Conformal Inverse Optimization

Bo Lin [1] **Erick Delage** [2] **Timothy C. Y. Chan** [1]

Abstract

Inverse optimization has been increasingly used to estimate unknown parameters in an optimization model based on decision data. We show that such a point estimation is insufficient in a prescriptive setting where the estimated parameters are used to prescribe new decisions. The prescribed decisions may be low-quality and misaligned with human intuition and thus are unlikely to be adopted. To tackle this challenge, we propose conformal inverse optimization, which seeks to learn an uncertainty set for the unknown parameters and then solve a robust optimization model to prescribe new decisions. Under mild assumptions, we show that the suggested decisions can achieve bounded out-of-sample optimality gaps, as evaluated using both the ground-truth parameters and the decision maker's perception of the unknown parameters. Our method demonstrates strong empirical performance compared to classic inverse optimization.

1. Introduction

Inverse optimization (IO) seeks to estimate unknown parameters in an optimization model based on decision data. The estimated parameters can then be used to prescribe future decisions. For this IO pipeline to succeed in practice, the prescribed decision should not only be of high-quality (as evaluated using the ground-truth parameters) but also align with human intuition (i.e., perceived to be of high-quality). The latter encourages algorithm adoption (Chen et al., 2023; Donahue et al., 2023), which is critical in many real-world applications, e.g., rideshare vehicle positioning (Liu et al., 2023), bin packing (Sun et al., 2022), and product assortment (Kesavan & Kushwaha, 2020; Kawaguchi, 2021).

As an example, rideshare platforms, e.g., Uber and Lyft, provide a shortest-path to the driver at the start of a trip based on real-time traffic data (Nguyen, 2015). The driver then relies on her perception of the road network formed through past experience to evaluate the path. The driver may deviate from the suggested path if it is perceived to be low-quality. Although seasoned drivers are often capable of identifying a better path due to their tacit knowledge of the road network (Merchán et al., 2022), such deviations impose operational challenges as it may cause rider safety concerns and affect downstream decisions such as arrival time estimation, trip pricing, and rider-driver matching (Hu et al., 2022). Therefore, the platform may be interested in leveraging historical paths taken by drivers to suggest high-quality paths for future trips, as evaluated using both the travel time and the driver's perception.

In this paper, we first show that the classic IO pipeline may generate decisions that are low-quality and misaligned with human intuition. We next propose conformal IO, which first learns an uncertainty set from decision data and then solves a robust optimization model with the learned uncertainty set to prescribe decisions. Finally, we prove that the proposed approach has provable guarantees on the actual and perceived solution quality. Our contributions are as follows.

- **A new framework**. We propose a new prescriptive IO pipeline that integrates i) a novel method to learn uncertainty sets from decision data and ii) a robust model for decision recommendation.
- **Theoretical guarantees**. We prove that, with high probability, the learned uncertainty set contains parameters that make future observed decisions optimal. This coverage guarantee leads to provable bounds on the optimality gap of the decisions from conformal IO, as evaluated using the ground-truth parameters and the decision maker's (DM's) perceived parameters.
- **Performance**. Through experiments, we demonstrate strong performance of conformal IO compared to classic IO and provide insights into modeling choices.

2. Literature Review

Inverse optimization. IO is a method to estimate unknown parameters in the objective function (Ahuja & Orlin, 2001; Chan et al., 2014) or constraint matrix (Bertsimas et al., 2015; Birge et al., 2017; Chan & Kaw, 2020) of an optimization model based on decision data. Early IO papers focus

[1] Department of Mechanical and Industrial Engineering, University of Toronto, Toronto, ON, Canada [2] GERAD & Department of Decision Sciences, HEC Montréal, Montréal, Québec, Canada. Correspondence to: Bo Lin <blin@mie.utoronto.ca>.

Preliminary work. Under review.

on deterministic settings where the observed decisions are assumed to be optimal to the specified optimization model. Recently there has been growing interest in applying IO in stochastic settings where the observed decisions are subject to measurement and execution errors, and bounded rationality (Esfahani & Kuhn, 2018). Progress has been made to provide estimators that are statistically consistent (Aswani et al., 2018; Birge et al., 2022), tractable (Chan et al., 2019), and robust to data corruption (Esfahani et al., 2018). See Chan et al. (2023b) for a comprehensive review.

Our paper is positioned in the stochastic stream because the observed decisions are assumed to be generated using noisy perceptions of the unknown parameters. Unlike existing methods that provide a point estimation of the unknown parameters, we learn an uncertainty set that can be used in a robust optimization model to prescribe new decisions.

Estimate, then optimize. Conformal IO belongs to a family of data-driven optimization methods called "estimate, then optimize" (Elmachtoub et al., 2023). Recent research suggests that even small estimation errors may be amplified in the optimization step, resulting in significant decision errors. This issue can be mitigated by training the estimation model with decision-aware losses (Wilder et al., 2019; Mandi et al., 2022; Elmachtoub & Grigas, 2022) and robustifying the optimization model (Sun et al., 2023; Chan et al., 2023a).

We take a similar approach as the second stream, yet deviate from them by i) utilizing decision data instead of observations of the unknown parameters, and ii) focusing on both the ground-truth and perceived solution quality, the latter of which has not been studied in this stream of literature.

Data-driven uncertainty set construction. Uncertainty set construction involves deciding the structure and size of the uncertainty set. Central to this problem are i) the tractability of the resulting robust model and ii) the price of robustness (Bertsimas & Sim, 2004). Early endeavours use prior knowledge about the parameter uncertainty to design sets that are polyhedral (Ben-Tal & Nemirovski, 1999), ellipsoidal (Ben-Tal & Nemirovski, 2000), cardinality constrained (Bertsimas & Sim, 2004), and norm constrained (Bertsimas et al., 2004). The size of these uncertainty sets is usually tied to the coverage level specified by the DM. More recently, data have become a critical ingredient to define new uncertainty set structures (Delage & Ye, 2010; Ben-Tal et al., 2013; Esfahani & Kuhn, 2018; Gao & Kleywegt, 2023) and calibrate the size of the uncertainty set (Bertsimas et al., 2018; Chenreddy et al., 2022; Sun et al., 2023).

Our paper is related to the work of Sun et al. (2023) who first use an ML model to predict the unknown parameters and then calibrate an uncertainty set around the prediction using prediction errors estimated from a validation set. However, this approach does not apply in our setting as it requires observations of the unknown parameters, which we do not have access to. Our paper presents the first approach to calibrating uncertainty sets using decision data, which in many applications is more readily observable than parameters.

Algorithm aversion. AI is being increasingly used to augment human decisions. However, humans may be reluctant to adopt algorithmic suggestions despite their superior performance—a phenomenon called algorithm aversion (Burton et al., 2020; Jussupow et al., 2020). Empirical studies have identified a range of factors that explain such phenomenon, including algorithm transparency (Kizilcec, 2016), algorithm accuracy (Yin et al., 2019; Dietvorst et al., 2015), lack of decision control (Dietvorst et al., 2018; Meissner & Keding, 2021), and lack of human inputs (Kawaguchi, 2021). Recent studies reveal that AI solutions are more likely to be adopted if they align with the users' intuition (Bauer et al., 2023; Chen et al., 2023; Donahue et al., 2023; Liu et al., 2023), which motivates this study.

We contribute to this stream of literature by providing a principled approach to generate decisions that are high-quality and intuitive, aiming to mitigate algorithm aversion.

3. Preliminaries

In this section, we first present the problem setup (Section 3.1) and then describe the challenges with the classic IO pipeline (Section 3.2). Finally, we discuss alternatives to mitigate the challenges and provide intuition on why robustifying the optimization model would help (Section 3.3).

3.1. Problem Setup

Consider a *forward optimization* problem

$$\mathbf{FO}(\boldsymbol{\theta}, \mathbf{u}) : \underset{\mathbf{x} \in \mathcal{X}(\mathbf{u})}{\text{minimize}} \; f(\boldsymbol{\theta}, \mathbf{x}) \tag{1a}$$

where $\mathbf{x} \in \mathbb{R}^n$ is the decision vector whose feasible region $\mathcal{X}(\mathbf{u})$ is non-empty and is parameterized by exogenous parameters $\mathbf{u} \in \mathbb{R}^m$, $\boldsymbol{\theta} \in \mathbb{R}^d$ is a parameter vector, and $f : \mathbb{R}^{n \times d} \to \mathbb{R}$ is the objective function. Suppose $\mathbf{u}$ is distributed according to $\mathbb{P}_{\mathbf{u}}$ supported on $\mathcal{U}$. There exists a ground-truth parameter vector $\boldsymbol{\theta}^*$ that is unknown to the DM. Instead, the DM obtains a decision $\hat{\mathbf{x}}$ by solving $\mathbf{FO}(\hat{\boldsymbol{\theta}}, \mathbf{u})$ where $\hat{\boldsymbol{\theta}}$ is a noisy perception of $\boldsymbol{\theta}^*$. We assume that, while the distribution $\mathbb{P}_{\boldsymbol{\theta}}$ of $\hat{\boldsymbol{\theta}}$ is unknown, it is supported on a known bounded set $\boldsymbol{\Theta} \subset \mathbb{R}^d$ and that $\boldsymbol{\theta}^*$ is within the support of $\mathbb{P}_{\boldsymbol{\theta}}$. Let $\mathbb{P}_{(\boldsymbol{\theta},\mathbf{u})}$ denote the joint distribution of $\hat{\boldsymbol{\theta}}$ and $\mathbf{u}$. Let $\tilde{f} : \boldsymbol{\Theta} \times \mathcal{U} \to \mathbb{R}$ be an oracle that returns the optimal value of $\mathbf{FO}$ and $\tilde{\mathbf{x}} : \boldsymbol{\Theta} \times \mathcal{U} \to \mathbb{R}^n$ be an oracle that returns an optimal solution to $\mathbf{FO}$ drawn uniformly at random from $\mathcal{X}^{\text{OPT}}(\boldsymbol{\theta}, \mathbf{u}) := \text{argmin}\,\{f(\boldsymbol{\theta}, \mathbf{x}) \,|\, \mathbf{x} \in \mathcal{X}(\mathbf{u})\}$.

Given a dataset of N decision and exogenous parameter pairs $\mathcal{D} = \{\hat{\mathbf{x}}_k, \mathbf{u}_k\}_{k \in [N]}$, we are interested in finding a decision policy $\bar{\mathbf{x}} : \mathcal{U} \to \mathbb{R}^n$ to suggest decisions for future $\mathbf{u}$. We require $\bar{\mathbf{x}}(\mathbf{u}) \in \mathcal{X}(\mathbf{u})$ for any $\mathbf{u} \in \mathcal{U}$. As discussed later, $\bar{\mathbf{x}}(\mathbf{u})$ is usually generated by solving an optimization model that may have multiple optimal solutions. So we consider randomized policies (e.g., uniformly sample from a set of optimal solutions). This is nonrestrictive because a deterministic policy can be recovered from a randomized policy that samples the deterministic solution with probability one. We use the following metrics to evaluate $\bar{\mathbf{x}}$.

Definition 3.1. The actual optimality gap (AOG) of a decision policy $\bar{\mathbf{x}}$ is defined as

$$\mathrm{AOG}(\bar{\mathbf{x}}) := \mathbb{E}\left[f\left(\boldsymbol{\theta}^*, \bar{\mathbf{x}}(\mathbf{u})\right) - \tilde{f}\left(\boldsymbol{\theta}^*, \mathbf{u}\right)\right] \tag{2}$$

where the expectation is taken over the joint distribution of the random variable $\mathbf{u}$ and the decision sampled using the possibly randomized policy $\bar{\mathbf{x}}$.

Definition 3.2. The perceived optimality gap (POG) of a decision policy $\bar{\mathbf{x}}$ is defined as

$$\mathrm{POG}(\bar{\mathbf{x}}) := \mathbb{E}\left[f\left(\hat{\boldsymbol{\theta}}, \bar{\mathbf{x}}(\mathbf{u})\right) - \tilde{f}\left(\hat{\boldsymbol{\theta}}, \mathbf{u}\right)\right]. \tag{3}$$

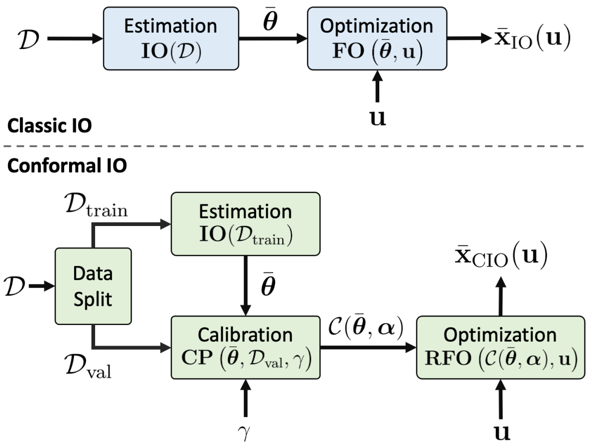
where the expectation is taken with respect to the randomness in $\hat{\boldsymbol{\theta}}$, $\mathbf{u}$, and possibly $\bar{\mathbf{x}}$.

AOG is an objective performance measure under the ground-truth parameters $\boldsymbol{\theta}^*$. Achieving a low AOG means that $\bar{\mathbf{x}}$ can generate high-quality decisions. In contrast, POG is a subjective measure that depends on the DM's perception of the problem. Achieving a low POG is critical to mitigate algorithm aversion (Burton et al., 2020).

3.2. An Inverse Optimization Pipeline

Finding $\bar{\mathbf{x}}$ is challenging for three reasons. First, unlike many machine learning problems where the prediction target is unconstrained, we require $\bar{\mathbf{x}}(\mathbf{u})$ to be a feasible solution to $\mathbf{FO}(\boldsymbol{\theta}, \mathbf{u})$ which may involve a large number of complex constraints. End-to-end supervised learning approaches that predict $\hat{\mathbf{x}}$ based on $\mathbf{u}$ can often fail as they typically do not provide feasibility guarantees. An optimization module is often needed to recover feasibility or produce feasible solutions based on $\mathbf{u}$ and some estimated $\boldsymbol{\theta}$. Second, we do not have access to $\boldsymbol{\theta}^*$ or $\hat{\boldsymbol{\theta}}$, which precludes using classic machine learning techniques to estimate the parameters. Finally, since AOG and POG are tied to $\boldsymbol{\theta}^*$ and $\hat{\boldsymbol{\theta}}$, respectively, it is unclear how to design loss functions for parameter estimation that lead to low AOG or POG. To make things even worse, these two metrics may not align with each other, so we are essentially dealing with a bi-objective problem.

