DATA-DRIVEN INVERSE OPTIMIZATION WITH INCOMPLETE INFORMATION

PEYMAN MOHAJERIN ESFAHANI, SOROOSH SHAFIEEZADEH-ABADEH,
GRANI A. HANASUSANTO, AND DANIEL KUHN

ABSTRACT. In data-driven inverse optimization an observer aims to learn the preferences of an agent who solves a parametric optimization problem depending on an exogenous signal. Thus, the observer seeks the agent’s objective function that best explains a historical sequence of signals and corresponding optimal actions. We formalize this inverse optimization problem as a distributionally robust program minimizing the worst-case risk that the estimated decision (i.e., the decision implied by a particular candidate objective) differs from the agent’s actual response to a random signal. We show that our framework offers attractive out-of-sample performance guarantees for different prediction errors and that the emerging inverse optimization problems can be reformulated as (or approximated by) tractable convex programs when the prediction error is measured in the space of objective values. A main strength of the proposed approach is that it naturally generalizes to situations where the observer has imperfect information, e.g., when the agent’s true objective function is not contained in the space of candidate objectives, when the agent suffers from bounded rationality or implementation errors, or when the observed signal-response pairs are corrupted by measurement noise.

1. Introduction

In inverse optimization an observer aims to learn the preferences of an agent who solves a parametric optimization problem depending on an exogenous signal. The observer knows the constraints imposed on the agent’s actions but is unaware of her objective function. By monitoring a sequence of signals and corresponding actions, the observer seeks to identify an objective function that makes the observed actions optimal in the agent’s optimization problem. This learning problem can be cast as an inverse optimization problem over candidate objective functions. The hope is that the solution of this inverse problem enables the observer to predict the agent’s future actions in response to new signals.

Inverse optimization has a wide spectrum of applications spanning several disciplines ranging from econometrics and operation research to engineering and biology. For example, a marketing executive aims to understand the purchasing behavior of consumers with unknown utility functions by monitoring sales figures [1, 11, 8], a transportation planner wishes to learn the route choice preferences of the passengers in a multimodal transport system by measuring traffic flows [3, 14, 15, 18], or a healthcare manager seeks to design clinically acceptable treatments in view of historical treatment plans [17]. It is even believed that the behavior of many biological systems is governed by a principle of optimality with respect to an unknown decision criterion, which can be inferred by tracking the system [12, 35].
Inverse optimization has also been applied in geoscience [29, 38], portfolio selection [10, 23], production planning [36], inventory management [16], network design and control [3, 15, 22, 18] etc.

The main thrust of the early literature on inverse optimization is to identify an objective function that explains a single observation. In the seminal paper [4] the agent solves a static (non-parametric) linear program and reveals her optimal decision to the observer, who then identifies the objective function closest to a prescribed nominal objective, under which the observed decision is optimal. This model was later extended to conic programs [23], integer programs [2, 21, 33, 37] and linearly constrained separable convex programs [39]. Another variant of this problem is considered in [2], where the observer identifies an admissible objective function for which the optimal value of the agent’s problem is closest to the observed optimal value corresponding to the unknown true objective.

This paper focuses on inverse optimization problems where the agent solves a parametric optimization problem repeatedly in response to random signals, and thus the observer has access to multiple signal-response observations. Data-driven inverse optimization problems of this type have only just started to attract attention, and to the best of our knowledge there are currently only three papers that study such problems. In [24] the observer seeks an objective function under which all observed decisions satisfy the Karush-Kuhn-Tucker (KKT) optimality conditions of the agent’s convex optimization problem. More precisely, as it is thinkable that the agent’s true objective function is not contained in the considered search space, the observer minimizes some norm of the KKT residuals corresponding to the observed decisions. A similar approach is pursued in [11], where the optimality conditions are expressed via variational inequalities that can be reformulated as tractable conic constraints using ideas from robust optimization. This approach has the additional benefit that it extends to inverse equilibrium problems. The most recent paper formulates the data-driven inverse optimization problem as a bilevel program that has conceptual appeal when the observations are noisy [7].

The contributions of this paper can be summarized as follows:

- We propose a unifying framework for data-driven inverse optimization. Specifically, we formulate the inverse optimization problem as a distributionally robust program minimizing the worst-case risk that the estimated decision (i.e., the decision implied by a particular candidate objective) differs from the agent’s actual response to a random signal. Here, the worst case is evaluated with respect to an ambiguity set of distributions that could have generated the historical signal-response pairs with high confidence.

- Prediction errors can be measured, e.g., in the space of decisions, the space of objective values or the space of the residuals of the optimality conditions. Using this flexibility, we show that the approaches propagated in [11] and [7] emerge as special cases of our framework.

- We demonstrate that our framework offers attractive out-of-sample performance guarantees for the different prediction errors.

- We prove that the emerging distributionally robust optimization problems can be reformulated as (or approximated by) tractable convex programs when the prediction error is measured in the space of objective values.

- We show that our distributionally robust approach naturally generalizes to situations where the observer has imperfect information, e.g., when the agent’s true objective function is not contained in the space of candidate objectives, when the agent suffers from bounded rationality or when the observed signal-response pairs are corrupted by measurement noise.
The proposed framework strikes a healthy balance between predictive power and tractability. When measuring prediction errors in the space of objective values, our method enjoys similar tractability properties as the variational inequality-based method in [11] but offers often stronger (but never weaker) predictions of the agents decisions. On the other hand, our method has computational advantages over the bilevel approach [7], which leads to a non-convex optimization problem, but results in weaker prediction errors in the space of decisions. We further remark that the bilevel approach offers an asymptotic guarantee on the predictability of the agent’s decision, while the variational inequality method enjoys a finite-sample guarantee. The size of the ambiguity set in the approach proposed here constitutes an additional degree of freedom that is unavailable in the existing methods. Our numerical experiments indicate that this extra freedom enables us to obtain stronger finite-sample guarantees.

**Notation.** The inner product of two vectors $s, t \in \mathbb{R}^m$ is denoted by $\langle s, t \rangle := s^T t$, and the dual of a norm $\| \cdot \|$ on $\mathbb{R}^m$ is defined through $\| \|_* := \sup_{\|s\|\leq 1} \langle t, s \rangle$. The dual of a proper (closed, solid, pointed) convex cone $C \subset \mathbb{R}^m$ is defined as $C^* := \{ t \in \mathbb{R}^m : \langle t, s \rangle \geq 0 \ \forall s \in C \}$, and the relation $s \succeq_C t$ is interpreted as $s - t \in C$. Similarly, for two symmetric matrices $Q, R \in \mathbb{R}^{m \times m}$ the relation $Q \succeq R$ ($Q \preceq R$) means that $Q - R$ is positive (negative) semidefinite. The identity matrix is denoted by $I$. The indicator function $\mathbb{1}_C$ of a logical expression $C$ is defined as $\mathbb{1}_C = 1$ if $C$ is true; $= 0$ otherwise. We denote by $\delta_\xi$ the Dirac distribution concentrating unit mass at $\xi \in \Xi$. The $N$-fold product of a distribution $\mathbb{P}$ on $\Xi$ is denoted by $\mathbb{P}^N$, which represents a distribution on the Cartesian product $\Xi^N$.

### 2. Inverse Optimization under Perfect Information

Consider an agent who first receives a random signal $s \in S \subset \mathbb{R}^m$ and then solves the following parametric optimization problem:

$$
\text{(1)} \quad \min_{x \in \mathbb{X}(s)} F(s, x).
$$

Note that both the objective function $F : \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}$ as well as the (multivalued) feasible set mapping $\mathbb{X} : \mathbb{R}^m \rightrightarrows \mathbb{R}^n$ depend on the signal. We assume for simplicity that (1) is a convex program for every $s \in S$ and that its minimum is always attained. Consider also an independent observer who monitors the signal $s$ as well as the agent’s response $x$, which is given by a minimizer of (1) for the fixed signal $s$. Throughout the paper, we assume that the observer is ignorant of the agent’s preferences encoded by the objective function $F$. Thus, a priori, the observer cannot predict the agent’s response to a particular signal. We assume, however, that $F$ is known to belong to a family $\mathcal{F} = \{ F_\theta : \theta \in \Theta \}$ of candidate objective functions that are all convex in $x$, where $\Theta$ represents a finite-dimensional parameter set (this somewhat restrictive assumption will be relaxed in Section 6).

If a candidate objective function $F_\theta$ is believed to approximate $F$, it can be used to predict the agent’s optimal response to a signal $s$ by solving a variant of problem (1), where $F$ is replaced with $F_\theta$.

**Definition 2.1 (Predicted Response).** We refer to $x_\theta(s) \in \arg \min_{x \in \mathbb{X}(s)} F_\theta(s, x)$ as a response to $s$ predicted by $\theta$. Note that $x_\theta(s)$ will typically differ from $x$, which is an optimal response to $s$ under the unknown true objective function $F$. We denote by $\mathbb{X}_\theta(s)$ the set of all responses to $s$ predicted by $\theta$.

In order to assess how well a particular candidate model $F_\theta$ predicts the agent’s decisions, the observer uses a parametric loss function $\ell_\theta : \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}_+$, which satisfies $\ell_\theta(s, x) = 0$ if $x \in \mathbb{X}_\theta(s)$ and $\ell_\theta(s, x) > 0$ otherwise. Specifically, the observer might try to learn the agent’s preferences by minimizing the loss function $\ell_\theta(s, x)$ over all candidate models encoded by $\theta \in \Theta$. However, this naïve
optimization problem is not well-defined because \( s \) and, \emph{a fortiori}, \( x \) are random objects. Indeed, the signal-response pairs \( \xi := (s, x) \) are governed by some probability distribution \( P \) supported on \( \Xi := \{(s, x) : s \in S, x \in X(s)\} \), which can be viewed as the graph of the feasible set mapping \( X \). Note that the marginal distribution of \( s \) captures the frequency of the exogenous signals, while the conditional distribution of \( x \) given \( s \) depends on the agent’s true objective function \( F \). The observer can therefore infer the agent’s preferences from \( P \) by solving the well-defined inverse optimization problem

\[
\begin{align*}
\text{minimize} & \quad \rho^F(\ell_\theta),
\end{align*}
\]

where \( \rho^F \) constitutes a (normalized, translation-invariant and monotone) risk-measure under \( P \) that penalizes positive losses. In the remainder we will sometimes refer to \( \rho^F(\ell_\theta) \) as the \emph{certainty equivalent} of the random loss \( \ell_\theta(s, x) \). If the agent’s true objective function is contained in \( F \), that is, if \( F = F_\theta^* \) for some \( \theta^* \in \Theta \), then the loss \( \ell_\theta(s, x) \) vanishes almost surely with respect to \( P \), implying that \( \theta^* \) is optimal in (2). If \( F \neq F_\theta^* \), on the other hand, the certainty equivalent of \( \ell_\theta(s, x) \) is nonnegative (typically strictly positive) due to the normalization and monotonicity properties of \( \rho^F \).

\textbf{Example 1} (Loss Functions). A possible loss function is

\[ \ell_\theta(s, x) := F_\theta(s, x) - \min_{y \in X(s)} F_\theta(s, y), \]

which can be interpreted as the suboptimality of the decision \( x \) in (1) with respect to the objective function \( F_\theta \). If \( F_\theta \) is differentiable with respect to \( x \), one may alternatively define

\[ \ell_\theta^{\nabla F}(s, x) := \max_{y \in X(s)} \langle \nabla_x F_\theta(s, x), x - y \rangle, \]

which measures to what extent \( x \) violates the first-order optimality condition of the convex program (1) where \( F \) is replaced with \( F_\theta \). Yet another possibility is to define the loss function as the squared distance of \( x \) from the set of responses to \( s \) predicted by \( \theta \), that is,

\[ \ell_\theta^P(s, x) := \min_{y \in X(s)} \|x - y\|^2. \]

To our best knowledge, the loss function (3a) has not yet been considered in inverse optimization. In contrast, (3b) has been used in \cite{11} for solving inverse variational inequality problems, and (3c) has been studied in \cite{7}. While the loss functions (3a) and (3c) admit intuitive interpretations as the prediction errors of the agent’s optimal costs and decisions, respectively, (3b) lacks a clear physical interpretation and seems to serve primarily as an approximation for (3a) or (3c); see also Proposition 2.2 below.

The following proposition establishes basic properties of the loss functions (3).

