Conformal Inverse Optimization for Adherence-aware Prescriptive Analytics

Timothy C. Y. Chan
Department of Mechanical & Industrial Engineering, University of Toronto, tcychan@mie.utoronto.ca

Erick Delage
Department of Decision Sciences, HEC Montréal, erick.delage@hec.ca,

Bo Lin
Department of Mechanical & Industrial Engineering, University of Toronto, blin@mie.utoronto.ca

Inverse optimization is increasingly used to estimate unknown parameters in an optimization model based on decision data. We show that such a point estimate alone is insufficient in a prescriptive setting where the estimated parameters are used to prescribe new decisions. The resulting decisions may be low-quality and misaligned with human intuition and thus are unlikely to be adopted. To tackle this challenge, we propose a novel decision recommendation pipeline, which seeks to learn an uncertainty set for the unknown parameters and then solve a robust optimization model to prescribe new decisions. We show that the suggested decisions can achieve bounded optimality gaps, as evaluated using both the ground-truth parameters and human perceptions. Our method demonstrates strong empirical performance compared to the standard inverse optimization pipeline. Finally, we perform a case study where we apply this new pipeline to provide delivery route recommendations in Toronto, Canada. Our approach achieves a significantly higher delivery path adherence rate than current industry practices without compromising service quality. Moreover, our method provides a better trade-off between absolute and perceived decision quality than baselines under various realistic scenarios, including cases with model mis-specification and data scarcity.

Key words: inverse optimization; robust optimization; prescriptive analytics; human-AI collaboration.

1. Introduction

Inverse optimization (IO) is often used to fit unknown parameters in an optimization model to decision data, so that the model with the fitted parameters can subsequently be used to prescribe prospective decisions. For this "estimate, then optimize" 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 property encourages algorithm adoption (Chen et al. 2023, Donahue et al. 2023), which is critical in many real-world applications, such as rideshare vehicle positioning (Liu et al. 2023), warehouse operations (Sun et al. 2022), and product assortment (Kawaguchi 2021). The essence of the issue is that an algorithm may present

an "optimal" decision, but a human may reject it and implement a decision that would be worse than if the algorithm presented a slightly suboptimal decision that is better aligned with human judgment (and thus more likely to be implemented).

For instance, Alibaba developed an algorithm to optimize the packing of items into boxes for delivery to its customers. However, they found that workers at their warehouse did not conform to algorithm-suggested packing plans for 5.8% of the packages and these packages experienced a 48.3% increase in processing time compared to the recommended plan (Sun et al. 2022). After adjusting algorithmic recommendations to incorporate possible human deviations, the deviation rate dropped significantly, equivalent to an estimated annual cost saving of 2.6 million US dollars. Similarly, when an automobile parts retailer utilized a data-driven tool to identify low-performing stock-keeping units for removal from local stores, recommendations were overridden over 50% of the time by local merchants, leading to an estimated 5.8% reduction in profitability (Kesavan and Kushwaha 2020). Deviations themselves might have other negative consequences beside degraded solution quality. For example, a rideshare driver's deviation from the recommended path may cause rider safety concerns (China Daily 2021, Global Times 2021) and affect other tasks that are critical to the platform, including arrival time estimation and trip pricing (Hu et al. 2022). Similar observations have been documented in last-mile delivery where delivery route modeling is an important input to order batching and assignment algorithms (Liu et al. 2021).

Despite the potential negative impact of human deviation from algorithmic solutions, completely eliminating the discretionary power of human decision makers may not be a realistic option for various reasons. In high-stake settings such as healthcare, algorithms are rarely used without human oversight due to safety concerns and regulatory requirements (Wilson and Daugherty 2018, Ge et al. 2023). In decentralized systems, local staff may be able to improve upon algorithmic recommendations because they possess information that is inaccessible to the centralized algorithms (Phillips et al. 2015, Kesavan and Kushwaha 2020). Therefore, to encourage algorithm adoption, one should focus on deriving not only high-quality decisions as evaluated using objective criteria, but decisions that are also perceived to be high-quality by a human decision maker. The latter metric is usually unobservable but may be inferred from historical human decisions.

Inverse optimization is commonly used to perform such estimation, assuming that the observed decisions are solutions to an optimization model in which some parameters are

subject to human perception. With few exceptions (Birge et al. 2022b, Yousefi 2023), IO provides a point estimate of these parameters to represent the aggregated perception of the decision makers who contributed to the dataset. While such IO methods have been successfully applied in many settings (Chan et al. 2014, Rönnqvist et al. 2017, Liu et al. 2022), we argue that such a point estimate may not always be effective. First, human perceptions may constitute a wide distribution for which a point estimate carries limited value due to a lack of consensus among the contributing decision makers. Second, the specified optimization model may not perfectly represent the human decision making process, potentially limiting the generalization power of such point estimates. Finally, even if the model is well-specified, many IO approaches do not provide a statistically consistent estimate and can be sensitive to noise in the input data (Shahmoradi and Lee 2022). While consistent IO estimators exist for strictly convex problems (Aswani et al. 2019), solving these IO models is computationally challenging. Furthermore, neither the solution method nor the consistency guarantees apply to discrete optimization problems, which are ubiquitous in practice. Therefore, in many real-world applications, using computationally tractable IO estimators that do not possess consistency guarantees is the only realistic option. Consideration of uncertainty in this estimate is necessary.

In this paper, we develop a new IO approach to learn from decision data and return not only a point estimate, but also an uncertainty set around that estimate. This uncertainty set is calibrated so that it contains a decision maker's perceived value of the target parameters with high probability. This set is then used in a robust optimization model to form a decision pipeline whose goal is to improve the alignment between data-driven decision recommendations and human perception. The calibration of the uncertainty set is inspired by the concept of conformal prediction (Shafer and Vovk 2008) from machine learning (ML), a distribution-free method to generate a set that contains the true prediction target with high probability. The robust optimization model can be interpreted as a process of finding a decision that is perceived to be of high-quality by most decision makers (as captured by the uncertainty set). Under mild technical assumptions, we show that decisions prescribed using this new IO pipeline enjoy theoretical guarantees on both absolute solution quality and perceived solution quality. Finally, we develop several algorithms to accelerate both the uncertainty set calibration and the solution of the robust optimization model, leading to a computationally efficient pipeline. Our contributions are summarized as follows.

1. *A new framework.* We propose a decision making pipeline to generate decisions that are of high quality and aligned with human intuition. Our pipeline integrates i) "conformal IO", a novel approach to constructing uncertainty sets based on decision data and ii) a robust optimization model for decision recommendation. Item i) may be of independent interest to the data-driven robust optimization community as it presents a new way to construct uncertainty sets without observations of the unknown parameters. While the resulting robust optimization model is non-convex (due to a norm constraint), we introduce a data-driven method to approximate it, with provable solution quality guarantees. This method can be generalized to solve other norm-constrained robust optimization problems.

2. *Theoretical guarantees.* We prove that the probability of the learned uncertainty set containing parameters that make future observed decisions optimal always exceeds a specified threshold (conservatively valid), and this probability converges to the threshold as the sample size goes to infinity (asymptotic exact). This coverage guarantee leads to provable bounds on the optimality gap of the decisions from conformal IO, as evaluated using both the ground-truth parameters and the decision maker's perceived parameters.

3. *Performance.* Through extensive numerical experiments, we i) empirically verify the effectiveness of conformal IO in constructing an uncertainty set that achieves the out-of-sample coverage desired by the model user, ii) demonstrate better performance of our conformal IO pipeline compared to the standard IO pipeline in terms of ground-truth and perceived solution quality, and iii) showcase that our data-driven approximation to the non-convex robust optimization model significantly accelerates the computational efficiency of our pipeline while maintaining high decision quality.

4. *Real-world case study.* We apply the proposed approach to provide route recommendations for food delivery couriers in Toronto, Canada, leading to a rich set of managerial insights. Compared to the current industry practice that recommends the shortest delivery route, our approach improves the path adherence rate by 8.75–62.5 percentage points while incurring comparable delivery time. When couriers have a low tolerance for sub-optimal (with respect to their own perceptions) path recommendations, we can achieve shorter average delivery time than recommending the shortest path (with respect to ground-truth parameters) because we prevent couriers from deviating and choosing a perceived optimal path that turns out to take more time in reality. This finding underscores the opportunity for a win-win: improved recommendation adherence and improved service quality. When we

benchmark against the standard IO pipeline, our pipeline offers a better trade-off between absolute and perceived decision quality. It also demonstrates more robust performance when the optimization model is mis-specified. In a third experiment, we compare our model against a personalization strategy that maintains one model for each courier, varying the number of data points observed for each courier. Our method incurs lower computational overhead and demonstrates stronger performance in data-poor regimes with comparable performance in data-rich regimes. This finding highlights the value of our approach when the platform enters a new market or is onboarding new couriers, for example.

2. Literature Review

Our paper is related to a broad range of literature including inverse optimization, "estimate, then optimize", data-driven robust optimization, and human-AI interaction. Below, we discuss how our work relates to and differs from each stream.

IO aims to impute unknown parameters in an optimization model based on decision data in both offline (Ahuja and Orlin 2001, Bertsimas et al. 2015, Chan and Kaw 2020, Birge et al. 2022a) and online settings (Bärmann et al. 2018, Dong et al. 2018, Besbes et al. 2023) where data are observed sequentially. Early IO papers focus on deterministic settings where the observed decisions are assumed to be optimal to the specified optimization model. Recent papers have focused on stochastic settings where the observed decisions are noisy. When the decision data are subject to execution errors, Aswani et al. (2018) propose an IO model that minimizes the predictability loss, leading to a statistically consistent estimator for strictly convex problems, i.e., the estimated parameters converge to the true parameters as the number of observations goes to infinity. When the decision data are subject to model mis-specification, measurement error, and bounded rationality, Mohajerin Esfahani et al. (2018) propose a distributionally robust suboptimality loss and derive generalization bounds for the IO estimate. Chan et al. (2019) study similar suboptimality losses and demonstrate their tractability given noisy decision data. We refer readers to Chan et al. (2023b) for a comprehensive review. In this paper, we focus on another source of noise — our decision data are generated using possibly biased human perception of the ground-truth parameters. Our IO pipeline utilizes the sub-optimality loss of Mohajerin Esfahani et al. (2018) and Chan et al. (2019) as we focus on developing a computationally

tractable pipeline that can be applied to both continuous and discrete optimization models. Moreover, we derive theoretical guarantees on absolute and perceived solution quality for our prescribed decisions, which has not been studied in the IO literature.

The above IO methods all return a point estimate of the target parameters, while our goal is to learn an uncertainty set. Recently, Birge et al. (2022b) and Yousefi (2023) present IO approaches for estimating a parameter distribution, which can also be used to construct an uncertainty set. The former approach relies on distributional assumptions regarding the unknown parameters, while the latter relies on an identifiability condition being satisfied. Both papers rely on posterior sampling algorithms to solve the IO problem, which is computationally demanding and cannot deal with more than a few parameters. In this paper, do not impose any distributional assumptions regarding the unknown parameters (which are usually high-dimensional), so the method by Birge et al. (2022b) is not applicable. Moreover, we focus on a class of problems whose objective function is linear in the unknown parameters, which violates the identifiability condition used by Yousefi (2023).

Our approach belongs to the family of "estimate, then optimize" methods. Recent studies suggest that even small estimation errors may be amplified in the optimization step, leading to 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 and Grigas 2022) and robustifying the optimization model (Sun et al. 2023, Chan et al. 2023a). We are similar to the second stream, but differ by i) using 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.

Uncertainty set construction for robust optimization has been extensively studied. Central to this problem are the tractability of the resulting robust model and the price of robustness. Early papers use prior knowledge about the parameter uncertainty to design sets that are polyhedral (Ben-Tal and Nemirovski 1999), ellipsoidal (Ben-Tal and Nemirovski 2000), cardinality constrained (Bertsimas and Sim 2004), and norm constrained (Bertsimas et al. 2004). Recently, data have become a critical ingredient in defining new uncertainty sets for distributionally robust optimization (Delage and Ye 2010, Mohajerin Esfahani et al. 2018, Gao and Kleywegt 2023) and calibrating the size of the uncertainty set (Bertsimas et al. 2018, Chenreddy et al. 2022). Closest to our work is Sun et al. (2023) who first use a

machine learning model to predict the unknown parameters and then calibrate an uncertainty set around the prediction. However, this approach does not apply in our setting as it requires observations of the unknown parameters, which we assume we do not have access to. Our paper presents the first approach to calibrating uncertainty sets using decision data alone, which in many applications is more readily observable than parameters.

Our paper contributes to the literature on human-AI interaction. AI techniques have demonstrated superior performance on various tasks, but humans are often reluctant to adopt these techniques—a phenomenon called "algorithm aversion" (Burton et al. 2020). Empirical studies have identified several reasons, including lack of algorithm transparency (Kizilcec 2016), poor performance (Yin et al. 2019), and the lack of human control (Dietvorst et al. 2018) or inputs (Kawaguchi 2021). In response, researchers have proposed algorithms that are interpretable (Ciocan and Mišić 2022, Bastani et al. 2021) and adherence-aware (Grand-Clément and Pauphilet 2024), leading to improved human-AI collaboration in many settings. Our paper shares the same goal, but focuses on improving the alignment between algorithmic recommendation and human perception, which has been identified as another major factor that affects algorithm adoption (Chen et al. 2023, Liu et al. 2023). Closest to our paper is Fu et al. (2023), who combine driver routing preferences learned using inverse reinforcement learning with the "theoretical shortest path" to prescribe last-mile delivery routes. In contrast, our method is not specific to last-mile delivery. Additionally, we explicitly characterize the concepts of actual and perceived optimality gaps and derive theoretical guarantees for our model's performance.

3. Preliminaries

In this section, we first present the problem setup (Section 3.1) and then describe the challenges with the standard IO pipeline (Section 3.2). Finally, we provide intuition on why robustifying the optimization model would help (Section 3.4). We use the following notational conventions. Vectors and matrices are denoted in bold and sets are in calligraphic font. We let $\mathbb{1}$ denote the indicator function and $[n] = \{1, 2, \ldots, n\}$ for integer n.

3.1. Problem Setup

Consider a *forward optimization problem*

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

where $\mathbf{x} \in \mathbb{R}^n$ is the decision vector whose feasible region $\mathcal{X}(\mathbf{u})$ is non-empty, compact and is parameterized by exogenous parameters $\mathbf{u} \in \mathbb{R}^m$, $\boldsymbol{\theta} \in \mathbb{R}^d$ is a non-zero parameter vector, and $f : \mathbb{R}^{n \times d} \to \mathbb{R}$ is the objective function. Suppose $\mathbf{u}$ is distributed according to $\mathbb{P}_{\mathbf{u}}$, which is supported on a closed and bounded set $\mathcal{U}$. There exists a ground-truth parameter vector $\boldsymbol{\theta}^*$ that is *unknown* to the decision maker. Instead, the decision maker obtains a decision $\hat{\mathbf{x}}$ by solving $\mathbf{FOP}\left(\hat{\boldsymbol{\theta}}, \mathbf{u}\right)$ 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 $\tilde{\mathbf{x}} : \boldsymbol{\Theta} \times \mathcal{U} \to \mathbb{R}^n$ be an oracle that returns an optimal solution to **FOP**, i.e., $\tilde{\mathbf{x}}(\boldsymbol{\theta}, \mathbf{u}) \in \mathcal{X}^{\mathrm{OPT}}(\boldsymbol{\theta}, \mathbf{u}) := \arg\min \{f(\boldsymbol{\theta}, \mathbf{x}) \,|\, \mathbf{x} \in \mathcal{X}(\mathbf{u})\}$.

In this paper, we focus on the case where f is linear in $\boldsymbol{\theta}$. That is the objective function of **FOP** can be written as

$$f(\boldsymbol{\theta}, \mathbf{x}) = \sum_{i \in [d]} \theta_i f_i(\mathbf{x}), \tag{2}$$

where $f_i : \mathbb{R}^n \to \mathbb{R}$ for all $i \in [d]$ are some continuous convex basis functions. This function generalizes the linear objective $f(\boldsymbol{\theta}, \mathbf{x}) = \boldsymbol{\theta}^\intercal \mathbf{x}$. Moreover, **FOP** with this objective can be interpreted as a multi-objective optimization model, whose inverse model has been applied to model routing preferences (Rönnqvist et al. 2017), radiation therapy treatment planning (Chan et al. 2014), and portfolio optimization (Dong and Zeng 2021). In this setting, the optimal solution to **FOP** is invariant to the scale of $\boldsymbol{\theta}$, i.e., if $\mathbf{x} \in \mathcal{X}^{\mathrm{OPT}}(\boldsymbol{\theta}, \mathbf{u})$, then $\mathbf{x} \in \mathcal{X}^{\mathrm{OPT}}(\beta\boldsymbol{\theta}, \mathbf{u})$ for any $\beta \in \mathbb{R}_+$. So, we set $\boldsymbol{\Theta} = \left\{\boldsymbol{\theta} \in \mathbb{R}^d \,\middle|\, \|\boldsymbol{\theta}\|_2 = 1\right\}$ without loss of generality.

3.1.1. Learning task. 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 designing a decision policy $\bar{\mathbf{x}} : \mathcal{U} \to \mathbb{R}^n$ that recommends a decision given a future $\mathbf{u}$. We require $\bar{\mathbf{x}}(\mathbf{u}) \in \mathcal{X}(\mathbf{u})$ for any $\mathbf{u} \in \mathcal{U}$.

We propose the following two metrics to evaluate $\bar{\mathbf{x}}$.

Definition 1. The *actual optimality gap* (AOG) of a decision policy $\bar{\mathbf{x}}$ is defined as

$$\mathrm{AOG}(\bar{\mathbf{x}}) := \mathbb{E}_{\mathbf{u}}\left[f\left(\boldsymbol{\theta}^*, \bar{\mathbf{x}}(\mathbf{u})\right) - \min_{\mathbf{x} \in \mathcal{X}(\mathbf{u})} f(\boldsymbol{\theta}^*, \mathbf{x})\right], \tag{3}$$

where the expectation is taken over the distribution of the random variable $\mathbf{u}$.