In light of the first two challenges, a classic IO pipeline (visualized in Figure 1) has been proposed to first obtain a point estimation $\bar{\boldsymbol{\theta}}$ of the unknown parameters and then employ a policy $\bar{\mathbf{x}}_{\mathrm{IO}}(\mathbf{u}) := \tilde{\mathbf{x}}(\bar{\boldsymbol{\theta}}, \mathbf{u})$ to prescribe decisions for any $\mathbf{u} \in \mathcal{U}$ (Rönnqvist et al., 2017; Babier et al., 2020). Specifically, we can estimate the parameters by solving the following *inverse optimization* problem



$\mathcal{D}$
Estimation
$\mathbf{IO}(\mathcal{D})$
$\bar{\boldsymbol{\theta}}$
Optimization
$\mathbf{FO}(\bar{\boldsymbol{\theta}}, \mathbf{u})$
$\bar{\mathbf{x}}_{\mathrm{IO}}(\mathbf{u})$
$\mathbf{u}$
Classic IO
Conformal IO
$\mathcal{D}_{\mathrm{train}}$
Estimation
$\mathbf{IO}(\mathcal{D}_{\mathrm{train}})$
$\mathcal{D}$
Data Split
$\bar{\boldsymbol{\theta}}$
$\bar{\mathbf{x}}_{\mathrm{CIO}}(\mathbf{u})$
Calibration
$\mathbf{CP}(\bar{\boldsymbol{\theta}}, \mathcal{D}_{\mathrm{val}}, \gamma)$
$\mathcal{C}(\bar{\boldsymbol{\theta}}, \boldsymbol{\alpha})$
Optimization
$\mathbf{RFO}(\mathcal{C}(\bar{\boldsymbol{\theta}}, \boldsymbol{\alpha}), \mathbf{u})$
$\mathcal{D}_{\mathrm{val}}$
γ
$\mathbf{u}$

Figure 1. Classic and conformal IO pipelines.

$$\mathbf{IO}(\mathcal{D}) : \underset{\boldsymbol{\theta}}{\text{minimize}} \ \frac{1}{N} \sum_{k \in [N]} \ell\left(\hat{\mathbf{x}}_k, \mathcal{X}^{\mathrm{OPT}}(\boldsymbol{\theta}, \mathbf{u}_k)\right), \tag{4}$$

where ℓ is a non-negative loss function that returns 0 when $\hat{\mathbf{x}}_k \in \mathcal{X}^{\mathrm{OPT}}(\boldsymbol{\theta}, \mathbf{u}_k)$. For instance, the following two loss functions are commonly used in the literature

Definition 3.3. The decision loss of $\boldsymbol{\theta}$ is given by

$$\ell_{\mathrm{D}}\left(\hat{\mathbf{x}}, \mathcal{X}^{\mathrm{OPT}}(\boldsymbol{\theta}, \mathbf{u})\right) = \min_{\mathbf{x} \in \mathcal{X}^{\mathrm{OPT}}(\boldsymbol{\theta}, \mathbf{u})} \|\mathbf{x} - \hat{\mathbf{x}}\|_2. \tag{5}$$

Definition 3.4. The sub-optimality loss of $\boldsymbol{\theta}$ is given by

$$\ell_{\mathrm{S}}\left(\hat{\mathbf{x}}, \mathcal{X}^{\mathrm{OPT}}(\boldsymbol{\theta}, \mathbf{u})\right) = \min_{\mathbf{x} \in \mathcal{X}^{\mathrm{OPT}}(\boldsymbol{\theta}, \mathbf{u})} f(\boldsymbol{\theta}, \hat{\mathbf{x}}) - f(\boldsymbol{\theta}, \mathbf{x}). \tag{6}$$

The decision loss, which penalizes the L_2 distance between the observed and suggested decisions, enjoys statistical consistency when the forward problem is convex (Aswani et al., 2018). The statistical consistency implies that the resulting policy can achieve zero AOG when i) $\mathbb{E}(\hat{\boldsymbol{\theta}}) = \boldsymbol{\theta}^*$ and ii) a large $\mathcal{D}$ is available. However, this may not be attainable in practice because $\mathbb{E}(\hat{\boldsymbol{\theta}}) \neq \boldsymbol{\theta}^*$ and $\mathbf{IO}(\mathcal{D})$ may be challenging to solve when $\mathcal{D}$ is large since it is non-convex even if the $\mathbf{FO}$ is convex. For many real-world problems, e.g., routing problems that involve a large number of discrete decisions and unknown parameters, the sub-optimality loss, which penalizes the optimality gap achieved by the observed decision under the estimated parameters, is often preferable because it offers better computational properties. As remarked by Esfahani et al. (2018), we encounter a situation similar to binary classification where it is preferable to minimize the convex hinge/cross-entropy loss instead of the 0-1 loss even if the 0-1 loss is the actual metric of interest. While such a

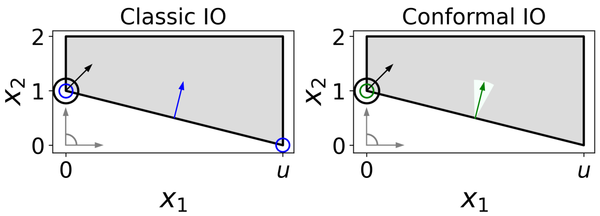


Figure 2. Illustration of the classic and conformal IO in Example 3.1. The gray areas are the feasible region $\mathcal{X}(u)$. The black arrows correspond to the ground-truth parameter $\boldsymbol{\theta}^*$. The gray arrows are the extreme rays of $\boldsymbol{\Theta}$. The blue and green arrows are the point estimations ($\bar{\boldsymbol{\theta}}$). The green area is the uncertainty set $\mathcal{C}(\bar{\boldsymbol{\theta}}, \alpha)$. The black circles are the optimal solution to $\mathbf{FO}(\boldsymbol{\theta}^*, u)$. The blue and green circles are the suggested decisions from the two pipelines. Note that $\bar{\mathbf{x}}_{\text{IO}}$ may suggest any decisions on the facet corresponding to the constraint $x_1 + ux_2 \geq u$, which are omitted for clarity.

trade-off is acceptable in some applications, we suggest that it is undesirable in our setting because the resulting policy can achieve arbitrarily large AOG and POG. To see this, consider the following example (visualized in Figure 2).

Example 3.1. *Let* $\mathbf{FO}(\theta, u)$ *be the following problem*

$$\begin{aligned} \text{minimize} \quad & \theta_1 x_1 + \theta_2 x_2 && \text{(7a)} \\ \text{subject to} \quad & x_1 + ux_2 \geq u && \text{(7b)} \\ & 0 \leq x_1 \leq u && \text{(7c)} \\ & 0 \leq x_2 \leq 2. && \text{(7d)} \end{aligned}$$

Let the ground-truth $\boldsymbol{\theta}^* = (\cos(\pi/4), \sin(\pi/4))$ *and* $\mathcal{U} = \{u\}$ *where* $u > 1$ *is a real constant. We are given a dataset* $\mathcal{D} = \{\hat{\mathbf{x}}_k, u\}_{k=1}^N$ *where* $\hat{\mathbf{x}}_k = \tilde{\mathbf{x}}(\hat{\boldsymbol{\theta}}_k, u)$ *with* $\hat{\boldsymbol{\theta}}_k$ *uniformly and independently drawn from* $\boldsymbol{\Theta} = \{(\cos\delta, \sin\delta) \mid \delta \in (0, \pi/2)\}$ *for all* $k \in [N]$.

Lemma 3.5. *w In Example 3.1, let* $\bar{\boldsymbol{\theta}}_N$ *denote an optimal solution to* $\mathbf{IO}(\mathcal{D})$ *with the sub-optimality loss* (6), *we have* $\mathbb{P}\left(\bar{\boldsymbol{\theta}}_N = \boldsymbol{\theta}_u\right) \to 1$ *as* $N \to \infty$, *where* $\boldsymbol{\theta}_u := \left(1/\sqrt{1+u^2}, u/\sqrt{1+u^2}\right)$.

Lemma 3.5 shows that, when using $\mathbf{IO}(\mathcal{D})$ with the sub-optimality loss to estimate the unknown parameter in Example 3.1, the probability of the estimated parameter being $\boldsymbol{\theta}_u$ converges to one asymptotically. This implies that asymptotically we are almost certain that $\bar{\mathbf{x}}_{\text{IO}}(u) = \tilde{\mathbf{x}}(\boldsymbol{\theta}_u, u)$, i.e. the policy that samples uniformly from the facet corresponding to the constraint $x_1 + ux_2 \geq u$. As a result, $\bar{\mathbf{x}}_{\text{IO}}$ can achieve arbitrarily large AOG and POG when u is set to a large enough value since decisions closer to the end of $(u, 0)$ are of low-quality and perceived to be of low-quality by most DMs. See Appendix A for complete statements and proofs. We then arrive at the following negative result for the classic IO pipeline.

Proposition 3.6. *Let* $\mathcal{D}$ *be a dataset,* $\bar{\boldsymbol{\theta}}$ *be an optimal solution to* $\mathbf{IO}(\mathcal{D})$ *using the sub-optimality loss* (6), *and* $\bar{\mathbf{x}}_{\text{IO}}(\mathbf{u}) = \tilde{\mathbf{x}}(\bar{\boldsymbol{\theta}}, \mathbf{u})$ *for any* $\mathbf{u} \in \mathcal{U}$, *then* $\bar{\mathbf{x}}_{\text{IO}}$ *can achieve arbitrarily large* AOG *and* POG.

Consequently, the suggested decisions are not useful to the DM as they are of poor quality, and they are unlikely to be adopted as they do not align with the DM's intuition.

3.3. Robustifying the Inverse Optimization Pipeline

A natural idea to improve the AOG and POG of $\bar{\mathbf{x}}$ is to robustify the decision pipeline. In particular, we may robustify i) the inverse problem, which has been studied by Esfahani et al. (2018), and ii) the decision recommendation problem, which is what we propose in this paper. In this section, we show that robustifying the inverse problem does not address the challenge of unbounded AOG and POG, but robustifying the forward problem does.

3.3.1. Robustifying the inverse problem.

Consider the following loss function.

Definition 3.7 (Esfahani et al. (2018))**.** The distributionally robust sub-optimality loss of $\boldsymbol{\theta}$ is given by

$$\ell_{\text{DR-S}}(\boldsymbol{\theta}) := \sup_{\mathbb{Q} \in \mathfrak{B}_r^p(\hat{\mathbb{P}}_{\mathbf{u},\hat{\mathbf{x}}})} \rho^{\mathbb{Q}} \left[\ell_{\text{S}}\left(\hat{\mathbf{x}}, \mathcal{X}^{\text{OPT}}(\boldsymbol{\theta}, \mathbf{u})\right)\right] \qquad (8)$$

where $\hat{\mathbb{P}}_{\mathbf{u},\hat{\mathbf{x}}}$ is the sample distribution of $\mathcal{D}$, $\mathfrak{B}_r^p(\hat{\mathbb{P}}_{\mathbf{u},\hat{\mathbf{x}}})$ is a p-Wasserstain ball of radius r centered at $\hat{\mathbb{P}}_{\mathbf{u},\hat{\mathbf{x}}}$, and $\rho^{\mathbb{Q}}$ is a risk measure, e.g., the value at risk.

The *distributionally robust inverse optimization* problem is

$$\mathbf{DRIO}(\mathcal{D}) : \underset{\boldsymbol{\theta} \in \boldsymbol{\Theta}}{\text{minimize}} \quad \ell_{\text{DR-S}}(\boldsymbol{\theta}). \qquad (9)$$

As shown by Esfahani et al. (2018), the estimated parameters from $\mathbf{DRIO}$ achieve bounded out-of-sample sub-optimality loss with a high probability. However, this does not imply bounded AOG and POG for the decision policy.

Lemma 3.8. *In Example 3.1,* $\boldsymbol{\theta}_u$ *is an optimal solution to* $\mathbf{DRIO}(\mathcal{D})$.

Lemma 3.8 shows that, in Example 3.1, the estimated parameter from $\mathbf{DRIO}(\mathcal{D})$ may still be $\boldsymbol{\theta}_u$. Hence, the decision policy is identical to $\bar{\mathbf{x}}_{\text{IO}}$ whose AOG and POG can be unbounded. The fundamental reason behind these negative results is the misalignment between the sub-optimality loss and the evaluation metrics. Achieving a low sub-optimality loss means that the suggested and observed decisions are of similar quality as evaluated using the estimated parameters. However, this does not speak to the similarity between these two decisions with respect to the DM's perceived parameters (POG) or the ground-truth parameters (AOG). Therefore, the out-of-sample guarantees on the sub-optimality loss do not translate into bounded AOG or POG.

3.3.2. Robustifying the forward problem.

Alternatively, we can consider robustifying the forward problem when prescribing new decisions. Specifically, we solve the following *robust forward optimization problem*

$$\mathbf{RFO}\left(\mathcal{C}(\bar{\boldsymbol{\theta}}, \boldsymbol{\alpha}), \mathbf{u}\right) : \underset{\mathbf{x}\in\mathcal{X}(\mathbf{u})}{\text{minimize}}\ \underset{\boldsymbol{\theta}\in\mathcal{C}(\bar{\boldsymbol{\theta}},\boldsymbol{\alpha})}{\text{maximize}}\ f(\boldsymbol{\theta}, \mathbf{x}) \quad (10a)$$

where $\mathcal{C}$ is an uncertainty set with $\bar{\boldsymbol{\theta}}$ being its center and $\boldsymbol{\alpha}$ representing parameters that control its shape/size. In this paper, since we generally do not know the scale of $\boldsymbol{\theta}$ (e.g., a driver's perceived travel cost on each road segment in a network), we focus on cases where the optimal decision to **FO** is invariant to the scale of $\boldsymbol{\theta}$, i.e., if $\mathbf{x} \in \mathcal{X}^{\text{OPT}}(\boldsymbol{\theta}, \mathbf{u})$, then $\mathbf{x} \in \mathcal{X}^{\text{OPT}}(\beta\boldsymbol{\theta}, \mathbf{u})$ for any $\beta \in \mathbb{R}_+$. So, we assume $\|\bar{\boldsymbol{\theta}}\|_2 = 1$ and focus on the following uncertainty set.

$$\mathcal{C}(\bar{\boldsymbol{\theta}}, \alpha) := \left\{\boldsymbol{\theta} \in \mathbb{R}^d \mid \|\boldsymbol{\theta}\|_2 = 1,\ \boldsymbol{\theta}^\intercal\bar{\boldsymbol{\theta}} \geq \cos\alpha\right\} \quad (11)$$

where $\alpha \in (0, \pi]$ represents the max angle between $\bar{\boldsymbol{\theta}}$ and any vector in the uncertainty set.

Remark 3.9. The scale-invariant condition apply to Example 3.1 and other cases where f is linear in $\boldsymbol{\theta}$. However, it does not imply that f is linear in $\boldsymbol{\theta}$. To see this, consider a forward problem with $f(\theta, x) = (\exp(\theta) - 1)x$ and feasible region $[-1, 1]$. The optimal solution is $x = -1$ when $\theta > 0$ and $x = 1$ when $\theta < 0$, so it is invariant to the scale of θ.

Now, using Example 3.1, we analyze the performance of a policy that utilizes **RFO** to prescribe decisions and provide insights into the conditions for bounding AOG and POG.

Lemma 3.10. *In Example 3.1, let* $\bar{\mathbf{x}}_{\text{CIO}}(u)$ *be an optimal solution to* $\mathbf{RFO}\left(\mathcal{C}(\bar{\boldsymbol{\theta}}_N, \alpha), u\right)$ *where* $\bar{\boldsymbol{\theta}}_N$ *is an optimal solution to* $\mathbf{IO}(\mathcal{D})$ *with the sub-optimality loss* (6)*. When* $\alpha \in (0, \pi/2)$*, we have* $\mathbb{P}\left[\text{AOG}(\mathbf{x}_{\text{CIO}}) = 0\right] \to 1$ *as* $N \to \infty$.

Lemma 3.11. *In Example 3.1, let* $\bar{\mathbf{x}}_{\text{CIO}}(u)$ *be an optimal solution to* $\mathbf{RFO}\left(\mathcal{C}(\bar{\boldsymbol{\theta}}_N, \alpha), u\right)$ *where* $\bar{\boldsymbol{\theta}}_N$ *is an optimal solution to* $\mathbf{IO}(\mathcal{D})$ *with the sub-optimality loss* (6)*. When* $\alpha \in (0, \pi/2)$*, we have* $\mathbb{P}\left[\text{POG}(\bar{\mathbf{x}}_{\text{CIO}}) < \pi/2\sqrt{2}\right] \to 1$ *as* $N \to \infty$.

Lemmas 3.10 and 3.11 show that, when using **RFO** to prescribe new decisions, the probability of achieving upper-bounded AOG and POG converges to one as N goes to infinity, as long as $\alpha \in (0, \pi/2)$. Interestingly, Lemmas 3.10 and 3.11 do not require the uncertainty set to contain $\boldsymbol{\theta}^*$ or most $\hat{\boldsymbol{\theta}}_k$. In fact, a very small α can help to bound AOG and POG. However, the performance of this approach still depends on the choice of α, which is non-trivial when $\mathbf{FO}(\boldsymbol{\theta}, \mathbf{u})$ is more complex than a two-dimensional linear program. We address this problem next.

4. Conformal Inverse Optimization

In this section, we present a principled approach to learn uncertainty sets that lead to provable performance guarantees. As presented later, the learned uncertainty set contains parameters that make the next DM's decision optimal with a specified probability. We call this approach conformal IO due to its connection to conformal prediction (Vovk et al., 2005), which aims to predict a set that contains the next prediction target with a specified probability. As illustrated in Figure 1, conformal IO has three steps: i) data split, ii) point estimation, and iii) uncertainty set calibration. We present these three steps in Section 4.1 and analyze the properties of the learned uncertainty set and conformal IO in Section 4.2.

4.1. Learning an Uncertainty Set

Data split. We first split $\mathcal{D}$ into training and validation sets, namely $\mathcal{D}_{\text{train}}$ and $\mathcal{D}_{\text{val}}$. Let $\mathcal{K}_{\text{train}}$ and $\mathcal{K}_{\text{val}}$ index $\mathcal{D}_{\text{train}}$ and $\mathcal{D}_{\text{val}}$, respectively, while $N_{\text{train}} = |\mathcal{D}_{\text{train}}|$ and $N_{\text{val}} = |\mathcal{D}_{\text{val}}|$.

Point estimation. Given the training set $\mathcal{D}_{\text{train}}$, we next apply data-driven IO techniques to obtain a point estimation $\bar{\boldsymbol{\theta}}$ of the unknown parameters. The most straightforward way is to solve $\mathbf{IO}(\mathcal{D}_{\text{train}})$ with any loss function. Alternatively, one may consider using end-to-end learning and optimization methods that do not require observations of the parameter vectors, e.g., the ones proposed by Berthet et al. (2020) and Tan et al. (2020). We remark that the point estimation can also come from other sources, for example, from a machine learning model that predicts the parameters. Our uncertainty set calibration method, which we introduce next, is independent of the point estimation method.