\textbf{Proposition 2.2} (Dominance Relations between Loss Functions). Assume that \( F_\theta(s, x) \) is convex and differentiable in \( x \), and define \( \gamma \geq 0 \) as the largest number satisfying the inequality

\[ F_\theta(s, y) - F_\theta(s, x) \geq \langle \nabla_x F_\theta(s, x), y - x \rangle + \frac{\gamma}{2} \|y - x\|^2 \quad \forall x, y \in X(s). \]

Note that (4) always holds for \( \gamma = 0 \); see \cite[§ 3.1.3]{13}. Then, the loss functions of Example 1 satisfy

\[ \ell_\theta^{\nabla F}(s, x) \geq \ell_\theta(s, x) \geq \frac{\gamma}{2} \ell_\theta^P(s, x) \quad \forall s \in S, x \in X(s). \]

Moreover, all three loss functions are non-negative and evaluate to zero if and only if \( x \in X_\theta(s) \).
Proof. Setting $\gamma = 0$ and minimizing both sides of (4) over $y \in \mathcal{X}(s)$ yields $\ell^\alpha_\theta \geq \ell_\theta$. Next, the first-order optimality condition of the convex program (1) with objective function $F_\theta$ requires that
\begin{equation}
(\nabla_y F_\theta(s, x), y - x) \geq 0 \quad \forall y \in \mathcal{X}(s)
\end{equation}
at any optimal point $x \in \mathcal{X}_\theta(s)$. Combining the inequalities (4) and (6) then yields
\begin{equation}
F_\theta(s, y) - F_\theta(s, x) \geq \frac{\gamma}{2} ||y - x||^2 \quad \forall x \in \mathcal{X}_\theta(s), \ y \in \mathcal{X}(s).
\end{equation}
Minimizing both sides of the above inequality over $x \in \mathcal{X}_\theta(s)$ yields $\ell_\theta \geq \frac{\gamma}{2} F_\theta$. Note that this inequality is only useful for $\gamma > 0$, in which case $\mathcal{X}_\theta(s)$ is in fact a singleton. Moreover, it is straightforward to verify that all loss functions are non-negative and evaluate to zero if and only if $x \in \mathcal{X}_\theta(s)$. In the case of $\ell^\gamma_V$, for instance, this equivalence holds because the first-order condition (6) is both necessary and sufficient for the optimality of $x$. We remark that (5) remains valid if $\gamma$ depends on $s$ and $\theta$. \hfill \Box

The following example suggests two risk measures that could be used in (2).

Example 2 (Risk Measures). A popular risk measure that the observer could use to quantify the risk of incurring a positive loss is the value-at-risk (VaR) at level $\alpha \in [0, 1]$, which is defined as

\begin{equation}
\text{VaR}_\alpha^\theta(\ell_\theta) = \inf_{\tau} \{ \tau : P[\ell_\theta(s, x) \leq \tau] \geq 1 - \alpha \}.
\end{equation}

Note that the VaR coincides with the upper $(1 - \alpha)$-quantile of the loss distribution. Alternatively, the observer could rely on the conditional value-at-risk (CVaR) at level $\alpha \in (0, 1]$, which is defined as

\begin{equation}
\text{CVaR}_\alpha^\theta(\ell_\theta) = \inf_{\tau} \tau + \frac{1}{\alpha} E[\max\{\ell_\theta(s, x) - \tau, 0\}],
\end{equation}

see [32]. For $\alpha = 1$, the CVaR reduces to the expected value, and for $\alpha \downarrow 0$, it converges to the essential supremum of the loss. If $\ell_\theta(s, x)$ has a continuous marginal distribution under $P$, then $\text{CVaR}_\alpha^\theta(\ell_\theta)$ coincides with the expected loss above $\text{VaR}_\alpha^\theta(\ell_\theta)$.

Remark 2.3 (Choice of Risk Measures). Recall that the minimum of (2) vanishes whenever $F \in \mathcal{F}$. In this case, any $\theta^\star$ with $F_{\theta^\star} = F$ is a minimizer of (2) irrespective of $\rho^P$. Thus, one might believe that the choice of the risk measure is immaterial for the inverse optimization problem. However, different risk measures may result in different solution sets. For example, if $\rho^P$ is the CVaR at level $\alpha \in (0, 1]$, then $\theta^\star$ is a minimizer of (2) if and only if $P[\ell_\theta(s, x) = 0] = 1$. In contrast, if $\rho^P$ is the VaR at level $\alpha \in [0, 1]$, then $\theta^\star$ is a minimizer of (2) if and only if $P[\ell_\theta(s, x) = 0] \geq 1 - \alpha$. Thus, the use of VaR leads to an inflated solution set. Moreover, the choice of the risk measure may impact the tractability and approximability of (2); see also Sections 4 and 5.

Unfortunately, the distribution $P$ of the signal-response pairs is unknown in practice, and therefore we lack essential information to evaluate the risk of $\ell_\theta(s, x)$ and, a fortiori, to solve the inverse optimization problem (2). Often, it is reasonable to assume that the observer has access to $N$ independent samples $\hat{\xi}_i := (\hat{s}_i, \hat{x}_i)$ from $P$, where $\hat{x}_i$ is known to be an optimal response to $\hat{s}_i$ under the agent’s true objective function $F$. We denote the set of all training samples by $\hat{\Xi}_N := \{\hat{x}_i\}_{i \leq N}$.

Loosely speaking, this paper addresses the question of how the observer can learn the agent’s preferences on the basis of $\hat{\Xi}_N$. More precisely, we seek a data-driven solution for the inverse optimization problem, i.e., a solution $\hat{\theta}_N \in \Theta$ constructed from the training dataset $\hat{\Xi}_N$ that is near-optimal in (2). Throughout the paper, we reserve the superscript $\cdot^N$ for random objects that depend on the training data and are therefore governed by the product distribution $P^N$. 
Example 3 (Sample Average Approximation). A simple yet powerful approach for constructing data-driven solutions is to approximate $P$ by the empirical distribution on the training dataset, that is,

$$
\hat{P}_N := \frac{1}{N} \sum_{i=1}^{N} \delta_{\xi_i}.
$$

This gives rise to the sample average approximation (SAA) problem

$$
\min_{\theta \in \Theta} \rho^{\hat{P}_N}(\ell_\theta).
$$

Any minimizer of (8) then constitutes a data-driven solution $\hat{\theta}_N$ for (2).

In the remainder of the paper we will propose different data-driven solutions $\hat{\theta}_N$ for (2) that admit rigorous out-of-sample performance guarantees. The out-of-sample risk of a data-driven solution $\hat{\theta}_N$ is defined as $\rho^P(\ell_{\hat{\theta}_N})$, that is, the risk of the loss $\ell_{\hat{\theta}_N}(s, x)$ evaluated under the true distribution $P$ of a test sample $(s, x)$ that is independent of the training samples. We emphasize that $\rho^P(\ell_{\hat{\theta}_N})$ inherits the dependence on the training data from the loss function $\ell_{\hat{\theta}_N}$ and thus constitutes a random variable. Any performance guarantee involves a certificate $\hat{J}_N \geq 0$, which constitutes a (hopefully small) upper confidence bound on the out-of-sample risk. We stress that $\hat{J}_N$ may depend on the training data. The concept of an out-of-sample performance guarantee is formalized in the following definition.

**Definition 2.4** (Out-of-Sample Performance Guarantee). We say that a data-driven solution $\hat{\theta}_N$ of the inverse optimization problem (2) enjoys an out-of-sample performance guarantee at confidence level $\beta \in [0, 1]$ if there exists a certificate $\hat{J}_N$ such that

$$
P^N [\rho^P(\ell_{\hat{\theta}_N}) \leq \hat{J}_N] \geq 1 - \beta.
$$

**Example 4** (Suboptimality Guarantee). Assume that $\ell_\theta(s, x)$ is defined as in (3a) and that the performance guarantee (9) holds for some data-driven solution $\hat{\theta}_N$. Then, the actual response $x$ to a test signal $s$ is near-optimal under the estimated objective function $F_{\hat{\theta}_N}$. Formally, the certainty equivalent suboptimality of $x$ under the objective function $F_{\hat{\theta}_N}$ is bounded above by $\hat{J}_N$ with confidence $1 - \beta$.

**Example 5** (Predictibility Guarantee). Assume that $\ell_\theta^P(s, x)$ is defined as in (3c) and that the performance guarantee (9) holds for some data-driven solution $\hat{\theta}_N$. Then, the responses to a test signal $s$ predicted by $\hat{\theta}_N$ are close to $x$, the best response to $s$ under the unknown objective function $F$. Formally, the certainty equivalent distance between $X_{\hat{\theta}_N}(s)$ and $x$ is bounded above by $\hat{J}_N$ with confidence $1 - \beta$.

**Remark 2.5** (Identifiability Guarantee). In addition to the suboptimality and predictability guarantees, the observer might seek a confidence bound of the form $P^N [\|\theta^* - \hat{\theta}_N\| \leq \hat{J}_N] \geq 1 - \beta$, which would guarantee that the data-driven solution $\hat{\theta}_N$ closely approximates an exact minimizer $\theta^*$ of the inverse optimization problem (2). Note that this bound differs structurally from (9) as it cannot be expressed in terms of a loss function $\ell_\theta(s, x)$. A more fundamental problem is that $\theta^*$ may not be unique. Indeed, any two candidate objective functions $F_\theta$ and $F_\theta'$ that are related through a strictly increasing transformation imply the same preferences. Thus, $\theta$ and $\theta'$ are indistinguishable in any inverse optimization problem. Unlike the identifiability guarantee, the suboptimality and predictability guarantees are always available.

3. Distributionally Robust Inverse Optimization

In this paper we advocate a distributionally robust approach to construct data-driven solutions for the inverse optimization problem (2). To this end, we first design an ambiguity set $\hat{P}$ that contains all
distributions of the signal-response pairs $\xi \in \Xi$ that could have generated the training data $\hat{\Xi}_N$ with sufficiently high confidence. Next, we replace the risk of $\ell_\theta$ under the unknown crisp distribution $P$ with the worst-case risk of $\ell_\theta$, where the worst case is taken over all distributions $Q \in \hat{\mathcal{P}}$. This gives rise to the distributionally robust inverse optimization problem

$$\text{(10)} \quad \minimize_{\theta \in \Theta} \sup_{Q \in \hat{\mathcal{P}}} \rho^Q(\ell_\theta).$$

Any minimizer of (10) then constitutes a data-driven solution $\hat{\theta}_N$ for (2). This data-driven solution may offer powerful performance guarantees if we construct the ambiguity set judiciously, e.g., if we define $\hat{\mathcal{P}}$ as a ball in the space of probability distributions with respect to the Wasserstein metric.

**Definition 3.1** (Wasserstein Metric). For any $p \geq 1$ we let $\mathcal{M}^p(\Xi)$ be the space of all probability distributions $Q$ supported on $\Xi$ with $\mathbb{E}^Q(\|\xi\|_p^p) = \int_{\Xi} \|\xi\|_p^p Q(d\xi) < \infty$. The $p$-Wasserstein distance between two distributions $Q_1, Q_2 \in \mathcal{M}^p(\Xi)$ is defined as

$$W_p(Q_1, Q_2) := \inf \left\{ \left( \int_{\Xi} \|\xi_1 - \xi_2\|_p^p \Pi(d\xi_1, d\xi_2) \right)^{1/p} : \Pi \text{ is a joint distribution of } \xi_1 \text{ and } \xi_2 \right\}.$$

The Wasserstein distance $W_p(Q_1, Q_2)$ can be viewed as the minimum cost for moving the distribution $Q_1$ to $Q_2$, where the cost of moving a unit mass from $\xi_1$ to $\xi_2$ amounts to $\|\xi_1 - \xi_2\|_p^p$. The joint distribution $\Pi$ of $\xi_1$ and $\xi_2$ is therefore naturally interpreted as a mass transportation plan.

In the sequel we let $\hat{\mathcal{P}}$ be a $p$-Wasserstein ball in $\mathcal{M}^p(\Xi)$ centered at the empirical distribution on the training dataset; see (7). Specifically, we define the $p$-Wasserstein ball of radius $\varepsilon$ around $\hat{P}_N$ as

$$\mathbb{B}_p^\varepsilon(\hat{P}_N) := \left\{ Q \in \mathcal{M}^p(\Xi) : W_p(Q, \hat{P}_N) \leq \varepsilon \right\}.$$  

The idea to model ambiguity through Wasserstein balls originates from [31]; see also [30]. Note that for $\varepsilon = 0$ the Wasserstein ball $\mathbb{B}_p^\varepsilon(\hat{P}_N)$ shrinks to the singleton set that contains only the empirical distribution. In this case, the distributionally robust inverse optimization problem (10) reduces to the SAA problem (8). In the remainder we will use $\hat{J}_N(\varepsilon)$ and $\hat{\theta}_N(\varepsilon)$ to denote the optimal value (i.e., the worst-case loss) and an optimal solution of problem (10) with $\hat{\mathcal{P}} = \mathbb{B}_p^\varepsilon(\hat{P}_N)$, respectively.