Definition 2. The *perceived optimality gap* (POG) of a decision policy $\bar{\mathbf{x}}$ is defined as

$$\mathrm{POG}(\bar{\mathbf{x}}) := \mathbb{E}_{\hat{\boldsymbol{\theta}},\mathbf{u}}\left[f\left(\hat{\boldsymbol{\theta}}, \bar{\mathbf{x}}(\mathbf{u})\right) - \min_{\mathbf{x} \in \mathcal{X}(\mathbf{u})} f\left(\hat{\boldsymbol{\theta}}, \mathbf{x}\right)\right], \tag{4}$$

where the expectation is taken with respect to $\hat{\boldsymbol{\theta}}$ and $\mathbf{u}$.

AOG is an *objective* performance metric that measures the absolute quality of the decision recommendation, i.e., with respect to $\boldsymbol{\theta}^*$. In contrast, POG is a *subjective* metric that measures the decision quality as perceived by human decision makers. In this paper, our goal is to design a decision policy $\bar{\mathbf{x}}$ that achieves low AOG and POG simultaneously, so the recommendations generated by $\bar{\mathbf{x}}$ are of high quality and likely to be implemented.

REMARK 1. Note that $\bar{\mathbf{x}}(\mathbf{u})$ is usually obtained by solving an optimization model, which may have multiple optimal solutions. In this case, we think of an optimal solution being drawn at random from the optimal solution set. As a result, we may think of $\bar{\mathbf{x}}$ being a "randomized" decision policy, and the expectations in Equations (3) and (4) would be taken with respect to the randomness in $\bar{\mathbf{x}}$ as well.

3.1.2. Assumptions. We introduce three mild technical assumptions regarding the data generation process and the forward optimization problem.

ASSUMPTION 1 **(I.I.D. Samples)**. *The dataset $\mathcal{D}$ is generated using $\hat{\mathbf{x}}_k := \tilde{\mathbf{x}}\left(\hat{\boldsymbol{\theta}}_k, \mathbf{u}_k\right)$ where $\left(\hat{\boldsymbol{\theta}}_k, \mathbf{u}_k\right)$ are i.i.d. samples from $\mathbb{P}_{(\boldsymbol{\theta},\mathbf{u})}$ for all $k \in [N]$.*

ASSUMPTION 2 **(Bounded Divergence)**. *There exists a constant $\sigma \in \mathbb{R}_+$ such that $\left\|\mathbb{E}\left(\hat{\boldsymbol{\theta}}\right) - \boldsymbol{\theta}^*\right\|_2 \leq \sigma$.*

ASSUMPTION 3 **(Bounded Inverse Feasible Set)**. *There exists a constant $\eta \in \mathbb{R}_+$ such that $\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_2 \leq \eta$ for any $\mathbf{u} \in \mathcal{U}$, $\hat{\mathbf{x}} \in \mathcal{X}(\mathbf{u})$, and $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \boldsymbol{\Theta}^{\mathrm{OPT}}(\mathbf{u}, \hat{\mathbf{x}})$, where*

$$\boldsymbol{\Theta}^{\mathrm{OPT}}(\mathbf{u}, \mathbf{x}) := \left\{\boldsymbol{\theta} \in \mathbb{R}^d \,\middle|\, \mathbf{x} \in \mathcal{X}^{\mathrm{OPT}}(\boldsymbol{\theta}, \mathbf{u}), \|\boldsymbol{\theta}\|_2 = 1\right\}. \tag{5}$$

Assumption 1 is standard in the data-driven optimization literature. Assumption 2 is mild, simply stating that the l_2 distance between the expected value of the perceived parameters and the ground-truth parameters is bounded. For example, if the support of $\boldsymbol{\theta}$ is bounded, then this assumption will hold. Assumption 3 is true by definition because $\boldsymbol{\Theta}^{\mathrm{OPT}}(\hat{\mathbf{x}}, \mathbf{u})$ is bounded due to the norm constraint. We state it simply to introduce η, which will show up in the bounds later.

3.2. Standard Inverse Optimization Pipeline

A natural approach to accomplishing the learning task set out in Section 3.1.1 is to use inverse optimization to first obtain a point estimate $\bar{\boldsymbol{\theta}}$ of the unknown parameters and

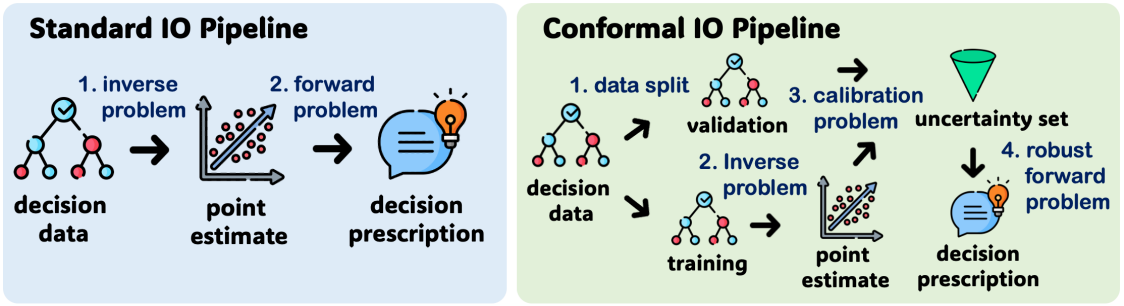


Figure 1 Standard inverse optimization and conformal inverse optimization pipelines.

then employ a policy $\bar{\mathbf{x}}_{\mathrm{IO}}(\mathbf{u}) := \tilde{\mathbf{x}}\left(\bar{\boldsymbol{\theta}}, \mathbf{u}\right)$ to prescribe decisions for any $\mathbf{u} \in \mathcal{U}$ (Rönnqvist et al. 2017, Babier et al. 2020), i.e., solving **FOP** with the estimated $\bar{\boldsymbol{\theta}}$ and $\mathbf{u}$ (see Figure 1). Specifically, given $\mathcal{D} := \{(\hat{\mathbf{x}}_k, \mathbf{u}_k) \,|\, k \in [N]\}$, we can estimate the parameters by solving the following *inverse optimization problem*

$$\mathbf{IOP}(\mathcal{D}) : \underset{\boldsymbol{\theta} \in \boldsymbol{\Theta}}{\text{minimize}} \; \frac{1}{N} \sum_{k \in [N]} \ell\left(\hat{\mathbf{x}}_k, \mathcal{X}^{\mathrm{OPT}}(\boldsymbol{\theta}, \mathbf{u}_k)\right), \tag{6}$$

where ℓ is a non-negative loss function that returns 0 only when $\hat{\mathbf{x}}_k \in \mathcal{X}^{\mathrm{OPT}}(\boldsymbol{\theta}, \mathbf{u}_k)$. For instance, a popular choice is the following sub-optimality loss, which penalizes the absolute optimality gap achieved by the observed decision under the estimated parameters.

Definition 3. The *sub-optimality loss* of $\boldsymbol{\theta}$ is given by

$$\ell_{\mathrm{S}}\left(\hat{\mathbf{x}}, \mathcal{X}^{\mathrm{OPT}}(\boldsymbol{\theta}, \mathbf{u})\right) = f(\boldsymbol{\theta}, \hat{\mathbf{x}}) - \min_{\mathbf{x} \in \mathcal{X}^{\mathrm{OPT}}(\boldsymbol{\theta}, \mathbf{u})} f(\boldsymbol{\theta}, \mathbf{x}). \tag{7}$$

While this loss function does not produce statistically consistent estimates of the parameters (Aswani et al. 2018), it is the most commonly used because it is tractable. In fact, when **FOP** is non-convex or the unknown parameters are high-dimensional, the sub-optimality loss is usually the only loss function that leads to a tractable **IOP**. As noted by Mohajerin Esfahani et al. (2018), this situation is similar to binary classification where it is preferable to minimize the convex cross-entropy loss instead of the 0-1 loss, even if the latter is the actual metric of interest. However, as the next example shows, this loss function can lead to decision policies with arbitrarily large AOG and POG.

3.3. Limitations of the Standard Inverse Optimization Pipeline: An Example

To build intuition, we examine the performance of the standard IO pipeline with the sub-optimality loss (7) via an illustrative example (visualized in Figure 2).

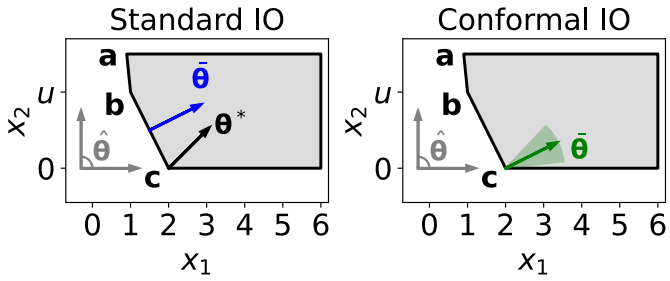


Figure 2 Illustration of standard and conformal IO pipelines. The gray areas are the feasible region $\mathcal{X}(u)$. The black arrows correspond to the ground-truth parameter $\boldsymbol{\theta}^*$. The gray arrows are the extreme rays of $\boldsymbol{\Theta}$. The blue and green arrows are the point estimates $\bar{\boldsymbol{\theta}}$. The green area is the uncertainty set $\mathcal{C}(\bar{\boldsymbol{\theta}}, \alpha)$.

Let $\mathbf{FOP}(\boldsymbol{\theta}, u)$ be the following problem

$$\begin{aligned} \text{minimize} \quad & \theta_1 x_1 + \theta_2 x_2 && \text{(8a)} \\ \text{subject to} \quad & (u-1)x_1 + x_2 \geq 2u - 1 && \text{(8b)} \\ & 2ux_1 + x_2 \geq 3u && \text{(8c)} \\ & 0 \leq x_1 \leq 6 && \text{(8d)} \\ & 0 \leq x_2 \leq u + 1. && \text{(8e)} \end{aligned}$$

Let the ground-truth parameter vector be $\boldsymbol{\theta}^* = (\cos(\pi/4), \sin(\pi/4))$ and $\mathcal{U} = \{u\}$ where $u > 2$ is a real constant. The optimal solution of **FOP** corresponding to $\boldsymbol{\theta}^*$ is point $\mathbf{c} = (2, 0)$, as shown in Figure 2. Now, suppose $\hat{\boldsymbol{\theta}}_k$ is uniformly and independently drawn from $\boldsymbol{\Theta} = \{(\cos\delta, \sin\delta) \,|\, \delta \in [0, \pi/2]\}$ for all $k \in [N]$, resulting in a dataset $\mathcal{D} = \{(\hat{\mathbf{x}}_k, u) \,|\, k \in [N]\}$, where $\hat{\mathbf{x}}_k = \tilde{\mathbf{x}}\left(\hat{\boldsymbol{\theta}}_k, u\right)$. Given the support $\boldsymbol{\Theta}$, the possible decisions in the dataset (assuming the solver only returns extreme points) are points $\mathbf{a} = (1 - 1/2u, u + 1)$, $\mathbf{b} = (1, u)$, and $\mathbf{c} = (2, 0)$. Let $N_{\mathbf{a}}$, $N_{\mathbf{b}}$, and $N_{\mathbf{c}}$ be the number of times $\mathbf{a}$, $\mathbf{b}$, $\mathbf{c}$ appear in $\mathcal{D}$, respectively.

Suppose $N_{\mathbf{a}}, N_{\mathbf{b}}, N_{\mathbf{c}} > 0$ and $N_{\mathbf{c}} > N_{\mathbf{a}}$ (which occurs with probability 1 as $N \to \infty$). Then the optimal solution to $\mathbf{IOP}(\mathcal{D})$ with the sub-optimality loss is $\bar{\boldsymbol{\theta}} = (\cos\delta_u, \sin\delta_u)$, where δ_u satisfies $\cos\delta_u = u/\sqrt{u^2+1}$, corresponding to the blue arrow orthogonal to Constraint (8b) in Figure 2. This implies that the decision policy derived by solving **FOP**, $\bar{\mathbf{x}}_{\text{IO}}(u) := \tilde{\mathbf{x}}\left(\bar{\boldsymbol{\theta}}, u\right)$, will randomly draw from $\{\mathbf{b}, \mathbf{c}\}$ as the recommendation. However, point $\mathbf{b}$ is sub-optimal with respect to $\boldsymbol{\theta}^*$ and most $\hat{\boldsymbol{\theta}}_k \in \Theta$. It can be shown that (see EC.1.1 for all calculations) the AOG and POG of $\bar{\mathbf{x}}_{\text{IO}}$ are $\sqrt{2}(u-1)/4$ and $(2\sqrt{u^2+1}+1)/\pi$, respectively, both of which can be arbitrarily large as u increases.

This example shows that **IOP** with the sub-optimality loss may provide a point estimate that is sufficiently different from the ground-truth $\boldsymbol{\theta}^*$ and most $\hat{\boldsymbol{\theta}} \sim \mathbb{P}_{\boldsymbol{\theta}}$ that in the downstream optimization phase, because estimation errors can be amplified, there can be considerable decision errors as measured by AOG and POG. A similar issue was highlighted in Elmachtoub and Grigas (2022). Existing literature typically relies on two strategies to tackle this issue: i) a decision-aware loss function for parameter estimation (Wilder et al. 2019, Mandi et al. 2022), and ii) a robust optimization model that accounts for the estimation errors for decision prescription (Sun et al. 2023, Chan et al. 2023a). Strategy i) is not applicable in our setting due to computational tractability and that we do not assume access to observations of the unknown parameters, which are inputs to most decision-aware loss functions. So, we propose a new decision pipeline along the lines of strategy ii).

3.4. Conformal Inverse Optimization Pipeline

In this section, we propose a new pipeline we refer to as *conformal inverse optimization*, due to its connection to the conformal prediction literature. As illustrated in Figure 1, in this new decision pipeline, for tractability, we still utilize **IOP** with the sub-optimality loss to obtain a point estimate of the unknown parameters. However, we do not solely rely on this point estimate to generate decisions. Instead, we calibrate an uncertainty set around this point estimate, which is then used in a robust optimization model for decision prescription. This robust optimization model hedges against deviations from the point estimate, leading to recommendations that are less sensitive to parameter estimation errors.

Specifically, we solve the following *robust forward optimization problem*.

$$\textbf{RFOP}\left(\mathcal{C}(\bar{\boldsymbol{\theta}}, \boldsymbol{\alpha}), \mathbf{u}\right) : \underset{\mathbf{x} \in \mathcal{X}(\mathbf{u})}{\text{minimize}} \ \max_{\boldsymbol{\theta} \in \mathcal{C}(\bar{\boldsymbol{\theta}}, \boldsymbol{\alpha})} f(\boldsymbol{\theta}, \mathbf{x}). \tag{9}$$

In this model, $\mathcal{C}$ is an uncertainty set centered at the point estimate $\bar{\boldsymbol{\theta}}$ with $\boldsymbol{\alpha}$ being parameters that control its size. Given the support $\boldsymbol{\Theta} = \left\{\boldsymbol{\theta} \in \mathbb{R}^d \,\middle|\, \|\boldsymbol{\theta}\|_2 = 1\right\}$, we define:

$$\mathcal{C}(\bar{\boldsymbol{\theta}}, \alpha) := \left\{\boldsymbol{\theta} \in \mathbb{R}^d \,|\, \|\boldsymbol{\theta}\|_2 = 1, \, \boldsymbol{\theta}^\intercal \bar{\boldsymbol{\theta}} \geq \cos\alpha\right\}. \tag{10}$$

This uncertainty set is a unit spherical cap in $\mathbb{R}^d$, i.e., the intersection of the unit sphere and a revolution cone with center $\bar{\boldsymbol{\theta}}$ and aperture angle α. In other words, $\mathcal{C}$ contains all the unit vectors in $\mathbb{R}^d$ that are within an angle α of the point estimate $\bar{\boldsymbol{\theta}}$.

3.5. Performance of the Conformal Inverse Optimization Pipeline: Example Revisited

Since we still utilize **IOP** for point estimation, our point estimate is $\bar{\boldsymbol{\theta}} = (\cos\delta, \sin\delta)$ where $0 \le \delta < \delta_u$, as illustrated in Section 3.3. When the aperture angle α satisifies $0 < \alpha \le \pi/2$, one can easily verify that the optimal solution to **RFOP** is always $\mathbf{c} = (2, 0)$, which is optimal with respect to the ground-truth $\boldsymbol{\theta}^*$ and most $\hat{\boldsymbol{\theta}} \sim \mathbb{P}_{\boldsymbol{\theta}}$. It can be shown that (see EC.1.2 for all calculations) the AOG and POG that result from this pipeline are 0 and upper bounded by $(4\sqrt{5} - \sqrt{17} - 12)/2\pi$, respectively, regardless of u. While this pipeline guarantees a bounded AOG and POG in this example, its performance still depends on the choice of α. We address this issue in the following section.

4. Conformal Inverse Optimization

Next, we present a principled approach to learning uncertainty sets based on decision data. In particular, the learned uncertainty set contains a parameter vector that makes the next decision maker's decision optimal with a specified probability. We call this approach *conformal inverse optimization* 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. Accordingly, we call the new decision pipeline the conformal IO pipeline. As illustrated in Figure 1, this new pipeline has four steps: i) data split, ii) point estimation, iii) uncertainty set calibration, iv) and decision prescription. Steps i)–iii) comprise conformal IO and are presented in Section 4.1. We analyze the properties of conformal IO in Section 4.2. Finally, in Section 4.3, we discuss methods that can be used to speed up step iv).

4.1. Learning an Uncertainty Set

4.1.1. Data split. We first split the decision dataset $\mathcal{D}$ into training ($\mathcal{D}_{\text{train}}$) and validation ($\mathcal{D}_{\text{val}}$) sets. Let $\mathcal{K}_{\text{train}}$ and $\mathcal{K}_{\text{val}}$ index the elements in $\mathcal{D}_{\text{train}}$ and $\mathcal{D}_{\text{val}}$, respectively, and $N_{\text{train}} = |\mathcal{D}_{\text{train}}|$ and $N_{\text{val}} = |\mathcal{D}_{\text{val}}|$.

4.1.2. Point estimation. Given the training set $\mathcal{D}_{\text{train}}$, we solve $\mathbf{IOP}(\mathcal{D}_{\text{train}})$ with the sub-optimality loss function to obtain a point estimate $\bar{\boldsymbol{\theta}}$. Other approaches are possible, such as using a different loss function in $\mathbf{IOP}(\mathcal{D}_{\text{train}})$ or employing end-to-end machine learning approaches by treating **FOP** as an optimization layer that returns a decision based on $\boldsymbol{\theta}$ and applying gradient-based methods, e.g., Berthet et al. (2020), to optimize $\boldsymbol{\theta}$. We will compare our approach with these others in Section 5. Note that our uncertainty set calibration method works with any point estimation approach.