Uncertainty set calibration. Given a point estimation $\bar{\boldsymbol{\theta}}$, we next calibrate an uncertainty set that, with a specified probability, contains parameters that make the next observed decision optimal. This property is critical for the results in Section 4.2 to hold. While we can naively achieve this by setting $\alpha = \pi$, the resulting **RFO** may generate overly conservative decisions. Hence, we are interested in learning the smallest uncertainty set that satisfies this condition. We solve the following *calibration problem*

$$\mathbf{CP}(\bar{\boldsymbol{\theta}}, \mathcal{D}_{\text{val}}, \gamma) :$$

$$\begin{aligned}
&\underset{\alpha, \{\boldsymbol{\theta}_k\}_{k\in\mathcal{K}_{\text{val}}}}{\text{minimize}} && \alpha && (12a)\\
&\text{subject to} && \hat{\mathbf{x}}_k \in \mathcal{X}^{\text{OPT}}(\boldsymbol{\theta}_k, \mathbf{u}_k),\ \forall k \in \mathcal{K}_{\text{val}} && (12b)\\
& && \sum_{k\in\mathcal{K}_{\text{val}}} \frac{\mathbb{1}\left[\boldsymbol{\theta}_k \in \mathcal{C}(\bar{\boldsymbol{\theta}}, \alpha)\right]}{N_{\text{val}} + 1} \geq \gamma && (12c)\\
& && \|\boldsymbol{\theta}_k\|_2 = 1,\ \forall k \in \mathcal{K}_{\text{val}} && (12d)\\
& && 0 \leq \alpha \leq \pi, && (12e)
\end{aligned}$$

where decision α controls the size of the uncertainty set, decisions $\boldsymbol{\theta}_k$ represent a possible parameter vector associated with data point $k \in \mathcal{K}_{\text{val}}$, $\gamma \in [0, 1]$ is a DM-specified confidence level, and $\mathbb{1}$ is an indicator function that returns 1 if the condition is true and 0 otherwise.

Objective (12a) minimizes the size of the uncertainty set. Constraints (12b) ensure that $\boldsymbol{\theta}_k$ can make the observed decision $\hat{\mathbf{x}}_k$ optimal for $k \in \mathcal{K}_{\text{val}}$. Constraint (12c) ensures that at least γ of the decisions in $\mathcal{D}_{\text{val}}$ can find a vector in $\mathcal{C}$ that makes it optimal. Constraints (12d) ensure that the parameter vectors are on the unit sphere as defined in Equation (11). Constraint (12e) specifies the range of α.

Remark 4.1 (Optimality Conditions). The representation of Constraints (12b) depends on the structure of **FO**. For example, when the **FO** is a linear program, Constraints (12b) can be replaced with the dual feasibility and strong duality constraints. When the **FO** is a general convex optimization problem, we can use the KKT conditions. For non-convex forward problems, we can replace Constraints (12b) with

$$f(\boldsymbol{\theta}_k, \hat{\mathbf{x}}_k) \leq f(\boldsymbol{\theta}_k, \mathbf{x}), \quad \forall \mathbf{x} \in \mathcal{X}(\mathbf{u}), \tag{13}$$

which can be generated on-the-fly in a cutting-plane fashion.

Remark 4.2 (Feasibility). For **CP** to be feasible, we require, for each observed decision, there exists a $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ that make it optimal. This condition holds for a range of problems, e.g., the shortest path problem, travelling salesman problem, and knapsack problem, even if the DM is subject to bounded rationality, i.e., the DM settles for suboptimal solutions due to cognitive/computational limitations. For problems where this condition is violated, we may pre-process $\mathcal{D}_{\text{val}}$ to project $\hat{\mathbf{x}}$ to a point in $\mathcal{X}(\mathbf{u})$ such that the condition is satisfied.

Solving **CP** is hard for two reasons. First, it is non-convex regardless of the forward problem structure due to Constraints (12d). Second, Constraints (12b) involve the optimality conditions of N_{val} problems, so the size of **CP** scales up quickly as N_{val} increases. Nevertheless, as highlighted in Section 4.2, considering a large $\mathcal{D}_{\text{val}}$ is critical to ensure desirable properties of the learned uncertainty set, imposing considerable computation burdens. Below we introduce a decomposition method to solve **CP** efficiently.

Theorem 4.3. *Let $\mathcal{D}_{\text{val}}$ be a dataset, $\gamma \in [0, 1]$, $\bar{\boldsymbol{\theta}} \in \mathbb{R}^d$, $\tau = \lceil \gamma(N_{\text{val}} + 1) \rceil$ and Γ_τ be an operator that returns the τ^{th} largest value in a set. The optimal solution to $\mathbf{CP}(\bar{\boldsymbol{\theta}}, \mathcal{D}_{\text{val}}, \gamma)$ is $\alpha_\gamma := \arccos\left(\Gamma_\tau\left(\{c_k\}_{k \in \mathcal{K}_{val}}\right)\right)$ with*

$$c_k := \underset{\boldsymbol{\theta}_k}{\text{maximize}} \quad \boldsymbol{\theta}_k^\intercal \bar{\boldsymbol{\theta}} \tag{14a}$$

$$\text{subject to} \quad \hat{\mathbf{x}}_k \in \mathcal{X}^{\text{OPT}}(\boldsymbol{\theta}_k, \mathbf{u}_k) \tag{14b}$$

$$\|\boldsymbol{\theta}_k\|_2 \leq 1. \tag{14c}$$

Theorem 4.3 states that we can solve **CP** by first solving N_{val} optimization problems whose size is independent of N_{val} and then find a quantile in a set of N_{val} elements. The first step is parallelizable and the second step can be done in $O\left(N_{\text{val}} \log(\tau)\right)$ time. Since Problem (14) is a maximization problem, we can replace the constraint $\|\boldsymbol{\theta}_k\|_2 = 1$ with constraint (14c), so Problem (14) is convex when the forward problem is convex.

Let $\boldsymbol{\Theta}^{\text{OPT}}(\mathbf{u}, \mathbf{x}) := \left\{\boldsymbol{\theta} \in \mathbb{R}^d \,\middle|\, \mathbf{x} \in \mathcal{X}^{\text{OPT}}(\boldsymbol{\theta}, \mathbf{u}), \|\boldsymbol{\theta}\|_2 = 1\right\}$.

Remark 4.4 (Alternative Formulation). In **CP**, we make a modeling choice of letting γ of the validation data points satisfy $\boldsymbol{\Theta}^{\text{OPT}}(\mathbf{u}_k, \hat{\mathbf{x}}_k) \cap \mathcal{C}(\bar{\boldsymbol{\theta}}, \alpha) \neq \varnothing$. An alternative is to let γ of the validation data points satisfy $\hat{\boldsymbol{\theta}}_k \in \mathcal{C}(\bar{\boldsymbol{\theta}}, \alpha)$, or equivalently $\boldsymbol{\Theta}^{\text{OPT}}(\mathbf{u}_k, \hat{\mathbf{x}}_k) \subseteq \mathcal{C}(\bar{\boldsymbol{\theta}}, \alpha)$, which generally leads to a larger α, and thus more conservative decisions. We make this choice for two reasons. First, as illustrated in Example 3.1, covering most $\hat{\boldsymbol{\theta}}_k$ is unnecessary for conformal IO to achieve bounded AOG and POG. Second, using the alternative formulation would make Problem (14) a minimization problem, so Constraint (14c) needs to be an equality to avoid the trivial solution $\boldsymbol{\theta}_k = \mathbf{0}$. Consequently, Problem (14) cannot be cast as an equivalent convex problem.

4.2. Properties of the Learned Uncertainty Set

Next, we analyze the properties of the learned uncertainty set and the performance of conformal IO. We make the following assumption that is standard in the literature.

Assumption 4.5 (I.I.D. Samples)**.** The validation set $\mathcal{D}_{\text{val}}$ is generated using $\hat{\mathbf{x}}_k := \tilde{\mathbf{x}}(\hat{\boldsymbol{\theta}}_k, \mathbf{u}_k)$ where $(\hat{\boldsymbol{\theta}}_k, \mathbf{u}_k)$ are i.i.d. samples from $\mathbb{P}_{(\boldsymbol{\theta}, \mathbf{u})}$ for all $k \in \mathcal{K}_{\text{val}}$.

Theorem 4.6 (Uncertainty Set Validity)**.** *Let $\mathcal{D}_{\text{val}}$ be a dataset that satisfies Assumption 4.5, $(\hat{\boldsymbol{\theta}}, \mathbf{u})$ be a new i.i.d. sample from $\mathbb{P}_{(\boldsymbol{\theta}, \mathbf{u})}$, $\hat{\mathbf{x}} = \tilde{\mathbf{x}}(\hat{\boldsymbol{\theta}}, \mathbf{u})$, $\hat{\boldsymbol{\Theta}} = \boldsymbol{\Theta}^{\text{OPT}}(\mathbf{u}, \hat{\mathbf{x}})$, and α_γ be an optimal solution to $\mathbf{CP}(\bar{\boldsymbol{\theta}}, \mathcal{D}_{\text{val}}, \gamma)$ where $\bar{\boldsymbol{\theta}} \in \mathbb{R}^d$. We have, for any $\gamma \in [0, N_{val}/(N_{val} + 1)]$, that*

$$\mathbb{P}\left(\hat{\boldsymbol{\Theta}} \cap \mathcal{C}(\bar{\boldsymbol{\theta}}, \alpha_\gamma) \neq \varnothing\right) \geq \gamma. \tag{15}$$

For any $\gamma \in [0, 1]$, with probability at least $1 - 1/N_{\text{val}}$,

$$\left|\mathbb{P}\left(\hat{\boldsymbol{\Theta}} \cap \mathcal{C}(\bar{\boldsymbol{\theta}}, \alpha_\gamma) \neq \varnothing\right) - \gamma\right| \leq \epsilon(N_{\text{val}}) \tag{16}$$

where

$$\epsilon(N_{\text{val}}) := \sqrt{\frac{8\log(N_{val} + 1) + 2\log N_{val}}{N_{val}}} + \frac{2}{N_{val}}. \tag{17}$$

Theorem 4.6 states that our learned uncertainty set is conservatively valid and asymptotically exact (Vovk et al., 2005). More specifically, first, our method will produce a set that contains a $\boldsymbol{\theta}$ that makes the next DM's decision optimal no less than γ of the time that it is used (conservatively valid). The probability in Inequality (15) is with respect to the joint distribution over $\mathcal{D}_{\text{val}}$ and the new sample. Second, once the set is given, we have high confidence that, the probability of the next DM's decision being covered is within $\epsilon(N_{\text{val}})$

from γ. The probability in Inequality (16) is with respect to the new sample, while the high confidence is with respect to the draw of the validation data set. Overall, we have the almost sure convergence of $\mathbb{P}\left(\hat{\mathbf{\Theta}} \cap \mathcal{C}(\bar{\boldsymbol{\theta}}, \alpha_\gamma) \neq \varnothing\right)$ to γ as N goes to infinity, i.e. the coverage is asymptotically exact.

Next, we relate the validity results to the performance of conformal IO. We need additional assumptions as follows.

Assumption 4.7 (Lipschitz Continuity)**.** Let $\hat{\mathcal{X}} := \cup_{\hat{\boldsymbol{\theta}}, \mathbf{u} \in \mathbf{\Theta} \times \mathcal{U}} \mathcal{X}^{\text{OPT}}(\hat{\boldsymbol{\theta}}, \mathbf{u})$. For any $\hat{\mathbf{x}} \in \hat{\mathcal{X}}$, there exists a constant $\nu(\hat{\mathbf{x}}) \in \mathbb{R}_+$ such that, for any $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \mathbf{\Theta}$, we have $f(\boldsymbol{\theta}, \hat{\mathbf{x}}) - f(\boldsymbol{\theta}', \hat{\mathbf{x}}) \leq \nu(\hat{\mathbf{x}}) \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_2$.

Assumption 4.8 (Bounded Inverse Feasible Set)**.** There exists a constant $\eta \in \mathbb{R}_+$ such that, for any $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \mathbf{\Theta}^{\text{OPT}}(\hat{\mathbf{x}}, \mathbf{u})$, for some $\hat{\mathbf{x}} \in \hat{\mathcal{X}}$ and $\mathbf{u} \in \mathcal{U}$, we have $\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_2 \leq \eta$.

Assumption 4.9 (Linearity)**.** Function f is linear in $\boldsymbol{\theta}$.

Assumption 4.10 (Bounded Divergence)**.** There exists a constant $\sigma \in \mathbb{R}_+$ such that $\|\mathbb{E}(\hat{\boldsymbol{\theta}}) - \boldsymbol{\theta}^*\|_2 \leq \sigma$.

Assumption 4.7 is satisfied by many problems. For example, since $\mathbf{\Theta}$ is bounded, when f is convex in $\boldsymbol{\theta}$, it is Lipschitz in $\boldsymbol{\theta}$ on $\mathbf{\Theta}$. Assumption 4.8 is mild because $\mathbf{\Theta}^{\text{OPT}}(\hat{\mathbf{x}}, \mathbf{u})$ is by definition bounded for any $\hat{\mathbf{x}} \in \hat{\mathcal{X}}$ and $\mathbf{u} \in \mathcal{U}$, and is usually much smaller than $\mathbf{\Theta}$. Assumption 4.9 restricts the class of objective functions, yet is satisfied by many optimization models, e.g., routing problems, the knapsack problem, etc. Assumption 4.10 states that the distance between the expected perceived parameters and the ground-truth parameters is upper bounded. It is reasonable in many real-world settings. For example, rideshare drivers' perceived travel cost ($\hat{\boldsymbol{\theta}}$) should not be too different from the travel time ($\boldsymbol{\theta}^*$) as the latter is an important factor that drivers consider. We note that Assumptions 4.9 and 4.10 are needed only for bounding AOG.

Theorem 4.11 (Conformal IO Achieves Bounded POG)**.** *Let* $\bar{\mathbf{x}}_{\text{CIO}}(\mathbf{u})$ *be an optimal solution to* **RFO** $\left(\mathcal{C}(\bar{\boldsymbol{\theta}}, \alpha_1), \mathbf{u}\right)$ *for any* $\mathbf{u} \in \mathcal{U}$*, where* $\bar{\boldsymbol{\theta}} \in \mathbb{R}^d$ *and* α_1 *are chosen such that, for a new sample* $(\boldsymbol{\theta}', \mathbf{u}')$ *from* $\mathbb{P}_{(\boldsymbol{\theta}, \mathbf{u})}$ *and* $\mathbf{x}' = \tilde{\mathbf{x}}(\boldsymbol{\theta}', \mathbf{u}')$, $\mathbb{P}\left(\mathcal{C}(\bar{\boldsymbol{\theta}}, \alpha_1) \cap \mathbf{\Theta}^{\text{OPT}}(\mathbf{u}', \mathbf{x}') \neq \varnothing\right) = 1$. *If Assumptions 4.7 and 4.8 hold, then*

$$\text{POG}(\bar{\mathbf{x}}_{\text{CIO}}) \leq (\eta - 2\cos 2\alpha_1 + 2)\mu + \eta\mu_{\text{CIO}} \tag{18}$$

where $\mu := \mathbb{E}[\nu(\tilde{\mathbf{x}}(\hat{\boldsymbol{\theta}}, \mathbf{u}))]$ *and* $\mu_{\text{CIO}} := \mathbb{E}(\nu[\bar{\mathbf{x}}_{\text{CIO}}(\mathbf{u})])$.

Corollary 4.12 (Conformal IO Achieves Bounded AOG)**.** *Let* $\bar{\mathbf{x}}_{\text{CIO}}(\mathbf{u})$ *be an optimal solution to* **RFO** $\left(\mathcal{C}(\bar{\boldsymbol{\theta}}, \alpha_1), \mathbf{u}\right)$ *for any* $\mathbf{u} \in \mathcal{U}$*, where* $\bar{\boldsymbol{\theta}} \in \mathbb{R}^d$ *and* α_1 *are chosen such that, for a new sample* $(\boldsymbol{\theta}', \mathbf{u}')$ *from* $\mathbb{P}_{(\boldsymbol{\theta}, \mathbf{u})}$ *and* $\mathbf{x}' = \tilde{\mathbf{x}}(\boldsymbol{\theta}', \mathbf{u}')$, $\mathbb{P}\left(\mathcal{C}(\bar{\boldsymbol{\theta}}, \alpha_1) \cap \mathbf{\Theta}^{\text{OPT}}(\mathbf{u}', \mathbf{x}') \neq \varnothing\right) = 1$. *If Assumptions 4.7–4.10 hold, then*

$$\text{AOG}(\bar{\mathbf{x}}_{\text{CIO}}) \leq (2 - 2\cos 2\alpha_1 + \eta + \sigma)\mu^* + (\eta + \sigma)\mu_{\text{CIO}} \tag{19}$$

where $\mu^* := \mathbb{E}\left(\nu[\tilde{\mathbf{x}}(\boldsymbol{\theta}^*, \mathbf{u})]\right)$.