If the unknown distribution $P$ of the signal-response pairs satisfies a popular light tail assumption, then the data-driven solution $\hat{\theta}_N(\varepsilon)$ offers attractive out-of-sample performance guarantees.

**Assumption 3.2** (Light-Tailed Distribution). There exists a light tail exponent $a > 1$ with

$$A := \mathbb{E}^P \left[ \exp(\|\xi\|^a) \right] = \int_{\Xi} \exp(\|\xi\|^a) P(d\xi) < \infty.$$

Assumption 3.2 essentially requires the tail of the distribution $P$ to decay at an exponential rate, which holds trivially, for instance, if $\Xi$ is compact. The following recent measure concentration result asserts that the unknown true distribution $P$ has Wasserstein distance larger than $\varepsilon$ from the empirical distribution $\hat{P}_N$ with a probability that is exponentially small in the number of training samples $N$ and $\varepsilon$. This bound provides the basis for establishing powerful out-of-sample guarantees.

**Theorem 3.3** (Measure Concentration [19, Theorem 2]). If Assumption 3.2 holds, we have

$$\mathbb{P}^N \left[ W_p(P, \hat{P}_N) \geq \varepsilon \right] \leq c_1 \left( \exp \left( -c_2 N \varepsilon^{(m+n)/p} \right) I_{\{\varepsilon \leq \varepsilon_0\}} + \exp \left( -c_3 N \varepsilon^a/p \right) I_{\{\varepsilon > \varepsilon_0\}} \right),$$

where $c_1, c_2, c_3$ are positive constants.
for all $N \geq 1$ and $\varepsilon > 0$, where $\varepsilon_0 := (c_3/c_2)^{p/(m+n-a)}$, while $c_1$, $c_2$, and $c_3$ are positive constants that depend on the distribution $P$ only through $a$, $A$, $m$, $n$ and $p$.

Theorem 3.3 provides an a priori estimate of the probability that the unknown data-generating distribution $P$ resides outside of the Wasserstein ball $B_p(\hat{P}_N)$. This estimate immediately translates to an out-of-sample performance guarantee for the data-driven solution $\hat{\theta}(\varepsilon)$ obtained from (10).

**Corollary 3.4 (Out-of-Sample Performance Guarantee).** The out-of-sample risk of $\hat{\theta}_N(\varepsilon)$ satisfies
\[
P^N \left[ P \left( \ell_{\hat{\theta}_N(\varepsilon)} \leq \hat{J}_N(\varepsilon) \right) \right] \geq 1 - \beta
\]
for every confidence level $\beta \in (0, 1)$ and for every Wasserstein radius $\varepsilon \geq \varepsilon_N(\beta)$, where
\[
\varepsilon_N(\beta) := \left( \frac{\log(c_1\beta^{-1})}{c_2N} \right)^{p/(m+n)} \mathbb{I}_{\{N \geq \frac{\log(c_1\beta^{-1})}{c_20^p(m+n)}\}} + \left( \frac{\log(c_1\beta^{-1})}{c_2N} \right)^{p/a} \mathbb{I}_{\{N < \frac{\log(c_1\beta^{-1})}{c_20^p(m+n)}\}}.
\]

**Proof.** Equating the right-hand side of (11) to $\beta$ and solving for $\varepsilon$ yields (12). By construction, the $p$-Wasserstein ball of radius $\varepsilon \geq \varepsilon_N(\beta)$ centered at $\hat{P}_N$ thus contains the true distribution $P$ with confidence at least $1 - \beta$. The claim then follows because $P \left( \ell_{\hat{\theta}_N(\varepsilon)} \leq \hat{J}_N(\varepsilon) \right)$ whenever $P \in B_p(\hat{P}_N)$. $\square$

Besides offering powerful out-of-sample guarantees, our distributionally robust approach to inverse optimization is attractive for the following reasons. First, problem (10) reduces to a tractable conic program for popular classes of candidate objective functions $F$ if the suboptimality criterion (3a) is used to quantify loss; see Sections 4 and 5. Furthermore, our distributionally robust approach naturally generalizes to situations where the observer has imperfect information, that is, when the agent’s true objective function may not be contained in $F$, when the agent suffers from bounded rationality or when the observed signal-response pairs are corrupted by measurement noise; see Section 6.

In summary, we have developed a unifying framework for constructing inverse optimization models, which are obtained by combining different loss functions, risk measures and ambiguity sets. Instantiating problem (10) with the loss function (3b), the essential supremum risk measure and a degenerate Wasserstein ball of radius zero, for example, results in an inverse optimization model studied in [11]. Alternatively, instantiating (10) with the loss function (3c), the expected value risk measure and a Wasserstein ball of radius zero leads to an inverse optimization model propagated in [7]. In this paper we focus on the new loss function (3a) because (3c) is known to cause computational intractability [7], while (3b) lacks an intuitive interpretation and primarily serves as an approximation for (3a) or (3c); see Example 1. However, there is no computational advantage in approximating (3a), and Proposition 2.2 suggests that (3a) provides a stronger tractable approximation for (3c) than (3b).

### 4. Linear and Piecewise Linear Objective Functions

On the one hand, the class $F$ of candidate models should be rich enough to contain the agent’s unknown true objective function $F$. On the other hand, $F$ should be small enough to ensure tractability of the distributionally robust inverse optimization problem (10) and to prevent degeneracy of its optimal solutions. A particular class $F$ that strikes this delicate balance and proves useful in many applications is the family of linear objective functions $F_\theta(s, x) := \langle \theta, x \rangle$, where $\theta \in \mathbb{R}^n$ ranges over the search space
\[
\Theta := \{ \theta \in \mathbb{R}^n : \|\theta\|_\infty = 1 \}.
\]
Note that the normalization $\|\theta\|_\infty = 1$ is non-restrictive because the objective functions corresponding to $\theta$ and $\kappa \theta$ imply the same preferences for any coefficient vector $\theta \neq 0$ and scaling factor $\kappa > 0$. We emphasize that we are free to define $\Theta$ as the unit sphere induced by any norm on $\mathbb{R}^n$. However, the $\infty$-norm stands out from a computational perspective. While all norm spheres are non-convex and therefore a priori unattractive as search spaces, the $\infty$-norm sphere decomposes into $2n$ polytopes—one for each facet. This polyhedral decomposition property allows us to optimize efficiently over $\Theta$.\(^1\)

When focusing on linear candidate objective functions, the loss function (3a) reduces to

$$
\ell_\theta(s,x) = \langle \theta, x \rangle - \min_{y \in \mathcal{X}(s)} \langle \theta, y \rangle = \max_{y \in \mathcal{X}(s)} \langle \theta, x - y \rangle,
$$

which is positive homogeneous and subadditive in $\theta$. The tractability results to be established below rely on the additional assumption that the signal space $\mathcal{S}$ and the feasible set $\mathcal{X}(s)$ are conic representable,

$$
\mathcal{S} := \{ s \in \mathbb{R}^n : Cs \succeq_C d \}, \quad \mathcal{X}(s) := \{ x \in \mathbb{R}^n : W x \succeq_K Hs + h \} \forall s \in \mathcal{S},
$$

where the relations ‘$\succeq_C$’ and ‘$\succeq_K$’ represent conic inequalities with respect to some proper convex cones $\mathcal{C}$ and $\mathcal{K}$ of appropriate dimensions, respectively. By construction, the loss function $\ell_\theta(s,x)$ is therefore concave in $(s,x)$ for every fixed $\theta$, see, e.g., [13, Section 3.2.5]. Moreover, the uncertainty set $\Xi$ comprising all possible signal-response pairs reduces to

$$
\Xi = \{(s,x) \in \mathbb{R}^n \times \mathbb{R}^n : Cs \succeq_C d, \ W x \succeq_K H s + h \}.
$$

We will henceforth assume that the convex set $\Xi$ is compact and has non-empty relative interior. We are now ready to state our first tractability result for the class of linear candidate objective functions.

**Theorem 4.1** (Tractability of (10) for Linear Candidate Objectives). Assume that $\mathcal{F}$ represents the class of linear candidate objectives, the loss function is defined as in (13), the signal space $\mathcal{S}$ and the feasible set $\mathcal{X}(s)$ are defined as in (14), and risk is measured by the CVaR at level $\alpha \in (0,1]$. Then, the distributionally robust inverse optimization problem (10) over the 1-Wasserstein ball is equivalent to the finite conic program\(^2\)

$$
\begin{align*}
\text{minimize} \quad & \tau + \frac{1}{\alpha} \left( \varepsilon \lambda + \frac{1}{N} \sum_{i=1}^N r_i \right) \\
\text{subject to} \quad & \theta \in \Theta, \quad \lambda \in \mathbb{R}^+ \cup \{0\}, \quad \tau, r_i \in \mathbb{R}, \quad \phi_{i1}, \phi_{i2} \in \mathcal{C}^*, \quad \mu_{i1}, \mu_{i2}, \gamma_i \in \mathcal{K}^* \forall i \leq N \\
& \langle C\hat{s}_i - d, \phi_{i1} \rangle + \langle W x_i - H\hat{s}_i - h, \mu_{i1} + \gamma_i \rangle \leq r_i + \tau \forall i \leq N \\
& \langle C\hat{s}_i - d, \phi_{i2} \rangle + \langle W x_i - H\hat{s}_i - h, \mu_{i2} \rangle \leq r_i, \quad \theta = W^T \mu_i \forall i \leq N \\
& \left\| \begin{pmatrix} C^T \phi_{i1} - H^T (\mu_{i1} + \gamma_i) \\ W^T (\mu_{i1} + \gamma_i) \end{pmatrix} \right\|_s \leq \lambda, \quad \left\| \begin{pmatrix} C^T \phi_{i2} - H^T \mu_{i2} \\ W^T \mu_{i2} \end{pmatrix} \right\|_s \leq \lambda \forall i \leq N.
\end{align*}
$$

**Proof.** By the definition of CVaR, the objective function of (10) can be expressed as

$$
\sup_{Q \in \mathcal{B}(\mathcal{F}_N)} \rho^Q(\ell_\theta) = \sup_{Q \in \mathcal{B}(\mathcal{F}_N)} \inf_{\tau} \tau + \frac{1}{\alpha} \mathbb{E}^P \left[ \max \{ \ell_\theta(s,x) - \tau, 0 \} \right]
$$

$$
= \inf_{\tau} \tau + \frac{1}{\alpha} \sup_{Q \in \mathcal{B}(\mathcal{F}_N)} \mathbb{E}^P \left[ \max \{ \ell_\theta(s,x) - \tau, 0 \} \right].
$$

\(^1\)If $\theta$ is known to be non-negative, it is convenient to set $\Theta := \{ \theta \in \mathbb{R}^n : \| \theta \|_1 = 1, \theta \geq 0 \}$, which is a convex polytope.

\(^2\)Strictly speaking, (16) constitutes a family of $2n$ finite conic programs because $\Theta$ is non-convex but decomposes into $2n$ convex polytopes.
The interchange of the maximization over \( Q \) and the minimization over \( \tau \) in the second line is justified by Sion’s minimax theorem [34], which applies because the Wasserstein ball \( \mathbb{B}_1^P(\hat{P}_N) \) contains only distributions supported on the compact set \( \Xi \) and is therefore weakly compact [5, Theorem 15.11].

