

Inverse Optimization of Convex Risk Functions

Jonathan Yu-Meng Li

Telfer School of Management, University of Ottawa, Ottawa, Ontario K1N6N5, Canada, Jonathan.Li@telfer.uOttawa.ca

The theory of convex risk functions has now been well established as the basis for identifying the families of risk functions that should be used in risk-averse optimization problems. Despite its theoretical appeal, the implementation of a convex risk function remains difficult, as there is little guidance regarding how a convex risk function should be chosen so that it also well represents a decision maker’s subjective risk preference. In this paper, we address this issue through the lens of inverse optimization. Specifically, given solution data from some (forward) risk-averse optimization problem (i.e., a risk minimization problem with known constraints), we develop an inverse optimization framework that generates a risk function that renders the solutions optimal for the forward problem. The framework incorporates the well-known properties of convex risk functions—namely, monotonicity, convexity, translation invariance, and law invariance—as the general information about candidate risk functions, as well as feedback from individuals—which include an initial estimate of the risk function and pairwise comparisons among random losses—as the more specific information. Our framework is particularly novel in that unlike classical inverse optimization, it does not require making any parametric assumption about the risk function (i.e., it is non-parametric). We show how the resulting inverse optimization problems can be reformulated as convex programs and are polynomially solvable if the corresponding forward problems are polynomially solvable. We illustrate the imputed risk functions in a portfolio selection problem and demonstrate their practical value using real-life data.

1. Introduction

The theory of convex risk functions, established since the work of Artzner et al. (1999) and later generalized by Föllmer and Schied (2002), Ruszczyński and Shapiro (2006), and others, has played a central role in the development of modern risk-averse optimization models. The work of Ruszczyński and Shapiro (2006) in particular brings to light the intimate relationship between convex risk functions and optimization theory and provides necessary tools for analyzing the tractability of risk-averse optimization problems involving convex risk functions. The unified scheme that Ruszczyński and Shapiro (2006) provided through convex analysis also explains the success of several convex risk functions that have now been widely applied for risk minimization, among which the most well known is arguably Conditional Value-at-Risk (CVaR) (Rockafellar and Uryasev (2000)).

It was however not for the purpose of optimization (i.e., risk minimization), at least not solely, that the theory was first established. Rather, the motivation lay in the need for alternative measures of risk that could better characterize how individuals perceive risk. For example, the property of convexity, which led to the term “convex” risk function, was postulated by the theory as an essential and universal characteristic of how risk-averse individuals would perceive risk, namely that diversification should not increase risk. The industry-standard measure of risk, Value-at-Risk (VaR), unfortunately does not satisfy the property of convexity, whereas CVaR, as its counterpart, which does satisfy convexity, has become a popular theory-supported alternative. Other properties of the theory that have also been widely referenced in justifying the choice of a measure for risk include monotonicity and translation invariance (Föllmer and Schied (2002)), law invariance (Kusuoka (2001)), positive homogeneity (Artzner et al. (1999)), and comonotonicity (Acerbi (2002)), among others. Each of these properties represents a certain well-grounded rationale for how risk might be perceived over random variables. Some are applicable fairly generally (e.g., monotonicity and law invariance), whereas some others can be domain dependent (e.g., positive homogeneity and comonotonicity).

However, despite the general attractive features of convex risk functions from the point of view of both optimization and risk modelling, very little guidance has been provided to date regarding how to choose a convex risk function that can also well represent a decision maker’s subjective perception of risk. In current practice, the choice of a convex risk function is mostly ad hoc and involves very little knowledge of decision makers’ true risk preferences. This raises the question of how ones’ risk preferences may be observed and how to generate a convex risk function that complies with the observed preferences. Delage and Li (2018) appear to be the first to address this question, by proposing a means to construct a convex risk function from the assessments provided by the decision maker, who compares pairs of risky random losses. Their work is closely related to the scheme of preference (or utility) elicitation (see, e.g., Clemen and Reilly (2014)), where queries are considered for extracting users’ preferences in establishing their utility functions. One of the main challenges facing this line of inquiry is that in reality decision makers may only be able to provide limited responses because of potential time and cognitive constraints, and thus the elicited preference information is often incomplete. This situation is formulated in Delage and Li (2018) as a preference robust optimization problem where a worst-case risk measure is sought that complies with a finite number of pairwise preference relations elicited from the decision maker. Similar ideas can be found also in the context of expected utility theory. Armbruster and Delage (2015) and Hu and Mehrotra (2015) consider the formulation of a worst-case expected utility function based on limited preference information, whereas Boutilier et al. (2006) considers a worst-case regret criterion over utility functions.

In this paper, we attempt to provide an alternative perspective on the search of a convex risk function that takes into account decision makers’ true risk preferences, namely through the lens of inverse optimization. The motivation is that in many current applications, it becomes possible to have access to the record of the decisions made by individuals, and the past decisions, if optimal, provide useful preference information. Such kinds of preference information may be viewed as a special form of pairwise preference relations, where the random variable chosen according to a made decision is considered preferable to the random variables that could be chosen by alternative decisions. In the case that the alternative decisions are finite, the pairwise preference relations are also finite, which can then be handled by existing frameworks such as [Delage and Li \(2018\)](#). However, this work emphasizes the case where the alternative decisions may be described through a convex set, which leads to infinitely many pairwise relations that existing frameworks cannot handle. Moreover, we also recognize that, in practice, even though individuals may perceive risk differently, they often start by agreeing upon some seemingly reasonable risk measure. They then adjust their measurement of risk after receiving more precise preference information. One such example is that many investors tend to follow the principle of *safety first* ([Roy \(1952\)](#)), which states that the top concern of an investor is to avoid a possible catastrophic event. They would thus naturally start by choosing a downside risk measure that they feel safe enough to apply (e.g., [Fabozzi \(2015\)](#)). For example, this could be the CVaR risk measure that has now been widely applied in various areas. Although staying aligned with the downside risk measure is desirable, decisions made by individuals are often inconsistent with (e.g., more aggressive than) what the risk measure prescribes. In this paper, we refer to such a risk measure as a reference risk function, which should be followed closely before more precise preference information can be revealed. A natural framework to address the above issues is the setup of inverse optimization; namely, given the solutions for some forward problem (i.e., a decision optimization problem with known constraints), the inverse problem seeks a risk function that renders the solutions optimal for the forward problem by minimally deviating from the reference risk function. Our formulation of the inverse problem will allow for incorporating preference information in both the forms of pairwise relations and “most preferable” decisions in convex sets of alternatives, and also the important properties of convex risk functions, namely the monotonicity, convexity, translation invariance, and law invariance. We show how the resulting inverse optimization problems can be tractably analyzed by applying conjugate duality theory ([Rockafellar \(1974\)](#)).

To the best of our knowledge, little has been discussed in the literature about inverse optimization for convex risk functions. [Bertsimas et al. \(2012\)](#) considered inverse optimization for a financial application involving the use of coherent risk measures, but they assumed that the measure is given a priori and focused instead on the estimation of parameters characterizing random returns and

risk budgets. [Iyengar and Kang \(2005\)](#) also applied inverse optimization to estimate parameters of expected returns in a financial problem. More generally, inverse optimization methods have been developed for linear programs ([Ahuja and Orlin \(2001\)](#), [Dempe and Lohse \(2006\)](#)), conic programs ([Iyengar and Kang \(2005\)](#)), and convex separable programs ([Zhang and Xu \(2010\)](#)) for estimating the parameters that characterize the programs. Early works also include [Burton and Toint \(1992\)](#), [Zhang and Liu \(1996\)](#), and [Hochbaum \(2003\)](#), who focused on network and combinatorial optimization problems (see [Heuberger \(2004\)](#) for a survey), whereas more recent works include [Schaefer \(2009\)](#) on integer programs, [Chan et al. \(2014\)](#) on multi-objective programs, [Ghate \(2015\)](#) on countably infinite linear programs, [Chan et al. \(2018\)](#) on the issue of sub-optimality of an observed solution, and [Keshavarz et al. \(2011\)](#), [Bertsimas et al. \(2014\)](#), [Aswani et al. \(2015\)](#), and [Mohajerin Esfahani et al. \(2015\)](#) on various issues related to the observations of multiple responses from an agent solving a parametric optimization problem.

In much of the literature referenced above, the problems are structured in a parametric fashion, and the goal is to estimate the parameters that characterize the forward problems from observed decisions. However, the parametric assumption is too limiting for the purpose of identifying a decision maker’s true risk function because it restricts the class of functions to which the true risk function may belong. It also provides no guarantee regarding the convergence to the true risk function even if some elicited information, such as pairwise preference relations, is available. In contrast, the inverse optimization formulations presented in this paper are parameter-free and search over the entire space of convex risk functions for the true risk function. With the collection of more elicited information, their solutions can converge to the true risk function, if it is a convex risk function. In this sense, our work broadens the scope of inverse optimization and opens the door for nonparametric approaches to function estimation through inverse optimization. However, we should note that [Bertsimas et al. \(2014\)](#) provide a kernel method for inverse optimization, which can also be considered non-parametric. While their method focuses on estimating a function characterizing all the subgradients associated with an unknown function, the method developed in this paper addresses directly the estimation of the unknown function. More detailed discussions along this line are provided later in [Section 3](#). We should also mention that the inverse problem considered in this paper generally falls into the class of inverse problems that focus on estimating the objective function of an optimization problem. Although this class of inverse problems is known to be tractable in a parametric setting, the class of inverse problems that seek to impute parameters defining the feasible region of an optimization problem is generally much less tractable (see, e.g., [Birge et al. \(2017\)](#)). This paper’s focus on the former may help explain why tractably solving the inverse problem, even in a non-parametric setting, is a reasonable hope.

Our formulation of the inverse problem does require, albeit implicitly, an assumption that may affect the scope of application of our inverse models, and we should point it out here. Namely, our formulation assumes that in the case where decisions are based on some probabilistic views (or beliefs) of the decision makers, these views (i.e., the probabilistic assessments of random outcomes), either stay unchanged over time or, if not, can always be disclosed together with the decisions made. This may not always be possible because some decision makers may only have a vague sense of probability driven by their intuition and find it cognitively too demanding to articulate how their views change over time. As extensively discussed in this paper, while it is possible to address the case where decision makers take no view or a constant (personal) view on the likelihood of outcomes, namely by applying non-law-invariant risk functions, it remains an open question as to how a (law-invariant) risk function can possibly be learned from past decisions when the decision makers did hold different probabilistic views over time but were not able to disclose them. This question is important because the scenario, as described, could happen in practice; and from a statistical point of view, the use of law-invariant risk function can be necessary. As a preliminary step, we have conducted some experiments, which we describe in this paper, to examine first how the performances of the solutions optimized based on imputed risk functions may be affected by the misspecification of probability distributions in our inverse models (i.e., distributions applied in the models are inconsistent with the actual views of the decision maker). We observe that, despite the misspecification, the performances can still be noticeably improved towards the optimal performances as more decision data are incorporated into the inverse models. However, it still requires the full development of a formal and rigorous theory to satisfactorily answer the question, and we leave this for future study.

One natural application of our inverse optimization framework is to identify a risk function that captures the risk preference of an investor from his/her past investment decisions (see, e.g., [Delage and Li \(2018\)](#)). While this is the application that we primarily focus on in this paper, we should point out that our framework can also be naturally applied to other settings that involve budget allocation decisions under uncertainty. For instance, [Haskell et al. \(2018\)](#) consider a setting of homeland security and the problem of learning the risk preference of the Department of Homeland Security (DHS), which makes decisions to allocate budget across a number of cities so as to protect them from potential terrorist attacks. While covering these other applications in depth may go beyond the scope of this paper, the framework established in this paper shall provide the basis for further exploring these other applications.

We briefly summarize our main contributions below:

1. We develop for the first time an inverse-optimization framework for convex risk functions that generates a risk function incorporating the following information: 1) the properties of monotonicity, convexity, translation invariance, and law invariance promoted in the theory of convex risk functions, 2) observable optimal solutions from forward problems, 3) a reference risk function, and 4) elicited pairwise preference relations.

2. We formulate the inverse optimization problem in a non-parametric fashion and show that for a large number of cases, the computational tractability of the inverse problem is largely determined by the forward problem: namely that the former is polynomially solvable if the latter is so.

3. Methodologically speaking, we show that to solve the inverse problem it suffices to search over a particular class of risk functions for an optimal solution. This class of risk functions is representable in terms of the random variables resulting from the observed decisions and a fixed set of parameters. Based on this representation, the inverse problem reduces to determining the values of the parameters in the representation, and these values correspond to the amounts of risk estimated for the chosen random variables. The number of parameters needed in the representation is set by the number of observed decisions. This offers an intuitive and computationally tractable interpretation of the (non-parametric) inverse problem.

4. We demonstrate the application of our framework in a portfolio selection problem and provide computational evidence that the imputed risk functions utilize well the preference information contained in observable solutions and a reference risk function. This leads to a solution that can be well justified in terms of both its performance evaluated based on the true risk function and the reference risk function. We also demonstrate how quickly the performances of the solutions optimized based on the imputed risk functions can converge to the performances of the solutions optimized based on the true risk function, as the number of observed decisions increases.

2. Forward and Inverse Problem of Risk Minimization

We begin by formalizing the forward problem of risk minimization and characterizing the problem using the theory of convex risk functions. We then proceed to the formulation of the inverse problem.

2.1. Forward problem of risk minimization

Our general setup of the forward problem follows closely the setup in the literature of choice over acts (i.e., random variables $Z : \Omega \rightarrow \mathbb{R}$) ([Savage \(1954\)](#)). In this setup, it is assumed that a decision maker's preference over random variables can be specified and that the forward problem seeks the most preferable random variables. In the special case where a probability measure \mathbb{P} can be identified over the sigma-algebra Σ of sample space Ω , the preference can be alternatively defined

over the distributions, denoted by F_Z , of the random variables. This special case is considered in the literature of choice over lotteries (i.e., distributions) (Von Neumann and Morgenstern (1944)).

We will first continue formalizing the forward problem under the general setting, and the special case will follow naturally as we proceed. Without loss of generality, we assume that any random variable Z represents some form of loss, by which we mean that it has the interpretation that for any $\omega \in \Omega$, the larger the value of $Z(\omega)$ is, the worse it is. We denote by \succeq a system of preference relations that is complete, transitive, and continuous¹, where $Z_1 \succeq Z_2$ denotes that Z_1 is preferred to Z_2 . A risk function ρ is a numerical representation that captures the preference relation \succeq in terms of the riskiness of random losses (i.e., a random loss Z is preferable if it is perceived less risky). In this paper, we will focus on the case where a risk function ρ is defined over random losses based on a sample space with finitely many outcomes $\Omega := \{\omega_i\}_{i=1}^M$. In this setting, any random loss Z can be represented also by a vector $\vec{Z} \in \mathbb{R}^{|\Omega|}$, where $(\vec{Z})_i = Z(\omega_i)$, and the random loss resulting from a decision $x \in \mathbb{R}^n$ can be written by $\vec{Z}(x) := (Z(x, \omega_1), \dots, Z(x, \omega_M))^T \in \mathbb{R}^{|\Omega|}$. If a random loss Z_2 is perceived at least as risky as Z_1 (i.e., $Z_1 \succeq Z_2$), the risk function $\rho: \mathbb{R}^{|\Omega|} \rightarrow \mathbb{R}$ should satisfy $\rho(\vec{Z}_1) \leq \rho(\vec{Z}_2)$. Accordingly, a solution x^* is optimal if and only if it satisfies $\rho(\vec{Z}(x^*)) \leq \rho(\vec{Z}(x))$, $\forall x \in \mathcal{X}$, and a risk minimization problem can be formulated as

$$\min_{x \in \mathcal{X}} \rho(\vec{Z}(x)). \quad (1)$$

Throughout this paper, we assume that the function $\vec{Z}(x)$ is convex in x (i.e., $Z(x, \omega_i)$ is convex in x for all $\omega_i \in \Omega$), and the feasible set $\mathcal{X} \subseteq \mathbb{R}^n$ is a convex set.

It is hypothesized in the theory of convex risk functions (Föllmer and Schied (2002)) that any risk function ρ that represents a reasonable or “rational” preference system would satisfy certain axioms. The most widely known ones are the following three:

1. (Monotonicity) $\rho(\vec{Z}_1) \leq \rho(\vec{Z}_2)$ for any $Z_1(\omega) \leq Z_2(\omega)$, $\forall \omega \in \Omega$,
2. (Convexity) $\rho(\lambda \vec{Z}_1 + (1 - \lambda) \vec{Z}_2) \leq \lambda \rho(\vec{Z}_1) + (1 - \lambda) \rho(\vec{Z}_2)$, where $0 \leq \lambda \leq 1$, and
3. (Translation Invariance) $\rho(\vec{Z} + c) = \rho(\vec{Z}) + c$.

The first axiom, monotonicity, captures the fact that any reasonable preference system would never prefer a random loss that is known to have higher loss for any possible outcome (i.e., any Z_1, Z_2 such that $Z_1(\omega) \leq Z_2(\omega)$, $\forall \omega \in \Omega$ must lead to $Z_1 \succeq Z_2$). The axiom of convexity describes the diversification preference, namely that any convex combination (diversification) $\lambda Z_1 + (1 - \lambda) Z_2$ must be (at least equally) preferable to non-diversified counterparts (i.e., Z_1 or Z_2). Lastly, translation invariance is necessary when a monetary interpretation of risk is required, which is the case in finance where a deterministic amount such as cash can always be used to offset the risk

¹ These are the conditions required to ensure that there exists a function ρ that captures the system of preference relations (Debreu (1954)).

by the same amount. The corresponding preference system is insensitive to any constant amount c added to or subtracted from all random losses (i.e., $Z_1 \succeq Z_2 \Rightarrow Z_1 + c \succeq Z_2 + c$). Based on these three axioms, we define the following class of functions that capture risk-averse preferences.

DEFINITION 1. (*Risk-averse functions*) Let \mathcal{R} denote the set of functions that satisfy axioms (1)–(3) and $\rho(\vec{0}) = 0$.

The condition $\rho(\vec{0}) = 0$ is imposed for the purpose of normalization so that risk estimated based on different risk functions is comparable. We now also formalize the special case where a probability measure \mathbb{P} can be defined over (Ω, Σ) . This naturally leads to the consideration of the following axiom for a risk function (see, e.g., Kusuoka (2001)).

4. (Law Invariance) $\rho(\vec{Z}_1) = \rho(\vec{Z}_2)$, for any $Z_1 \sim_{\mathbb{P}} Z_2$ (distributionally equivalent).

The above axiom immediately implies that the risk function in this case is essentially a function of distributions. We define the following class of functions when the distributions for all random variables are available.

DEFINITION 2. (*Law-invariant risk-averse functions*) Let $\mathcal{R}_F \subset \mathcal{R}$ denote the set of risk-averse functions that are law invariant. Without loss of generality, we can equivalently write $\rho(\vec{Z})$ as $\rho(F_Z)$ (i.e., a function of distributions F_Z).

Depending on the decision maker’s knowledge about the distribution, one may decide which class of risk functions (i.e., risk-averse versus law-invariant risk-averse) is more appropriate to assume in defining the forward problem. In particular, we provide the following examples, which cover three possible cases: the case with no distribution, the case with ambiguous distributions, and the case with a specific distribution. The first two cases can be treated as special cases of our general setup (i.e., preference over acts), whereas the third case invokes the property of law invariance.

EXAMPLE 1. (General case) If a decision maker’s choices are made by knowing only that there are $|\Omega|$ possible outcomes for the uncertain losses, we may assume that the individual solves a forward problem based on a certain risk-averse function $\rho \in \mathcal{R}$.

EXAMPLE 2. (Distributional ambiguity) If a decision maker actually has in mind a certain distribution-based convex risk measure $\hat{\rho}(\cdot; q)$, where q denotes the distribution, for evaluating risk but is concerned about the estimation error associated with an empirical distribution \hat{p} , we may assume that she is ambiguity-averse and that her choices follow a distributionally robust version of the risk measure. Namely, it can take the following form with uncertain probability q :

$$\rho^\uparrow(Z) = \sup_q \left\{ \hat{\rho}(Z; q) \mid D(q, \hat{p}) \leq d, \bar{1}^\top q = 1, q \geq 0 \right\},$$

where D measures the difference between two distributions and is usually defined based on some ϕ -divergence function (see Ben-Tal et al. (2013) for more details). It is not hard to confirm that in

this case $\rho^\dagger \in \mathcal{R}$. Of course, in reality we do not know the structure of $\hat{\rho}$ and D and may assume only that the true risk function is a risk-averse function $\rho \in \mathcal{R}$.

EXAMPLE 3. (Known Distribution) Suppose that a decision maker can identify which distribution q to use in $\hat{\rho}(\cdot; q)$ and that the information about the distribution is available. In this case, we may assume that the individual solves a certain law-invariant risk-averse function $\rho \in \mathcal{R}_F$ based on the distribution q .

REMARK 1. One scenario that can also be of practical interest but is not covered above is the case where a decision maker solves some distribution-based risk measure $\hat{\rho}(\cdot; q)$ but the individual does not reveal which distribution q is used. While it remains possible to address this case if q stays unchanged over time, namely by assuming that the true risk function is a risk-averse function $\rho \in \mathcal{R}$ because $\hat{\rho}(\cdot; q) \in \mathcal{R}$ for a fixed q , it becomes less clear how to address the case when the distribution q may differ from one time point to another. In the latter, the observed decisions can actually appear inconsistent with the assumption $\rho \in \mathcal{R}$. This can be easily seen by considering for instance the case where the decision maker's true risk function is the simple expected value function. We may observe between two random variables X_i and X_j that at one point $\mathbb{E}_{q_1}[X_i] > \mathbb{E}_{q_1}[X_j]$ and at the other point $\mathbb{E}_{q_2}[X_i] < \mathbb{E}_{q_2}[X_j]$ based on two different distributions q_1 and q_2 . Clearly, there exists no $\rho \in \mathcal{R}$ that can capture such preferences. For this reason, we find it necessary to assume throughout this paper that in the case where the decision makers applied different distributions, these distributions can always be disclosed together with the decisions made. Later in the numerical section, Section 5.2, we will revisit this assumption and address the case where the disclosed distributions may not be fully accurate.

2.2. Inverse problem of risk minimization

In the inverse problem, the risk function ρ is unknown, but one has access to decisions made according to forward problems as defined in the previous section. The goal is to generate a risk function ρ that renders the observed decisions as optimal as possible in the forward problem. Specifically, let $(x^t, \vec{Z}^t(\cdot), \mathcal{X}^t)$ represent each observation, which denotes that x^t was made with respect to the random vector $\vec{Z}^t(\cdot)$ and the feasible region \mathcal{X}^t . We can write down the following optimality condition that characterizes the risk function ρ through the observed decisions.

Optimality Condition: Given a list of observations $\{(x^t, \vec{Z}^t(\cdot), \mathcal{X}^t)\}_{t \in \mathcal{T}}$, where $|\mathcal{T}| < \infty$, the set of risk functions that render the decisions optimal admits

$$\mathcal{R}_{inv} := \left\{ \rho \mid \rho(\vec{Z}^t(x^t)) \leq \rho(\vec{Z}^t(x)), \forall x \in \mathcal{X}^t, t \in \mathcal{T} \right\}.$$

Note that the risk function ρ does not depend on t (i.e., the decision maker's risk preference is assumed to stay constant when the past decisions were made). In the case where $|\mathcal{T}|$ is small, the

above set may not be sufficient to build a meaningful inverse problem, since it can contain some degenerate form of functions. It would thus be necessary to assume that some “prior” knowledge about the risk function can be acquired. The most direct way to acquire such knowledge is through preference elicitation (e.g., [Clemen and Reilly \(2014\)](#)) where the decision maker would be asked to make comparisons among a selective list of random variables. We can also borrow the concept of reference solution from the literature of inverse optimization, which stands for a solution that can be used as a reference while searching for an alternative better solution.

Elicited Preference Relations: Given a list of pairs of random losses $\{(L_k, U_k)\}_{k \in \mathcal{K}}$, where $|\mathcal{K}| < \infty$, that satisfy $L_k \succeq U_k$ for $k \in \mathcal{K}$, we define the set

$$\mathcal{R}_{el}(\{(L_k, U_k)\}_{k \in \mathcal{K}}) := \left\{ \rho \mid \rho(\vec{L}_k) \leq \rho(\vec{U}_k), \forall k \in \mathcal{K} \right\}.$$

Reference Risk Function: Given a reference risk function $\tilde{\rho} \in \mathcal{R}$ and a parameter $\epsilon \geq 0$ that describes the maximum discrepancy between the candidate risk function and the reference risk function, the following set of risk functions can be defined accordingly:

$$\mathcal{R}_{ref}(\epsilon) := \{ \rho \mid \|\rho - \tilde{\rho}\|_{\infty} \leq \epsilon \},$$

where $\|\cdot\|_{\infty}$ stands for the infinity norm defined over functions $\rho: \mathbb{R}^{|\Omega|} \rightarrow \mathbb{R}$.

REMARK 2. The constraints defining \mathcal{R}_{inv} can also be viewed as a special case of the constraints inferred from preference elicitation. While the random losses $\vec{Z}^t(x)$ considered in \mathcal{R}_{inv} are “forward-problem” dependent, in preference elicitation any pair of random losses may be considered for comparison. We take such a unified perspective in formulating the inverse problem.

REMARK 3. It is worth mentioning here that it is also possible to consider norms other than the infinity norm in the definition of \mathcal{R}_{ref} . For instance, one might consider L^2 norm $\|\rho\|_{\phi} := (\int \rho(z)^2 \phi(dz))^{1/2}$ for some probability measure ϕ . However, it may not be clear whether this is practically useful because one may find it hard to specify the measure ϕ and to interpret the norm. In the case where ϕ is discrete, which requires only comparing ρ and $\tilde{\rho}$ over finite points, the analysis presented in this paper can easily accommodate such a case.

With the above definitions of \mathcal{R}_{inv} , \mathcal{R}_{el} , and \mathcal{R}_{ref} , one may consider different criteria to determine which risk function described by the above sets is the optimal choice. In particular, we consider the following four criteria. The first three are primarily concerned about the fitting of risk levels, whereas the fourth one addresses the fitting of prescribed decisions. The first criterion follows most closely the spirit of classical inverse optimization.

1. Minimizing the deviation from the reference risk function $\tilde{\rho}$:

$$\begin{aligned} & \inf_{\rho, \epsilon \in \mathbb{R}} \epsilon \\ & \text{subject to } \rho \in \mathcal{R} \text{ (or } \mathcal{R}_F \text{ in the case of law invariance),} \\ & \rho \in \mathcal{R}_{ref}(\epsilon) \cap \mathcal{R}_{inv} \cap \mathcal{R}_{el}(\{(L_k, U_k)\}_{k \in \mathcal{K}}). \end{aligned} \quad (2)$$

This model assumes that in addition to capturing the preferences implied by the observed decisions, there is a practical need to stay aligned with the reference risk function whenever possible. As mentioned in the introduction, the reference risk function can for example be the CVaR risk measure used to implement the safety-first principle. In general, one can apply the model to generate an alternative risk function that not only renders the observed decisions optimal but also maximally aligns with the chosen downside risk measure.

One may argue that in some cases the observed decisions might not be “rigorously” optimal given that human beings are not perfectly rational. We can accommodate such a possibility by replacing the set \mathcal{R}_{inv} with the following set based on sub-optimality:

$$\mathcal{R}_{inv}(\gamma) = \left\{ \rho \mid \rho(\vec{Z}^t(x^t)) \leq \rho(\vec{Z}^t(x)) + \gamma_t, \forall x \in \mathcal{X}^t, t \in \mathcal{T} \right\}.$$

Thus, if observed decisions are known to be sub-optimal, the following model may be considered that seeks to close the optimality gap.

2. Minimizing the sub-optimality of observed decisions:

$$\begin{aligned} & \inf_{\rho, \gamma \in \mathbb{R}^{|\mathcal{T}|}} \sum_{t \in \mathcal{T}} \gamma_t \\ & \text{subject to } \rho \in \mathcal{R} \text{ (or } \mathcal{R}_F \text{ in the case of law invariance),} \\ & \rho \in \mathcal{R}_{inv}(\gamma) \cap \mathcal{R}_{el}(\{(L_k, U_k)\}_{k \in \mathcal{K}}) \cap \mathcal{R}_{ref}(\epsilon^*), \end{aligned} \quad (3)$$

where $\epsilon^* \in (0, \infty]$ is fixed beforehand.

Here, one’s priority is to ensure that the imputed risk function will render the observed decisions as favorable as possible. One may choose to set $\epsilon^* = \infty$ if needed. The central idea behind this model is that the observed decisions, albeit sub-optimal, still closely follow the decision maker’s true preference. However, if the accuracy of the decision data is in doubt, one may instead fix $\gamma = \gamma^*$ to some value γ^* larger than the optimal solution in (3). Moreover, if the concern is about potential risk underestimation, the following model provides a means to examine the worst possible risk.

3. For any \vec{Z} , seek a worst-case risk function ² :

$$\begin{aligned} \rho(\vec{Z}) := & \sup_{\rho'} \rho'(\vec{Z}) \\ \text{subject to } & \rho' \in \mathcal{R} \text{ (or } \mathcal{R}_F \text{ in the case of law invariance),} \\ & \rho' \in \mathcal{R}_{inv}(\gamma^*) \cap \mathcal{R}_{el}(\{(L_k, U_k)\}_{k \in \mathcal{K}}) \cap \mathcal{R}_{ref}(\epsilon^*), \end{aligned} \quad (4)$$

where $\gamma^* \in [0, \infty)^{|\mathcal{T}|}$ and $\epsilon^* \in (0, \infty]$ are fixed beforehand.

The last criterion we consider is closely related to the idea behind the second criterion, where the goal is to reconcile the decision data. The difference is that not only do we like to ensure the optimal solution generated from the imputed risk function is aligned with the decision maker's risk preference, but we also want the solution itself to be close to the observed decision. To formalize this, we define the following set parameterized by a decision x' :

$$\tilde{\mathcal{R}}_{inv}(x') = \left\{ \rho \mid \rho(\vec{Z}^T(x')) \leq \rho(\vec{Z}^T(x)), \forall x \in \mathcal{X}^T \right\}.$$

Note that to facilitate our later discussion in Section 4, we consider here only the case of single observation indexed by T (indicating the most recent observation). Clearly, by fixing $x' := x^T$, the above set reduces to \mathcal{R}_{inv} . Our last criterion can be modelled as follows.

4. Minimizing the deviation from an observed decision:

$$\begin{aligned} \inf_{\rho, x' \in \mathcal{X}} & \|x' - x^T\| \\ \text{subject to } & \rho \in \mathcal{R} \text{ (or } \mathcal{R}_F \text{ in the case of law invariance),} \\ & \rho \in \tilde{\mathcal{R}}_{inv}(x') \cap \mathcal{R}_{ref}(\epsilon^*), \end{aligned} \quad (5)$$

where $\|\cdot\|$ is an arbitrary norm and $\epsilon^* \in (0, \infty]$ is fixed beforehand.

More detailed discussion about the case of multiple observations will be given in Section 4. The above model is particularly useful when there is a preference for the status quo decision x^T . For instance, an investor can prefer that the portfolio generated from the imputed risk function does not differ much from his most current portfolio.

3. Solving the Inverse Problems

In this section, we address first the inverse problems (2)–(4). In particular, we will base our discussions primarily on the inverse problem (2), which best highlights the fundamental complexity behind all the inverse problems. Once we walk through the steps it takes to resolve the complexity,

² Although it may not be immediately obvious, one can actually confirm that the worst-case function ρ is itself a feasible solution to the constraints in (4) (see Lemma EC.3.1) and hence the problem can be equivalently stated as: seek a function $\rho \in \mathcal{R} \text{ (or } \mathcal{R}_F) \cap \mathcal{R}_{inv}(\gamma^*) \cap \mathcal{R}_{el}(\{(L_k, U_k)\}_{k \in \mathcal{K}}) \cap \mathcal{R}_{ref}(\epsilon^*)$ such that $\rho \geq \rho'$, $\forall \rho' \in \mathcal{R} \text{ (or } \mathcal{R}_F) \cap \mathcal{R}_{inv}(\gamma^*) \cap \mathcal{R}_{el}(\{(L_k, U_k)\}_{k \in \mathcal{K}}) \cap \mathcal{R}_{ref}(\epsilon^*)$.

it will also be clear how to solve inverse problems (3)–(4). To put the problem into perspective, let us recast first the inverse problem (2) into

$$\begin{aligned} & \inf_{\rho} \|\rho - \tilde{\rho}\|_{\infty} \\ & \text{subject to } \rho \in \mathcal{R} \text{ (or } \mathcal{R}_F \text{ in the case of law invariance),} \end{aligned} \quad (6)$$

$$\rho(\vec{W}^t) \leq \rho(\vec{W}), \forall \vec{W} \in \mathcal{W}^t, t \in \mathcal{T}, \quad (7)$$

$$\rho(L_k) \leq \rho(U_k), \forall k \in \mathcal{K}, \quad (8)$$

where $\vec{W}^t := \vec{Z}^t(x^t)$ and $\mathcal{W}^t := \left\{ \vec{Z}^t(x) \mid x \in \mathcal{X}^t \right\}$.