Theorem 4.11 and Corollary 4.12 state that, when the uncertainty set contains a $\boldsymbol{\theta}$ that makes the next DM's decision optimal almost surely, conformal IO achieves upper-bounded POG and AOG. Such uncertainty sets exist because for any $\bar{\boldsymbol{\theta}} \in \mathbb{R}^d$, we can simply set $\alpha = \pi$ to achieve 100% coverage, although the resulting bounds can be large. Instead, we can solve **CP** to calibrate an uncertainty set that achieves close-to-100% coverage using a large validation set. We may also consider adding a small $\Delta_\alpha \in \mathbb{R}_+$ to the α_γ obtained by solving **CP**. Such extra protection can be useful in special cases where all the observed decisions are consistent with one parameter vector (i.e. $\alpha_\gamma = 0$) as it helps to break ties when prescribing new decisions and thus lowers μ_{CIO} (recall Example 3.1). Moreover, we show numerically in Section 5 that, when using $\gamma < 100\%$, conformal IO still demonstrates favorable performance compared to classic IO. In fact, using a $\gamma < \gamma_{\text{max}}$ might yield better performance than using $\gamma = \gamma_{\text{max}}$.

5. Numerical Studies

We next present numerical experiments with shortest path problem (linear program) and knapsack problem (integer program) instances to compare the performance of conformal and classic IO. See Appendix C.2 for the formulations.

5.1. Experiment Setup

Shortest path problem data. We use a 5×5 grid network $G(\mathcal{N}, \mathcal{E})$ where $\mathcal{N}$ and $\mathcal{E}$ indicate the node and edge sets, respectively. For each edge $(i, j) \in \mathcal{E}$, we draw a ground-truth travel cost θ^{ij} uniformly from $[1, 10]$. For each driver $k \in [N]$, we randomly select two distinct nodes $u_k^o, u_k^d \in \mathcal{N}$ as her origin and destination, respectively. We generate her perceived travel cost on edge (i, j) as

$$\hat{\theta}_k^{ij} = (\theta^{ij} * p_k^{ij} + \epsilon_k^{ij})^+ + \epsilon_0 \tag{20}$$

where p_k^{ij} are uniformly drawn from $[1/2, 2]$, ϵ_k^{ij} are drawn from a normal distribution with mean 0 and standard deviation 1, and ϵ_0 is set to 0.1 to ensure $\hat{\theta}_k^{ij}$ is positive.

Knapsack problem data. We consider a knapsack problem of $d = 10$ items. The ground-truth value θ^i and weight w^i of item $i \in [10]$ are drawn uniformly from $[1, 10]$. For each DM $k \in [N]$, we generate a budget $u_k = q_k \sum_i w^i$ where q_k is uniformly drawn from $[1/5, 5]$. DM k's perceived value of item i is generated using Equation (20). We assume all DMs have access to the ground-truth item weights.

Implementation. For both problems, we solve a data-driven IO problem with $\mathbf{\Theta} = \left\{\boldsymbol{\theta} \in \mathbb{R}_+^d \,\middle|\, \|\boldsymbol{\theta} - 1\|_2 \leq d/4\right\}$ and the sub-optimality loss to obtain point estimations. We use a cutting plane method detailed in Appendix C.3 to solve

IO. The calibration problems and their solution methods are in Appendix C.4. When recommending new decisions, the **RFO** is solved with a cutting plane method detailed in Appendix C.5. We implement the conformal IO pipeline with $\gamma \in \{10\%, 30\%, 50\%, 70\%, \gamma_{\text{max}}\}$. The computational setup is summarized in Appendix C.1. We set $\alpha_{\text{min}} = 0.1$.

Evaluation. We perform a 60/20/20 train-validation-test split. The classic IO pipeline uses the union of the training and validation sets. The conformal IO pipeline uses the training set for point estimation and the validation set for calibration. Both methods have access to the same amount of data and are evaluated on the same test set. We repeat this process ten times with different random seeds.

5.2. Results

The value of robustness. As shown in Figure 3, conformal IO typically achieves lower POG and AOG than the classic IO. On average, when varying γ, conformal IO improves the AOG by 20.1–30.4% and the POG by 15.0–23.2% for the shortest path problem, and improves the the AOG by 40.3–57.0% and the POG by 13.5–20.1% for the knapsack problem. When performing head-to-head comparisons, conformal IO outperforms the classic IO 70–80% of the time in AOG and 70–90% of the time in POG for the shortest path problem, and 100% of the time in both AOG and POG for the knapsack problem. The solutions generated by conformal IO are not only of higher quality, but also perceived to be higher quality, and thus are more likely to be adopted.

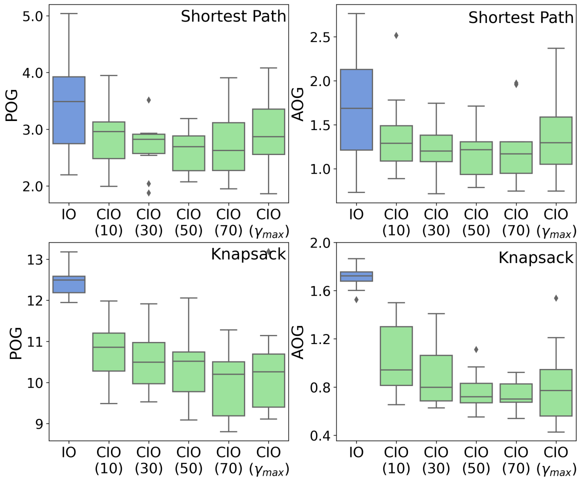


Figure 3. Performance of classic (blue) and conformal IO (green).

The choice of confidence level. We observe that the performance of conformal IO, as measured by both AOG and POG, improves quickly as the value of γ increases from 0 to 50% with diminishing marginal benefits. The performance remains stable when increasing the value of γ from 50% to 100%. When implementing conformal IO, it is possible to improve its out-of-sample performance by carefully tuning the confidence level using a standard cross-validation approach. However, this requires an additional validation dataset. If such a dataset is unavailable, setting γ to a relatively large value usually yields decent performance, which aligns with our theoretical analysis.

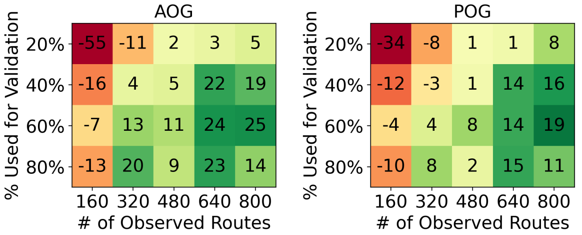


Figure 4. Percentage reduction in test AOG and POG when using the conformal IO (wins in green) vs classic IO (wins in red).

The impact of the train-validation data split. Another important parameter within the conformal IO pipeline is N_{val}. Intuitively, both the point estimation and uncertainty set calibration can benefit from more data. However, when the dataset is small, we need to strike a balance between these two steps aiming to achieve lower AOG and POG. To shed light on this choice, we implement conformal IO for the shortest path problem under different dataset sizes ($N_{\text{train}} + N_{\text{val}} \in \{160, 320, \ldots, 800\}$, corresponding to 20–100% of the dataset used in the previous analysis) and train-validation split ratios ($N_{\text{val}}/(N_{\text{train}}+N_{\text{val}}) \in \{20\%, 40\%, 60\%, 80\%\}$). We set $\gamma = \gamma_{\text{max}}$ for simplicity, which disadvantages our approach. As shown in Figure 4, when the given dataset is very small (160), there is no benefit of using conformal IO simply because we do not have enough data to obtain a good point estimation and a good uncertainty set at the same time. However, the performance of classic IO quickly plateaus as the dataset grows. When given a mid- or large-sized dataset, we can generally benefit from using more data points in the calibration step, echoing our theoretical analysis.

6. Conclusion

In this paper, we propose conformal IO, a novel IO pipeline for recommending high-quality decisions that align with human intuition. We present the first approach to learning uncertainty sets from decision data, which is then utilized in a robust model to prescribe new decisions. Under mild conditions, we prove that conformal IO achieves bounded optimality gaps, with respect to the ground-truth parameters and the DM's perceived parameters. This suggests that decisions may be more likely to be adopted compared to decisions from the classic IO pipeline. Our computational experiments demonstrate the strong performance of conformal IO compared to the classic IO approach.

References

Ahuja, R. K. and Orlin, J. B. Inverse optimization. *Operations Research*, 49(5):771–783, 2001.

Aswani, A., Shen, Z.-J., and Siddiq, A. Inverse optimization with noisy data. *Operations Research*, 66(3):870–892, 2018.

Babier, A., Mahmood, R., McNiven, A. L., Diamant, A., and Chan, T. C. Knowledge-based automated planning with three-dimensional generative adversarial networks. *Medical Physics*, 47(2):297–306, 2020.

Bauer, K., von Zahn, M., and Hinz, O. Please take over: XAI, delegation of authority, and domain knowledge. SSRN, 2023. Available at http://dx.doi.org/10.2139/ssrn.4512594.

Ben-Tal, A. and Nemirovski, A. Robust solutions of uncertain linear programs. *Operations Research Letters*, 25(1): 1–13, 1999.

Ben-Tal, A. and Nemirovski, A. Robust solutions of linear programming problems contaminated with uncertain data. *Mathematical Programming*, 88:411–424, 2000.

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

Berthet, Q., Blondel, M., Teboul, O., Cuturi, M., Vert, J.-P., and Bach, F. Learning with differentiable pertubed optimizers. In *Advances in Neural Information Processing Systems*, volume 33, pp. 9508–9519, 2020.

Bertsimas, D. and Sim, M. The price of robustness. *Operations Research*, 52(1):35–53, 2004.

Bertsimas, D., Pachamanova, D., and Sim, M. Robust linear optimization under general norms. *Operations Research Letters*, 32(6):510–516, 2004.

Bertsimas, D., Gupta, V., and Paschalidis, I. C. Data-driven estimation in equilibrium using inverse optimization. *Mathematical Programming*, 153:595–633, 2015.

Bertsimas, D., Gupta, V., and Kallus, N. Data-driven robust optimization. *Mathematical Programming*, 167:235–292, 2018.

Birge, J. R., Hortaçsu, A., and Pavlin, J. M. Inverse optimization for the recovery of market structure from market outcomes: An application to the miso electricity market. *Operations Research*, 65(4):837–855, 2017.

Birge, J. R., Li, X., and Sun, C. Stochastic inverse optimization, 2022. Available at https://xiaocheng-li.github.io/files/Stochastic_Inverse_Optimization.pdf. Accessed: 2023-01-20.

Burton, J. W., Stein, M.-K., and Jensen, T. B. A systematic review of algorithm aversion in augmented decision making. *Journal of Behavioral Decision Making*, 33(2): 220–239, 2020.

Chan, T. C., Mahmood, R., O'Connor, D. L., Stone, D., Unger, S., Wong, R. K., and Zhu, I. Y. Got (optimal) milk? pooling donations in human milk banks with machine learning and optimization. *Manufacturing & Service Operations Management*, 0(0), 2023a.

Chan, T. C. Y. and Kaw, N. Inverse optimization for the recovery of constraint parameters. *European Journal of Operational Research*, 282(2):415–427, 2020.

Chan, T. C. Y., Craig, T., Lee, T., and Sharpe, M. B. Generalized inverse multiobjective optimization with application to cancer therapy. *Operations Research*, 62(3):680–695, 2014.

Chan, T. C. Y., Lee, T., and Terekhov, D. Inverse optimization: Closed-form solutions, geometry, and goodness of fit. *Management Science*, 65(3):1115–1135, 2019.

Chan, T. C. Y., Mahmood, R., and Zhu, I. Y. Inverse optimization: Theory and applications. *Operations Research*, 0(0), 2023b.

Chen, V., Liao, Q. V., Wortman Vaughan, J., and Bansal, G. Understanding the role of human intuition on reliance in human-AI decision-making with explanations. In *Proceedings of the ACM on Human-Computer Interaction*, volume 7, pp. 1–32, 2023.

Chenreddy, A. R., Bandi, N., and Delage, E. Data-driven conditional robust optimization. In *Advances in Neural Information Processing Systems*, volume 35, pp. 9525–9537, 2022.

Delage, E. and Ye, Y. Distributionally robust optimization under moment uncertainty with application to data-driven problems. *Operations Research*, 58(3):595–612, 2010.

Dietvorst, B. J., Simmons, J. P., and Massey, C. Algorithm aversion: people erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, 144(1):114, 2015.

Dietvorst, B. J., Simmons, J. P., and Massey, C. Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Management Science*, 64(3):1155–1170, 2018.

Donahue, K., Kollias, K., and Gollapudi, S. When are two lists better than one?: Benefits and harms in joint decision-making. *arXiv preprint arXiv:2308.11721*, 2023.

Elmachtoub, A. N. and Grigas, P. Smart "predict, then optimize". *Management Science*, 68(1):9–26, 2022.

Elmachtoub, A. N., Lam, H., Zhang, H., and Zhao, Y. Estimate-then-optimize versus integrated-estimation-optimization: A stochastic dominance perspective. *arXiv preprint arXiv:2304.06833*, 2023.

Esfahani, P. M. and Kuhn, D. Data-driven distributionally robust optimization using the wasserstein metric: performance guarantees and tractable reformulations. *Mathematical Programming*, 171(1-2):115–166, 2018.

Esfahani, P. M., Shafieezadeh-Abadeh, S., Hanasusanto, G. A., and Kuhn, D. Data-driven inverse optimization with imperfect information. *Mathematical Programming*, 167:191–234, 2018.

Gao, R. and Kleywegt, A. Distributionally robust stochastic optimization with wasserstein distance. *Mathematics of Operations Research*, 48(2):603–655, 2023.

Hu, X., Cirit, O., Binaykiya, T., and Hora, R. DeepETA: How uber predicts arrival times using deep learning. Uber Engineering Blog, 2022. Available at https://www.uber.com/en-CA/blog/deepeta-how-uber-predicts-arrival-times/. Accessed: 2024-01-19.

Jussupow, E., Benbasat, I., and Heinzl, A. Why are we averse towards algorithms? a comprehensive literature review on algorithm aversion. In *Proceedings of the 28th European Conference on Information Systems*, 2020.

Kawaguchi, K. When will workers follow an algorithm? a field experiment with a retail business. *Management Science*, 67(3):1670–1695, 2021.

Kesavan, S. and Kushwaha, T. Field experiment on the profit implications of merchants' discretionary power to override data-driven decision-making tools. *Management Science*, 66(11):5182–5190, 2020.

Kizilcec, R. F. How much information? effects of transparency on trust in an algorithmic interface. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pp. 2390–2395, 2016.

Liu, M., Tang, X., Xia, S., Zhang, S., Zhu, Y., and Meng, Q. Algorithm aversion: Evidence from ridesharing drivers. *Management Science*, 0(0), 2023.

Mandi, J., Bucarey, V., Tchomba, M. M. K., and Guns, T. Decision-focused learning: through the lens of learning to rank. In *International Conference on Machine Learning*, pp. 14935–14947. PMLR, 2022.

Meissner, P. and Keding, C. The human factor in AI-based decision-making. *MIT Sloan Management Review*, 63(1): 1–5, 2021.

Merchán, D., Arora, J., Pachon, J., Konduri, K., Winkenbach, M., Parks, S., and Noszek, J. 2021 Amazon last mile routing research challenge: Data set. *Transportation Science*, 0(0), 2022.

Mohri, M., Rostamizadeh, A., and Talwalkar, A. *Foundations of machine learning*. MIT press, 2018.

Nguyen, T. ETA phone home: How uber engineers an efficient route. Uber Engineering Blog, 2015. Available at https://www.uber.com/en-CA/blog/engineering-routing-engine/. Accessed: 2024-01-19.

Rönnqvist, M., Svenson, G., Flisberg, P., and Jönsson, L.-E. Calibrated route finder: Improving the safety, environmental consciousness, and cost effectiveness of truck routing in sweden. *Interfaces*, 47(5):372–395, 2017.

Shafer, G. and Vovk, V. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9(3), 2008.

Sun, C., Liu, L., and Li, X. Predict-then-calibrate: A new perspective of robust contextual LP. In *Advances in Neural Information Processing Systems*, 2023.

Sun, J., Zhang, D. J., Hu, H., and Van Mieghem, J. A. Predicting human discretion to adjust algorithmic prescription: A large-scale field experiment in warehouse operations. *Management Science*, 68(2):846–865, 2022.

Tan, Y., Terekhov, D., and Delong, A. Learning linear programs from optimal decisions. In *Advances in Neural Information Processing Systems*, volume 33, pp. 19738–19749, 2020.

Vovk, V., Gammerman, A., and Shafer, G. *Algorithmic learning in a random world*, volume 29. Springer, 2005.

Wainwright, M. J. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

Wilder, B., Dilkina, B., and Tambe, M. Melding the data-decisions pipeline: Decision-focused learning for combinatorial optimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 1658–1665, 2019.

Yin, M., Wortman Vaughan, J., and Wallach, H. Understanding the effect of accuracy on trust in machine learning models. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pp. 1–12, 2019.