The subordinate worst-case expectation problem in the second line of (17) constitutes a semi-infinite linear program. As the corresponding integrand is given by the maximum of

\[
\max \langle \theta, x - y \rangle - \tau - \lambda \| (s, x) - (\hat{s}_i, \hat{x}_i) \| \leq r_i \quad \forall i \leq N
\]

the worst-case expectation problem admits a strong dual semi-infinite linear program of the form

\[
\inf_{\lambda \geq 0, r_i} \varepsilon \lambda + \frac{1}{N} \sum_{i=1}^{N} r_i \\
\text{s.t.} \sup_{(s, x) \in \Xi} \sup_{y \in \Xi(s)} \langle \theta, x - y \rangle - \tau - \lambda \| (s, x) - (\hat{s}_i, \hat{x}_i) \| \leq r_i \quad \forall i \leq N
\]

(18)

see Theorem 4.3 in [27] for a detailed derivation of (18) for more general integrands. By the definitions of \( \Xi(s) \) and \( \Xi \) in (14) and (15), respectively, the \( i \)-th member of the first constraint group in (18) holds if and only if the optimal value of the conic program

\[
\sup_{s, x, y} \langle \theta, x - y \rangle - \tau - \lambda \| (s, x) - (\hat{s}_i, \hat{x}_i) \| \\
\text{s.t.} \quad C s \succeq d, \quad W x \succeq_{\mathcal{K}} H s + h, \quad W y \succeq_{\mathcal{K}} H s + h
\]

is smaller or equal to \( r_i \). The dual of this conic program is given by

\[
\inf \langle C \hat{s}_i - d, \phi_{i1} \rangle + \langle W \hat{x}_i - H \hat{s}_i - h, \mu_{i1} + \gamma_i \rangle - \tau \\
\text{s.t.} \quad \phi_{i1} \in C^*, \quad \mu_{i1}, \gamma_i \in \mathcal{K}^* \\
\left\| \begin{pmatrix} C^T \phi_{i1} - H^T (\mu_{i1} + \gamma_i) \\ W^T (\mu_{i1} + \gamma_i) \end{pmatrix} \right\|_* \leq \lambda, \quad \theta = W^T \gamma_i,
\]

and strong duality holds because the uncertainty set \( \Xi \) is assumed to have a non-empty relative interior. Thus, the \( i \)-th member of the first constraint group in (18) holds if and only if there exist \( \phi_{i1} \in C^* \) and \( \mu_{i1}, \gamma_i \in \mathcal{K}^* \) such that \( \theta = W^T \gamma_i \),

\[
\langle C \hat{s}_i - d, \phi_{i1} \rangle + \langle W \hat{x}_i - H \hat{s}_i - h, \mu_{i1} + \gamma_i \rangle \leq r_i + \tau \quad \text{and} \quad \left\| \begin{pmatrix} C^T \phi_{i1} - H^T (\mu_{i1} + \gamma_i) \\ W^T (\mu_{i1} + \gamma_i) \end{pmatrix} \right\|_* \leq \lambda.
\]

A similar reasoning shows that the \( i \)-th member of the second constraint group in (18) holds if and only if there exist \( \phi_{i2} \in C^* \) and \( \mu_{i2}, \gamma_{i2} \in \mathcal{K}^* \) such that

\[
\langle C \hat{s}_i - d, \phi_{i2} \rangle + \langle W \hat{x}_i - H \hat{s}_i - h, \mu_{i2} \rangle \leq r_i \quad \text{and} \quad \left\| \begin{pmatrix} C^T \phi_{i2} - H^T \mu_{i2} \\ W^T \mu_{i2} \end{pmatrix} \right\|_* \leq \lambda.
\]

3If \( (\hat{s}_i, \hat{x}_i) \in \Xi \), this constraint simplifies to \( r_i \geq 0 \). Situations where \( (\hat{s}_i, \hat{x}_i) \notin \Xi \) will be considered in Section 6.
In summary, the worst-case expectation in the second line of (17) thus coincides with the optimal value of the finite conic program

$$\inf \epsilon \lambda + \frac{1}{N} \sum_{i=1}^{N} r_i$$

s.t. $$\lambda \in \mathbb{R}_+, \ r_i \in \mathbb{R}, \ \phi_{i1}, \phi_{i2} \in C^*, \ \mu_{i1}, \mu_{i2}, \gamma_i \in K^* \ \forall i \leq N$$

$$\langle C\delta_i - d, \phi_{i1} \rangle + \langle W\delta_i - H\delta_i - h, \mu_{i1} + \gamma_i \rangle \leq r_i + \tau \ \forall i \leq N$$

$$\langle C\delta_i - d, \phi_{i2} \rangle + \langle W\delta_i - H\delta_i - h, \mu_{i2} \rangle \leq \tau, \ \theta = W^T \gamma_i \ \forall i \leq N$$

$$\left\| \left( \frac{C^T \phi_{i1} - H^T (\mu_{i1} + \gamma_i)}{W^T (\mu_{i1} + \gamma_i)} \right) \right\| \leq \lambda, \ \left\| \left( \frac{C^T \phi_{i2} - H^T \mu_{i2}}{W^T \mu_{i2}} \right) \right\| \leq \lambda \ \forall i \leq N.$$ 

The claim then follows by substituting this conic program into (17). \hfill \Box

For stress test experiments it is often desirable to know the extremal distribution that achieves the worst-case risk in (10). The following theorem shows that this extremal distribution can be constructed systematically for any fixed $$\theta \in \Theta$$ by solving finite convex optimization problem akin to (16).

**Theorem 4.2.** Under the assumptions of Theorem 4.1, the worst-case risk in (16) corresponding to a fixed $$\theta \in \Theta$$ coincides with the optimal value of a finite convex program, i.e.,

$$\sup_{q \in \mathbb{B}^n(\tilde{p}_N)} \text{CVaR}^Q_{\alpha}(\ell_\theta) = \max \frac{1}{\alpha N} \sum_{i=1}^{N} \pi_{i1} \ell_{\theta} \left( \frac{p_{i1}}{\pi_{i1}}, \frac{q_{i1}}{\pi_{i1}} \right)$$

s.t. $$\pi_{ij} \in \mathbb{R}_+, \ p_{ij} \in \mathbb{R}^m, \ q_{ij} \in \mathbb{R}^n \ \forall i \leq N, \ j \leq 2$$

$$\frac{p_{i1}}{\pi_{i1}} \in \mathbb{S}, \ \frac{q_{i1}}{\pi_{i1}} \in \mathbb{X} \left( \frac{p_{i1}}{\pi_{i1}} \right) \ \forall i \leq N, \ j \leq 2$$

$$\pi_{i1}, \pi_{i2} = 1, \ \frac{1}{N} \sum_{i=1}^{N} \pi_{i1} = \alpha \ \forall i \leq N$$

$$\frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{2} \pi_{ij} \left\| \left( \frac{p_{ij}}{\pi_{ij}} - \tilde{s}_i \right) \right\| \leq \varepsilon.$$ 

(19)

For any optimal solution $$\{ \pi^*_i, p^*_{ij}, q^*_{ij} \}$$ of this convex program, the discrete distribution

$$Q^* := \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{2} \pi^*_i \delta_{\xi^*_ij} \quad \text{with} \quad \xi^*_ij := \left( \frac{p^*_{ij}}{\pi^*_ij}, \frac{q^*_{ij}}{\pi^*_ij} \right)^T$$

belongs to the Wasserstein ball $$\mathbb{B}^1(\tilde{p}_N)$$ and attains the supremum on the left hand side of (19).

**Proof.** As the loss function (13) is proper and jointly concave in $$x$$ and $$s$$, we can use a similar reasoning as in [27, Theorem 4.5] to show that the convex program on the right hand side of (19) coincides with the strong dual of (16) for any fixed $$\theta \in \Theta$$. This convex program is solvable because $$\Xi$$ was assumed to be compact, and therefore the distribution $$Q^*$$ is well-defined; see also [27, Corollary 4.7]. It remains to be shown that $$\text{CVaR}^Q_{\alpha}(\ell_\theta)$$ is no smaller than (16). Indeed, by the definition of $$\text{CVaR}$$ we have

$$\text{CVaR}^Q_{\alpha}(\ell_\theta) = \inf_{\tau \in \mathbb{R}} \tau + \frac{1}{\alpha N} \sum_{i=1}^{N} \sum_{j=1}^{2} \pi^*_{ij} \max \left\{ \ell_{\theta} \left( \frac{p^*_{ij}}{\pi^*_ij}, \frac{q^*_{ij}}{\pi^*_ij} \right) - \tau, 0 \right\}$$

$$= \sup_{0 \leq \gamma_{ij} \leq \pi^*_{ij}} \left\{ \frac{1}{\alpha N} \sum_{i=1}^{N} \sum_{j=1}^{2} \gamma_{ij} \ell_{\theta} \left( \frac{p^*_{ij}}{\pi^*_ij}, \frac{q^*_{ij}}{\pi^*_ij} \right) : \alpha = \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{2} \gamma_{ij} \right\}$$
\[
\frac{1}{
\alpha N \sum_{i=1}^{N} \pi_{i1}^* \ell_\theta \left( \frac{p_{i1}^*}{\gamma_{i1}^*}, \frac{q_{i1}^*}{\pi_{i1}^*} \right) \}
\leq \sup_{Q \in \mathbb{B}_2(F_N)} \text{CVaR}_Q^\alpha(\ell_\theta).
\]

Here, the second equality follows from strong linear programming duality, while the inequality follows from the feasibility of the solution \(\gamma_{i1} = \pi_{i1}^*\) and \(\gamma_{i2} = 0\) for \(i \leq N\). \(\square\)

In the remainder of this section we investigate to what extent our findings for linear candidate objective functions generalize to the richer class of piecewise linear objective functions with \(J\) pieces. Specifically, we focus on candidate objective functions representable as

\[F_\theta(s, x) := \max_{i \leq J} \langle \theta_i, x \rangle,\]

where \(\theta := \{\theta_j\}_{j \leq J}\) ranges over a judiciously chosen search space \(\Theta \subset \mathbb{R}^{nJ}\). In order to avoid trivial solutions for the inverse optimization problem (10), none of the linear pieces may be flat. This requirement can be enforced through the normalization constraints \(\|\theta_j\|_\infty \geq 1\) \(\forall j \leq J\). Note that these inequality constraints allow some linear pieces to have significantly steeper slopes than others, which is desirable as it increases the expressiveness of the underlying class of candidate objectives. Unfortunately, however, the non-convex search space \(\Theta\) implied by these normalization constraints constitutes an irreducible union of \((2n)^J\) half spaces. Thus, optimizing over \(\Theta\) may be time-consuming if \(J\) is large. A computationally more attractive alternative is to define \(\Theta\) as the union of polynomially many half spaces in \(\mathbb{R}^n\).

When focusing on piecewise linear candidate objective functions, the loss function (3a) reduces to

\[
\ell_\theta(s, x) = \max_{j \leq J} \langle \theta_j, x \rangle - \min_{y \in X(s)} \max_{k \leq J} \langle \theta_k, y \rangle = \max_{j \leq J} \min_{y \in X(s)} \langle \theta_j, x \rangle - \langle \theta_k, y \rangle.
\]

By construction, \(\ell_\theta(s, x)\) is therefore given by the pointwise maximum of \(J\) functions that are proper and concave in \((s, x)\) for every fixed \(\theta\). Next, we prove a finite reduction theorem for the inverse optimization problem (10) over the class of piecewise linear candidate objective functions.

**Theorem 4.3** (Finite Reduction of (10) for Piecewise Linear Candidate Objectives). Assume that \(F\) represents the class of piecewise linear candidate objectives, the loss function is defined as in (20), the signal space \(S\) and the feasible set \(X(s)\) are defined as in (14), and risk is measured by the CVaR at level \(\alpha \in (0, 1]\). Then, the distributionally robust inverse optimization problem (10) over the 1-Wasserstein ball is equivalent to the finite nonlinear program

\[
\begin{align*}
\text{minimize} & \quad \tau + \frac{1}{2} \left( \varepsilon \lambda + \frac{1}{N} \sum_{i=1}^{N} s_i \right) \\
\text{subject to} & \quad \theta \in \Theta, \ \lambda \in \mathbb{R}^+, \ \tau, r_i \in \mathbb{R}, \ \phi_{ij1}, \phi_{ij2} \in C^*, \ \mu_{ij1}, \mu_{ij2}, \gamma_{ij1}, \gamma_{ij2} \in K^* \quad \forall j \leq J, \ \forall i \leq N \\
& \quad \langle C \tilde{s}_i - d, \phi_{ij1} \rangle + \langle W \tilde{x}_i - H \tilde{s}_i - h, \mu_{ij1} + \gamma_{ij1} \rangle \leq r_i + \tau \quad \forall j \leq J, \ \forall i \leq N \\
& \quad \langle C \tilde{s}_i - d, \phi_{ij2} \rangle + \langle W \tilde{x}_i - H \tilde{s}_i - h, \mu_{ij2} \rangle \leq r_i \quad \forall i \leq N \\
& \quad \left\| \begin{pmatrix} C \phi_{ij1} - H \phi_{ij2} \\
W \mu_{ij1} + \theta_j \end{pmatrix} \right\|_* \leq \lambda \quad \forall j \leq J, \ \forall i \leq N \\
& \quad W \gamma_{ij1} \in \text{Conv}(\theta_1, \ldots, \theta_J) \quad \forall j \leq J, \ \forall i \leq N,
\end{align*}
\]

where \(\text{Conv}(\theta_1, \ldots, \theta_J)\) denotes the convex hull spanned by the vectors \(\theta_1, \ldots, \theta_J\).

