R}_{\alpha_{t-1};P(\cdot|x_{1:t-2})}\big(\mathsf{AV@R}_{\alpha_t;P(\cdot|x_{1:t-1})}(Y)\big)\Big).\end{aligned} \tag{13}$$

The nested Average Value-at-Risk can be bounded by the Average Value-at-Risk. Indeed, it follows from Xin and Shapiro [38, Proposition 4.2] that $\mathsf{nAV@R}_{\alpha_{1:T};P}(Y) \leq \mathsf{AV@R}_{\alpha}(Y)$ provided that the risk level $\alpha \in \mathbb{R}$ satisfies $\alpha \geq 1-(1-\alpha_1)\cdot\cdots\cdot(1-\alpha_T)$ and all risk levels α_t are deterministic.

4 The distance adapted to nested risk measures

Generalizing the concept of distance from probability spaces to filtered probability spaces corresponds to generalizing the distance from random variables to stochastic processes. As a metric for probability measures we recall the Wasserstein distance first here, which we then generalize to a metric of stochastic processes.

4.1 Wasserstein metric

Consider the Polish space (Ξ, d) and probability measures

$$P,\ \tilde{P}\colon \mathcal{F} \to [0,1]$$

on the Borel sigma algebra $\mathcal{F} := \mathscr{B}(\Xi)$.

Definition 13 (*Wasserstein metric*) Let P and $\tilde{P}$ be probability measures on Ξ and $r \in [1,\infty)$ and $c : \Xi \times \Xi \to \mathbb{R}$ a lower semicontinuous cost function. The *Wasserstein metric of order r* with respect to the lower semicontinuous cost function $c\colon \Xi \times \Xi \to [0,\infty)$ is

$$w_r(P,\tilde{P};\ c) := \inf_{\pi} \left(\mathbb{E}_\pi\, c^r\right)^{1/r} = \inf_{\pi} \left(\iint_{\Xi\times\Xi} c(x,y)^r\, \pi(\mathrm{d}x,\mathrm{d}y) \right)^{1/r}, \tag{14}$$

where the infimum in (14) is among all bivariate probability measures $\pi \in \mathcal{P}(\Xi \times \Xi)$ with marginals P and $\tilde{P}$, i.e.,

$$\pi(A \times \Xi) = P(A), \quad A \in \mathscr{B}(\Xi) \quad \text{and} \tag{15}$$

$$\pi(\Xi \times B) = \tilde{P}(B), \quad B \in \mathscr{B}(\Xi). \tag{16}$$

For the Wasserstein distance of order $r = 1$ we shall also write simply $w(P,\tilde{P})$.

Remark 14 The Wasserstein metric introduced in (14) is based on a cost functions $c(\cdot,\cdot)$ (cf. also Villani [37]). This setting slightly generalizes the usual definition, which is based on the distance function d of the space (Ξ, d) in lieu of c. In what follows, this extension will be essential. The Wasserstein distance is finite for the cost function $c = d$ if $P \in \mathcal{P} := \left\{P\colon \int_\Xi d(x,y)^r\, P(\mathrm{d}x) < \infty \text{ for some } y \in \Xi\right\}$, cf. Villani [37, Sect. 7.1]. In what follows we shall always assume that $P, \tilde{P} \in \mathcal{P}$ for some appropriate $r \geq 1$ without stating this explicitly.

4.2 The nested distance

The Wasserstein metric w_r introduced in Definition 13 is of course well defined for measures P and $\tilde{P}$ on the product space $(\Xi_{1:T}, d)$. The nested distance generalizes the Wasserstein metric by involving the filtration in addition. The filtration carries the information revealed over time. The filtration considered here is the coordinate filtration, and for this we may introduce the nested distance on coordinate basis as well, i.e., sequentially by defining the process stage by stage.

Definition 15 (*Cost process, nested distance*) Let P and $\tilde{P}$ be probability measures on $\Xi_{1:T}$, let $r \in [1, \infty)$ and let $c\colon \Xi_{1:T} \times \Xi_{1:T} \to \mathbb{R}$ be a lower semicontinuous (lsc.) function.

(i) Cost process c_t for $t = T$ down to 0:

(a) The cost function c_T on $\Xi_{1:T} \times \Xi_{1:T}$ at terminal time T is

$$c_T\big(x_{1:T}, y_{1:T}\big) := c\big(x_{1:T}, y_{1:T}\big).$$

We shall refer to c_T also as the *terminal* cost function.

(b) The cost functions c_t for $t < T$ are defined in a backwards recursive way by

$$\begin{aligned} &c_{t-1}\big(x_{1:T}, y_{1:T}\big) \\ &\quad := w_r\left(P_t\left(\cdot \mid x_{1:t-1}\right), \tilde{P}_t\left(\cdot \mid y_{1:t-1}\right); c_t\right), \quad t = T, \ldots, 1, \end{aligned} \tag{17}$$

where w_r is the Wasserstein metric of order r.

(c) The *cost-process* is the stochastic process $c = (c_t)_{t=0}^T$.

(ii) The nested distance: let $c = (c_t)_{t=0}^T$ be the cost process with terminal cost

$$c_T(\cdot) = d(\cdot), \tag{18}$$

the distance of the space $\Xi_{1:T}$ (cf. (1)). The *nested distance* of order $r \geq 1$ of the measures P and $\tilde{P}$ is

$$\mathsf{dl}_r\big(P, \tilde{P}\big) := c_0. \tag{19}$$

Remark 16 The function c_t is defined for $(x_{1:T}, y_{1:T}) \in \Xi_{1:T} \times \Xi_{1:T}$, but its definition in (17) notably involves only the truncated states $(x_{1:t}, y_{1:t}) \in \Xi_{1:t} \times \Xi_{1:t}$. The cost function c_t thus is unambiguously defined for $(x_{1:t}, y_{1:t})$, irrespective of future realization $(x_{t+1:T}, y_{t+1:T})$. It follows that c_t is $\mathcal{F}_t \otimes \mathcal{F}_t$ measurable and the cost process $(c_t)_{t=0}^T$ is adapted to the filtration $\mathcal{F} \otimes \mathcal{F}$.

In particular, c_0 is independent of the formal argument $(x_{1:T}, y_{1:T})$ [the string $x_{1:0}$ is empty for $t = 0$ in (17)] so that c_0 is a number ($c_0 = \mathsf{dl}_r\big(P, \tilde{P}\big) \in \mathbb{R}$) and the nested distance is well-defined by (19).

Remark 17 It is a consequence of Hölder's inequality that $w_r\big(P, \tilde{P}\big) \leq w_{r'}\big(P, \tilde{P}\big)$ whenever $r \leq r'$. By monotonicity of (17) we thus get that

$$\mathsf{dl}_r\big(P, \tilde{P}\big) \leq \mathsf{dl}_{r'}\big(P, \tilde{P}\big) \qquad (r \leq r'). \tag{20}$$

Remark 18 (*Relation to Wasserstein metric*) For $T = 1$ we have $\Xi_{1:T} = \Xi_1$ and there are no intermediary stages present. In this case, the nested distance reduces to the usual Wasserstein metric and it holds that

$$\mathsf{dl}_r\big(P, \tilde{P}\big) = w_r\big(P, \tilde{P};\, d\big) \qquad (T = 1).$$

Remark 19 As for the Wasserstein distance we also write $\mathsf{dl}(P, \tilde{P})$ if the order is $r = 1$ (cf. Remark 14).

An important case in practice is given by costs which occur sequentially at every stage and total costs are accumulated over time. The cost process reflects this additive property, as the following proposition outlines.

Proposition 20 (Additive cost functions) *Suppose the terminal cost function is of particular form*

$$c_T(x, y) = \ell_r(x, y) = \left(\sum_{t=1}^{T} d_t(x_t, y_t)^r\right)^{1/r}, \tag{21}$$

where d_t, $t = 1, \ldots, T$, are distance functions on $\Xi_t \times \Xi_t$ and $r \geq 1$. Then the process

$$\tilde{c}_t := \left(c_t^r - \sum_{j=1}^{t-1} d_j^r\right)^{1/r} \tag{22}$$

satisfies the recursive equations

$$\tilde{c}_{t-1}^r = d_{t-1}^r + \Big(w_r\Big(P_t\left(\cdot \mid x_{1:t-1}\right),\ \tilde{P}_t\left(\cdot \mid y_{1:t-1}\right);\ \tilde{c}_t\Big)\Big)^r \tag{23}$$

with $\tilde{c}_T = d_T$.

Further, the nested distance is

$$\mathrm{dl}_{\mathrm{r}}\big(P, \tilde{P}\big) = \tilde{c}_0.$$

Remark 21 The recursive equation (23) is actually the initial attempt in defining a distance on the nested spaces $\Xi_t \times \mathcal{P}(\Xi_{t-1})$ for the particular case $r = 1$, where $\mathcal{P}(\Xi_{t-1})$ is the set of probability measures on Ξ_{t-1}. We refer to Pflug [19] for the initial and complete discussion on nested spaces and nested distances.