A. Omitted Statements and Proofs in Section 3

A.1. Poof of Lemma 3.5

Proof. We first show that $\hat{\mathbf{x}} \in \{(0,1),(u,0)\}$ almost surely. Let $\delta_u := \arccos\left(1/\sqrt{1+u^2}\right)$, so $\cos\delta_u = 1/\sqrt{1+u^2}$ and $\sin\delta_u = u/\sqrt{1+u^2}$. It is easy to verify that, when $\hat{\boldsymbol{\theta}}_k \in \boldsymbol{\Theta}_1 := \{(\cos\delta,\sin\delta) \mid \delta \in (0,\delta_u]\}$, we have $\hat{\mathbf{x}}_k = \tilde{\mathbf{x}}(\hat{\boldsymbol{\theta}}_k, u) = (0,1)$ almost surely; When $\hat{\boldsymbol{\theta}}_k \in \boldsymbol{\Theta}_2 := \{(\cos\delta,\sin\delta) \mid \delta \in (\delta_u,\pi/2)\}$, we have $\hat{\mathbf{x}}_k = \tilde{\mathbf{x}}(\hat{\boldsymbol{\theta}}_k, u) = (u,0)$ almost surely. Since $\hat{\theta}_k$ is uniformly distributed in $\hat{\boldsymbol{\theta}} \in \boldsymbol{\Theta} = \boldsymbol{\Theta}_1 \cup \boldsymbol{\Theta}_2$, the distribution of $\hat{\mathbf{x}}_k$ is

$$\hat{\mathbf{x}}_k = \left\{ \begin{array}{l} (0,1), \text{ w.p. } 2\delta_u/\pi \\ (u,0), \text{ w.p. } (\pi-2\delta_u)/\pi. \end{array} \right. \tag{21}$$

Given a sample set $\mathcal{D} = \{\mathbf{u}_k, \hat{\mathbf{x}}_k\}_{k\in[N]}$, let N_1 and N_2, respectively, denote the numbers of $(0,1)$ and $(u,0)$ in $\mathcal{D}$. We next show that when $N_1 > 0$ and $N_2 > 0$, θ_u is the unique optimal solution to $\mathbf{IO}(\mathcal{D})$. Specifically, in Example 3.1, $\mathbf{IO}(\mathcal{D})$ is presented as follows.

$$\bar{\boldsymbol{\theta}}_N := \underset{\boldsymbol{\theta}\in\boldsymbol{\Theta}}{\operatorname{argmin}} \quad \frac{N_1}{N} l_1(\boldsymbol{\theta}) + \frac{N_2}{N} l_2(\boldsymbol{\theta}) \tag{22}$$

where

$$l_1(\boldsymbol{\theta}) = \begin{cases} 0, & \text{if } \boldsymbol{\theta} \in \boldsymbol{\Theta}_1, \\ \theta_2 - u\theta_1, & \text{if } \boldsymbol{\theta} \in \boldsymbol{\Theta}_2, \end{cases} \tag{23}$$

and

$$l_2(\boldsymbol{\theta}) = \begin{cases} u\theta_1 - \theta_2, & \text{if } \boldsymbol{\theta} \in \boldsymbol{\Theta}_1, \\ 0, & \text{if } \boldsymbol{\theta} \in \boldsymbol{\Theta}_2. \end{cases} \tag{24}$$

A simple calculation gives that when $N_1 > 0$ and $N_2 > 0$, the minimum is 0 which occurs uniquely at $\boldsymbol{\theta} = (\cos\delta_u, \sin\delta_u)$; When $N_2 = 0$, the minimum is 0 which occurs when $\boldsymbol{\theta} \in \boldsymbol{\Theta}_1$; When $N_1 = 0$, the minimum is 0 which occurs when $\boldsymbol{\theta} \in \boldsymbol{\Theta}_2$. Therefore, we have

$$\mathbb{P}(N_1 N_2 > 0) \leq \mathbb{P}\left(\bar{\boldsymbol{\theta}}_N = (\cos\delta_u, \sin\delta_u)\right) \leq 1. \tag{25}$$

Given the probability distribution given in Equation (21) and that $\mathcal{D}$ is generated using i.i.d. samples from $\mathbb{P}_\theta$, we have

$$\mathbb{P}(N_1 N_2 > 0) = 1 - \left(\frac{2\delta_u}{\pi}\right)^N - \left(1 - \frac{2\delta_u}{\pi}\right)^N, \tag{26}$$

which converges to 1 as N goes to infinity. Therefore, we conclude that $\mathbb{P}(\bar{\boldsymbol{\theta}}_N = (\cos\delta_u, \sin\delta_u))$ converges to 1 as N goes to infinity. □

A.2. Proof of Proposition 3.6

Proof. According to Lemma 3.5, in Example 3.1, when using **IO** with the sub-optimality loss (6), the probability of the estimated parameter being $\bar{\boldsymbol{\theta}} = \boldsymbol{\theta}_u := (\cos\delta_u, \sin\delta_u)$ goes to one as N goes to infinity, where $\delta_u := \arccos(1/\sqrt{1+u^2})$. In fact, as long as our decision dataset contains both $(0,1)$ and $(u,0)$, the estimated parameter will be $\bar{\boldsymbol{\theta}} = \boldsymbol{\theta}_u$. Hence, it suffices to show that the decision policy based on this estimation, i.e. $\bar{\mathbf{x}}_{\text{IO}}(u) := \tilde{\mathbf{x}}(\boldsymbol{\theta}_u, u)$, can achieve unbounded AOG and POG as we change u, which we prove in the following two lemmas.

Lemma A.1. *In Example 3.1, let* $\bar{\mathbf{x}}_{\text{IO}}(u) = \tilde{\mathbf{x}}(\boldsymbol{\theta}_u, u)$. *For any* $v \in \mathbb{R}_+$ *there exists some* $\bar{u} > 1$ *such that* $\text{AOG}(\bar{\mathbf{x}}_{\text{IO}}) > v$ *for any* $u > \bar{u}$.

Proof. According to the definition of $\tilde{\mathbf{x}}$, we know that $\bar{\mathbf{x}}_{\text{IO}}(u)$ is uniformly drawn from

$$\mathcal{X}^{\text{OPT}}(\boldsymbol{\theta}_u, u) = \left\{ \left(\frac{ut}{\sqrt{u^2+1}}, 1 - \frac{t}{\sqrt{u^2+1}} \right) \middle| \, t \in \left[0, \sqrt{u^2+1}\right] \right\}. \tag{27}$$

Since the ground-truth $\boldsymbol{\theta}^* = (\cos(\pi/4), \sin(\pi/4))$, the true optimal solution is $\mathbf{x}^* = (0,1)$ with $\tilde{f}(\boldsymbol{\theta}^*, u) = \sqrt{2}/2$. Hence, we have

$$\text{AOG}(\bar{\mathbf{x}}_{\text{IO}}) = \int_0^{\sqrt{u^2+1}} \frac{\sqrt{2}}{2\sqrt{u^2+1}} \left(1 - \frac{t}{\sqrt{u^2+1}} + \frac{ut}{\sqrt{u^2+1}}\right) dt - \frac{\sqrt{2}}{2} = \frac{\sqrt{2}(u-1)}{4} \tag{28}$$

Therefore, for any $v \in \mathbb{R}_+$, there exists $\bar{u} = 2\sqrt{2}v + 1$ such that $\text{AOG}(\bar{\mathbf{x}}_{\text{IO}}) > v$ for any $u > \bar{u}$.

Lemma A.2. *In Example 3.1, let $\bar{\mathbf{x}}_{\text{IO}}(u) = \tilde{\mathbf{x}}(\boldsymbol{\theta}_u, u)$. for any $v \in \mathbb{R}_+$ there exists some $\bar{u} > 1$ such that $\text{POG}(\bar{\mathbf{x}}_{\text{IO}}) > v$ for any $u > \bar{u}$.*

Proof. According to the definition of $\tilde{\mathbf{x}}$, $\bar{\mathbf{x}}_{\text{IO}}(u)$ is uniformly drawn from

$$\mathcal{X}^{\text{OPT}}(\boldsymbol{\theta}_u, u) = \left\{ \left(\frac{ut}{\sqrt{u^2+1}}, 1 - \frac{t}{\sqrt{u^2+1}} \right) \middle| \, t \in \left[0, \sqrt{u^2+1}\right] \right\}. \tag{29}$$

It is easy to verify that, when $\hat{\boldsymbol{\theta}} \in \boldsymbol{\Theta}_1 := \{(\cos\delta, \sin\delta) \,|\, \delta \in (0, \delta_u]\}$, we have $\hat{\mathbf{x}}_k = \tilde{\mathbf{x}}(\hat{\boldsymbol{\theta}}, u) = (0,1)$ with $\tilde{f}(\hat{\boldsymbol{\theta}}, u) = \hat{\theta}_2$ almost surely; When $\hat{\boldsymbol{\theta}} \in \boldsymbol{\Theta}_2 := \{(\cos\delta, \sin\delta) \,|\, \delta \in (\delta_u, \pi/2\}$, we have $\hat{\mathbf{x}}_k = \tilde{\mathbf{x}}(\hat{\boldsymbol{\theta}}, u) = (u, 0)$ with $\tilde{f}(\hat{\boldsymbol{\theta}}, u) = u\hat{\theta}_1$ almost surely. Since the optimal solution drawn from $\mathcal{X}^{\text{OPT}}(\boldsymbol{\theta}_u, u)$ is independent of the DM's perception $\hat{\boldsymbol{\theta}}$, we have

$$\begin{aligned}
\text{POG}(\bar{\mathbf{x}}_{\text{IO}}) = & \int_0^{\delta_u} \int_0^{\sqrt{u^2+1}} \frac{1}{\sqrt{u^2+1}} \left[\frac{ut}{\sqrt{u^2+1}} \cos\delta + \left(1 - \frac{t}{\sqrt{u^2+1}}\right) \sin\delta - \sin\delta \right] dt\, d\delta \\
& + \int_{\delta_u}^{\pi/2} \int_0^{\sqrt{u^2+1}} \frac{1}{\sqrt{u^2+1}} \left[\frac{ut}{\sqrt{u^2+1}} \cos\delta + \left(1 - \frac{t}{\sqrt{u^2+1}}\right) \sin\delta - u\cos\delta \right] dt\, d\delta \\
= & \frac{1}{2} \int_0^{\delta_u} (u\cos\delta - \sin\delta)\, d\delta + \frac{1}{2} \int_{\delta_u}^{\pi/2} (-u\cos\delta + \sin\delta)\, d\delta \\
= & \sqrt{1+u^2} - \frac{u+1}{2}. \\
> & \frac{u-1}{2}
\end{aligned}$$

The inequality holds because $\sqrt{1+u^2} > u$. Therefore, we have, for any $v \in \mathbb{R}_+$, there exists $\bar{u} = 2v + 1$ such that $\text{POG}(\bar{\mathbf{x}}_{\text{IO}}) > v$ for any $u > \bar{u}$. □

Based on Lemmas A.1 and A.2, we conclude that $\bar{\mathbf{x}}_{\text{IO}}$ can achieve unbounded AOG and POG. □

A.3. Proof of Lemma 3.8

Proof. According to the proof of Lemma 3.5, we know that the distribution of $\hat{\mathbf{x}}_k$ is

$$\hat{\mathbf{x}}_k = \begin{cases} (0,1), & \text{w.p. } 2\delta_u/\pi \\ (u,0), & \text{w.p. } (\pi - 2\delta_u)/\pi. \end{cases} \tag{30}$$

Moreover, when setting $\boldsymbol{\theta} = \boldsymbol{\theta}_u$, the sub-optimality losses associated with both (0, 1) and $(u, 0)$ are zero. Hence, the minimum of the distributionally robust sub-optimality loss equals zero, which occurs at $\boldsymbol{\theta} = \boldsymbol{\theta}_u$, because the distributionally robust sub-optimality loss is a weighted sum of the sub-optimality losses associated with (0, 1) and $(u, 0)$.

□

A.4. Proof of Lemma 3.10

Proof. We first show that when $\bar{\boldsymbol{\theta}} = \boldsymbol{\theta}_u$ and $\alpha \in (0, \pi/2)$, $\mathbf{RFO}\left(\mathcal{C}(\bar{\boldsymbol{\theta}}, \alpha), u\right)$ has a unique optimal solution $(0,1)$. Let $\mathbf{x}_1 = (0,1)$, $\mathbf{x}_2 = (0,2)$, $\mathbf{x}_3 = (u,0)$ and $\mathbf{x}_4 = (u,2)$ denote the four extreme points of the feasible region $\mathcal{X}(u)$, respectively, and

$$R(\mathbf{x}) := \max_{\boldsymbol{\theta} \in \mathcal{C}(\boldsymbol{\theta}_u, \alpha)} \theta_1 x_1 + \theta_2 x_2. \tag{31}$$

Since **FO** is a linear program, it suffices to show that, when $\alpha \in (0, \pi/2)$, $R(\mathbf{x}_1) < \min\{R(\mathbf{x}_2), R(\mathbf{x}_3), R(\mathbf{x}_4)\}$ because, if there exists an optimal solution that is not an extreme point, then there must exist another extreme point $\mathbf{x}_i$ such that $R(\mathbf{x}_1) = R(\mathbf{x}_i)$ where $i \neq 1$. Next, we compare $R(\mathbf{x}_1)$ with $R(\mathbf{x}_2)$, $R(\mathbf{x}_3)$, and $R(\mathbf{x}_4)$.

It is easy to verify that

$$R(\mathbf{x}_1) = \begin{cases} \sin(\delta_u + \alpha), & \text{if } \alpha \in (0, \pi/2 - \delta_u], \\ 1, & \text{if } \alpha \in (\pi/2 - \delta_u, \pi/2). \end{cases} \tag{32}$$

For $\mathbf{x}_2$, we have

$$R(\mathbf{x}_2) = \begin{cases} 2\sin(\delta_u + \alpha), & \text{if } \alpha \in (0, \pi/2 - \delta_u], \\ 2, & \text{if } \alpha \in (\pi/2 - \delta_u, \pi/2). \end{cases} \tag{33}$$

Hence, we have $R(\mathbf{x}_1) < R(\mathbf{x}_2)$ when $\alpha \in (0, \pi/2)$.

For $\mathbf{x}_3$, we have

$$R(\mathbf{x}_3) = \begin{cases} u\cos(\delta_u - \alpha), & \text{if } \alpha \in (0, \delta_u], \\ u, & \text{if } \alpha \in (\delta_u, \pi/2). \end{cases} \tag{34}$$

Since $u > 1$, we have $\pi/2 - \delta_u < \pi/4 < \delta_u < \pi/2$. We will show that $R(\mathbf{x}_1) < R(\mathbf{x}_3)$ when α is in $(0, \pi/2 - \delta_u)$, $[\pi/2 - \delta_u, \delta_u)$, and $[\delta_u, \pi/2)$. When $\alpha \in (0, \pi/2 - \delta_u)$, we have

$$\begin{aligned} R(\mathbf{x}_1) &= \sin(\delta_u + \alpha) \\ &= \sin\delta_u \cos\alpha + \cos\delta_u \sin\alpha \\ &= \frac{u}{\sqrt{1+u^2}}\cos\alpha + \frac{1}{\sqrt{1+u^2}}\sin\alpha \\ &< \frac{u}{\sqrt{1+u^2}}\cos\alpha + \frac{u^2}{\sqrt{1+u^2}}\sin\alpha \\ &= u\left(\frac{1}{\sqrt{1+u^2}}\cos\alpha + \frac{u}{\sqrt{1+u^2}}\sin\alpha\right) \\ &= u\left(\cos\delta_u\cos\alpha + \sin\delta_u\sin\alpha\right) \\ &= u\cos(\delta_u - \alpha) \\ &= R(\mathbf{x}_3). \end{aligned}$$

The second line holds due to the sum of angles identity. The third line holds due to the definition of δ_u. The fourth line holds because $u > 1$. The fifth line is obtained by simple manipulation. The sixth line holds due to the definition of δ_u. The seventh line holds due to the sum of angles identity.

When $\alpha \in [\pi/2 - \delta_u, \delta_u)$, we have

$$\begin{aligned} R(\mathbf{x}_1) &= 1 \\ &< 1 + \frac{u-1}{u^2+1} \\ &= \frac{u}{\sqrt{u^2+1}}\frac{1}{\sqrt{u^2+1}} + \frac{u^2}{\sqrt{u^2+1}}\frac{1}{\sqrt{u^2+1}} \\ &= u\cos\delta_u\frac{1}{\sqrt{u^2+1}} + u\sin\delta_u\frac{1}{\sqrt{u^2+1}} \\ &< u\cos\delta_u\cos\alpha + u\sin\delta_u\sin\alpha \\ &= u\cos(\delta_u - \alpha) \\ &= R(\mathbf{x}_3). \end{aligned}$$

The second line holds because $u > 1$. The third line is obtained through simple manipulation. The forth line holds due to the definition of δ_u. For the fifth line, we know that $\alpha \in [\pi/2 - \delta_u, \delta_u) \subseteq [\pi/4 - \pi/2]$ where $\cos\alpha$ is strictly decreasing in α and where $\sin\alpha$ is strictly increasing in α. Therefore, $\cos\alpha < \cos\delta_u = 1/\sqrt{u^2+1}$ and $\sin\alpha \leq \sin(\pi/2 - \delta_u) = \cos\delta_u = 1/\sqrt{u^2+1}$. Hence, the fifth line holds. The sixth line holds due to the sum of angles identity.