---

4Our treatment of piecewise linear functions also accounts for piecewise affine functions because we are free, for instance, to constrain the first component of \(x\) to a constant.
Proof. The proof widely parallels that of Theorem 4.1. Evidently, the objective function of the inverse optimization problem (10) still coincides with (17), and the main burden of the proof is to derive a finite-dimensional reformulation of the subordinate worst-case expectation problem

\[
\text{sup}_{Q \in \mathbb{P}^r_{\mathcal{Y}}} \mathbb{E}^P \left[ \max \{ f_\theta(s, x) - \tau, 0 \} \right].
\]

Note that the corresponding integrand is given by the maximum of 0 and \( \max_{y \in \Xi(s)} \min_{k \leq J} \langle \theta_j, x \rangle - \langle \theta_j, y \rangle - \tau \) for \( j \leq J \). As each of these \( J + 1 \) functions is proper and concave in \( (s, x) \), we can proceed as in [27, Theorem 4.3] to show that this worst-case expectation problem has the same optimal value as its strong semi-infinite dual, which is given by

\[
\inf_{\lambda \geq 0, r_i} \varepsilon \lambda + \frac{1}{N} \sum_{i=1}^N r_i \quad \text{s.t.} \quad \sup_{(s,x) \in \Xi, y \in \Xi(s)} \min_{k \leq J} \langle \theta_j, x \rangle - \langle \theta_j, y \rangle - \tau - \lambda \| (s, x) - (\hat{s}_i, \hat{x}_i) \| \leq r_i \quad \forall i \leq N, \forall j \leq J
\]

\[
\sup_{(s,x) \in \Xi, y \in \Xi(s)} -\lambda \| (s, x) - (\hat{s}_i, \hat{x}_i) \| \leq r_i \quad \forall i \leq N.
\]

Next, observe that the \((i, j)\)-th member of the first constraint group in (23) holds if and only if

\[
\sup_{(s,x) \in \Xi, y \in \Xi(s)} \min_{k \leq J} \langle \theta_j, x \rangle - \langle \theta_j, y \rangle - \tau - \lambda \| (s, x) - (\hat{s}_i, \hat{x}_i) \| \leq r_i
\]

\[
\iff \sup_{(s,x) \in \Xi, y \in \Xi(s)} \inf_{\nu \in \Delta^J} \langle \theta_j, x \rangle - \sum_{k \leq J} \nu_k \langle \theta_j, y \rangle - \tau - \lambda \| (s, x) - (\hat{s}_i, \hat{x}_i) \| \leq r_i
\]

\[
\iff \inf_{\nu \in \Delta^J} \sup_{(s,x) \in \Xi, y \in \Xi(s)} \langle \theta_j, x \rangle - \sum_{k \leq J} \nu_k \langle \theta_j, y \rangle - \tau - \lambda \| (s, x) - (\hat{s}_i, \hat{x}_i) \| \leq r_i
\]

\[
\iff \exists \nu \in \Delta^J : \sup_{(s,x) \in \Xi, y \in \Xi(s)} \langle \theta_j, x \rangle - \sum_{k \leq J} \nu_k \langle \theta_j, y \rangle - \tau - \lambda \| (s, x) - (\hat{s}_i, \hat{x}_i) \| \leq r_i,
\]

where \( \Delta^J := \{ \nu \in \mathbb{R}^J_+ : \langle e, \nu \rangle = 1 \} \) stands for the probability simplex in \( \mathbb{R}^J \). The interchange of the supremum and the infimum operators in the second equivalence is justified by the classical minimax theorem. Recalling the definitions of \( \Xi(s) \) and \( \Xi \) in (14) and (15), respectively, we can employ standard conic duality arguments as in the proof of Theorem 4.1 to show that the \((i, j)\)-th member of the first constraint group in (23) holds if and only if there exist \( \phi_{ij1} \in \mathcal{C}^r, \mu_{ij1}, \gamma_{ij1} \in \mathcal{K}^r \) with

\[
\langle C\hat{s}_i - d, \phi_{ij1} \rangle + \langle W\hat{x}_i - H\hat{s}_i - h, \mu_{ij1} + \gamma_{ij1} \rangle \leq r_i + \tau \quad \forall j \leq J, \forall i \leq N,
\]

\[
\left\| \begin{pmatrix} C^* \phi_{ij1} - H^T (\mu_{ij1} + \gamma_{ij1}) \\ W^T \mu_{ij1} + \theta \end{pmatrix} \right\|_* \leq \lambda, \quad \forall j \leq J, \forall i \leq N.
\]

A similar reasoning shows that the \( i \)-th member of the second constraint group in (23) holds if and only if there exist \( \phi_{i2} \in \mathcal{C}^r, \mu_{i2}, \gamma_{i2} \in \mathcal{K}^r \) such that\(^5\)

\[
\langle C\hat{s}_i - d, \phi_{i2} \rangle + \langle W\hat{x}_i - H\hat{s}_i - h, \mu_{i2} \rangle \leq r_i, \quad \left\| \begin{pmatrix} C^* \phi_{i2} - H^T \mu_{i2} \end{pmatrix} \right\|_* \leq \lambda \quad \forall i \leq N.
\]

\(^5\)If \( (\hat{s}_i, \hat{x}_i) \in \Xi \), this constraint simplifies to \( r_i \geq 0 \). Situations where \( (\hat{s}_i, \hat{x}_i) \not\in \Xi \) will be considered in Section 6.
As in Section 4, we assume that the signal space \( S \) is a conic program and is therefore tractable for common choices of the cones \( C \). Specifically, we assume again that (14) holds. In this setting the agent’s decision problem (1) constitutes the class of quadratic candidate objectives, the loss function is defined as in (Intractability of (10) for Quadratic Candidate Objectives) Theorem 5.1.

Theorem 5.1. (Intractability of (10) for Quadratic Candidate Objectives). Assume that \( F \) represents the class of quadratic candidate objectives, the loss function is defined as in (25), and risk is measured by the CVaR at level \( \alpha \in (0, 1] \). Then, evaluating the objective function of (10) for a fixed \( \theta \in \Theta \) is NP-hard even if \( N = 1 \) (there is only a single data point), \( \alpha = 1 \) (the observer is risk-neutral), the signal space \( S \) is a singleton, the feasible set \( \mathcal{X}(s) \) is a polytope independent of \( s \), and \( Q_{xs} = 0 \).

In summary, the worst-case expectation (22) reduces to the optimal value of the finite conic program

\[
\inf \ e\lambda + \frac{1}{N} \sum_{i=1}^{N} r_i \\
\text{s.t.} \quad \lambda \in \mathbb{R}_+, \ r_i \in \mathbb{R}, \ \phi_{ij1}, \phi_{i2} \in \mathcal{C}^*, \ \mu_{ij1}, \mu_{i2}, \gamma_{ij1}, \gamma_{i2} \in \mathcal{K}^* \\
\langle C\hat{s}_i - d, \phi_{ij1} \rangle + \langle W\hat{x}_i - H\hat{s}_i - h, \mu_{ij1} + \gamma_{ij1} \rangle \leq r_i + \tau \\
\forall j \leq J, \ \forall i \leq N \\
\langle C\hat{s}_i - d, \phi_{i2} \rangle + \langle W\hat{x}_i - H\hat{s}_i - h, \mu_{i2} \rangle \leq r_i \\
\forall i \leq N \\
\left\| \left( CT\phi_{ij1} - HT(\mu_{ij1} + \gamma_{ij1}) \right) \right\|_* \leq \lambda, \quad \left\| \left( CT\phi_{i2} - HT\mu_{i2} \right) \right\|_* \leq \lambda \\
\forall j \leq J, \ \forall i \leq N \\
W\gamma_{ij1} \in \text{Conv}(\theta_1, \cdots, \theta_J) \\
\forall j \leq J, \ \forall i \leq N.
\]

The claim then follows from substituting this conic program into (17).

We emphasize that problem (21) is non-convex because it involves convex combinations of the decision variables \( \theta_j \), which must be expressed through bilinear terms in these decision variables and the associated convex weights. Thus, (21) may not be solved efficiently to global optimality. However, it can be solved approximately with a fast block coordinate descent algorithm that keeps either the \( \theta_j \) variables or their convex weights fixed in each iteration and thus solves only convex subproblems.

5. Quadratic Objective Functions

Optimization problems with quadratic objectives abound in control [6], statistics [20], finance [26] and many other application domains. Algorithms for inverse optimization that can learn quadratic objective functions from signal-response pairs are therefore of great practical interest. This motivates us to consider the class \( \mathcal{F} \) of quadratic candidate objective functions of the form

\[
\mathcal{F}_\theta(s, x) := \langle x, Q_{xx} x \rangle + \langle x, Q_{xs} s \rangle + \langle q, x \rangle,
\]

which are encoded by a parameter \( \theta := (Q_{xx}, Q_{xs}, q) \) that ranges over

\[
\Theta := \left\{ \theta = (Q_{xx}, Q_{xs}, q) \in \mathbb{R}^{n \times n} \times \mathbb{R}^{n \times m} \times \mathbb{R}^n : Q_{xx} \succeq I \right\}.
\]

The normalization constraint \( Q_{xx} \succeq I \) implies that \( \mathcal{F}_\theta(s, x) \) is strongly convex in \( x \). When focusing on quadratic candidate objective functions, the loss function (3a) reduces to

\[
\ell_\theta(s, x) = \langle x, Q_{xx} x + Q_{xs} s + q \rangle - \min_{y \in \mathcal{X}(s)} \langle y, Q_{xx} y + Q_{xs} s + q \rangle \\
= \max_{y \in \mathcal{X}(s)} \langle x, Q_{xx} x + Q_{xs} s + q \rangle - \langle y, Q_{xx} y + Q_{xs} s + q \rangle.
\]

As in Section 4, we assume that the signal space \( S \) and the feasible set \( \mathcal{X}(s) \) are conic representable. Specifically, we assume again that (14) holds. In this setting the agent’s decision problem (1) constitutes a conic program and is therefore tractable for common choices of the cones \( \mathcal{C} \) and \( \mathcal{K} \). In contrast, the inverse optimization problem (10) is hard. In fact, it is already hard to evaluate the objective function of (10) for a fixed \( \theta \). As we work with quadratic objectives, throughout this section we use the 2-norm on the signal-response space and the 2-Wasserstein metric to measure distances of distributions.

Theorem 5.1. (Intractability of (10) for Quadratic Candidate Objectives). Assume that \( \mathcal{F} \) represents the class of quadratic candidate objectives, the loss function is defined as in (25), and risk is measured by the CVaR at level \( \alpha \in (0, 1] \). Then, evaluating the objective function of (10) for a fixed \( \theta \in \Theta \) is NP-hard even if \( N = 1 \) (there is only a single data point), \( \alpha = 1 \) (the observer is risk-neutral), the signal space \( S \) is a singleton, the feasible set \( \mathcal{X}(s) \) is a polytope independent of \( s \), and \( Q_{xs} = 0 \).
Proof. The proof relies on a reduction from the NP-hard quadratic maximization problem [28].