4.1.3. Uncertainty set calibration. Given a point estimate $\bar{\boldsymbol{\theta}}$, we construct an uncertainty set using the other dataset $\mathcal{D}_{\text{val}}$ that, with a specified probability, contains parameters that make the next unseen decision optimal. This property is critical for the results in Section 4.2 to hold. Note that we can naively achieve a probability of 1 by setting $\alpha = \pi$, but the resulting **RFOP** would generate overly conservative decisions. Hence, we are interested in learning the smallest uncertainty set that achieves the desired probability. The uncertainty set *calibration problem* is

$$\mathbf{CP}(\bar{\boldsymbol{\theta}}, \mathcal{D}_{\text{val}}, \gamma): \underset{\alpha, \{\boldsymbol{\theta}_k\}_{k\in\mathcal{K}_{\text{val}}}}{\text{minimize}} \quad \alpha \tag{11a}$$

$$\text{subject to} \quad \hat{\mathbf{x}}_k \in \mathcal{X}^{\text{OPT}}(\boldsymbol{\theta}_k, \mathbf{u}_k), \ \forall k \in \mathcal{K}_{\text{val}} \tag{11b}$$

$$\sum_{k\in\mathcal{K}_{\text{val}}} \mathbb{1}\left[\boldsymbol{\theta}_k \in \mathcal{C}(\bar{\boldsymbol{\theta}}, \alpha)\right] \geq \gamma(N_{\text{val}} + 1) \tag{11c}$$

$$\|\boldsymbol{\theta}_k\|_2 = 1, \ \forall k \in \mathcal{K}_{\text{val}} \tag{11d}$$

$$0 < \alpha \leq \pi, \tag{11e}$$

where α controls the size of the uncertainty set, $\boldsymbol{\theta}_k$ represents a parameter vector associated with data point $k \in \mathcal{K}_{\text{val}}$, and $\gamma \in [0, 1]$ is a user-specified confidence level (e.g., 90%). The objective function (11a) minimizes the size of the uncertainty set, i.e., the aperture angle of the spherical cap. Constraints (11b) ensure that $\boldsymbol{\theta}_k$ makes the observed decision $\hat{\mathbf{x}}_k$ optimal for $k \in \mathcal{K}_{\text{val}}$. Constraint (11c) ensures that at least γ of all decisions in $\mathcal{D}_{\text{val}}$ are optimal with respect to some vector in $\mathcal{C}$. Constraints (11d) ensure that the parameter vectors are on the unit sphere as required in the definition of $\mathcal{C}$ in Equation (10).

Remark 2 (Optimality Conditions). The form of Constraints (11b) depends on the structure of **FOP**. For example, when the **FOP** is convex, we can use the KKT conditions. For non-convex forward problems, we can replace Constraints (11b) with $f(\boldsymbol{\theta}_k, \hat{\mathbf{x}}_k) \leq f(\boldsymbol{\theta}_k, \mathbf{x})$ for all $\mathbf{x} \in \mathcal{X}(\mathbf{u})$, which can be generated on-the-fly in a cutting-plane fashion.

Remark 3 (Feasibility). For **CP** to be feasible, we require, for each observed decision, that there exists a $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ that makes it optimal (Constraints (11b)). This is a mild assumption. First, it holds for a range of problems, e.g., the shortest path problem (set the cost of chosen arcs to be zero), knapsack problem (set value of excluded items to zero), even if the decision maker is subject to bounded rationality. Second, if $\mathcal{D}_{\text{val}}$ violates this assumption, **CP** can be easily modified to accommodate. For example, Constraints (11b)

could be rewritten as ϵ-optimality for an appropriately chosen ϵ, or with a weighted value of ϵ to be minimized in the objective along with α. These are standard approaches for addressing IO with noisy data (Chan et al. 2023b).

Solving **CP** appears to be a challenging task due to its non-convexity, which stems from Constraints (11d), and the rapid growth in its size as N_{val} increases, driven by Constraints (11b). However, we can efficiently solve **CP** based on the following insights.

THEOREM 1. *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 $c_k := \max_{\boldsymbol{\theta}_k}\left\{\boldsymbol{\theta}_k^\intercal \bar{\boldsymbol{\theta}} \,\middle|\, \hat{\mathbf{x}}_k \in \mathcal{X}^{\text{OPT}}(\boldsymbol{\theta}_k, \mathbf{u}_k),\ \|\boldsymbol{\theta}_k\|_2 = 1\right\}$.*

Theorem 1 states that we can solve **CP** by i) solving N_{val} optimization problems whose size is independent of N_{val} and ii) finding a quantile in a set of N_{val} elements. Step i) is parallelizable and Step ii) can be done in $O\left(N_{\text{val}}\log(\tau)\right)$ time. Since the problem required for evaluating c_k is a maximization problem, we can replace the constraint $\|\boldsymbol{\theta}_k\|_2 = 1$ with $\|\boldsymbol{\theta}_k\|_2 \le 1$ if $\bar{\boldsymbol{\theta}} \in \mathbb{R}^d_+$, so this problem is convex when **FOP** is convex.

4.2. Properties of the Learned Uncertainty Set

THEOREM 2 **(Set Validity)**. *Let $\mathcal{D}_{\text{val}}$ be a dataset that satisfies Assumption 1, $(\hat{\boldsymbol{\theta}}, \mathbf{u})$ be a new i.i.d. sample from $\mathbb{P}_{(\boldsymbol{\theta},\mathbf{u})}$, $\hat{\mathbf{x}} \in \tilde{\mathbf{x}}(\hat{\boldsymbol{\theta}}, \mathbf{u})$, $\hat{\boldsymbol{\Theta}} = \boldsymbol{\Theta}^{\text{OPT}}(\mathbf{u}, \hat{\mathbf{x}})$, and α_γ be an optimal solution to $\mathbf{CP}(\bar{\boldsymbol{\theta}}, \mathcal{D}_{\text{val}}, \gamma)$ where $\bar{\boldsymbol{\theta}} \in \mathbb{R}^d$. For any $\gamma \in [0, N_{val}/(N_{val}+1)]$,*

$$\mathbb{P}\left(\hat{\boldsymbol{\Theta}} \cap \mathcal{C}(\bar{\boldsymbol{\theta}}, \alpha_\gamma) \neq \varnothing\right) \ge \gamma. \tag{12}$$

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

$$\left|\mathbb{P}\left(\hat{\boldsymbol{\Theta}} \cap \mathcal{C}(\bar{\boldsymbol{\theta}}, \alpha_\gamma) \neq \varnothing\right) - \gamma\right| \le \sqrt{\frac{8\log(N_{val}+1) + 2\log N_{val}}{N_{val}}} + \frac{2}{N_{val}}. \tag{13}$$

Theorem 2 states that our learned uncertainty set is conservatively valid and asymptotically exact. More specifically, first, our method will produce a set that contains a $\boldsymbol{\theta}$ that makes the next decision maker's decision optimal no less than γ of the time that it is used (conservatively valid). The probability in Inequality (12) 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 decision maker's decision being covered is within $\epsilon(N_{\text{val}})$ from γ. The probability in Inequality (13) 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 $\mathbb{P}\left(\hat{\boldsymbol{\Theta}} \cap \mathcal{C}(\bar{\boldsymbol{\theta}}, \alpha_\gamma) \neq \varnothing\right)$ to γ as N_{val} goes to infinity.

Now, we relate the validity results to the performance of conformal IO. The following Lemma is an immediate result of the objective function f being linear in $\boldsymbol{\theta}$.

LEMMA 1. *For any* $\left(\hat{\boldsymbol{\theta}}, \mathbf{u}\right) \in \boldsymbol{\Theta} \times \mathcal{U}$ *and* $\hat{\mathbf{x}} \in \tilde{\mathbf{x}}\left(\hat{\boldsymbol{\theta}}, \mathbf{u}\right)$*, there exists a constant* $\nu(\hat{\mathbf{x}}) \in \mathbb{R}_+$ *such that, for any* $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \boldsymbol{\Theta}$*, we have* $f(\boldsymbol{\theta}, \hat{\mathbf{x}}) - f(\boldsymbol{\theta}', \hat{\mathbf{x}}) \leq \nu(\hat{\mathbf{x}}) \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_2$.

THEOREM 3 **(POG and AOG Bounds)**. *Let* $\bar{\mathbf{x}}_{\text{CIO}}(\mathbf{u})$ *be an optimal solution to* **RFOP** $\left(\mathcal{C}(\bar{\boldsymbol{\theta}}, \alpha_1), \mathbf{u}\right)$ *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* $\mathbf{x}' \in \tilde{\mathbf{x}}(\boldsymbol{\theta}', \mathbf{u}')$*,* $\mathbb{P}\left(\mathcal{C}(\bar{\boldsymbol{\theta}}, \alpha_1) \cap \boldsymbol{\Theta}^{\text{OPT}}(\mathbf{u}', \mathbf{x}') \neq \varnothing\right) = 1$*. If Assumptions 3–2 hold, then*

$$\text{POG}(\bar{\mathbf{x}}_{\text{CIO}}) \leq (\eta - 2\cos 2\alpha_1 + 2)\mu + \eta\mu_{\text{CIO}}, \tag{14}$$

and

$$\text{AOG}(\bar{\mathbf{x}}_{\text{CIO}}) \leq (2 - 2\cos 2\alpha_1 + \eta + \sigma)\mu^* + (\eta + \sigma)\mu_{\text{CIO}}, \tag{15}$$

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

Theorem 3 state that, when the uncertainty set contains a $\boldsymbol{\theta}$ that makes the next observed 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**. Moreover, we show numerically in Section 5 that, when using $\gamma < 100\%$, conformal IO still demonstrates favorable performance compared to standard IO. We note that the bounds in Theorem 3 depend on several problem-dependent constants. To demonstrate the tightness of these bounds, we visualize their numerical values for the example in Section 3.3 in EC.1.3. Our bounds closely follows the AOG and POG achieved by the conformal IO pipeline, which outperforms the standard IO pipeline by a large margin especially when u is large.

4.3. Acceleration Scheme for Prescribing New Decisions

Once an uncertainty set $\mathcal{C}\left(\bar{\boldsymbol{\theta}}, \alpha_\gamma\right)$ is calibrated and a new contex $\mathbf{u}$ is given, we then solve **RFOP** to prescribe new decisions. In this section, we discuss the computational challenges of solving this problem and methods to accelerate its solution process.

Let $\mathbf{f}(\mathbf{x}) := (f_1(\mathbf{x}), f_2(\mathbf{x}), \ldots, f_d(\mathbf{x}))$, then $\mathbf{RFOP}\left(\mathcal{C}\left(\bar{\boldsymbol{\theta}}, \alpha_\gamma\right), \mathbf{u}\right)$ can be written as

$$\underset{\mathbf{x}\in\mathcal{X}(\mathbf{u})}{\text{minimize}}\ h(\mathbf{x}), \tag{16}$$

where $h(\mathbf{x}) := \max_{\boldsymbol{\theta}\in\mathbb{R}^d} \left\{\boldsymbol{\theta}^\intercal \mathbf{f}(\mathbf{x}) \,\middle|\, \bar{\boldsymbol{\theta}}^\intercal \boldsymbol{\theta} \geq \cos\alpha_\gamma,\ \|\boldsymbol{\theta}\|_2 = 1\right\}$. Problem (16) can be solved with a general purpose cutting plane algorithm that iteratively solves the inner maximization problem to identify cuts to be added to the outer minimization problem (detailed in EC.3.5). While this algorithm works well for small problem instances, the computation might be prohibitively expensive when the unknown vector $\boldsymbol{\theta}$ is high dimensional because the cut generation problem is non-convex due to the constraint $\|\boldsymbol{\theta}\|_2 = 1$ and the algorithm may take a large number of iterations to converge.

Note that in many applications, we know $\bar{\boldsymbol{\theta}}$ and $\mathbf{f}(\mathbf{x})$ are both non-negative. For example, in routing problems where the objective function is $f(\boldsymbol{\theta}, \mathbf{x}) = \boldsymbol{\theta}^\intercal \mathbf{x}$, the travel costs represented by $\boldsymbol{\theta}$ are typically non-negative and the routing decisions represented by $\mathbf{x}$ are binary (Zattoni Scroccaro et al. 2024). In these settings, we can replace the constraint $\|\boldsymbol{\theta}\|_2 = 1$ with $\|\boldsymbol{\theta}\|_2 \leq 1$, leading to the following reformulation of Problem (16).

PROPOSITION 1. *Given a fixed $\mathbf{u} \in \mathcal{U}$, assuming that $\bar{\boldsymbol{\theta}} \in \mathbb{R}^d_+$ and $\mathbf{f}(\mathbf{x}) \in \mathbb{R}^d_+$ for any $\mathbf{x} \in \mathcal{X}(\mathbf{u})$, then Problem (16) can be formulated as*

$$\underset{\mathbf{x}\in\mathcal{X}(\mathbf{u}),\, \lambda\in\mathbb{R}_+}{\text{minimize}} \quad \left\|\mathbf{f}(\mathbf{x}) + \lambda\bar{\boldsymbol{\theta}}\right\|_2 - \lambda\cos\alpha_\gamma. \tag{17}$$

Going forward, we assume that $\bar{\boldsymbol{\theta}}$ and $\mathbf{f}(\mathbf{x})$ are non-negative. While this reformulation accelerates the solution process, Problem (17) may still be difficult to solve. For example, when **FOP** involves discrete decisions, Problem (17) is a quadratic mixed integer program.

In light of this challenge, we develop an approximation to Problem (17). In particular, we approximate the l_2 norm using a *polyhedral norm* $|||\cdot|||$. A polyhedral norm is such that the unit ball $\left\{\boldsymbol{\theta}' \in \mathbb{R}^d \,\middle|\, |||\boldsymbol{\theta}'||| \leq 1\right\}$ associated with the norm is a polyhedron. A benefit of this approach is that we can dualize the inner problem and generate an approximation to **RFOP** with similar computational tractability as **FOP**. Specifically, we solve

$$\underset{\mathbf{x}\in\mathcal{X}(\mathbf{u})}{\text{minimize}}\ g(\mathbf{x}), \tag{18}$$

where

$$g(\mathbf{x}) := \max_{\boldsymbol{\theta}\in\mathbb{R}^d_+} \left\{\boldsymbol{\theta}^\intercal \mathbf{f}(\mathbf{x}) \,\middle|\, \bar{\boldsymbol{\theta}}^\intercal \boldsymbol{\theta} \geq \cos\alpha_\gamma,\ |||\boldsymbol{\theta}||| \leq 1\right\}. \tag{19}$$

The next proposition bounds the optimality loss of using an optimal solution to (18), as a function of the approximation error of the polyhedral norm to the l_2 norm.

PROPOSITION 2. *Let* $|||\cdot|||$ *be a norm that satisfies* $|||\boldsymbol{\theta}||| - \epsilon \le \|\boldsymbol{\theta}\|_2 \le |||\boldsymbol{\theta}||| + \epsilon$ *for any* $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ *and some* $\epsilon \in \mathbb{R}_+$, $\mathbf{x}^*$ *and* $\mathbf{x}'$ *be optimal solutions to Problems* (17) *and* (18), *respectively, and* $v := \max\{\|\mathbf{f}(\mathbf{x})\|_2 \,|\, \mathbf{x} \in \mathcal{X}(\mathbf{u})\}$, *which is finite because the basis functions* f_i *are continuous and the feasible set* $\mathcal{X}(\mathbf{u})$ *is compact. Assuming that* $\bar{\boldsymbol{\theta}} \in \mathbb{R}^d_+$, $\alpha_\gamma \in (0, \pi/2)$, *and* $\mathbf{f}(\mathbf{x}) \in \mathbb{R}^d_+$ *for any* $\mathbf{x} \in \mathcal{X}(\mathbf{u})$, *then* $h(\mathbf{x}') - h(\mathbf{x}^*) \le v(\epsilon^2 + 2\epsilon)/\sin\alpha_\gamma$.

Indeed, the idea of approximating the l_2 norm using a polyhedral norm has been used to improve the computational tractability of other norm-constrained optimization models. Existing literature has proposed the D-norm (Bertsimas et al. 2004), which approximates a vector's l_2 norm as a weighted sum of its largest entries, and the D_p norm (Chen et al. 2021), which derives an approximation as the maximal inner product of the vector of interest and any vectors that satisfy a set of constraints on their l_1 and l_∞ norms. While both the D-norm and D_p-norm enjoy global bounds on the approximation error, our experiments suggest that their local approximation errors for $\boldsymbol{\theta} \in \left\{\boldsymbol{\theta}' \in \mathbb{R}^d \,\middle|\, \bar{\boldsymbol{\theta}}^\intercal \boldsymbol{\theta}' \ge \cos\alpha_\gamma,\ \|\boldsymbol{\theta}\|_2 \le 1\right\}$ were large and the resulting decisions were of poor quality. Thus, we propose the following data-driven norm, which can be written as a linear combination of polyhedral norms and whose parameters can be tuned to refine the approximation locally.

DEFINITION 4 (DATA-DRIVEN NORM). Given a set of polyhedral norms $\{|||\cdot|||_t\}_{t\in[n_{\text{norm}}]}$. Let $\boldsymbol{\beta} \in \mathbb{R}^{n_{\text{norm}}}_+$. A *data-driven norm* of a vector $\boldsymbol{\theta} \in \mathbb{R}^d$ is defined as

$$|||\boldsymbol{\theta}|||_{\boldsymbol{\beta}} := \sum_{t\in[n_{\text{norm}}]} \beta_t \, |||\boldsymbol{\theta}|||_t \,.$$

We require $\boldsymbol{\beta}$ to be non-negative so $|||\boldsymbol{\theta}|||_{\boldsymbol{\beta}}$ is non-negative. Moreover, it is easy to verify that $|||\cdot|||_{\boldsymbol{\beta}}$ satisfies the triangle inequality and $|||\mathbf{0}|||_{\boldsymbol{\beta}} = 0$. So, $|||\cdot|||_{\boldsymbol{\beta}}$ is a norm.