The above problem cannot be solved by traditional analysis for inverse optimization due to the non-parametric nature of the risk function ρ . In a more specialized setting where one removes the optimality constraints (7) and replaces the norm $\|\cdot\|_{\infty}$ by a norm that requires only comparing ρ and $\tilde{\rho}$ over finite points (see Remark 3), the problem might be solvable based on the non-parametric method developed in Delage and Li (2018). This method hinges on the observation that if only finitely many points need to be compared (e.g., (8) involving only $|\mathcal{K}| < \infty$ pairs), the problem can be reduced to a finite-dimensional convex program whose size grows polynomially with the number of points needing comparison. Unfortunately, in the setting of the above inverse problem, because of the infinity norm and the optimality constraints (7) it necessarily involves infinitely many points (e.g., $\forall \vec{W} \in \mathcal{W}^t$) that must be compared and thus renders the method of Delage and Li (2018) inapplicable.

While these difficulties may put in doubt the tractability of the inverse problem (6)–(8), our key finding is that it is possible to bypass the difficulties by new analysis based on conjugate duality theory (see, e.g., Rockafellar (1974)). Our goal here is to present from a high-level perspective the key analysis steps to approach the problem. The technical details of the theory and proofs can be found in Appendix EC.1 and EC.4.

3.1. Imputing risk-averse functions

We start by considering the general case where the observed decisions to a forward problem were made over acts, as described in Example 1. That is, we intend to solve the inverse problem (6)–(8) with $\rho \in \mathcal{R}$. From here on, we make the following assumption about the reference risk function $\tilde{\rho}$.

DEFINITION 3. A risk function ρ is called a coherent risk measure if $\rho \in \mathcal{R}$ and it further satisfies $\rho(\lambda \vec{Z}) = \lambda \rho(\vec{Z})$ for any $\lambda \geq 0$ (scale invariance).

ASSUMPTION 1. The reference risk function $\tilde{\rho} \in \mathcal{R}$ is a coherent risk measure.

The above assumption is not stringent because most risk measures applied in practice are coherent risk measures. Moreover, we apply the following well-known representation result of coherent

risk measures and assume that such a representation is available (see Appendix EC.2 for the representations of several popular risk measures).

THEOREM 1. (*Artzner et al. (1999)*) Any coherent risk measure admits the supremum representation of

$$\rho(\vec{Z}) = \sup_{p \in \mathcal{C}} p^\top \vec{Z}, \quad (9)$$

where \mathcal{C} is a non-empty, closed, convex set of probability measures (i.e., $\mathcal{C} \subseteq \Delta \subseteq \mathbb{R}^{|\Omega|}$ with $\Delta := \{p \in \mathbb{R}^{|\Omega|} \mid \mathbf{1}^\top p = 1, p \geq 0\}$).

For convenience, in this paper we say that the coherent risk measure ρ is supported by the set \mathcal{C} . Our first observation to solving the inverse problem (6)–(8) is that it is possible to identify a subset of risk functions that is “sufficiently” large to contain an optimal solution to the inverse problem. We rely on the following definition to characterize this subset of risk functions.

DEFINITION 4. Given a set of random losses $\{\vec{Z}_j\}_{j \in \bar{\mathcal{J}}}$ for some $\bar{\mathcal{J}}$ and a set $\bar{\mathcal{C}} \subseteq \Delta \subseteq \mathbb{R}^{|\Omega|}$, we say that a function ρ is supported by the pair $(\{\vec{Z}_j\}_{j \in \bar{\mathcal{J}}}, \bar{\mathcal{C}})$ if it belongs to the following set of functions:

$$\mathcal{L}(\{\vec{Z}_j\}_{j \in \bar{\mathcal{J}}}, \bar{\mathcal{C}}) := \left\{ \rho_\delta \left| \exists \delta \in \mathbb{R}^{|\bar{\mathcal{J}}|}, \forall \vec{Z}, \rho_\delta(\vec{Z}) = \sup_{p \in \bar{\mathcal{C}}} p^\top \vec{Z} - \max_{j \in \bar{\mathcal{J}}} \{p^\top \vec{Z}_j - \delta_j\} \right. \right\}.$$

We also need the following definition that will be applied throughout the rest of this paper.

DEFINITION 5. Let $\{\vec{X}_j\}_{j \in \mathcal{J}}$ be the random losses in the union $\{\vec{W}^t\}_{t \in \mathcal{T}} \cup \{\vec{L}_k\}_{k \in \mathcal{K}} \cup \{\vec{U}_k\}_{k \in \mathcal{K}} \cup \vec{0}$. Without loss of generality, we assume $\{\vec{X}_t\}_{t \in \mathcal{T}} = \{\vec{W}^t\}_{t \in \mathcal{T}}$.

PROPOSITION 1. Given that Assumption 1 holds and the set of optimal solutions is non-empty, there must exist a function $\rho_\delta \in \mathcal{L}(\{\vec{X}_j\}_{j \in \mathcal{J}}, \mathcal{C})$, where \mathcal{C} is the set that supports the reference risk function $\tilde{\rho}$, which is optimal to the inverse problem (6)–(8). Moreover, given any optimal solution ρ_0 to the inverse problem (6)–(8), there always exists a function $\rho_\delta \in \mathcal{L}(\{\vec{X}_j\}_{j \in \mathcal{J}}, \mathcal{C})$ that is also optimal and bounds from above the solution ρ_0 :

$$\rho_\delta(\vec{Z}) \geq \rho_0(\vec{Z}), \forall \vec{Z}, \quad (10)$$

namely, by setting $\delta_j = \rho_0(\vec{X}_j)$, $\forall j \in \mathcal{J}$.

Proof of Proposition 1 The general strategy of the proof is to show that if there exists a risk function $\rho_0 \in \mathcal{R}$ that is optimal to the inverse problem with some optimal value $u^* < \infty$, there must exist a risk function $\rho_\delta \in \mathcal{L}(\{\vec{X}_j\}_{j \in \mathcal{J}}, \mathcal{C})$ that is also optimal to the problem, namely by setting

$$\delta_j = \rho_0(\vec{X}_j), \forall j \in \mathcal{J}.$$

We leave the details of confirming this claim and that it implies (10) to Appendix EC.4. \square

The above result implies firstly that there is no loss of optimality if we restrict our search of an optimal solution to the set $\mathcal{L}(\{\vec{X}_j\}_{j \in \mathcal{J}}, \mathcal{C})$. This significantly reduces the complexity of solving the inverse problem (6)–(8) because the search over the set $\mathcal{L}(\{\vec{X}_j\}_{j \in \mathcal{J}}, \mathcal{C})$ can be effectively done by searching the space of parameter δ . This provides the assurance that even though one cannot identify a decision maker's risk function through any parametric form, it is still possible to learn the risk function by tuning only a finite number of parameters (i.e., $|\mathcal{J}|$ many parameters). Moreover, we know from the inequality (10) that in the case where the optimal solution to the inverse problem is not unique, the optimal solution found in the set $\mathcal{L}(\{\vec{X}_j\}_{j \in \mathcal{J}}, \mathcal{C})$ would be the most “robust” because it provides the most conservative estimate of risk. What can appear counterintuitive is that the representation of ρ_δ does not depend on the feasible set $\mathcal{W}^t = \left\{ \vec{Z}^t(x) \mid x \in \mathcal{X}^t \right\}$, and one may wonder how an imputed risk function ρ_δ then takes into account the information about the set \mathcal{W}^t . The short answer is that the information would be incorporated into the risk function ρ_δ when it comes to the point of determining the value of the parameter δ ³. To show how the value of δ can be calculated, we first need the following intermediate result. The result shows that searching over the set $\mathcal{L}(\{\vec{X}_j\}_{j \in \mathcal{J}}, \mathcal{C})$ can be equivalently formulated as a finite-dimensional system of constraints over the parameter δ where the solution δ corresponds to how function values may be assigned over the set $\{\vec{X}_j\}_{j \in \mathcal{J}}$ that supports the risk function.

PROPOSITION 2. *Given any $\rho_\delta \in \mathcal{L}(\{\vec{Z}_j\}_{j \in \bar{\mathcal{J}}}, \bar{\mathcal{C}})$, there must exist $y_j \in \bar{\mathcal{C}}$, $j \in \bar{\mathcal{J}}$ such that ρ_δ satisfies the following system of constraints:*

$$\begin{aligned} \rho_\delta(\vec{Z}_j) + y_j^\top (\vec{Z}_i - \vec{Z}_j) &\leq \rho_\delta(\vec{Z}_i), \forall i \neq j, \\ y_j &\in \bar{\mathcal{C}}, \forall j \in \bar{\mathcal{J}}. \end{aligned}$$

Conversely, given any solution $\{y_j^\}_{j \in \bar{\mathcal{J}}}$, $\{\delta_j^*\}_{j \in \bar{\mathcal{J}}}$ that satisfies the system below:*

$$\begin{aligned} \delta_j + y_j^\top (\vec{Z}_i - \vec{Z}_j) &\leq \delta_i, \forall i \neq j, \\ y_j &\in \bar{\mathcal{C}}, \forall j \in \bar{\mathcal{J}}, \end{aligned} \tag{11}$$

there must exist a $\rho_\delta \in \mathcal{L}(\{\vec{Z}_j\}_{j \in \bar{\mathcal{J}}}, \bar{\mathcal{C}})$ that satisfies $\rho_\delta(\vec{Z}_j) = \delta_j^$, namely the function*

$$\rho_{\delta^*}(\vec{Z}) = \sup_{y \in \bar{\mathcal{C}}} y^\top \vec{Z} - \max_{j \in \bar{\mathcal{J}}} \left\{ y^\top \vec{Z}_j - \delta_j^* \right\}. \tag{12}$$

³ To provide a better grasp of this, let us suppose for now that one is able to efficiently determine if δ belongs to the following set:

$$\Theta := \left\{ \delta \in \mathbb{R}^{|\mathcal{J}|} \mid \exists \rho \in \mathcal{R} \cap \mathcal{R}_{inv} \cap \mathcal{R}_{el}(\{(L_k, U_k)_{k \in \mathcal{K}}\}) \text{ such that } \rho(\vec{X}_j) = \delta_j, \forall j \in \mathcal{J} \right\}.$$

Then, one can quickly confirm that because of (10) there exists $\delta \in \Theta$ such that ρ_δ will necessarily satisfy the optimality condition (i.e., $\rho_\delta \in \mathcal{R}_{inv}$). Indeed, we have $\rho_\delta(\vec{W}^t) = \rho_0(\vec{W}^t) \leq \rho_0(\vec{Z}) \leq \rho_\delta(\vec{Z})$, $\forall \vec{Z} \in \mathcal{W}^t$. This explains why the complexity actually lies in dealing with the set Θ . It is not necessarily clear, however how one can tractably search over the set Θ . The fact that $\mathcal{L}(\{\vec{X}_j\}_{j \in \mathcal{J}}, \mathcal{C}) \supseteq \{\rho_\delta \mid \delta \in \Theta\}$ suggests that we may first consider the relaxed problem of searching over the set $\mathcal{L}(\{\vec{X}_j\}_{j \in \mathcal{J}}, \mathcal{C})$.

We next show that by restricting ourselves to the search in the set $\mathcal{L}(\{\vec{X}_j\}_{j \in \mathcal{J}}, \mathcal{C})$, we can also identify how to search in the subset $\mathcal{L}(\{\vec{X}_j\}_{j \in \mathcal{J}}, \mathcal{C}) \cap \mathcal{R}_{inv} \cap \mathcal{R}(\{(L_k, U_k)\}_{k \in \mathcal{K}})$. Namely, it is equivalent to adding additional constraints to the system (11). This final system comprises finite-dimensional convex constraints.

PROPOSITION 3. *Given any $\rho_\delta \in \mathcal{L}(\{\vec{X}_j\}_{j \in \mathcal{J}}, \mathcal{C})$, the function further satisfies $\rho_\delta \in \mathcal{R}_{inv} \cap \mathcal{R}(\{(L_k, U_k)\}_{k \in \mathcal{K}})$ if and only if there exists $y_j \in \mathcal{C}$, $j \in \mathcal{J}$ such that δ satisfies the following system:*

$$\delta_j + y_j^\top (\vec{X}_i - \vec{X}_j) \leq \delta_i, \forall i, j \in \mathcal{J}, i \neq j, \quad (13)$$

$$y_j \in \mathcal{C}, \forall j \in \mathcal{J}, \quad (14)$$

$$y_t^\top \vec{X}_t \leq h_t(y_t), \forall t \in \mathcal{T}, \quad (15)$$

$$\delta_i \leq \delta_j, \forall (i, j) \in \mathcal{B}, \quad (16)$$

where $h_t(y) := \min_x \left\{ y^\top \vec{Z}^t(x) \mid x \in \mathcal{X}^t \right\}$ and $\mathcal{B} := \left\{ (i, j) \in \{1, 2, \dots, \mathcal{J}\}^2 \mid (\vec{X}_i, \vec{X}_j) \in \{(\vec{L}_k, \vec{U}_k)\}_{k \in \mathcal{K}} \right\}$.

One can thus resort to solving the above system (13)–(16) to determine more efficiently the values of the parameters δ_j , $j \in \mathcal{J}$, so that its corresponding risk function ρ_δ would necessarily satisfy all the imposed conditions. It is also clear at this point that the above system indeed incorporates the information about the set $\mathcal{W}^t = \left\{ \vec{Z}^t(x) \mid x \in \mathcal{X}^t \right\}$ (i.e., in the definition of $h_t(y)$).

As the final step, we discuss how to ensure that the risk function ρ_δ also minimizes the objective function $\|\rho_\delta - \tilde{\rho}\|_\infty$. It turns out that the absolute difference between $\rho_\delta(\vec{Z})$ and $\tilde{\rho}(\vec{Z})$ at any random loss \vec{Z} can always be bounded by the difference at one of the random losses from the set $\{\vec{X}_j\}_{j \in \mathcal{J}}$ (which supports the risk function ρ_δ). This is due to the piecewise linear structure embedded in the representation of ρ_δ (i.e., the term $\max_{j \in \mathcal{J}} \{p^\top \vec{X}_j - \delta_j\}$ ⁴). As a result, we need only to seek a risk function ρ_δ that minimizes the absolute differences over finite points (i.e., $\max_{j \in \mathcal{J}} |\rho_\delta(\vec{X}_j) - \tilde{\rho}(\vec{X}_j)|$). Combining this with the above observation that ρ_δ can be found by solving a finite-dimensional convex system over the parameter δ , we arrive at the conclusion that to solve the inverse problem (2), one only needs to solve a convex program over the parameter δ . This main result is presented below, and its detailed proof can be found in Appendix EC.4. We note here that the steps we present to analyze the problem are particularly important from a methodological perspective. This for example enables us to unravel the more complicated case presented in the next section.

THEOREM 2. *Given that Assumption 1 holds and that the set of optimal solutions is non-empty, the inverse optimization problem (2) can be solved by a risk function $\rho_\delta \in \mathcal{L}(\{\vec{X}_j\}_{j \in \mathcal{J}}, \mathcal{C})$, where \mathcal{C} is the support set of the reference risk function $\tilde{\rho}$ and the parameter δ is calculated by solving*

$$\begin{aligned} \min_{\delta \in \mathbb{R}^{|\mathcal{J}|}, y_j \in \mathbb{R}^{|\Omega|}} \quad & \max_{j \in \mathcal{J}} |\delta_j - \tilde{\rho}(\vec{X}_j)| \\ \text{subject to} \quad & (13), (14), (15), (16). \end{aligned} \quad (17)$$

⁴ In particular, the linearity of each piece in the term has the implication that the largest difference can always be found at the support points \vec{X}_j , $j \in \mathcal{J}$.

The problem (17) is a convex optimization problem and is polynomially solvable if

1. given any $t \in \mathcal{T}$, the forward problem $\min_{x \in \mathcal{X}^t} \rho(\vec{Z}^t(x))$ is polynomially solvable in the case where $\rho(\vec{Z}^t(x)) := \sum_{i=1}^{|\Omega|} y_i Z^t(x, \omega_i)$ for any $y \in \mathbb{R}_+^{|\Omega|}$, and
2. the support set \mathcal{C} for $\tilde{\rho}$ is equipped with an oracle that can for any $p \in \mathbb{R}^{|\Omega|}$ either confirm that $p \in \mathcal{C}$ or provide a hyperplane that separates p from \mathcal{C} in polynomial time.

Intuitively, the above theorem summarizes a two-step procedure to identify an optimal risk function to the inverse problem (2). First, it computes the parameter δ by solving (17) so that it can determine the function values $\rho_\delta(\vec{Z})$ over finite points $\{\vec{X}_j\}_{j \in \mathcal{J}}$, namely by setting $\rho_\delta(\vec{X}_j) = \delta_j$, $j \in \mathcal{J}$. Then, it interpolates (and extrapolates) other function values (i.e., $\rho_\delta(\vec{Z})$ for any \vec{Z}), based on the structure of ρ_δ (i.e., the definition of the set $\mathcal{L}(\{\vec{X}_j\}_{j \in \mathcal{J}}, \mathcal{C})$). It is guaranteed by the theorem that a risk function ρ_δ interpolated (and extrapolated) as such will necessarily satisfy all the imposed conditions and reach the optimal value. The main computational complexity of the procedure lies in solving the problem (17), whose complexity, roughly speaking, is in the same order of the complexity of the forward problem. It is thus assured by the theorem that one can always efficiently learn decision makers' risk preference from their past decisions as long as the forward problems that they solved are amenable to efficient solution methods. This is the case, for example, when one tries to learn about investors' risk functions from their past investment decisions, given that many portfolio selection problems (i.e., the forward problem) can be solved efficiently. Note that the condition about the oracle is very mild, which is usually required for proving any general tractability result. Note also that the problem (17) can be solved efficiently as a conic program (Nemirovski (2007)), under mild regularity conditions, if the forward problem and the set \mathcal{C} are conic representable.

To help deepen one's understanding and intuition about how the function values are interpolated (and extrapolated), we present the following two alternative formulations of ρ_δ .

COROLLARY 1. *The risk function $\rho_\delta \in \mathcal{L}(\{\vec{X}_j\}_{j \in \mathcal{J}}, \mathcal{C})$ can be equivalently formulated as*

$$\begin{aligned} \rho_\delta(\vec{Z}) &= \sup_{p, s \in \mathbb{R}} p^\top \vec{Z} - s \\ \text{subject to } & p^\top \vec{X}_j - s \leq \delta_j, \forall j \in \mathcal{J}, \\ & p \in \mathcal{C}. \end{aligned}$$

It can also be formulated as

$$\begin{aligned} \rho_\delta(\vec{Z}) &= \inf_{t \in \mathbb{R}} t \\ \text{subject to } & \vec{Z} - t \in \mathcal{A}, \end{aligned}$$

where $\mathcal{A} := \left\{ \vec{Z} \mid \rho_\delta(\vec{Z}) \leq 0 \right\}$. In the case $\mathcal{C} := \left\{ p \in \Delta \mid \vec{R}_\nu^\top p \leq b_\nu, \nu \in \mathcal{V} \right\}$ and $|\mathcal{V}| < \infty$, we have

$$\mathcal{A} = \left\{ \vec{Z} \mid \exists \theta \in \mathbb{R}_+^{|\mathcal{J}|}, \vartheta \in \mathbb{R}_+^{|\mathcal{V}|}, \vec{Z} \leq \sum_{j \in \mathcal{J}} \theta_j (\vec{X}_j - \delta_j) + \sum_{\nu \in \mathcal{V}} \vartheta_\nu (\vec{R}_\nu - b_\nu), \sum_{j \in \mathcal{J}} \theta_j = 1 \right\}.$$

The first formulation of ρ_δ indicates that once the parameter δ is determined, the function ρ_δ then evaluates any random loss \vec{Z} by a linear function $p^\top \vec{Z} - s$ that bounds from below all δs (i.e., all assigned function values over $\{\vec{X}_j\}_{j \in \mathcal{J}}$) as tightly as possible subject to its sub-gradient p bounded by \mathcal{C} . In the second formulation, the definition of the set \mathcal{A} is well known in the risk theory (see, e.g., [Artzner et al. \(1999\)](#)), which stands for an “acceptance set” (i.e., the set of random losses with acceptable (non-positive) risk). In particular, here the function value $\rho_\delta(\vec{Z})$ can be translated into the minimum amount of cash (i.e., constant t) required to render the final loss $\vec{Z} - t$ acceptable (i.e., lying in the set \mathcal{A}). This sub-level set, as shown above in the second formulation of \mathcal{A} , can be viewed as a monotone convex set generated from the non-negative span of the sets $\{\vec{X}_j - \delta_j\}_{j \in \mathcal{J}}$ and $\{\vec{R}_\nu - b_\nu\}_{\nu \in \mathcal{V}}$. This sheds light on how the whole contour $\left\{ \vec{Z} \mid \rho_\delta(\vec{Z}) = 0 \right\}$ (which determines $\rho_\delta(\vec{Z})$ for any \vec{Z} value) is interpolated (and extrapolated) once δ is determined: namely, having set $\rho_\delta(\vec{X}_j) = \delta_j, \forall j \in \mathcal{J}$ over the finite points is equivalent to fixing first the points $\{\vec{X}_j - \delta_j\}_{j \in \mathcal{J}}$ on the boundary of \mathcal{A} . Then, the rest of the boundary of \mathcal{A} is interpolated by spanning a cone from $\{\vec{R}_\nu - b_\nu\}_{\nu \in \mathcal{V}}$ at every point in the monotone convex hull of $\{\vec{X}_j - \delta_j\}_{j \in \mathcal{J}}$.

Figure 1 provides an illustration, in two states, of the updates of ρ_δ as more information is acquired. In each of the plots, the level set $\left\{ \vec{Z} \mid \rho_\delta(\vec{Z}) = 0 \right\}$ is drawn over losses \vec{Z} bounded between -1 and 1. Note that because of the property of translation invariance, this level set completely characterizes the risk function ρ_δ ⁵. In all three plots, we consider a decision maker optimizing a linear function $\vec{Z}^t(x) = Z^t x \in \mathbb{R}^2$, where $Z^t \in \mathbb{R}^{2 \times 5}$, subject to budget constraints $\mathcal{X} := \left\{ x \in \mathbb{R}^5 \mid \vec{1}^\top x = 1, x \geq 0 \right\}$. In the first plot, we have $Z^1 = \begin{pmatrix} 0.5 & 0.225 & 0.275 & 0.5 & 0.65 \\ 0.15 & 0.05 & -0.1 & -0.15 & -0.1 \end{pmatrix}$ and the observed decision $x^1(3) = 1$ and 0 otherwise. Here, the risk function ρ_δ is updated according to the largest possible δ_1 that is feasible. To verify this, one can see from the dash lines the range of feasible subgradients that render the decision x^1 optimal. For any other convex function that assigns $Z^1 x^1$ a value larger than δ_1 (i.e., the level set is “farther” from $Z^1 x^1$), its subgradient at the point $Z^1 x^1 - \delta_1$ will fall outside the range of feasible subgradients. In the next plot, an additional observation is considered. We have $Z^2 = \begin{pmatrix} 0.55 & 0.425 & 0.5 & 0.75 & 0.90 \\ -0.3 & -0.35 & -0.5 & -0.55 & -0.5 \end{pmatrix}$ and the observed decision $x^2(3) = 1$ and 0 otherwise. One can see that δ_1 , which is feasible in the first plot, is no longer feasible. This is because to make the subgradients at $Z^2 x^2 - \delta_2$ feasible (i.e., rendering x^2 optimal) and retain the convexity of the sublevel set \mathcal{A} , we must reduce δ_1 . One would not be able to find

⁵ That is, $\rho_\delta(\vec{X}) = c \Rightarrow \rho_\delta(\vec{X} - c) = 0$, which implies that any other level set is just the zero-level set shifted along $\vec{1}$ by c .

any other level set that is farther from Z^1x^1 (and resp. Z^2x^2) than δ^1 (and resp. δ^2) and retains the convexity of \mathcal{A} . In the last plot, we assume that a CVaR_β measure with $\beta = 0.4$ is provided as the reference risk function $\tilde{\rho}$, which is supported by the set $\mathcal{C} := \{p \mid \bar{\Gamma}^\top p = 1, 0 \leq p \leq \beta\}$. Now the values of δ_1 and δ_2 can only be feasible if the subgradients at $Z^1x^1 - \delta_1$ and $Z^2x^2 - \delta_2$ are “confined” by \mathcal{C} . Moreover, the set \mathcal{A} is “extrapolated” by \mathcal{C} also (e.g., over the points $\{\vec{Z} \mid Z(\omega_1) < 0\}$).

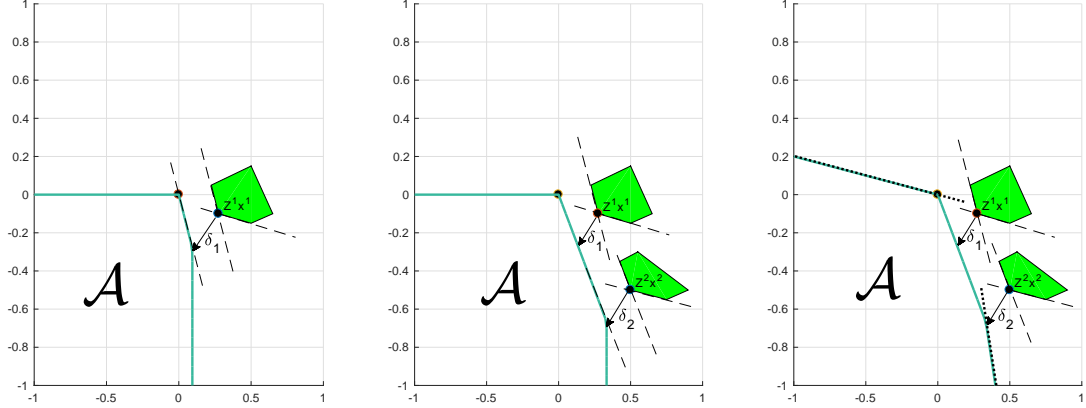


Figure 1 The updates of the set $\mathcal{A} = \{\vec{Z} \mid \rho_\delta(\vec{Z}) \leq 0\}$: with one observed decision (Left), with two observed decisions (Center), and with $\text{CVaR}_{0.4}$ additionally as the reference risk function (Right)

It is valuable to point out here the similarities and differences between our approach and other non-parametric approaches. In particular, [Bertsimas et al. \(2014\)](#) also consider a non-parametric inverse problem, albeit in a very different setting from ours, and show that it is possible to reduce the problem by restricting the search of an optimal solution to a smaller class of functions supported by finite points only. More specifically, they show that to seek a subgradient function $y : \mathbb{R}^{|\Omega|} \rightarrow \mathbb{R}^{|\Omega|}$ that satisfies the optimality conditions

$$y(\vec{W}^t)^\top (\vec{W} - \vec{W}^t) \geq 0, \forall \vec{W} \in \mathcal{W}^t, t \in \mathcal{T}, \quad (18)$$

it suffices to search in the set $\left\{ y : \mathbb{R}^{|\Omega|} \rightarrow \mathbb{R}^{|\Omega|} \mid \exists \alpha_{i,t} \in \mathbb{R}, y_i(\vec{Z}) = \sum_{t \in \mathcal{T}} \alpha_{i,t} k(\vec{W}^t, \vec{Z}), i = 1, \dots, |\Omega| \right\}$, where $k : \mathbb{R}^{|\Omega|} \times \mathbb{R}^{|\Omega|} \rightarrow \mathbb{R}$ is a kernel function. The insight here is that such a class of functions provides sufficient flexibility to locally fit any given values $y(\vec{W}^t)$, $t \in \mathcal{T}$. There is however a fundamental difficulty to apply this method in our setting: namely that it is not amenable to incorporating global properties such as convexity. On the other hand, the class of functions we consider (i.e., $\mathcal{L}(\{\vec{X}_j\}_{j \in \mathcal{J}}, \mathcal{C})$) can both provide the flexibility for local fitting and incorporate the global properties of convex risk functions. We should note, however, that their kernel approach does not rely on convex analysis as we do in this paper and hence might be useful to handle the case of non-convexity where local and global optimality do not coincide.

Although in a different context there is another non-parametric approach for convex interpolation, which has been successfully applied for instance in regression under the name of convex regression (Boyd and Vandenberghe (2004)). This approach seeks a convex function that best fits observed function values over finite points $\{\vec{X}_j\}_{j \in \mathcal{J}}$. The approach involves solving the constraints (13) also to determine the function values over finite points, and it uses a piecewise linear function to interpolate other function values. In this vein, closer to the context of this paper is the work of Delage and Li (2018) mentioned earlier in the section (also Armbruster and Delage (2015)), which in fact involves also solving the constraints (13) to determine function values over finite points and from there to identify the worst-case convex function. Indeed, as these works and our work are both concerned about convexity, the constraint (13) naturally arises. Our work in this sense can be viewed as a generalization of these works: namely that our work further addresses the comparison of function values over infinite points by expanding the system (13) and interpolating via a richer class of functions $\mathcal{L}(\{\vec{X}_j\}_{j \in \mathcal{J}}, \mathcal{C})$.

The discussions up to this point have laid enough ground work for solving other inverse problems. In particular, to solve the inverse problem (3), one needs only note that the property of translation invariance enables us to formulate the constraint in $\mathcal{R}_{inv}(\gamma)$ equivalently as

$$\mathcal{R}_{inv}(\gamma) = \left\{ \rho \mid \rho(\vec{Z}^t(x^t) - \gamma_t) \leq \rho(\vec{Z}^t(x)), \forall x \in \mathcal{X}^t, t \in \mathcal{T} \right\}.$$

Based on the same analysis, we arrive at the following.

COROLLARY 2. *Under the same assumption in Theorem 2, the inverse optimization problem (3) can be solved by a risk function $\rho_\delta \in \mathcal{L}(\{\vec{X}_j\}_{j \in \mathcal{J}}, \mathcal{C})$, where \mathcal{C} is the support set of the reference risk function $\tilde{\rho}$ and the parameter δ is calculated by solving*

$$\begin{aligned} \min_{\delta \in \mathbb{R}^{|\mathcal{J}|}, y_j \in \mathbb{R}^{|\Omega|}, \gamma \in \mathbb{R}^{|\mathcal{T}|}} \quad & \sum_{t \in \mathcal{T}} \gamma_t \\ \text{subject to} \quad & y_t^\top \vec{X}_t \leq h_t(y_t) + \gamma_t, \quad t \in \mathcal{T}, \\ & |\delta_j - \tilde{\rho}(\vec{X}_j)| \leq \epsilon^*, \quad \forall j \in \mathcal{J}, \\ & (13), (14), (16). \end{aligned}$$

To solve the problem (4), we know already from Lemma EC.3.1 that there exists a feasible function ρ that bounds from above all other feasible functions. Following the proof of Proposition 1, one can further confirm that there must exist $\rho_\delta \in \mathcal{L}(\{\vec{X}_j\}_{j \in \mathcal{J}}, \mathcal{C})$ that bounds from above the risk function ρ . We can thus conclude that it suffices to search in the set $\mathcal{L}(\{\vec{X}_j\}_{j \in \mathcal{J}}, \mathcal{C})$ for a worst-case risk function by maximizing all δ s. This can be formulated equivalently as the following problem.

COROLLARY 3. *Under the same assumption in Theorem 2, the inverse optimization problem (4) can be solved by a risk function $\rho_\delta \in \mathcal{L}(\{\vec{X}_j\}_{j \in \mathcal{J}}, \mathcal{C})$, where \mathcal{C} is the support set of the reference risk function $\tilde{\rho}$ and the parameter δ is calculated by solving*

$$\begin{aligned} & \max_{\delta \in \mathbb{R}^{|\mathcal{J}|}, y_j \in \mathbb{R}^{|\Omega|}} \sum_{j \in \mathcal{J}} \delta_j \\ & \text{subject to } y_t^\top \vec{X}_t \leq h_t(y_t) + \gamma_t^*, \quad \forall t \in \mathcal{T}, \\ & |\delta_j - \tilde{\rho}(\vec{X}_j)| \leq \epsilon^*, \quad \forall j \in \mathcal{J}, \\ & (13), (14), (16). \end{aligned}$$

3.2. Imputing permutation-invariant risk-averse functions

We introduce in this section the notion of permutation invariance, which will enable us to identify a subclass of risk-averse functions that can be of practical interest. The notion also provides the basis for the discussion of law invariance in Section 3.3. We say that an operator $\sigma : \mathbb{R}^{|\Omega|} \rightarrow \mathbb{R}^{|\Omega|}$ is a permutation operator over $\vec{Z} \in \mathbb{R}^{|\Omega|}$ if it satisfies $(\sigma(\vec{Z}))_i = (\vec{Z})_{g^{-1}(i)}$ for any $\vec{Z} \in \mathbb{R}^{|\Omega|}$, where $g : \{1, \dots, |\Omega|\} \rightarrow \{1, \dots, |\Omega|\}$ is a bijective function that permutes over $|\Omega|$ elements. We denote by Σ the set of all permutation operators.

DEFINITION 6. (*Permutation-invariant risk-averse functions*) Let $\overline{\overline{\mathcal{R}}}$ denote the set of permutation-invariant risk-averse functions defined by

$$\overline{\overline{\mathcal{R}}} := \left\{ \rho \in \mathcal{R} \mid \rho(\vec{Z}) = \rho(\sigma(\vec{Z})), \forall \sigma \in \Sigma, \forall \vec{Z} \in \mathbb{R}^{|\Omega|} \right\}.$$

Here we should revisit Example 2 in Section 2.1, which provides an important class of permutation-invariant risk-averse functions. Namely, any distributionally robust risk measure ρ^\uparrow defined based on a phi-divergence criteria and an empirical distribution $\hat{p} = \frac{1}{M} \vec{1}$ satisfies $\rho^\uparrow \in \overline{\overline{\mathcal{R}}}$. This is not hard to confirm once one recognizes that the function $D(q, \hat{p})$ satisfies $D(q, \hat{p}) = D(\sigma(q), \hat{p})$, $\forall \sigma \in \Sigma$. Indeed, although a distributionally robust risk measure provides more conservative estimates of risk, in principle it should not be sensitive to the ordering of \vec{Z} given that ρ^\uparrow are built based on samples. More generally, in any case where the decision maker is found to be insensitive to the ordering, one may consider solving the inverse problems by replacing $\rho \in \mathcal{R}$ with $\rho \in \overline{\overline{\mathcal{R}}}$. Moreover, we should assume in the inverse problems that the reference risk function $\tilde{\rho}$ employed is also permutation invariant. As shown in Lemma EC.3.2, it is equivalent to making the following assumption.