Proof From (17) we have that

$$c_{t-1}\left(x_{1:T}, y_{1:T}\right)^r = w_r \left(P_t\left(\cdot \mid x_{1:t-1}\right),\ \tilde{P}_t\left(\cdot \mid y_{1:t-1}\right);\ c_t\right)^r.$$

As d_j are $\mathcal{F}_{t-1}$-measurable for for every $j < t$ it follows further that

$$c_{t-1}\left(x_{1:T}, y_{1:T}\right)^r = \sum_{j=1}^{t-1} d_j^r + w_r \left(P_t\left(\cdot \mid x_{1:t-1}\right),\ \tilde{P}_t\left(\cdot \mid y_{1:t-1}\right);\ \left(c_t^r - \sum_{j=1}^{t-1} d_j^r\right)^{1/r}\right)^r$$

and hence

$$\begin{aligned}
\tilde{c}_{t-1}^r\left(x_{1:T}, y_{1:T}\right)^r &= c_{t-1}\left(x_{1:T}, y_{1:T}\right)^r - \sum_{j=1}^{t-2} d_j^r \\
&= d_{t-1}^r + w_r \left(P_t\left(\cdot \mid x_{1:t-1}\right),\ \tilde{P}_t\left(\cdot \mid y_{1:t-1}\right);\ \left(c_t^r - \sum_{j=1}^{t-1} d_j^r\right)^{1/r}\right)^r \\
&= d_{t-1}^r + w_r \left(P_t\left(\cdot \mid x_{1:t-1}\right),\ \tilde{P}_t\left(\cdot \mid y_{1:t-1}\right);\ \tilde{c}_t\right)^r,
\end{aligned}$$

which is the assertion. □

Remark 22 It is evident that the assertion of the previous statement holds as well in case of cost functions which are nonanticipative and of the form $c_T(x, y) = \left(\sum_{t=1}^T d_t(x_{1:t}, y_{1:t})^r\right)^{1/r}$.

4.3 Characterization as a martingale

For the measure P we have given the nested expressions (3) and (4) based on kernels explicitly. In the same way one may glue together the kernels which are optimal in (17) to compute the nested distance and cost process. To this end denote the optimal kernels on $\Xi_t \times \Xi_t$ obtained in (17) by $\pi_t(\cdot \times \cdot \mid x_{1:t}, y_{1:t})$. A well-known result of Brenier [3,4] (see also McCann [17]) asserts that the Wasserstein problem (14) attains the infimum at a unique bivariate measure π for the quadratic cost function $c(x, y) = \|x - y\|^2$, if both measures P and $\tilde{P}$ have finite variance and do not give mass to small sets (cf. Villani [37, Theorem 2.12]); the measures $\pi_t(\cdot \times \cdot \mid x_{1:t}, y_{1:t})$ thus exist.

The global measure governing all kernels then is

$$\begin{aligned}
\pi\left(A \times B\right) := \iint_{A_1 \times B_1} \Big(\iint_{A_2 \times B_2} \ldots \Big(\iint_{A_T \times B_T} \pi_T(\mathrm{d}x_T, \mathrm{d}y_T \mid x_{1:T-1}, y_{1:T-1}) \Big) \\
\ldots \pi_2(\mathrm{d}x_2, \mathrm{d}y_2 \mid x_1, y_1) \Big) \pi_1(\mathrm{d}x_1, \mathrm{d}y_1),
\end{aligned} \tag{24}$$

where $A = A_1 \times \cdots \times A_T$ and $B = B_1 \times \cdots \times B_T$. The measure π is a bivariate measure on the entire space $\Xi_{1:T} \times \Xi_{1:T}$. The σ-algebra on $\Xi_{1:T}$ is generated by rectangles of the form $A = A_1 \times \cdots \times A_T \subset \Xi_1 \times \cdots \times \Xi_T$ and hence it is sufficient to consider the sets A and B in product form.

We have the following alternative characterization of the governing bivariate measure (24).

Proposition 23 *The conditional marginals of the measure π defined in* (24) *satisfy*

$$\pi\left(A \times \Xi_{1:T} \mid x_{1:t}, y_{1:t}\right) = P\left(A \mid x_{1:t}\right), \qquad A \in \mathcal{F}_T \quad \textit{and} \tag{25}$$

$$\pi\left(\Xi_{1:T} \times B \mid x_{1:t}, y_{1:t}\right) = \tilde{P}\left(B \mid y_{1:t}\right), \qquad B \in \mathcal{F}_T, \tag{26}$$

for every $t \in \{0, \ldots T-1\}$.

Proof The most inner integral in (24) satisfies

$$\iint\limits_{A_T \times \Xi_T} \pi(\mathrm{d}x_T, \mathrm{d}y_T \,|\, x_{1:T-1}, y_{1:T-1}) = \pi(A_T \times \Xi_T \,|\, x_{1:T-1}, y_{1:T-1}) = P(A_T \,|\, x_{1:T-1})$$

by construction of the measure $\pi(\cdot, \cdot \,|\, x_{1:T-1}, y_{1:T-1})$. This is (25) for the terminal time $t = T-1$.

Suppose now, by backwards inductions, that the marginal (25) is valid for $t+1$. Then

$$\begin{aligned}
&\pi\left(A_{t+1:T} \times \Xi_{t+1:T} \,|\, x_{1:t}, y_{1:t}\right) \\
&\quad = \iint\limits_{A_{t+1} \times \Xi_{t+1}} \cdots \iint\limits_{A_T \times \Xi_T} \pi(\mathrm{d}x_T, \mathrm{d}y_T \,|\, x_{1:T-1}, y_{1:T-1}) \ldots \pi(\mathrm{d}x_{t+1}, \mathrm{d}y_{t+1} \,|\, x_{1:t}, y_{1:t}) \\
&\quad = \iint\limits_{A_{t+1} \times \Xi_{t+1}} \pi(A_{t+2:T} \times \Xi_{t+2:T} \,|\, x_{1:t+1}, y_{1:t+1}) \pi(\mathrm{d}x_{t+1}, \mathrm{d}y_{t+1} \,|\, x_{1:t}, y_{1:t}) \\
&\quad = \iint\limits_{A_{t+1} \times \Xi_{t+1}} P(A_{t+2:T} \,|\, x_{1:t+1}) \pi(\mathrm{d}x_{t+1}, \mathrm{d}y_{t+1} \,|\, x_{1:t}, y_{1:t}) \\
&\quad = \int_{A_{t+1}} P(A_{t+2:T} \,|\, x_{1:t+1}) P(\mathrm{d}x_{t+1} \,|\, x_{1:t}) \\
&\quad = P(A_{t+1:T} \,|\, x_{1:t}),
\end{aligned}$$

where we have used the decomposition (24), the induction hypothesis, the decomposition (4) and the setting $A_{t+1:T} := A_{t+1} \times A_{t+2:T}$. We conclude that identity (25) is valid for all t and $A = A_1 \times \cdots \times A_T$. As above, the σ-algebra on $\Xi_{1:T}$ is generated by rectangles of the form $A = A_1 \times \cdots \times A_T \subset \Xi_1 \times \cdots \times \Xi_T$ and hence it is sufficient to consider the sets A in product form.

The remaining identity (26) follows analogously. □

The process $(c_t)_{t=0}^T$ given in Definition 15 is constructed by recursively averaging with respect to the conditional measures of π given in (24). We thus have the following characterization as a martingale.

Theorem 24 (Martingale characterization) *Let $\pi(\cdot,\cdot)$ be the measure defined in (24) and $r \geq 1$. Then the cost process $c = (c_t^r)_{t=1}^T$ is a martingale with respect to π and the canonical filtration, i.e.,*

$$c_t^r = \mathbb{E}_\pi \left(c_{t+1}^r \mid \mathcal{F}_t \otimes \mathcal{F}_t \right).$$

Proof By definition of the process c_t in (17) we have that

$$c_{t-1}(x_{1:T}, y_{1:T})^r = \iint\limits_{\Xi_t \times \Xi_t} c_t(x_{1:T}, y_{1:T})^r \pi(\mathrm{d}x_t, \mathrm{d}y_t \mid x_{1:t-1}, y_{1:t-1}),$$

where $\pi(\cdot,\cdot \mid x_{1:t-1}, y_{1:t-1})$ is the measure with marginals $P(\cdot \mid x_{1:t-1})$ and $\tilde{P}(\cdot \mid y_{1:t-1})$, resp., for which the Wasserstein distance attains the infimum in (17). This is the conditional martingale property for the fibers $(x_{1:t-1}, y_{1:t-1})$. The assertion follows as the measure π in (24) combines these optimal, conditional measures. □

Corollary 25 (Alternative characterization) *The nested distance is given by*

$$\mathsf{dl}_r\left(P, \tilde{P}\right) = \inf_\pi \left(\mathbb{E}_\pi\, d^r\right)^{1/r} = \inf_\pi \left(\iint\limits_{\Xi \times \Xi} d(x,y)^r \pi(\mathrm{d}x, \mathrm{d}y) \right)^{1/r},$$

where the infimum is among all probability measures $\pi \in \mathcal{P}(\Xi \times \Xi)$ satisfying the conditional marginal constraints (25)–(26). The infimum is attained for the measure π defined in (24).