There are many possible norms that could be used to form the data-driven norm, including the l_1, l_∞, D-norm and D_p-norm. We demonstrate in our numerical results that simply using l_1 and l_∞ achieves strong empirical performance, approaching the best possible performance (obtained using an exact cutting plane algorithm). We write the specific data-driven norm that uses l_1 and l_∞ as building blocks as

$$|||\boldsymbol{\theta}|||_{\beta_1,\beta_2} := \beta_1\|\boldsymbol{\theta}\|_1 + \beta_2\|\boldsymbol{\theta}\|_\infty. \tag{20}$$

To approximate the l_2 norm with $|||\boldsymbol{\theta}|||_{\beta_1,\beta_2}$, we need to fit the model parameters β_1 and β_2 using data collected around the region of interest, i.e., $\left\{\boldsymbol{\theta} \in \mathbb{R}^d_+ \,\middle|\, \bar{\boldsymbol{\theta}}^\intercal \boldsymbol{\theta} \ge \cos\alpha_\gamma,\ \|\boldsymbol{\theta}\|_2 \le 1\right\}$.

Generating the training data for this linear regression model can be a daunting computational task when $\boldsymbol{\theta}$ is high dimensional. It typically involves i) generating random vectors in a d-dimensional space and ii) accepting the vector if it falls inside the specified region. The probability of accepting in step ii) decreases quickly as the dimensionality increases; see discussions in Arun and Venkatapathi (2021). To address this challenge, we build a training data set for fitting the data-driven norm using the cost vectors generated in the uncertainty set calibration step, i.e., in the solution of **CP** (see Section 4.1). Recall that solving **CP** can be done by solving N_{val} optimization problems (Theorem 1), where each is an inverse optimization problem with one data point from the validation decision dataset. The resulting cost vectors are guaranteed to be around the sub-region of interest. To enhance this training set with data that has additional variation in the l_2 norm, we randomly sample cost vector pairs and create a linear combination of them. This complete procedure is detailed in EC.3.6. Once the training data is generated, $|||\cdot|||_{\beta_1,\beta_2}$ can be fit as a constrained linear regression model with non-negative parameters and a zero intercept.

Next, we specialize the bound in Proposition 2 for the data-driven norm $|||\cdot|||_{\beta_1,\beta_2}$.

COROLLARY 1. *For any fixed* $\bar{\boldsymbol{\theta}} \in \mathbb{R}^d_+$, $\mathbf{u} \in \mathcal{U}$, *and* $\alpha_\gamma \in (0, \pi/2)$, *let* $\mathbf{x}^*$ *be an optimal solution to Problem* (17) *and* $\mathbf{x}'$ *be an optimal solution to Problem* (18) *with* $|||\cdot|||_{\beta_1,\beta_2}$. *If* $\beta_1\beta_2 \geq 1/4(2-\cos^2\alpha_\gamma)$ *and* $\left|\left|\left|\bar{\boldsymbol{\theta}}\right|\right|\right|_{\beta_1,\beta_2} \leq 1/\cos\alpha_\gamma$, *then* $h(\mathbf{x}') - h(\mathbf{x}^*) \leq v \sin\alpha_\gamma$.

Corollary 1 presents an upper bound on the solution quality achieved by our data-driven approximation model. The bound requires $\beta_1\beta_2$ to be larger than a threshold. If both β_1 and β_2 are small, then the resulting polyhedron would be much larger than the unit l_2 ball, leading to overly conservative cost estimation for a decision $\mathbf{x}$. In principle, one should include this threshold constraint in the parameter estimation problem. However, we found that fitting the regression without this constraint always resulted in parameters that satisfied the inequality. The intuition is that small $\beta_1\beta_2$ generally results in poor l_2 approximation error, which is the training loss used for parameter estimation. Additionally, the bound also requires that $\left|\left|\left|\bar{\boldsymbol{\theta}}\right|\right|\right|_{\beta_1,\beta_2}$ be below a threshold. If β_1 and β_2 are too large, the resulting polyhedron would be much smaller than the unit l_2 ball, leading to an underestimation of the decision cost. Similarly, we find empirically that our fitted parameters always satisfy this constraint as $\bar{\boldsymbol{\theta}}$ is the "center" of the original uncertainty set, around which the data-driven norm typically achieves small approximation errors, i.e., $\left|\left|\left|\bar{\boldsymbol{\theta}}\right|\right|\right|_{\beta_1,\beta_2} \approx 1$. Thus, we

ignore both constraints so we can leverage off-the-shelf ML packages to fit the data-driven norm. If the fitted parameters violate these constraints, one can simply fit the model again by solving a constrained optimization model with the constraint directly embedded.

The bound in Corollary 1 is useful in providing direct insight into the performance of the general bound. We expect the bound in Proposition 2 to increase as α_γ increases, reflecting the fact that it is generally harder to achieve high approximation accuracy and thus high decision quality when the uncertainty set becomes larger. Also, we expect the bound to go to zero when α_γ goes to zero. But without an explicit relationship between ϵ and α_γ, these behaviors are difficult to establish. With $|||\cdot|||_{\beta_1\beta_2}$, Corollary 1 provides a clear picture of the bound's behavior as a function of α: increasing in α_γ and converging to zero as α_γ goes to zero, as expected.

Finally, we present a complete formulation of Problem (18) with $|||\cdot|||_{\beta_1,\beta_2}$. The resulting model is a linear program. Note that a similar model can be written for a general data-driven norm $|||\cdot|||_{\boldsymbol{\beta}}$; it will be a linear program as well, but the structure will depend on the constituent polyhedral norms of $|||\cdot|||_{\boldsymbol{\beta}}$, in particular how the inner problem gets dualized.

PROPOSITION 3. *Problem (18) with the data-driven norm* $|||\cdot|||_{\beta_1,\beta_2}$ *is*

$$\begin{aligned}
\underset{\mathbf{x},\boldsymbol{\phi},\lambda,\zeta}{\text{minimize}} \quad & \zeta - \lambda\cos\alpha_\gamma \\
\text{subject to} \quad & \beta_1\zeta + \phi_i - \bar{\theta}_i\lambda \geq f_i(\mathbf{x}), \quad \forall i \in [d] \\
& \beta_2\zeta - \mathbf{1}^\intercal\boldsymbol{\phi} \geq 0 \\
& \mathbf{x} \in \mathcal{X}(\mathbf{u}) \\
& \boldsymbol{\phi} \in \mathbb{R}^d_+ \\
& \lambda, \zeta \geq 0.
\end{aligned}$$

5. Numerical Studies

We perform computational studies using synthetic problem instances, which allows us to compare the performance of the standard and conformal IO pipelines. We introduce the experimental design in Section 5.1 and present the numerical results in Section 5.2.

5.1. Experiment Setup

We consider two forward problems: (1) a shortest path problem on a 5×5 grid (linear program) and (2) a knapsack problem with 10 items (integer program). See EC.3.2 for

their formulations. For both problems, we generate a ground-truth $\boldsymbol{\theta}^*$ and a dataset of $N = 1000$ decisions, corresponding to distinct decision makers. For each decision maker $k \in [N]$, we generate her perceptions as $\hat{\theta}_i^k = \max\{\theta_i^{*k} p_i^k + \epsilon_i^k, 0\} + 0.1$ for $i \in [d]$ where p_i^k is drawn from $[1/2, 2]$ and ϵ_i^k is drawn from a standard normal distribution. For the shortest path problem, $\mathbf{u}^k$ indicate a random origin-destination pair. For the knapsack problem, item weights $w_i, \forall i \in [d]$ are drawn from $[1, 10]$ and are shared among decision makers. Each decision maker $k \in [N]$ has a budget $u^k = q^k \sum_i w_i$ where q^k is drawn from $[1/5, 5]$.

Conformal IO is compatible with any point estimation methods (Step 2 in Section 4.1). To our knowledge, i) solving **IOP** with the sub-optimality loss and ii) the PFYL approach from Berthet et al. (2020) are the only two methods that can perform this task *at scale*. We thus implement conformal IO with both methods. They also serve as our baselines. We call both i) and ii) "standard IO" to emphasize that they rely on a point estimation for decision prescription, although PFYL is not an IO approach. See EC.3 for implementation details. In all experiments, conformal IO uses the training set for point estimation and the validation set for calibration, while standard IO uses the union of the training and validation sets for point estimation. So, the two pipelines have access to the same amount of data and are evaluated on the same test set. Experiments are based on 60-20-20 train-validation-test splits and are repeated 10 times with different random seeds.

5.2. Experiment Results

5.2.1. Uncertainty set validity. We first evaluate the out-of-sample coverage achieved by the uncertainty set learned using conformal IO under different target levels γ and sample sizes N_{val}. As shown in Figure 3, when the validation set is small ($N_{\text{val}} = 10$), we always achieve the specified target but $\mathcal{C}(\bar{\boldsymbol{\theta}}, \alpha_\gamma)$ tends to over-cover. When using larger validation sets ($N_{\text{val}} \geq 100$), our coverage level gets closer to the specified γ. These empirical findings echo our theoretical analysis (Theorem 2).

5.2.2. Solution quality with respect to AOG and POG. We next compare the out-of-sample AOG and POG achieved by the standard and conformal IO pipelines. As shown in Figure 4, conformal IO typically achieves lower POG and AOG. 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 AOG by 40.3–57.0% and the POG by 13.5–20.1% for the knapsack problem. The solutions generated by conformal IO are not only of higher quality, but also perceived to be higher quality.

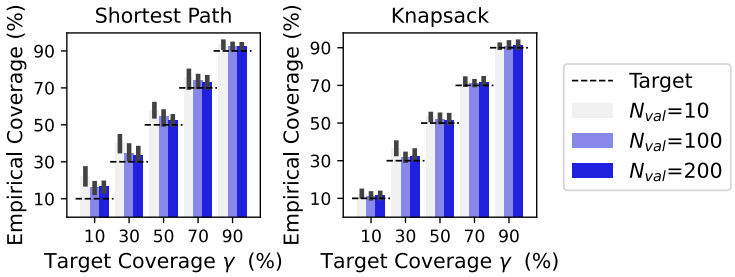


Figure 3 Empirical coverage achieved by the learned uncertainty set (error bar = range).

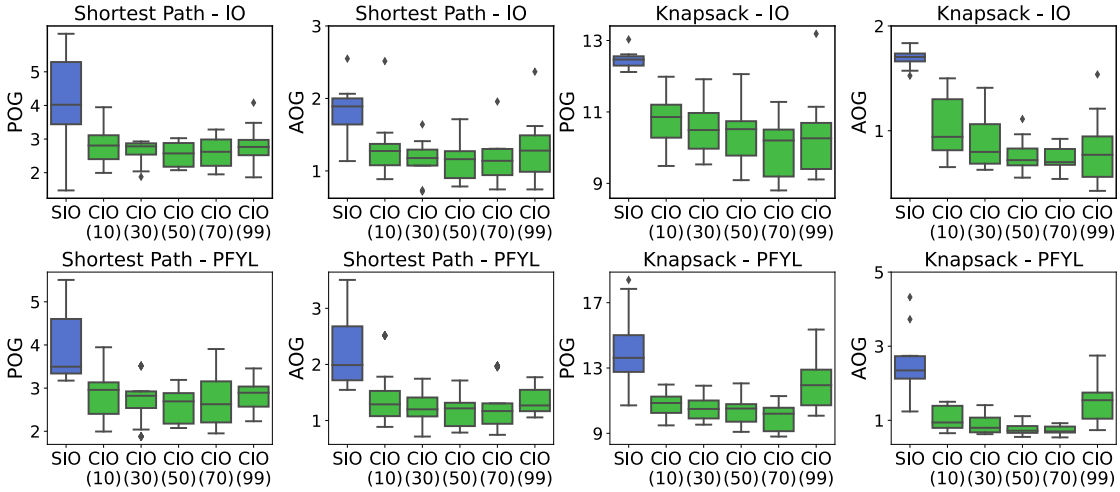


Figure 4 Performance profile of standard (SIO) and conformal IO (CIO) pipelines. The numbers in parentheses represent the coverage level γ used in the Conformal IO pipeline.

5.2.3. Choice of hyper-parameters in conformal IO. We provide empirical evidence that sheds light on the choice of two important hyper-parameters: i) *confidence level* γ, and ii) *train-validation split ratio*. Regarding γ, as shown in Figure 4, the performance of conformal IO improves quickly as γ increases from 0 to 50% and remains stable and even worsens slightly after that. Hence, it is possible to improve the performance of conformal IO by carefully tuning γ. Regarding the train-validation split, intuitively, both the point estimation and uncertainty set calibration steps can benefit from more data. However, when data is limited, we need to strike a balance in how much data is given to each step. To investigate, we implement conformal IO for the shortest path problem under different dataset sizes $N_{\text{train}} + N_{\text{val}}$ and train-validation split ratios $N_{\text{val}}/(N_{\text{train}} + N_{\text{val}})$. As shown in Figure 5, when the dataset is small (160), there is no benefit of using conformal IO because there is not enough data to obtain both a good point estimate and a good uncertainty set at the same time. However, as the dataset grows, conformal IO can achieve significantly

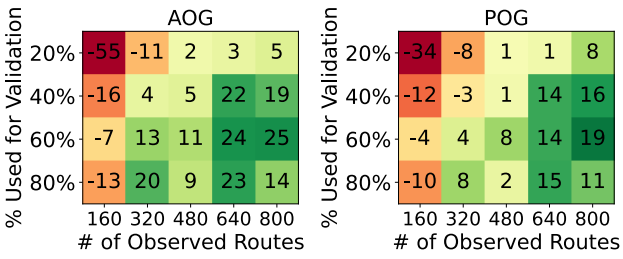


Figure 5 Percentage reduction in test AOG and POG when using the conformal IO vs classic IO.

lower AOG and POG than standard IO. Furthermore, for medium- to large-sized datasets, more data should be given to the calibration step, which reinforces our theoretical analysis.

5.2.4. Computational efficiency. As shown in Table 1, standard and conformal IO require similar "training" times. In standard IO, training refers to generating the point estimate by solving the **IOP**. In conformal IO, training time is the time it takes to solve both the **IOP** and **CP**. When **FOP** is an integer program (knapsack), the training of conformal IO is faster than standard IO because it replaces a relatively large inverse integer program (associated with $\mathcal{D}_{\text{train}} \cup \mathcal{D}_{\text{val}}$), which is notoriously difficult to solve (Bodur et al. 2022), with a smaller inverse integer program (associated with $\mathcal{D}_{\text{train}}$) and a set of small calibration problems (Theorem 1). At prediction time, our method achieves lower AOG and POG at the cost of solving a more challenging **RFOP**. Since these problems are small, we can solve them exactly using a cutting plane method described in EC.3.5.

Table 1 Average (standard deviation) computational time of standard and conformal IO pipelines in seconds.

	Training		Prediction (per decision)	
Problem	Standard IO	Conformal IO	**FOP**	**RFOP**
Shortest Path	0.18 (0.02)	0.27 (0.03)	0.01 (0.00)	0.63 (0.12)
Knapsack	2.47 (0.37)	1.95 (0.32)	0.01 (0.00)	0.44 (0.15)

Next, we investigate the performance of the data-driven approximation models described in Section 4.3. We focus on the shortest path problem as our approximation requires the cost vector to be non-negative. We first present the l_2 norm approximation error achieved by our data-driven norm compared to other norms from the literature. We then investigate if the reduced approximation error can be translated into better decisions.

Approximation accuracy of the data-driven norm. As presented in Table 2, the out-of-sample MAE achieved by our data-driven norm is 87% lower than the second lowest (D norm) and is orders of magnitude lower than the D_p, l_1 and l_∞ norms.

Table 2 Mean (standard deviation) out-of-sample l_2 norm approximation MAE in %.

	Data-driven Norm	D Norm	D_p Norm	l_1 Norm	l_∞ Norm
MAE	0.51 (0.001)	3.93 (0.00)	23.82 (0.00)	708.75 (0.002)	84.22 (0.001)

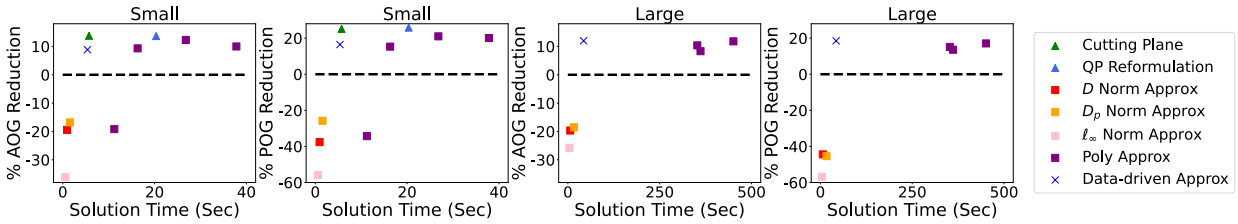


Figure 6 The trade-off between decision quality and solution time. The ℓ_1 approximation model are omitted as they are far worse than others. The small and large instances are on 5×5 and 10×10 grids, respectively.

Tradeoff between solution time and solution quality. We compare our data-driven approximation model against approximation models that utilize the D, D_p, l_1, and l_∞ norms. In addition, we implement the polyhedral outer approximation method proposed by Kocuk (2021). This method can approximate the second-order cone described by the l_2 norm constraint to an arbitrary accuracy ϵ with a polyhedral cone in an extended space. We vary ϵ in $\{0.1, 0.01, 0.001, 0.0001\}$. To obtain the exact solution, we implement the cutting plane algorithm and the quadratic program reformulation presented in Section 4.3. For each method, we report the average percentage reduction in out-of-sample AOG and POG with respect to the standard IO pipeline (y-axis) and the solution time required to generate 200 route recommendations (x-axis) in Figure 6. For small problem instances, our approach is the only norm-based approximation method that can achieve AOG and POG reductions compared to standard IO. Some of the polyhedral approximation models of Kocuk (2021) can achieve similar AOG and POG reduction, but take 2-5 times as long to solve. The only competitor to our approach for small problem sizes is the exact cutting plane method, which achieves comparable AOG and POG with similar solution time. For large problem instances, however, exact approaches were not able to find a feasible solution. In these cases, our data-driven norm approximation model essentially Pareto-dominated the other models: 1) it achieved similar solution quality using less than 10% of the solution time compared to the polyhedral approximation model, and 2) at similar computational expense was the only norm-approximation model to achieve AOG and POG reductions.