When $\alpha \in [\delta_u, \pi/2)$, we have $R(\mathbf{x}_1) = 1 < u = R(\mathbf{x}_3)$.

Hence, $R(\mathbf{x}_1) < R(\mathbf{x}_3)$ when $\alpha \in (0, \pi/2)$.

For $\mathbf{x}_4$, we have

$$R(\mathbf{x}_4) = \max_{\delta \in \mathcal{C}(\delta_u, \alpha)} u\cos\delta + 2\sin\delta. \tag{35}$$

Let δ_1^* denote the optimal solution to the maximization problem for calculating $R(\mathbf{x}_1)$. It is easy to verify that $\delta_1^* \in (0, \pi/2)$ when $\alpha \in (0, \pi/2)$. So $\cos\delta_1^* > 0$ and $\sin\delta_1^* > 0$. Hence, we have

$$R(\mathbf{x}_4) = \max_{\delta \in \mathcal{C}(\delta_u, \alpha)} u\cos\delta + 2\sin\delta \geq u\cos\delta_1^* + 2\sin\delta_1^* > \sin\delta_1^* = R(\mathbf{x}_1). \tag{36}$$

The first inequality holds because δ_1^* may not be the maximizer of the problem associated with $\mathbf{x}_4$. The second inequality holds because $u > 1$, $\cos\delta_1^* > 0$, and $\sin\delta_1^* > 0$.

Hence, when $\alpha \in (0, \pi/2)$, **RFO** $(\mathcal{C}(\delta_u, \alpha), u)$ has a unique optimal solution $\mathbf{x}_1$, so $\bar{\mathbf{x}}_{\text{CIO}}(u) = (0, 1)$ almost surely. Given that $(0, 1)$ is also the optimal solution to $\mathbf{FO}(\delta^*, u)$, we have $\text{AOG}(\bar{\mathbf{x}}_{\text{CIO}}) = 0$ when $\bar{\boldsymbol{\theta}}_N = \boldsymbol{\theta}_u$ and $\alpha \in (0, \pi/2)$. According to Lemma 3.5, we know that $\mathbb{P}(\bar{\boldsymbol{\theta}}_N = \boldsymbol{\theta}_u) \to 1$ as $N \to \infty$. So we conclude that, when $\alpha \in (0, \pi/2)$, we have $\mathbb{P}\left[\text{AOG}(\bar{\mathbf{x}}_{\text{CIO}}) = 0\right] \to 1$ as $N \to \infty$.

□

A.5. Proof of Lemma 3.11

Proof. As shown in the proof of Lemma 3.10, when $\alpha \in (0, \pi/2)$, the **RFO** $(\mathcal{C}(\boldsymbol{\theta}_u, \alpha), u)$ has a unique optimal solution $(0, 1)$. So $\bar{\mathbf{x}}_{\text{CIO}}(u) = (0, 1)$ almost surely, when $\bar{\boldsymbol{\theta}}_N = \boldsymbol{\theta}_u$ and $\alpha \in (0, \pi/2)$. It is easy to verify that, when $\hat{\boldsymbol{\theta}} \in \boldsymbol{\Theta}_1 := \{(\cos\delta, \sin\delta) \mid \delta \in (0, \delta_u]\}$, we have $\hat{\mathbf{x}}_k = \tilde{\mathbf{x}}(\hat{\boldsymbol{\theta}}, u) = (0, 1)$ almost surely; When $\hat{\boldsymbol{\theta}} \in \boldsymbol{\Theta}_2 := \{(\cos\delta, \sin\delta) \mid \delta \in (\delta_u, \pi/2)\}$, we have $\hat{\mathbf{x}}_k = \tilde{\mathbf{x}}(\hat{\boldsymbol{\theta}}, u) = (u, 0)$ almost surely. Hence, we have

$$\text{POG}(\bar{\mathbf{x}}_{\text{CIO}}) = \int_0^{\delta_u} \frac{\pi}{2} \times 0 \, d\delta + \int_{\delta_u}^{\pi/2} \frac{\pi}{2} \times \sin\delta \, d\delta = -\frac{\pi}{2}\cos\delta \Big|_{\delta_u}^{\pi/2} = \frac{\pi}{2\sqrt{1+u^2}} < \frac{\pi}{2\sqrt{2}}. \tag{37}$$

The inequality holds because $u > 1$.

According to Lemma 3.5, we know that $\mathbb{P}(\bar{\boldsymbol{\theta}}_N = \boldsymbol{\theta}_u) \to 1$ as $N \to \infty$. So we conclude that, when $\alpha \in (0, \pi/2)$, we have $\mathbb{P}\left[\text{POG}(\bar{\mathbf{x}}_{\text{CIO}}) < \pi/2\sqrt{2}\right] \to 1$ as $N \to \infty$. □

B. Proof of Statements in Section 4

B.1. Definitions

Definition B.1 (Empirical Rademacher Complexity)**.** Let $\mathcal{F}$ be a class of functions mapping from $\mathcal{Z} = \{Z_1, Z_2, \ldots, Z_m\}$ to $[a, b]$ and $\mathcal{D}$ be a fixed sample of size N with elements in $\mathcal{Z}$, then the empirical Rademacher Complexity of $\mathcal{F}$ with respect to the sample $\mathcal{D}$ is defined as

$$\hat{\mathfrak{R}}_{\mathcal{D}}(\mathcal{F}) := \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{f \in \mathcal{F}} \frac{1}{N} \sum_{i \in [N]} \sigma_i f(Z_i)\right] \tag{38}$$

where $\boldsymbol{\sigma} = (\sigma_1, \sigma_2, \ldots, \sigma_N)^\intercal$ with σ_i's being independent uniform random variables taking values in $\{-1, 1\}$.

Definition B.2 (Rademacher Complexity)**.** Let $\mathbb{P}$ denote the distribution according to which samples are drawn. For any integer $N \geq 1$, the Rademacher complexity of a function class $\mathcal{F}$ is the expectation of the empirical Rademacher complexity over the samples of size N drawn from $\mathbb{P}$:

$$\mathfrak{R}_N(\mathcal{F}) := \mathbb{E}_{\mathcal{D} \sim \mathbb{P}^N}\left[\hat{\mathfrak{R}}_{\mathcal{D}}(\mathcal{F})\right] \tag{39}$$

Definition B.3 (Growth Function)**.** Let $\mathcal{H}$ be a class of functions that take values in $\{-1, 1\}$. The growth function $\Pi_{\mathcal{H}} : \mathbb{N} \to \mathbb{N}$ for $\mathcal{H}$ is defined as

$$\Pi_{\mathcal{H}}(N) := \max_{(Z_1, Z_2, \ldots, Z_N) \in \mathcal{Z}^N} |\{(h(Z_1), h(Z_2), \ldots, h(Z_N)) \,|\, h \in \mathcal{H}\}| \tag{40}$$

which measures the maximum number of distinct ways in which N data points in $\mathcal{Z}$ can be classified using the function class $\mathcal{H}$.

B.2. Useful Lemmas

Lemma B.4 (Corollary 3.1 in Mohri et al. (2018))**.** *Let $\mathcal{H}$ be a class of functions taking values in $\{1, -1\}$, then, for any integer $N \geq 1$, the following holds*

$$\mathfrak{R}_N(\mathcal{H}) \leq \sqrt{\frac{2 \log \Pi_{\mathcal{H}}(N)}{N}}. \tag{41}$$

Lemma B.5 (Theorem 4.10 in Wainwright (2019))**.** *For any b-uniformly bounded class of functions $\mathcal{F}$, any positive integer $N \geq 1$, and any scalar $\delta \geq 0$, with probability at least $1 - \exp\left(-N\delta^2/(2b^2)\right)$, we have*

$$\sup_{f \in \mathcal{F}} \left| \frac{1}{N} \sum_{i \in [N]} f(X_i) - \mathbb{E}\left[f(X_i)\right] \right| \leq 2\mathfrak{R}_N(\mathcal{F}) + \delta \tag{42}$$

where $\mathfrak{R}(\mathcal{F})$ denotes the Rademacher complexity of the function class $\mathcal{F}$.

B.3. Proof of Theorem 4.3

Proof. For convenience, we define $\hat{\boldsymbol{\Theta}}_k := \boldsymbol{\Theta}^{\text{OPT}}(\mathbf{u}_k, \hat{\mathbf{x}}_k)$ for any $k \in [N]$.

We first present the extensive formulation of Problem (12). When $\alpha \in [0, \pi]$, $\cos \alpha$ is a strictly decreasing in α. Therefore, minimizing α is equivalent to maximizing the value of $\cos \alpha$. We can replace the decision variable α in Problem (12) with a new decision variable $c := \cos \alpha$ with an additional constraint t with $-1 \leq c \leq 1$. In addition, we introduce a new set of decision variables $y_k \in \{0, 1\}$ that indicate if $\hat{\boldsymbol{\Theta}}_k$ intersects with the learned uncertainty set ($= 1$) or not ($= 0$) for any $k \in \mathcal{K}_{\text{val}}$. Problem (12) can be presented as follows.

$$\begin{aligned}
\underset{c, \{\boldsymbol{\theta}_k\}_{k \in \mathcal{K}_{\text{val}}}, \{y_k\}_{k \in \mathcal{K}_{\text{val}}}}{\text{maximize}} \quad & c && \text{(43a)} \\
\text{subject to} \quad & \hat{\mathbf{x}}_k \in \mathcal{X}^{\text{OPT}}(\boldsymbol{\theta}_k, \mathbf{u}_k), \quad \forall k \in \mathcal{K}_{\text{val}} && \text{(43b)} \\
& \boldsymbol{\theta}_k^{\intercal} \bar{\boldsymbol{\theta}} \geq c + 2(y_k - 1), \quad \forall k \in \mathcal{K}_{\text{val}} && \text{(43c)} \\
& \sum_{k \in \mathcal{K}_{\text{val}}} y_k \geq \lceil \gamma(N_{\text{val}} + 1) \rceil && \text{(43d)} \\
& \|\boldsymbol{\theta}_k\|_2 = 1, \quad \forall k \in \mathcal{K}_{\text{val}} && \text{(43e)} \\
& -1 \leq c \leq 1 && \text{(43f)} \\
& y_k \in \{0, 1\}, \quad \forall k \in \mathcal{K}_{\text{val}}. && \text{(43g)}
\end{aligned}$$

Constraints (43b) ensure that $\boldsymbol{\theta}_k$ is a member of $\hat{\boldsymbol{\Theta}}_k$ for any $k \in \mathcal{K}_{\text{val}}$. Constraints (43c) decide if $\boldsymbol{\theta}_k$ should be taken into account when calculating the maximal cosine value c based on if $\hat{\boldsymbol{\Theta}}_k$ intersects with $\mathcal{C}$. Constraint (43d) ensures that $\mathcal{C}$ intersects with at least $\lceil \gamma(N_{\text{val}} + 1) \rceil$ inverse feasible sets. Constraint (43e) enforces $\boldsymbol{\theta}_k$ to be on the unit sphere as defined in Equation (11). Constraints (43f)–(43g) specify the ranges of the decision variables.

Observing that the objective of Problem (43) is to maximize c and that decision variables $\boldsymbol{\theta}_k$ of different data points only interact in Constraints (43c). We can re-write Problem (43) as

$$
\begin{aligned}
\text{maximize} \quad & c && \text{(44a)}\\
\text{subject to} \quad & c \leq c_k - 2(y_k - 1), \quad \forall k \in \mathcal{K}_{\text{val}} && \text{(44b)}\\
& \sum_{k \in \mathcal{K}_{\text{val}}} y_k \geq \lceil \gamma(N_{\text{val}} + 1) \rceil && \text{(44c)}\\
& -1 \leq c \leq 1 && \text{(44d)}\\
& y_k \in \{0, 1\}, \quad \forall k \in \mathcal{K}_{\text{val}}, && \text{(44e)}
\end{aligned}
$$

where

$$
\begin{aligned}
c_k := \underset{\boldsymbol{\theta}_k}{\text{maximize}} \quad & \boldsymbol{\theta}_k^{\intercal} \bar{\boldsymbol{\theta}} && \text{(45a)}\\
\text{subject to} \quad & \hat{\mathbf{x}}_k \in \mathcal{X}^{\text{OPT}}(\boldsymbol{\theta}_k, \mathbf{u}_k) && \text{(45b)}\\
& \|\boldsymbol{\theta}_k\|_2 \leq 1. && \text{(45c)}
\end{aligned}
$$

Note that we replace Constraints (43e) with Constraints (45c) because the objective of Problem (45) is to maximize the inner product of $\boldsymbol{\theta}_k$ and $\bar{\boldsymbol{\theta}}$, so the maximum only occurs when $\|\boldsymbol{\theta}_k\|_2 = 1$. We further observe that the optimal solution to Problem (44a) is to set $y_k = 1$ for all k such that $c_k \geq \Gamma_\tau\left(\{c_k\}_{k \in \mathcal{K}_{\text{val}}}\right)$ and $y_k = 0$ otherwise. Therefore, the optimal objective value of Problem (44a) is $c = \Gamma_\tau\left(\{c_k\}_{k \in \mathcal{K}_{\text{val}}}\right)$ corresponding to $\alpha_\gamma = \arccos \Gamma_\tau\left(\{c_k\}_{k \in \mathcal{K}_{\text{val}}}\right)$.

□

B.4. Proof of Theorem 4.6

Proof. We first prove the learned uncertainty set is conservatively valid. Following the conformal prediction language used by Vovk et al. (2005), we define a conformality measure of each data point,i.e. an observed decision and exogenous parameter pair, $A_{\bar{\boldsymbol{\theta}}} : \mathbb{R}^n \times \mathcal{U} \to \mathbb{R}_+$ as follows

$$
\begin{aligned}
A_{\bar{\boldsymbol{\theta}}}(\hat{\mathbf{x}}, \mathbf{u}) := \underset{\boldsymbol{\theta}}{\text{maximize}} \quad & \boldsymbol{\theta}^{\intercal} \bar{\boldsymbol{\theta}} && \text{(46a)}\\
\text{subject to} \quad & \boldsymbol{\theta} \in \boldsymbol{\Theta}^{\text{OPT}}(\hat{\mathbf{x}}, \mathbf{u}) && \text{(46b)}\\
& \|\boldsymbol{\theta}\|_2 \leq 1. && \text{(46c)}
\end{aligned}
$$

We note that $c_k = A_{\bar{\boldsymbol{\theta}}}(\hat{\mathbf{x}}_k, \mathbf{u}_k)$ for any $k \in \mathcal{K}_{\text{val}}$ where c_k is defined in Theorem 4.3. Let $\tau = \lceil \gamma(N_{\text{val}} + 1) \rceil$, and $\mathcal{A} := \{A_{\bar{\boldsymbol{\theta}}}(\hat{\mathbf{x}}_k, \mathbf{u}_k)\}_{k \in \mathcal{K}_{\text{val}}}$, or equivalently, $\mathcal{A} := \{c_k\}_{k \in \mathcal{K}_{\text{val}}}$. Due to the definition of $\mathcal{C}\left(\bar{\boldsymbol{\theta}}, \alpha\right)$ and that α is chosen such that $\cos \alpha = \Gamma^\tau(\mathcal{A})$, the event "$\boldsymbol{\Theta}^{\text{OPT}}(\hat{\mathbf{x}}, \mathbf{u}) \cap \mathcal{C}(\bar{\boldsymbol{\theta}}, \alpha) \neq \varnothing$" is equivalent to "$A_{\bar{\boldsymbol{\theta}}}(\hat{\mathbf{x}}, \mathbf{u}) \geq \Gamma^\tau(\mathcal{A})$", so

$$
\mathbb{P}\left(\boldsymbol{\Theta}^{\text{OPT}}(\hat{\mathbf{x}}, \mathbf{u}) \cap \mathcal{C}(\bar{\boldsymbol{\theta}}, \alpha) \neq \varnothing\right) = \mathbb{P}\left(A_{\bar{\boldsymbol{\theta}}}(\hat{\mathbf{x}}, \mathbf{u}) \geq \Gamma^\tau(\mathcal{A})\right). \tag{47}
$$

Assumption 4.5 implies that the dataset $\mathcal{D}' = \mathcal{D}_{\text{val}} \cup \{(\hat{\mathbf{x}}, \mathbf{u})\}$ is exchangeable, i.e. the ordering of the data points in $\mathcal{D}'$ does not affect its joint probability distribution (Shafer & Vovk, 2008). Therefore, the rank (from high to low) of $A_{\bar{\boldsymbol{\theta}}}(\hat{\mathbf{x}}, \mathbf{u})$ in $\mathcal{A}' := \mathcal{A} \cup \{A_{\bar{\boldsymbol{\theta}}}(\hat{\mathbf{x}}, \mathbf{u})\}$ is uniformly distributed in $\{1, 2, \ldots, N_{\text{val}} + 1\}$. So, we have

$$
\begin{aligned}
\gamma &\leq \mathbb{P}\left\{A_{\bar{\boldsymbol{\theta}}}(\hat{\mathbf{x}}, \mathbf{u}) \geq \Gamma^\tau(\mathcal{A}')\right\}\\
&= 1 - \mathbb{P}\left\{A_{\bar{\boldsymbol{\theta}}}(\hat{\mathbf{x}}, \mathbf{u}) < \Gamma^\tau(\mathcal{A}')\right\}\\
&= 1 - \mathbb{P}\left\{A_{\bar{\boldsymbol{\theta}}}(\hat{\mathbf{x}}, \mathbf{u}) < \Gamma^\tau(\mathcal{A})\right\}\\
&= \mathbb{P}\left\{A_{\bar{\boldsymbol{\theta}}}(\hat{\mathbf{x}}, \mathbf{u}) \geq \Gamma^\tau(\mathcal{A})\right\}\\
&= \mathbb{P}\left\{\boldsymbol{\Theta}^{\text{OPT}}(\hat{\mathbf{x}}, \mathbf{u}) \cap \mathcal{C}(\bar{\boldsymbol{\theta}}, \alpha) \neq \varnothing\right\}.
\end{aligned}
$$

The first line holds due to the definition of τ. We obtain the second line by taking the complement of the event in the first line (inside the probability). The third line holds because $A_{\bar{\boldsymbol{\theta}}}(\hat{\mathbf{x}}, \mathbf{u})$ can never be strictly smaller than itself, so any elements in $\mathcal{A}'$ that are strictly smaller than $A_{\bar{\boldsymbol{\theta}}}(\hat{\mathbf{x}}, \mathbf{u})$ are in $\mathcal{A}$. Note that this line holds only when $\tau \leq N_{\text{val}}$, which occurs when $\gamma \leq N_{\text{val}}/(N_{\text{val}}+1)$, because $\mathcal{A}$ only has N_{val} elements. We obtain the third line by taking the complement of the event in the second line (inside the probability). The last line holds due to Equation (47). We note that all the probabilities are over the joint distribution of $\mathcal{D}_{\text{val}}$ and the new sample, i.e. $\mathcal{D}'$.