ASSUMPTION 2. *The support set of the reference risk function $\tilde{\rho}$, now denoted by $\overline{\overline{\mathcal{C}}}$, satisfies $p \in \overline{\overline{\mathcal{C}}} \Leftrightarrow \sigma(p) \in \overline{\overline{\mathcal{C}}}$.*

It is technically involved however to solve the inverse problems that take into account all the possible permutations. In particular, difficulty arises when one seeks a risk function satisfying the optimality condition that now takes the form

$$\rho(\sigma(\vec{W}^t)) \leq \rho(\sigma'(\vec{W}^t)), \forall \vec{W} \in \mathcal{W}^t, \forall \sigma, \sigma' \in \Sigma, t \in \mathcal{T},$$

where the set $\left\{ \sigma'(\vec{W}) \mid \sigma' \in \Sigma, \vec{W} \in \mathcal{W}^t \right\}$ is non-convex. However, as detailed in Appendix EC.4, Proposition EC.1, one can resolve this difficulty by following closely the steps presented in the previous section. Namely, one can apply the following definition to search over the subset $\bar{\mathcal{L}}(\{\vec{X}_j\}_{j \in \mathcal{J}}, \bar{\mathcal{C}})$, which necessarily contains an optimal solution to the inverse problem, and there again the conjugate duality theory comes to our rescue.

DEFINITION 7. Given a support set $\bar{\mathcal{C}}$ that is permutation invariant, we say that a function $\bar{\rho}_\delta$ is permutation invariant and supported by the pair $(\{\vec{Z}_j\}_{j \in \mathcal{J}}, \bar{\mathcal{C}})$ if it belongs to the following set of functions

$$\bar{\mathcal{L}}(\{\vec{Z}_j\}_{j \in \mathcal{J}}, \bar{\mathcal{C}}) := \left\{ \bar{\rho}_\delta \mid \exists \delta \in \mathcal{R}^{|\mathcal{J}|}, \bar{\rho}_\delta(\vec{Z}) = \sup_{p \in \bar{\mathcal{C}}} p^\top \vec{Z} - \max_{\sigma \in \Sigma, j \in \mathcal{J}} \left\{ p^\top \sigma(\vec{Z}_j) - \delta_j \right\} \right\}.$$

The figure below provides some intuition of the above functions, where we continue the examples presented in Figure 1. In particular, we see that the sub-level set \mathcal{A} now takes a symmetric shape.

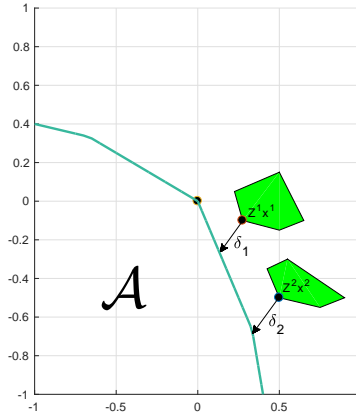


Figure 2 The update of the set $\mathcal{A} = \left\{ \vec{Z} \mid \bar{\rho}_\delta(\vec{Z}) \leq 0 \right\}$ in Figure 1: with $\bar{\rho}_\delta$ now further satisfying permutation-invariance

However, there is also the difficulty of handling the size of the inverse problems, which grows exponentially with respect to the input data of \vec{X}_j because of the need to take into account all the permutations (e.g. $\sigma(\vec{X}_j)$, $\forall \sigma \in \Sigma$). We detail also in Appendix EC.4 how to reduce the problems to programs that grow only polynomially in the size of the input data (e.g., $|\mathcal{J}|$).

PROPOSITION 4. *Given that Assumption 2 holds and that the set of optimal solutions is non-empty, the inverse problem (6)–(8) with $\mathcal{R} := \overline{\mathcal{R}}$ can be solved by a risk function $\overline{\rho}_\delta \in \overline{\mathcal{L}}(\{\vec{X}_j\}_{j \in \mathcal{J}}, \overline{\mathcal{C}})$, where $\overline{\mathcal{C}}$ is the support set of the reference risk function $\tilde{\rho}$ and the parameter δ is calculated by solving*

$$\min_{\delta, y_j, v_{i,j}, w_{i,j}} \max_{j \in \mathcal{J}} |\delta_j - \tilde{\rho}(\vec{X}_j)| \quad (19)$$

$$\text{subject to } \vec{1}^\top v_{i,j} + \vec{1}^\top w_{i,j} \leq \delta_i - \delta_j + y_j^\top \vec{X}_j, \quad \forall j \in \mathcal{J}, \forall i \neq j, \quad (20)$$

$$\begin{aligned} \vec{X}_i y_j^\top - v_{i,j} \vec{1}^\top - \vec{1} w_{i,j}^\top &\leq 0, \quad \forall j \in \mathcal{J}, \forall i \neq j, \\ y_j &\in \overline{\mathcal{C}}, \quad \forall j \in \mathcal{J}, \\ y_t^\top \vec{X}_t &\leq h_t(y_t), \quad \forall t \in \mathcal{T}, \\ \delta_i &\leq \delta_j, \quad \forall (i, j) \in \mathcal{B}, \end{aligned} \quad (21)$$

where $\delta \in \mathbb{R}^{|\mathcal{J}|}$, $y_j \in \mathbb{R}^{|\Omega|}$, $v_{i,j} \in \mathbb{R}^{|\Omega|}$, $w_{i,j} \in \mathbb{R}^{|\Omega|}$, the set $\mathcal{B} := \left\{ (i, j) \in \{1, 2, \dots, \mathcal{J}\}^2 \mid (\vec{X}_i, \vec{X}_j) \in \{(\vec{L}_k, \vec{U}_k)\}_{k \in \mathcal{K}} \right\}$, and h_t denotes the function $h_t(y) := \min_x \left\{ y^\top \vec{Z}^t(x) \mid x \in \mathcal{X}^t \right\}$.

Moreover, the supremum representation in $\overline{\mathcal{L}}(\{\vec{X}_j\}_{j \in \mathcal{J}}, \overline{\mathcal{C}})$ can be reduced to

$$\begin{aligned} \sup_{p \in \mathbb{R}^{|\Omega|}, t \in \mathbb{R}, v_j \in \mathbb{R}^{|\Omega|}, w_j \in \mathbb{R}^{|\Omega|}} p^\top \vec{Z} - t \\ \text{subject to } \vec{1}^\top v_j + \vec{1}^\top w_j &\leq t + \delta_j, \quad \forall j \in \mathcal{J}, \\ \vec{X}_j p^\top - v_j \vec{1}^\top - \vec{1} w_j^\top &\leq 0, \quad \forall j \in \mathcal{J}, \\ p &\in \overline{\mathcal{C}}. \end{aligned} \quad (22)$$

The fact that the size of the above programs (19) and (22) grows only polynomially, rather than exponentially, with respect to the input data should provide a strong incentive to consider applying the above models. In particular, we should emphasize here that the above models offer an opportunity to learn a decision maker’s risk function potentially much faster, in terms of requiring fewer observations of past decisions to reach certain learning performance, than the models presented in the previous section. Indeed, recall from the representation of the risk function $\overline{\rho}_\delta$ that for each random loss \vec{X}_j chosen by an observed decision, the representation will automatically incorporate exponentially many more points (i.e., all the permutations of \vec{X}_j) and carry over the preference relations learned from \vec{X}_j to all its permutations. In other words, the above models allow for incorporating “exponentially” more preference information with the “cost” of taking only polynomially longer time to solve the models. This should stress the importance of always checking first whether a decision maker is sensitive or not to the ordering (i.e., $\sigma(Z)$ for different σ). Following the same analysis, we can derive similar results for the inverse problems (3)–(4). For brevity, we defer them to Appendix EC.3, Corollary EC.1 and EC.2.

3.3. Imputing law-invariant risk-averse functions

As mentioned in Section 2.1, imposing the condition of law invariance on a risk-averse function is equivalent to considering it as a function of distributions. That is, we necessarily assume by default that the distributions of all random variables are available, as discussed also in Example 3, Section 2.1. For this reason, we should start by assuming that all elements involved in the inverse problems are distribution-based.

ASSUMPTION 3. *Each entry of the random loss $\vec{Z}^t(x)$ admits the form of $(\vec{Z}^t(x))_i = Z(x, \xi^t(\omega_i))$, where $\xi^t : \Omega \rightarrow \mathbb{R}^m$. The random vector ξ^t has finite support $\{\xi_1^t, \dots, \xi_{\tau_0^t}^t\}$ and a probability distribution F_{ξ^t} that satisfies $\mathbb{P}(\xi^t = \xi_o^t) = \bar{p}_o^{\xi^t}$ for $o = 1, \dots, \tau_0^t$.*

We also make the following assumption about the reference risk function $\tilde{\rho}$.

ASSUMPTION 4. *The reference risk function $\tilde{\rho}$ is law invariant (i.e., $\tilde{\rho} \in \mathcal{R}_F$).*

The key to solving the inverse problems that account for distributions lies in identifying the connection between the condition of law invariance and permutation invariance discussed in the previous section. Namely, supposing that the probability measure \mathbb{P} is uniform, one can observe that for any two random losses Z_1, Z_2 that share the same distribution, their vector representations $\vec{Z}_1, \vec{Z}_2 \in \mathbb{R}^{|\Omega|}$ must satisfy $\vec{Z}_1 = \sigma(\vec{Z}_2)$ for some $\sigma \in \Sigma$. That is, in the case of uniform probability measure, a law-invariant risk measure must satisfy $\rho \in \overline{\mathcal{R}}$. To exploit this fact further, we make the following mild assumption.

ASSUMPTION 5. *All probability distributions of random losses take rational numbers as probability values.*

In this case, given any discrete probability distribution F_Z specified by a pair of support and probability vector $(\vec{S}, \bar{p}) \in \mathbb{R}^\tau \times \mathbb{R}^\tau$ (i.e., $F_Z = \sum_{o=1}^\tau \bar{p}_o \mathbf{Dirac}((\vec{S})_o)$, where \mathbf{Dirac} is the Dirac measure with all its weight on $(\vec{S})_o$), one can always equivalently express the probability value \bar{p}_o , $o = 1, \dots, \tau$ by a ratio n_o/M , $n_o \in \{1, \dots, M\}$ for some $M \in \mathbb{Z}^+$. The random loss $Z \sim F_Z$ can thus be equivalently defined as a mapping from an outcome space Ω with M uniformly distributed outcomes to \mathbb{R} that satisfies $Z(\omega) \in \{(\vec{S})_1, \dots, (\vec{S})_\tau\}$ and $|\{\omega \in \Omega \mid Z(\omega) = (\vec{S})_o\}| = \bar{p}_o M$, $o = 1, \dots, \tau$. However, it might be costly to implement such a procedure because the constant M might need to be large and thus significantly increases the size of the optimization problems (20) and (22). In the following proposition, we show that the optimization problems can always be further reduced to programs whose sizes depend (almost) only on the size of the supports of distributions (i.e., $|\text{supp}(F_Z)| = \tau$), rather than the size of the outcome space (i.e., M). The proof is deferred to Appendix EC.4.

PROPOSITION 5. Let $\{F_j\}_{j \in \mathcal{J}}$ be the distributions of the random losses in the support set $\{\vec{X}_j\}_{j \in \mathcal{J}}$, and each distribution F_j be specified by a pair of $(\vec{S}_j, \vec{p}^j) \in \mathbb{R}^{\tau_j} \times \mathbb{R}^{\tau_j}$ such that $F_j = \sum_{o=1}^{\tau_j} \vec{p}_o^j \mathbf{Dirac}((\vec{S}_j)_o)$. Given that Assumption 3, 4, and 5 hold and that the set of optimal solutions is non-empty, the inverse problem (6)–(8) with \mathcal{R}_F can be solved by a risk function $\rho_\delta^F \in \mathcal{L}_F(\{F_j\}_{j \in \mathcal{J}}, \mathcal{C}_0)$, where

$$\mathcal{L}_F(\{F_j\}_{j \in \mathcal{J}}, \mathcal{C}_0) := \{\rho_\delta^F \mid \exists \delta \in \mathbb{R}^{|\mathcal{J}|}, \rho_\delta^F(F_Z) = (23)\},$$

$$\begin{aligned} & \sup_{p \in \mathbb{R}^\tau, v_j \in \mathbb{R}^{\tau_j}, w_j \in \mathbb{R}^\tau, t \in \mathbb{R}} p^\top \vec{S} - t \\ & \text{subject to } \vec{1}^\top v_j + \vec{1}^\top w_j \leq t + \delta_j, \quad \forall j \in \mathcal{J}, \\ & \quad \vec{S}_j p^\top - \Lambda_j \circ (v_j \vec{1}^\top) - \vec{1} w_j^\top \leq 0, \quad \forall j \in \mathcal{J}, \\ & \quad p \in \mathcal{C}_0 \subseteq \mathbb{R}_+^\tau, \end{aligned} \tag{23}$$

◦ is the Hadamard product; the coefficient Λ_j is calculated by $(\Lambda_j)_{m,n} = \bar{p}_n / \bar{p}_m^j$, $n = 1, \dots, \tau$, $m = 1, \dots, \tau_j$; and the parameter δ is calculated by solving the following optimization problem

$$\begin{aligned} & \min_{\delta, y_j, v_{i,j}, w_{i,j}} \max_{j \in \mathcal{J}} |\delta_j - \tilde{\rho}(\vec{X}_j)| \\ & \text{subject to } \vec{1}^\top v_{i,j} + \vec{1}^\top w_{i,j} \leq \delta_i - \delta_j + y_j^\top \vec{S}_j, \quad \forall j \in \mathcal{J}, \forall i \neq j, \end{aligned} \tag{24}$$

$$\vec{S}_i y_j^\top - \Lambda_{i,j} \circ (v_{i,j} \vec{1}^\top) - \vec{1} w_{i,j}^\top \leq 0, \quad \forall j \in \mathcal{J}, \forall i \neq j, \tag{25}$$

$$y_j \in \mathcal{C}_j \subseteq \mathbb{R}_+^{\tau_j}, \quad \forall j \in \mathcal{J},$$

$$y_t^\top \vec{S}_t \leq h_t(y_t), \quad \forall t \in \mathcal{T}, \tag{26}$$

$$\delta_i \leq \delta_j, \quad \forall (i, j) \in \mathcal{B},$$

where $\delta \in \mathbb{R}^{|\mathcal{J}|}$, $y_j \in \mathbb{R}^{\tau_j}$, $v_{i,j} \in \mathbb{R}^{\tau_i}$, $w_{i,j} \in \mathbb{R}^{\tau_j}$; $\vec{S}_t := (Z(x^t, \xi_1^t), \dots, Z(x^t, \xi_{\tau_0^t}^t))^\top$; the set $\mathcal{B} := \{(i, j) \in \{1, 2, \dots, \mathcal{J}\}^2 \mid (\vec{X}_i, \vec{X}_j) \in \{(\vec{L}_k, \vec{U}_k)\}_{k \in \mathcal{K}}\}$; and h_t denotes the function $h_t(y) := \min_x \left\{ \sum_{o=1}^{\tau_0^t} y_o Z(x, \xi_o^t) \mid x \in \mathcal{X}^t \right\}$. The coefficient $\Lambda_{i,j}$ is calculated by $(\Lambda_{i,j})_{m,n} = \bar{p}_n^j / \bar{p}_m^i$, $n = 1, \dots, \tau_j$, $m = 1, \dots, \tau_i$.

Moreover, the above set \mathcal{C}_j , $j \in \{0\} \cup \mathcal{J}$ can be derived from the set $\bar{\mathcal{C}}$ (i.e., the support set of the reference risk function $\tilde{\rho}$ in the case where \mathbb{P} is uniform) using

$$\mathcal{C}_j = \left\{ y \mid \mathcal{H}_{F_j}((\lambda_{F_j})^{-1} \circ y) \in \bar{\mathcal{C}} \right\}, \tag{27}$$

where $F_0 := F_Z$, $y \in \mathbb{R}^{\tau_j}$, $\lambda_{F_j} := (\bar{p}_1^j |\Omega|, \dots, \bar{p}_{\tau_j}^j |\Omega|)^\top$ and $(\lambda_{F_j})^{-1} \circ \lambda_{F_j} = \vec{1}$, and $\mathcal{H}_{F_j} : \mathbb{R}^{\tau_j} \rightarrow \mathbb{R}^{|\Omega|}$ stands for an operator associated with F_j that generates a vector in $\mathbb{R}^{|\Omega|}$ from a vector in the dimension of $|\text{supp}(F_j)|$. Specifically, it replicates each entry \tilde{y}_o of a given vector $(\tilde{y}_1, \dots, \tilde{y}_{\tau_j})^\top \in \mathbb{R}^{\tau_j}$ by $\bar{p}_o^j |\Omega|$ many times, where we denote the replications by \tilde{y}_o , and generates a vector $(Y(\omega_1), \dots, Y(\omega_{|\Omega|}))$ in $\mathbb{R}^{|\Omega|}$ by concatenating the replication vectors (i.e., $(\tilde{y}_1^\top, \dots, \tilde{y}_{\tau_j}^\top) := (\tilde{y}_1, \dots, \tilde{y}_1, \tilde{y}_2, \dots, \tilde{y}_2, \dots, \tilde{y}_{\tau_j}, \dots, \tilde{y}_{\tau_j})$).

It is not hard to confirm that the convex programs presented above are similar to those presented in the previous section (see Proposition 4). In particular, the programs here would recover the ones presented in Proposition 4 if we set the size of the support to be the same across all the distributions and assign equal weight to each support. In other words, there is simply more freedom in these programs to describe distributions in terms of both support and probabilities. More importantly, this extra freedom adds no additional computational burden (in the sense that incorporating this information only affects the parameter values, but not the size, of the above convex programs). Thus, this may provide a further reason to encourage the decision maker to disclose the probability information (see Remark 1). Similar results for the inverse problems (3)–(4) can be found in Appendix EC.3, Corollary EC.3 and Corollary EC.4.

It is the complexity of the set \mathcal{C}_j that appears to remain dependent on the size of the outcome space $|\Omega|$. As shown in the examples below, such a dependency can often be removed by considering more specifically the exact form of the reference risk function $\tilde{\rho}$. This would thus allow the whole problem in Proposition 5 to be recast independently from the exact construction of the sample space. In particular, we consider below the implementation of several popular risk measures as the reference risk function $\tilde{\rho}$ and derive their corresponding set \mathcal{C}_j using the relation (27). The distribution F_j here is generally expressed by $F_j = \sum_{o=1}^{\tau_j} \bar{p}_o \mathbf{Dirac}((\vec{S})_o)$. The support set $\bar{\mathcal{C}}$ required in (27) can be identified through the dual representation of these risk measures (see Appendix EC.2) with the use of a uniform probability measure (i.e., $\mathbb{P}(\{\omega_i\}) = \frac{1}{M}$, $\forall \omega_i \in \Omega$).

EXAMPLE 4. (Risk measures and their corresponding set of support \mathcal{C}_j)

1. **(Maximum loss)** The risk function is defined by $\tilde{\rho}(Z) = \max_i \{Z(\omega_i)\}$, and its corresponding set of support \mathcal{C}_j is simply

$$\mathcal{C}_j = \left\{ q \in \mathbb{R}_+^{\tau_j} \mid q^\top \vec{1} = 1 \right\}.$$

2. **(Expectation)** The risk function $\tilde{\rho}(Z) = \mathbb{E}[Z]$ has the following set of support:

$$\mathcal{C}_j = \left\{ q \in \mathbb{R}_+^{\tau_j} \mid q = \bar{p} \right\}.$$

3. **(Mean-upper-semideviation)** The risk function is defined by

$$\tilde{\rho}(Z) = \mathbb{E}[Z] + \gamma \mathbb{E}[|Z - \mathbb{E}[Z]|], \quad \gamma \in [0, \frac{1}{2}],$$

and its set of support takes the form

$$\mathcal{C}_j = \left\{ q \in \mathbb{R}_+^{\tau_j} \mid q_o = \bar{p}_o \left(1 + \gamma \left(h_o - \sum_{o=1}^{\tau_j} \bar{p}_o h_o \right) \right), \quad o = 1, \dots, \tau_j, \quad \sum_{o=1}^{\tau_j} \bar{p}_o |h_o|^t \leq 1, \quad h \geq 0 \right\}.$$

4. **(Conditional value-at-risk (CVaR))** The risk function can be defined by

$$\tilde{\rho}(Z) = \frac{1}{1-\alpha} \int_{\alpha}^1 F_Z^{-1}(t) dt,$$

where F_Z^{-1} stands for the generalized inverse distribution function, and its corresponding set of support admits

$$\mathcal{C}_j = \left\{ q \in \mathbb{R}_+^{\tau_j} \mid q_o \leq \frac{1}{1-\alpha} \bar{p}_o, o = 1, \dots, \tau_j, q^\top \vec{1} = 1 \right\}.$$

5. **(Spectral risk measures)** As a generalization of CVaR, the risk function is defined by

$$\tilde{\rho}(Z) = \int_0^1 F_Z^{-1}(t) \phi(t) dt,$$

where the function $\phi: [0, 1] \rightarrow [0, 1]$ is non-decreasing and satisfies $\int_0^1 \phi(t) dt = 1$. The function ϕ is also known as the risk spectrum. We should note that it is not possible to derive the reduced set \mathcal{C}_j that is independent of the sample space Ω for general spectral risk measures because one can always seek a more “detailed” spectrum ϕ by increasing the size of the sample space. Even so, for practical purposes, a “step-wise” spectrum, $\phi^-(p) := \sum_{k=1}^K \bar{\phi}_k \mathbf{1}_{(p_{k-1}, p_k]}(p)$ for some $0 < \bar{\phi}_1 < \dots < \bar{\phi}_K$ and $0 = p_0 < p_1 < \dots < p_K = 1$, is usually sufficient, which can approximate any general spectrum to a pre-determined precision. By assuming that p_k takes rational numbers as values, we prove at the end of Appendix EC.4 that, based on the representation $\bar{\mathcal{C}} = \text{Conv}(\{\sigma(\bar{\phi}), \sigma \in \Sigma\})$ from Example EC.2.6, the set of support \mathcal{C}_j can be reduced to

$$\mathcal{C}_j = \left\{ q \mid q = \bar{Q} \bar{\phi}, \bar{Q} \vec{1} = \bar{p}, \bar{Q}^\top \vec{1} = p^\phi, \bar{Q} \geq 0 \right\},$$

where $\bar{Q} \in \mathbb{R}^{\tau_j \times K}$ and $(p^\phi)_k := (p_k - p_{k-1})$, $k = 1, \dots, K$.

4. Solving the Inverse Problem in Decision Space

There is a distinct difference, from the complexity point of view, between the inverse problem (5) and the inverse problems addressed in the previous section. In particular, in this setting both the risk function ρ and the decision variable x are variables and the problem becomes non-convex even in the simplest case where ρ is linear (i.e., $\rho(\vec{Z}) = y^\top \vec{Z}$ for some unknown y). The goal of this section is to discover cases where the problem can be solved in a relatively efficient way, by which we mean from a practical standpoint. In particular, we assume the following structure for the problem. Note that for simplicity we suppress the index T in $\vec{Z}^T(x)$ and \mathcal{X}^T in the rest of this section.

ASSUMPTION 6. *The loss function $\vec{Z}(x)$ is linear; namely, it generally admits $\vec{Z}(x) = Zx$ for some $Z \in \mathbb{R}^{|\Omega| \times n}$, and in the case of law invariance it has finite support $\{\xi_o^\top x\}_{o=1}^{\tau_0}$ and a probability distribution F_ξ that satisfies $\mathbb{P}(\xi = \xi_o) = \bar{p}_o^\xi$ for $o = 1, \dots, \tau_0$. In addition, the feasible set \mathcal{X} is a non-empty, bounded full-dimensional polytope that takes the form $\mathcal{X} := \{x \mid Ax \geq b\}$ for some $A \in \mathbb{R}^{N \times n}$ and $b \in \mathbb{R}^N$.*

We show in the following proposition that the inverse problem (5) reduces to solving an inverse linear optimization problem.

PROPOSITION 6. *In the case where Assumption 1 and 6 hold, the inverse problem (5) can be solved by a risk function $\rho_\delta \in \mathcal{L}(\{\vec{X}_j\}_{j \in \{1,2\}}, \mathcal{C})$, where \mathcal{C} is the support set of the reference risk function $\tilde{\rho}$, and $\vec{X}_1 = Zx^*$, $\delta_1 = y^{*\top} Zx^*$, $\vec{X}_2 = \vec{0}$, $\delta_2 = 0$. The values of x^* and y^* are calculated by solving*

$$\begin{aligned} \min_{x' \in \mathbb{R}^n, y \in \mathbb{R}^{|\Omega|}} \quad & \|x' - x^T\| \\ \text{subject to} \quad & x' \in \arg \min_x \{y^\top Zx \mid Ax \geq b\}, \\ & y \in \mathcal{C}. \end{aligned} \tag{28}$$

Moreover, the above problem can be solved by the following mixed-integer program (MIP):

$$\begin{aligned} \min_{x' \in \mathbb{R}^n, y \in \mathbb{R}^{|\Omega|}, u \in \mathbb{R}^N, \eta \in \mathbb{R}^N} \quad & \|x' - x^T\| \\ \text{subject to} \quad & Ax' - b \geq 0, \end{aligned} \tag{29a}$$

$$A^\top u = Z^\top y, \tag{29b}$$

$$Ax' - b \leq M\eta, \tag{29c}$$

$$u \leq M(1 - \eta), \tag{29d}$$

$$u \geq 0, y \in \mathcal{C}, \eta \in \{0, 1\}^N, \tag{29e}$$

for a sufficiently large M .

The problem (28) has also been studied in the context of inverse linear optimization, and it is known that it may be solved in closed-form in the special case where y is unconstrained and the rank of Z equals the dimension of x (Chan et al. (2018)). Unfortunately, the problem in the general form (28) is known to be difficult to solve, as the set of feasible solutions to (28) is non-convex. Nevertheless, the MIP program that we present above can be very relevant from a practical standpoint. In particular, in the case where the 1-norm or ∞ -norm is applied in the objective function, this class of programs can often be solved efficiently as mixed-integer linear programs (MILPs) on a large scale using commercial solvers such as Gurobi or Cplex.

Next, we consider the case where ρ satisfies permutation-invariance. Because we must take into account all possible permutations, the inverse problem (5) in this case is more involved and cannot be reduced as cleanly as (28). Nevertheless, we show in the following proposition that it remains possible to solve the problem through a MIP program where the number of binary variables grows in the order of $O(|\Omega|^2)$.

PROPOSITION 7. *In the case where Assumption 2 and 6 hold, the inverse problem (5) can be solved by a risk function $\bar{\rho}_\delta \in \bar{\mathcal{L}}(\{\vec{X}_j\}_{j \in \{1,2\}}, \bar{\mathcal{C}})$, where $\bar{\mathcal{C}}$ is the support set of the reference risk function $\tilde{\rho}$ and $\vec{X}_1 = Zx^*$, $\delta_1 = y^{*\top} Zx^*$, $\vec{X}_2 = \vec{0}$, and $\delta_2 = 0$. The values of x^* and y^* are calculated by solving the following mixed-integer program (MIP):*

$$\begin{aligned} \min_{x' \in \mathbb{R}^n, y \in \mathbb{R}^{|\Omega|}, u \in \mathbb{R}^N, \eta \in \mathbb{R}^N, \nu_{ij}} \quad & \|x' - x^T\| \\ \text{subject to} \quad & (Zx')_i \leq (Zx')_j + M\nu_{ij}, \quad \forall i, j \in \{1, \dots, |\Omega|\}, i \neq j, \end{aligned} \quad (30a)$$

$$(Zx')_j \leq (Zx')_i + M(1 - \nu_{ij}), \quad \forall i, j \in \{1, \dots, |\Omega|\}, i \neq j, \quad (30b)$$

$$y_i \leq y_j + M\nu_{ij}, \quad \forall i, j \in \{1, \dots, |\Omega|\}, i \neq j, \quad (30c)$$

$$y_j \leq y_i + M(1 - \nu_{ij}), \quad \forall i, j \in \{1, \dots, |\Omega|\}, i \neq j, \quad (30d)$$

$$Ax' - b \geq 0,$$

$$A^\top u = Z^\top y,$$

$$Ax' - b \leq M\eta,$$

$$u \leq M(1 - \eta),$$

$$u \geq 0, y \in \bar{\mathcal{C}}, \nu_{ij} \in \{0, 1\}, \eta \in \{0, 1\}^N,$$

for a sufficiently large M .

As detailed in Appendix EC.4, by carefully examining constraints that take into account all the permutations, the constraints boil down to requiring an ordering matching condition between Zx' and subgradient y (i.e., $(Zx')_{i'} \leq \dots \leq (Zx')_{i''} \Leftrightarrow y_{i'} \leq \dots \leq y_{i''}$). We can equivalently state this condition by the first four constraints (30a)–(30d) using binary variables. To address the case of law invariance, we employ the same technique used in Section 3.3 to first map the distribution $F_{\xi^\top x}$ to a random variable $Zx \in \mathbb{R}^{|\Omega|}$ over an outcome space Ω with a uniform probability measure. Based on this constructed random variable, we can then apply the previous proposition to formulate a MIP program. We show in the following proposition how the program can be further reduced to a program depending only on the size of the support $\{\xi_o^\top x\}_{o=1}^{\tau_0}$.

PROPOSITION 8. *In the case where Assumption 4, 5, and 6 hold, the inverse problem (5) can be solved by a risk function $\rho_\delta^F \in \mathcal{L}_F(\{F_j\}_{j \in \{1,2\}}, \mathcal{C}_0)$, where $F_1 = \sum_{o=1}^{\tau_0} \bar{p}_o^\xi \mathbf{Dirac}((\xi_o^\top x^*)_o)$, $\delta_1 = y^{*\top} \Xi x^*$, $F_2 = \mathbf{Dirac}(0)$, $\delta_2 = 0$, and $\Xi := [\xi_1, \dots, \xi_{\tau_0}]^\top$. The values of x^* and y^* are calculated by solving the following mixed-integer program (MIP):*

$$\begin{aligned} \min_{x' \in \mathbb{R}^n, y \in \mathbb{R}^{\tau_0}, u \in \mathbb{R}^N, \eta \in \mathbb{R}^N, \nu_{ij}} \quad & \|x' - x^T\| \\ \text{subject to} \quad & \xi_i^\top x' \leq \xi_j^\top x' + M\nu_{ij}, \quad \forall i, j \in \{1, \dots, \tau_0\}, i \neq j, \end{aligned}$$

$$\begin{aligned}\xi_j^\top x' &\leq \xi_i^\top x' + M(1 - \nu_{ij}), \quad \forall i, j \in \{1, \dots, \tau_0\}, i \neq j, \\ y_i &\leq y_j + M\nu_{ij}, \quad \forall i, j \in \{1, \dots, \tau_0\}, i \neq j,\end{aligned}\tag{31}$$

$$y_j \leq y_i + M(1 - \nu_{ij}), \quad \forall i, j \in \{1, \dots, \tau_0\}, i \neq j,\tag{32}$$

$$Ax' - b \geq 0,$$

$$A^\top u = \Xi^\top (\bar{p}^\xi \circ y),$$

$$Ax' - b \leq M\eta,$$

$$u \leq M(1 - \eta),$$

$$u \geq 0, y \in \mathcal{C}_1, \nu_{ij} \in \{0, 1\}, \eta \in \{0, 1\}^N,$$

for a sufficiently large M .

Moreover, the set \mathcal{C}_0 can be derived from (27) and the set \mathcal{C}_1 can be derived from $\mathcal{C}_1 := \left\{ y \mid \mathcal{H}_{F_1}((\frac{1}{|\Omega|})y) \in \bar{\mathcal{C}} \right\}$, where $\bar{\mathcal{C}}$ is the support set of the reference risk function \bar{p} in the case of uniform probability measure.

All the MIP programs presented above could be solved in seconds in our experiments conducted in the next section. It is natural to consider extending the inverse problem (5) more generally to incorporate multiple observations (or preference elicitation relations). Unfortunately, not only do these more general cases quickly become intractable to analyze, but it also appears unreasonable to assume that their solutions admit any particular structure as observed in the above propositions. To solve these problems directly, one faces the complexity of bilinear constraints that are highly interdependent, which is computationally intractable even on a small scale.

Although it appears not possible to solve the problems in general, we believe the true value of the above MIP programs lies in providing a reasonable means to “extrapolate” an optimal decision from the observed sub-optimal decision. What we meant by extrapolation here is that the decision is extrapolated from the optimality condition of some risk-averse function over the set of feasible decisions. This guarantee should make the extrapolated decision a more ideal candidate to incorporate in the inverse models than the original observed decision. Hence, we recommend applying the MIP programs in either one of the following two ways when there is a need to incorporate more observations:

1. First, solve the MIP programs based on the latest observation of decision x^T . Obviously, if the observed decision x^T is already optimal, then the MIP programs will return x^T . Otherwise, replace x^T by the solution x^* generated from the programs and then run the inverse model (3) discussed in the previous section by setting the optimality condition associated with x^* to be tight (i.e., $\gamma_T = 0$).