5.2.5. Summary. The main takeaways from this section are: (1) conformal IO generates uncertainty sets that satisfy the target coverage levels, (2) the conformal IO pipeline

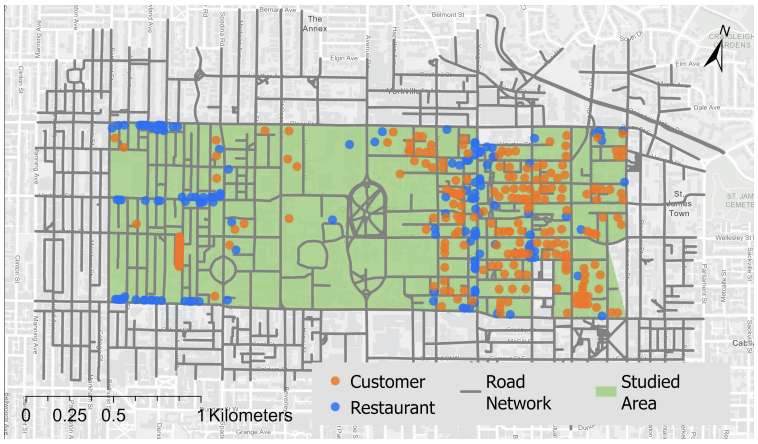


Figure 7 Road network around the area of study in Toronto, Canada (forward sortation areas M5S and M4Y).

produces decisions with higher absolute and perceived quality than the standard IO pipeline, and (3) replacing **RFOP** in the conformal IO pipeline with our data-driven norm approximation model significantly enhances its computational tractability while still improving absolute and perceived decision quality compared to the standard IO pipeline.

6. Case Study

Finally, we present a case study applying conformal IO to provide delivery path recommendations in Toronto, Canada. The primary goal of food delivery platforms is to fulfill customer demand in a timely manner. However, a theoretical shortest path is often suboptimal from the courier's perspective (Fu et al. 2023). So, platforms often need to trade delivery time for path adherence, which is crucial as it impacts many downstream operations, e.g., order batching, which relies on accurate route modeling (Liu et al. 2021). This raises a key question: how much additional travel time must the platform bear to achieve a certain adherence rate? In this section, we perform numerical experiments to quantify this trade-off. We focus on bike delivery, which has become increasingly popular in urban areas as it is low-cost, low-emission, and often faster than car delivery (DoorDash 2024).

6.1. Data

We focus on delivery trips whose origin and destination are within the forward sortation areas M5S and M4Y in downtown Toronto, as visualized in Figure 7. This area is densely populated and has well-connected cycling infrastructure (Lin et al. 2021).

Road network. We retrieve the centerline road network from the Toronto Open Data Portal (City of Toronto 2020). We focus on the road network within 500 meters of the

studied area's boundary so couriers can utilize road segments outside the area of interest. The road network is represented as a directed graph $\mathcal{G}(\mathcal{N}, \mathcal{E})$, where $\mathcal{N}$ and $\mathcal{E}$ represent the sets of 1,163 nodes (road intersections) and 2,450 edges (road segments), respectively.

Road segment features. We collect features related to length, traffic volume, and "bike friendliness" for each edge. The edge lengths $l_{ij} \in \mathbb{R}_+$ are retrieved from the centerline road network. We obtain the daily motor traffic volume on each edge from a well-established and validated transportation model (Travel Modelling Group 2016). We categorize traffic volume into light ($\leq$ 8,000 vehicles per day), medium (between 8,000 and 20,000 vehicles per day), and heavy volume ($\geq$ 20,000 vehicles per day) to be consistent with the route choice model proposed by Zimmermann et al. (2017), which we introduce later. Let $m_{ij} \in \{0,1\}$ and $h_{ij} \in \{0,1\}$ denote if edge $(i,j) \in \mathcal{E}$ has medium or heavy volume, respectively. To measure "bike friendliness", we follow Lin et al. (2021) to classify edges into low-stress (i.e., edges that are perceived as safe for biking for a casual cyclist) and high-stress based on detailed road network data, including road geometry, traffic speed, and the presence of bike infrastructure. Let $s_{ij} \in \{0,1\}$ denote if edge (i,j) is high-stress (1) or not (0).

Perceived travel cost. We consider a set of bike couriers indexed by $t \in [N_{\text{courier}}]$, each having an unobservable travel cost perception $\hat{c}_{ij}^t \in \mathbb{R}_+$ for each edge $(i,j) \in \mathcal{E}$. We generate travel cost perceptions by adapting a route choice model fitted by Zimmermann et al. (2017) using real data. Specifically, we generate $\hat{c}_{ij}^t \in \mathbb{R}_+$ according to

$$\hat{c}_{ij}^t = \left(\hat{\theta}_0^t + \hat{\theta}_1^t l_{ij} + \hat{\theta}_2^t s_{ij} l_{ij} + \hat{\theta}_3^t m_{ij} l_{ij} + \hat{\theta}_4^t h_{ij} l_{ij}\right)^{\delta}. \tag{21}$$

Parameters $\hat{\boldsymbol{\theta}}^t = (\hat{\theta}_0^t, \hat{\theta}_1^t, \ldots, \hat{\theta}_4^t)$ reflect the routing preference of courier t, which are drawn from a multivariate normal distribution whose mean is estimated by Zimmermann et al. (2017) and whose covariance matrix is an identity matrix. We set $\delta = 1$ by default and then vary δ in Section 6.3.2 to investigate the performance under model mis-specification.

Delivery routes. We generate N_{trip} delivery routes for each courier, totaling $N = N_{\text{trip}} \cdot N_{\text{courier}}$ observed routes. Each route $k \in [N]$ has an origin o_k and a destination d_k that are, respectively, sampled from 147 restaurants and 230 residential buildings queried from the study area using OpenStreetMap (2017). We ensure that the origin and destination of each trip are at least one kilometer apart to justify the use of a food delivery service. Once the origin and destination pair is selected, we generate a route $\hat{\mathbf{x}}_k$ by solving a shortest path problem from o_k to d_k based on the courier's perceived travel cost $\hat{\mathbf{c}}^t$. We set $N_{\text{trip}} = 1$ and vary this value in Section 6.3.3 to study the benefit of personalized route recommendations.

6.2. Experimental Setup

Forward problem. Given a dataset of delivery routes $\{o_k, d_k, \hat{\mathbf{x}}_k\}_{k\in[N]}$ and edge features $\{l_{ij}, m_{ij}, h_{ij}, s_{ij}\}_{(i,j)\in\mathcal{E}}$, let $\boldsymbol{F}$ be a feature matrix where each column corresponds to an edge, $\boldsymbol{A}$ be the node-edge incidence matrix of graph $\mathcal{G}$, and $\mathbf{e}_{od}$ be a vector with a 1 in the entry corresponding to origin o and a -1 in the entry corresponding to destination d, with all other entries being 0. Since the travel cost is linear in edge features as (Equation (21)), we formulate the following multi-objective shortest path problem between o and d as **FOP**:

$$\underset{\mathbf{x}\in\mathcal{X}(o,d)}{\text{minimize}} \quad \boldsymbol{\theta}^\intercal \boldsymbol{F}\mathbf{x}, \tag{22}$$

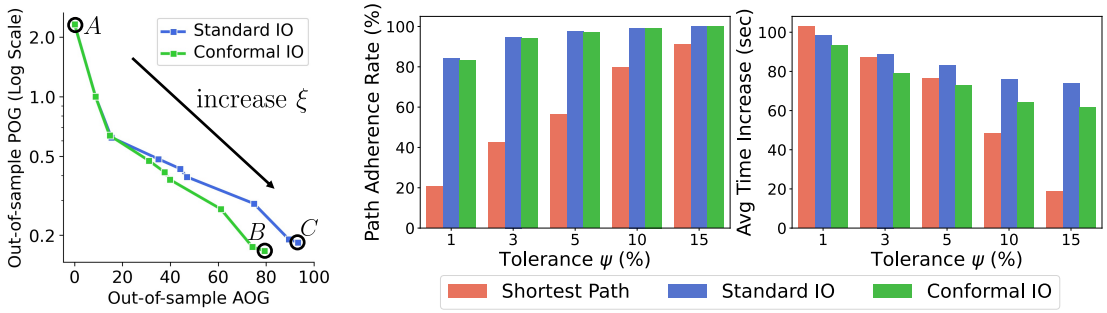
where $\mathcal{X}(o,d) := \left\{\mathbf{x}\in\{0,1\}^{|\mathcal{E}|} \,\middle|\, \boldsymbol{A}\mathbf{x} = \mathbf{e}_{od}\right\}$ represents the set of paths from o to d on graph $\mathcal{G}$. This forward problem has the same structure as defined in Equation (1). So, we can apply the standard and conformal IO methods described previously.

Remark 4. Alternatively, one can formulate the **FOP** as a shortest path problem, where $\boldsymbol{\theta}\in\mathbb{R}_+^{|\mathcal{E}|}$ represents the travel cost on each road segment. However, formulation (22) is more common in the literature (Zattoni Scroccaro et al. 2024, Rönnqvist et al. 2017) because (1) its inverse model is more tractable due to the reduced problem size, (2) the estimated objective function is more interpretable, as it characterizes the relationship between travel cost and road features, and (3) the learned multi-objective weights can be applied across different road networks, whereas the shortest path objective is specific to a single network.

Path prescription. We use the following procedure to generate delivery paths targeting different levels of service quality and path adherence. When using standard IO to prescribe a path from origin o to destination d, we add a constraint $\mathbf{l}^\intercal\mathbf{x} \le l^*_{od}(1+\xi)$ to **FOP**, where $\mathbf{l}$ denotes the vector of edge lengths, l^*_{od} is the length of the shortest path from o to d, and $\xi\in\mathbb{R}_+$ is the maximum percentage increase in travel time that the platform allows. As ξ increases, we expect the generated path to achieve a higher adherence rate (approximated by POG) and lower service quality (measured by AOG), since the model can operate in a larger feasible region, tailoring the recommendation to the courier's preferences. Similarly, we add the same constraint to **RFOP** when using the conformal IO pipeline. We vary ξ from 0% to 20% to provide a spectrum of possible model performance. Note that when $\xi = 0\%$, both pipelines would provide the shortest path as the recommendation.

6.3. Results

6.3.1. Trade-off between path adherence and service quality. As shown in Figure 8a, compared to the shortest path policy (point A), conformal IO (point B) can reduce POG by up to 93% while only increasing AOG by 80 units, corresponding to an 8% increase in travel time. For any level of POG reduction, conformal IO consistently incurs a smaller AOG than standard IO. This performance gap widens as ξ increases.



(a) (b)

Figure 8 Panel (a) shows the AOG and POG achieved by the two pipelines when varying ξ. Point A corresponds to the shortest path policy. Points B and C correspond to setting ξ to 20% for the conformal and standard IO pipelines, respectively. Panel (b) shows the path adherence rates and average delivery times achieved at points A, B, and C, when varying couriers' tolerance for suboptimal recommendations (ψ).

Next, we investigate how the POG reduction translates into improvements in path adherence when fixing $\xi = 20\%$ for both pipelines (points B and C). We assume that couriers will follow the recommended path only if it is within $\psi\%$ of their perceived optimality. We vary ψ in $\{1, 3, 5, 10, 15\}$. As shown in Figure 8b, for every value of ψ, conformal IO results in a smaller travel time increase compared to standard IO, while achieving similar adherence rates. Compared to the shortest path policy, conformal IO achieves significant increases in adherence rates across all values of ψ. However, the impact on travel times depends on ψ. Most notably, when ψ is small ($\leq 5\%$), there is no trade-off: conformal IO improves adherence rates by 40.8 to 62.5 percentage points while simultaneously reducing delivery times. In other words, when couriers have a low tolerance for suboptimal recommendations (i.e., they strongly believe that their estimate of $\boldsymbol{\theta}$ is closer to the truth), a system using conformal IO can produce shorter travel times than the shortest path recommendation. For higher tolerance levels ($\psi \geq 10\%$), there is a trade-off, where conformal IO still improves adherence rates by 8.75 to 19.25 percentage points, while increasing the travel time

by only 1–2 minutes. Overall, these results highlight the potential to significantly improve path adherence without compromising service quality in last-mile delivery.

6.3.2. Performance under model mis-specification. A potential concern of applying IO is model mis-specification, meaning the specified forward problem may not align with the problem that decision makers solve to derive their decisions. To investigate the performance of standard and conformal IO under model mis-specification, we vary the value of δ in $\{1, 2, \ldots, 5\}$ in Equation (21), allowing couriers' perceived travel costs to depend on higher order interactions between different road features. However, when applying the standard and conformal IO pipelines, we assume the underlying **FOP** is as defined in Equation (22), i.e. $\delta = 1$. We repeat the experiment in Section 6.3.1 for each value of δ. As shown in Figure 9, conformal IO consistently provides a better trade-off between path adherence (POG) and service quality (AOG) than standard IO, even when the model is mis-specified ($\delta \geq 2$). Moreover, the performance gap widens as the level of mis-specification δ increases. This suggests that it is beneficial to use conformal IO in practice, as the true forward problem is typically inaccessible, necessitating the use of a "wrong" model.

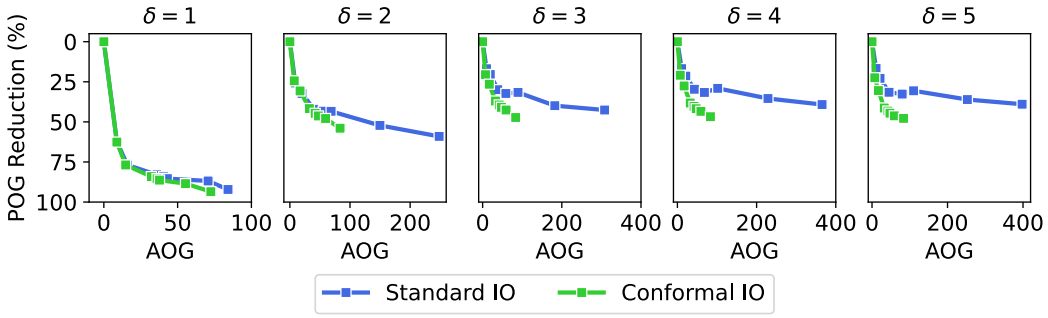


Figure 9 AOG-POG tradeoff under different levels of model mis-specification.

6.3.3. Data pooling vs personalization. So far, our case study has focused on using one route recommendation model for all couriers, which aligns with current industry practice (Liu and Jiang 2022). An alternative approach is to fit an IO model for each courier using their individual historical routes. Next, we compare the performance of this personalization strategy (PIO) against standard and conformal IO, both of which use data from multiple users. We vary N_{trip} in $\{10, 20, 30, 100\}$ to study the effect of sample size. As shown in Figure 10, the conformal IO pipeline consistently offers a better trade-off between

delivery path adherence (POG) and service quality (AOG) than standard IO, regardless of observations for each courier (N_{trips}). In the "small data" regime ($N_{\text{trips}} \leq 10$), conformal IO significantly outperforms PIO. Therefore, when launching services in a new city or onboarding a new courier in an existing one, it is advantageous to use conformal IO to leverage the "shared" knowledge from all couriers. In the "big data" regime, conformal IO performs similarly to PIO. Furthermore, conformal IO incurs much lower computational overhead, as it requires performing hyperparameter tuning and model training for only one model across all couriers, while PIO requires separate processes for each courier.

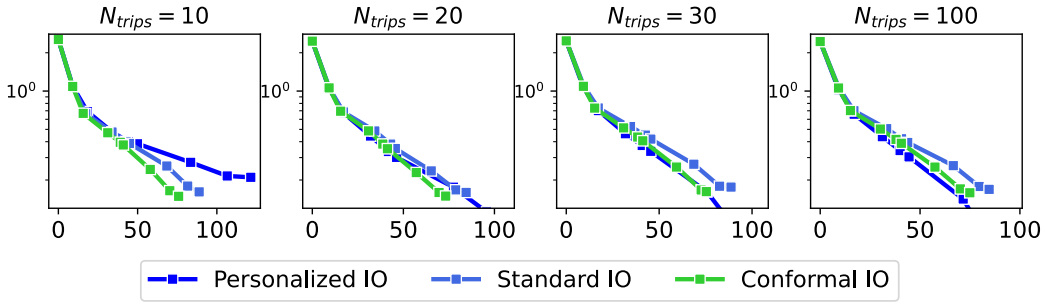


Figure 10 The AOG and POG tradeoff when varying the number of observed routes per courier.

6.3.4. Summary. Our results highlight the potential to improve delivery path adherence without compromising service quality in last-mile delivery. Conformal IO showcases robust performance under model mis-specification, a common challenge in IO applications. It is particularly useful in the "small data" regime, e.g., when the platform is launching services in a new city or onboarding a new courier. In the "big data" regime, conformal IO closely matches the performance of PIO while requiring less computational overhead.

7. Conclusion

In this paper, we propose conformal IO, a novel 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 both the ground-truth parameters and the decision maker's perceived parameters. This suggests that decisions based on conformal IO may be more likely to be adopted compared to decisions from standard IO. We demonstrate strong performance of conformal IO using synthetic problem instances and a real-world case study.

Acknowledgments

The authors are grateful to the anonymous NeurIPS reviewers for their valuable feedback and insightful comments on an earlier version of this paper.

References

Ahuja RK, Orlin JB (2001) Inverse optimization. Operations Research 49(5):771–783.

Arun I, Venkatapathi M (2021) An $O(n)$ algorithm for generating uniform random vectors in n-dimensional cones. arXiv e-prints arXiv–2101.

Aswani A, Shen ZJ, Siddiq A (2018) Inverse optimization with noisy data. Operations Research 66(3):870–892.

Aswani A, Shen ZJM, Siddiq A (2019) Data-driven incentive design in the medicare shared savings program. Operations Research 67(4):1002–1026.

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

Bärmann A, Martin A, Pokutta S, Schneider O (2018) An online-learning approach to inverse optimization. arXiv preprint arXiv:1810.12997 .

Bastani H, Bastani O, Sinchaisri WP (2021) Improving human decision-making with machine learning. arXiv preprint arXiv:2108.08454 5.

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

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

Berthet Q, Blondel M, Teboul O, Cuturi M, Vert JP, Bach F (2020) Learning with differentiable pertubed optimizers. Advances in Neural Information Processing Systems, volume 33, 9508–9519.