Proof Let $\pi(\cdot \mid \cdot)$ satisfy the marginals (25)– 26). Then every conditional measure $\pi(\cdot,\cdot \mid x_{1:t-1}, y_{1:t-1})$ satisfies the constraints (15)–(16) to compute the Wasserstein distance. It follows that $\mathsf{dl}_r(P, \tilde{P})^r \leq \mathbb{E}_\pi\, d^r$.

The measure π defined in (24) satisfies the constraints (25)–(26) as well. However, we have from Theorem 24 that c_t^r is a martingale. The assertion follows from the tower property of the conditional expectation, as $c_T^r = d^r$ and

$$\begin{aligned}
\mathsf{dl}_r(P, \tilde{P})^r = c_0^r &= \mathbb{E}_\pi \left(\ldots \mathbb{E}_\pi \left(c_{t+1}^r \mid \mathcal{F}_t \otimes \mathcal{F}_t \right) \cdots \mid \mathcal{F}_1 \otimes \mathcal{F}_1 \right) \\
&= \mathbb{E}_\pi \left(\ldots \mathbb{E}_\pi \left(\ldots \mathbb{E}_\pi \left(d^r \mid \mathcal{F}_T \otimes \mathcal{F}_T \right) \cdots \mid \mathcal{F}_t \otimes \mathcal{F}_t \right) \cdots \mid \mathcal{F}_1 \otimes \mathcal{F}_1 \right) \\
&= \mathbb{E}_\pi\, d^r;
\end{aligned}$$

hence the result. □

For additive cost functions the distance of the individual stages have to be taken care of. The following corollary describes the process in analogy to Proposition 20 above.

Corollary 26 (Additive cost functions) *Let $\pi(\cdot,\cdot)$ be the optimal measure (24) and c_T the additive cost function (21) for $r \geq 1$. Then the process*

$$\tilde{c}_t^r + \sum_{j=1}^{t-1} d_j^r$$

is a martingale with respect to the measure π (cf. (22)).

Proof This is immediate as $\tilde{c}_t^r = c_t^r - \sum_{j=1}^{t-1} d_j^r$ by definition of the process (22) and as c_t is a martingale by Theorem 24. □

5 Continuity properties

The risk functionals defined in (6) above are continuous with respect to the Wasserstein distance. We generalize the results here and verify that nested risk functionals are continuous with respect to the nested distance. This section elaborates the modulus of continuity.

Proposition 27 (Continuity of risk functionals) *Let $\mathcal{R}_{\mathcal{S}}$ be a general risk functional according to (8). Suppose that the random variables $Y, \tilde{Y} : \Xi \to \mathbb{R}$ satisfy*

$$Y(x) - \tilde{Y}(y) \leq L \cdot d(x,y)^{\beta} \tag{27}$$

for some $\beta \leq 1$. Then

$$\begin{aligned} \mathcal{R}_{\mathcal{S};P}(Y) - \mathcal{R}_{\mathcal{S};\tilde{P}}(\tilde{Y}) &\leq L \cdot \sup_{\sigma \in \mathcal{S}} \|\sigma\|_q \cdot w_{\beta r}(P,\tilde{P})^{\beta} \\ &\leq L \cdot \sup_{\sigma \in \mathcal{S}} \|\sigma\|_q \cdot w_r(P,\tilde{P})^{\beta}, \end{aligned}$$

where $q \in (1,\infty]$ is the Hölder conjugate exponent of r (the order of the Wasserstein metric) for which $\frac{1}{q} + \frac{1}{r} = 1$.

Proof Let $\zeta \geq 0$ with $\mathbb{E}\,\zeta = 1$ be chosen so that the supremum in (8) is attained up to $\varepsilon > 0$, i.e., $\mathbb{E}_P\, Y\zeta > \mathcal{R}_{\mathcal{S};P}(Y) - \varepsilon$. Let π have marginals P and $\tilde{P}$ and extend ζ and $\tilde{Y}$ to the product space by setting $\zeta(x,y) := \zeta(x)$ and $\tilde{Y}(x,y) := \tilde{Y}(y)$. Note that $\mathbb{E}_\pi\, \zeta = \mathbb{E}_P\, \zeta = 1$ and thus

$$\mathcal{R}_{\mathcal{S};\tilde{P}}(\tilde{Y}) = \mathcal{R}_{\mathcal{S};\pi}(\tilde{Y}) \geq \mathbb{E}_\pi\, \tilde{Y}\zeta.$$

It follows from Hölder's inequality that

$$\begin{aligned}\mathcal{R}_{\mathcal{S};P}(Y) - \varepsilon - \mathcal{R}_{\mathcal{S};\tilde{P}}(\tilde{Y}) &\leq \iint_{\Xi\times\Xi} \Big(Y(x) - \tilde{Y}(y)\Big)\,\zeta(x,y)\pi(\mathrm{d}x,\mathrm{d}y)\\ &\leq L \iint_{\Xi\times\Xi} d\,(x,y)^{\beta}\,\zeta(x,y)\pi(\mathrm{d}x,\mathrm{d}y)\\ &\leq L\left(\iint_{\Xi\times\Xi} d\,(x,y)^{\beta r}\,\pi(\mathrm{d}x,\mathrm{d}y)\right)^{1/r}\left(\mathbb{E}_{\pi}\,\zeta^{q}\right)^{1/q}. \quad (28)\end{aligned}$$

Now recall that $\zeta(x,y) = \zeta(x)$ depends only on the first variable and thus $(\mathbb{E}_{\pi}\,\zeta^{q})^{1/q} = (\mathbb{E}_{P}\,\zeta^{q})^{1/q} = \|\sigma\|_{q}$, where $\sigma(\cdot) := F_{\zeta}^{-1}(\cdot) \in \mathcal{S}$ is the generalized inverse distribution function. We obtain the desired result by taking the infimum in (28) over all possible measures with marginals P and $\tilde{P}$ and after letting $\varepsilon \to 0$.

For the remaining inequality observe that

$$\left(\mathbb{E}_{\pi}\,d^{\beta r}\right)^{1/\beta r} = \|d\|_{\beta r} \leq \|d\|_{r} = \left(\mathbb{E}_{\pi}\,d^{r}\right)^{1/r}$$

by Hölder's inequality, so that

$$(28) \leq L\left(\iint_{\Xi\times\Xi} d\,(x,y)^{r}\,\pi(\mathrm{d}x,\mathrm{d}y)\right)^{\beta/r} \cdot \sup_{\sigma\in\mathcal{S}} \|\sigma\|_{q}\,.$$

This is the assertion. □

Corollary 28 (Continuity of the Average Value-at-Risk) *Suppose that* $Y(x) - \tilde{Y}(y) \leq L \cdot d(x,y)$. *Then*

$$\mathrm{AV@R}_{\alpha;P}(Y) - \mathrm{AV@R}_{\alpha;\tilde{P}}(\tilde{Y}) \leq \frac{L}{1-\alpha}\,w\left(P,\tilde{P};\,d\right).$$

Proof This is a special case of Proposition 27 for $r = 1$ and $q = \infty$ (cf. Example 6). □

Theorem 29 (Continuity of nested risk functionals) *Suppose that the function* $Y\colon \Xi \to \mathbb{R}$ *is Hölder continuous with constant* L *and exponent* $\beta \leq 1$,

$$|Y(x) - Y(y)| \leq L \cdot d(x,y)^{\beta}.$$

Then the nested risk functional $\mathcal{R}_{\mathcal{S}_{1:T}}(Y)$ *is continuous with respect to the nested distance, it holds that*

$$\left|\mathcal{R}_{\mathcal{S}_{1:T};P}(Y) - \mathcal{R}_{\mathcal{S}_{1:T};\tilde{P}}(Y)\right| \leq \sup_{\sigma\in\mathcal{S}_t,\,t=1,\ldots T} \|\sigma_1\|_{q} \cdot \ldots \|\sigma_T\|_{q} \cdot L \cdot \mathrm{dl}_{r}\left(P,\tilde{P}\right)^{\beta}.$$