We next prove that the learned uncertainty set is asymptotically exact. Let $z_k := (\mathbf{u}_k, \hat{\mathbf{x}}_k)$, $\mathcal{Z} := \{z_k\}_{k \in \mathcal{K}_{\text{val}}}$. We define a function class

$$\mathcal{H} = \left\{ h(z, \alpha) = \mathbb{1}\left[\boldsymbol{\Theta}^{\text{OPT}}(\hat{\mathbf{x}}, \mathbf{u}) \cap \mathcal{C}(\bar{\boldsymbol{\theta}}, \alpha)\right] \,\middle|\, \alpha \in (0, \pi) \right\}. \tag{48}$$

Let $\Pi_{\mathcal{H}}$ denote the growth function of $\mathcal{H}$ as defined in Definition B.3. It is easy to verify that

$$\Pi_{\mathcal{H}}(N_{\text{val}}) = N_{\text{val}} + 1 \tag{49}$$

because the value of $h(z, \alpha)$ is monotonically increasing in α for any fixed $z \in \mathcal{Z}$, so changing the value of α can only leads to $N_{\text{val}} + 1$ different outcomes for a fixed dataset $\mathcal{Z}$.

Therefore, according to Lemma B.4, we have

$$\mathfrak{R}_{N_{\text{val}}}(\mathcal{H}) \leq \sqrt{\frac{2\log(N_{\text{val}}+1)}{N_{\text{val}}}} \tag{50}$$

where $\mathfrak{R}_{N_{\text{val}}}(\mathcal{H})$ denotes the Rademacher complexity of $\mathcal{H}$ when sample size is N_{val}, as defined in Definition B.2.

We know that the value of α is chosen such that it is the smallest value that satisfies

$$\frac{1}{N_{\text{val}}} \sum_{k \in \mathcal{K}_{\text{val}}} h(z_k, \alpha) = \frac{1}{N_{\text{val}}} \sum_{k \in \mathcal{K}_{\text{val}}} \mathbb{1}\left[\boldsymbol{\Theta}^{\text{OPT}}(\hat{\mathbf{x}}_k, \mathbf{u}_k) \cap \mathcal{C}(\bar{\boldsymbol{\theta}}, \alpha)\right] = \frac{\lceil \gamma(N_{\text{val}}+1) \rceil}{N_{\text{val}}}, \tag{51}$$

so we have

$$\gamma \leq \frac{1}{N_{\text{val}}} \sum_{k \in \mathcal{K}_{\text{val}}} h(z_k, \alpha) \leq \gamma + \frac{2}{N_{\text{val}}}. \tag{52}$$

The second inequality holds because

$$\frac{\lceil \gamma(N_{\text{val}}+1) \rceil}{N_{\text{val}}} = \frac{\lfloor \gamma N_{\text{val}} \rfloor + \lceil \gamma N_{\text{val}} - \lfloor \gamma N_{\text{val}} \rfloor + \gamma \rceil}{N_{\text{val}}} \leq \frac{\gamma N_{\text{val}} + \lceil \gamma N_{\text{val}} - \lfloor \gamma N_{\text{val}} \rfloor + \gamma \rceil}{N_{\text{val}}} \leq \gamma + \frac{2}{N_{\text{val}}}. \tag{53}$$

Since $\mathcal{D}_{\text{val}}$ is i.i.d. sampled, for any fixed α, $\sum_{k \in \mathcal{K}_{\text{val}}} h(z_k, \alpha)/N_{\text{val}}$ provides a sample average approximation to $\mathbb{E}\left[h(z, \alpha)\right]$, which can be interpreted as $\mathbb{P}\left(\boldsymbol{\Theta}^{\text{OPT}}(\hat{\mathbf{x}}, \mathbf{u}) \cap \mathcal{C}(\bar{\boldsymbol{\theta}}, \alpha)\right)$ for any new sample $(\hat{\boldsymbol{\theta}}, \mathbf{u})$ from $\mathbb{P}_{\hat{\boldsymbol{\theta}}, \mathbf{u}}$ and $\hat{\mathbf{x}} = \tilde{\mathbf{x}}(\hat{\boldsymbol{\theta}}, \mathbf{u})$.

By applying Lemma B.5, we have, with probability at least $\delta = 1 - 1/N_{\text{val}}$,

$$\left| \frac{1}{N_{\text{val}}} \sum_{k \in \mathcal{K}_{\text{val}}} h(z_k, \alpha) - \mathbb{E}\left[h(z, \alpha)\right] \right| \leq 2\mathfrak{R}_{N_{\text{val}}}(\mathcal{H}) + \sqrt{\frac{2\log N_{\text{val}}}{N_{\text{val}}}}. \tag{54}$$

By combing (50)–(54), we have, with probability at least $1 - 1/N_{\text{val}}$,

$$\left| \mathbb{P}\left(\boldsymbol{\Theta}^{\text{OPT}}(\hat{\mathbf{x}}, \mathbf{u}) \cap \mathcal{C}(\bar{\boldsymbol{\theta}}, \alpha)\right) - \gamma \right| \leq \sqrt{\frac{8\log(N_{\text{val}}+1) + 2\log N_{\text{val}}}{N_{\text{val}}}} + \frac{2}{N_{\text{val}}}. \tag{55}$$

□

B.5. Proof of Theorem 4.11

Proof. We first bound the perceived optimality gap of a sampled DM. Let $(\hat{\boldsymbol{\theta}}, \mathbf{u})$ be a sample from $\mathbb{P}_{(\boldsymbol{\theta},\mathbf{u})}$, $\hat{\mathbf{x}} = \tilde{\mathbf{x}}(\hat{\boldsymbol{\theta}}, \mathbf{u})$, $\hat{\boldsymbol{\theta}}_{\text{CIO}}(\mathbf{u})$ denote the optimal solution to the inner maximization problem in $\textbf{RFO}\left(\mathcal{C}(\bar{\boldsymbol{\theta}}, \alpha_1), \mathbf{u}\right)$ when the outer decision variable is set to $\hat{\mathbf{x}}$, $\bar{\boldsymbol{\theta}}_{\text{CIO}}(\mathbf{u})$ denote the optimal solution to the inner maximization problem in $\textbf{RFO}\left(\mathcal{C}(\bar{\boldsymbol{\theta}}, \alpha_1), \mathbf{u}\right)$ when the outer decision variable is set to $\bar{\mathbf{x}}_{\text{CIO}}(\mathbf{u})$, If $\boldsymbol{\Theta}^{\text{OPT}}(\hat{\mathbf{x}}, \mathbf{u}) \cap \mathcal{C}(\bar{\boldsymbol{\theta}}, \alpha_1) \neq \varnothing$, let $\tilde{\boldsymbol{\theta}}$ be an element of $\boldsymbol{\Theta}^{\text{OPT}}(\hat{\mathbf{x}}, \mathbf{u}) \cap \mathcal{C}(\bar{\boldsymbol{\theta}}, \alpha_1)$, we have

$$
\begin{aligned}
f\left(\hat{\boldsymbol{\theta}}, \bar{\mathbf{x}}_{\text{CIO}}(\mathbf{u})\right) - f\left(\hat{\boldsymbol{\theta}}, \hat{\mathbf{x}}\right) &\leq f\left(\tilde{\boldsymbol{\theta}}, \bar{\mathbf{x}}_{\text{CIO}}(\mathbf{u})\right) - f\left(\tilde{\boldsymbol{\theta}}, \hat{\mathbf{x}}\right) + \left[\nu(\hat{\mathbf{x}}) + \nu\left(\bar{\mathbf{x}}_{\text{CIO}}(\mathbf{u})\right)\right] \left\|\hat{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}}\right\|_2 \\
&\leq f\left(\tilde{\boldsymbol{\theta}}, \bar{\mathbf{x}}_{\text{CIO}}(\mathbf{u})\right) - f\left(\tilde{\boldsymbol{\theta}}, \hat{\mathbf{x}}\right) + \eta\left[\nu(\hat{\mathbf{x}}) + \nu\left(\bar{\mathbf{x}}_{\text{CIO}}(\mathbf{u})\right)\right] \\
&\leq f\left(\bar{\boldsymbol{\theta}}_{\text{CIO}}(\mathbf{u}), \bar{\mathbf{x}}_{\text{CIO}}(\mathbf{u})\right) - f\left(\tilde{\boldsymbol{\theta}}, \hat{\mathbf{x}}\right) + \eta\left[\nu(\hat{\mathbf{x}}) + \nu\left(\bar{\mathbf{x}}_{\text{CIO}}(\mathbf{u})\right)\right] \\
&\leq f\left(\hat{\boldsymbol{\theta}}_{\text{CIO}}(\mathbf{u}), \hat{\mathbf{x}}\right) - f\left(\tilde{\boldsymbol{\theta}}, \hat{\mathbf{x}}\right) + \eta\left[\nu(\hat{\mathbf{x}}) + \nu\left(\bar{\mathbf{x}}_{\text{CIO}}(\mathbf{u})\right)\right] \\
&\leq \nu(\hat{\mathbf{x}}) \left\|\hat{\boldsymbol{\theta}}_{\text{CIO}}(\mathbf{u}) - \tilde{\boldsymbol{\theta}}\right\|_2 + \eta\left[\nu(\hat{\mathbf{x}}) + \nu\left(\bar{\mathbf{x}}_{\text{CIO}}(\mathbf{u})\right)\right] \\
&\leq 2\nu(\hat{\mathbf{x}})(1 - \cos 2\alpha_1) + \eta\left(\nu(\hat{\mathbf{x}}) + \nu\left[\bar{\mathbf{x}}_{\text{CIO}}(\mathbf{u})\right]\right) \\
&= \nu(\hat{\mathbf{x}})(\eta - 2\cos 2\alpha_1 + 2) + \eta\nu\left[\bar{\mathbf{x}}_{\text{CIO}}(\mathbf{u})\right].
\end{aligned}
$$

The first line holds due to Assumption 4.7. The second line holds due to assumption 4.8. The third line holds due to the definition of $\bar{\boldsymbol{\theta}}_{\text{CIO}}(u)$. The fourth line holds because $\left(\bar{\mathbf{x}}_{\text{CIO}}(\mathbf{u}), \bar{\boldsymbol{\theta}}_{\text{CIO}}(\mathbf{u})\right)$ is an optimal solution to $\textbf{RFO}\left(\mathcal{C}(\bar{\boldsymbol{\theta}}, \alpha_1), \mathbf{u}\right)$. The fifth line holds due to Assumption 4.7. The sixth line holds because both $\hat{\boldsymbol{\theta}}_{\text{CIO}}(\mathbf{u})$ and $\tilde{\boldsymbol{\theta}}$ are in $\mathcal{C}(\bar{\boldsymbol{\theta}}, \alpha_1)$ so the angle between them is no larger than $2\alpha_1$. Since both $\hat{\boldsymbol{\theta}}_{\text{CIO}}(\mathbf{u})$ and $\tilde{\boldsymbol{\theta}}$ are on the unit sphere, the L_2 distance between them are bounded by $2(1 - \cos 2\alpha_1)$.

Since α_1 is chosen such that $\mathbb{P}\left(\boldsymbol{\Theta}^{\text{OPT}}(\hat{\mathbf{x}}, \mathbf{u}) \cap \mathcal{C}(\bar{\boldsymbol{\theta}}, \alpha_1)\right) = 1$, we have

$$
\begin{aligned}
\text{POG}(\bar{\mathbf{x}}_{\text{CIO}}) &= \mathbb{E}\left[f\left(\hat{\boldsymbol{\theta}}, \bar{\mathbf{x}}_{\text{CIO}}(\mathbf{u})\right) - f\left(\hat{\boldsymbol{\theta}}, \hat{\mathbf{x}}\right)\right] \\
&\leq \mathbb{E}\left\{\nu(\hat{\mathbf{x}})(\eta - 2\cos 2\alpha_1 + 2) + \eta\nu\left[\bar{\mathbf{x}}_{\text{CIO}}(\mathbf{u})\right]\right\} \\
&= \mu(\eta - 2\cos 2\alpha_1 + 2) + \eta\mu_{\text{CIO}}
\end{aligned}
$$

where $\mu := \mathbb{E}\left[\nu(\hat{\mathbf{x}})\right]$ and $\mu_{\text{CIO}} := \mathbb{E}\left(\nu\left[\bar{\mathbf{x}}_{\text{CIO}}(\mathbf{u})\right]\right)$.

□

B.6. Proof of Corollary 4.12

Proof. We first derive an upper bound on the optimality gap of the suggested decision $\bar{\mathbf{x}}_{\text{CIO}}(\mathbf{u})$ as evaluated using $\boldsymbol{\theta}^*$ for any $\mathbf{u} \in \mathcal{U}$. Let $(\hat{\boldsymbol{\theta}}, \mathbf{u})$ be a sample from $\mathbb{P}_{(\boldsymbol{\theta},\mathbf{u})}$, $\hat{\mathbf{x}} = \tilde{\mathbf{x}}(\hat{\boldsymbol{\theta}}, \mathbf{u})$, and $\tilde{\boldsymbol{\theta}}$ be an element of $\boldsymbol{\Theta}^{\text{OPT}}(\hat{\mathbf{x}}, \mathbf{u}) \cap \mathcal{C}(\bar{\boldsymbol{\theta}}, \alpha_1)$, which is non-empty almost surely because α_1 is chosen such that $\mathbb{P}\left(\boldsymbol{\Theta}^{\text{OPT}}(\hat{\mathbf{x}}, \mathbf{u}) \cap \mathcal{C}(\bar{\boldsymbol{\theta}}, \alpha_1)\right) = 1$. Let $\bar{\boldsymbol{\theta}}_{\text{CIO}}(\mathbf{u})$ denote the optimal solution to the inner maximization problem in $\textbf{RFO}\left(\mathcal{C}(\bar{\boldsymbol{\theta}}, \alpha_1), \mathbf{u}\right)$ when the outer decision variable is set to $\bar{\mathbf{x}}_{\text{CIO}}(\mathbf{u})$. For any $\mathbf{u} \in \mathcal{U}$, let $\mathbf{x}^*(\mathbf{u}) := \tilde{\mathbf{x}}(\boldsymbol{\theta}^*, \mathbf{u})$ and $\boldsymbol{\theta}^*_{\text{CIO}}(\mathbf{u})$ denote the optimal solution to the inner maximization problem in