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

Bertsimas D, Gupta V, Paschalidis IC (2015) Data-driven estimation in equilibrium using inverse optimization. Mathematical Programming 153:595–633.

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

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

Besbes O, Fonseca Y, Lobel I (2023) Contextual inverse optimization: Offline and online learning. Operations Research .

Birge JR, Chan TCY, Pavlin JM, Zhu IY (2022a) Spatial price integration in commodity markets with capacitated transportation networks. Operations Research 70(3):1739–1761.

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

Bodur M, Chan TC, Zhu IY (2022) Inverse mixed integer optimization: Polyhedral insights and trust region methods. INFORMS Journal on Computing 34(3):1471–1488.

Burton JW, Stein MK, Jensen TB (2020) A systematic review of algorithm aversion in augmented decision making. Journal of Behavioral Decision Making 33(2):220–239.

Chan TCY, Craig T, Lee T, Sharpe MB (2014) Generalized inverse multiobjective optimization with application to cancer therapy. Operations Research 62(3):680–695.

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

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

Chan TCY, Mahmood R, O'Connor DL, Stone D, Unger S, Wong RK, Zhu IY (2023a) Got (optimal) milk? pooling donations in human milk banks with machine learning and optimization. Manufacturing & Service Operations Management 0(0).

Chan TCY, Mahmood R, Zhu IY (2023b) Inverse optimization: Theory and applications. Operations Research 0(0).

Chen L, Ramachandra A, Rujeerapaiboon N, Sim M (2021) Robust data-driven CARA optimization. SSRN Available at http://dx.doi.org/10.2139/ssrn.3842446.

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

Chenreddy AR, Bandi N, Delage E (2022) Data-driven conditional robust optimization. Advances in Neural Information Processing Systems, volume 35, 9525–9537.

China Daily (2021) Driver accused of role in passenger's death appeals sentence. URL https://global.chinadaily.com.cn/a/202109/28/WS6152dc7ea310cdd39bc6c2dc.html, accessed: 2024-08-01.

Ciocan DF, Mišić VV (2022) Interpretable optimal stopping. Management Science 68(3):1616–1638.

City of Toronto (2020) City of Toronto open data. https://www.toronto.ca/city-government/data-research-maps/open-data/, accessed: 2020-09-15.

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

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

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

Dong C, Chen Y, Zeng B (2018) Generalized inverse optimization through online learning. Advances in Neural Information Processing Systems, volume 31.

Dong C, Zeng B (2021) Wasserstein distributionally robust inverse multiobjective optimization. Proceedings of the AAAI Conference on Artificial Intelligence, 5914–5921, number 7 in 35.

DoorDash (2024) Bike delivery strategies. URL https://dasher.doordash.com/en-ca/blog/bike-delivery-strategies, accessed: 2024-06-12.

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

Fu G, Zhang P, Lei D, Qi W, Shen ZJM (2023) Learning for guiding: A framework for unlocking trust and improving performance in last-mile delivery. SSRN Available at: http://dx.doi.org/10.2139/ssrn.4639706.

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

Ge H, Bastani H, Bastani O (2023) Rethinking fairness for human-AI collaboration. arXiv preprint arXiv:2310.03647 .

Global Times (2021) Panicked woman rider jumps out of car, breaking bones; car-hailing platform probed. URL https://www.globaltimes.cn/page/202106/1226697.shtml, accessed: 2024-08-01.

Grand-Clément J, Pauphilet J (2024) The best decisions are not the best advice: Making adherence-aware recommendations. Management Science 0(0).

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

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

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

Kizilcec RF (2016) How much information? effects of transparency on trust in an algorithmic interface. Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, 2390–2395.

Kocuk B (2021) Rational polyhedral outer-approximations of the second-order cone. Discrete Optimization 40:100643.

Lin B, Chan TC, Saxe S (2021) The impact of COVID-19 cycling infrastructure on low-stress cycling accessibility: A case study in the City of Toronto. Findings .

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

Liu S, He L, Max Shen ZJ (2021) On-time last-mile delivery: Order assignment with travel-time predictors. Management Science 67(7):4095–4119.

Liu S, Jiang H (2022) Personalized route recommendation for ride-hailing with deep inverse reinforcement learning and real-time traffic conditions. Transportation Research Part E: Logistics and Transportation Review 164:102780.

Liu S, Siddiq A, Zhang J (2022) Planning bike lanes with data: Ridership, congestion, and path selection. SSRN Available at http://dx.doi.org/10.2139/ssrn.4055703.

Mandi J, Bucarey V, Tchomba MMK, Guns T (2022) Decision-focused learning: through the lens of learning to rank. International Conference on Machine Learning, 14935–14947 (PMLR).

Mohajerin Esfahani P, Shafieezadeh-Abadeh S, Hanasusanto GA, Kuhn D (2018) Data-driven inverse optimization with imperfect information. Mathematical Programming 167:191–234.

Mohri M, Rostamizadeh A, Talwalkar A (2018) Foundations of machine learning (MIT press).

OpenStreetMap (2017) Planet dump retrieved from https://planet.osm.org . https://www.openstreetmap.org.

Phillips R, Şimşek AS, Van Ryzin G (2015) The effectiveness of field price discretion: Empirical evidence from auto lending. Management Science 61(8):1741–1759.

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

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

Shahmoradi Z, Lee T (2022) Quantile inverse optimization: Improving stability in inverse linear programming. Operations Research 70(4):2538–2562.

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

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

Tang B, Khalil EB (2022) Pyepo: A pytorch-based end-to-end predict-then-optimize library for linear and integer programming. arXiv preprint arXiv:2206.14234 .

Travel Modelling Group (2016) GTAModel V4 introduction. https://tmg.utoronto.ca/doc/1.6/gtamodel/index.html, accessed: 2020-11-20.

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

Wainwright MJ (2019) High-dimensional statistics: A non-asymptotic viewpoint, volume 48 (Cambridge University Press).

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

Wilson HJ, Daugherty PR (2018) Collaborative intelligence: Humans and AI are joining forces. Harvard Business Review 96(4):114–123.

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

Yousefi N (2023) Inverse Optimization and its Applications in Measuring Clinical Pathway Concordance. Ph.D. thesis, University of Toronto (Canada).

Zattoni Scroccaro P, van Beek P, Mohajerin Esfahani P, Atasoy B (2024) Inverse optimization for routing problems. Transportation Science 0.

Zimmermann M, Mai T, Frejinger E (2017) Bike route choice modeling using gps data without choice sets of paths. Transportation Research Part C: Emerging Technologies 75:183–196.

Electronic Companion

EC.1. Example Calculation Details

In this section, we present details regarding the example introduced in Section 3.3. We present the performance of the standard inverse optimization and conformal inverse optimization pipelines in EC.1.1 and EC.1.2, respectively.

We use the following notations in this section. Let $N_{\mathbf{a}}, N_{\mathbf{b}}, N_{\mathbf{c}} > 0$ denote the number of times points $\mathbf{a} = (1 - 1/2u, u + 1)$, $\mathbf{b} = (1, u)$, and $\mathbf{c} = (2, 0)$ appear in the dataset $\mathcal{D}$, respectively. Let $\delta_u \in (0, \pi/2)$ satisfy $\cos \delta_u = u/\sqrt{u^2+1}$ and δ'_u satisfy $\cos \delta'_u = 2u/\sqrt{4u^2+1}$. In other words, vectors $(\cos \delta_u, \sin \delta_u)$ and $(\cos \delta'_u, \sin \delta'_u)$ are orthogonal to Constraints (8b) and (8c), respectively.

EC.1.1. Performance of the Standard IO Pipeline

Step I: We first characterize the dataset $\mathcal{D}$. Since human perceptions $\hat{\boldsymbol{\theta}}_k$ are uniformly drawn at random from $\boldsymbol{\Theta}\{(\cos\delta, \sin\delta) \,|\, \delta \in [0, \pi/2]\}$ for any $k \in [N]$, we have

$$\hat{\mathbf{x}}_k = \begin{cases} \mathbf{a}, & \text{w.p. } 2\delta'_u/\pi \\ \mathbf{b}, & \text{w.p. } 2(\delta_u - \delta'_u)/\pi \\ \mathbf{c}, & \text{w.p. } (\pi - 2\delta_u)/\pi. \end{cases} \tag{EC.1}$$

where $\hat{\mathbf{x}}_k = \tilde{\mathbf{x}}\left(\hat{\boldsymbol{\theta}}_k, u\right)$. Given that $0 < \delta'_u < \delta_u < \pi/4$, it is easy to verify that, as N goes to infinity, we have $N_{\mathbf{a}}, N_{\mathbf{b}}, N_{\mathbf{c}} > 0$ and $N_{\mathbf{c}} > N_{\mathbf{a}}$ almost surely.

Step II: Next, we characterize the point estimate. We can write $\mathbf{IOP}(\mathcal{D})$ with the sub-optimality loss as

$$\delta^* = \operatorname*{arg\,min}_{\delta \in [0, \pi/2]} \ \ell(\delta),$$

where

$$\ell(\delta) = N_{\mathbf{a}}\ell_{\mathbf{a}}(\delta) + N_{\mathbf{b}}\ell_{\mathbf{b}}(\delta) + N_{\mathbf{c}}\ell_{\mathbf{c}}(\delta),$$

$$\ell_{\mathbf{a}}(\delta) = \begin{cases} 0, & \text{if } \delta \in [0, \delta'_u) \\ -\frac{1}{2u}\cos\delta + \sin\delta, & \text{if } \delta \in [\delta'_u, \delta_u) \\ -(\frac{1}{2u} + 1)\cos\delta + (u+1)\sin\delta, & \text{if } \delta \in [\delta_u, \pi/2] \end{cases},$$

$$\ell_{\mathbf{b}}(\delta) = \begin{cases} \frac{1}{2u}\cos\delta - \sin\delta, & \text{if } \delta \in [0, \delta'_u) \\ 0, & \text{if } \delta \in [\delta'_u, \delta_u) \\ -\cos\delta + u\sin\delta, & \text{if } \delta \in [\delta_u, \pi/2] \end{cases},$$

and

$$\ell_{\mathbf{c}}(\delta) = \begin{cases} (\frac{1}{2u}+1)\cos\delta - (u+1)\sin\delta, & \text{if } \delta \in [0, \delta'_u) \\ \cos\delta - u\sin\delta, & \text{if } \delta \in [\delta'_u, \delta_u) \\ 0, & \text{if } \delta \in [\delta_u, \pi/2] \end{cases}.$$

When $\delta \in (\delta_u, \pi/2]$, $\ell_{\mathbf{c}}$ is a constant and $\ell_{\mathbf{a}}$ and $\ell_{\mathbf{b}}$ are strictly increasing δ. Thus, ℓ is also strictly increasing in δ. The optimal solution to **IOP** $\delta^* \notin (\delta_u, \pi/2]$. Similarly, when $\delta \in [0, \delta'_u]$, ℓ is strictly decreasing in δ, so $\delta^* \notin [0, \delta'_u)$. When $\delta \in [\delta'_u, \delta_u]$, we have

$$\ell(\delta) = \left(N_{\mathbf{c}} - \frac{1}{2u}N_{\mathbf{a}}\right)\cos\delta + (N_{\mathbf{a}} - uN_{\mathbf{c}})\sin\delta$$

which is strictly decreasing in δ. Therefore, we have $\bar{\boldsymbol{\theta}} = (\cos\delta_u, \sin\delta_u)$.

Step III: Finally, we calculate the AOG and POG of $\tilde{\mathbf{x}}_{IO}$. Since $\bar{\boldsymbol{\theta}} = (\cos\delta_u, \sin\delta_u)$, the standard inverse optimization policy will sample uniformly from $\{\mathbf{b}, \mathbf{c}\}$ as a recommendation, assuming that the solver only returns extreme points. We note that the following results still hold even if the solver sample from "all" optimal solutions, i.e., the facet associated with Constraint (8b). Given this policy, we have

$$\mathrm{AOG}(\mathbf{x}_{\mathrm{IO}}) = \frac{1}{2}\left[\frac{\sqrt{2}}{2}(1+u) - \sqrt{2}\right] = \frac{\sqrt{2}(u-1)}{4},$$

and

$$\begin{aligned} \mathrm{POG}(\mathbf{x}_{\mathrm{IO}}) = & \frac{1}{2}\int_0^{\delta'_u}\frac{2}{\pi}\left[\frac{1}{2u}\cos\delta - \sin\delta\right]d\delta + \frac{1}{2}\int_{\delta_u}^{\pi/2}\frac{2}{\pi}\left[-\cos\delta + u\sin\delta\right]d\delta \\ & \frac{1}{2}\int_0^{\delta'_u}\frac{2}{\pi}\left[(1+\frac{1}{2u})\cos\delta - (u+1)\sin\delta\right]d\delta + \frac{1}{2}\int_{\delta'_u}^{\delta_u}\frac{2}{\pi}\left[\cos\delta - u\sin\delta\right]d\delta \\ = & \frac{2}{\pi}\left[\sqrt{u^2+1} + \frac{4u^2+u+1}{2u\sqrt{4u^2+1}} - \frac{1}{2}\right] \\ > & \frac{2\sqrt{u^2+1}-1}{\pi} \end{aligned}$$

Both $\mathrm{AOG}(\mathbf{x}_{\mathrm{IO}})$ and $\mathrm{POG}(\mathbf{x}_{\mathrm{IO}})$ are strictly increasing in u. So, they can be arbitrarily large simultaneously as u increases.

EC.1.2. Performance of the Conformal IO Pipeline

Step I: We first derive the optimal solution to **RFOP**. Let

$$R(\mathbf{x}) := \max_{\boldsymbol{\theta}\in\mathcal{C}(\boldsymbol{\theta}_u,\alpha)} \theta_1 x_1 + \theta_2 x_2 \tag{EC.2}$$

denote the worst-case cost of $\mathbf{x}\in\mathcal{X}(u)$, i.e. the optimal value of the inner maximization problem in **RFOP**. Next we calculate the worst-case cost for the three candidate solutions $\mathbf{a}$, $\mathbf{b}$, and $\mathbf{c}$.

We have

$$R(\mathbf{a}) = \begin{cases} \left(1-\frac{1}{2u}\right)\cos(\delta_u+\alpha) + (u+1)\sin(\delta_u+\alpha), & \text{if } \alpha\in(0,\pi/2-\delta_u] \\ u+1, & \text{if } \alpha\in[\pi/2-\delta_u,\pi/2]. \end{cases},$$

$$R(\mathbf{b}) = \begin{cases} \cos(\delta_u+\alpha) + u\sin(\delta_u+\alpha)., & \text{if } \alpha\in(0,\pi/2-\delta_u] \\ u, & \text{if } \alpha\in[\pi/2-\delta_u,\pi/2] \end{cases},$$

and

$$R(\mathbf{c}) = \begin{cases} 2\cos(\delta_u-\alpha), & \text{if } \alpha\in(0,\delta_u] \\ 2, & \text{if } \alpha\in[\delta_u,\pi/2]. \end{cases}$$

We first compare $\mathbf{a}$ *and* $\mathbf{b}$. When $\alpha\in(0,\pi/2-\delta_u]$, we have

$$R(\mathbf{a}) - R(\mathbf{b}) = -\frac{1}{2u}\cos(\delta_u+\alpha) + \sin(\delta_u+\alpha) = \frac{1}{\sqrt{u^2+1}}\left[\frac{1}{2}\cos\delta_u + \left(u-\frac{1}{2u}\right)\sin\delta_u\right] > 0.$$

When $\alpha\in(\pi/2-\delta_u,\pi/2]$, we have $R(\mathbf{a}) = u+1 > u = R(\mathbf{b})$. So, $\mathbf{a}$ can not be an optimal solution to **RFOP**.

We next compare $\mathbf{b}$ *and* $\mathbf{c}$. When $\alpha\in(0,\delta_u)$, we have

$$R(\mathbf{b}) - R(\mathbf{c}) = \frac{u^2-3}{\sqrt{u^2+1}}\sin\alpha > 0.$$

The inequality holds because $u>0$.

When $\alpha\in[\delta_u,\pi/2-\delta_u)$, we know that $R(\mathbf{b})$ first increases then decreases in α because

$$\frac{\partial R(\mathbf{b})}{\partial\alpha} = -\sin(\delta_u+\alpha) + u\cos(\delta_u+\alpha) = \sqrt{u^2+1}\cos(\alpha+2\delta_u)$$

So,

$$R(\mathbf{b}) \geq \max\{\cos 2\delta_u + u \sin 2\delta_u, u\} = \max\left\{\frac{3u^2-1}{u^2+1}, u\right\} > 2 = R(\mathbf{c}).$$

When $\alpha \in [\pi/2 - \delta_u, \pi/2]$, we have $R(\mathbf{b}) = u > 2 = R(\mathbf{c})$.

Therefore, we know that $\mathbf{c}$ is an unique optimal solution to **RFOP**.

Step II: Next, we calculate the AOG and POG. Given the calculations in Step II, conformal inverse optimization pipeline always returns $\mathbf{c}$ as the recommendation. Given this policy, we have,

$$\mathrm{AOG}(\bar{\mathbf{x}}_{\mathrm{CIO}}) = 0,$$

and

$$\begin{aligned}
\mathrm{POG}(\bar{\mathbf{x}}_{\mathrm{CIO}}) =& \int_0^{\delta_u'} \frac{2}{\pi}\left[\left(1+\frac{1}{2u}\right)\cos\delta - (u+1)\sin\delta\right] d\delta + \int_{\delta_u'}^{\delta_u} \frac{2}{\pi}\left[\cos\delta - u\sin\delta\right] d\delta \\
=& \frac{\pi}{2}\left[\sqrt{u^2+1} - (u+1) + \frac{2u}{\sqrt{4u^2+1}} + \frac{1}{2u\sqrt{4u^2+1}}\right] \\
<& \frac{2\sqrt{5} + \sqrt{17}/2 - 6}{\pi}.
\end{aligned}$$

The inequality holds because the function in the second line is a decreasing function in $u > 2$, and we obtain the final line by plugging in $u = 2$.

Therefore, we conclude that the conformal IO pipeline can achive upper-bounded AOG and POG simultaneously. These bounds are independent of u.