Proof We infer from Proposition 27 with $\tilde{Y} = Y$ that

$$\mathcal{R}_{\mathcal{S}_T;P(\cdot|\,x_{1:T-1})}(Y) - \mathcal{R}_{\mathcal{S}_T;\tilde{P}(\cdot|\,y_{1:T-1})}(Y) \leq L \cdot \sup_{\sigma_T \in \mathcal{S}_T} \|\sigma_T\|_q \cdot w_r \left(P\left(\cdot|\,x_{1:T-1}\right), \tilde{P}\left(\cdot|\,y_{1:T-1}\right); c_T \right)^{\beta}, \tag{29}$$

where the terminal cost function is the distance as in the definition of the nested distance (cf. (18)),

$$c_T = d.$$

Define the random variables

$$Y_{T-1}(x_{1:T-1}) := \mathcal{R}_{\mathcal{S}_T;P(\cdot|\,x_{1:T-1})}(Y) \quad \text{and} \quad \tilde{Y}_{T-1}(y_{1:T-1}) := \mathcal{R}_{\mathcal{S}_T;\tilde{P}(\cdot|\,y_{1:T-1})}(Y),$$

so that we have

$$Y_{T-1}(x_{1:T-1}) - \tilde{Y}_{T-1}(y_{1:T-1}) \leq L \cdot \sup_{\sigma_T \in \mathcal{S}_T} \|\sigma_T\|_q \cdot c_{T-1}\left(x_{1:T},\, y_{1:T}\right)^{\beta}$$

by (29) and the definition of the process c_t in (17). The random variables Y_{T-1} and $\tilde{Y}_{T-1}$ thus satisfy the condition (27) with respect to the cost function c_{T-1}. So we may again apply Proposition 27 to the measures $P\left(\cdot \mid x_{1:T-2}\right)$ and $\tilde{P}\left(\cdot \mid y_{1:T-2}\right)$ and repeating this procedure for $t = T-2$ down to $t = 0$ gives

$$\mathcal{R}_{\mathcal{S}_{1:T};P}(Y) - \mathcal{R}_{\mathcal{S}_{1:T};\tilde{P}}(Y) \leq \sup_{\sigma_t \in \mathcal{S}_t,\, t=1,\ldots T} \|\sigma_1\|_q \cdot \ldots \|\sigma_T\|_q \cdot L \cdot c_0^{\beta},$$

with terminal cost function $c_T = d$. We have that $c_0 = \mathsf{dl}_r\left(P, \tilde{P}\right)$ and thus

$$\mathcal{R}_{\mathcal{S}_{1:T};P}(Y) - \mathcal{R}_{\mathcal{S}_{1:T};\tilde{P}}(Y) \leq \sup_{\sigma_t \in \mathcal{S}_t,\, t=1,\ldots T} \|\sigma_1\|_q \cdot \ldots \|\sigma_T\|_q \cdot L \cdot \mathsf{dl}_r\left(P, \tilde{P}\right)^{\beta}.$$

The result follows finally by exchanging the probability measures P and $\tilde{P}$. □

Corollary 30 (Continuity of the nested Average Value-at-Risk) *Suppose that Y is Lipschitz continuous with constant L. Then the nested Average Value-at-Risk, $\mathsf{nAV@R}$, is continuous with respect to the nested distance dl. More precisely, it holds that*

$$\left| \mathsf{nAV@R}_{\alpha_{1:T};P}(Y) - \mathsf{nAV@R}_{\alpha_{1:T};\tilde{P}}(Y) \right| \leq \frac{L}{1-\alpha}\, \mathsf{dl}_r\left(P, \tilde{P}\right)$$

for every $r \geq 1$, where $\alpha \geq 1 - (1-\alpha_1) \cdot \ldots (1-\alpha_T)$ (cf. (13)).

Proof The statement for $r = 1$ is immediate by the definition of the nested Average Value-at-Risk, Corollary 28 and Theorem 29. The statement for general $r \geq 1$ follows from (20). □

6 Dynamic equations and the martingale property

In what follows we consider multistage optimization problems with cost function

$$Q\colon \mathcal{Z}_{0:T} \times \Xi_{1:T} \to \mathbb{R},$$

where a sequence of subsequent decisions $z_t \in \mathcal{Z}_t$, $t = 0, \dots T$, is chosen from $\mathcal{Z}_{0:T} = \mathcal{Z}_0 \times \cdots \times \mathcal{Z}_T$. To account for risk-averse decision making under uncertainty we involve risk functionals at each stage.

Definition 31 (*Policy*) The random variable $z_t\colon \Xi \to \mathcal{Z}_t$ is a random *policy* or *decision* at time t, $t = 0, \dots T$. The decision z_t is *nonanticipative* (or *adapted*) if $z_t\colon \Xi \to \mathcal{Z}_t$ is $\mathcal{F}_t$-measurable for every $t = 0, \dots, T$, abbreviated by $z_t \lhd \mathcal{F}_t$. The function $z\colon \Xi \to \mathcal{Z}_{1:T}$ with $z(x)_t := z_t(x)$ is nonanticipative (adapted; in short, $z \lhd \mathcal{F}$), if each component z_t is nonanticipative for every $t = 0, \dots T$.

Remark 32 It is a consequence of the Doob–Dynkin lemma that z_t is nonanticipative if it depends solely on the information available at time $t \in \{0, \dots, T\}$, i.e., if $z_t(x_{1:T}) = \tilde{z}_t(x_{1:t})$ for some measurable function $\tilde{z}_t\colon \Xi_{1:t} \to \mathcal{Z}_t$ (cf. Kallenberg [13, Lemma 1.13] or Shiryaev [36, Theorem II.4.3]). As the filtration $\mathcal{F} = \left(\mathcal{F}_t\right)_{t=0}^T$ is the coordinate filtration it follows that every nonanticipativative random decision $z \lhd \mathcal{F}$ can be written explicitly as

$$z_{0:T}(x_{1:T}) = \begin{pmatrix} z_0 \\ z_1(x_1) \\ z_2(x_1, x_2) \\ \vdots \\ z_T(x_1, \dots, x_T) \end{pmatrix}$$

for adequate, measurable functions $z_t\colon \Xi_{0:t} \to \mathcal{Z}_t$.

Definition 33 (*Multistage optimization*) Let $Q\colon \mathcal{Z}_{0:T} \times \Xi_{1:T} \to \mathbb{R} \cup \{\infty\}$ be a lsc. cost function. The risk-averse multistage optimization problem is

$$\inf_{z_{0:T} \lhd \mathcal{F}_{0:T}} \mathcal{R}_{\mathcal{S}_{1:T}}\Big(Q\big(z_{0:T}(\cdot); \cdot\big)\Big), \tag{30}$$

where the infimum is among all adapted policies $z \lhd \mathcal{F}$. We emphasize and indicated the random component in (30) by '$\cdot$'.

Remark 34 To avoid confusions or ambiguities regarding the arguments of the function Q we separate the arguments $z \in \mathcal{Z}_{0:T}$ and $x \in \Xi_{1:T}$ explicitly and write

$Q(z; x)$. This will turn out helpful in what follows, for example in expressions as $Q(z_{0:t-1}, z_{t:T};\ x_{1:t}, x_{t+1:T})$.

Remark 35 Constraints of the form $z_{0:t}(x_{1:t}) \in \mathscr{Z}_t(x_{1:t}) \subseteq \mathcal{Z}_t$ for some multifunction $\mathscr{Z}_t(\cdot)$ appear naturally in applications involving optimization under uncertainty. They are easily incorporated in the problem formulation (30) just by employing the function $Q(z_{0:T}, x_{1.T}) + \mathbb{1}_{\mathscr{Z}_t(x_{1:t})}(z_{0:T})$ instead of Q, where $\mathbb{1}_A(\cdot)$ is the characteristic function of the set A. This setting is not advisable for real world implementations, but convenient for the conceptual treatment envisaged here.

The multistage problem (30) thus consists in finding optimal functions $z_0,\ z_1(\cdot), \ldots, z_T(\cdot)$ (only z_0 is deterministic) and therefore can be considered as *optimization on function spaces*.

6.1 The essential infimum

We shall make use of the following interchangeability principle, also cf. Ruszczyński and Shapiro [31, Remark 7.1] and Shapiro [34, Proposition 6.60]. For $z \in \mathcal{Z}$ fixed, the mapping $x \mapsto Q(z, x)$ is a random variable for which we write $Q(z, \cdot)$. In what follows we discuss the expression $\inf_z Q(z, \cdot)$ and its measurability. We refer to Karatzas and Shreve [14, Appendix A] for a formal definition of the essential infimum $\operatorname{ess\,inf}_{z \in \mathcal{Z}} Q(z, \cdot)$, which is a measurable random variable as well.

Proposition 36 *Let $\mathcal{Z}$ be a vector space and consider all policies with values $z(\cdot) \in \mathcal{Z}$. Then there exists a sequence $z_n(\cdot)$ of simple functions so that*

$$\lim_{n\to\infty} Q\big(z_n(\cdot), \cdot\big) = \operatorname*{ess\,inf}_{z(\cdot)\in\mathcal{Z}} Q(z(\cdot), \cdot) \qquad \textit{almost surely} \tag{31}$$

and $Q\big(z_n(\cdot), \cdot\big)$ is nonincreasing.