$\mathbf{RFO}\left(\mathcal{C}(\bar{\boldsymbol{\theta}}, \alpha_1), \mathbf{u}\right)$ when the outer decision variable is set to $\mathbf{x}^*(\mathbf{u})$, we have

$$
\begin{aligned}
f\left(\boldsymbol{\theta}^*, \bar{\mathbf{x}}_{\text{CIO}}(\mathbf{u})\right) - f\left(\boldsymbol{\theta}^*, \mathbf{x}^*(\mathbf{u})\right) &\le f\left(\mathbb{E}(\hat{\boldsymbol{\theta}}), \bar{\mathbf{x}}_{\text{CIO}}(\mathbf{u})\right) - f\left(\mathbb{E}(\hat{\boldsymbol{\theta}}), \mathbf{x}^*(\mathbf{u})\right) + \left(\nu\left[\bar{\mathbf{x}}_{\text{CIO}}(\mathbf{u})\right] + \nu\left[\mathbf{x}^*(\mathbf{u})\right]\right)\|\boldsymbol{\theta}^* - \mathbb{E}(\hat{\boldsymbol{\theta}})\|_2 \\
&\le f\left(\mathbb{E}(\hat{\boldsymbol{\theta}}), \bar{\mathbf{x}}_{\text{CIO}}(\mathbf{u})\right) - f\left(\mathbb{E}(\hat{\boldsymbol{\theta}}), \mathbf{x}^*(\mathbf{u})\right) + \sigma\left(\nu\left[\bar{\mathbf{x}}_{\text{CIO}}(\mathbf{u})\right] + \nu\left[\mathbf{x}^*(\mathbf{u})\right]\right) \\
&= \mathbb{E}\left[f\left(\hat{\boldsymbol{\theta}}, \bar{\mathbf{x}}_{\text{CIO}}(\mathbf{u})\right) - f\left(\hat{\boldsymbol{\theta}}, \mathbf{x}^*(\mathbf{u})\right)\right] + \sigma\left(\nu\left[\bar{\mathbf{x}}_{\text{CIO}}(\mathbf{u})\right] + \nu\left(\mathbf{x}^*(\mathbf{u})\right)\right) \\
&\le \mathbb{E}\Big[f\left(\tilde{\boldsymbol{\theta}}, \bar{\mathbf{x}}_{\text{CIO}}(\mathbf{u})\right) - f\left(\tilde{\boldsymbol{\theta}}, \mathbf{x}^*(\mathbf{u})\right) + \left(\nu\left[\bar{\mathbf{x}}_{\text{CIO}}(\mathbf{u})\right] + \nu\left[\hat{\mathbf{x}}\right]\right)\|\hat{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}}\|_2\Big] \\
&\quad + \sigma\left(\nu\left[\bar{\mathbf{x}}_{\text{CIO}}(\mathbf{u})\right] + \nu\left(\mathbf{x}^*(\mathbf{u})\right)\right) \\
&\le \mathbb{E}\Big[f\left(\tilde{\boldsymbol{\theta}}, \bar{\mathbf{x}}_{\text{CIO}}(\mathbf{u})\right) - f\left(\tilde{\boldsymbol{\theta}}, \mathbf{x}^*(\mathbf{u})\right) + \left(\nu\left[\bar{\mathbf{x}}_{\text{CIO}}(\mathbf{u})\right] + \nu\left[\hat{\mathbf{x}}\right]\right)\eta\Big] \\
&\quad + \sigma\left(\nu\left[\bar{\mathbf{x}}_{\text{CIO}}(\mathbf{u})\right] + \nu\left(\mathbf{x}^*(\mathbf{u})\right)\right) \\
&\le \mathbb{E}\left[f\left(\bar{\boldsymbol{\theta}}_{\text{CIO}}(\mathbf{u}), \bar{\mathbf{x}}_{\text{CIO}}(\mathbf{u})\right) - f\left(\tilde{\boldsymbol{\theta}}, \mathbf{x}^*(\mathbf{u})\right)\right] + (\eta + \sigma)\left(\nu\left[\bar{\mathbf{x}}_{\text{CIO}}(\mathbf{u})\right] + \nu\left(\mathbf{x}^*(\mathbf{u})\right)\right) \\
&\le \mathbb{E}\left[f\left(\boldsymbol{\theta}^*_{\text{CIO}}(\mathbf{u}), \mathbf{x}^*(\mathbf{u})\right) - f\left(\tilde{\boldsymbol{\theta}}, \mathbf{x}^*(\mathbf{u})\right)\right] + (\eta + \sigma)\left(\nu\left[\bar{\mathbf{x}}_{\text{CIO}}(\mathbf{u})\right] + \nu\left(\mathbf{x}^*(\mathbf{u})\right)\right) \\
&\le \mathbb{E}\left[\nu(\mathbf{x}^*(\mathbf{u}))\|\boldsymbol{\theta}^*_{\text{CIO}}(\mathbf{u}) - \tilde{\boldsymbol{\theta}}\|_2\right] + (\eta + \sigma)\left(\nu\left[\bar{\mathbf{x}}_{\text{CIO}}(\mathbf{u})\right] + \nu\left(\mathbf{x}^*(\mathbf{u})\right)\right) \\
&\le 2\nu\left(\mathbf{x}^*(\mathbf{u})\right)(1 - \cos 2\alpha_1) + (\eta + \sigma)\left(\nu\left[\bar{\mathbf{x}}_{\text{CIO}}(\mathbf{u})\right] + \nu\left(\mathbf{x}^*(\mathbf{u})\right)\right) \\
&\le (2 - 2\cos 2\alpha_1 + \eta + \sigma)\nu\left(\mathbf{x}^*(\mathbf{u})\right) + (\eta + \sigma)\nu\left[\bar{\mathbf{x}}_{\text{CIO}}(\mathbf{u})\right]
\end{aligned}
$$

The first line holds because of Assumptions 4.7. The second line holds due to Assumption 4.10. The third line holds because f is linear in $\boldsymbol{\theta}$. The expectation is taken over $\mathbb{P}_{\boldsymbol{\theta}}$. The fourth line holds due to Assumption 4.7. The fifth line holds due to Assumption 4.8. The sixth line holds because of the definition of $\bar{\boldsymbol{\theta}}_{\text{CIO}}(\mathbf{u})$. The seventh line holds because $\left(\bar{\mathbf{x}}_{\text{CIO}}(\mathbf{u}), \bar{\boldsymbol{\theta}}_{\text{CIO}}(\mathbf{u})\right)$ is an optimal solution to $\mathbf{RFO}\left(\mathcal{C}(\bar{\boldsymbol{\theta}}, \alpha_1), \mathbf{u}\right)$. The eigth line holds due to Assumption 4.7. The ninth line holds since both $\boldsymbol{\theta}^*_{\text{CIO}}(\mathbf{u})$ and $\tilde{\boldsymbol{\theta}}$ are on the unit sphere and the angle between them is no greater than $2\alpha_1$, then the L_2 distance between them is upper bounded by $2(1 - \cos 2\alpha_1)$.

Next, we bound the AOG of $\bar{\mathbf{x}}_{\text{CIO}}$. We have

$$
\begin{aligned}
\text{AOG}(\bar{\mathbf{x}}_{\text{CIO}}) &= \mathbb{E}\left[f\left(\boldsymbol{\theta}^*, \bar{\mathbf{x}}_{\text{CIO}}(\mathbf{u})\right) - f\left(\boldsymbol{\theta}^*, \mathbf{x}^*(\mathbf{u})\right)\right] \\
&\le \mathbb{E}\left[(2 - 2\cos 2\alpha_1 + \eta + \sigma)\nu\left(\mathbf{x}^*(\mathbf{u})\right) + (\eta + \sigma)\nu\left[\bar{\mathbf{x}}_{\text{CIO}}(\mathbf{u})\right]\right] \\
&= (2 - 2\cos 2\alpha_1 + \eta + \sigma)\mu^* + (\eta + \sigma)\mu_{\text{CIO}}
\end{aligned}
$$

where $\mu^* := \mathbb{E}\left(\nu[\mathbf{x}^*(\mathbf{u})]\right)$. □

C. Numerical Experiment Details

C.1. Computational Setup

All the algorithms are implemented and test using Python 3.9.1 on a MacBook Pro with an Apple M1 Pro processor and 16 GB of RAM. Optimization models are implemented with Gurobi 9.5.2.

C.2. Forward Problem

C.2.1. SHORTEST-PATH

Let $\mathcal{E}^+(i)$ and $\mathcal{E}^-(i)$ denote the sets of edges that enter and leave node $i \in \mathcal{N}$, respectively. Let u^o and u^d denote the origin and destination of the trip, respectively. We define $x_{ij} \in \mathcal{E}$ as binary decision variables that take 1 if road (i, j) is traversed for any $(i, j) \in \mathcal{E}$. The shortest path problem is presented as follows.

$$\underset{\mathbf{x}}{\text{minimize}} \quad \sum_{(i,j)\in\mathcal{E}} \theta_{ij} x_{ij} \tag{56a}$$

$$\text{subject to} \quad \sum_{(j,i)\in\mathcal{E}^{+}(i)} x_{ji} - \sum_{(i,j)\in\mathcal{E}^{-}(i)} x_{ij} = \begin{cases} 1, & \text{if } i = u^d \\ -1, & \text{if } i = o^d \\ 0, & \text{otherwise} \end{cases}, \quad \forall i \in \mathcal{N} \tag{56b}$$

$$x_{ij} \in \{0, 1\}, \quad (i, j) \in \mathcal{E}. \tag{56c}$$

The objective function minimizes the total travel cost. The first set of constraints are the flow-balancing constraints that make sure we can find a path from u_o to u_d. The second set of constraints specify the range of our decision variables. Note that the constriant matrix is totally unimodular, so we can replace the binary constraints with $0 \leq x_{ij} \leq 1$ for any $(i, j) \in \mathcal{E}$ when implementing this model.

C.2.2. KNAPSACK

We define binary decision variables x_i that indicate if item $i \in [d]$ is selected $(= 1)$ or not $(= 0)$. The knapsack problem is presented as follows.

$$\underset{\mathbf{x}}{\text{maximize}} \quad \sum_{i\in[d]} \theta_i x_i \tag{57a}$$

$$\text{subject to} \quad \sum_{i\in[d]} w_i x_i \leq u \tag{57b}$$

$$x_i \in \{0, 1\}, \forall i \in [d]. \tag{57c}$$

The objective maximizes the total value of the selected items. The first constraint enforces a total budget for item selection. The second set of constraints specify the range of our decision variables.

C.3. Solving the Data-driven Inverse Optimization Problem

For all problem instances, we solve the following inverse problem to obtain a point estimation of parameters.

$$\underset{\boldsymbol{\theta}\in\mathbb{R}^{|\mathcal{E}|}, \boldsymbol{\epsilon}\in\mathbb{R}_{+}^{n_{\text{train}}}}{\text{minimize}} \quad \frac{1}{N_{\text{train}}} \sum_{k\in\mathcal{K}_{\text{train}}} l_k \tag{58a}$$

$$\text{subject to} \quad l_k \geq \boldsymbol{\theta}^{\intercal}\hat{\mathbf{x}}_k - \boldsymbol{\theta}^{\intercal}\mathbf{x}, \quad \forall \mathbf{x} \in \mathcal{X}_k, \ k \in \mathcal{K}_{\text{train}} \tag{58b}$$

$$\|\boldsymbol{\theta} - \mathbf{1}\|_1 \leq \frac{|\mathcal{E}|}{4}. \tag{58c}$$

This problem is initialized without Constraints (58b), which were added iteratively using a cutting-plane method. Specifically, in each iteration, after solving Problem (58), let $\boldsymbol{\theta}'$ and $\{l'_k\}_{k\in\mathcal{K}_{\text{train}}}$ be the optimal solution. For each data point $k \in \mathcal{K}_{\text{train}}$, we solve the following sub-problem

$$\underset{\mathbf{x}_k\in\mathcal{X}(\mathbf{u}_k)}{\text{minimize}} \quad \boldsymbol{\theta}'^{\intercal}\mathbf{x}_k. \tag{59}$$

Let $\mathbf{x}'_k$ be the optimal solution to the sub-problem. If $l'_k < \boldsymbol{\theta}'^{\intercal}\hat{\mathbf{x}}_k - \boldsymbol{\theta}'^{\intercal}\mathbf{x}'_k$, we add the following cut to Problem (58)

$$l_k \geq \boldsymbol{\theta}^{\intercal}\hat{\mathbf{x}}_k - \boldsymbol{\theta}^{\intercal}\mathbf{x}'_k. \tag{60}$$

We keep running this procedure until no cut is added to the master Problem (58).

C.4. Solving the Calibration Problem

C.4.1. Shortest Path

For each data point in the validation set, we calculate the value of c_k by solving the following problem

$$\begin{aligned}
\underset{\boldsymbol{\theta}\in\mathbb{R}^{|\mathcal{E}|},\mathbf{w}\in\mathbb{R}^{\mathcal{N}},\mathbf{v}\in\mathbb{R}_+^{|\mathcal{E}|}}{\text{maximize}} \quad & \bar{\boldsymbol{\theta}}^\intercal\boldsymbol{\theta} && \text{(61a)}\\
\text{subject to} \quad & w_{d_k} - w_{o_k} - \sum_{(i,j)\in\mathcal{E}} v_{ij} = \boldsymbol{\theta}^\intercal\hat{\mathbf{x}}_k && \text{(61b)}\\
& w_j - w_i - v_{ij} \le c_{ij}, \quad \forall(i,j)\in\mathcal{E} && \text{(61c)}\\
& \|\boldsymbol{\theta}\|_2 \le 1. && \text{(61d)}
\end{aligned}$$

where $\mathbf{w}\in\mathbb{R}^{\mathcal{N}}$ and $\mathbf{v}\in\mathbb{R}_+^{|\mathcal{E}|}$, respectively, denote the dual variables associated with the flow-balancing constraints and the capacity constraints in the primal problem. The first constraint enforces strong duality. The second set of constraints are the dual feasibility constraints. The last constraint ensures the optimal solution is on the unit sphere. Note that we do not need to enforce $\|\boldsymbol{\theta}\|_2 = 1$ because this is a maximization problem.

C.4.2. Knapsack

For each data point in the validation set, we calculate the value of c_k by solving the following calibration problem

$$\begin{aligned}
\underset{\boldsymbol{\theta}\in\mathbb{R}^{d}}{\text{maximize}} \quad & \bar{\boldsymbol{\theta}}^\intercal\boldsymbol{\theta} && \text{(62a)}\\
\text{subject to} \quad & \boldsymbol{\theta}^\intercal\hat{\mathbf{x}}_k \ge \boldsymbol{\theta}^\intercal\mathbf{x}, \quad \forall\mathbf{x}\in\mathcal{X}(\mathbf{u}_k) && \text{(62b)}\\
& \|\boldsymbol{\theta}\|_2 \le 1. && \text{(62c)}
\end{aligned}$$

We initialize this problem without Constraints (62b). In each iteration, after solving the calibration problem, let $\boldsymbol{\theta}'$ denote the optimal solution. We solve $\mathbf{FO}(\boldsymbol{\theta}', \mathbf{u}_k)$ and let $\mathbf{x}'$ denote the optimal solution. If $\boldsymbol{\theta}'^\intercal\mathbf{x}' > \boldsymbol{\theta}'^\intercal\hat{\mathbf{x}}_k$, we then add the corresponding cut to the model. We keep running this process until no cut is added.

C.5. Solving the Robust Forward Problem

Let $\alpha = \cos^{-1}(\Gamma_k(\{c_k\}_{k\in\mathcal{K}_{\text{val}}}))$. We next solve the following robust model to recommend a new decision to prescribe a decision given a $u\in\mathcal{U}$.

$$\begin{aligned}
\underset{\mathbf{x}\in\mathcal{X}(\mathbf{u})}{\text{minimize}}\ \underset{\boldsymbol{\theta}\in\mathbb{R}^{|\mathcal{E}|}}{\text{maximize}} \quad & \boldsymbol{\theta}^\intercal\mathbf{x} && \text{(63a)}\\
\text{subject to} \quad & \bar{\boldsymbol{\theta}}^\intercal\boldsymbol{\theta} \ge \cos(\alpha) && \text{(63b)}\\
& \|\boldsymbol{\theta}\|_2 \le 1. && \text{(63c)}
\end{aligned}$$

We initialize this problem as follows.

$$\begin{aligned}
\underset{\mathbf{x}\in\mathcal{X}(\mathbf{u}),\Omega\in\mathbb{R}_+}{\text{minimize}} \quad & \Omega && \text{(64a)}\\
\text{subject to} \quad & \boldsymbol{\theta}^\intercal\mathbf{x} \le \Omega, \quad \forall\boldsymbol{\theta}\in\tilde{\boldsymbol{\Theta}}. && \text{(64b)}
\end{aligned}$$

We initialize $\tilde{\boldsymbol{\Theta}} = \varnothing$. We first solve Problem (64), let $\mathbf{x}'$ and Ω' denote the optimal solution. Then we solve the following sub-problem

$$\begin{aligned}
\underset{\boldsymbol{\theta}\in\mathbb{R}^{|\mathcal{E}|}}{\text{maximize}} \quad & \boldsymbol{\theta}^\intercal\mathbf{x}' && \text{(65a)}\\
\text{subject to} \quad & \bar{\boldsymbol{\theta}}^\intercal\boldsymbol{\theta} \ge \cos(\alpha) && \text{(65b)}\\
& \|\boldsymbol{\theta}\|_2 \le 1. && \text{(65c)}
\end{aligned}$$

Let $\boldsymbol{\theta}'$ denote the optimal solution to the sub-problem. If $\boldsymbol{\theta}^\intercal\mathbf{x}' > \Omega'$, then we add $\boldsymbol{\theta}'$ to $\tilde{\boldsymbol{\Theta}}$ and re-solve Problem (64). We keep running this procedure until no new solution is added to $\tilde{\boldsymbol{\Theta}}$.