EC.1.3. Numerical Values of Solution Quality Bounds

In Figure EC.1, we present the numerical values of our bounds from Theorem 3 for the example introduced in Section 3.3. We observe that our bounds closely follow the conformal inverse optimization pipeline, which outperforms the standard inverse optimization pipeline by a large margin, especially when u is set to a large value.

EC.2. Proof of Statements in Section 4

In this section, we present proofs that are omitted in Section 4. We first present useful definitions and lemmas from the literature in Sections EC.2.1 and EC.2.2, respectively. We then present the proofs of our theoretical results in subsequent sections.

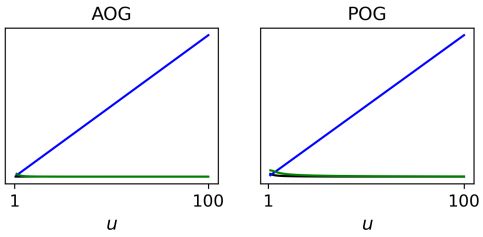


Figure EC.1 Illustration of the AOG (left) and POG (right) bounds for example intrudced in Section 3.3. The black lines represent the AOG and POG achieved by $\bar{\mathbf{x}}_{\text{CIO}}$ when varying u from 2 to 50. The green lines are the bounds derived in Theorem 3, respectively. The blue lines are the AOG and POG achieved by $\bar{\mathbf{x}}_{\text{IO}}$.

EC.2.1. Definitions

Definition EC.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

$$\hat{\mathfrak{R}}_{\mathcal{D}}(\mathcal{F}) := \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{f\in\mathcal{F}} \frac{1}{N}\sum_{i\in[N]} \sigma_i f(Z_i)\right] \tag{EC.3}$$

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 EC.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 $\mathbb{P}$:

$$\mathfrak{R}_N(\mathcal{F}) := \mathbb{E}_{\mathcal{D}\sim\mathbb{P}^N}\left[\hat{\mathfrak{R}}_{\mathcal{D}}(\mathcal{F})\right] \tag{EC.4}$$

Definition EC.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

$$\Pi_{\mathcal{H}}(N) := \max_{(Z_1, Z_2, \ldots, Z_N)\in\mathcal{Z}^N} |\{(h(Z_1), h(Z_2), \ldots, h(Z_N)) \,|\, h \in \mathcal{H}\}| \tag{EC.5}$$

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

EC.2.2. Useful Lemmas

LEMMA EC.1 **(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*

$$\mathfrak{R}_N(\mathcal{H}) \leq \sqrt{\frac{2\log \Pi_{\mathcal{H}}(N)}{N}}. \tag{EC.6}$$

LEMMA EC.2 **(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*

$$\sup_{f\in\mathcal{F}} \left| \frac{1}{N}\sum_{i\in[N]} f(X_i) - \mathbb{E}\left[f(X_i)\right] \right| \leq 2\mathfrak{R}_N(\mathcal{F}) + \delta \tag{EC.7}$$

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

EC.2.3. Proof of Theorem 1

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

We first present the extensive formulation of Problem (11). When $\alpha \in [0,\pi]$, $\cos\alpha$ is strictly decreasing in α. Therefore, minimizing α is equivalent to maximizing $\cos\alpha$. We can replace the decision variable α in Problem (11) 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 binary decision variables $y_k \in \{0,1\}$ that indicate if $\hat{\mathbf{\Theta}}_k$ intersects with the learned uncertainty set $(=1)$ or not $(=0)$ for any $k \in \mathcal{K}_{\mathrm{val}}$. Problem (11) can be reformulated as follows.

$$\begin{aligned}
\underset{c,\{\boldsymbol{\theta}_k\}_{k\in\mathcal{K}_{\mathrm{val}}},\{y_k\}_{k\in\mathcal{K}_{\mathrm{val}}}}{\text{maximize}} \quad & c && \text{(EC.8a)}\\
\text{subject to} \quad & \hat{\mathbf{x}}_k \in \mathcal{X}^{\mathrm{OPT}}(\boldsymbol{\theta}_k, \mathbf{u}_k), \quad \forall k \in \mathcal{K}_{\mathrm{val}} && \text{(EC.8b)}\\
& \boldsymbol{\theta}_k^\intercal \bar{\boldsymbol{\theta}} \geq c + 2(y_k - 1), \quad \forall k \in \mathcal{K}_{\mathrm{val}} && \text{(EC.8c)}\\
& \sum_{k\in\mathcal{K}_{\mathrm{val}}} y_k \geq \lceil \gamma(N_{\mathrm{val}}+1) \rceil && \text{(EC.8d)}\\
& \|\boldsymbol{\theta}_k\|_2 = 1, \quad \forall k \in \mathcal{K}_{\mathrm{val}} && \text{(EC.8e)}\\
& -1 \leq c \leq 1 && \text{(EC.8f)}\\
& y_k \in \{0,1\}, \quad \forall k \in \mathcal{K}_{\mathrm{val}}. && \text{(EC.8g)}
\end{aligned}$$

Constraints (EC.8b) ensure that $\boldsymbol{\theta}_k$ is a member of $\hat{\mathbf{\Theta}}_k$ for any $k \in \mathcal{K}_{\mathrm{val}}$. Constraints (EC.8c) decide if $\boldsymbol{\theta}_k$ should be taken into account when calculating the maximal cosine value c

based on if $\hat{\mathbf{\Theta}}_k$ intersects with $\mathcal{C}$. Constraint (EC.8d) ensures that $\mathcal{C}$ intersects with at least $\lceil \gamma(N_{\text{val}}+1) \rceil$ inverse feasible sets. Constraint (EC.8e) enforces $\boldsymbol{\theta}_k$ to be on the unit sphere as defined in Equation (10). Constraints (EC.8f)–(EC.8g) specify the ranges of the decision variables.

Observing that the objective of Problem (EC.8) is to maximize c and that decision variables $\boldsymbol{\theta}_k$ of different data points only interact in Constraints (EC.8c). We can re-write Problem (EC.8) as

$$\begin{aligned}
\text{maximize} \quad & c && \text{(EC.9a)}\\
\text{subject to} \quad & c \le c_k - 2(y_k - 1), \quad \forall k \in \mathcal{K}_{\text{val}} && \text{(EC.9b)}\\
& \sum_{k \in \mathcal{K}_{\text{val}}} y_k \ge \lceil \gamma(N_{\text{val}}+1) \rceil && \text{(EC.9c)}\\
& -1 \le c \le 1 && \text{(EC.9d)}\\
& y_k \in \{0,1\}, \quad \forall k \in \mathcal{K}_{\text{val}}, && \text{(EC.9e)}
\end{aligned}$$

where

$$\begin{aligned}
c_k := \underset{\boldsymbol{\theta}_k}{\text{maximize}} \quad & \boldsymbol{\theta}_k^\intercal \bar{\boldsymbol{\theta}} && \text{(EC.10a)}\\
\text{subject to} \quad & \hat{\mathbf{x}}_k \in \mathcal{X}^{\text{OPT}}(\boldsymbol{\theta}_k, \mathbf{u}_k) && \text{(EC.10b)}\\
& \|\boldsymbol{\theta}_k\|_2 \le 1. && \text{(EC.10c)}
\end{aligned}$$

Note that we replace Constraints (EC.8e) with Constraints (EC.10c) because the objective of Problem (EC.10) 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 (EC.9a) is to set $y_k = 1$ for all k such that $c_k \ge \Gamma_\tau\left(\{c_k\}_{k \in \mathcal{K}_{\text{val}}}\right)$ and $y_k = 0$ otherwise. Therefore, the optimal objective value of Problem (EC.9a) is $c = \Gamma_\tau\left(\{c_k\}_{k \in \mathcal{K}_{\text{val}}}\right)$ corresponding to $\alpha_\gamma = \arccos \Gamma_\tau\left(\{c_k\}_{k \in \mathcal{K}_{\text{val}}}\right)$. □

EC.2.4. Proof of Theorem 2

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

$$A_{\bar{\boldsymbol{\theta}}}(\hat{\mathbf{x}}, \mathbf{u}) := \underset{\boldsymbol{\theta}}{\text{maximize}} \quad \boldsymbol{\theta}^\intercal \bar{\boldsymbol{\theta}} \tag{EC.11a}$$

$$\text{subject to} \quad \boldsymbol{\theta} \in \boldsymbol{\Theta}^{\text{OPT}}(\hat{\mathbf{x}}, \mathbf{u}) \tag{EC.11b}$$

$$\|\boldsymbol{\theta}\|_2 \leq 1. \tag{EC.11c}$$

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 1. 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

$$\mathbb{P}\left(\boldsymbol{\Theta}^{\text{OPT}}(\hat{\mathbf{x}}, \mathbf{u}) \cap \mathcal{C}(\bar{\boldsymbol{\theta}}, \alpha) \neq \varnothing\right) = \mathbb{P}\left(A_{\bar{\boldsymbol{\theta}}}(\hat{\mathbf{x}}, \mathbf{u}) \geq \Gamma^\tau(\mathcal{A})\right). \tag{EC.12}$$

Assumption 1 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 and 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 (EC.12). 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 $z_k := (\mathbf{u}_k, \hat{\mathbf{x}}_k)$ and $\mathcal{Z} := \{z_k\}_{k\in\mathcal{K}_{\text{val}}}$. We define a function class

$$\mathcal{H} = \left\{ h(z,\alpha) = \mathbb{1}\left[\boldsymbol{\Theta}^{\text{OPT}}(\hat{\mathbf{x}},\mathbf{u}) \cap \mathcal{C}(\bar{\boldsymbol{\theta}},\alpha)\right] \,\middle|\, \alpha \in (0,\pi) \right\}. \tag{EC.13}$$

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

$$\Pi_{\mathcal{H}}(N_{\text{val}}) = N_{\text{val}} + 1 \tag{EC.14}$$

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 EC.1, we have

$$\mathfrak{R}_{N_{\text{val}}}(\mathcal{H}) \le \sqrt{\frac{2\log(N_{\text{val}}+1)}{N_{\text{val}}}} \tag{EC.15}$$

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 EC.2.

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

$$\frac{1}{N_{\text{val}}} \sum_{k\in\mathcal{K}_{\text{val}}} h(z_k,\alpha) = \frac{1}{N_{\text{val}}} \sum_{k\in\mathcal{K}_{\text{val}}} \mathbb{1}\left[\boldsymbol{\Theta}^{\text{OPT}}(\hat{\mathbf{x}}_k,\mathbf{u}_k) \cap \mathcal{C}(\bar{\boldsymbol{\theta}},\alpha)\right] = \frac{\lceil \gamma(N_{\text{val}}+1)\rceil}{N_{\text{val}}}, \tag{EC.16}$$

so we have

$$\gamma \le \frac{1}{N_{\text{val}}} \sum_{k\in\mathcal{K}_{\text{val}}} h(z_k,\alpha) \le \gamma + \frac{2}{N_{\text{val}}}. \tag{EC.17}$$

The second inequality holds because

$$\frac{\lceil \gamma(N_{\text{val}}+1)\rceil}{N_{\text{val}}} = \frac{\lfloor \gamma N_{\text{val}}\rfloor + \lceil \gamma N_{\text{val}} - \lfloor \gamma N_{\text{val}}\rfloor + \gamma\rceil}{N_{\text{val}}} \le \frac{\gamma N_{\text{val}} + \lceil \gamma N_{\text{val}} - \lfloor \gamma N_{\text{val}}\rfloor + \gamma\rceil}{N_{\text{val}}} \le \gamma + \frac{2}{N_{\text{val}}}. \tag{EC.18}$$

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 EC.2, we have, with probability at least $\delta = 1 - 1/N_{\text{val}}$,

$$\left| \frac{1}{N_{\text{val}}} \sum_{k\in\mathcal{K}_{\text{val}}} h(z_k,\alpha) - \mathbb{E}\left[h(z,\alpha)\right] \right| \le 2\mathfrak{R}_{N_{\text{val}}}(\mathcal{H}) + \sqrt{\frac{2\log N_{\text{val}}}{N_{\text{val}}}}. \tag{EC.19}$$

By combing (EC.15)–(EC.19), we have, with probability at least $1 - 1/N_{\text{val}}$,

$$\left|\mathbb{P}\left(\boldsymbol{\Theta}^{\text{OPT}}(\hat{\mathbf{x}},\mathbf{u}) \cap \mathcal{C}(\bar{\boldsymbol{\theta}},\alpha)\right) - \gamma\right| \le \sqrt{\frac{8\log(N_{\text{val}}+1) + 2\log N_{\text{val}}}{N_{\text{val}}}} + \frac{2}{N_{\text{val}}}. \tag{EC.20}$$

□

EC.2.5. Proof of Lemma 1

Proof. For any fixed $\mathbf{x}$, we have

$$\begin{aligned} f(\boldsymbol{\theta}, \hat{\mathbf{x}}) - f(\boldsymbol{\theta}', \hat{\mathbf{x}}) &= \sum_{i\in[d]} (\theta_i - \theta_i') f_i(\hat{\mathbf{x}}) \\ &\leq \sqrt{\sum_{i\in[d]} f_i^2(\hat{\mathbf{x}})} \sqrt{\sum_{i\in[d]} (\theta_i - \theta_i')^2} \\ &= \nu(\hat{\mathbf{x}}) \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_2 \end{aligned}$$

where $\nu(\hat{\mathbf{x}}) := \sqrt{\sum_{i\in[d]} f_i^2(\hat{\mathbf{x}})}$. The inequality follows the Cauchy-Schwartz inequality. □

EC.2.6. Proof of Theorem 3

Proof. *To derive the POG bound for $\bar{\mathbf{x}}_{CIO}$*, we first bound the perceived optimality gap of a sampled DM. Let $(\hat{\boldsymbol{\theta}}, \mathbf{u})$ be a sample from $\mathbb{P}_{(\boldsymbol{\theta},\mathbf{u})}$, $\hat{\mathbf{x}} = \tilde{\mathbf{x}}(\hat{\boldsymbol{\theta}}, \mathbf{u})$, $\hat{\boldsymbol{\theta}}_{\mathrm{CIO}}(\mathbf{u})$ 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 $\hat{\mathbf{x}}$, $\bar{\boldsymbol{\theta}}_{\mathrm{CIO}}(\mathbf{u})$ 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 $\bar{\mathbf{x}}_{\mathrm{CIO}}(\mathbf{u})$, If $\boldsymbol{\Theta}^{\mathrm{OPT}}(\hat{\mathbf{x}}, \mathbf{u}) \cap \mathcal{C}(\bar{\boldsymbol{\theta}}, \alpha_1) \neq \varnothing$, let $\tilde{\boldsymbol{\theta}}$ be an element of $\boldsymbol{\Theta}^{\mathrm{OPT}}(\hat{\mathbf{x}}, \mathbf{u}) \cap \mathcal{C}(\bar{\boldsymbol{\theta}}, \alpha_1)$, we have

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

The first line holds due to Lemma 1. The second line holds due to assumption 3. The third line holds due to the definition of $\bar{\boldsymbol{\theta}}_{\mathrm{CIO}}(u)$. The fourth line holds because $\left(\bar{\mathbf{x}}_{\mathrm{CIO}}(\mathbf{u}), \bar{\boldsymbol{\theta}}_{\mathrm{CIO}}(\mathbf{u})\right)$ is an optimal solution to $\mathbf{RFO}\left(\mathcal{C}(\bar{\boldsymbol{\theta}}, \alpha_1), \mathbf{u}\right)$. The fifth line holds due to Lemma 1. The sixth line holds because both $\hat{\boldsymbol{\theta}}_{\mathrm{CIO}}(\mathbf{u})$ 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 $\hat{\boldsymbol{\theta}}_{\mathrm{CIO}}(\mathbf{u})$ 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 $\mathbb{P}\left(\boldsymbol{\Theta}^{\mathrm{OPT}}(\hat{\mathbf{x}},\mathbf{u})\cap\mathcal{C}(\bar{\boldsymbol{\theta}},\alpha_1)\right)=1$, we have