2. Same as above, but run the inverse model (3) without imposing $\gamma_T = 0$.

Based on the results in the above propositions, in either case there must exist at least one feasible risk function (even in case 1) that renders the decision x^* optimal. The difference is that the first approach essentially puts absolute priority on the fitting to the decision x^* , whereas the second approach treats the sub-optimality of each observed decision equally important to improve. Thus, if one's main interest is to fit perfectly the decision (based on most current information), one can apply the first approach. Otherwise, if the interest is to seek a potentially better decision as the input to the other inverse models, one may apply the second approach.

5. Numerical Study

In this section, we illustrate the use of inverse optimization on a portfolio selection problem. We simulate the situation where a fund manager is required to construct a portfolio that aligns with a client's personal preference but has fairly limited opportunity to assess the client's risk preference. We assume that the client's true risk function satisfies the conditions of monotonicity, convexity, and translation invariance and that his/her past investments were made according to the following forward risk minimization problem:

$$\min_x \left\{ \rho \left(- \sum_i x_i \vec{R}_i \right) \mid \vec{1}^\top x = 1, x \geq 0 \right\},$$

where $\vec{R}_i \in \mathbb{R}^{|\Omega|}$ denotes the random returns of an asset i and x_i stands for the proportion of the total wealth invested in the asset i . The non-negativity constraint $x \geq 0$ assumes that the client considers only long positions.

REMARK 4. Although throughout this paper we have assumed the ability to specify the constraints that define the forward problem, we should mention here a few words about the possibility that the constraints may be misspecified. For instance, although the long-only constraint is common in the practice of portfolio management, at times the client may actually be open to taking some short positions without the manager's awareness. The question then arises as to how well in this case our inverse models capture the true risk function (by assuming the long-only constraint in the forward problem) and if it is possible to detect potential misspecification. As discussed with more details in Appendix EC.5, we find it generally not possible to conclude, solely based on observed decisions, if there is any constraint misspecification. However, even without this confirmation, we discuss in the appendix why and how our inverse models may still be effective in capturing the true risk function.

In Section 5.1, we demonstrate the case with limited observations of made decisions. In Section 5.2, we consider more generally the cases where multiple observations are available. All computations are carried out in Matlab 2014a using GUROBI 5.0 as an optimization solver. YALMIP (Lofberg (2004)) is used to implement our models in Matlab.

5.1. The case of single observation

In this section we first consider the case of small $|\mathcal{T}|$, particularly $|\mathcal{T}| = 1$. We assume in all experiments that the client's true risk preference is captured by the risk measure, optimized-certainty-equivalent (OCE), which was first introduced by [Ben-Tal and Teboulle \(2007\)](#). This class of risk measures was widely referred to in both the literature of optimization (e.g., [Natarajan et al. \(2010\)](#)) and risk theory (e.g., [Drapeau and Kupper \(2013\)](#)), given its generality and its one-to-one correspondence to a (dis-)utility function. Namely, it is defined as

$$\rho_{OCE}(\vec{Z}) := \inf_{t \in \mathbb{R}} \left\{ t + \mathbb{E}[u(\vec{Z} - t)] \right\},$$

where the function $u : \mathbb{R} \rightarrow [-\infty, \infty)$ stands for a proper, closed, convex, and nondecreasing disutility function that satisfies $u(0) = 0$ and $\partial u(0) \ni 1$ and ∂u denotes the subdifferential of u . We assume that the client's risk preference is captured by the OCE risk measure, denoted by $\rho_{OCE}^s(\vec{Z})$, with the exponential disutility function $u_s(x) := \frac{1}{s}(e^{sx} - 1)$, where s is a parameter that controls the level of risk aversion. This class of disutility function is fairly standard in the literature (see, e.g., [Natarajan et al. \(2010\)](#)).

We assume further that one chooses a reference risk function $\tilde{\rho}$ by following the safety-first principle but is not fully ignorant of the upside of uncertain outcomes. Specifically, the reference risk function takes the following form of spectral risk measure, where a CVaR-90% is chosen to capture downside risk and a small weight $\lambda = 0.2$ is put on the average:

$$\rho_{Spec}(\vec{Z}) := \lambda \mathbb{E}[\vec{Z}] + (1 - \lambda) \rho_{CVaR-90\%}(\vec{Z}).$$

We ran our experiments against the dataset of daily historical returns from 335 companies that are part of the S&P500 index during the period from January 1997 to November 2013. We conducted 5000 experiments, each consisting of randomly choosing a time window of 60 days and 5 stocks from the 335 companies. The first thirty days of data were used for in-sample calculation, whereas the second thirty days were for out-of-sample evaluation.

The following steps were taken to simulate how the manager may employ imputed risk functions. First, to simulate the past investment, we solved the forward problem based on the OCE risk measure ρ_{OCE}^s with different choices of the parameter s . Then, we fed the obtained portfolio x_{OCE}^s together with the pre-specified spectral risk measure ρ_{Spec} into the model in [Proposition 5](#) to generate an imputed convex risk function ρ_{IC}^s . Finally, we solved the forward risk minimization problem again based on the imputed risk function ρ_{IC}^s to obtain a portfolio x_{IC}^s . Note that in solving the forward problem, it is possible that the optimal portfolio may not be unique. Throughout our experiments, we added a regularization term $\hat{\lambda} \|x\|_2$ to the objective function with a small weight

$\hat{\lambda} = 10^{-6}$ so as to ensure the uniqueness of the optimal solution. Intuitively, we looked for the most diversified portfolio among the optimal portfolios, as the L2 norm is known to encourage the diversification (DeMiguel et al. (2009)).

We compared both in-sample and out-of-sample performances of the portfolios x_{OCE}^s , x_{Spec} , and x_{IC}^s optimized respectively based on the OCE risk measures ρ_{OCE}^s , the spectral risk measure ρ_{Spec} , and the imputed risk functions ρ_{IC}^s . In establishing the outcome space Ω and the associated distribution used in any of these risk functions, we used a uniform distribution constructed based on the first thirty days of joint returns in each sample. In comparing the out-of-sample performances, we additionally compute the optimal out-of-sample portfolios \hat{x}_{OCE}^s and \hat{x}_{Spec} optimized respectively based on the risk measure ρ_{OCE}^s and ρ_{Spec} using the out-of-sample data (i.e., the second thirty days of joint return in each sample). We benchmark the performances of the in-sample portfolios (i.e., x_{OCE}^s , x_{Spec} , and x_{IC}^s) against the optimal out-of-sample portfolios \hat{x}_{OCE}^s and \hat{x}_{Spec} .

Table 1 and 2 present respectively the in-sample and out-of-sample results in terms of the averages. In reading the tables, when an entry corresponds to the portfolio x_{IC}^s or x_{OCE}^s and/or the measure ρ_{OCE}^s parameterized by s , the s -value on the top of each column is the value specifying the parameter. All values in the tables are calculated by averaging the performances over 5000 experiments. We also provide other statistics in terms of boxplot in Figure 3 and 4. Note that the plus sign “+” in the boxplots refers to the average.

portfolio	ρ_{OCE}^s (in lost p.p. relative to x_{OCE}^s)				ρ_{Spec} (in lost p.p. relative to x_{Spec})			
	$s = 0.1$	$s = 1$	$s = 10$	$s = 100$	$s = 0.1$	$s = 1$	$s = 10$	$s = 100$
x_{Spec}	0.33	0.25	0.16	0.04	0.00	0.00	0.00	0.00
x_{IC}^s	0.05	0.03	0.02	0.01	1.27	0.95	0.39	0.04
x_{OCE}^s	0.00	0.00	0.00	0.00	1.86	1.26	0.49	0.08

Table 1 Comparison of the average in-sample performances in lost percentage points (lower is less risky) of the portfolios x_{Spec} , x_{IC}^s , and x_{OCE}^s with respect to different choices of the parameter s . Each portfolio in the left is evaluated based on the true risk measure ρ_{OCE}^s (relative to the performances of x_{OCE}^s), whereas in the right is evaluated based on the reference risk function ρ_{Spec} (relative to the performances of x_{Spec}).

It is not surprising to see in Table 1 that, in terms of in-sample performance, the best-performing portfolios are those optimized according to the measures used for performance evaluation. Note that the portfolios optimized based on the reference risk function (i.e., x_{Spec}) can be deemed unsatisfactory when evaluated according to the true risk measure ρ_{OCE}^s . They underperform the optimal portfolios x_{OCE}^s by an amount up to 33 basis points (i.e., 0.33 p.p.), which can be difficult to justify in terms of their alignment with the performances desired by the client. On the other hand, the portfolios optimized based on the imputed risk functions (i.e., x_{IC}^s) perform more closely to the

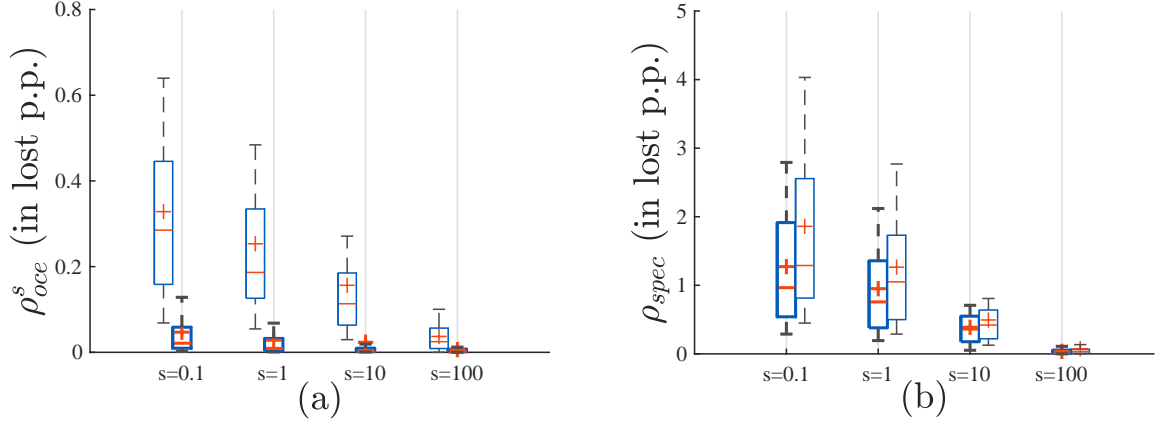


Figure 3 Comparison of the in-sample performances in terms of boxplot. (a) Evaluations based on the true risk measure ρ_{OCE}^s (relative to the performances of x_{OCE}^s), where each pair of boxplots consists of x_{IC}^s (right) and x_{Spec}^s (left). (b) Evaluations based on the reference risk function ρ_{Spec}^s (relative to the performances of x_{Spec}^s), where x_{IC}^s (left) and x_{OCE}^s (right).

optimal portfolios x_{OCE}^s with less than 5 basis points' difference. Note that although by construction the imputed risk functions ρ_{IC}^s guarantee the optimality of the portfolios x_{OCE}^s , minimizing ρ_{IC}^s in the forward problem does not necessarily lead to the same optimal solution (i.e., $x_{IC}^s \neq x_{OCE}^s$). Even so, the benefit of incorporating the solution x_{OCE}^s into the imputed risk function ρ_{IC}^s is still clear when one considers the improvement of x_{IC}^s over x_{Spec}^s in terms of the true risk (i.e., ρ_{OCE}^s). It is expected also from our formulation of the inverse problem that the imputed risk function ρ_{IC}^s should not differ too significantly from the spectral risk measure ρ_{Spec}^s . We can see that the results evaluated based on ρ_{Spec}^s provide the evidence for that (i.e., that the portfolio x_{IC}^s also performs more similarly to the optimal portfolio x_{Spec}^s in this case than the portfolio x_{OCE}^s). This also confirms the effectiveness of the imputed risk functions ρ_{IC}^s to take into account the information contained in the reference risk function ρ_{Spec}^s . Moreover, from Figure 3 we can further see that the improvements of x_{IC}^s over x_{Spec}^s (in terms of ρ_{OCE}^s) and x_{OCE}^s (in terms of ρ_{Spec}^s) are also evident across all the statistics presented in the boxplots. The observation that the performances of x_{IC}^s in fact dominate the performances of the others indicates a clear gain from employing an imputed risk function.

The out-of-sample results presented in Table 2 are calculated by subtracting the performances of the optimal out-of-sample portfolios \hat{x}_{OCE}^s and \hat{x}_{Spec}^s from the performances of the in-sample portfolios x_{OCE}^s , x_{Spec}^s , and x_{IC}^s . Like Table 1, the best performances (i.e., the lowest values) are bold in Table 2. Perhaps quite surprisingly, in the cases of $s = 0.1, 1, 10$, the in-sample portfolio x_{OCE}^s actually underperforms the portfolio x_{Spec}^s when evaluated according to the true risk measure ρ_{OCE}^s . This means that in these cases the portfolios optimized based on ρ_{OCE}^s do not generalize well to out-of-sample data. It is possible to explain this by drawing the connection between the

portfolio	ρ_{OCE}^s (in lost p.p. relative to \hat{x}_{OCE}^s)				ρ_{Spec}^s (in lost p.p. relative to \hat{x}_{Spec}^s)			
	$s = 0.1$	$s = 1$	$s = 10$	$s = 100$	$s = 0.1$	$s = 1$	$s = 10$	$s = 100$
x_{Spec}	0.55	0.42	0.37	1.10	1.12	1.26	1.12	1.11
x_{IC}^s	0.57	0.47	0.37	0.99	2.25	1.79	1.13	1.02
x_{OCE}^s	0.57	0.49	0.39	0.95	2.91	2.08	1.18	0.99

Table 2 Comparison of the average out-of-sample performances in lost percentage points (lower is less risky) of the portfolios x_{Spec} , x_{IC}^s , and x_{OCE}^s with respect to different choices of the parameter s . Each portfolio in the left is evaluated based on the true risk measure ρ_{OCE}^s (relative to the optimal out-of-sample portfolio \hat{x}_{OCE}^s), whereas in the right is evaluated based on the reference risk function ρ_{Spec}^s (relative to the optimal out-of-sample portfolio \hat{x}_{Spec}^s).

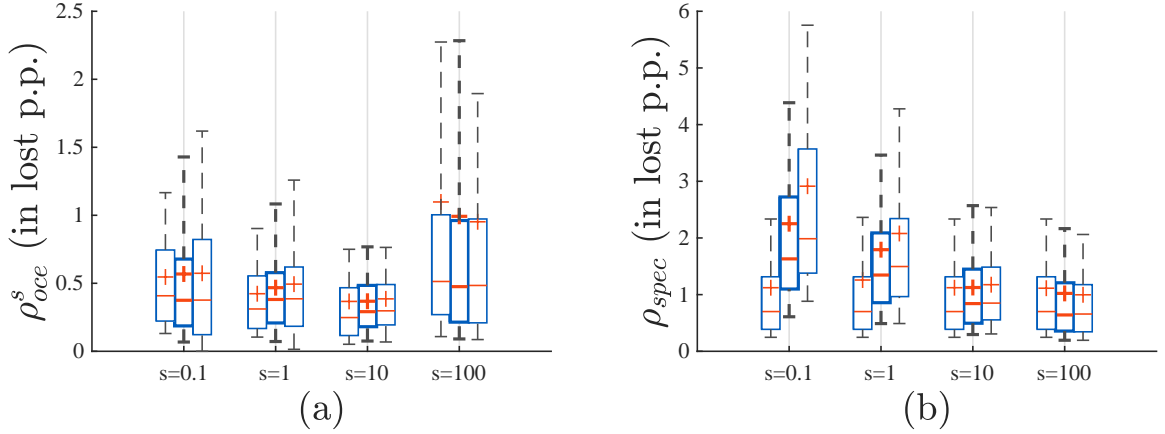


Figure 4 Comparison of the out-of-sample performances in terms of boxplots, where each group of boxplots consists of x_{IC}^s (center), x_{Spec} (left), and x_{OCE}^s (right). (a) Evaluations based on the true risk measure ρ_{OCE}^s (relative to the optimal out-of-sample portfolio \hat{x}_{OCE}^s). (b) Evaluations based on the reference risk function ρ_{Spec}^s (relative to the optimal out-of-sample portfolio \hat{x}_{Spec}^s).

risk measure ρ_{OCE}^s and expectation, namely that ρ_{OCE}^s is close to expectation when s is small. It is well known in portfolio optimization that portfolios optimized based on sample averages are highly unstable and suffer from poor out-of-sample performances. Indeed, we can see from Figure 4 (a) that, in the cases $s = 0.1$ and $s = 1$, the performances of x_{OCE}^s are not only higher in terms of the average but also in terms of the spread, in comparison with that of x_{Spec} . We believe the reason why portfolios optimized based on ρ_{Spec}^s appear more stable is that ρ_{Spec}^s provides more conservative estimates of risk than ρ_{OCE}^s for small s . This can be seen by comparing Figure 4 (a) and (b), where the values of the former are clearly smaller than the latter, except the case of $s = 100$ where the two are more similar. Actually, we see in the case of $s = 100$, which is the most risk-averse case, the portfolio x_{OCE}^s turns to outperform x_{Spec} also in terms of both risk measures ρ_{OCE}^s and ρ_{Spec}^s . These observations appear to align with the common belief in robust optimization that solutions optimized based on more conservative estimates are likely to enjoy more stable performances.

Interestingly, in all cases, the performance of the in-sample portfolio x_{IC}^s consistently falls between the performances of the two portfolios x_{Spec} and x_{OCE}^s in most of the statistics. Hence, in the cases where the portfolios optimized based on the true risk measure ρ_{OCE}^s may suffer from the instability issue, the imputed risk function ρ_{IC}^s can also be considered as a useful surrogate for generating more stable portfolios. That the imputed risk function ρ_{IC}^s resembles only partially the true risk function ρ_{OCE}^s (through an observed decision) and is more similar to the reference risk function ρ_{Spec} otherwise appears to provide a mechanism to dampen the potential over-fitting issue. This also provides a further reason why one should consider choosing a more conservative risk measure as the reference risk function: not only is it practically more sensible (as mentioned in the introduction), but it also provides the basis for the bias-variance tradeoff (i.e., biased towards more conservative estimates with the hope to reduce the variance). Note that from Figure 4 (a), we see that in the cases $s = 0.1$ and $s = 1$, the performance of x_{Spec} is worse than that of x_{OCE}^s in the lower quantiles and this can be the price paid for using more conservative risk measures (i.e., ρ_{Spec}). The imputed risk function ρ_{IC}^s partially corrects the bias (from the observed decision) and reaches a finer level of bias-variance tradeoff. One may consider also tuning the parameter of the reference risk function (e.g., λ in ρ_{Spec}) to explore different tradeoff levels.

5.2. The case of multiple observations

In this section, we consider incorporating multiple observations in the inverse models and focus on studying the resulting performances as $|\mathcal{T}| \rightarrow \infty$. We use the historical returns of the same 335 companies considered in the previous section. Following closely the experiment setup in [Delage and Li \(2018\)](#), here we consider weekly returns and assume that at every given time point the client applies the latest 13 weeks of joint returns to construct plausible scenarios (i.e., $|\Omega| = 13$). In each of our experiments, we randomly draw five assets from the pool of 335 and $|\mathcal{T}|$ samples of their 13-week returns $R^t \in \mathbb{R}^{13 \times 5}$, $t = 1, \dots, |\mathcal{T}|$. Based on each sample R^t , we solve the forward problem using the true risk function, which gives us the observations $(x^t, R^t x, \mathcal{X})$, $t = 1, \dots, |\mathcal{T}|$, where $\mathcal{X} := \left\{x \mid \bar{1}^\top x = 1, x \geq 0\right\}$. We then apply suitable inverse models, as detailed below, to generate the risk functions. To evaluate the performances resulting from the use of the imputed risk functions, we randomly select an alternative sample of 13-week returns \hat{R} of the 5 assets, solve the forward problem based on the sample \hat{R} and the imputed risk function, and evaluate the performance of the resulting portfolio based on the true risk function. We repeat such an experiment 1000 times for evaluating average performance.

In particular, we simulate both the case where the client has incomplete knowledge about the underlying distribution and is ambiguity-averse (like Example 2) and the case where the client is able to identify and reveal the distribution (like Example 3). Finally, we simulate the case where the client did not optimally make the decisions according to the true risk function.

The case of distributional ambiguity We assume in this scenario that the client’s true risk preference is captured by the following form of distributionally robust risk measure

$$\rho_{OCE}^{s,\uparrow}(\vec{Z}) := \sup_{q \in \Delta} \left\{ \rho_{OCE}^s(\vec{Z}; q) \mid \sum_{i=1}^{|\Omega|} |q_i - \hat{q}_i| \leq d \right\},$$

where variation distance is applied to measure the difference between distributions and we set $d = 0.1$. Recall that the notation $\rho_{OCE}^s(\vec{Z}; q)$ refers to the use of distribution q in evaluating the OCE risk measure and that \hat{q} stands for the empirical distribution.

Here, the measure of variation distance is chosen because of its simplicity, which is perhaps easier to interpret than other distance measures. Of course, in our simulation, the manager has no knowledge about this form (neither the function ρ_{OCE}^s nor the uncertainty set of q) and has access only to decisions made according to the function $\rho_{OCE}^{s,\uparrow}(\vec{Z})$. Confirming that the client is insensitive to the ordering of outcomes, the manager can choose to implement the inverse model (4) and assume that the client’s true risk function is permutation invariant (i.e., setting $\rho \in \mathcal{R} := \overline{\mathcal{R}}$ in (4)). The convex program presented in Corollary EC.2 in Appendix EC.3 is implemented.

Figure 5 shows the average performances based on the true risk function $\rho_{OCE}^{s,\uparrow}(\vec{Z})$ for various risk aversion parameter s . We also provide the performance evaluation for the famous $(1/N)$ -investment rule (i.e., setting $x_d = (1/5)$, $d = 1, \dots, 5$), which has often been considered as a popular rule-of-thumb for dealing with ambiguity. The boxplots presented in this figure (and also Figure 6, 7, and 8) provide the statistics of $\rho(\vec{Z}_S(x_{IC}^s(S))) - \rho(\vec{Z}_S(x_{OCE}^s(S))) + \frac{1}{|S|} \sum_{S \in \mathcal{S}} \rho(\vec{Z}_S(x_{OCE}^s(S)))$, where ρ represents the true risk function (i.e., $\rho := \rho_{OCE}^{s,\uparrow}$ in Figure 5), $\vec{Z}_S(\cdot)$ is the total return estimated based on each sample of historical returns (denoted by $S \in \mathcal{S}$), and $x_{IC}^s(S)$ (resp. $x_{OCE}^s(S)$) stands for the portfolio optimized based on the imputed risk function (resp. the true risk function) and each sample S . One can see in Figure 5 that as the number of observations $|\mathcal{T}|$ increases, the performance of portfolios optimized based on the imputed risk functions $\overline{\rho}_\delta$ converge to the performance of portfolios optimized based on the true risk function $\rho_{OCE}^{s,\uparrow}(\vec{Z})$. The rate of convergence is fast for a small number of samples (i.e., the performance is improved most rapidly when $|\mathcal{T}|$ is small), and it generally takes about 10–50 samples to reach a good accuracy (i.e., with less than 0.1 p.p. difference). This is encouraging because it indicates how the risk preference can be efficiently captured even without complete knowledge of the underlying distribution. One can also see that the $(1/N)$ -investment rule unfortunately provides a poor proxy of one’s optimal decision and that the cost of its naive form is high. In this sense, the inverse optimization approach can already serve as a useful alternative even when, for instance, only a single observation is available $|\mathcal{T}| = 1$.

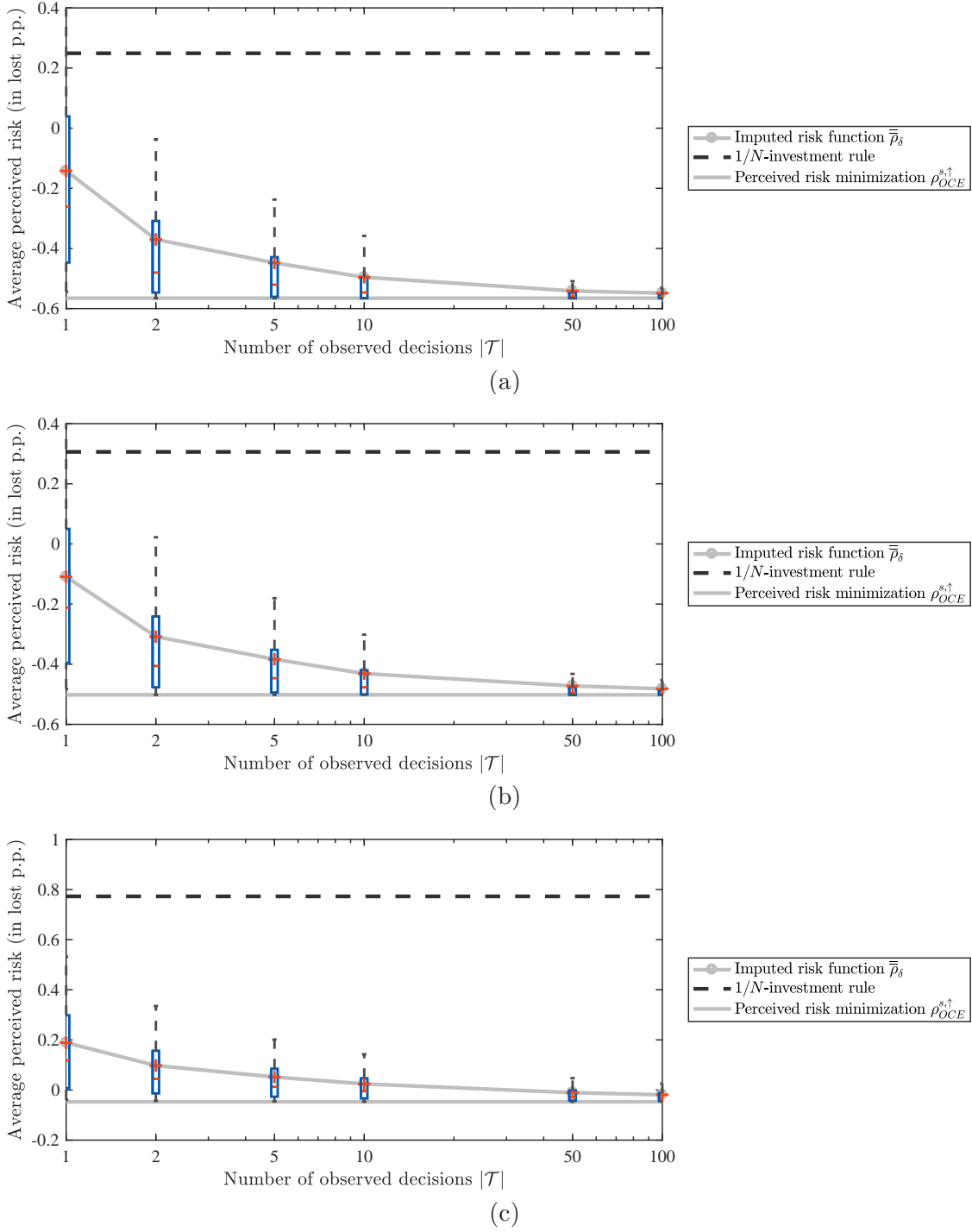
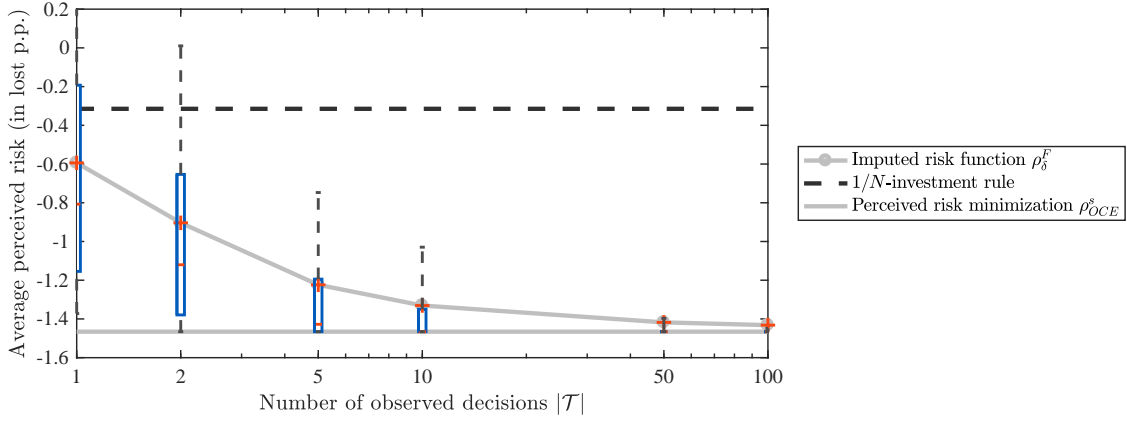


Figure 5 Comparison of the average perceived risk (in lost percentile points) for the portfolio obtained using either imputed risk function $\bar{\rho}_\delta$ or 1/N-investment rule with up to 100 observed decisions. We also report the best average perceived risk that could be obtained if the representation of this perception was exactly known. In particular, the true risk function $\rho_{OCE}^{s,\uparrow}$ above is set by (a) $s = 0.1$, (b) $s = 1$, and (c) $s = 10$.

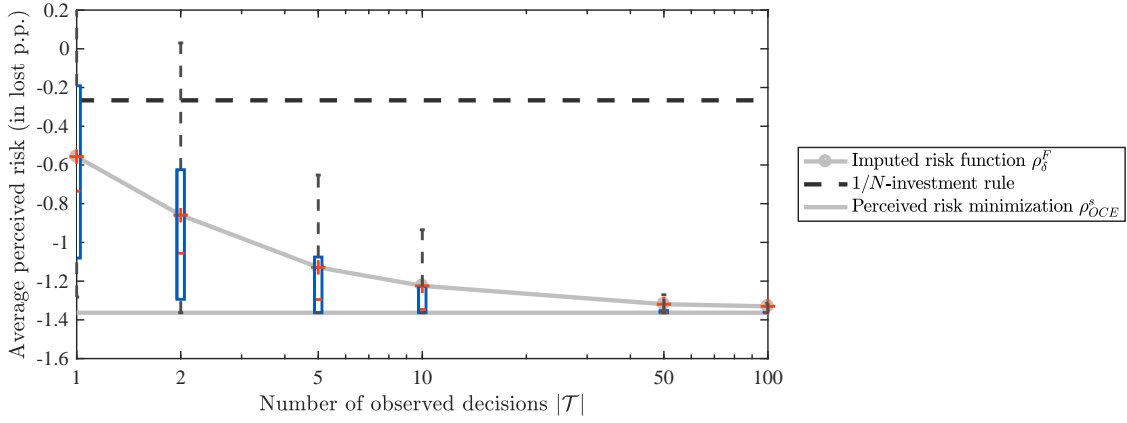
The case of known distributions In this scenario, we assume that the client’s true risk function is the OCE risk measure and that she can identify and reveal the distribution q used for evaluating the risk function $\rho_{OCE}^s(\vec{Z}; q)$. The manager in this case can choose to solve the inverse model (4) and assume that the client’s true risk function is law invariant (i.e., setting $\rho \in \mathcal{R}_F$ in (4)). The convex program presented in Corollary EC.4 in Appendix EC.3 is implemented. In each of our experiments, we sample uniformly from the probability simplex Δ a distribution q to simulate the distribution provided by the client.

Figure 6 shows the average performances evaluated based on the true risk function $\rho_{OCE}^s(\vec{Z})$ for various risk aversion parameters s . We see that the rate of convergence here is similar to what we observed for the case of distributional ambiguity. The only minor difference we noticed is that in the case of large samples $|\mathcal{T}| = 100$, the convergence seems to be slighter stronger in the case of known distributions (i.e., it is more evident that the average performance gets closer to the lower bound), particularly in Figure 6 (a) and (b). While we cannot comment with certainty, this might indicate the benefit of incorporating exact distribution information.