**Quadratic Maximization**

**Instance.** A positive definite matrix $Q = Q^T \succeq I$.

**Goal.** Evaluate $\max_{\|x\|_\infty \leq 1} \langle x, Qx \rangle$.

Given an input $Q \succeq I$ to the quadratic maximization problem, we construct an instance of the inverse optimization problem (10) with $N = 1$, $\alpha = 1$, and Wasserstein radius $\epsilon = \sqrt{n}$, where

\[ \hat{s}_1 := 0, \quad \hat{x}_1 := 0, \quad Q_{xx} := Q, \quad Q_{xs} := 0, \quad S := \{0\}, \quad X(s) := \{x \in \mathbb{R}^n : \|x\|_\infty \leq 1\}. \]

Under this parametrization, the objective function of (10) reduces to

\[
\sup_{Q \in \mathbb{B}_2^2(\hat{P}_N)} \rho^Q(\ell_\theta) = \sup_{Q \in \mathbb{B}_2^2(\hat{P}_N)} \mathbb{E}_Q \left[ \max_{y \in X(s)} \langle x, Qxx + Qxss + q \rangle - \langle y, Qxx + Qxss + q \rangle \right]
\]

\[
= \sup_{Q \in \mathbb{B}_2^2(\hat{P}_N)} \mathbb{E}_Q \left[ \max_{y \in X(s)} \langle x, Qx \rangle - \langle y, Qy \rangle \right]
\]

\[
\leq \sup_{s \in S, x \in X(s)} \max_{y \in X(s)} \langle x, Qx \rangle - \langle y, Qy \rangle = \sup_{\|x\|_\infty \leq 1} \langle x, Qx \rangle,
\]

where the inequality in the third line follows from the inclusion $\mathbb{B}_2^2(\hat{P}_N) \subset \mathcal{M}^2(\Xi)$, while the last equality holds because the innermost maximum is attained at $y = 0$. As $(s, x) \in \Xi$ if and only if $s = 0$ and $\|x\|_\infty \leq 1$, we conclude that $(s, x) \in \Xi$ implies $\|x\|_2 \leq \sqrt{n}$ and $\|\langle s, x \rangle\|_2^2 \leq n = \epsilon^2$. Moreover, as the empirical distribution $\hat{P}_N$ coincides with the Dirac point measure at 0, the Wasserstein ball $\mathbb{B}_2^2(\hat{P}_N)$ thus contains all distributions supported on $\Xi$, implying that the inequality in the above expression is in fact an equality. Hence, evaluating the objective function of (10) is tantamount to solving the NP-hard quadratic maximization problem. This observation completes the proof. \[\square\]

**Corollary 5.2** (Intractability of (10) for Signal-Dependent Linear Candidate Objectives). Under the assumptions of Theorem 5.1, evaluating the objective function of (10) for a fixed $\theta \in \Theta$ remains NP-hard if the condition $Q_{xs} = 0$ is replaced with $Q_{xx} = 0$.

**Proof.** The proof is similar to that of Theorem 5.1 and omitted for brevity. \[\square\]

Corollary 5.2 asserts that the inverse optimization problem (10) is intractable even if we focus on linear candidate objectives that may depend on the exogenous signal $s$. This finding contrasts with the tractability Theorem 4.1 for candidate objectives independent of $s$. The intractability results portrayed in Theorem 5.1 and Corollary 5.2 motivate us to devise a safe conic approximation for the inverse optimization problem (10) with quadratic candidate objective functions.

**Theorem 5.3** (Safe Conic Approximation of (10) for Quadratic Candidate Objectives). Assume that $\mathcal{F}$ represents the class of quadratic candidate objectives, the loss function is defined as in (25), the signal space $\mathcal{S}$ and the feasible set $\mathcal{X}(s)$ are defined as in (14), and risk is measured by the CVaR at level $\alpha \in (0, 1]$. Then, the following conic program provides a safe approximation for the distributionally
robust inverse optimization problem (10) over the 2-Wasserstein ball:

\[
\begin{align*}
\text{minimize} & \quad \tau + \frac{1}{2} \left( \epsilon^2 \lambda + \frac{1}{N} \sum_{i=1}^{N} s_i \right) \\
\text{subject to} & \quad \theta \in \Theta, \quad \lambda \in \mathbb{R}_+, \quad \tau, r, \rho, \mu, \phi, \gamma \in \mathbb{R}^n, \quad \phi_1, \phi_2 \in C^*, \quad \mu_1, \mu_2, \gamma_i \in K^* \\
& \quad \chi_{i1}, \chi_{i2} \in \mathbb{R}^m, \quad \zeta_{i1}, \eta_{i1}, \zeta_{i2} \in \mathbb{R}^n \\
& \quad \chi_{i1} = \frac{1}{2} (-C^T \phi_1 + H^T (\mu_1 + \gamma_{i1}) - 2 \lambda \hat{s}_i) \quad \forall i \leq N \\
& \quad \zeta_{i1} = \frac{1}{2} (q - W^T \mu_1 - 2 \lambda \hat{s}_{i}) \quad \forall i \leq N \\
& \quad \rho_{i1} = \tau + r_i + \lambda \left( \langle \hat{x}_i, \hat{x}_i \rangle + \langle \hat{s}_i, \hat{s}_i \rangle \right) + \langle d, \phi_1 \rangle + \langle h, \mu_i + \gamma_i \rangle \quad \forall i \leq N \\
& \quad \chi_{i2} = \frac{1}{2} (-C^T \phi_2 + H^T \mu_2 - 2 \lambda \hat{s}_i) \quad \forall i \leq N \\
& \quad \zeta_{i2} = \frac{1}{2} (-W^T \mu_2 - 2 \lambda \hat{s}_i), \quad \eta_{i2} = -\frac{1}{2} W^T \gamma_{i2} \quad \forall i \leq N \\
& \quad \rho_{i2} = r_i + \lambda \left( \langle \hat{x}_i, \hat{x}_i \rangle + \langle \hat{s}_i, \hat{s}_i \rangle \right) + \langle d, \phi_2 \rangle + \langle h, \mu_i \rangle \quad \forall i \leq N \\
& \quad \begin{bmatrix} 
\lambda & -\frac{1}{2} Q_{xs} & \frac{1}{2} Q_{1s} \\
-\frac{1}{2} Q_{xs} \lambda & -Q_{xx} & 0 \\
\frac{1}{2} Q_{xs} & 0 & Q_{xx} \\
\zeta_{i1} & \frac{1}{2} Q_{ys} \lambda & -Q_{ys} \eta_{i1} \\
\zeta_{i1} & -\frac{1}{2} Q_{ys} \eta_{i1} & \zeta_{i2} \\
0 & \zeta_{i2} & \rho_{i2} 
\end{bmatrix} \succeq 0, \quad \begin{bmatrix} 
\lambda & 0 & \chi_{i2} \\
0 & \lambda & \zeta_{i2} \\
\chi_{i2} & \zeta_{i2} & \rho_{i2} 
\end{bmatrix} \succeq 0 \quad \forall i \leq N.
\end{align*}
\]

Proof. As in the proof of Theorem 4.1 one can show that the objective function of the inverse optimization problem (10) coincides with a variant of (17) that involves the 2-Wasserstein ball. In the remainder, we derive a safe conic approximation for the subordinate worst-case expectation problem

\[
\sup_{Q \in \mathbb{R}^2_+} \mathbb{E}^P \left[ \max \{ \ell_P(s, x) - \tau, 0 \} \right].
\]

Duality arguments borrowed from [27, Theorem 4.3] imply that the above infinite-dimensional linear program admits a strong dual of the form

\[
\begin{align*}
\inf_{\lambda \geq 0, r_i} & \quad \epsilon^2 \lambda + \frac{1}{N} \sum_{i=1}^{N} r_i \\
\text{s.t.} & \quad \sup_{s \in S} \sup_{x \in \mathbb{X}(s)} \sup_{y \in \mathbb{Y}(s)} \left( \langle x, Q_{xx} x + Q_{xx} s + q \rangle - \langle y, Q_{xy} y + Q_{xs} s + q \rangle - \tau \right) - \lambda \| (s, x) - (\hat{s}_i, \hat{x}_i) \|^2 \leq r_i \quad \forall i \leq N \\
& \quad -\lambda \| (s, x) - (\hat{s}_i, \hat{x}_i) \|^2 \leq r_i \quad \forall i \leq N.
\end{align*}
\]

By using the definitions of \( S \) and \( \mathbb{X}(s) \) given in (14), the \( i \)-th member of the first constraint group in (28) is satisfied if and only if the optimal value of the maximization problem

\[
\sup_{s, x, y} \quad \langle x, Q_{xx} x + Q_{xx} s + q \rangle - \langle y, Q_{xy} y + Q_{xs} s + q \rangle - \tau - \lambda \| (s, x) - (\hat{s}_i, \hat{x}_i) \|^2
\]

(29)

does not exceed \( r_i \). The Lagrangian of the conic quadratic program (29) is defined as

\[
L(s, x, y; \phi_i, \mu_i, \gamma_i) := \langle x, Q_{xx} x + Q_{xx} s + q \rangle - \langle y, Q_{xy} y + Q_{xs} s + q \rangle - \tau - \lambda \| (s, x) - (\hat{s}_i, \hat{x}_i) \|^2 + \langle C_s - d, \phi_i \rangle + \langle W - H_s - h, \mu_i \rangle + \langle W_y - H_s - h, \gamma_i \rangle,
\]

and therefore (29) can be expressed as a max-min problem of the form

\[
\sup_{s, x, y} \inf_{\phi_i \in C^*} \inf_{\mu_i, \gamma_i \in K^*} L(s, x, y; \phi_i, \mu_i, \gamma_i) \leq \inf_{\phi_i \in C^*} \inf_{\mu_i, \gamma_i \in K^*} \sup_{s, x, y} L(s, x, y; \phi_i, \mu_i, \gamma_i),
\]
where the inequality follows from weak duality. Note that strong duality holds (meaning that the inequality collapses to an equality) if the Lagrangian is concave in \((s, x, y)\). A sufficient condition for strong duality is that \(\lambda\) exceeds the largest eigenvalue of \(Q_{xx}\). We conclude that the \(i\)-th member of the first constraint group in (28) holds if there exist \(\phi_{i1} \in \mathcal{C}^*\) and \(\mu_{i1}, \gamma_{i1} \in \mathcal{K}^*\) with \(\sup_{s,x,y} L(s, x; y; \phi_{i1}, \mu_{i1}, \gamma_{i1}) \leq r_i\) As the Lagrangian constitutes a quadratic function, this statement is satisfied if and only if there are \(\phi_{i1} \in \mathcal{C}^*\), \(\mu_{i1}, \gamma_{i1} \in \mathcal{K}^*\), \(\chi_{i1} \in \mathbb{R}^m\), \(\eta_{i1} \in \mathbb{R}^n\) and \(\rho_{i1} \in \mathbb{R}\) with

\[
\begin{align*}
\chi_{i1} &= \frac{1}{2}(-CT\phi_{i1} + H^T(\mu_{i1} + \gamma_{i1}) - 2\lambda \hat{s}_i) \\
\zeta_{i1} &= \frac{1}{2}(-q - W^T\mu_{i1} - 2\lambda \hat{x}_i), \quad \eta_{i1} = \frac{1}{2}(q - W^T\gamma_{i1}) \\
\rho_{i1} &= \tau + r_i + \lambda \left(\langle \hat{x}_i, \hat{x}_i \rangle + \langle \hat{s}_i, \hat{s}_i \rangle\right) + \langle d, \phi_{i1} \rangle + \langle h, \mu_i + \gamma_i \rangle
\end{align*}
\]

\[
\begin{bmatrix}
\lambda \\
-\frac{1}{2}Q_{xx}^* \\
\frac{1}{2}Q_{xx} \\
\chi_{i1} \\
\zeta_{i1} \\
\eta_{i1} \\
\rho_{i1}
\end{bmatrix} \succeq 0.
\]

Similarly, it can be shown that the \(i\)-th member of the second constraint group in (28) is satisfied if and only if there exist \(\phi_{i2} \in \mathcal{C}^*\), \(\mu_{i2} \in \mathcal{K}^*\), \(\chi_{i2} \in \mathbb{R}^m\), \(\zeta_{i2} \in \mathbb{R}^n\) and \(\rho_{i2} \in \mathbb{R}\) such that\(^6\)

\[
\begin{align*}
\chi_{i2} &= \frac{1}{2}(-CT\phi_{i2} + H^T\mu_{i2} - 2\lambda P_s \hat{s}_i) \\
\zeta_{i2} &= \frac{1}{2}(-W^T\mu_{i2} - 2\lambda \hat{x}_i), \quad \eta_{i2} = -\frac{1}{2}W^T\gamma_{i2} \\
\rho_{i2} &= \tau + r_i + \lambda \left(\langle \hat{x}_i, \hat{x}_i \rangle + \langle \hat{s}_i, \hat{s}_i \rangle\right) + \langle d, \phi_{i2} \rangle + \langle h, \mu_i \rangle
\end{align*}
\]

\[
\begin{bmatrix}
\lambda \\
-\frac{1}{2}Q_{xx}^* \\
\frac{1}{2}Q_{xx} \\
\chi_{i2} \\
\zeta_{i2} \\
\eta_{i2} \\
\rho_{i2}
\end{bmatrix} \succeq 0.
\]