Proof Denote the set of simple functions $z(\cdot) = \sum_{i=1}^k a_i\, \mathbb{1}_{A_i}(\cdot)$ by s. For $z(\cdot)$ and $z'(\cdot)$ simple functions define

$$z''(x) := \begin{cases} z(x) & \text{if } Q\big(z(x), x\big) \le Q\big(z'(x), x\big), \\ z'(x) & \text{else,} \end{cases} \tag{32}$$

which is a simple function again and measurable. (The maximization (32) actually defines a directed set or preorder on s.) It holds that $Q(z''(\cdot), \cdot) \le Q(z'(\cdot), \cdot)$ and $Q(z''(\cdot), \cdot) \le Q(z(\cdot), \cdot)$ and the set $\{Q(z(\cdot), \cdot) \colon z \in s\}$ thus is closed under pairwise minimization. It follows from Karatzas and Shreve [14, Theorem A.3] that there is a sequence $z_n(\cdot)$ of simple functions so that

$$\operatorname*{ess\,inf}_{z(\cdot)} Q\big(z(\cdot), \cdot\big) = \lim_{n\to\infty} Q\big(z_n(\cdot), \cdot\big) \qquad \text{almost everywhere}$$

and thus the assertion. □

Corollary 37 *Let s be a set of policies containing all simple functions and suppose that*

$$x \mapsto Q(z, x) \tag{33}$$

is upper semicontinuous for every $z \in \mathcal{Z}$. Then there exists a sequence $z_n(\cdot)$ of policies so that

$$\lim_{n\to\infty} Q\big(z_n(\cdot), \cdot\big) = \inf_{z\in\mathcal{Z}} Q(z, \cdot) \qquad \textit{almost everywhere.}$$

Proof The set s contains the constant functions and thus

$$\inf_{z\in\mathcal{Z}} Q(z, x) = \inf_{z(\cdot)\in\mathcal{Z}} Q\big(z(x), x\big) \qquad \text{for every } x.$$

We have that $\big\{x\colon \inf_{z\in\mathcal{Z}} Q(z, x) < \alpha\big\} = \bigcup_{z\in\mathcal{R}} \{x\colon Q(z, x) < \alpha\}$ for every $\alpha \in \mathbb{R}$ so that the additional assumptions ensure that $x \mapsto \inf_{z\in\mathcal{Z}} Q(z, x)$ is measurable. The assertion thus follows as

$$\inf_{z(\cdot)\in\mathcal{Z}} Q\big(z(\cdot), \cdot\big) = \operatorname*{ess\,inf}_{z(\cdot)\in\mathcal{Z}} Q(z(\cdot), \cdot) = \lim_{n\to\infty} Q\big(z_n(\cdot), \cdot\big),$$

where $z_n(\cdot)$ is the sequence found in Proposition 36. □

Convention 38 In what follows we shall always understand the measurable version when writing $\inf_{z\in\mathcal{Z}} Q(z, \cdot)$, i.e., we set

$$\inf_{z\in\mathcal{Z}} Q(z, \cdot) := \operatorname*{ess\,inf}_{z(\cdot)\in\mathcal{Z}} Q(z(\cdot), \cdot). \tag{34}$$

We shall further assume that Q is uniformly bounded from below, $Q \geq c$ for some $c \in \mathbb{R}$, to ensure that risk functionals are well-defined when applied to a random variable (34).

The preceding Corollary 37 provides general conditions so that (34) in the convention is void and automatically valid in these cases.

Proposition 39 (Risk functional at the essential infimum, cf. Ruszczyński and Shapiro [31, Remark 7.1] and Shapiro et al. [35]) *Suppose that $\mathcal{R}$ is continuous at* $\inf_z Q(z, \cdot)$ *with respect to convergence in L^p. Then it holds that*

$$\inf_{z(\cdot)\in\mathcal{Z}} \mathcal{R}\big(Q(z(\cdot), \cdot)\big) = \mathcal{R}\left(\inf_{z\in\mathcal{Z}} Q\big(z, \cdot\big)\right).$$

Proof The result is a consequence Lebesgue's dominated convergence theorem in view of our setting (34) and the representation as nonincreasing limit given in (31). □

6.2 Martingale properties of the value process

Section 4.3, in particular Theorem 24, characterizes the nested distance as a martingale process. This concept extends to the value process of the stochastic optimization problem when generalizing the concept of martingales. We incorporate risk awareness in the definition of the martingale term first and characterize the optimal solution of the multistage stochastic optimization problem as a martingale with respect to the risk functionals involved.

The notion of a risk-martingale already exists in the literature. Peng [18] considers continuous-time martingales with respect to a risk measure called *G-expectation* and Zhang [39] considers an optimal control problem with respect to G-expectation, where, similar to Theorem 46, the optimal value process satisfies a martingale property with respect to the underlying risk measure or nonlinear expectation. We also refer the interested reader to Zhang [39] and the references therein.

Definition 40 (*Risk martingale*) The stochastic process $v = (v_t)_{t=0}^T$ is a *submartingale* (*supermartingale*, resp.) *with respect to the risk functionals* $\mathcal{R}_{\mathcal{S}t}$ (an $\mathcal{R}$-submartingale, for short), if

$$v_t \leq \mathcal{R}_{\mathcal{S}t+1}(v_{t+1}) \text{ a.s.} \qquad (v_t \geq \mathcal{R}_{\mathcal{S}t+1}(v_{t+1}) \text{ a.s., resp.}) \tag{35}$$

for very $t \in \{0, 1, \dots T\}$. The process v_t is an $\mathcal{R}$-martingale, if (35) holds with equality.

For the expectation, $\mathcal{R} = \mathbb{E}$, the notion of an $\mathcal{R}$-martingale (sub-, supermartingale, resp.) coincides with the usual term martingale (sub-, supermartingale, resp.).

Remark 41 A process $v = (v_t)_{t=0}^T$, which is an $\mathcal{R}$-submartingale, satisfies in addition

$$v_s \leq \mathcal{R}_{S_{s+1:t}}(v_t), \qquad 0 \leq s < t < T.$$

This follows as the risk functionals $\mathcal{R}_{\mathcal{S}_t}$ are monotonic (Axiom A1) and from the recursive definition of the nested risk functional given in Definition 8.

Definition 42 (*The value process*) Let $z_{0:T} : \Xi \to \mathcal{Z}_{0:T}$ be an adapted policy. Then the *controlled value process* $v_t^z(x_{1:t})$ is defined as

$$v_t^z(x_{1:t}) := v_t(z_{0:T} \mid x_{1:t}) := \mathcal{R}_{S_{t+1:T}}\Big(Q\big(z_{0:t-1}(x_{1:t}), z_{t:T}(x_{1:t}, \cdot); x_{1:t}, \cdot\big) \mid x_{1:t}\Big) \tag{36}$$

and the *optimal value process* is defined by

$$\begin{aligned} v_t(z_{0:t-1} \mid x_{1:t}) &:= \inf_{z_{t:T} \lhd \mathcal{F}_{t:T}} v_t^z(x_{1:t}) \\ &= \inf_{z_{t:T} \lhd \mathcal{F}_{t:T}} \mathcal{R}_{S_{t+1:T}}\Big(Q\big(z_{0:t-1}(x_{1:t}), z_{t:T}(x_{1:t}, \cdot); x_{1:t}, \cdot\big) \mid x_{1:t}\Big), \qquad t = 0, \dots, T, \end{aligned} \tag{37}$$

where the infimum in (37) is among all adapted processes $z_{t:T}(x_{1:T}) = \begin{pmatrix} z_t(x_1, \ldots, x_t) \\ \vdots \\ z_T(x_1, \ldots, x_t, \ldots, x_T) \end{pmatrix}$.

Theorem 43 *Let $z = (z_t)_{t=0}^T$ be an adapted policy. Then the controlled value process $v_t^z(x_{1:t})$ is an $\mathcal{R}$-martingale with terminal value*

$$v_T^z = Q\big(z_{0:T}(\cdot); \cdot\big). \tag{38}$$

Furthermore let $z_{0:T}^* \colon \Xi \to \mathcal{Z}_{0:T}$ be an optimal policy in the multistage stochastic optimization problem (30) and $v^* = (v_t^*)_{t=0}^T$ the value process (36) associated with the policy $z_{0:T}^*$. Then v^* is an $\mathcal{R}$-martingale and the starting value v_0^* is the solution of the optimization problem (30).

Proof Choosing $t = T$ in the defining equation (36) gives $v_T^z(x_{1:T}) = Q\big(z_{0:T}(x_{1:T}); x_{1:T}\big)$ and thus (38).