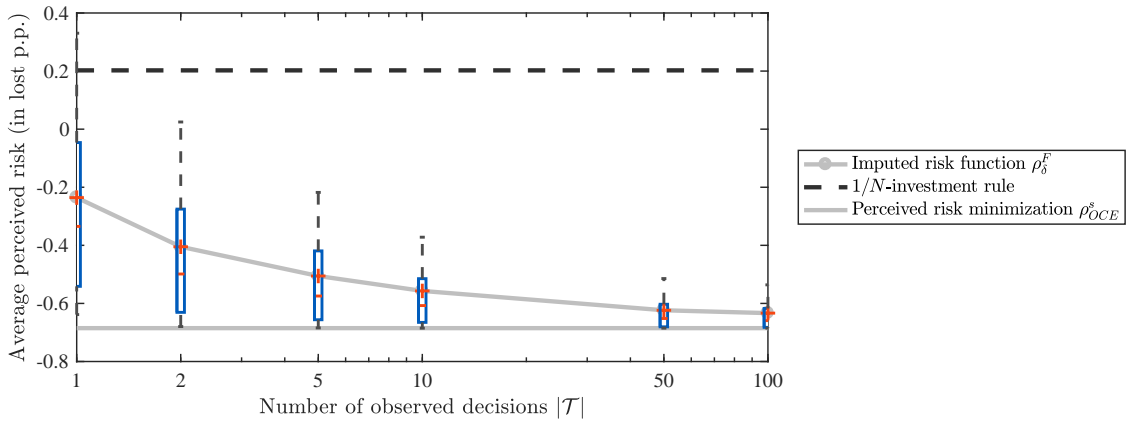
Although we have mentioned earlier that it is important to have the decision maker reveal the distributions in use, one may wonder what would result from some unintended inaccuracy. As a preliminary step to answer this question, we simulate here a situation where the distributions provided by the client for inputting into our inverse models might be different from the actual ones she used for solving the forward problem. As mentioned in Remark 1, it would not be reasonable to consider distributions that are too different from the ones used by the decision maker (i.e., reflecting completely different views of which outcome is more or less likely to occur). Here, we assume that the distributions input into the inverse models are at least similar to those used by the decision maker in that they both assign the probability to each historical return in a non-decreasing fashion with respect to the time of the observed returns. This is an assumption commonly made in finance for example. In other words, the distribution q should satisfy $q \in \Delta_e = \{p \in \Delta \mid p_1 \leq \dots \leq p_{|\Omega|}\}$ where p_i is now indexed according to the time of each observed return (i.e., $p_{|\Omega|}$ is now the probability assigned to the most recent observed return). We stress that the risk function itself does not depend on the time (i.e., it stays constant). In each of the experiments we randomly sample one distribution q' from Δ_e for solving the forward problem to generate observed decisions and then sample another distribution q'' from Δ_e for solving the inverse model. The sampling is done by generating a vector uniformly from the probability simplex Δ and then sorting the vector. Note that in this case we apply first the inverse model (3) because the observed decisions now may not necessarily be optimal with respect to q'' . Once we find the minimum gaps γ , we substitute them into the model (4) and solve (4) to obtain the parameter δ . The convex program presented in Corollary EC.3 in Appendix EC.3 is additionally implemented.



(a)



(b)



(c)

Figure 6 Comparison of the average perceived risk (in lost percentile points) for the portfolio obtained using either imputed risk function ρ_δ^F or $1/N$ -investment rule with up to 100 observed decisions. We also report the best average perceived risk that could be obtained if the representation of this perception was exactly known. In particular, the above true risk function ρ_{OCE}^s is set by (a) $s = 0.1$, (b) $s = 1$, and (c) $s = 10$.

Figure 7 shows the average performance evaluated based on the true risk function $\rho_{OCE}^s(\vec{Z}; q)$ with the true distributions (i.e., $q := q'$). Interestingly, even in this case one can see that the imputed risk functions still greatly benefit from the observed decisions. While the performance does not converge as exactly as the case of known distributions in, for instance, the case $|\mathcal{T}| = 100$, it does exhibit strong convergence rate overall.

The case of sub-optimal observed decisions We consider now the case where the observed decisions are sub-optimal. Here we use the same setting described in the case of distributional ambiguity as the testbed. In each of the experiments, after obtaining the observed decision x^t from optimizing the true risk function (i.e., the distributionally robust risk measure), we generate a sub-optimal decision \hat{x}^t by perturbing x^t as follows:

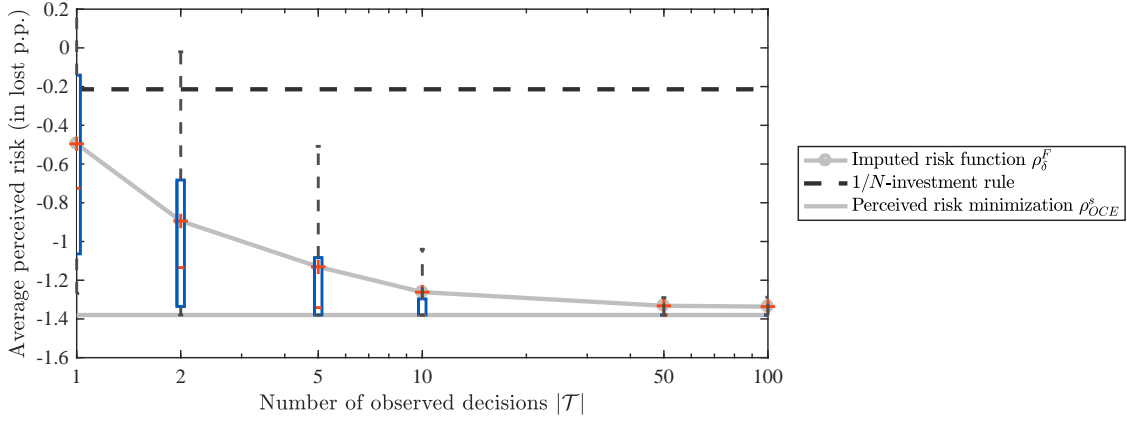
$$\hat{x}^t = \tilde{\lambda}(\frac{1}{5})\vec{1} + (1 - \tilde{\lambda})x^t,$$

where $\tilde{\lambda}$ is uniformly generated from $[0, 0.1]$. That is, we take a convex combination between x^t and a portfolio assigning equal weight to each asset.

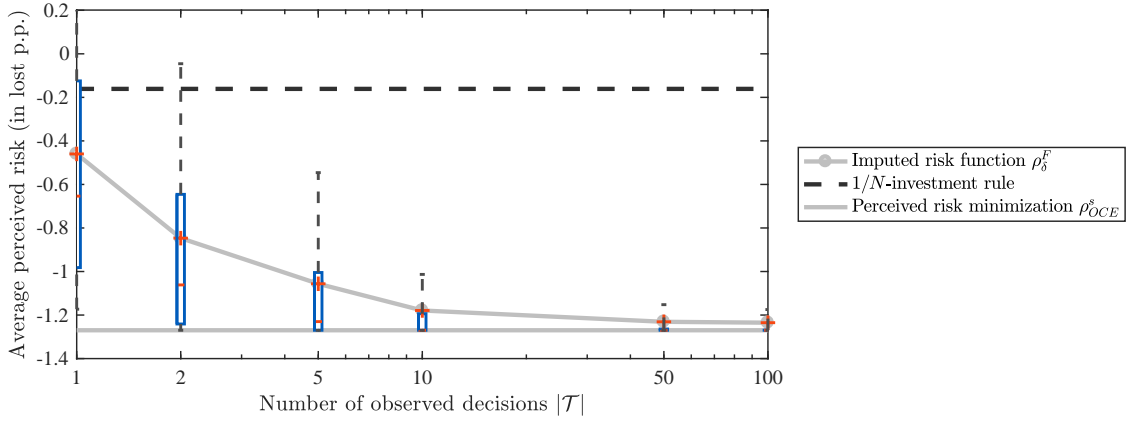
To address the sub-optimality of decisions \hat{x}^t , we consider solving both the inverse model (3) and the inverse model (5). In particular, to implement the inverse model (5), for each sub-optimal decision we apply first the program (30a) to seek an alternative optimal decision \tilde{x}^t and then follow the first approach discussed in the end of Section 4 to incorporate multiple observations.

Figure 8 presents the results of average performances for the portfolios generated from the two approaches (i.e., (3) versus first solving (5) then solving (3)). There appears no noticeable difference between the performances of the portfolios generated from the two approaches. However, in Figure 9 one can see that the optimal portfolios \tilde{x}^t generated based on the model (5) are far closer to the observed sub-optimal decisions \hat{x}^t than the optimal portfolios generated based on the model (3) only. Hence, if the client has a concern about what portfolio x specifically the model would suggest to invest and whether it is aligned with her past investment decision, the model (5) might be easier to justify for that purpose.

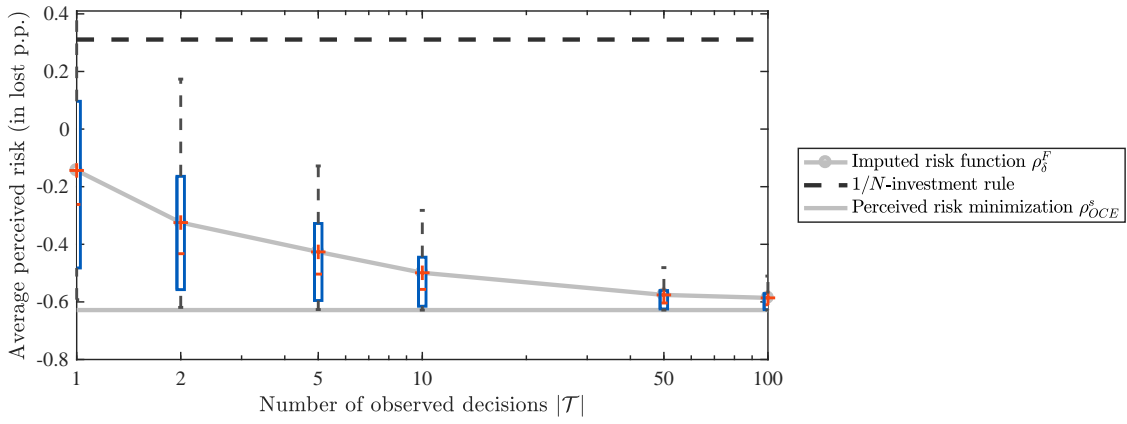
Computation time We present in Table 3 the computation time taken in an experiment for the case of distributional ambiguity. We ran the inverse model (4) for different numbers of assets and outcomes. As seen, the time is not sensitive to the number of assets but grows more noticeably in the number of outcomes. The rate of growth in general is fairly consistent with how the size of the convex program like (20) grows in the number of observed decisions and outcomes. Namely, the program has $O(|\Omega||\mathcal{T}|^2)$ decision variables and $O(|\Omega|^2|\mathcal{T}|^2)$ constraints, given that the set $\{\vec{X}_j\}_{j \in \mathcal{J}}$ consists of only random losses selected by observed decisions x^t , $t \in \mathcal{T}$ (and zero vector $\vec{0}$).



(a)



(b)



(c)

Figure 7 Comparison of the average perceived risk (in lost percentile points) for the portfolio obtained using either imputed risk function ρ_δ^F or $1/N$ -investment rule with up to 100 observed decisions. We also report the best average perceived risk that could be obtained if the representation of this perception was exactly known. In particular, the above true risk function ρ_{OCE}^s is set by (a) $s = 0.1$, (b) $s = 1$, and (c) $s = 10$.

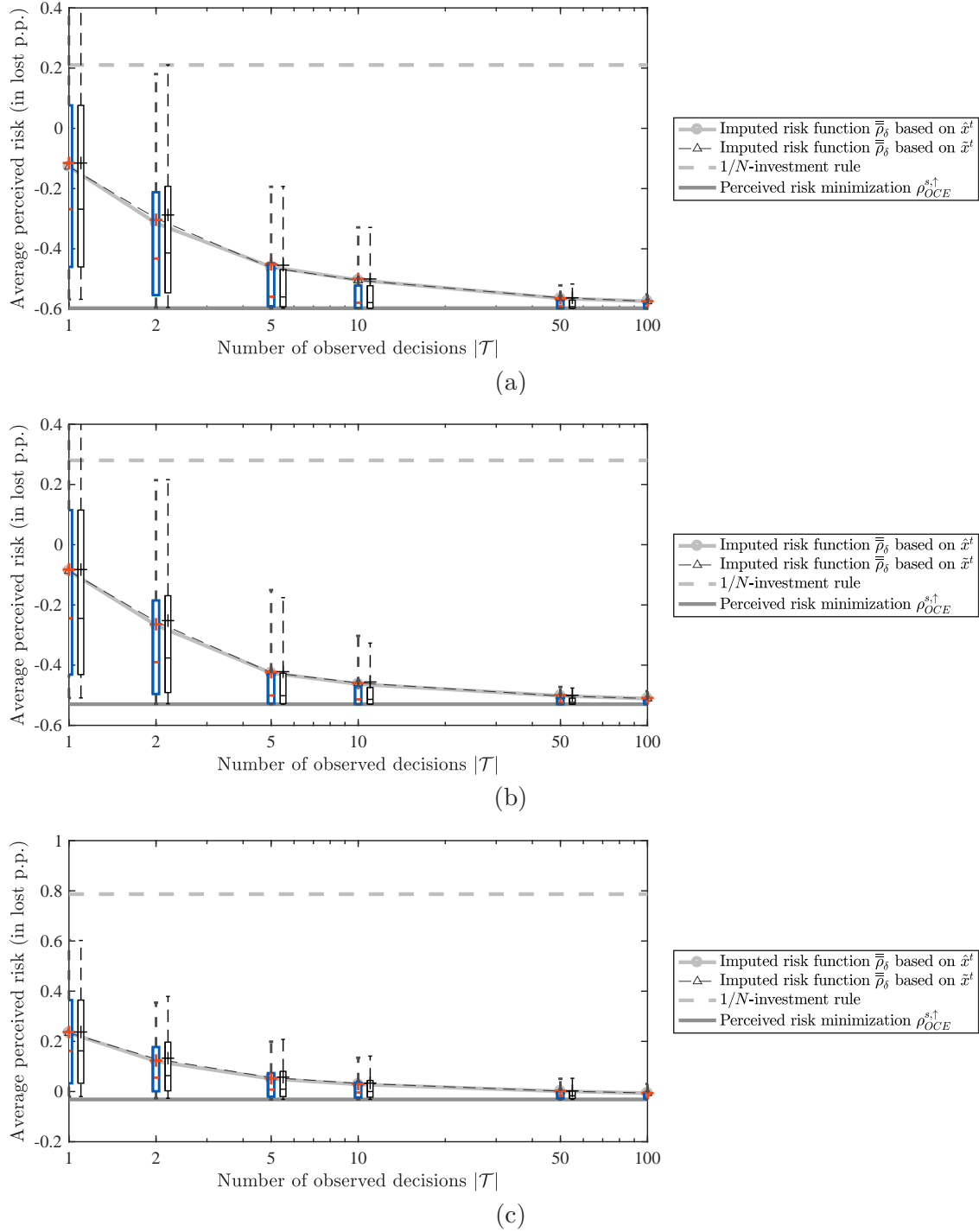


Figure 8 Comparison of the average perceived risk (in lost percentile points) for the portfolio obtained using either imputed risk function $\bar{\rho}_\delta$ or 1/N-investment rule with up to 100 observed decisions. We also report the best average perceived risk that could be obtained if the representation of this perception was exactly known. In particular, the true risk function above $\rho_{OCE}^{s,\uparrow}$ is set by (a) $s = 0.1$, (b) $s = 1$, and (c) $s = 10$. Each pair of boxplots consists of a boxplot based on \hat{x}^t (left) and based on \tilde{x}^t (right).

Number of observed decisions	1	5	10	50	100
Number of assets					
5	0.01	0.07	0.61	22.96	159.00
100	0.01	0.11	0.62	22.05	160.00
300	0.01	0.08	0.52	28.12	165.00
Number of observed decisions	1	5	10	50	100
Number of outcomes					
13	0.01	0.07	0.61	22.96	159.00
26	0.02	0.63	4.28	152.14	863.97
39	0.07	1.06	5.69	373.91	8050.00

Table 3 Computation time in seconds: with fixed number of outcomes $|\Omega|=13$ (Top), and with fixed number of assets, 5 assets (Bottom).

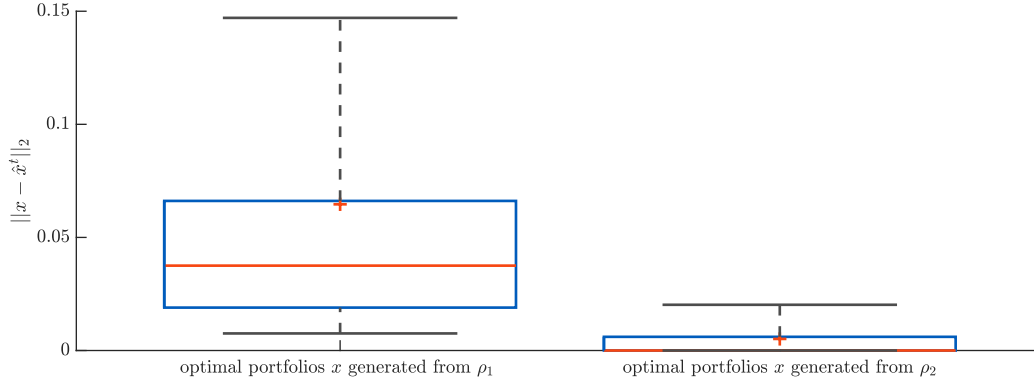


Figure 9 The L2-norm distances between the optimal portfolios generated from imputed risk functions and the observed portfolios \hat{x}^t out of all experiments with sub-optimal observed decisions: ρ_1 denotes risk functions generated from the inverse model (3) (left), and ρ_2 denotes risk functions generated from the inverse model (5) (right).

Out-of-Sample Performance Finally, we conducted out-of-sample testing based also on the setting described in the case of distributional ambiguity. In particular, in each of the experiments we further collected the realized return of each portfolio in the week that follows right after the weeks used for portfolio optimization. Then, we calculated for each portfolio the realized risk (i.e., calculating the distributionally robust risk measure $\rho_{OCE}^{s,\uparrow}$ based on the collected samples). The results can be found in Figure 10. We observe that the sign of convergence remains strong in these results. This provides further evidence of how the imputed risk function closely resembles the true risk function in general. Another observation is that the convergence of the out-of-sample performance in general is not as “regular” as the convergence of the in-sample performance. In particular, in the case of high risk-aversion (i.e., $s = 10$), the realized risk of the portfolios generated based on the imputed risk function already reach a similar level as that of the portfolios generated based on the true risk function, with only a small number of observations. After reaching that

level, the performance then slightly goes up and down as more observations are acquired. We believe that this is related to the fact that the inverse model used in the experiments for the case of distributional ambiguity is the one that seeks the worst-case risk function. As the worst-case function may encourage a more conservative choice of a portfolio, it thus might not be that surprising why such a portfolio already performs well, even with a small number of observed decisions, when evaluated based on a highly risk-averse OCE risk measure.

6. Conclusions

We proposed in this paper a non-parametric inverse optimization framework for risk averse optimization problems involving convex risk functions. The framework allows one to impute a risk function based on preference information acquired from multiple sources, including preference elicitation, observed decisions made for a forward problem, and a reference risk function. The framework can be flexibly applied to generate a risk function based on different criteria. We showed that in general the solution can be efficiently identified by solving convex programs and demonstrated in numerical experiments that the imputed risk function that incorporates the information of past decisions could indeed generate risk estimates that are significantly closer to the true risk level. In addition to the portfolio selection application considered in this paper, other applications such as homeland security ([Haskell et al. \(2018\)](#)) could also benefit from the development of our framework. We leave the study of these other applications as our future work.

Acknowledgments

The author gratefully acknowledges that this research has been supported by the Canadian Natural Sciences and Engineering Research Council under [Grant RGPIN-2014-05602]. The author is indebted to the support of Li Lily Liu in preparing this manuscript. The author thanks the associate editor and three reviewers for their detailed comments and suggestions that greatly improved the quality of the paper.

References

- Acerbi, C. 2002. Spectral measures of risk: A coherent representation of subjective risk aversion. *Journal of Banking & Finance* **26**(7) 1505–1518.
- Ahuja, R. K., J. B. Orlin. 2001. Inverse optimization. *Operations Research* **49**(5) 771–783.
- Armbruster, B., E. Delage. 2015. Decision making under uncertainty when preference information is incomplete. *Management Science* **61**(1) 111–128.
- Artzner, P., F. Delbaen, J. M. Eber, D. Heath. 1999. Coherent Measures of Risk. *Mathematical Finance* **9**(3).
- Aswani, A., Z.-J. M. Shen, A. Siddiq. 2015. Inverse optimization with noisy data. ArXiv preprint arXiv:1507.03266.

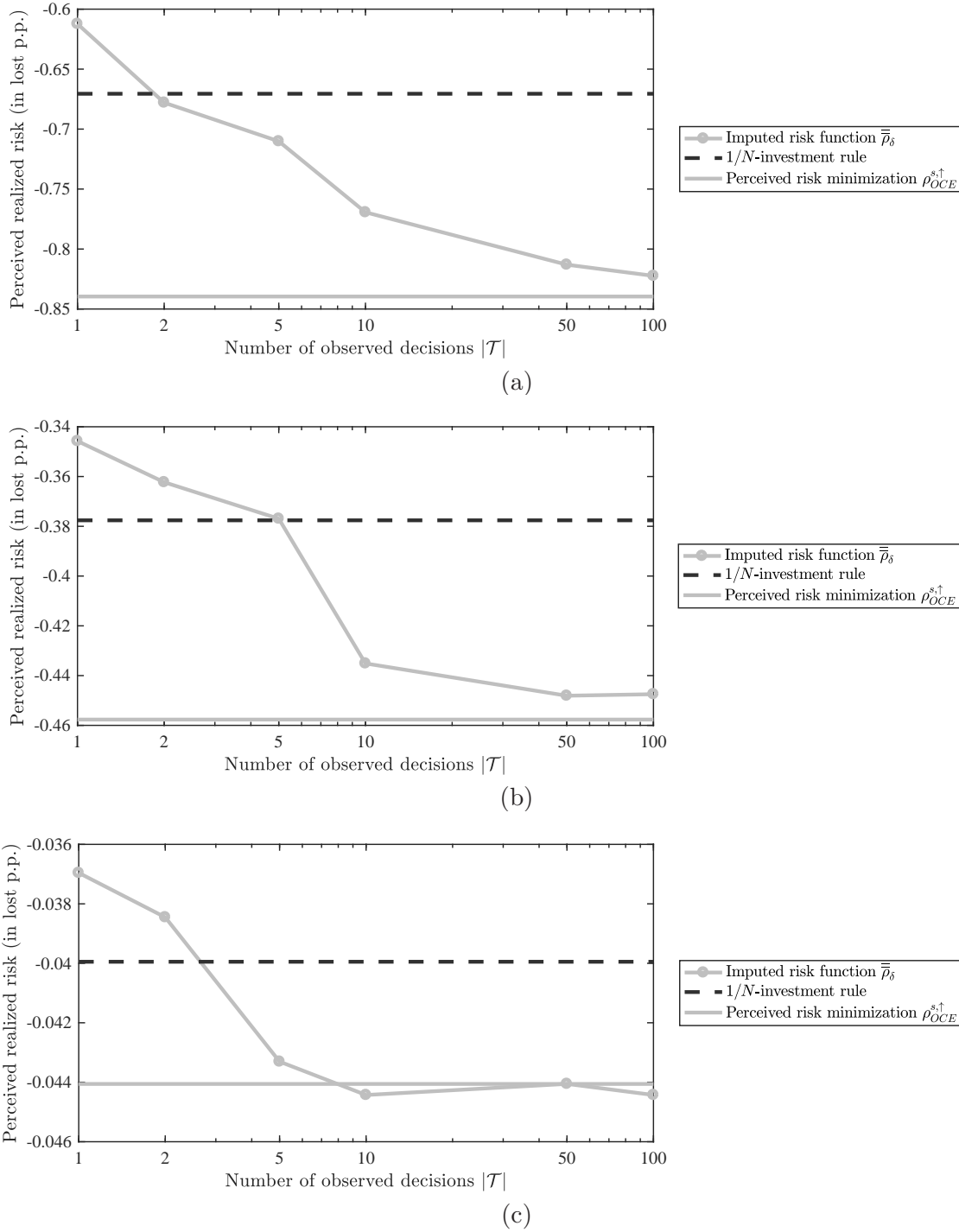


Figure 10 Comparison of the perceived realized risk (in lost percentile points) for the portfolio obtained using either imputed risk function $\bar{\rho}_\delta$ or 1/N-investment rule with up to 100 observed decisions. We also report the best average realized risk that could be obtained if the representation of this perception was exactly known. In particular, the above true risk function $\rho_{OCE}^{s,\uparrow}$ is set by (a) $s = 0.1$, (b) $s = 1$, and (c) $s = 10$.

Ben-Tal, A., D. den Hertog, A. De Waegenare, B. Melenberg, G. Rennen. 2013. Robust solutions of optimization problems affected by uncertain probabilities. *Management Science* **59**(2) 341–357.

-
- Ben-Tal, A., M. Teboulle. 2007. An old-new concept of convex risk measures: The optimized certainty equivalent. *Mathematical Finance* **17**(3) 449–476.
- Bertsimas, D., V. Gupta, I. C. Paschalidis. 2014. Data-driven estimation in equilibrium using inverse optimization. *Mathematical Programming A* 1–39.
- Bertsimas, D., V. Gupta, I. Ch. Paschalidis. 2012. Inverse optimization: A new perspective on the blacklitterman model. *Operations Research* **60**(6) 1389–1403.
- Birge, J. R., A. Hortaçsu, J. M. Pavlin. 2017. Inverse optimization for the recovery of market structure from market outcomes: an application to the miso electricity market. *Operations Research* **65**(4) 837–1113.
- Birkhoff, G. 1946. Three observations on linear algebra. *Universidad Nacional de Tucumán Revista Serie A* **5** 147–151.
- Boutilier, C., R. Patrascu, P. Poupart, D. Schuurmans. 2006. Constraint-based optimization and utility elicitation using the minimax decision criterion. *Artificial Intelligence* **170**(8-9) 686–713.
- Boyd, Stephen, Lieven Vandenberghe. 2004. *Convex Optimization*. Cambridge University Press, New York, NY, USA.
- Burton, D., P. L. Toint. 1992. On an instance of the inverse shortest paths problem. *Mathematical Programming* **53** 45–61.
- Chan, T. C., T. Craig, T. Lee, M. B. Sharpe. 2014. Generalized inverse multiobjective optimization with application to cancer therapy. *Operations Research* **62** 680–695.
- Chan, T. C. Y., T. Lee, D. Terekhov. 2018. Inverse optimization: Closed-form solutions, geometry, and goodness of fit. *Management Science* doi.org/10.1287/mnsc.2017.2992.
- Clemen, R.T., T. Reilly. 2014. Making hard decisions with decisiontools suite. *South-Western, Mason, OH* **3rd ed.**
- Debreu, G. 1954. Representation of a preference ordering by a numerical function. *Decision Processes* 159–165.
- Delage, E., J. Y. Li. 2018. Minimizing risk exposure when the choice of a risk measure is ambiguous. *Management Science* **64**(1) 327–344.
- DeMiguel, V., L. Garlappi, F. J. Nogales, R. Uppal. 2009. A generalized approach to portfolio optimization: improving performance by constraining portfolio norms. *Management Science* **55**(5) 798–812.
- Dempe, S., S. Lohse. 2006. Inverse linear programming. *Recent Advances in Optimization*. Springer, 19–28.
- Drapeau, S., M. Kupper. 2013. Risk preferences and their robust representation. *Mathematics of Operations Research* 1–208.
- Fabozzi, Frank J. 2015. *Capital Markets: Institutions, Instruments, and Risk Management, Fifth Edition*, vol. 1. The MIT Press.

- Föllmer, H., A. Schied. 2002. Convex measures of risk and trading constraints. *Finance and Stochastics* **6**(4) 429–447.
- Ghate, A. 2015. Inverse optimization in countably infinite linear programs. *Operations Research Letters* **43**(3) 231–235.
- Grötschel, M., L. Lovász, A. Schrijver. 1981. The ellipsoid method and its consequences in combinatorial optimization. *Combinatorica* **1**(169–197).
- Haskell, W. B., W. Huang, H. Xu. 2018. Preference elicitation and robust optimization with multi-attribute quasi-concave choice functions. *arXiv:1805.06632* .
- Heuberger, C. 2004. Inverse combinatorial optimization: A survey on problems, methods, and results. *Journal of Combinatorial Optimization* **8** 329–361.
- Hochbaum, D. S. 2003. Efficient algorithms for the inverse spanning-tree problem. *Operations Research* **51** 785–797.
- Hu, J., S. Mehrotra. 2015. Robust decision making over a set of random targets or risk-averse utilities with an application to portfolio optimization. *IIIE Transactions* (47) 1–15.
- Iyengar, G., W. Kang. 2005. Inverse conic programming with applications. *Operations Research Letters* **33**(3) 319–330.
- Keshavarz, A., Y. Wang, S. Boyd. 2011. Imputing a convex objective function. *IEEE International Symposium on Intelligent Control*. 613–619.
- Kusuoka, S. 2001. On law invariant coherent risk measures. Shigeo Kusuoka, Toru Maruyama, eds., *Advances in Mathematical Economics, Advances in Mathematical Economics*, vol. 3. Springer Japan, 83–95.
- Lofberg, J. 2004. Yalmip : A toolbox for modeling and optimization in matlab. *IEEE International Symposium on Computer Aided Control Systems Design*. 284 –289.
- Luo, Z-Q, J-S Pang, D. Ralph. 1996. *Mathematical Programs with Equilibrium Constraints*. Cambridge University Press, Cambridge, UK.
- Mohajerin Esfahani, P., S. Shafieezadeh-Abadeh, G. Adiwena Hanasusanto, D. Kuhn. 2015. Data-driven inverse optimization with incomplete information. ArXiv e-print arXiv:1512.05489.
- Natarajan, K., M. Sim, J. Uichanco. 2010. Tractable robust expected utility and risk models for portfolio optimization. *Mathematical Finance* **20**(4) 695–731.
- Nemirovski, A. 2007. Advances in convex optimization: Conic programming. *Proceedings of the International Congress of Mathematicians*, vol. I. European Mathematical Society, Zürich, 413–444.
- Rockafellar, R. T. 1974. *Conjugate duality and optimization*. Regional Conference Series in Applied Mathematics, SIAM, Philadelphia, PA.
- Rockafellar, R. T., S. Uryasev. 2000. Optimization of conditional value-at-risk. *Journal of Risk* **2** 21–41.

-
- Roy, Arthur D. 1952. Safety first and the holding of assets. *Econometrica* **20**(3) 431–450.
- Ruszczynski, Andrzej, A. Shapiro. 2006. Optimization of convex risk functions. *Mathematics of Operations Research* **31**(3) 433–452.
- Savage, L. J. 1954. *The Foundations of Statistics*. New York: Wiley.
- Schaefer, A. J. 2009. Inverse integer programming. *Optimization Letters* **3** 483–489.
- Sion, M. 1958. On general minimax theorems. *Pacific Journal of Mathematics* **8** 171–176.
- Von Neumann, J, O Morgenstern. 1944. *Theory of Games and Economic Behavior*. Princeton University Press, Princeton, NJ.
- Zhang, J., Z. Liu. 1996. Calculating some inverse linear programming problems. *Journal of Computational and Applied Mathematics* **72**(2) 261–273.
- Zhang, J., C. Xu. 2010. Inverse optimization for linearly constrained convex separable programming problems. *European Journal of Operational Research* **200** 671–679.

Additional Material

EC.1. Results about conjugate duality theory and the function

$$\rho_\delta \in \mathcal{L}(\{\vec{Z}_j\}_{j \in \bar{\mathcal{J}}}, \bar{\mathcal{C}})$$

Recall that the conjugate ρ^* of a function $\rho: \mathbb{R}^{|\Omega|} \rightarrow \mathbb{R}$ is defined as $\rho^*(p) = \sup_{\vec{Z}} \{p^\top \vec{Z} - \rho(\vec{Z})\}$, and the biconjugate ρ^{**} of ρ is defined as $\rho^{**}(\vec{Z}) = \sup_p \{p^\top \vec{Z} - \rho^*(p)\}$.

THEOREM EC.1.1. (*Conjugate Duality Theory (see, e.g., [Rockafellar \(1974\)](#) for detailed proofs)*) Given any function $\rho: \mathbb{R}^{|\Omega|} \rightarrow \mathbb{R}$, the biconjugate ρ^{**} satisfies $\rho^{**} \leq \rho$, and if ρ is convex, then the following must hold.

1. $\rho = \rho^{**}$, and
2. given any $\vec{Z}^* \in \text{dom}(\rho)$ such that $\rho(\vec{Z}^*)$ is subdifferentiable at \vec{Z}^* , the subdifferential satisfies $\partial \rho(\vec{Z}^*) = \partial \rho^{**}(\vec{Z}^*) = \arg \max_p \{p^\top \vec{Z}^* - \rho^*(p)\}$.

In the following lemma, we provide the conjugate of the risk function $\rho_\delta \in \mathcal{L}(\{\vec{Z}_j\}_{j \in \bar{\mathcal{J}}}, \bar{\mathcal{C}})$.

LEMMA EC.1.1. *The conjugate function of the risk function $\rho_\delta \in \mathcal{L}(\{\vec{Z}_j\}_{j \in \bar{\mathcal{J}}}, \bar{\mathcal{C}})$ admits the form*

$$\rho_\delta^*(y) = \begin{cases} \max_{j \in \bar{\mathcal{J}}} \{y^\top \vec{Z}_j - \delta_j^*\} & \text{if } y \in \bar{\mathcal{C}} \\ \infty & \text{if } y \notin \bar{\mathcal{C}} \end{cases},$$

where $\delta_j^* := \rho_\delta(\vec{Z}_j)$, $\forall j \in \bar{\mathcal{J}}$.