Replacing the semi-infinite constraints in (28) with the respective semidefinite approximations shows that the worst-case expectation (27) is bounded above by the optimal value of the conic program

\[
\begin{align*}
\inf_{\epsilon^2 \lambda + \frac{r_i}{\epsilon}} & \\
\text{s.t.} & \lambda \in \mathbb{R}_+, \ r_i, \rho_{i1}, \rho_{i2} \in \mathbb{R}, \ \phi_{i1}, \phi_{i2} \in \mathcal{C}^*, \ \mu_{i1}, \mu_{i2}, \gamma_{i1} \in \mathcal{K}^* \\
& \chi_{i1}, \chi_{i2} \in \mathbb{R}^m, \ \zeta_{i1}, \zeta_{i2} \in \mathbb{R}^n \\
& \chi_{i1} = \frac{1}{2}(-CT\phi_{i1} + H^T(\mu_{i1} + \gamma_{i1}) - 2\lambda \hat{s}_i) \quad \forall i \leq N \\
& \zeta_{i1} = \frac{1}{2}(-q - W^T\mu_{i1} - 2\lambda \hat{x}_i), \quad \eta_{i1} = \frac{1}{2}(q - W^T\gamma_{i1}) \quad \forall i \leq N \\
& \rho_{i1} = \tau + r_i + \lambda \left(\langle \hat{x}_i, \hat{x}_i \rangle + \langle \hat{s}_i, \hat{s}_i \rangle\right) + \langle d, \phi_{i1} \rangle + \langle h, \mu_i + \gamma_i \rangle \quad \forall i \leq N \\
& \chi_{i2} = \frac{1}{2}(-CT\phi_{i2} + H^T\mu_{i2} - 2\lambda \hat{s}_i) \quad \forall i \leq N \\
& \zeta_{i2} = \frac{1}{2}(-W^T\mu_{i2} - 2\lambda \hat{x}_i), \quad \eta_{i2} = -\frac{1}{2}W^T\gamma_{i2} \quad \forall i \leq N \\
& \rho_{i2} = \tau + r_i + \lambda \left(\langle \hat{x}_i, \hat{x}_i \rangle + \langle \hat{s}_i, \hat{s}_i \rangle\right) + \langle d, \phi_{i2} \rangle + \langle h, \mu_i \rangle \quad \forall i \leq N \\
& \begin{bmatrix}
\lambda \\
-\frac{1}{2}Q_{xx}^* \\
\frac{1}{2}Q_{xx} \\
\chi_{i1} \\
\zeta_{i1} \\
\eta_{i1} \\
\rho_{i1}
\end{bmatrix} \succeq 0, \quad \begin{bmatrix}
\lambda \\
0 \\
\chi_{i2} \\
\zeta_{i2} \\
\rho_{i2}
\end{bmatrix} \succeq 0 \quad \forall i \leq N.
\end{align*}
\]

The claim then follows by substituting (30) into a suitable worst-case CVaR formula akin to (17). \(\Box\)

\(^6\)If \((\hat{s}_i, \hat{x}_i) \in \Xi\), this constraint simplifies to \(r_i \geq 0\). Situations where \((\hat{s}_i, \hat{x}_i) \notin \Xi\) will be considered in Section 6.
Remark 5.4 (Exactness for Small Wasserstein Radii). In spite of the hardness results outlined in Theorems 5.1 and 5.2, the tractable approximation of Theorem 5.3 is sometimes exact for small Wasserstein radii. Specifically, one can prove that (26) is equivalent to (10) for sufficiently small $\varepsilon$ if the observer is risk neutral ($\alpha = 0$) and optimizes over the class of signal-dependent linear candidate objectives (i.e., the class of quadratic candidate objectives with $Q_{xx} = 0$).

6. INVERSE OPTIMIZATION UNDER IMPERFECT INFORMATION

The proposed framework for inverse optimization described in Sections 2 and 3 is predicated on the assumption of perfect information. Specifically, it is assumed that the agent is able to determine and implement the best response $x$ to any given signal $s$ and that the observer can measure $s$ and $x$ precisely. Moreover, it is assumed that the family $\mathcal{F}$ of candidate objective functions contains the agent’s true objective function $F$. In practice, however, any of the following complications may arise.

(i) The observer might face model uncertainty in the sense that the chosen class $\mathcal{F}$ could fail to contain the true objective function $F$. (ii) The agent might suffer from bounded rationality, meaning that he might settle for a suboptimal decision $x$ due to cognitive or computational limitations. Even if the best response can be computed exactly, the agent may suffer from implementation errors, meaning that an error-free implementation of the desired best response may not be possible. (iii) The exact signal $s$ and corresponding response $x$ might be corrupted by measurement noise.

To some extent, all of these complications are inevitable in any real application. However, the proposed approach to inverse optimization enables us to address all of the above forms of imperfect information in a systematic and theoretically satisfactory manner.

6.1. Model Uncertainty. If the class of candidate models $\mathcal{F} = \{F_\theta : \theta \in \Theta\}$ underlying the observer’s inverse optimization problem is not rich enough to contain the agent’s true objective function $F$ (or a monotonically increasing function of $F$ that implies the same preferences), then there exists typically no $\theta \in \Theta$ such that the loss functions (3) vanish on all training samples. In this case, the distributionally robust inverse optimization problem (10) determines the candidate model $F_\theta^{\star}$ from within $\mathcal{F}$ that best approximates $F$. Even though an exact recovery of $F$ is impossible, the out-of-sample guarantees established in Section 3 and the tractable reformulations derived in Sections 4 and 5 remain valid.

6.2. Bounded Rationality. Agents with cognitive and/or computational limitations may not be able to solve (1) to global optimality. Such agents may respond to a given signal $s$ with a $\kappa$-optimal solution $x_\kappa$, that is, a decision $x_\kappa \in X(s)$ with $F(s, x_\kappa) \leq \min_{x \in X(s)} F(s, x) + \kappa$. Here, the parameter $\kappa \geq 0$ quantifies the agent’s rationality. Indeed, if $\kappa > 0$, the agent accepts an extra cost of $\kappa$ for the freedom of choosing a $\kappa$-optimal decision, which is presumably cheaper to find than a global minimizer. If the agent’s decisions are known to be $\kappa$-optimal, it makes sense to replace the loss function (3a) with

$$
\ell_\theta(s, x) := \max \{F_\theta(s, x) - \min_{y \in X(s)} F_\theta(s, y) - \kappa, 0\}.
$$

By construction, $\ell_\theta(s, x) = 0$ whenever $x$ is a $\kappa$-optimal response to the signal $s$, and $\ell_\theta(s, x) > 0$ otherwise. For this loss function the out-of-sample performance guarantee offered by Corollary 3.4 still holds. Moreover, when $\mathcal{F}$ represents the class of linear (quadratic) candidate objectives, then problem (10) is equivalent to (can be conservatively approximated by) a finite conic program.
Corollary 6.1 (Bounded Rationality). Assume that the loss function is defined as in (31), the set \( S \) and \( \Xi(s) \) are defined as in (14), and risk is measured by the CVaR at level \( \alpha \in (0, 1] \).

(i) If \( F \) represents the class of linear candidate objectives from Section 4, then the inverse optimization problem (10) over the 1-Wasserstein ball is equivalent to a variant of the conic program (16) where \( \tau \in \mathbb{R}_+ \) and the first constraint group is replaced with

\[
\langle C\hat{s}_i - d, \phi_i^1 \rangle + \langle W\hat{x}_i - H\hat{s}_i - h, \mu_i^1 + \gamma_i^1 \rangle \leq r_i + \tau + \kappa \quad \forall i \leq N.
\]

(ii) If \( F \) represents the class of quadratic candidate objectives from Section 5, then the inverse optimization problem (10) over the 2-Wasserstein ball is equivalent to a variant of the conic program (26) where \( \tau \in \mathbb{R}_+ \) and the defining equation of \( \rho_i \) is replaced with

\[
\rho_i^1 = \tau + r_i + \kappa + \lambda \left( \langle \tilde{x}_i, \tilde{x}_i \rangle \right) + \langle \tilde{s}_i, \tilde{s}_i \rangle + \langle d, \phi_i^1 \rangle + \langle h, \mu_i + \gamma_i \rangle \quad \forall i \leq N.
\]

Proof. The proofs of the two assertions are almost identical to those of Theorem 4.1 and Theorem 5.3, respectively. Details are omitted for brevity. \( \square \)

6.3. Noisy Observations. The observed signal-response pairs could be noisy due to measurement errors. Moreover, the actual response \( x \) to a given signal \( s \) could differ from the intended response calculated from (1) due to implementation errors \([9]\). When the observations are corrupted by noise, we can think of the training samples \( \tilde{s}_i = (\tilde{s}_i, \tilde{x}_i) \) as random perturbations of some unobservable pure samples \( \xi_i = (s_i, \tilde{x}_i) \), where \( \tilde{x} \) is an exact optimal response to an unperturbed signal \( s \). Note that, in contrast to the pure samples \( \xi_i \), the noisy samples \( \tilde{s}_i \) may materialize outside of the uncertainty set \( \Xi \). We will henceforth assume that the noisy samples \( \tilde{s}_i \) have marginal distribution \( \tilde{P} \), while the corresponding unperturbed samples \( s_i \) are governed by a different marginal distribution \( \hat{P} \). As usual, we denote the empirical distribution on the (noisy) training samples by \( \hat{P}_N \).

Recall that, by construction, all distributions in the Wasserstein ball \( B_2^p(\hat{P}_N) \) are supported on \( \Xi \) irrespective of \( \varepsilon \). As the noisy training samples \( \tilde{s}_i \) may materialize outside of \( \Xi \), however, the empirical distribution \( \hat{P}_N \) is no longer guaranteed to be contained in \( B_2^p(\hat{P}_N) \). This implies that the Wasserstein ball \( B_2^p(\hat{P}_N) \) is nonempty only if \( \varepsilon \) exceeds the \( p \)-Wasserstein distance between \( \hat{P}_N \) and the set of all distributions supported on \( \Xi \), that is, if

\[
\inf_{Q \in M^p(\Xi)} W_p(Q, \hat{P}_N) \leq \varepsilon. \tag{32}
\]

If condition (32) is violated, the inner maximization problem in (10) is infeasible, which in turn implies that the outer minimization problem is unbounded. In this case the distributionally robust inverse optimization problem (10) is no longer solvable. In the remainder of this section we assume that (32) holds, which implies that the tractability Theorems 4.1, 4.3 and 5.3 as well as the existence Theorem 4.2 for extremal distributions remain valid. Indeed, recall that all of these theorems were proved without assuming that the training samples are contained in \( \Xi \).

In order to establish out-of-sample performance guarantees, we need an assumption about the noise distribution. Specifically, we assume here that the noise is small in the sense that the \( p \)-Wasserstein distance between \( P \) and \( \tilde{P} \) is bounded by a known constant \( \kappa \geq 0 \). This is the case, e.g., if the noise realizations are bounded by \( \kappa \) with certainty. As in the perfect information setting of Section 2, we use \( \tilde{J}_N(\varepsilon) \) and \( \tilde{\theta}_N(\varepsilon) \) to denote the optimal value and an optimal solution of problem (10) with \( \tilde{P} = B_2^p(\hat{P}_N) \), respectively, both of which are functions of the noisy training samples. In contrast to
Section 2, however, the certainty-equivalent loss of \( \hat{\theta}_N(\varepsilon) \) must be assessed under a pure test sample drawn from \( \hat{P} \) instead of a noisy test sample drawn from \( P \). This approach is mathematically sound as the loss function loses its meaning (or may even be negative) for signal-response pairs outside of \( \Xi \). It also has conceptual appeal because the observer wishes to predict the agent’s true preferences unaffected by noise that carries no information. Thus, we aim to control \( \rho_n(\ell_{\theta_N}(\varepsilon)) \) instead of \( \rho_n(\ell_{\hat{\theta}_N}(\varepsilon)) \).