Apply $\mathcal{R}_{\mathcal{S}_T}$ and it follows from (38) that

$$\begin{aligned} \mathcal{R}_{\mathcal{S}_T}\Big(v_T^z \mid x_{1:T-1}\Big) &= \mathcal{R}_{\mathcal{S}_T}\Big(Q\big(z_{0:T}(x_{1:T}); x_{1:T}\big) \mid x_{1:T-1}\Big) \\ &= \mathcal{R}_{\mathcal{S}_T}\Big(Q\big(z_{0:T-1}(x_{1:T-1}), z_{T:T}(x_{1:T-1}, \cdot); x_{1:T-1}, \cdot\big) \mid x_{1:T-1}\Big) \\ &= v_{T-1}^z(x_{1:T-1}), \end{aligned}$$

as z is adapted. This is the desired martingale property for $t = T - 1$, see (35). Apply next $\mathcal{R}_{\mathcal{S}_{T-1}}$ to the latter equation and observe that

$$\begin{aligned} v_{T-2}^z(x_{1:T-2}) &= \mathcal{R}_{\mathcal{S}_{T-1:T}}\Big(Q\big(z_{0:T-1}(x_{1:T-1}); x_{1:T-1}\big) \mid x_{1:T-2}\Big) \\ &= \mathcal{R}_{\mathcal{S}_{T-1:T}}\Big(v_{T-1}^z(x_{1:T-1}) \mid x_{1:T-2}\Big), \end{aligned}$$

which is the assertion for $t = T - 2$. The general assertion is immediate by repeatedly applying the risk functional corresponding to the individual stage. □

We will now show that the optimal value process is also an $\mathcal{R}$-martingale. First we make the following observation.

Remark 44 The optimal value process at initial time $t = 0$ is

$$v_0^* := \inf_{z_{0:T} \lhd \mathcal{F}_{0:T}} \mathcal{R}_{\mathcal{S}_{1:T}}\Big(Q\big(z_{0:T}(\cdot); \cdot\big)\Big),$$

this value coincides with the risk-averse multistage stochastic program (30) given in Definition 33. The quantity v_0^* is a deterministic number and not random. In addition, we have for $t = T$ that

$$v_T(z_{0:T-1} \mid x_{1:T}) = \inf_{z_T \lhd \mathcal{F}_T} Q\Big(z_{0:T-1}(x_{1:T}), z_T(x_{1:T}); x_{1:T}\Big),$$

so that the terminal value function does not involve a risk measure any longer and the terminal optimization problem is deterministic, i.e., not random either.

Theorem 45 (Submartingale characterization of the value process) *The optimal value process v_t (cf. (37)) is an $\mathcal{R}$-submartingale.*

Proof We have that

$$\begin{aligned}
&\mathcal{R}_{\mathcal{S}_{t+1}}(v_{t+1}) \\
&\quad = \mathcal{R}_{\mathcal{S}_{t+1}}\left(\inf_{z_{t+1:T} \lhd \mathcal{F}_{t+1:T}} \mathcal{R}_{\mathcal{S}_{t+2:T}}\Big(Q\big(z_{0:t}(x_{1:t+1}), z_{t+1:T}(x_{1:t+1}, \cdot); x_{1:t+1}, \cdot\big) \mid x_{1:t+1}\Big)\right) \\
&\quad = \inf_{z_{t+1:T} \lhd \mathcal{F}_{t+1:T}} \mathcal{R}_{\mathcal{S}_{t+1}}\left(\mathcal{R}_{\mathcal{S}_{t+2:T}}\Big(Q\big(z_{0:t}(x_{1:t+1}), z_{t+1:T}(x_{1:t+1}, \cdot); x_{1:t+1}, \cdot\big) \mid x_{1:t+1}\Big)\right) \qquad (39) \\
&\quad = \inf_{z_{t+1:T} \lhd \mathcal{F}_{t+1:T}} \mathcal{R}_{\mathcal{S}_{t+1:T}}\Big(Q\big(z_{0:t}(x_{1:t+1}), z_{t+1:T}(x_{1:t+1}, \cdot); x_{1:t+1}, \cdot\big) \mid x_{1:t+1}\Big), \qquad (40)
\end{aligned}$$

where we have employed (34) in (39).

The result follows now, as the optimal value process (37) is the infimum among all $z_{t:T} \lhd \mathcal{F}_{t:T}$, while the infimum in (40) is among $z_{t+1:T} \lhd \mathcal{F}_{t+1:T}$, which is one dimension less. □

Dynamic optimization employs verification theorems which give sufficient conditions for a solution to the optimal control problem, cf. Fleming and Soner [8, Theorems 5.1 and 5.2]. The following theorem provides the corresponding statement for the risk-averse multistage stochastic problem.

Theorem 46 (Martingale characterization, dynamic equations, verification theorem) *For the value process it holds that*

$$v_0^* = \inf_{z_{0:t} \lhd \mathcal{F}_{0:t}} \mathcal{R}_{\mathcal{S}_{1:t}}\big(v_t(z_{0:t})\big), \qquad t \in \{0, 1, \ldots, T\}.$$

More generally, for $s < t$ the recursive equations

$$v_s(z_{0:s-1}) = \inf_{z_{s:t} \lhd \mathcal{F}_{s:t}} \mathcal{R}_{\mathcal{S}_{s+1:t}}\big(v_t(z_{1:s}, z_{s+1:t})\big) \qquad (41)$$

hold true.

Proof Applying the conditional risk functional $\mathcal{R}_{\mathcal{S}_t}(\cdot \mid x_{1:t-1})$ to (37) gives

$$\begin{aligned}
&\mathcal{R}_{\mathcal{S}_t}\Big(v_t(z_{0:t-1} \mid x_{1:t}) \mid x_{1:t-1}\Big) \\
&\quad = \mathcal{R}_{\mathcal{S}_t}\left(\inf_{z_{t:T} \lhd \mathcal{F}_{t:T}} \mathcal{R}_{\mathcal{S}_{t+1:T}}\Big(Q\big(z_{0:t-1}(x_{1:t}), z_{t:T}(x_{1:t}, \cdot); x_{1:t}, \cdot\big) \mid x_{1:t}\Big) \mid x_{1:t-1}\right) \qquad (42)
\end{aligned}$$

$$= \inf_{z_{t:T} \lhd \mathcal{F}_{t:T}} \mathcal{R}_{\mathcal{S}_t}\Big(\mathcal{R}_{\mathcal{S}_{t+1:T}}\Big(Q\big(z_{0:t-1}(x_{1:t}), z_{t:T}(x_{1:t}, \cdot); x_{1:t}, \cdot\big) \mid x_{1:t}\Big) \mid x_{1:t-1}\Big)$$

$$= \inf_{z_{t:T} \lhd \mathcal{F}_{t:T}} \mathcal{R}_{\mathcal{S}_{t:T}}\Big(Q\big(z_{0:t-1}(x_{1:t}), z_{t:T}(x_{1:t}, \cdot); x_{1:t}, \cdot\big) \mid x_{1:t-1}\Big)\Big), \tag{43}$$

where we have used the monotonicity Axiom A1 and Proposition 3 to obtain "$\leq$" in (43). The converse inequality "$\geq$" involves the Lebesgue Dominated Convergence Theorem and is a consequence of Proposition 39.

At this stage take the infimum with respect to $z_{t-1} \lhd \mathcal{F}_{t-1}$ and thus

$$\begin{aligned}
&\inf_{z_{t-1} \lhd \mathcal{F}_{t-1}} \mathcal{R}_{\mathcal{S}_t}\left(v_t(z_{0:t-1} \mid x_{1:t}) | x_{1:t-1}\right) \\
&\quad = \inf_{z_{t-1:T} \lhd \mathcal{F}_{t-1:T}} \mathcal{R}_{\mathcal{S}_{t:T}}\Big(Q\big(z_{0:t-1}(x_{1:t}), z_{t:T}(x_{1:t}, \cdot); x_{1:t}, \cdot\big) \mid x_{1:t-1}\Big)\Big) \\
&\quad = v_{t-1}(z_{0:t-2} \mid x_{1:t-1}),
\end{aligned}$$

which is the martingale property of the value process $v(z)$. The remaining equation (41) follows in line with Remark 41. □

The dynamic equations derived in this section can be employed to characterize optimal solution of the multistage stochastic optimization problem. The conceptual advantage lies in the fact that each stage can be considered for its own. For this the dynamic equations can be employed in algorithms to improve suboptimal policies at each stage individually.

7 Continuity of risk-averse multistage programs

The value of the risk-averse multistage stochastic optimization problem (30) depends on the probability measure P. We shall make this explicit by writing

$$v_P := \inf_{z_{0:T} \lhd \mathcal{F}_{0:T}} \mathcal{R}_{\mathcal{S}_{1:T};P}\Big(Q\big(z_{0:T}(\cdot); \cdot\big)\Big). \tag{44}$$

It is known that the *risk-neutral* version of the multistage problem (44) is continuous with respect to changing the probability measure.