Proof of Lemma [EC.1.1](#) By definition, we have

$$\rho_\delta^*(y) = \sup_{\vec{Z}} \{y^\top \vec{Z} - \rho_\delta(\vec{Z})\} \geq \max_{j \in \bar{\mathcal{J}}} \{y^\top \vec{Z}_j - \rho_\delta(\vec{Z}_j)\} = \max_{j \in \bar{\mathcal{J}}} \{y^\top \vec{Z}_j - \delta_j^*\}.$$

To show the other direction, note first that

$$\begin{aligned} \rho_\delta^*(y) &= \sup_{\vec{Z}} \{y^\top \vec{Z} - \rho_\delta(\vec{Z})\} \\ &= \sup_{\vec{Z}} \left\{ y^\top \vec{Z} - \left(\sup_{p \in \bar{\mathcal{C}}} p^\top \vec{Z} - \max_{j \in \bar{\mathcal{J}}} \{p^\top \vec{Z}_j - \delta_j\} \right) \right\} \\ &= \sup_{\vec{Z}} \inf_{p \in \bar{\mathcal{C}}} \left\{ (y - p)^\top \vec{Z} + \max_{j \in \bar{\mathcal{J}}} \{p^\top \vec{Z}_j - \delta_j\} \right\} \\ &= \inf_{p \in \bar{\mathcal{C}}} \sup_{\vec{Z}} \left\{ (y - p)^\top \vec{Z} + \max_{j \in \bar{\mathcal{J}}} \{p^\top \vec{Z}_j - \delta_j\} \right\} \\ &= \max_{j \in \bar{\mathcal{J}}} \{y^\top \vec{Z}_j - \delta_j\}, \text{ if } y \in \bar{\mathcal{C}}, \text{ or } \infty \text{ otherwise,} \end{aligned}$$

where the fourth equality follows the Sion's minimax theorem ([Sion \(1958\)](#)), as the function is linear in \vec{Z} , convex in p , and the set $\bar{\mathcal{C}}$ is compact.

We can then show that

$$\rho_\delta^*(y) = \max_{j \in \bar{\mathcal{J}}} \left\{ y^\top \vec{Z}_j - \delta_j \right\} \leq \max_{j \in \bar{\mathcal{J}}} \left\{ y^\top \vec{Z}_j - \delta_j^* \right\}, \text{ for } y \in \bar{\mathcal{C}},$$

since

$$\delta_j^* = \rho_\delta(\vec{Z}_j) = \sup_{p \in \bar{\mathcal{C}}} \left\{ p^\top \vec{Z}_j - \max_{i \in \bar{\mathcal{J}}} \left\{ p^\top \vec{Z}_i - \delta_i \right\} \right\} \leq \sup_{p \in \bar{\mathcal{C}}} \left\{ p^\top \vec{Z}_j - (p^\top \vec{Z}_j - \delta_j) \right\} = \delta_j.$$

□

EC.2. Dual representation of convex risk functions

THEOREM EC.2.1. (see, e.g., [Ruszczyński and Shapiro \(2006\)](#)) A risk function ρ satisfies the axiom of monotonicity, convexity, and translation invariance if and only if it admits the representation

$$\rho(\vec{Z}) = \sup_{p \in \Delta} \{p^\top \vec{Z} - \rho^*(p)\}, \quad (\text{EC.1})$$

where $\Delta := \left\{ p \in \mathbb{R}^{|\Omega|} \mid \mathbf{1}^\top p = 1, p \geq 0 \right\}$.

The following is a list of risk functions that have been considered in this paper as the candidates for the reference risk function $\tilde{\rho}$.

Risk function	Formulation
Maximum loss	$\max_i \{Z(\omega_i)\}$
Expectation	$\mathbb{E}[Z]$
Mean-absolute-deviation	$\mathbb{E}[Z] + \gamma \mathbb{E}[Z - \mathbb{E}[Z]], \gamma \in [0, \frac{1}{2}]$
Mean-upper-semideviation	$\mathbb{E}[Z] + \gamma (\mathbb{E}[(Z - \mathbb{E}[Z] ^+)^s])^{1/s}, \gamma \in [0, 1], s \geq 1$
Conditional Value-at-Risk (CVaR)	$\frac{1}{1-\alpha} \int_\alpha^1 F_Z^{-1}(t) dt, \alpha \in [0, 1)$
Spectral risk measures	$\int_0^1 F_Z^{-1}(t) \phi(t) dt,$ ϕ is nonnegative, non-decreasing, and $\int_0^1 \phi(t) dt = 1$

Table EC.1 Several well-known risk functions, where F_Z^{-1} stands for the generalized inverse distribution function.

All the risk functions in Table [EC.1](#) are law invariant convex risk functions, and we provide their dual representations based on [\(EC.1\)](#). In Example [EC.2.2](#) to [EC.2.5](#), we denote by $\bar{p} \in \mathbb{R}^{|\Omega|}$ the probability mass function (i.e., $\bar{p}_i = \mathbb{P}(\{\omega_i\}), \omega_i \in \Omega$).

EXAMPLE EC.2.1. (Maximum loss) Its dual representation is simply [\(EC.1\)](#) with $\rho^*(p) = 0$.

EXAMPLE EC.2.2. (Expectation) Its dual representation is also trivial: [\(EC.1\)](#) with $\rho^*(p) = 0$ and Δ replaced by $\mathcal{C} := \{q \mid q = \bar{p}\}$.

EXAMPLE EC.2.3. (Mean-absolute-deviation) Its dual representation has been studied in [Ruszczyński and Shapiro \(2006\)](#): namely, [\(EC.1\)](#) with $\rho^*(p) = 0$ and Δ replaced by $\mathcal{C} := \{q \mid q_i = \bar{p}_i(1 + \gamma(h_i - \sum_i \bar{p}_i h_i)), \|h\|_\infty \leq 1\}$ and $h \in \mathbb{R}^{|\Omega|}$.

EXAMPLE EC.2.4. (Mean-upper-semideviation) The derivation of its dual representation is similar to that of the previous example (see, e.g., [Ruszczyński and Shapiro \(2006\)](#)); namely, the representation is given by

$$\rho(\vec{Z}) = \max_{p \in \mathcal{C}} p^\top \vec{Z},$$

where $\mathcal{C} := \{q \mid q_i = \bar{p}_i(1 + \gamma(h_i - \sum_i \bar{p}_i h_i)), \sum_i \bar{p}_i |h_i|^t \leq 1, h \geq 0\}$, $h \in \mathbb{R}^{|\Omega|}$ and $t = \frac{s}{s-1}$.

EXAMPLE EC.2.5. (Conditional Value-at-Risk (CVaR)) The following dual representation of CVaR is fairly standard:

$$\rho(\vec{Z}) = \max_{p \in \mathcal{C}} p^\top \vec{Z},$$

where $\mathcal{C} := \left\{ q \mid q_i \leq \frac{1}{1-\alpha} \bar{p}_i, 1^\top q = 1, q \geq 0 \right\}$.

Note that in the following example we assume that the probability mass function \bar{p} is uniform so that we can write the dual representation in terms of permutation operators. This provides the basis to derive a more general representation in Example 4.

EXAMPLE EC.2.6. (Spectral risk measures) Given that $\mathbb{P}(\{\omega_i\}) = 1/M$ for $\omega_i \in \Omega$, any spectral risk measure can be equivalently written as ⁶:

$$\rho(\vec{Z}) = \phi^\top \varphi(\vec{Z}),$$

where $\varphi : \mathbb{R}^M \rightarrow \mathbb{R}^M$ denotes an ordering operator such that $\varphi(\vec{Z})_1 \leq \dots \leq \varphi(\vec{Z})_M$, and $\phi \in \mathbb{R}^M$ satisfies $\phi \geq 0$, $\sum_{i=1}^M \phi_i = 1$, and $\phi_1 \leq \dots \leq \phi_M$. It is not difficult to verify that the following dual representation attains the same optimal value as the one above:

$$\rho(\vec{Z}) = \max_{p, \sigma} \left\{ p^\top \vec{Z} \mid p = \sigma(\phi), \sigma \in \Sigma \right\},$$

where σ is an operator that permutes a M -dimensional vector, and Σ is the set of all such operators. It can be further reformulated as follows using the convex hull operator $\text{Conv}(\cdot)$:

$$\rho(\vec{Z}) = \max_{p \in \mathbb{R}_+^M \cap \mathcal{C}} p^\top \vec{Z},$$

where $\mathcal{C} := \{q \mid q \in \text{Conv}(\{\sigma(\phi), \sigma \in \Sigma\})\}$.

⁶ Indeed, letting $\bar{g} : \{1, \dots, M\} \rightarrow \{1, \dots, M\}$ be an one-to-one mapping such that $Z(\omega_{\bar{g}(1)}) \leq Z(\omega_{\bar{g}(2)}) \leq \dots \leq Z(\omega_{\bar{g}(M)})$, we have for any $t \in (\frac{j-1}{M}, \frac{j}{M}]$, $j \in \{1, \dots, M\}$, $F_Z^{-1}(t) = Z(\omega_{\bar{g}(j)})$ must hold. This can be verified via the definition $F_Z^{-1}(t) = \inf\{z : \sum_{Z(\omega_i) \leq z} \mathbb{P}(\{\omega_i\}) \geq t\}$. Namely, since $\mathbb{P}(\{\omega_i\}) = \frac{1}{M}$ we must have $\sum_{Z(\omega_i) \leq Z(\omega_{\bar{g}(j)})} \mathbb{P}(\{\omega_i\}) \geq t^*$, $t^* \in (\frac{j-1}{M}, \frac{j}{M}]$. And there must not exist $z < Z(\omega_{\bar{g}(j)})$ such that $\sum_{Z(\omega_i) \leq z} \mathbb{P}(\{\omega_i\}) \geq t^*$ because such z must be $z \in \{Z(\omega_{\bar{g}(k)})\}_{k=1, \dots, j-1}$ and it contradicts the fact that $\sum_{Z(\omega_i) \leq Z(\omega_{\bar{g}(k)})} \mathbb{P}(\{\omega_i\}) \leq \frac{k}{M}$ for any $k \in \{1, \dots, j-1\}$. Hence, we can write $\rho(\vec{Z}) = \sum_{j=1}^M (\int_{\frac{j-1}{M}}^{\frac{j}{M}} F_Z^{-1}(t) \phi(t) dt) = \sum_{j=1}^M (\int_{\frac{j-1}{M}}^{\frac{j}{M}} \phi(t) dt) Z(\omega_{\bar{g}(j)}) = \sum_{j=1}^M \phi_j Z(\omega_{\bar{g}(j)})$, where $\phi_j := \int_{\frac{j-1}{M}}^{\frac{j}{M}} \phi(t) dt$.

EC.3. Additional Results

LEMMA EC.3.1. *Let $\rho := \sup_{\rho' \in \mathcal{R}_0} \rho'$, where $\mathcal{R}_0 := \mathcal{R}$ (or \mathcal{R}_F) $\cap \mathcal{R}_{inv}(\gamma^*) \cap \mathcal{R}_{el}(\{(L_k, U_k)\}_{k \in \mathcal{K}}) \cap \mathcal{R}_{ref}(\epsilon^*)$ for some fixed $\gamma^* \in [0, \infty)$ and $\epsilon^* \in (0, \infty]$. We have $\rho \in \mathcal{R}_0$.*

Proof of Lemma EC.3.1 We skip the steps of confirming that $\rho \in \mathcal{R}$ (or \mathcal{R}_F) and $\rho \in \mathcal{R}_{el}(\{(L_k, U_k)\}_{k \in \mathcal{K}})$ because these steps can be found in Proposition 1 in [Delage and Li \(2018\)](#). We are left to show that $\rho \in \mathcal{R}_{inv}(\gamma^*) \cap \mathcal{R}_{ref}(\epsilon^*)$. We can confirm that $\rho \in \mathcal{R}_{inv}(\gamma^*)$, because

$$\rho(\vec{W}^t) = \sup_{\rho' \in \mathcal{R}_0} \rho'(\vec{W}^t) \leq \sup_{\rho' \in \mathcal{R}_0} \rho'(\vec{W}) + \gamma^* = \rho(\vec{W}) + \gamma^*, \quad \forall \vec{W} \in \mathcal{W}^t,$$

where we use the fact that the condition $\rho'(\vec{W}^t) \leq \rho'(\vec{W}) + \gamma^*$, $\forall \vec{W} \in \mathcal{W}^t$ holds for any $\rho' \in \mathcal{R}_0$.

We can also confirm that $\rho \in \mathcal{R}_{ref}(\epsilon^*)$, because

$$\rho(\vec{Z}) = \sup_{\rho' \in \mathcal{R}_0} \rho'(\vec{Z}) \leq \tilde{\rho}(\vec{Z}) + \epsilon, \quad \forall \vec{Z},$$

and

$$\rho(\vec{Z}) = \sup_{\rho' \in \mathcal{R}_0} \rho'(\vec{Z}) \geq \rho''(\vec{Z}) \geq \tilde{\rho}(\vec{Z}) - \epsilon, \quad \forall \vec{Z},$$

where ρ'' is some function satisfying $\rho'' \in \mathcal{R}_0$. In the last inequality of the first line we use the fact that the condition $-\epsilon \leq \rho' - \tilde{\rho} \leq \epsilon$ holds for any $\rho' \in \mathcal{R}_0$. \square

LEMMA EC.3.2. *A risk function ρ satisfies $\rho \in \overline{\overline{\mathcal{R}}}$ if and only if it admits a supremum representation $\rho(\vec{Z}) = \sup_p \left\{ p^\top \vec{Z} - \rho^*(p) \right\}$ that satisfies $\rho^*(p) = \rho^*(\sigma(p))$, $\forall \sigma \in \Sigma$. As a result, a coherent risk measure (9) is permutation invariant if and only if its support set \mathcal{C} satisfies $p \in \mathcal{C} \Leftrightarrow \sigma(p) \in \mathcal{C}$.*

Proof of Lemma EC.3.2 We can first show that ρ is indeed permutation invariant given $\rho^*(\sigma(p)) = \rho^*(p)$, since

$$\rho(\sigma(\vec{Z})) = \sup_p p^\top \sigma(\vec{Z}) - \rho^*(p) = \sup_p \sigma^{-1}(p)^\top \vec{Z} - \rho^*(p) = \sup_p p^\top \vec{Z} - \rho^*(\sigma(p)) = \rho(\vec{Z}). \quad (\text{EC.2})$$

To show the other direction, we apply the definition of the conjugate and have for any $\sigma \in \Sigma$,

$$\rho^*(\sigma(p)) = \sup_{\vec{Z}} \vec{Z}^\top \sigma(p) - \rho(\vec{Z}) = \sup_{\vec{Z}} \sigma^{-1}(\vec{Z})^\top p - \rho(\vec{Z}) = \sup_{\vec{Z}} \vec{Z}^\top p - \rho(\sigma(\vec{Z})),$$

and hence for any permutation-invariant ρ (i.e., $\rho(\sigma(\vec{Z})) = \rho(\vec{Z})$), we must have $\rho^*(\sigma(p)) = \rho^*(p)$. In the case of coherent risk measure, the fact that $\rho^*(p)$ satisfies $\rho^*(p) = 0$ for $p \in \mathcal{C}$ and ∞ otherwise implies immediately the symmetry of \mathcal{C} . \square

COROLLARY EC.1. *Under the same assumption in Proposition 4, the inverse optimization problem (3) can be solved by the risk function (22), and the parameter δ is calculated by solving the convex program*

$$\begin{aligned}
& \min_{\delta, y_j, v_{i,j}, w_{i,j}, \gamma} \sum_{t \in \mathcal{T}} \gamma_t \\
& \text{subject to } \bar{\mathbf{I}}^\top v_{i,j} + \bar{\mathbf{I}}^\top w_{i,j} \leq \delta_i - \delta_j + y_j^\top \bar{\mathbf{X}}_j, \quad \forall j \in \mathcal{J}, \forall i \neq j, \\
& \quad \bar{\mathbf{X}}_i y_j^\top - v_{i,j} \bar{\mathbf{I}}^\top - \bar{\mathbf{I}} w_{i,j}^\top \leq 0, \quad \forall j \in \mathcal{J}, \forall i \neq j, \\
& \quad y_j \in \bar{\mathcal{C}}, \quad \forall j \in \mathcal{J}, \\
& \quad y_t^\top \bar{\mathbf{X}}_t \leq h_t(y_t) + \gamma_t, \quad \forall t \in \mathcal{T}, \\
& \quad \delta_i \leq \delta_j, \quad \forall (i, j) \in \mathcal{B}, \\
& \quad |\delta_j - \bar{\rho}(\bar{\mathbf{X}}_j)| \leq \epsilon^*, \quad \forall j \in \mathcal{J},
\end{aligned}$$

where $\delta \in \mathbb{R}^{|\mathcal{J}|}$, $y_j \in \mathbb{R}^M$, $v_{i,j} \in \mathbb{R}^M$, $w_{i,j} \in \mathbb{R}^M$, and $\gamma \in \mathbb{R}^{|\mathcal{T}|}$; the set $\mathcal{B} := \left\{ (i, j) \in \{1, 2, \dots, \mathcal{J}\}^2 \mid (\bar{\mathbf{X}}_i, \bar{\mathbf{X}}_j) \in \{(\bar{\mathbf{L}}_k, \bar{\mathbf{U}}_k)\}_{k \in \mathcal{K}} \right\}$; and h_t denotes the function $h_t(y) := \min_x \left\{ y^\top \bar{\mathbf{Z}}^t(x) \mid x \in \mathcal{X}^t \right\}$.

COROLLARY EC.2. *Under the same assumption in Proposition 4, the inverse optimization problem (4) can be solved by the risk function (22), and the parameter δ is calculated by solving the convex program*

$$\begin{aligned}
& \max_{\delta, y_j, v_{i,j}, w_{i,j}} \sum_{j \in \mathcal{J}} \delta_j \\
& \text{subject to } \bar{\mathbf{I}}^\top v_{i,j} + \bar{\mathbf{I}}^\top w_{i,j} \leq \delta_i - \delta_j + y_j^\top \bar{\mathbf{X}}_j, \quad \forall j \in \mathcal{J}, \forall i \neq j, \\
& \quad \bar{\mathbf{X}}_i y_j^\top - v_{i,j} \bar{\mathbf{I}}^\top - \bar{\mathbf{I}} w_{i,j}^\top \leq 0, \quad \forall j \in \mathcal{J}, \forall i \neq j, \\
& \quad y_j \in \bar{\mathcal{C}}, \quad \forall j \in \mathcal{J}, \\
& \quad y_t^\top \bar{\mathbf{X}}_t \leq h_t(y_t) + \gamma_t^*, \quad \forall t \in \mathcal{T}, \\
& \quad \delta_i \leq \delta_j, \quad \forall (i, j) \in \mathcal{B}, \\
& \quad |\delta_j - \bar{\rho}(\bar{\mathbf{X}}_j)| \leq \epsilon^*, \quad \forall j \in \mathcal{J},
\end{aligned}$$

where $\delta \in \mathbb{R}^{|\mathcal{J}|}$, $y_j \in \mathbb{R}^M$, $v_{i,j} \in \mathbb{R}^M$, $w_{i,j} \in \mathbb{R}^M$; the set $\mathcal{B} := \left\{ (i, j) \in \{1, 2, \dots, \mathcal{J}\}^2 \mid (\bar{\mathbf{X}}_i, \bar{\mathbf{X}}_j) \in \{(\bar{\mathbf{L}}_k, \bar{\mathbf{U}}_k)\}_{k \in \mathcal{K}} \right\}$; and h_t denotes the function $h_t(y) := \min_x \left\{ y^\top \bar{\mathbf{Z}}^t(x) \mid x \in \mathcal{X}^t \right\}$.

COROLLARY EC.3. *Under the same assumption in Proposition 5, the inverse optimization problem (3) can be solved by the risk function (23), and the parameter δ is calculated by solving the convex program*

$$\min_{\delta, y_j, v_{i,j}, w_{i,j}, \gamma} \sum_{t \in \mathcal{T}} \gamma_t$$

$$\begin{aligned}
& \text{subject to } \bar{\mathbf{I}}^\top v_{i,j} + \bar{\mathbf{I}}^\top w_{i,j} \leq \delta_i - \delta_j + y_j^\top \vec{S}_j, \quad \forall j \in \mathcal{J}, \forall i \neq j, \\
& \quad \vec{S}_i y_j^\top - \Lambda_{i,j} \circ (v_{i,j} \bar{\mathbf{I}}^\top) - \bar{\mathbf{I}} w_{i,j}^\top \leq 0, \quad \forall j \in \mathcal{J}, \forall i \neq j, \\
& \quad y_j \in \mathcal{C}_j \subseteq \mathbb{R}_+^{\tau_j}, \quad \forall j \in \mathcal{J}, \\
& \quad y_t^\top \vec{S}_t \leq h_t(y_t) + \gamma_t, \quad \forall t \in \mathcal{T}, \\
& \quad \delta_i \leq \delta_j, \quad \forall (i,j) \in \mathcal{B}, \\
& \quad |\delta_j - \tilde{\rho}(\vec{X}_j)| \leq \epsilon^*, \quad \forall j \in \mathcal{J},
\end{aligned}$$

where $\delta \in \mathbb{R}^{|\mathcal{J}|}$, $y_j \in \mathbb{R}^{\tau_j}$, $v_{i,j} \in \mathbb{R}^{\tau_i}$, $w_{i,j} \in \mathbb{R}^{\tau_j}$, and $\gamma \in \mathbb{R}^{|\mathcal{T}|}$; $\vec{S}_t := (Z(x^t, \xi_1^t), \dots, Z(x^t, \xi_{\tau_0^t}^t))^\top$; the set $\mathcal{B} := \left\{ (i,j) \in \{1, 2, \dots, \mathcal{J}\}^2 \mid (\vec{X}_i, \vec{X}_j) \in \{(\vec{L}_k, \vec{U}_k)\}_{k \in \mathcal{K}} \right\}$; and h_t denotes the function $h_t(y) := \min_x \left\{ \sum_{o=1}^{\tau_0^t} y_o Z(x, \xi_o^t) \mid x \in \mathcal{X}^t \right\}$. The coefficient $\Lambda_{i,j}$ is calculated by $(\Lambda_{i,j})_{m,n} = \bar{p}_n^j / \bar{p}_m^i$, $n = 1, \dots, \tau_j$, $m = 1, \dots, \tau_i$. Moreover, the above sets \mathcal{C}_j , $j \in \mathcal{J}$, can be derived as indicated in Proposition 5.

COROLLARY EC.4. Under the same assumption in Proposition 5, the inverse optimization problem (4) can be solved by the risk function (23), and the parameter δ is calculated by solving the convex program

$$\begin{aligned}
& \max_{\delta, y_j, v_{i,j}, w_{i,j}} \sum_{j \in \mathcal{J}} \delta_j \\
& \text{subject to } \bar{\mathbf{I}}^\top v_{i,j} + \bar{\mathbf{I}}^\top w_{i,j} \leq \delta_i - \delta_j + y_j^\top \vec{S}_j, \quad \forall j \in \mathcal{J}, \forall i \neq j, \\
& \quad \vec{S}_i y_j^\top - \Lambda_{i,j} \circ (v_{i,j} \bar{\mathbf{I}}^\top) - \bar{\mathbf{I}} w_{i,j}^\top \leq 0, \quad \forall j \in \mathcal{J}, \forall i \neq j, \\
& \quad y_j \in \mathcal{C}_j \subseteq \mathbb{R}_+^{\tau_j}, \quad \forall j \in \mathcal{J}, \\
& \quad y_t^\top \vec{S}_t \leq h_t(y_t) + \gamma_t^*, \quad \forall t \in \mathcal{T}, \\
& \quad \delta_i \leq \delta_j, \quad \forall (i,j) \in \mathcal{B}, \\
& \quad |\delta_j - \tilde{\rho}(\vec{X}_j)| \leq \epsilon^*, \quad \forall j \in \mathcal{J},
\end{aligned}$$

where $\delta \in \mathbb{R}^{|\mathcal{J}|}$, $y_j \in \mathbb{R}^{\tau_j}$, $v_{i,j} \in \mathbb{R}^{\tau_i}$, and $w_{i,j} \in \mathbb{R}^{\tau_j}$; $\vec{S}_t := (Z(x^t, \xi_1^t), \dots, Z(x^t, \xi_{\tau_0^t}^t))^\top$; the set $\mathcal{B} := \left\{ (i,j) \in \{1, 2, \dots, \mathcal{J}\}^2 \mid (\vec{X}_i, \vec{X}_j) \in \{(\vec{L}_k, \vec{U}_k)\}_{k \in \mathcal{K}} \right\}$; and h_t denotes the function $h_t(y) := \min_x \left\{ \sum_{o=1}^{\tau_0^t} y_o Z(x, \xi_o^t) \mid x \in \mathcal{X}^t \right\}$. The coefficient $\Lambda_{i,j}$ is calculated by $(\Lambda_{i,j})_{m,n} = \bar{p}_n^j / \bar{p}_m^i$, $n = 1, \dots, \tau_j$, $m = 1, \dots, \tau_i$. Moreover, the above sets \mathcal{C}_j , $j \in \mathcal{J}$, can be derived as indicated in Proposition 5.

EC.4. Proofs

Proof of Proposition 1

Proof of Proposition 1 The proof consists of two parts. In the first part of the proof, we show that for any optimal solution ρ_0 , the domain of its conjugate function must admit the set \mathcal{C} . Then, in the second part we confirm the optimality of ρ_δ where $\delta_j = \rho_0(\vec{X}_j)$, $\forall j \in \mathcal{J}$.

(Part I) Prove that any optimal solution ρ_0 admits the following representation

$$\rho_0(\vec{Z}) = \sup_{p \in \mathcal{C}} p^\top \vec{Z} - \rho_0^*(p), \quad (\text{EC.3})$$

where $\rho_0^*(p) := \sup_{\vec{Z}} p^\top \vec{Z} - \rho_0(\vec{Z})$.

We start by invoking the theory of conjugate duality (Theorem EC.1.1 (1.)), which states that given that $\rho_0 \in \mathcal{R}$ we can always represent ρ_0 by

$$\rho_0(\vec{Z}) = \sup_p p^\top \vec{Z} - \rho_0^*(p). \quad (\text{EC.4})$$

The rest is to show that the conjugate ρ_0^* must satisfy $\rho_0^*(p) = \infty$ for any $p \in \mathbb{R}^{|\Omega|} \setminus \mathcal{C}$. We prove this by contradiction. Suppose that there exists a solution $p_0 \in \mathbb{R}^{|\Omega|} \setminus \mathcal{C}$ such that $\rho_0^*(p_0) < \infty$. The fact that \mathcal{C} is a closed convex set implies that (by hyperplane separation theorem) there must exist a nonzero vector $\vec{R} \in \mathbb{R}^{|\Omega|}$ and $b \in \mathbb{R}$ such that

$$p^\top \vec{R} \leq b, \quad \forall p \in \mathcal{C}, \quad (\text{EC.5})$$

$$p_0^\top \vec{R} > b \quad (\text{or equivalently } p_0^\top \vec{R} = b + \epsilon \text{ for some } \epsilon > 0) \quad (\text{EC.6})$$

hold. Let us consider a random loss $\vec{Z} := \lambda \vec{R}$ for some $\lambda > 0$. Based on the dual representation of the reference risk function $\tilde{\rho}$, we have

$$\tilde{\rho}(\lambda \vec{R}) = \sup_{p \in \mathcal{C}} \lambda p^\top \vec{R} \leq \lambda b \quad (\text{due to (EC.5)}). \quad (\text{EC.7})$$

Based on (EC.4), we also have for the fixed p_0

$$\rho_0(\lambda \vec{R}) \geq p_0^\top (\lambda \vec{R}) - \rho_0^*(p_0) = \lambda b + \lambda \epsilon - \rho_0^*(p_0) \quad (\text{due to (EC.6)}). \quad (\text{EC.8})$$

By subtracting (EC.7) from (EC.8), we arrive at

$$\rho_0(\lambda \vec{R}) - \tilde{\rho}(\lambda \vec{R}) \geq \lambda \epsilon - \rho_0^*(p_0) \rightarrow \infty,$$

as $\lambda \rightarrow \infty$. This contradicts the fact that ρ_0 gives the optimal value $u^* := \|\rho_0 - \tilde{\rho}\|_\infty < \infty$.

(Part II) Verify the optimality of ρ_δ , where $\delta_j = \rho_0(\vec{X}_j)$, $\forall j \in \mathcal{J}$.

To proceed, we need to prove first the following inequalities between ρ_0 and ρ_δ :

$$\begin{aligned}
\rho_0(\vec{Z}) &= \sup_{p \in \mathcal{C}} p^\top \vec{Z} - \rho_0^*(p) \quad (\text{due to (EC.4)}) \\
&= \sup_{p \in \mathcal{C}} p^\top \vec{Z} - \left\{ \sup_{\vec{Z}} p^\top \vec{Z} - \rho_0(\vec{Z}) \right\} \\
&\leq \sup_{p \in \mathcal{C}} p^\top \vec{Z} - \left\{ \max_{j \in \mathcal{J}} \left\{ p^\top \vec{X}_j - \rho_0(\vec{X}_j) \right\} \right\} \\
&= \rho_\delta(\vec{Z}), \quad \forall \vec{Z},
\end{aligned} \tag{EC.9}$$

and

$$\begin{aligned}
\rho_\delta(\vec{X}_i) &= \sup_{p \in \mathcal{C}} p^\top \vec{X}_i - \left\{ \max_{j \in \mathcal{J}} \left\{ p^\top \vec{X}_j - \rho_0(\vec{X}_j) \right\} \right\} \\
&\leq \sup_{p \in \mathcal{C}} p^\top \vec{X}_i - \left\{ p^\top \vec{X}_i - \rho_0(\vec{X}_i) \right\} \\
&= \rho_0(\vec{X}_i), \quad \forall i \in \mathcal{J}.
\end{aligned} \tag{EC.10}$$

The inequalities (EC.9) and (EC.10) imply, first, that $\rho_\delta(\vec{0}) = \rho_0(\vec{0}) = 0$ (i.e., ρ_δ satisfies the normalization condition). One can confirm also that based on the representation theory (Theorem EC.2.1), ρ_δ must be monotonic, convex, and translation invariant by its construction. Hence, $\rho_\delta \in \mathcal{R}$. The inequalities (EC.9) and (EC.10) imply also that ρ_δ satisfies the constraints (7) and (8) because they lead to the following:

$$\rho_\delta(\vec{L}_k) \leq \rho_0(\vec{L}_k) \leq \rho_0(\vec{U}_k) \leq \rho_\delta(\vec{U}_k), \quad \forall k \in \mathcal{K},$$

$$\rho_\delta(\vec{W}^t) \leq \rho_0(\vec{W}^t) \leq \rho_0(\vec{W}) \leq \rho_\delta(\vec{W}), \quad \forall \vec{W} \in \mathcal{W}^t, \quad \forall t \in \mathcal{T}.$$

Finally, we verify that ρ_δ reaches the optimal value of u^* (i.e., $\|\rho_\delta - \tilde{\rho}\| = u^*$). We have

$$\begin{aligned}
\rho_\delta(\vec{Z}) &= \sup_{y \in \mathcal{C}} y^\top \vec{Z} - \max_{j \in \mathcal{J}} \left\{ y^\top \vec{X}_j - \rho_0(\vec{X}_j) \right\} \\
&\leq \sup_{y \in \mathcal{C}} y^\top \vec{Z} - \max_{j \in \mathcal{J}} \left\{ y^\top \vec{X}_j - (\tilde{\rho}(\vec{X}_j) + u^*) \right\} \quad (\text{since } \rho_0(\vec{X}_j) \leq \tilde{\rho}(\vec{X}_j) + u^*, \quad \forall j \in \mathcal{J}) \\
&\leq \sup_{y \in \mathcal{C}} y^\top \vec{Z} - (y^\top \vec{0} - (\tilde{\rho}(\vec{0}) + u^*)) = \tilde{\rho}(\vec{Z}) + u^*, \quad \forall \vec{Z},
\end{aligned} \tag{EC.11}$$

and also have

$$\begin{aligned}
\rho_\delta(\vec{Z}) &= \sup_{y \in \mathcal{C}} y^\top \vec{Z} - \max_{j \in \mathcal{J}} \left\{ y^\top \vec{X}_j - \rho_0(\vec{X}_j) \right\} \\
&\geq \sup_{y \in \mathcal{C}} y^\top \vec{Z} - \max_{j \in \mathcal{J}} \left\{ y^\top \vec{X}_j - (\tilde{\rho}(\vec{X}_j) - u^*) \right\} \quad (\text{since } \rho_0(\vec{X}_j) \geq \tilde{\rho}(\vec{X}_j) - u^*, \quad \forall j \in \mathcal{J}) \\
&= \sup_{y \in \mathcal{C}} y^\top \vec{Z} - u^* - \max_{j \in \mathcal{J}} \left\{ y^\top \vec{X}_j - \tilde{\rho}(\vec{X}_j) \right\} \\
&\geq \sup_{y \in \mathcal{C}} y^\top \vec{Z} - u^* = \tilde{\rho}(\vec{Z}) - u^*, \quad \forall \vec{Z},
\end{aligned} \tag{EC.12}$$

where the second inequality follows the dual representation of $\tilde{\rho}$; namely for any $j \in \mathcal{J}$,

$$\tilde{\rho}(\vec{X}_j) = \sup_{y \in \mathcal{C}} y^\top \vec{X}_j \Leftrightarrow y^\top \vec{X}_j - \tilde{\rho}(\vec{X}_j) \leq 0, \forall y \in \mathcal{C}.$$

□

Proof of Proposition 2

Proof of Proposition 2 Because $\rho_\delta \in \mathcal{R}$, we know from the theory of conjugate duality (Theorem EC.1.1 (1.)) that

$$\rho_\delta(\vec{Z}) = \rho_\delta^{**}(\vec{Z}), \forall \vec{Z}$$

must hold. Obviously, the condition can be equivalently stated in terms of the following two inequalities:

$$\begin{aligned} \rho_\delta(\vec{Z}) &\geq \rho_\delta^{**}(\vec{Z}), \forall \vec{Z}, \\ \rho_\delta(\vec{Z}) &\leq \rho_\delta^{**}(\vec{Z}), \forall \vec{Z}. \end{aligned}$$

The first inequality, however, is always satisfied given that it is implied by the construction of bi-conjugate function (see Theorem EC.1.1). Therefore, it suffices to proceed with the second inequality.