**Corollary 6.2** (Out-of-Sample Performance Guarantee with Noisy Measurements). If \( P \) satisfies the light tail Assumption 3.2, then the out-of-sample risk of \( \theta_N(\varepsilon) \) satisfies

\[
P^N \left[ \rho_n(\ell_{\hat{\theta}_N}(\varepsilon)) \leq \tilde{J}_N(\varepsilon) \right] \geq 1 - \beta
\]

for every confidence level \( \beta \in (0,1) \) and for every \( \varepsilon \geq \kappa + \varepsilon_N(\beta) \), where \( \varepsilon_N(\beta) \) is defined as in (12). We emphasize that the confidence of the performance bound must be computed as usual with respect to the distribution \( P^N \) of the (noisy) training samples underlying \( \tilde{J}_N(\varepsilon) \) and \( \hat{\theta}_N(\varepsilon) \).

**Proof.** Select \( \varepsilon \geq \kappa + \varepsilon_N(\beta) \). We first show that the distribution \( \hat{P} \) of the unperturbed test samples resides within the \( p \)-Wasserstein ball of radius \( \varepsilon \) around the empirical distribution \( \hat{P}_N \) of the noisy training samples with confidence \( 1 - \beta \). Indeed, the triangle inequality implies

\[
W_p(\hat{P}, \hat{P}_N) \leq W_p(\hat{P}, P) + W_p(P, \hat{P}_N).
\]

By assumption, the first term on the right hand side is bounded by \( \kappa \) with certainty. Theorem 3.3 further guarantees that the second term is bounded above by \( \varepsilon_N(\beta) \) with confidence \( 1 - \beta \). We may thus conclude that \( \hat{P} \in \mathcal{B}_p(\hat{P}_N) \) and, \textit{a fortiori}, that

\[
\rho_n(\ell_{\hat{\theta}_N}(\varepsilon)) \leq \sup_{Q \in \mathcal{B}_p(\hat{P}_N)} \rho_n(\ell_{\hat{\theta}_N}(\varepsilon)) = \tilde{J}_N(\varepsilon)
\]

with probability \( 1 - \beta \). This observation completes the proof. \( \square \)

### 7. Numerical Experiments

All optimization problems are implemented in MATLAB using the YALMIP interface [25]. All semi-definite programs are solved with Sedumi 1.3 and all other optimization problems with CPLEX 12.6.

#### 7.1. Learning a Linear Objective Function.

We illustrate our approach first through an expository example, where an agent receives a signal \( s \) that is uniformly distributed on the hypercube \( S = [0.5, 10]^3 \) and then minimizes a linear objective function \( F(s, x) = \langle \theta^*, x \rangle \) over the polytope \( \Xi(s) = \{ x \in \mathbb{R}^2 : \cos(\frac{\pi}{5}) x_1 + \sin(\frac{\pi}{5}) x_2 \leq s_k \forall k \leq 36 \} \). The gradient \( \theta^* \) of the objective function is known to belong to \( \Theta = \{ \theta \in \mathbb{R}^n : ||\theta||_\infty = 1 \} \). An observer who is aware of the agent’s feasible set but ignorant of \( \theta^* \) and, \textit{a fortiori}, the distribution \( P \) of the signal-response pair \( \xi = (s, x) \), must infer the agent’s preferences based on \( N = 20 \) independent training samples from \( P \).

The observer approaches this task by solving the distributionally robust inverse optimization problem (10) with the suboptimality loss function (13) tailored to linear candidate objectives. We assume that risk is measured using the CVaR at level \( \alpha = 0.9 \), the perceived ambiguity of \( P \) is modeled through a 1-Wasserstein ball around the empirical distribution on the training dataset, and we use the \( \infty \)-norm on \( \Xi \). Thus, Theorem 4.1 applies, and (10) reduces to the tractable linear program (16), whose minimum \( \tilde{J}_N \) is attained at some data-driven solution \( \tilde{\theta}_N \). The observer then uses \( F_{\tilde{\theta}_N} \) to approximate the
agent’s unknown true objective function $F_\theta^*$ and to predict the agent’s response $x$ to a new signal $s$ independent of the training data. The following indicators capture different quality attributes of $\hat{\theta}_N$:  

- **Suboptimality:** The certainty equivalent suboptimality of the observed response $x$ to a new signal $s$ under the estimated objective function $F_{\hat{\theta}_N}$, that is, $\rho^P(\ell_{\hat{\theta}_N})$.
- **Predictability:** The expected distance between the observed response $x$ to a new signal $s$ and the set of minimizers implied by $F_{\hat{\theta}_N}$, that is, $\mathbb{E}^P[\min\{\|x - y\| : y \in X_{\hat{\theta}_N}(s)\}]$.
- **Identifiability:** The distance between the estimated and the true gradient of the agent’s objective function, that is, $\|\theta^* - \hat{\theta}_N\|$.

In each simulation run we first select $\theta^*$ uniformly at random from $\Theta$ and then generate 20 independent training samples from the distribution $P$ implied by $\theta^*$. Next, we solve (16) to compute $\hat{\theta}_N$ and $\hat{J}_N$ as well as the corresponding suboptimality, predictability and identifiability indicators. We use 200 test samples to estimate the risk measures and expected values underlying these indicators via the sample average approximation. All results are averaged over 100 independent simulation runs.

The suboptimality, predictability and identifiability indicators for all Wasserstein radii $\varepsilon \in [10^{-4}, 10^0]$ are visualized in Figures 1(a), (b) and (c), respectively. While the suboptimality and predictability virtually drop to zero for $\varepsilon \lesssim \infty$, the identifiability saturates at a positive value ($\sim 0.05$). This is expected as any vertex of $X(s)$ has an interior angle of $\pi/18$ radians, implying that there are always multiple gradients $\theta$ that make a vertex optimal. Thus, it is fundamentally impossible to identify $\theta^*$.

In the second experiment the observer has only access to noisy samples $\tilde{\xi}_i = (\tilde{s}_i, \tilde{x}_i)$, $i \leq N$. Specifically, we assume that the observed decisions $\tilde{x}_i$ differ from their unperturbed counterparts $\hat{x}_i$ by additive noise that is uniformly distributed on $[-0.1, 0.1]^2$, while the signals can be measured exactly. Note that $\tilde{x}_i$ is therefore either suboptimal or infeasible in the agent’s decision problem corresponding to the signal $\tilde{s}_i = \tilde{s}_i$. As in Section 6, we denote the distribution of the noisy observations by $\tilde{P}$. The observer then solves the distributionally robust inverse optimization problem (10) using a 1-Wasserstein ball around the empirical distribution of the *noisy training samples*. As the noise may push these samples outside of $\Xi$, the Wasserstein ball may fail to contain their empirical distribution. Moreover, the ball is empty if its radius violates the condition (32), in which case the inner maximization problem in (10) is infeasible and the outer minimization problem is unbounded. Thus, problem (10) is only solvable for sufficiently large Wasserstein radii. Conditional on (10) being solvable, the suboptimality, predictability and identifiability indicators are computed via sample average approximation using 200 *unperturbed test samples* from $\tilde{P}$. All results are averaged over 100 independent simulation runs.

The empirical probability of solvability over the 100 simulation runs is visualized by the red dashed curves in Figures 1(d)–(f). By construction, this probability is non-decreasing in $\varepsilon$ and saturates at 1 when $\varepsilon$ exceeds the $\infty$-norm of the most extreme noise realizations (i.e., at $\varepsilon = 0.1$). The suboptimality and predictability indicators shown in Figures 1(d) and (e), respectively, are averaged over all those simulation runs for which (10) is solvable. Thus, they are well-defined only for $\varepsilon \gtrsim 10^{-3}$. Besides that, as in the first experiment, they are virtually zero for $\varepsilon \lesssim 10^{-1}$. We conclude that, in spite of the measurement noise, the distributionally robust inverse optimization approach successfully learns the agent’s preferences as long as $\varepsilon$ is sufficiently large but not too large. The identifiability indicator shown in Figure 1(f) is again bounded away from zero for the same reasons as in the first experiment.
7.2. Consumer Behavior. A fundamental problem in marketing is to understand the purchasing decisions of consumers, which is an essential prerequisite for estimating demand functions. Assume that there are $n = 10$ products whose prices $s = \{s_i\}_{i=1}^n$ are independent and uniformly distributed on $[2, 6]$. Upon observing $s$, consumers select a basket of goods $x = \{x_i\}_{i=1}^n \in X(s) = \mathbb{R}_+$ that minimizes their disutility $F(s, x) = \langle s, x \rangle - U(x)$, where $\langle s, x \rangle$ represents the purchasing costs, while the concave quadratic function $U(x) = \langle x, Q^*_x x \rangle + \langle q^*, x \rangle$ captures the utility of consumption. The utility function is encoded by a parameter $\theta^* = (Q^*_x, q^*)$ that belongs to the set

$$\Theta = \{ \theta = (Q_{xx}, q) \in \mathbb{R}^{n \times n} \times \mathbb{R}_+^n : Q = Q^T \succeq 0 \}.$$ 

An observer who is ignorant of $\theta^*$ and, a fortiori, the distribution $\mathbb{P}$ of the signal-response pair $\xi = (s, x)$, aims to learn the consumer’s utility function based on $N = 20$ independent training samples from $\mathbb{P}$. This problem setup is inspired by [24]. We assume that the observer approaches her goal by solving problem (10) with the loss function (25) tailored to quadratic candidate objectives (in a special case where $Q_{xx}$ is fixed to $I$). Moreover, she measures risk using the CVaR at level $\alpha = 0.9$ and models the perceived ambiguity of $\mathbb{P}$ through a 2-Wasserstein ball around the empirical distribution on the training dataset. Distances in $\Xi$ are measured using the 2-norm. By Theorem 5.3, problem (10) thus reduces to the tractable semidefinite program (26), whose minimum $J_N$ is attained at $\hat{\theta}_N$.

In each simulation run we first draw $q^*$ from the uniform distribution on $[8, 12]^n$ and $Q^*_{xx}$ from the uniform distribution on $\{Q \in \mathbb{R}^{n \times n} : I \preceq Q = Q^T \preceq 3 I \}$. Then, we generate 20 independent training
Figures 2(a) and 2(d) show the suboptimality and predictability of $\hat{\theta}_N$ for (squared) Wasserstein radii $\varepsilon^2 \in [10^{-6}, 10^{-1}]$. For all $\varepsilon^2 \lesssim 10^{-2}$ both indicators remain below 0.005 and are therefore essentially zero for practical purposes. Thus, our approach predicts the consumer’s preferences almost exactly.

In the second experiment the observer has only access to noisy training samples, which are constructed from the unperturbed samples as in Section 7.1. Thus, prices can be measured exactly, while consumption decision are perturbed by additive noise that is uniformly distributed on $[-0.1, 0.1]^n$. As in Section 7.1, the distributionally robust inverse optimization problem (10) is only solvable if the Wasserstein radius is sufficiently large; see (32).

Figures 2(b) and 2(e) visualize the empirical probability of solvability over the 100 simulation runs (dashed red curves) as well as the suboptimality and predictability indicators averaged over all those simulation runs for which (10) is solvable (solid blue curves). The best results are obtained for Wasserstein radii that are just large enough for (10) to be solvable, but even for an optimal choice of $\varepsilon$ the solution quality is significantly worse than in the noise-free setting. This suggests that inverse optimization with noisy measurements is fundamentally more challenging. To corroborate this conjecture,
we re-estimate $\theta^*$ using the state-of-the-art variational inequality-based inverse optimization approach proposed in [11]. The resulting estimator achieves a suboptimality of 2.6 and a predictability of 3.3.

In the third experiment the observer faces model uncertainty (but has access to noise-free training samples). Specifically, we assume that the agent’s utility function is now given by $U(x) = \langle e, \sqrt{Ax} \rangle$, where $e$ denotes the vector of ones in $\mathbb{R}^n$, while $A$ is a random matrix drawn from the uniform distribution on $[0, 1]^{n \times n}$; see also [24]. Thus, it is impossible to recover the agent’s true objective function by solving an inverse optimization problem over the class of quadratic candidate objectives. As usual, all results are averaged over 100 simulation runs with independently selected input utility functions.

The suboptimality and predictability indicators corresponding to the data-driven solution $\hat{\theta}_N$ of (10) are shown in Figures 2(c) and 2(f), respectively. Even though the problem is solvable for every $\varepsilon \geq 0$, the best result are obtained for strictly positive Wasserstein radii, which enable the observer to combat overfitting to the training samples. In this experiment, the variational inequality-based estimator from [11] achieves a suboptimality of 1.4 and a predictability of 0.15.

Acknowledgments: This research was supported by the Swiss National Science Foundation grant BSCGI0_157733.

References