The following main result elaborates continuity of the *risk-averse* problem with respect to the nested distance and gives the modulus of continuity explicitly.

Theorem 47 Continuity of the risk-averse multistage stochastic optimization problem *Suppose that*

$$x \mapsto Q(z; x), \qquad z \in \mathcal{Z},$$

is uniformly Lipschitz, i.e.,

$$|Q(z; x) - Q(z; y)| \leq L \cdot d(x, y) \quad \textit{for all } x, y \in \Xi_{1:T} \textit{ and } z \in \mathcal{Z} \tag{45}$$

and

$$z \mapsto Q(z; x) \qquad (x \in \Xi_{1:T})$$

is convex for every x fixed. Then the risk-averse optimization problem (30) is continuous with respect to changing the probability measure. More specifically, we have that

$$\left|v_P - v_{\tilde{P}}\right| \leq \sup_{\sigma \in \mathcal{S}_t,\, t=1,\dots T} \|\sigma_1\|_q \cdot \ldots \|\sigma_T\|_q \cdot L \cdot \mathsf{dl}_r(P, \tilde{P}),$$

where the exponents r and q are Hölder conjugates, $\frac{1}{r} + \frac{1}{q} = 1$, and $P,\ \tilde{P} \in \mathcal{P}$ (cf. Remark 14).

Remark 48 The assumption on Lipschitz continuity of the function Q notably insures the Convention 38 as Q is particularly usc., cf. (33).

Proof of Theorem 47 To compare with the second problem $v_{\tilde{P}}$ define the new policy

$$\tilde{z}_t(y_{1:t}) := \mathbb{E}_\pi\big(z_t(x) \mid \mathsf{pr}_t(x, y) = y_{1:t}\big),$$

where $\mathsf{pr}_t(x_{1:t}, y_{1:t}) := y_{1:t}$ is the projection onto the second marginal and consider the specific random variables

$$Y_t(x_{1:t}) := \mathcal{R}_{\mathcal{S}_{t+1:T}; P(\cdot|x_{1:t})}\Big(Q\big(z_{0:t}, z_{t+1:T}(x_{1:t}, \cdot); x_{1:t}, \cdot\big)\Big) \tag{46}$$

and

$$\tilde{Y}_t(y_{1:t}) := \mathcal{R}_{\mathcal{S}_{t+1:T}; \tilde{P}(\cdot|y_{1:t})}\Big(Q\big(z_{0:t}, \tilde{z}_{t+1:T}(y_{1:t}, \cdot); y_{1:t}, \cdot\big)\Big), \tag{47}$$

where $z_{0:t} \in \mathcal{Z}$ is fixed and '$\cdot$' indicates the random component.

For $\varepsilon > 0$ pick a policy $z = \big(z_{0:t}(x_{1:t})\big)_{t=1}^T$ so that

$$v_P > \mathcal{R}_{\mathcal{S}_{1:T}; P}\big(Q(z_{0:T}(\cdot); \cdot)\big) - \varepsilon. \tag{48}$$

Further, let the measure $\pi(\cdot, \cdot)$ have conditional marginals $P(\cdot)$ and $\tilde{P}(\cdot)$ with respect to the nested distance, cf. (24).

In line with Definition 15 we set $c_T := d$ and proceed by backwards induction from $t = T$ down to $t = 0$.

Base case: Note that $Y_T(x_{1:T}) = Q(z_{0:T}; x_{1:T})$ and $\tilde{Y}_T(y_{1:T}) = Q(z_{0:T}; y_{1:T})$. By Lipschitz continuity (45) it holds that $\tilde{Y}_T(y_{1:T}) - Y_T(x_{1:T}) \leq L \cdot d(x_{1:T}, y_{1:T})$. This is the statement

$$\tilde{Y}_t(y_{1:t}) - Y_t(x_{1:t}) \leq L \cdot \sup_{\sigma \in \mathcal{S}_{t+1:T}} \|\sigma_{t+1}\|_q \cdot \ldots \|\sigma_T\|_q \cdot c_t(x_{1:t}, y_{1:t}) \tag{49}$$

for the case $t = T$ (and by setting the empty product to $\prod_{t \in \emptyset} f_t := 1$).

Inductive step: In what follows we shall employ the statement (49) as induction hypothesis and deduce the statement for $t-1$ instead of t. From Jensen's inequality we infer that

$$Q\left(\tilde{z}(y);\, y\right) = Q\left(\mathbb{E}_\pi\left(z(x) \mid \mathsf{pr}_t(x,\, y) = y\right);\, y\right) \leq \mathbb{E}_\pi\left(Q\big(z(x);\, y\big) \mid \mathsf{pr}(x,\, y) = y\right). \tag{50}$$

To be more specific we emphasize that z is a vector of functions, $z = (z_t)_{t=0}^T$ and further, each z_t is a function of the variables $x_1, \ldots, x_t$, $z_t = z_t(x_{1:t})$. Jensen's inequality applies to each function z_t and each argument x_t separately, so that the inequality (50) is actually the result of applying Jensen's inequality t times repeatedly at each stage t.

Now let ζ be chosen so that $\mathbb{E}\,\tilde{Y}_t\zeta > \mathcal{R}_{\mathcal{S}_{t:T};\tilde{P}(\cdot|y_{1:t-1})}(\tilde{Y}_t) - \varepsilon'$ and $\mathsf{AV@R}_\alpha(\zeta) \leq \frac{1}{1-\alpha}\int_\alpha^1 \sigma(u)\mathrm{d}u$ $(\alpha \in (0,1))$ for some $\sigma(\cdot) \in \mathcal{S}_t$. As the risk functional is recursive we deduce from (46) and (47) that

$$\begin{aligned}
\tilde{Y}_{t-1} - Y_{t-1} - \varepsilon' &= \mathcal{R}_{\mathcal{S}_{t:T};\tilde{P}(\cdot|y_{1:t-1})}(\tilde{Y}_t) - \varepsilon' - \mathcal{R}_{\mathcal{S}_{t:T};P(\cdot|x_{1:t-1})}(Y_t) \\
&\leq \mathbb{E}_\pi\, Q\big(\tilde{z}(y);\, y\big)\zeta(y) - \mathbb{E}_\pi\, Q\big(z(x);\, x\big)\zeta(y) \\
&\leq \mathbb{E}_\pi\, \mathbb{E}_\pi\left(Q\big(z(x);\, y\big) \mid \mathsf{pr}(x,\, y) = y\right)\zeta(y) - \mathbb{E}_\pi\, Q\left(z(x);\, x\right)\zeta(y),
\end{aligned}$$

where we have used (50). By the tower property of the conditional expectation, Lipschitz continuity (45) and Hölder's inequality it follows further that

$$\begin{aligned}
\tilde{Y}_{t-1} - Y_{t-1} - \varepsilon' &\leq \mathbb{E}_\pi\, Q\big(z(x);\, y\big)\zeta(y) - \mathbb{E}_\pi\, Q\left(z(x);\, x\right)\zeta(y) \\
&\leq \mathbb{E}_\pi\, \zeta(y) c_t(x_{1:t},\, y_{1:t}) \\
&\leq L \sup_{\sigma \in \mathcal{S}_{t:T}} \|\sigma_t\|_q \cdot \ldots \|\sigma_T\|_q\, w_r\left(P(\cdot \mid x_{1:t-1}),\, \tilde{P}(\cdot \mid y_{1:t-1});\, c_t\right) \\
&= L \sup_{\sigma \in \mathcal{S}_{t:T}} \|\sigma_t\|_q \cdot \ldots \|\sigma_T\|_q \cdot c_t(x_{1:t-1},\, y_{1:t-1}),
\end{aligned}$$

as π has conditional marginals $\tilde{P}(\cdot \mid y_{1:t-1})$ and $P(\cdot \mid x_{1:t-1})$. By letting $\varepsilon' \to 0$ we get the assertion (49) for $t-1$. By repeatedly applying the previous reasoning we thus get that

$$\tilde{Y}_0 - Y_0 \leq L \sup_{\sigma \in \mathcal{S}_{1:T}} \|\sigma_1\|_q \cdot \ldots \|\sigma_T\|_q \cdot \mathsf{dl}_r(P,\, \tilde{P}). \tag{51}$$

Now note that $v_P > Y_0 - \varepsilon$ by (48) and we thus have found a policy $\tilde{z}$ so that $v_{\tilde{P}} \leq \tilde{Y}_0$. It follows with (51) that

$$v_{\tilde{P}} - v_P \leq \tilde{Y}_0 - (Y_0 - \varepsilon) \leq L \sup_{\sigma \in \mathcal{S}_{1:T}} \|\sigma_1\|_q \cdot \ldots \|\sigma_T\|_q \cdot \mathsf{dl}_r(P,\, \tilde{P}) + \varepsilon.$$

The result finally follows by letting $\varepsilon \to 0$ and by interchanging the role of P and $\tilde{P}$. □

8 Summary

This paper addresses risk-averse stochastic optimization problems. To define the risk functionals based on partial observations we introduce conditional risk measures first. They are defined on fibers and can be composed to nested risk measures. We demonstrate that these nested risk measures are continuous and we establish the modulus of continuity. As a consequence, the optimization problems are continuous as well, these problems inherit the modulus of continuity from the risk functionals.