Consider the above second inequality with $\vec{Z} = \vec{Z}_j$, $j \in \bar{\mathcal{J}}$. By expanding the biconjugate ρ_δ^{**} based on its definition and using the conjugate function ρ_δ^* derived in Lemma EC.1.1, we have the following:

$$\begin{aligned} \rho_\delta^{**}(\vec{Z}_j) &\geq \rho_\delta(\vec{Z}_j), \forall j \in \bar{\mathcal{J}} \\ \Leftrightarrow \sup_y \left\{ y^\top \vec{Z}_j - \rho_\delta^*(y) \right\} &\geq \rho_\delta(\vec{Z}_j), \forall j \in \bar{\mathcal{J}} \\ \Leftrightarrow \sup_{y \in \bar{\mathcal{C}}} \left\{ y^\top \vec{Z}_j - \max_{i \in \bar{\mathcal{J}}} \left\{ y^\top \vec{Z}_i - \rho_\delta(\vec{Z}_i) \right\} \right\} &\geq \rho_\delta(\vec{Z}_j), \forall j \in \bar{\mathcal{J}} \\ \Leftrightarrow \exists y_j \in \bar{\mathcal{C}}: y_j^\top \vec{Z}_j - \max_{i \in \bar{\mathcal{J}}} \left\{ y_j^\top \vec{Z}_i - \rho_\delta(\vec{Z}_i) \right\} &\geq \rho_\delta(\vec{Z}_j), \forall j \in \bar{\mathcal{J}} \\ \Leftrightarrow \exists y_j \in \bar{\mathcal{C}}: y_j^\top \vec{Z}_j - y_j^\top \vec{Z}_i + \rho_\delta(\vec{Z}_i) &\geq \rho_\delta(\vec{Z}_j), \forall i \neq j, \end{aligned}$$

where in the fourth line we apply the fact that $\bar{\mathcal{C}}$ is closed and bounded (i.e., compact). This completes the first part of the proof.

To prove the other direction, note that for any feasible solution $\{\delta_j^*\}_{j \in \bar{\mathcal{J}}}$ and $\{y_j^*\}_{j \in \bar{\mathcal{J}}}$ we can always construct a $(\{\vec{Z}_j\}_{j \in \bar{\mathcal{J}}}, \bar{\mathcal{C}})$ -supported risk function (12). It always satisfies $\rho_\delta(\vec{Z}_j) \leq \delta_j^*$ because

$$\rho_\delta(\vec{Z}_j) \leq \sup_{y \in \bar{\mathcal{C}}} y^\top \vec{Z}_j - y^\top \vec{Z}_j + \delta_j^* = \delta_j^*,$$

and satisfies $\rho_\delta(\vec{Z}_j) \geq \delta_j^*$ because

$$\rho_\delta(\vec{Z}_j) \geq y_j^{*\top} \vec{Z}_j - \max_{i \in \bar{\mathcal{J}}} \left\{ y_j^{*\top} \vec{Z}_i - \delta_i^* \right\} \geq \delta_j^* \text{ (due to the feasibility of } \{\delta_j^*\}_{j \in \bar{\mathcal{J}}} \text{ and } \{y_j^*\}_{j \in \bar{\mathcal{J}}}).$$

This proves the existence of a risk function $\rho_\delta \in \mathcal{L}(\{\vec{Z}_j\}_{j \in \bar{\mathcal{J}}}, \bar{\mathcal{C}})$ that satisfies $\rho_\delta(\vec{Z}_j) = \delta_j^*$. □

Proof of Proposition 3

Proof of Proposition 3 We seek a function $\rho_\delta \in \mathcal{L}(\{\vec{X}_j\}_{j \in \mathcal{J}}, \mathcal{C})$ that further satisfies the system below:

$$x^t \in \arg \min_{x \in \mathcal{X}^t} \rho_\delta(\vec{Z}^t(x)), \forall t \in \mathcal{T}, \quad (\text{EC.13})$$

$$\rho_\delta(\vec{L}_k) \leq \rho_\delta(\vec{U}_k), \forall k \in \mathcal{K}. \quad (\text{EC.14})$$

Due to the monotonicity of $\rho_\delta \in \mathcal{R}$, we can equivalently write the optimization problem in (EC.13) as

$$\min_{(x, \vec{W}) \in \Pi^t} \rho_\delta(\vec{W}), \text{ where } \Pi^t := \left\{ (x, \vec{W}) \mid \vec{W} \geq \vec{Z}^t(x), x \in \mathcal{X}^t \right\}.$$

Because Π^t is convex and $\vec{W}^t \in \{\vec{X}_j\}_{j \in \mathcal{J}}$, we can characterize the optimality condition for the problem; namely, there must exist a subgradient $y \in \partial \rho_\delta(\vec{W}^t)$ such that

$$y^\top (\vec{W} - \vec{W}^t) \geq 0, \forall (x, \vec{W}) \in \Pi^t \Leftrightarrow y^\top \vec{W}^t \leq \min_{x \in \mathcal{X}^t} y^\top \vec{Z}^t(x).$$

Moreover, applying the theory of conjugate duality (Theorem EC.1.1) (2.), which states that

$$\partial \rho_\delta(\vec{W}^t) = \partial \rho_\delta^{**}(\vec{W}^t) = \arg \max_y \left\{ y^\top \vec{W}^t - \rho_\delta^*(y) \right\},$$

we can further characterize the set of subgradients $\partial \rho_\delta(\vec{W}^t)$ in terms of inequality constraints, given that

$$\begin{aligned} & y \in \partial \rho_\delta^{**}(\vec{W}^t) \\ & \Leftrightarrow \{y : y^\top \vec{W}^t - \rho_\delta^*(y) \geq \rho_\delta(\vec{W}^t)\} \\ & \Leftrightarrow \{y : y^\top \vec{W}^t - \max_{j \in \mathcal{J}} \{y^\top \vec{X}_j - \rho_\delta(\vec{X}_j)\} \geq \rho_\delta(\vec{W}^t), y \in \mathcal{C}\} \\ & \Leftrightarrow \{y : y^\top \vec{W}^t - y^\top \vec{X}_j + \rho_\delta(\vec{X}_j) \geq \rho_\delta(\vec{W}^t), \forall j \in \mathcal{J}, y \in \mathcal{C}\}, \end{aligned} \quad (\text{EC.15})$$

where in the third line the conjugate $\rho_\delta^*(y)$ derived in Lemma EC.1.1 is applied.

Finally, applying Proposition 2, we can equivalently describe the set of risk functions ρ_δ by the system of constraints (11). Thus by replacing $\rho_\delta(\vec{X}_j)$ with δ_j in (EC.14) and (EC.15), we arrive at the final formulation. \square

Proof of Theorem 2

Proof of Theorem 2 To complete the proof, we need only to show that given that $\rho_\delta \in \mathcal{L}(\{\vec{X}_j\}_{j \in \mathcal{J}}, \mathcal{C})$ the objective function $\|\rho_\delta - \tilde{\rho}\|_\infty$ can be reduced to

$$\max_{j \in \mathcal{J}} \left\{ |\rho_\delta(\vec{X}_j) - \tilde{\rho}(\vec{X}_j)| \right\}. \quad (\text{EC.16})$$

To confirm this, let $u^* = (\text{EC.16})$. Obviously by the definition of u^* , we can also write

$$-u^* \leq \rho_\delta(\vec{X}_j) - \tilde{\rho}(\vec{X}_j) \leq u^*, \forall j \in \mathcal{J}.$$

The direction $\|\rho_\delta - \tilde{\rho}\|_\infty \geq u^*$ is clear. To prove the other direction (i.e., $\|\rho_\delta - \tilde{\rho}\|_\infty \leq u^*$), note first that because $\rho_\delta \in \mathcal{R}$, we can apply the theory of conjugate duality (Theorem EC.1.1 (1.)) and use the conjugate ρ_δ^* derived in Lemma EC.1.1 to write ρ_δ as

$$\rho_\delta(\vec{Z}) = \rho_\delta^{**}(\vec{Z}) = \sup_{y \in \mathcal{C}} y^\top \vec{Z} - \max_{j \in \mathcal{J}} \left\{ y^\top \vec{X}_j - \rho_\delta(\vec{X}_j) \right\}.$$

Based on this representation, we can then derive

$$\begin{aligned} \rho_\delta(\vec{Z}) &= \sup_{y \in \mathcal{C}} y^\top \vec{Z} - \max_{j \in \mathcal{J}} \left\{ y^\top \vec{X}_j - \rho_\delta(\vec{X}_j) \right\} \\ &\leq \sup_{y \in \mathcal{C}} y^\top \vec{Z} - \max_{j \in \mathcal{J}} \left\{ y^\top \vec{X}_j - (\tilde{\rho}(\vec{X}_j) + u^*) \right\} \quad (\text{since } \rho_\delta(\vec{X}_j) \leq \tilde{\rho}(\vec{X}_j) + u^*, \forall j \in \mathcal{J}) \\ &\leq \sup_{y \in \mathcal{C}} y^\top \vec{Z} - (y^\top \vec{0} - (\tilde{\rho}(\vec{0}) + u^*)) = \tilde{\rho}(\vec{Z}) + u^*, \forall \vec{Z}, \end{aligned} \tag{EC.17}$$

and also have

$$\begin{aligned} \rho_\delta(\vec{Z}) &= \sup_{y \in \mathcal{C}} y^\top \vec{Z} - \max_{j \in \mathcal{J}} \left\{ y^\top \vec{X}_j - \rho_\delta(\vec{X}_j) \right\} \\ &\geq \sup_{y \in \mathcal{C}} y^\top \vec{Z} - \max_{j \in \mathcal{J}} \left\{ y^\top \vec{X}_j - (\tilde{\rho}(\vec{X}_j) - u^*) \right\} \quad (\text{since } \rho_\delta(\vec{X}_j) \geq \tilde{\rho}(\vec{X}_j) - u^*, \forall j \in \mathcal{J}) \\ &= \sup_{y \in \mathcal{C}} y^\top \vec{Z} - u^* - \max_{j \in \mathcal{J}} \left\{ y^\top \vec{X}_j - \tilde{\rho}(\vec{X}_j) \right\} \\ &\geq \sup_{y \in \mathcal{C}} y^\top \vec{Z} - u^* = \tilde{\rho}(\vec{Z}) - u^*, \forall \vec{Z}, \end{aligned} \tag{EC.18}$$

where the second inequality follows the supremum representation of $\tilde{\rho}$; namely for any $j \in \mathcal{J}$,

$$\tilde{\rho}(\vec{X}_j) = \sup_{y \in \mathcal{C}} y^\top \vec{X}_j \Leftrightarrow y^\top \vec{X}_j - \tilde{\rho}(\vec{X}_j) \leq 0, \forall y \in \mathcal{C}.$$

We thus conclude that to determine the parameter δ , it suffices to solve the problem (17). Since the problem is a convex program, we can apply the famous result of Grötschel et al. (1981), which states that for a convex program like (17), it can be solved in polynomial time by using the ellipsoid method if and only if for any $z^* := (\delta^*, y_j^*)$ it takes polynomial time to either confirm that z^* is in the feasible set \mathcal{Z} or generate a hyperplane that separates z^* from \mathcal{Z} . Hence, if the function $h_t(y_t)$ can be evaluated in polynomial time (i.e., the forward problem can be solved in polynomial time), and the oracle for the set \mathcal{C} exists, it can be shown fairly straightforwardly that it also takes polynomial time to confirm $z^* \in \mathcal{Z}$ or separate z^* from \mathcal{Z} . This completes the proof. \square

Proof of Corollary 1

Proof of Corollary 1 By introducing a dummy variable s that bounds from above $\max_{j \in \mathcal{J}} \{p^\top \vec{X}_j - \delta_j\}$, we have the first formulation. The second formulation with the linear constraints is the dual of the first formulation that can be obtained by applying linear duality theory. Note that strong duality always holds here because, obviously, there always exists a feasible solution to the first formulation (e.g., by setting $s^* = \max_{j \in \mathcal{J}} \{p^{*\top} \vec{X}_j - \delta_j\}$ for any feasible $p^* \in \mathcal{C}$).

□

Proof of Proposition 4 As mentioned in Section 3.2, in the case of permutation invariance one must deal with the issue that the optimality conditions are defined over non-convex sets $\{\sigma(\vec{W}) \mid \sigma \in \Sigma, \vec{W} \in \mathcal{W}^t\}$ and the issue that the size of $\{\sigma(\vec{X}_j)\}_{\sigma \in \Sigma, j \in \mathcal{J}}$ grows exponentially.

We first show how the inverse problem (6)–(8) with $\mathcal{R} := \overline{\mathcal{R}}$ can still be formulated as a convex program by following closely the steps established in Section 3.1. This bypasses the issue of non-convexity. We summarize this first step in the following proposition. After proving this proposition, we then continue to prove how to reduce the size of the problem presented in the proposition from exponentially many to polynomially many.

PROPOSITION EC.1. *Given that Assumption 2 holds and that the set of optimal solutions is non-empty, the inverse problem (6)–(8) with $\mathcal{R} := \overline{\mathcal{R}}$ can be solved a risk function $\overline{\rho}_\delta \in \overline{\mathcal{L}}(\{\vec{X}_j\}_{j \in \mathcal{J}}, \overline{\mathcal{C}})$, and the parameter δ is calculated by solving*

$$\begin{aligned}
 & \min_{u, \delta, y_{\sigma, j}} \quad u \\
 & \text{subject to} \quad -u \leq \delta_j - \tilde{\rho}(\vec{X}_j) \leq u, \quad \forall j \in \mathcal{J}, \\
 & \quad \delta_j + y_{\sigma, j}^\top (\sigma'(\vec{X}_i) - \sigma(\vec{X}_j)) \leq \delta_i, \quad \forall (\sigma, j) \in \Sigma \times \mathcal{J}, \forall i \neq j, \forall \sigma' \in \Sigma, \\
 & \quad y_{\sigma, j} \in \overline{\mathcal{C}}, \quad \forall (\sigma, j) \in \Sigma \times \mathcal{J}, \\
 & \quad y_{\sigma^*, t}^\top \vec{X}_t \leq h_t(y_{\sigma^*, t}), \quad \forall t \in \mathcal{T}, \\
 & \quad \delta_i \leq \delta_j, \quad \forall (i, j) \in \mathcal{B},
 \end{aligned} \tag{EC.19}$$

where σ^* in $y_{\sigma^*, t}$ corresponds to the permutation such that $\sigma^*(\vec{X}) = \vec{X}$; $u \in \mathbb{R}$, $\delta \in \mathbb{R}^{|\mathcal{J}|}$, $y_{\sigma, j} \in \mathbb{R}^M$; the set $\mathcal{B} := \{(i, j) \in \{1, 2, \dots, \mathcal{J}\}^2 \mid (\vec{X}_i, \vec{X}_j) \in \{(\vec{L}_k, \vec{U}_k)\}_{k \in \mathcal{K}}\}$; and h_t denotes the function $h_t(y) := \min_x \{y^\top \vec{Z}^t(x) \mid x \in \mathcal{X}^t\}$.

Proof of Proposition EC.1 (Step 1) Like Proposition 1, we claim that the set $\overline{\mathcal{L}}(\{\vec{X}_j\}_{j \in \mathcal{J}}, \overline{\mathcal{C}})$ contains an optimal solution to the problem. We prove this by showing that if there exists a permutation-invariant risk function $\overline{\rho}_0$ that is optimal for the problem with the optimal value

$u^* < \infty$, there must exist a risk function $\bar{\rho}_\delta \in \bar{\mathcal{L}}(\{\vec{X}_j\}_{j \in \mathcal{J}}, \bar{\mathcal{C}})$ that is also optimal to the problem, namely by setting

$$\delta_j = \bar{\rho}_0(\vec{X}_j).$$

This can be proved by following exactly the steps in Proposition 1. Namely, by letting $\{\vec{X}_j\}_{j \in \mathcal{J}}$ in $\mathcal{L}(\{\vec{X}_j\}_{j \in \mathcal{J}}, \mathcal{C})$ of Proposition 1 now represent all random losses in $\{\sigma(\vec{X}_j)\}_{\sigma \in \Sigma, j \in \mathcal{J}}$ and letting \mathcal{C} now take the form of $\bar{\mathcal{C}}$, we can conclude based on the result in Proposition 1 that the following function $\rho_\delta \in \mathcal{L}(\{\sigma(\vec{X}_j)\}_{\sigma \in \Sigma, j \in \mathcal{J}}, \bar{\mathcal{C}})$ must also be optimal:

$$\rho_\delta(\vec{Z}) := \sup_{p \in \bar{\mathcal{C}}} p^\top \vec{Z} - \max_{\sigma \in \Sigma, j \in \mathcal{J}} \left\{ p^\top \sigma(\vec{X}_j) - \bar{\rho}_0(\sigma(\vec{X}_j)) \right\}.$$

Note that due to the fact that $\bar{\rho}_0$ is permutation invariant, the above function can be simplified to

$$\rho_\delta(\vec{Z}) = \sup_{p \in \bar{\mathcal{C}}} p^\top \vec{Z} - \max_{\sigma \in \Sigma, j \in \mathcal{J}} \left\{ p^\top \sigma(\vec{X}_j) - \bar{\rho}_0(\vec{X}_j) \right\} = \bar{\rho}_\delta(\vec{Z}).$$

This completes the first step of the proof.

(Step 2) Following Proposition 2, we can boil down the problem of search over the set $\bar{\mathcal{L}}(\{\vec{X}_j\}_{j \in \mathcal{J}}, \bar{\mathcal{C}})$ into identifying feasible solutions $y_{\sigma,j}$ and δ_j in the following systems:

$$\begin{aligned} \delta_j + y_{\sigma,j}^\top (\sigma'(\vec{X}_i) - \sigma(\vec{X}_j)) &\leq \delta_i, \forall i, j \in \mathcal{J}, \forall \sigma, \sigma' \in \Sigma, \\ y_{\sigma,j} &\in \bar{\mathcal{C}}, \forall j \in \mathcal{J}, \forall \sigma \in \Sigma. \end{aligned} \tag{EC.20}$$

(Step 3) We will follow closely the steps in Proposition 3. However, one can see that because we seek only a function $\bar{\rho}_\delta \in \bar{\mathcal{L}}(\{\vec{X}_j\}_{j \in \mathcal{J}}, \bar{\mathcal{C}})$, which is permutation invariant, we can reduce the constraints

$$x^t \in \arg \min_{x \in \mathcal{X}^t} \bar{\rho}_\delta(\sigma(\vec{Z}^t(x))), \forall t \in \mathcal{T}, \forall \sigma \in \Sigma, \tag{EC.21}$$

$$\bar{\rho}_\delta(\sigma'(\vec{L}_k)) \leq \bar{\rho}_\delta(\sigma(\vec{U}_k)), \forall k \in \mathcal{K}, \forall \sigma', \sigma \in \Sigma \tag{EC.22}$$

into

$$x^t \in \arg \min_{x \in \mathcal{X}^t} \bar{\rho}_\delta(\vec{Z}^t(x)), \forall t \in \mathcal{T}, \tag{EC.23}$$

$$\bar{\rho}_\delta(\vec{L}_k) \leq \bar{\rho}_\delta(\vec{U}_k), \forall k \in \mathcal{K}. \tag{EC.24}$$

We will follow the same arguments used in Proposition 3 to further reduce the constraint (EC.23). Note that although it looks quite repetitive, one should be careful to derive it based on structure of $\bar{\rho}_\delta$. First, the monotonicity of $\bar{\rho}_\delta$ allows us to equivalently write the constraint (EC.23) as

$$\min_{(x, \vec{W}) \in \Pi^t} \bar{\rho}_\delta(\vec{W}), \text{ where } \Pi^t := \left\{ (x, \vec{W}) \mid \vec{W} \geq \vec{Z}^t(x), x \in \mathcal{X}^t \right\}.$$

Because Π^t here is convex and $\vec{W}^t \in \{\vec{X}_j\}_{j \in \mathcal{J}}$, we can now write down the optimality condition that there must exist a subgradient $y \in \partial \bar{\rho}_\delta(\vec{W}^t)$ such that

$$y^\top (\vec{W} - \vec{W}^t) \geq 0, \forall (x, \vec{W}) \in \Pi^t \Leftrightarrow y^\top \vec{W}^t \leq \min_{x \in \mathcal{X}^t} y^\top \vec{Z}^t(x).$$

Because $\bar{\rho}_\delta$ is obviously convex, we again apply the theory of conjugate duality (Theorem [EC.1.1](#)) (2.), which states that

$$\partial \bar{\rho}_\delta(\vec{W}^t) = \partial \bar{\rho}_\delta^{**}(\vec{W}^t) = \arg \max_y \left\{ y^\top \vec{W}^t - \bar{\rho}_\delta^*(y) \right\}.$$

We now further characterize the set of subgradients $\partial \bar{\rho}_\delta(\vec{W}^t)$ by

$$\begin{aligned} & y \in \partial \bar{\rho}_\delta^{**}(\vec{W}^t) \\ \Leftrightarrow & \{y : y^\top \vec{W}^t - \bar{\rho}_\delta^*(y) \geq \bar{\rho}_\delta(\vec{W}^t)\} \\ \Leftrightarrow & \{y : y^\top \vec{W}^t - \max_{\sigma \in \Sigma, j \in \mathcal{J}} \left\{ y^\top \sigma(\vec{X}_j) - \bar{\rho}_\delta(\sigma(\vec{X}_j)) \right\} \geq \bar{\rho}_\delta(\vec{W}^t), y \in \bar{\mathcal{C}}\} \\ \Leftrightarrow & \{y : y^\top \vec{W}^t - y^\top \sigma(\vec{X}_j) + \bar{\rho}_\delta(\sigma(\vec{X}_j)) \geq \bar{\rho}_\delta(\vec{W}^t), \forall j \in \mathcal{J}, \forall \sigma \in \Sigma, \forall y \in \bar{\mathcal{C}}\}, \end{aligned} \quad (\text{EC.25})$$

where in the third line the conjugate $\bar{\rho}_\delta^*(y)$ derived in Lemma [EC.1.1](#) is applied.

(Step 4) Now, following the proof of Theorem [2](#), it is straightforward to see that the objective function in this case can be reduced to

$$\max_{\sigma \in \Sigma, j \in \mathcal{J}} \left\{ |\bar{\rho}_\delta(\sigma(\vec{X}_j)) - \tilde{\rho}(\vec{X}_j)| \right\}. \quad (\text{EC.26})$$

Moreover, in applying Step 2 we can equivalently describe the set of risk functions $\bar{\rho}_\delta$ by the system of constraints [\(EC.20\)](#). Thus by replacing $\bar{\rho}_\delta(\sigma(\vec{X}_j))$ with δ_j in [\(EC.24\)](#), [\(EC.25\)](#) and [\(EC.26\)](#), we arrive at the final formulation. \square

(Continue the proof of Proposition 4) Now, we show how to reduce the optimization problems presented Proposition [EC.1](#) (including the problem in the definition of $\bar{\mathcal{L}}(\{\vec{X}_j\}_{j \in \mathcal{J}}, \bar{\mathcal{C}})$). We consider first the reduction of the problem [\(EC.19\)](#). Note first that the constraints associated with $y_{\sigma^*, t}$ can be equivalently written as, with $y_{\sigma^*, t}$ replaced by y_t ,

$$\begin{aligned} & y_t^\top \vec{X}_t \leq h_t(y_t), \forall t \in \mathcal{T}, \\ & \delta_t + y_t^\top (\sigma(\vec{X}_i) - \vec{X}_t) \leq \delta_i, \forall t \in \mathcal{T}, \forall i \in \mathcal{J} \setminus \{t\}, \forall \sigma \in \Sigma, \\ & y_t \in \bar{\mathcal{C}}, \forall t \in \mathcal{T}. \end{aligned} \quad (\text{EC.27})$$

We show that the other constraints, namely the second and the third constraint in (EC.19) in general can also be reduced to

$$\begin{aligned} \delta_j + y_j^\top (\sigma'(\vec{X}_i) - \vec{X}_j) &\leq \delta_i, \forall j \in \mathcal{J}, \forall i \neq j, \forall \sigma' \in \Sigma, \\ y_j &\in \bar{\mathcal{C}}, \forall j \in \mathcal{J}. \end{aligned} \quad (\text{EC.28})$$

We prove this by showing that given any feasible solution $(u^*, \delta^*, y_{\sigma,j})$ of the problem (EC.19) we can always construct a feasible solution $(u^*, \delta^*, \bar{y}_{\sigma'',j})$ with $\bar{y}_{\sigma'',j}$ satisfying

$$\bar{y}_{\sigma'',j} := \frac{1}{|\Sigma|} \sigma'' \left(\sum_{\sigma \in \Sigma} \sigma^{-1}(y_{\sigma,j}) \right), \forall \sigma'' \in \Sigma,$$

which gives the same optimal value u^* .

To verify its feasibility for the second constraint of (EC.19), by substitution we have

$$\bar{y}_{\sigma'',j}^\top (\sigma'(\vec{X}_i) - \sigma''(\vec{X}_j)) = \frac{1}{|\Sigma|} \left(\sum_{\sigma \in \Sigma} y_{\sigma,j}^\top (\sigma(\sigma''^{-1}(\sigma'(\vec{X}_i))) - \vec{X}_j) \right) \leq \frac{1}{|\Sigma|} \left(\sum_{\sigma \in \Sigma} (\delta_i - \delta_j) \right) = \delta_i - \delta_j,$$

where the last inequality is due to the feasibility of $y_{\sigma,j}$. For the third constraint of (EC.19), feasibility can be verified as follows. Since $y_{\sigma,j} \in \bar{\mathcal{C}}$, we have $\sigma^{-1}(y_{\sigma,j}) \in \bar{\mathcal{C}}$. We also have $\sum_{\sigma \in \Sigma} \frac{1}{|\Sigma|} \sigma^{-1}(y_{\sigma,j}) \in \bar{\mathcal{C}}$ because the summation is a convex combination and the set $\bar{\mathcal{C}}$ is convex. Given this, we also have $\bar{y}_{\sigma'',j} \in \bar{\mathcal{C}}$ by the definition of $\bar{\mathcal{C}}$. Hence, we can replace $y_{\sigma,j}$ by $\sigma(y_j)$ for some $y_j \in \mathbb{R}^M$ in the second and third constraints in (EC.19) and arrive at the reduction (EC.28).

Now, both the constraint (EC.27) and (EC.28) can be re-arranged into the following general form

$$y_j^\top \sigma(\vec{X}_i) \leq \delta_i - \delta_j + y_j^\top \vec{X}_j, \forall \sigma \in \Sigma, \forall j \in \mathcal{J}. \quad (\text{EC.29})$$

We show in general how the constraint in the form of $y_j^\top \sigma(\vec{X}) \leq b$, $\forall \sigma \in \Sigma$ for some \vec{X} and b can be reduced, which can then be applied to reduce (EC.27) and (EC.28). Recall first that a permutation matrix Q_σ is a matrix that satisfies $\sigma(\vec{X}) = Q_\sigma \vec{X}$ and $(Q_\sigma)_{m,n} \in \{0, 1\}$ and $Q_\sigma^\top \vec{1} = \vec{1}$, $Q_\sigma \vec{1} = \vec{1}$. Hence, $y_j^\top \sigma(\vec{X}) \leq b$, $\forall \sigma \in \Sigma$ can be re-written as $\max_{\sigma \in \Sigma} y_j^\top Q_\sigma \vec{X} \leq b$ and also as

$$\max_{Q \in \text{Conv}(\{Q_\sigma\}_{\sigma \in \Sigma})} y_j^\top Q \vec{X} \leq b.$$

Applying the result of Birkhoff (1946), we can reformulate the convex hull of all permutation matrices into linear constraints and arrive at the following formulation:

$$\max_Q \left\{ y_j^\top Q \vec{X} \mid Q^\top \vec{1} = \vec{1}, Q \vec{1} = \vec{1}, Q \in \mathbb{R}_+^{M \times M} \right\} \leq b.$$

By deriving the dual problem of the above linear program, we have

$$\begin{aligned} \min_{v,w} \left\{ \vec{1}^\top v + \vec{1}^\top w \mid \vec{X} y^\top - v \vec{1}^\top - \vec{1} w^\top \leq 0 \right\} &\leq b \\ \Leftrightarrow \exists v, w : \vec{1}^\top v + \vec{1}^\top w \leq b, \vec{X} y^\top - v \vec{1}^\top - \vec{1} w^\top \leq 0. \end{aligned}$$

Strong duality holds for the above linear programs because there always exists a permutation matrix satisfying the above constraints. We apply this dualization procedure to (EC.29), which leads to the final formulation of the problem (20).

Finally, we consider the reduction of the problem in the definition of $\bar{\mathcal{L}}(\{\vec{X}_j\}_{j \in \mathcal{J}}, \bar{\mathcal{C}})$. The optimization problem can be equivalently formulated as

$$\begin{aligned} & \sup_{p \in \bar{\mathcal{C}}, t} p^\top \vec{Z} - t \\ & \text{subject to } p^\top \sigma(\vec{X}_j) \leq t + \delta_j, \quad \forall \sigma \in \Sigma, \forall j \in \mathcal{J}. \end{aligned} \quad (\text{EC.30})$$

We can apply the same dualization procedure above to reduce again the constraint (EC.30), which leads to the final formulation. \square

Proof of Proposition 5

Proof of Proposition 5 Given that Assumption 5 holds, we can always convert the probability values specified in the distribution of $Z(x, \xi^t)$ (that satisfies Assumption 3) and the set of distributions $\{F_j\}_{j \in \mathcal{J}}$ to ratios in the form of n/M for some fixed $M \in \mathbb{Z}^+$ and $n \in \{1, \dots, M\}$. By considering an outcome space with M uniformly distributed outcomes, we can equivalently define the random loss $\vec{Z}(x)$ as a mapping from $\Omega := \{\omega_i\}_{i=1}^M$ to \mathbb{R} that satisfies $Z(x, \xi(\omega_i)) \in \{Z(x, \xi_o)\}_{o=1}^{\tau_o}$ and $|\{\omega_i \mid Z(x, \xi(\omega_i)) = Z(x, \xi_o)\}| = \bar{p}_o^\xi M$, and similarly \vec{X}_j as a mapping $X_j : \Omega \rightarrow \mathbb{R}$ that satisfies $X_j(\omega_i) \in \{(\vec{S}_j)_o\}_{o=1}^{\tau_j}$ and $|\{\omega_i \mid X_j(\omega_i) = (\vec{S}_j)_o\}| = \bar{p}_o^j M$.

Suppose now that the optimization problems (20) and (22) are formulated based on the above definition of random losses. We show in what follows how the problems can be further reduced. Note first that by replacing the objective function $p^\top \vec{Z} - t$ in (22) with a new variable s , we can reformulate its first constraint into

$$\bar{\mathbf{I}}^\top v_j + \bar{\mathbf{I}}^\top w_j - p^\top \vec{Z} \leq \delta_j - s, \quad \forall j \in \mathcal{J}.$$

We can then write all the constraints on the variable p in (22) by

$$p \in \mathcal{G}(\vec{Z}, s, \{\delta_i\}_{i \in \mathcal{J}}), \quad (\text{EC.31})$$

where $\mathcal{G}(\vec{Y}, t, \{\delta_i\}_{i \in \mathcal{J}})$ is a parameterized set represented by the following system of constraints on y : $\exists v_i, w_i$ such that

$$\bar{\mathbf{I}}^\top v_i + \bar{\mathbf{I}}^\top w_i - y^\top \vec{Y} \leq \delta_i - t, \quad \forall i \in \mathcal{J}, \quad (\text{EC.32})$$

$$\vec{X}_i y^\top - v_i \bar{\mathbf{I}}^\top - \bar{\mathbf{I}} w_i^\top \leq 0, \quad \forall i \in \mathcal{J}, \quad (\text{EC.33})$$

$$y \in \bar{\mathcal{C}}. \quad (\text{EC.34})$$

Given any fixed $\{\delta_j^*\}_{j \in \mathcal{J}}$, the constraints in the optimization problem (20) can also be equivalently written as

$$y_t \in \mathcal{G}(\vec{X}_t, \delta_t^*, \{\delta_i^*\}_{i \in \mathcal{J}}) \cap \left\{ y \mid y^\top \vec{X}_t \leq h_t(y) \right\}, \forall t \in \mathcal{T}, \quad (\text{EC.35})$$

$$y_j \in \mathcal{G}(\vec{X}_j, \delta_j^*, \{\delta_i^*\}_{i \in \mathcal{J}}), \forall j \in \mathcal{J} \setminus \mathcal{T}. \quad (\text{EC.36})$$

We present only the reduction of the constraints (EC.35) with (EC.32)–(EC.34), given that the same steps can be applied to reduce the constraints (EC.36) (with (EC.32)–(EC.34)) and (EC.31) (with (EC.32)–(EC.34)).

Because it suffices to consider (EC.35) for any fixed t , from here on we consider only $t = 1$ and drop the index t for the variables to simplify the presentation. Given a fixed set of $\{\delta_j^*\}_{j \in \mathcal{J}}$, let y^*, v_i^*, w_i^* denote a feasible solution that satisfies (EC.35) and (EC.32)–(EC.34). For $o = 1, \dots, \tau_0$, let $\mathcal{I}_o^{(1)}$ denote the set of indices n of \vec{X}_1 such that $(\vec{X}_1)_n = (\vec{S}_1)_o$, and therefore $|\mathcal{I}_o^{(1)}| = \bar{p}_o^{(1)} M$. We claim that the solution v_i^* together with the following $y^{**} \in \mathbb{R}^M$, $w_i^{**} \in \mathbb{R}^M$ that satisfies for any $n \in \mathcal{I}_o^{(1)}$,

$$(y^{**})_n = \frac{1}{|\mathcal{I}_o^{(1)}|} \sum_{a \in \mathcal{I}_o^{(1)}} (y^*)_a, \quad (w_i^{**})_n = \frac{1}{|\mathcal{I}_o^{(1)}|} \sum_{a \in \mathcal{I}_o^{(1)}} (w_i^*)_a,$$

$o = 1, \dots, \tau_0$, will also satisfy (EC.35) and (EC.32)–(EC.34).

The constraint (EC.35) and (EC.32)–(EC.33) can be verified fairly straightforwardly by the direct substitution. To verify the third constraint (EC.34), we construct a sequence of solutions $y_*^{(1)}, \dots, y_*^{(\tau_0)}$ that satisfies $y_*^{(\tau_0)} = y^{**}$ and that $y_*^{(\tau_0)} \in \bar{\mathcal{C}}$. For $o = 1, \dots, \tau_0$, let Σ_o denote the set of all permutation operators σ that satisfy $(\sigma(y^*))_a = (y^*)_a$, $\forall a \notin \mathcal{I}_o^{(1)}$. In other words, the set consists of all permutations that permute only the entries $a \in \mathcal{I}_o^{(1)}$. Set $o = 1$ and we construct $y_*^{(1)}$ by $y_*^{(1)} := \sum_{\sigma \in \Sigma_1} \frac{1}{|\mathcal{I}_1^{(1)}|} \sigma(y^*)$. One can confirm that $y_*^{(1)}$ satisfies

$$(y_*^{(1)})_{\tilde{n}} = \frac{1}{|\mathcal{I}_1^{(1)}|} \sum_{a \in \mathcal{I}_1^{(1)}} (y^*)_a$$

for $\tilde{n} \in \mathcal{I}_1^{(1)}$ and $(y_*^{(1)})_{\tilde{n}} = (y^*)_{\tilde{n}}$ otherwise. Given that $y^* \in \bar{\mathcal{C}}$, we must have $\sigma(y^*) \in \bar{\mathcal{C}}$ by the definition of $\bar{\mathcal{C}}$ and therefore $y_*^{(1)} \in \bar{\mathcal{C}}$ because the summation is a convex combination. For $o \geq 2$, we can construct $y_*^{(o)} = \sum_{\sigma \in \Sigma_o} \frac{1}{|\mathcal{I}_o^{(1)}|} \sigma(y_*^{(o-1)})$. If $y_*^{(o-1)} \in \bar{\mathcal{C}}$, $y_*^{(o)} \in \bar{\mathcal{C}}$ must hold and $y_*^{(o)}$ satisfies that for any $\tilde{n} \in \mathcal{I}_{o'}^{(1)}$, $o' = 1, \dots, o$,

$$(y_*^{(o)})_{\tilde{n}} = \frac{1}{|\mathcal{I}_{o'}^{(1)}|} \sum_{a \in \mathcal{I}_{o'}^{(1)}} (y^*)_a,$$

and $(y_*^{(o)})_{\tilde{n}} = (y^*)_{\tilde{n}}$ otherwise. By induction, $y_*^{(\tau_0)} \in \bar{\mathcal{C}}$ and $y_*^{(\tau_0)} = y^{**}$.

Hence, we can reduce the constraints (EC.35) and (EC.32)–(EC.34) by imposing for some $\tilde{y} \in \mathbb{R}^{\tau_0}$, $\tilde{w} \in \mathbb{R}^{\tau_0}$ that for $a \in \mathcal{I}_o^{(1)}$, $o = 1, \dots, \tau_0$, $(y)_a = (\tilde{y})_o$ and $(w_i)_a = (\tilde{w}_i)_o$. This leads to the following constraints

$$\vec{1}^\top v_i + \vec{1}^\top (\lambda_1 \circ \tilde{w}_i) - (\lambda_1 \circ \tilde{y})^\top \vec{S}_1 \leq \delta_i - \delta_1, \quad (\text{EC.37})$$

$$\vec{X}_i \tilde{y}^\top - v_i \vec{1}^\top - \vec{1} \tilde{w}_i^\top \leq 0, \quad (\text{EC.38})$$

$$\mathcal{H}_{F_1}(\tilde{y}) \in \overline{\mathcal{C}}, \quad (\text{EC.39})$$

$$(\lambda_1 \circ \tilde{y})^\top \vec{S}_1 \leq \min_{x \in \mathcal{X}} (\lambda_1 \circ \tilde{y})^\top \vec{Z}'(x), \quad (\text{EC.40})$$

where $\lambda_1 := (|\mathcal{I}_1^{(1)}|, \dots, |\mathcal{I}_{\tau_0}^{(1)}|)$ and $\vec{Z}'(x) := (Z(x, \xi_1), \dots, Z(x, \xi_{\tau_0}))^\top$.

We now show that the above four constraints can be further reduced. Let $\mathcal{I}_o^{(i)}$ denote the set of indices n of \vec{X}_i such that $(\vec{X}_i)_n = (\vec{S}_i)_o$, $o = 1, \dots, \tau_i$, and therefore $|\mathcal{I}_o^{(i)}| = \bar{p}_o^i M$.

It is not difficult to see that for any $(v_i)_a$ such that $a \in \mathcal{I}_o^{(i)}$ the constraints associated with $(v_i)_a$ are identical in (EC.38). Because reducing $(v_i)_a$, for any a , is always feasible for (EC.37), if there exists any $(v_i^*)_a \neq (v_i^*)_b$ for $a, b \in \mathcal{I}_o^{(i)}$, we can always make them equal by reducing the larger one (without violating any constraint). We can thus conclude that we can always impose for some $\tilde{v}_i \in \mathbb{R}^{\tau_i}$ that $(v_i)_a = (\tilde{v}_i)_o$ for any $a \in \mathcal{I}_o^{(i)}$. This leads to the reformulation of the first constraint (EC.37) into

$$\vec{1}^\top (\lambda_i \circ \tilde{v}_i) + \vec{1}^\top (\lambda_1 \circ \tilde{w}_i) - (\lambda_1 \circ \tilde{y})^\top \vec{S}_1 \leq \delta_i - \delta_1,$$

where $\lambda_i := (|\mathcal{I}_1^{(i)}|, \dots, |\mathcal{I}_{\tau_i}^{(i)}|)$ and (EC.38) into

$$\vec{S}_i \tilde{y}^\top - \tilde{v}_i \vec{1}^\top - \vec{1} \tilde{w}_i^\top \leq 0.$$

Letting $(\lambda_i \circ \tilde{v}_i) = \hat{v}_i$, $(\lambda_1 \circ \tilde{w}_i) = \hat{w}_i$, and $(\lambda_1 \circ \tilde{y}) = \hat{y}$, we have (EC.38) become

$$\vec{S}_i ((\lambda_1)^{-1} \circ \hat{y})^\top - ((\lambda_i)^{-1} \circ \hat{v}_i) \vec{1}^\top - \vec{1} ((\lambda_1)^{-1} \circ \hat{w}_i)^\top \leq 0, \quad (\text{EC.41})$$

and (EC.39) become

$$\mathcal{H}_{F_1}((\lambda_1)^{-1} \circ \hat{y}) \in \overline{\mathcal{C}},$$

and (EC.37), (EC.40) reduce respectively to (24) (with $j = 1$) and (26) (with $t = 1$).

Finally, multiplying (EC.41) with $(\vec{1} \lambda_1^\top)$, we have

$$(\vec{1} \lambda_1^\top) \circ (\vec{S}_i ((\lambda_1)^{-1} \circ \hat{y})^\top - ((\lambda_i)^{-1} \circ \hat{v}_i) \vec{1}^\top - \vec{1} ((\lambda_1)^{-1} \circ \hat{w}_i)^\top) \leq 0,$$

which leads to the final formulation for (25) (with $j = 1$). \square

Proof of Proposition 6

Proof of Proposition 6 Following Proposition 3, we can equivalently formulate the problem as

$$\begin{aligned}
& \min_{x', \delta_Z, y_0, y_Z} ||x' - x^T|| \\
& \text{subject to } y_0^\top Zx' \leq \delta_Z, \\
& \delta_Z - y_Z^\top Zx' \leq 0, \\
& y_Z^\top Zx' \leq \min_x \{y_Z^\top Zx \mid Ax \geq b\}, \\
& y_0, y_Z \in \mathcal{C}.
\end{aligned}$$

Observe first that given any solution x' , y_Z , one can always set $y_0 = y_Z$ and $\delta_Z = y_Z^\top Zx'$ to satisfy the first two constraints. The third constraint, by definition, is equivalent to the constraint (28). Hence, the above problem is indeed equivalent to the first optimization problem in Proposition 6.

We can equivalently state the constraint (28) in terms of the KKT condition for linear programs, which is

$$Ax' - b \geq 0, \tag{EC.42}$$

$$A^\top u = Z^\top y,$$

$$u \geq 0, \tag{EC.43}$$

$$(Ax' - b)^\top u = 0. \tag{EC.44}$$

It is known that constraints (EC.42), (EC.43), and (EC.44) can be equivalently stated as the linear complementarity constraints $(Ax' - b)_i \cdot u_i = 0, \forall i$ (Luo et al. (1996)). By introducing binary variables $\eta_i \in \{0, 1\}$, we can apply the Big-M method to equivalently formulate these constraints as the constraints (29a)–(29e) provided that the constant M is sufficiently large.

Finally, because $\delta_Z = y_Z^\top Zx'$, we can obtain the risk function ρ_δ by setting $\vec{X}_1 := Zx^*$, $\vec{X}_2 = \vec{0}$, and $\delta_1 := \delta_Z^* = y_Z^{*\top} Zx^*$, $\delta_2 = 0$ with the optimal solution x^* and y_Z^* . \square

Proof of Proposition 7

Proof of Proposition 7 Following the proof in Proposition 4 (in particular, (EC.28)), we can equivalently formulate the problem as

$$\begin{aligned}
& \min_{x', \delta_Z, y_0, y_Z} ||x' - x^T|| \\
& \text{subject to } y_0^\top \sigma(Zx') \leq \delta_Z, \quad \forall \sigma \in \Sigma, \\
& \delta_Z - y_Z^\top Zx' \leq 0, \\
& y_Z^\top (\sigma(Zx') - Zx') \leq 0, \quad \forall \sigma \in \Sigma, \\
& y_Z^\top Zx' \leq \min_x \{y_Z^\top Zx \mid Ax \geq b\}, \\
& y_0, y_Z \in \bar{\mathcal{C}}.
\end{aligned} \tag{EC.45}$$

To show how one may reduce the problem, let us start by focusing on the constraint (EC.45):

$$y_Z^\top \sigma(Zx') \leq y_Z^\top Zx', \forall \sigma \in \Sigma.$$

By taking a closer look at this inequality, we can recognize that it is an instance of the rearrangement inequality, which states that y_Z is feasible to the constraints if and only if its ordering matches the ordering of Zx' :

$$(Zx')_i \leq (Zx')_j \Leftrightarrow (y_Z)_i \leq (y_Z)_j, \forall i \neq j.$$

We can thus apply the Big-M method to reformulate the above condition as the constraints (30a), (30b), (30c), and (30d) with sufficiently large M.

Moreover, for any $y_Z \in \bar{\mathcal{C}}$ satisfying the above matching condition, we can observe that

$$y_Z^\top Zx' \geq \left(\frac{1}{|\Omega|}\mathbf{1}\right)^\top Zx'$$

must hold. To see this, one may assume without loss of generality that $(Zx')_1 \leq \dots \leq (Zx')_{|\Omega|}$ and to minimize $y_Z^\top Zx'$ over $y_Z \in \bar{\mathcal{C}}$ that satisfies $(y_Z)_1 \leq \dots \leq (y_Z)_{|\Omega|}$, one can confirm that $y_Z = \frac{1}{|\Omega|}\mathbf{1}$ gives the minimum. This implies that given any y_Z satisfying (EC.45), there always exists a feasible y_0 and δ_Z for the first two constraints, namely by setting $y_0 = \frac{1}{|\Omega|}\mathbf{1}$ and $\delta_Z = y_Z^\top Zx'$. Hence, we can remove the first two constraints and arrive at the final formulation.

Finally, we can obtain the risk function $\bar{\rho}_\delta$ by setting $\vec{X}_1 := Zx^*$, $\vec{X}_2 = \vec{0}$, and $\delta_1 := \delta_Z^* = y_Z^{*\top} Zx^*$ and $\delta_2 = 0$ with the optimal solution x^* and y_Z^* . \square

Proof of Proposition 8

Proof of Proposition 8 Like the proof of Proposition 5, we first construct random variables in a outcome space Ω endowed with a uniform probability measure. We can then apply Proposition 7, and thereafter we show how the problem can be further reduced.

Given the finite supports $\{\xi_o^\top x\}_{o=1}^{\tau_0}$ and the associated probability values $\{\bar{p}_o^\xi\}_{o=1}^{\tau_0}$, we define random loss Zx as a mapping from Ω to \mathbb{R} that satisfies $(Zx)_i \in \{\xi_o^\top x\}_{o=1}^{\tau_0}$ and $|\{i \mid (Zx)_i = \xi_o^\top x\}| = \bar{p}_o^\xi M$. Let \mathcal{I}_o denote the set of indices i of $(Zx)_i$ such that $(Zx)_i = \xi_o^\top x$ and therefore $|\mathcal{I}_o| = \bar{p}_o^\xi M$. Using this definition of Zx , we can apply Proposition 7 to formulate the MIP model (30c). We claim that given any feasible solution $(x'^*, y^*, u^*, \eta^*, \nu_{i,j}^*)$ to (30c), one can always retain feasibility after replacing y^* by an alternative y^{**} that satisfies for any $i \in \mathcal{I}_o$,

$$(y^{**})_i = \frac{1}{|\mathcal{I}_o|} \sum_{a \in \mathcal{I}_o} (y^*)_a,$$

$o = 1, \dots, \tau_0$.

First, given the definition of \mathcal{I}_o , clearly for any y^* that are ordered the same way as Zx'^* (i.e., $(Zx'^*)_i \leq (Zx'^*)_j \Leftrightarrow y_i^* \leq y_j^*$), it must hold also that $(Zx'^*)_i \leq (Zx'^*)_j \Leftrightarrow y_i^{**} \leq y_j^{**}$. It is also clear

that $A^\top u^* = Z^\top y^* = Z^\top y^{**}$. Finally, to check $y^{**} \in \bar{\mathcal{C}}$, one can find the necessary arguments to prove it in the proof of Proposition 5.

Hence, this implies that without loss of generality, we can impose for some $\tilde{y} \in \mathbb{R}^{\tau_0}$ that for $a \in \mathcal{I}_o$, $o = 1, \dots, \tau_0$, $(y)_a = (\tilde{y})_o$ must hold. This leads to the constraints (31) and (32) in \tilde{y} and the constraint $A^\top u = \Xi^\top(\bar{p}_o^\xi M \circ \tilde{y})$ and $\mathcal{H}_{F_\xi}(\tilde{y}) \in \bar{\mathcal{C}}$. By setting $y = M\tilde{y}$, we arrive at the final formulation. \square

Proof of Example 4 (5. Spectral risk measures)

Proof of Example 4 (5. Spectral risk measures) Let $M \in \mathbb{Z}^+$ be a constant such that p_k , $k = 1, \dots, K$, and \bar{p}_o , $o = 1, \dots, \tau_j$, can be expressed in the form of n/M , $n \in \{1, \dots, M\}$. First, given the stepwise spectrum $\phi^-(p)$, to apply the representation \mathcal{C} in Example EC.2.6 (i.e., $\bar{\mathcal{C}}$), we have $\phi_j = \int_{\frac{j-1}{M}}^{\frac{j}{M}} \phi^-(t) dt = \frac{\bar{\phi}_k}{M}$ for any $j \in \{1, \dots, M\}$ such that $p_{k-1} \leq \frac{j-1}{M} < \frac{j}{M} \leq p_k$, and therefore $|\left\{j \mid \phi_j = \frac{\bar{\phi}_k}{M}\right\}| = (p_k - p_{k-1})M$. To see how the constraint $\mathcal{H}_{F_j}((\lambda_{F_j})^{-1} \circ y) \in \bar{\mathcal{C}}$ can be reduced, we apply first the result of Birkhoff (1946) to reformulate $\bar{\mathcal{C}}$ into

$$\bar{\mathcal{C}} = \left\{q \mid q = Q\phi, Q\vec{1} = \vec{1}, Q^\top \vec{1} = \vec{1}, Q \geq 0\right\}.$$

Let $\mathcal{I}_o^{(j)}$, $o = 1, \dots, \tau_j$, denote the set of indices n of $\mathcal{H}_{F_j}((\lambda_{F_j})^{-1} \circ y)$ such that $(\mathcal{H}_{F_j}((\lambda_{F_j})^{-1} \circ y))_n = ((\lambda_{F_j})^{-1} \circ y)_o$ and therefore $|\mathcal{I}_o^{(j)}| = (\lambda_{F_j})_o$. It is obvious that the constraint $\mathcal{H}_{F_j}((\lambda_{F_j})^{-1} \circ y) \in \bar{\mathcal{C}}$ has a feasible solution if and only if the following set of constraints

$$q_i = q_j, \forall i, j \in \mathcal{I}_o^{(j)}, o = 1, \dots, \tau_j, q = Q\phi, Q\vec{1} = \vec{1}, Q^\top \vec{1} = \vec{1}, Q \geq 0 \quad (\text{EC.46})$$

has a feasible solution. We show first how (EC.46) can be reduced. Let (q^*, Q^*) denote a feasible solution for the above constraints. We claim that the solution q^* together with the following construction of Q^{**}

$$(Q^{**})_{(\tilde{n}, :)} = \frac{1}{|\mathcal{I}_o^{(j)}|} \sum_{n \in \mathcal{I}_o^{(j)}} (Q^*)_{(n, :)}, \tilde{n} \in \mathcal{I}_o^{(j)}, o = 1, \dots, \tau_j$$

will also be feasible. The notation $(V)_{(k, :)}$ (respectively $(V)_{(:, k)}$) refers to the k th-row (respectively k th-column) of the matrix V . The claim can be fairly straightforward to verify by direct substitution, which gives $q^* = Q^{**}\phi$, $Q^{**}\vec{1} = \vec{1}$, and $Q^{**\top} \vec{1} = \vec{1}$.

Hence, we can reduce (EC.46) by imposing that for any $\tilde{n} \in \mathcal{I}_o^{(j)}$, $o = 1, \dots, \tau_j$, $(Q)_{(\tilde{n}, :)} = (\bar{Q})_{(o, :)}$ for some $\bar{Q} \in \mathbb{R}^{\tau_j \times M}$, which leads to

$$q_i = \bar{q}_o, \forall i \in \mathcal{I}_o^{(j)}, o = 1, \dots, \tau_j, \bar{q} = \bar{Q}\phi, \bar{Q}\vec{1} = \vec{1}, ((\lambda_{F_j} \vec{1}^\top) \circ \bar{Q})^\top \vec{1} = \vec{1}, \bar{Q} \geq 0, \quad (\text{EC.47})$$

where $\bar{q} \in \mathbb{R}^{\tau_j}$. Moreover, the constraint $\mathcal{H}_{F_j}((\lambda_{F_j})^{-1} \circ y) \in \bar{\mathcal{C}}$ can be equivalently written as $y \in \mathbb{R}_+^{\tau_j} \cap \mathcal{C}$, where

$$\mathcal{C} := \left\{ \bar{q} \mid \bar{q} = ((\lambda_{F_j} \bar{\mathbf{1}}^\top) \circ \bar{Q})\phi, \bar{Q}\bar{\mathbf{1}} = \bar{\mathbf{1}}, ((\lambda_{F_j} \bar{\mathbf{1}}^\top) \circ \bar{Q})^\top \bar{\mathbf{1}} = \bar{\mathbf{1}}, \bar{Q} \geq 0 \right\}.$$

Letting $\hat{Q} = (\lambda_{F_j} \bar{\mathbf{1}}^\top) \circ \bar{Q}$, we have

$$\mathcal{C} = \left\{ \bar{q} \mid \bar{q} = \hat{Q}\phi, \hat{Q}\bar{\mathbf{1}} = \lambda_{F_j}, \hat{Q}^\top \bar{\mathbf{1}} = \bar{\mathbf{1}}, \hat{Q} \geq 0 \right\}. \quad (\text{EC.48})$$

Next, let $\mathcal{I}_k^{(\phi)}$ denote the set of indices j of ϕ such that $\phi_j = \frac{\bar{\phi}_k}{M}$ for $k = 1, \dots, K$ and therefore $|\mathcal{I}_k^{(\phi)}| = (p_k - p_{k-1})M$. Given this, we show that the constraints in (EC.48) can be further reduced. Let \bar{q}^*, \hat{Q}^* be a feasible solution for (EC.48). We claim that \bar{q}^* together with the following construction of \hat{Q}^{**}

$$\hat{Q}_{(:, \tilde{n})}^{**} := \frac{1}{|\mathcal{I}_k^{(\phi)}|} \sum_{n \in \mathcal{I}_k^{(\phi)}} \hat{Q}_{(:, n)}^*, \tilde{n} \in \mathcal{I}_k^{(\phi)}, k = 1, \dots, K$$

is also feasible for the constraints. Similarly, the claim can be verified by the direct substitution, which gives $\hat{Q}^{**}\phi = \bar{q}^*$, $\hat{Q}^{**}\bar{\mathbf{1}} = \lambda_{F_j}$, and $(\hat{Q}^{**})^\top \bar{\mathbf{1}} = \bar{\mathbf{1}}$.

Hence, we can also impose that for any $\tilde{n} \in \mathcal{I}_k^{(\phi)}$, $k = 1, \dots, K$, $\hat{Q}(:, \tilde{n}) = \tilde{Q}(:, k)$ for some $\tilde{Q} \in \mathbb{R}^{\tau_j \times K}$ in (EC.48), which leads to the reformulation of the first and second constraint in \mathcal{C} into

$$\bar{q} = ((\bar{\mathbf{1}}\lambda_\phi^\top) \circ \tilde{Q})(\frac{1}{M}\bar{\phi}), \text{ and } ((\bar{\mathbf{1}}\lambda_\phi^\top) \circ \tilde{Q})\bar{\mathbf{1}} = \lambda_{F_j},$$

where $(\lambda_\phi)_k := (p_k - p_{k-1})M$, and therefore also the set \mathcal{C} into

$$\mathcal{C} = \left\{ \bar{q} \mid \bar{q} = (\frac{1}{M})(\bar{\mathbf{1}}\lambda_\phi^\top) \circ \tilde{Q}\bar{\phi}, ((\bar{\mathbf{1}}\lambda_\phi^\top) \circ \tilde{Q})\bar{\mathbf{1}} = \lambda_{F_j}, \tilde{Q}^\top \bar{\mathbf{1}} = \bar{\mathbf{1}}, \tilde{Q} \geq 0 \right\}.$$

Letting $\dot{Q} = (\frac{1}{M})(\bar{\mathbf{1}}\lambda_\phi^\top) \circ \tilde{Q}$, we arrive at the final reduced form. \square

EC.5. Further discussion about the issue of constraint misspecification raised in Remark 4

As mentioned in Remark 4, it is possible that the forward problem assumed in our inverse models does not well represent the true problem that the decision maker solved. In this section, we attempt to discuss this issue more formally and provide some ideas that might help (partially) resolve the issue.

The forward problem in this paper is completely characterized by the feasible sets \mathcal{X}^t , $t \in \mathcal{T}$, and from this point on we assume that the sets take the form of $\mathcal{X}^t := \{x \mid g_j^t(x) \leq 0, j = 1, \dots, J^t\}$, $t \in \mathcal{T}$. To differentiate the feasible sets \mathcal{X}^t assumed in our inverse models from the “true” feasible sets (i.e., the ones based on which the past decisions x^t , $t \in \mathcal{T}$ were optimized), we denote by $\bar{\mathcal{X}}^t$, $t \in \mathcal{T}$ the true feasible sets and thus $x^t \in \bar{\mathcal{X}}^t$, $t \in \mathcal{T}$ follows. To facilitate our discussion, we assume that the decision makers are rational ⁷ (i.e., the past decisions were optimally made with respect to the true risk function ρ^*). We can thus write down the following optimality condition that must hold for the observed decisions x^t , $t \in \mathcal{T}$ with respect to the true risk function ρ^* :

$$\rho^*(\bar{Z}^t(x^t)) \leq \rho^*(\bar{Z}^t(x)), \forall x \in \bar{\mathcal{X}}^t, t \in \mathcal{T}. \quad (\text{EC.49})$$

Recall that our inverse models take as input the observed decisions x^t , $t \in \mathcal{T}$ and the feasible sets \mathcal{X}^t , $t \in \mathcal{T}$ and that they seek a risk function ρ from the following set:

$$\mathcal{R}_{inv} = \left\{ \rho \mid \rho(\bar{Z}^t(x^t)) \leq \rho(\bar{Z}^t(x)), \forall x \in \mathcal{X}^t, t \in \mathcal{T} \right\}. \quad (\text{EC.50})$$

The question here is how the discrepancy between \mathcal{X}^t and $\bar{\mathcal{X}}^t$ would affect the risk function ρ generated from the inverse models and if the models can actually detect such a discrepancy. A trivial case where we can easily draw the conclusion, without even running the inverse models, is that if there are some observed decisions that are simply not feasible with respect to \mathcal{X}^t (i.e., $x^t \notin \mathcal{X}^t$), then one can easily tell which constraint is misspecified by checking which one is violated (i.e., $g_j^t(x^t) > 0$ for some $j \in \{1, \dots, J^t\}$ and $t \in \mathcal{T}$).

We thus focus primarily on the case where all the observed decisions are feasible with respect to \mathcal{X}^t (i.e., $x^t \in \mathcal{X}^t$, $t \in \mathcal{T}$) but the feasible sets \mathcal{X}^t may be misspecified (i.e., $\mathcal{X}^t \neq \bar{\mathcal{X}}^t$). The inverse models in this case could either return a message of infeasibility (i.e., $\mathcal{R}_{inv} = \emptyset$) or a feasible risk function ρ that can fit all the observed decisions with respect to \mathcal{X}^t . The case of infeasibility would allow one to detect the sets \mathcal{X}^t being misspecified, because we know from (EC.49) and (EC.50) that if the sets \mathcal{X}^t are correctly specified (i.e., $\mathcal{X}^t = \bar{\mathcal{X}}^t$), then $\rho^* \in \mathcal{R}_{inv}$ (i.e., $\mathcal{R}_{inv} \neq \emptyset$). In the case

⁷ This assumption is in fact necessary; otherwise in the case where there does not exist a risk function ρ that fits all the observed decisions, one cannot tell if it is because of the sub-optimality of the observed decisions or the misspecification of constraints.

where the inverse models do return a feasible risk function ρ , it appears, however, as discussed below, generally not possible to detect any misspecification of \mathcal{X}^t .

In particular, let us consider the case motivated by our portfolio management example where the client might not necessarily take into account the long-only constraint. In this case, the feasible set \mathcal{X}^t is only a subset of the true feasible set $\bar{\mathcal{X}}^t$ (i.e., $\mathcal{X}^t \subset \bar{\mathcal{X}}^t$), and we know from (EC.49) that the following must hold:

$$\rho^*(\vec{Z}^t(x^t)) \leq \rho^*(\vec{Z}^t(x)), \forall x \in \mathcal{X}^t, t \in \mathcal{T}. \quad (\text{EC.51})$$

That is, we have $\rho^* \in \mathcal{R}_{inv}$. In other words, despite $\mathcal{X}^t \neq \bar{\mathcal{X}}^t$, in this case the inverse models can still narrow down a smaller set of risk functions that includes the true risk function ρ^* as a feasible candidate. This also demonstrates why one would not be able to detect the discrepancy between \mathcal{X}^t and $\bar{\mathcal{X}}^t$, given that it is not even possible to rule out ρ^* from the set \mathcal{R}_{inv} . The only difference in this case between applying the set \mathcal{X}^t and the true feasible set $\bar{\mathcal{X}}^t$ in (EC.50) is that the latter might allow for a faster rate of convergence of the set \mathcal{R}_{inv} to the true risk function ρ^* . But it appears not possible to detect such a difference without knowing what the true risk function is. In other cases where $\mathcal{X}^t \neq \bar{\mathcal{X}}^t$ but $\mathcal{X}^t \not\subset \bar{\mathcal{X}}^t$, it is possible that the true risk function ρ^* may no longer be feasible (i.e., $\rho^* \notin \mathcal{R}_{inv}$). It remains unclear, however, how one can detect such infeasibility, given that all one knows from the output of the inverse models is that there exists a risk function ρ that perfectly fits all the observed decisions.

Although in general it appears not possible to detect $\rho^* \notin \mathcal{R}_{inv}$ (and thus also $\mathcal{X}^t \neq \bar{\mathcal{X}}^t$), we provide here some idea as to how one may quantify potential risk underestimation because of constraint misspecification. The idea here can also be applied to the case where one is able to confirm $\rho^* \notin \mathcal{R}_{inv}$ (e.g., the case $\mathcal{R}_{inv} = \emptyset$ as mentioned earlier) and tries to fix the issue. This idea comes from the observation made earlier that if $\mathcal{X}^t \subset \bar{\mathcal{X}}^t$, our inverse models would necessarily still include the true risk function ρ^* as a feasible candidate. So, given the sets of potentially misspecified feasible sets $\mathcal{X}^t, t \in \mathcal{T}$ and observed decisions $x^t, t \in \mathcal{T}$, what one can do is to construct alternative feasible sets \mathcal{X}_*^t that satisfy

$$x^t \in \mathcal{X}_*^t \subset \mathcal{X}^t \cap \bar{\mathcal{X}}^t, t \in \mathcal{T}, \quad (\text{EC.52})$$

(i.e., each \mathcal{X}_*^t is a subset of both the potentially misspecified set and the true feasible set). By replacing the sets $\mathcal{X}^t, t \in \mathcal{T}$ with the subsets $\mathcal{X}_*^t, t \in \mathcal{T}$ in (EC.50), one can apply instead the following set in the inverse models:

$$\mathcal{R}_{inv}^* = \left\{ \rho \mid \rho(\vec{Z}^t(x^t)) \leq \rho(\vec{Z}^t(x)), \forall x \in \mathcal{X}_*^t, t \in \mathcal{T} \right\}. \quad (\text{EC.53})$$

The condition $\mathcal{X}_*^t \subset \mathcal{X}^t$ implies that $\mathcal{R}_{inv} \subset \mathcal{R}_{inv}^*$, and the condition $\mathcal{X}_*^t \subset \bar{\mathcal{X}}^t$ implies that the above set always contains the true risk function ρ^* (i.e., $\rho^* \in \mathcal{R}_{inv}^*$). With these guarantees, one

can apply our inverse model (4) to generate a risk function ρ^\uparrow that bounds from above the true risk function ρ^* (i.e., $\rho^\uparrow \geq \rho^*$) and any risk function generated from the original set \mathcal{R}_{inv} (i.e., $\rho^\uparrow \geq \rho, \forall \rho \in \mathcal{R}_{inv}$). Hence, ρ^\uparrow provides a means to measure potential risk underestimation from constraint misspecification. More specifically, one may compare the risk estimates obtained from a risk function that is generated with respect to potentially misspecified constraints against ρ^\uparrow . If the difference is small, this may be regarded as a signal that the impact of constraint misspecification is less of a concern. Otherwise, the difference provides the upper bound on the potential risk underestimation from constraint misspecification. Ideally, one should seek the largest \mathcal{X}_*^t that satisfies (EC.52), given that the larger the constructed set \mathcal{X}_*^t is, the tighter the bound ρ^\uparrow is (and hence the more informative the bound is).

Here, we give two examples of how the set \mathcal{X}_*^t may be constructed. In the first one, we assume the form of the constraints $g_j^t(x)$ is known but it is only some of the “right-hand-side” parameters that are prone to misspecification: $g_j^t(x) := h_j^t(x) - b_j^t \leq 0, j = 1, \dots, J^t$, where $h_j^t(x)$ is known but b_j^t may be misspecified for some j and t . For example, it is common also in portfolio management that upper bounds must be imposed over the amount of investment for each asset (i.e., $g_j^t(x) = x_j - b_j^t \leq 0, j = 1, \dots, \dim(x), t \in \mathcal{T}$) for diversification purposes, but the values of the bounds may not be known exactly. To construct the set \mathcal{X}_*^t in this case, one can lower the values of the parameters b_j^t that are prone to misspecification, so that they guarantee $x^t \in \mathcal{X}_*^t \subset \bar{\mathcal{X}}^t$ (e.g., by setting $b_j^t := h_j^t(x^t)$). In the second example, we assume that the true feasible set $\bar{\mathcal{X}}^t$ does not change over time (i.e., $\bar{\mathcal{X}}^t = \bar{\mathcal{X}}, \forall t \in \mathcal{T}$). One can then always construct the convex hull of the observed decisions $\mathcal{X}_* := \text{Conv}(\{x^t\}_{t \in \mathcal{T}})$ so that $\mathcal{X}_* \subset \bar{\mathcal{X}}$. These examples are meant for demonstrating the general principle that one may apply for constructing the set \mathcal{X}_*^t .