All results come along with characterizations as generalized martingales. It is demonstrated that the underlying distance is a usual martingale with respect to the natural filtration. The value functions are shown to follow a generalized, risk-averse martingale pattern as well.

Acknowledgements We would like to thank Prof. Shapiro for proposing to elaborate the continuity relations of nested risk measures with respect to the nested distance.

References

1. Ambrosio, L., Gigli, N., Savaré, G.: Gradient Flows in Metric Spaces and in the Space of Probability Measures, 2nd edn. Birkhäuser Verlag, Basel (2005). https://doi.org/10.1007/978-3-7643-8722-8
2. Artzner, P., Delbaen, F., Heath, D.: Thinking coherently. Risk **10**, 68–71 (1997)
3. Brenier, Y.: Décomposition polaire et réarrangement monotone des champs de vecteurs. Comptes Rendus l'Acad. Sci. Paris Sér. I Math. **305**(19), 805–808 (1987)
4. Brenier, Y.: Polar factorization and monotone rearrangement of vector-valued functions. Commun. Pure Appl. Math. **44**(4), 375–417 (1991). https://doi.org/10.1002/cpa.3160440402
5. Dellacherie, C., Meyer, P.-A.: Probabilities and Potential. North-Holland Publishing Co., Amsterdam (1988). https://projecteuclid.org/euclid.bams/1183546371
6. Dentcheva, D., Ruszczyński, A.: Time-coherent risk measures for continuous-time Markov chains. SIAM J. Financ. Math. **9**(2), 690–715 (2018a). https://doi.org/10.1137/16m1063794
7. Dentcheva, D., Ruszczyński, A.: Risk forms: representation, disintegration, and application to partially observable two-stage systems, unpublished (2018b). https://arxiv.org/abs/1807.02300
8. Fleming, W.H., Soner, H.M.: Controlled Markov Processes and Viscosity Solutions, 2nd edn. Springer, Berlin (2006). https://doi.org/10.1007/0-387-31071-1
9. Föllmer, H., Schied, A.: *Stochastic Finance: An Introduction in Discrete Time*. de Gruyter Studies in Mathematics 27. Berlin, Boston, De Gruyter (2004). ISBN 978-3-11-046345-3. 10.1515/9783110218053. http://books.google.com/books?id=cL-bZSOrqWoC
10. Girardeau, P., Leclère, V., Philpott, A.B.: On the convergence of decomposition methods for multistage stochastic convex programs. Math. Oper. Res. **40**(1), 1–16 (2014). https://doi.org/10.1287/moor.2014.0664
11. Goulart, F.C., da Costa, B.F.P.: Nested distance for stagewise-independent processes, unpublished (2017). https://arxiv.org/pdf/1711.10633.pdf
12. Jouini E, Schachermayer W, Touzi N (2006) Law invariant risk measures have the Fatou property. In S. Kusuoka, A. Yamazaki(ed) Advances in Mathematical Economics, volume 9 of Kusuoka, Shigeo and Yamazaki, Akira chapter 4, pp 49–71. Springer, Japan. https://doi.org/10.1007/4-431-34342-3
13. Kallenberg, O.: Foundations of Modern Probability. Springer, New York (2002). https://doi.org/10.1007/b98838
14. Karatzas, I., Shreve, S.E.: Methods of Mathematical Finance. Stochastic Modelling and Applied Probability. Springer, Berlin (1998). https://doi.org/10.1007/b98840
15. Kusuoka, S.: On law invariant coherent risk measures. In: Kusuoka, S., Maruyama, T. (eds.) Advances in Mathematical Economics. Springer, Tokyo (2001). https://doi.org/10.1007/978-4-431-67891-5
16. Maggioni, F., Allevi, E., Bertocchi, M.: Measures of information in multistage stochastic programming. STOPROG (2012). https://doi.org/10.5200/stoprog.2012.14

17. McCann, R.J.: Polar factorization of maps on Riemannian manifolds. Geom. Funct. Anal. **11**(3), 589–608 (2001). https://doi.org/10.1007/PL00001679
18. Peng, S.: Nonlinear expectations, nonlinear evaluations and risk measures. In: *Lecture Notes in Mathematics*, pp. 165–253. Springer, Berlin Heidelberg, (2004). https://doi.org/10.1007/b100122
19. Pflug, GCh.: Version-independence and nested distributions in multistage stochastic optimization. SIAM J. Optim. **20**, 1406–1420 (2009). https://doi.org/10.1137/080718401
20. Pflug, G. Ch., Pichler, A.: *Multistage Stochastic Optimization*. Springer Series in Operations Research and Financial Engineering. Springer, Berlin (2014). ISBN 978-3-319-08842-6. https://doi.org/10.1007/978-3-319-08843-3. https://books.google.com/books?id=q_VWBQAAQBAJ
21. Pflug, GCh., Römisch, W.: Modeling, Measuring and Managing Risk. World Scientific, NJ (2007). https://doi.org/10.1142/9789812708724
22. Philpott, A.B., de Matos, V.L.: Dynamic sampling algorithms for multi-stage stochastic programs with risk aversion. Eur. J. Oper. Res. **218**(2), 470–483 (2012). https://doi.org/10.1016/j.ejor.2011.10.056
23. Philpott, A.B., de Matos, V.L., Finardi, E.: On solving multistage stochastic programs with coherent risk measures. Oper. Res. **61**(4), 957–970 (2013). https://doi.org/10.1287/opre.2013.1175
24. Pichler, A.: The natural Banach space for version independent risk measures. Insur. Math. Econ. **53**(2), 405–415 (2003). https://doi.org/10.1016/j.insmatheco.2013.07.005
25. Pichler, A., Shapiro, A.: Minimal representations of insurance prices. Insur. Math. Econ. **62**, 184–193 (2015). https://doi.org/10.1016/j.insmatheco.2015.03.011
26. Pichler, A., Shapiro, A.: Risk averse stochastic programming: time consistency and optimal stopping (2018). arXiv:1808.10807
27. Riedel, F.: Dynamic coherent risk measures. Stoch. Process. Appl. **112**(2), 185–200 (2004). https://doi.org/10.1016/j.spa.2004.03.004
28. Rockafellar, R.T., Wets, R.J.-B.: Nonanticipativity and L^1-martingales in stochastic optimization problems. Math. Program. Study **6**, 170–187 (1976)
29. Römisch, W., Guigues, V.: Sampling-based decomposition methods for multistage stochastic programs based on extended polyhedral risk measures. SIAM J. Optim. **22**(2), 286–312 (2012). https://doi.org/10.1137/100811696
30. Ruszczyński, A.: Risk-averse dynamic programming for Markov decision processes. Math. Program. Ser. B **125**, 235–261 (2010). https://doi.org/10.1007/s10107-010-0393-3
31. Ruszczyński, A., Shapiro, A.: Conditional risk mappings. Math. Oper. Res. **31**(3), 544–561 (2006). https://doi.org/10.1287/moor.1060.0204
32. Shapiro, A.: On Kusuoka representation of law invariant risk measures. Math. Oper. Res. **38**(1), 142–152 (2013). https://doi.org/10.1287/moor.1120.0563
33. Shapiro, A.: Rectangular sets of probability measures. Oper. Res. **64**(2), 528–541 (2016). https://doi.org/10.1287/opre.2015.1466
34. Shapiro, A.: Interchangeability principle and dynamic equations in risk averse stochastic programming. Oper. Res. Lett. **45**(4), 377–381 (2017). https://doi.org/10.1016/j.orl.2017.05.008
35. Shapiro, A., Dentcheva, D., Ruszczyński, A.: In: *Lectures on Stochastic Programming*. MOS-SIAM Series on Optimization. SIAM, second edition (2014). https://doi.org/10.1137/1.9780898718751
36. Shiryaev, A.N.: Probability. Springer, New York (1996). https://doi.org/10.1007/978-1-4757-2539-1
37. Villani, C.: *Topics in Optimal Transportation*, vol. 58 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence (2003). ISBN 0-821-83312-X. https://doi.org/10.1090/gsm/058. http://books.google.com/books?id=GqRXYFxe0l0C
38. Xin, L., Shapiro, A.: Bounds for nested law invariant coherent risk measures. Oper. Res. Lett. **40**, 431–435 (2012). https://doi.org/10.1016/j.orl.2012.09.002
39. Zhang, J.: Backward Stochastic Differential Equations. Springer, New York (2017). https://doi.org/10.1007/978-1-4939-7256-2