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

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

To derive the AOG bound for $\bar{\mathbf{x}}_{CIO}$, we first derive an upper bound on the optimality gap of the suggested decision $\bar{\mathbf{x}}_{\mathrm{CIO}}(\mathbf{u})$ 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 $\boldsymbol{\Theta}^{\mathrm{OPT}}(\hat{\mathbf{x}},\mathbf{u})\cap\mathcal{C}(\bar{\boldsymbol{\theta}},\alpha_1)$, which is non-empty almost surely because α_1 is chosen such that $\mathbb{P}\left(\boldsymbol{\Theta}^{\mathrm{OPT}}(\hat{\mathbf{x}},\mathbf{u})\cap\mathcal{C}(\bar{\boldsymbol{\theta}},\alpha_1)\right)=1$. Let $\bar{\boldsymbol{\theta}}_{\mathrm{CIO}}(\mathbf{u})$ 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 $\bar{\mathbf{x}}_{\mathrm{CIO}}(\mathbf{u})$. For any $\mathbf{u}\in\mathcal{U}$, let $\mathbf{x}^*(\mathbf{u}):=\tilde{\mathbf{x}}(\boldsymbol{\theta}^*,\mathbf{u})$ and $\boldsymbol{\theta}^*_{\mathrm{CIO}}(\mathbf{u})$ 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}}_{\mathrm{CIO}}(\mathbf{u})\right)-f\left(\boldsymbol{\theta}^*,\mathbf{x}^*(\mathbf{u})\right)\\
\le\ & f\left(\mathbb{E}(\hat{\boldsymbol{\theta}}),\bar{\mathbf{x}}_{\mathrm{CIO}}(\mathbf{u})\right)-f\left(\mathbb{E}(\hat{\boldsymbol{\theta}}),\mathbf{x}^*(\mathbf{u})\right)+\left(\nu\left[\bar{\mathbf{x}}_{\mathrm{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}}_{\mathrm{CIO}}(\mathbf{u})\right)-f\left(\mathbb{E}(\hat{\boldsymbol{\theta}}),\mathbf{x}^*(\mathbf{u})\right)+\sigma\left(\nu\left[\bar{\mathbf{x}}_{\mathrm{CIO}}(\mathbf{u})\right]+\nu\left[\mathbf{x}^*(\mathbf{u})\right]\right)\\
=\ & \mathbb{E}\left[f\left(\hat{\boldsymbol{\theta}},\bar{\mathbf{x}}_{\mathrm{CIO}}(\mathbf{u})\right)-f\left(\hat{\boldsymbol{\theta}},\mathbf{x}^*(\mathbf{u})\right)\right]+\sigma\left(\nu\left[\bar{\mathbf{x}}_{\mathrm{CIO}}(\mathbf{u})\right]+\nu\left(\mathbf{x}^*(\mathbf{u})\right)\right)\\
\le\ & \mathbb{E}\left[f\left(\tilde{\boldsymbol{\theta}},\bar{\mathbf{x}}_{\mathrm{CIO}}(\mathbf{u})\right)-f\left(\tilde{\boldsymbol{\theta}},\mathbf{x}^*(\mathbf{u})\right)+\left(\nu\left[\bar{\mathbf{x}}_{\mathrm{CIO}}(\mathbf{u})\right]+\nu\left[\hat{\mathbf{x}}\right]\right)\|\hat{\boldsymbol{\theta}}-\tilde{\boldsymbol{\theta}}\|_2\right]\\
&+\sigma\left(\nu\left[\bar{\mathbf{x}}_{\mathrm{CIO}}(\mathbf{u})\right]+\nu\left(\mathbf{x}^*(\mathbf{u})\right)\right)\\
\le\ & \mathbb{E}\left[f\left(\tilde{\boldsymbol{\theta}},\bar{\mathbf{x}}_{\mathrm{CIO}}(\mathbf{u})\right)-f\left(\tilde{\boldsymbol{\theta}},\mathbf{x}^*(\mathbf{u})\right)+\left(\nu\left[\bar{\mathbf{x}}_{\mathrm{CIO}}(\mathbf{u})\right]+\nu\left[\hat{\mathbf{x}}\right]\right)\eta\right]\\
&+\sigma\left(\nu\left[\bar{\mathbf{x}}_{\mathrm{CIO}}(\mathbf{u})\right]+\nu\left(\mathbf{x}^*(\mathbf{u})\right)\right)\\
\le\ & \mathbb{E}\left[f\left(\bar{\boldsymbol{\theta}}_{\mathrm{CIO}}(\mathbf{u}),\bar{\mathbf{x}}_{\mathrm{CIO}}(\mathbf{u})\right)-f\left(\tilde{\boldsymbol{\theta}},\mathbf{x}^*(\mathbf{u})\right)\right]+(\eta+\sigma)\left(\nu\left[\bar{\mathbf{x}}_{\mathrm{CIO}}(\mathbf{u})\right]+\nu\left(\mathbf{x}^*(\mathbf{u})\right)\right)\\
\le\ & \mathbb{E}\left[f\left(\boldsymbol{\theta}^*_{\mathrm{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}}_{\mathrm{CIO}}(\mathbf{u})\right]+\nu\left(\mathbf{x}^*(\mathbf{u})\right)\right)\\
\le\ & \mathbb{E}\left[\nu(\mathbf{x}^*(\mathbf{u}))\|\boldsymbol{\theta}^*_{\mathrm{CIO}}(\mathbf{u})-\tilde{\boldsymbol{\theta}}\|_2\right]+(\eta+\sigma)\left(\nu\left[\bar{\mathbf{x}}_{\mathrm{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}}_{\mathrm{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}}_{\mathrm{CIO}}(\mathbf{u})\right]
\end{aligned}$$

The first line holds because of Lemma 1. The second line holds due to Assumption 2. 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 Lemma 1. The fifth line holds due to Assumption 3. 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 Lemma 1. 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 $\mu^* := \mathbb{E}\left(\nu[\mathbf{x}^*(\mathbf{u})]\right)$. $\square$

EC.2.7. Proof of Proposition 1

Proof. We first show that the constraint $\|\boldsymbol{\theta}\|_2 = 1$ can be replaced with $\|\boldsymbol{\theta}\|_2 \le 1$ in $h(\mathbf{x})$. For any $\mathbf{x} \in \mathcal{X}(\mathbf{u})$, we show that there exists some optimal solution $\boldsymbol{\theta}^*$ to the maximization problem after the substitution that satisfies (1) $\boldsymbol{\theta}^* \in \mathbb{R}^d_+$ and (2) $\|\boldsymbol{\theta}^*\|_2 = 1$. So, it is also optimal to the original problem with the constraint $\|\boldsymbol{\theta}\|_2 = 1$.

(1) *We show that there exists an optimal $\boldsymbol{\theta}^* \in \mathbb{R}^d_+$ by construction.* Let $\boldsymbol{\theta}' = (\theta'_1, \theta'_2, \ldots, \theta'_d)$ denote an optimal solution to the maximization problem after the substitution. We assume there exists some $j \in [d]$ such that $\theta'_j < 0$. So, $\boldsymbol{\theta}' \notin \mathbb{R}^d_+$. We construct $\boldsymbol{\theta}^* = (|\theta'_1|, |\theta'_2|, \ldots, |\theta'_d|) \in \mathbb{R}^d_+$. Since $\bar{\boldsymbol{\theta}} \in \mathbb{R}^d_+$, it is easy to verify that $\bar{\boldsymbol{\theta}}^\intercal \boldsymbol{\theta}^* \ge \bar{\boldsymbol{\theta}}^\intercal \boldsymbol{\theta}' \ge \cos\alpha_\gamma$. Moreover, $\|\boldsymbol{\theta}^*\|_2 = \|\boldsymbol{\theta}'\|_2 \le 1$, meaning that $\boldsymbol{\theta}^*$ is a feasible solution to the maximization problem. Additionally, since $\mathbf{f}(\mathbf{x}) \in \mathbb{R}^d_+$, we have $\boldsymbol{\theta}^{*\intercal}\mathbf{f}(\mathbf{x}) \ge \boldsymbol{\theta}'^\intercal \mathbf{f}(\mathbf{x})$. So, $\boldsymbol{\theta}^*$ is also an optimal solution.

(2) *We show that there exists an optimal $\boldsymbol{\theta}^*$ such that $\|\boldsymbol{\theta}^*\|_2 = 1$ by construction.* Let $\boldsymbol{\theta}''$ be an optimal solution to the maximization problem after the substitution. We construct $\boldsymbol{\theta}^* = \boldsymbol{\theta}''/\|\boldsymbol{\theta}''\|_2$. So, $\|\boldsymbol{\theta}^*\|_2 = 1$. If $\|\boldsymbol{\theta}''\|_2 < 1$, since $\bar{\boldsymbol{\theta}} \in \mathbb{R}^d_+$, we have $\bar{\boldsymbol{\theta}}^\intercal \boldsymbol{\theta}^* = \bar{\boldsymbol{\theta}}^\intercal \boldsymbol{\theta}''/\|\boldsymbol{\theta}''\|_2 > \bar{\boldsymbol{\theta}}^\intercal \boldsymbol{\theta}'' \ge \cos\alpha_\gamma$, suggesting that $\boldsymbol{\theta}^*$ is a feasible solution to the maximization problem in $h(\mathbf{x})$. Moreover, since $\mathbf{f}(\mathbf{x}) \in \mathbb{R}^d_+$, we have $\boldsymbol{\theta}^{*\intercal}\mathbf{f}(\mathbf{x}) = \boldsymbol{\theta}''^\intercal \mathbf{f}(\mathbf{x})/\|\boldsymbol{\theta}''\|_2 \ge \boldsymbol{\theta}''^\intercal \mathbf{f}(\mathbf{x})$. Hence, $\boldsymbol{\theta}^*$ is also an optimal solution.

Next, we derive the dual of the inner maximization problem. Let λ be the Lagrangian multiplier associated with the constraint $\bar{\boldsymbol{\theta}}^\intercal \boldsymbol{\theta} \geq \cos \alpha_\gamma$ and $\mathbf{f}(\mathbf{x}) := (f_1(\mathbf{x}), f_2(\mathbf{x}), \ldots, f_d(\mathbf{x}))^\intercal$. For any $\mathbf{x} \in \mathcal{X}(\mathbf{u})$, we have

$$
\begin{aligned}
&\max_{\boldsymbol{\theta}} \left\{ \boldsymbol{\theta}^\intercal \mathbf{f}(\mathbf{x}) \,\middle|\, \bar{\boldsymbol{\theta}}^\intercal \boldsymbol{\theta} \geq \cos \alpha_\gamma,\ \|\boldsymbol{\theta}\|_2 \leq 1 \right\} \\
&= \max_{\boldsymbol{\theta}: \|\boldsymbol{\theta}\|_2 \leq 1} \min_{\lambda \geq 0} \boldsymbol{\theta}^\intercal \mathbf{f}(\mathbf{x}) + \lambda \left(\bar{\boldsymbol{\theta}}^\intercal \boldsymbol{\theta} - \cos \alpha_\gamma \right) \\
&= \min_{\lambda \geq 0} \max_{\boldsymbol{\theta}: \|\boldsymbol{\theta}\|_2 \leq 1} \boldsymbol{\theta}^\intercal \mathbf{f}(\mathbf{x}) + \lambda \left(\bar{\boldsymbol{\theta}}^\intercal \boldsymbol{\theta} - \cos \alpha_\gamma \right) \\
&= \min_{\lambda \geq 0} \ -\lambda \cos \alpha_\gamma + \max_{\boldsymbol{\theta}: \|\boldsymbol{\theta}\|_2 \leq 1} \boldsymbol{\theta}^\intercal \left[\mathbf{f}(\mathbf{x}) + \lambda \bar{\boldsymbol{\theta}} \right] \\
&= \min_{\lambda \geq 0} \ \left\| \mathbf{f}(\mathbf{x}) + \lambda \bar{\boldsymbol{\theta}} \right\|_2 - \lambda \cos \alpha_\gamma.
\end{aligned}
$$

We obtain the second line by applying the Lagrangian relaxation. Strong duality holds because the inner maximization problem is convex and satisfies the Slater's condition, i.e., there exists a strictly feasible solution $\boldsymbol{\theta} = \bar{\boldsymbol{\theta}} \cos(\alpha/2)$. The third line holds because the objective function in the second line is linear in both $\boldsymbol{\theta}$ and λ. We obtain the fourth line by re-arranging terms. The final line holds because the inner maximization problem by definition calculates the dual norm of the l_2 norm which is equivalent to the l_2 norm.

As a result, Problem (16) can be formulated as

$$
\underset{\mathbf{x} \in \mathcal{X}(\mathbf{u}),\, \lambda \geq 0}{\text{minimize}} \ \left\| \mathbf{f}(\mathbf{x}) + \lambda \bar{\boldsymbol{\theta}} \right\|_2 - \lambda \cos \alpha_\gamma.
$$

□

EC.2.8. Proof of Proposition 2

Proof. We start by defining notations that will be useful in this proof. Since $\bar{\boldsymbol{\theta}} \in \mathbb{R}^d_+$ and $\mathbf{f}(\mathbf{x}) \in \mathbb{R}^d_+$ for any $\mathbf{x} \in \mathcal{X}(\mathbf{u})$, according to the first step in the proof of Proposition 1, we can reformulate Problem (16) as

$$
\begin{aligned}
\underset{\mathbf{x} \in \mathcal{X}(\mathbf{u})}{\text{minimize}} \ \max_{\boldsymbol{\theta} \in \mathbb{R}^d} \quad & \boldsymbol{\theta}^\intercal \mathbf{f}(\mathbf{x}) && \text{(EC.21a)} \\
\text{subject to} \quad & \bar{\boldsymbol{\theta}}^\intercal \boldsymbol{\theta} \geq \cos \alpha_\gamma && \text{(EC.21b)} \\
& \|\boldsymbol{\theta}\|_2^2 \leq 1. && \text{(EC.21c)}
\end{aligned}
$$

Note that here we replace the constraint $\|\boldsymbol{\theta}\|_2 \leq 1$ with $\|\boldsymbol{\theta}\|_2^2 \leq 1$ as it helps to simplify the computation. It is easy to verify that these two formulations are equivalent. For notational convenience, we generalize the definition of function h as

$$h(\mathbf{x}, \kappa) := \max_{\boldsymbol{\theta}} \left\{ \boldsymbol{\theta}^\intercal \mathbf{f}(\mathbf{x}) \,\middle|\, \bar{\boldsymbol{\theta}}^\intercal \boldsymbol{\theta} \geq \cos \alpha_\gamma,\ \|\boldsymbol{\theta}\|_2^2 \leq \kappa \right\}. \tag{EC.22}$$

We require $\kappa \geq \cos^2 \alpha_\gamma$, so the feasible region is non-empty. Function $h(\mathbf{x}, \kappa)$ can be interpreted as the cost of decision $\mathbf{x}$ as evaluated by a generalized version of Problem (EC.21) with the right-hand side of Constraints (EC.21c) being κ. When $\kappa = 1$, it recovers the original function $h(\mathbf{x})$. Moreover, it is easy to verify that for any fixed $\mathbf{x} \in \mathcal{X}(\mathbf{u})$, $h(\mathbf{x}, \kappa)$ is increasing and concave in κ.

Step I: We first derive an upper bound on $h(\mathbf{x}, 1 + \Delta_\kappa) - h(\mathbf{x}, 1)$ for any $\Delta_\kappa \in \mathbb{R}_+$. Let $\lambda \in \mathbb{R}_+$ and $\pi \in \mathbb{R}_+$ be the dual decision variables associated with Constraints (EC.21b) and (EC.21c), respectively. We have, for any fixed $\mathbf{x} \in \mathbb{R}^n$, that

$$\begin{aligned}
h(\mathbf{x}, \kappa) = &\quad \max_{\boldsymbol{\theta}} \left\{ \boldsymbol{\theta}^\intercal \mathbf{f}(\mathbf{x}) \,\middle|\, \bar{\boldsymbol{\theta}}^\intercal \boldsymbol{\theta} \geq \cos \alpha_\gamma,\ \|\boldsymbol{\theta}\|_2 \leq \kappa \right\}, \\
= &\quad \min_{\lambda \in \mathbb{R}_+,\, \pi \in \mathbb{R}_+} \ \kappa\pi - \lambda \cos \alpha_\gamma + \max_{\boldsymbol{\theta} \in \mathbb{R}^d} \left(\mathbf{f}(\mathbf{x}) + \lambda \bar{\boldsymbol{\theta}} \right)^\intercal \boldsymbol{\theta} - \pi \|\boldsymbol{\theta}\|_2^2 \\
= &\quad \min_{\lambda \in \mathbb{R}_+,\, \pi \in \mathbb{R}_+} \ \kappa\pi + \frac{1}{4\pi} \left\| \mathbf{f}(\mathbf{x}) + \lambda \bar{\boldsymbol{\theta}} \right\|_2^2 - \lambda \cos \alpha_\gamma.
\end{aligned}$$

Let (λ^*, π^*) be the optimal dual solution. We have, for any $\Delta_\kappa \in \mathbb{R}_+$, that

$$h(\mathbf{x}, \kappa + \Delta_\kappa) \leq h(\mathbf{x}, \kappa) + \Delta_\kappa \pi^*. \tag{EC.23}$$

Now, we search for the optimal dual solution (λ^*, π^*). For notational convenience, let $\phi = \bar{\boldsymbol{\theta}}^\intercal \mathbf{f}(\mathbf{x})$ and $\omega = \mathbf{f}(\mathbf{x})^\intercal \mathbf{f}(\mathbf{x})$. We know $\phi \geq 0$ and $\omega \geq 0$ since $\mathbf{f}(\mathbf{x}), \bar{\boldsymbol{\theta}} \in \mathbb{R}_+^d$. Moreover, $\omega \geq \phi^2$, as follows from the Cauchy-Schwarz inequality. Using the first order condition, we have, for any fixed $\lambda \in \mathbb{R}_+$, that the optimal π is

$$\pi^*(\lambda) = \frac{1}{2} \left\| \mathbf{f}(\mathbf{x}) + \lambda \bar{\boldsymbol{\theta}} \right\|_2 = \frac{1}{2} \sqrt{\frac{\lambda^2 + 2\phi\lambda + \omega}{\kappa}}.$$

Then the dual problem becomes

$$\min_{\lambda \in \mathbb{R}_+} \ \bar{h}(\lambda),$$

where $\bar{h}(\lambda) = \sqrt{\kappa(\lambda^2 + 2\phi\lambda + \omega)} - \lambda \cos \alpha_\gamma$.

Next, we derive the optimal λ^* by examining the first order condition. We have

$$\begin{aligned}
\frac{\partial \bar{h}}{\partial \lambda} =& \frac{\sqrt{\kappa}(\lambda+\phi)}{\sqrt{\lambda^2+2\phi\lambda+\omega}} - \cos\alpha_\gamma \geq 0 \\
\Leftrightarrow \quad & (\kappa-\cos^2\alpha_\gamma)\lambda^2 + 2(\kappa-\cos^2\alpha_\gamma)\phi\lambda + \phi^2 - (\cos\alpha_\gamma)^2\omega \geq 0 \\
\Leftrightarrow \quad & (\kappa-\cos^2\alpha_\gamma)(\lambda+\phi)^2 + (\cos\alpha_\gamma)^2(\phi^2-\omega) \geq 0 \\
\Leftrightarrow \quad & (\lambda+\phi)^2 \geq \frac{(\cos\alpha_\gamma)^2(\omega-\phi^2)}{\kappa-\cos^2\alpha_\gamma}.
\end{aligned}$$

We consider the following two cases.

Case I: When $\cos\alpha_\gamma\sqrt{\omega-\phi^2}/\sqrt{\kappa-\cos^2\alpha_\gamma} \geq \phi$, we have

$$\lambda^* = \frac{\cos\alpha_\gamma\sqrt{\omega-\phi^2}}{\sqrt{\kappa-\cos^2\alpha_\gamma}} - \phi,$$

and

$$\pi^* = \sqrt{\frac{\omega-\phi^2}{4(\kappa-\cos^2\alpha_\gamma)}} \leq \sqrt{\frac{\omega}{4(\kappa-\cos^2\alpha_\gamma)}} \leq \frac{v}{\sqrt{4(\kappa-\cos^2\alpha_\gamma)}},$$

where $v = \max\{\|\mathbf{f}(\mathbf{x})\|_2 \,|\, \mathbf{x}\in\mathcal{X}(\mathbf{u})\}$, which is finite because the basis functions f_i are continuous and the feasible set $\mathcal{X}(\mathbf{u})$ is compact.

Case II: When $\cos\alpha_\gamma\sqrt{\omega-\phi^2}/\sqrt{\kappa-\cos^2\alpha_\gamma} < \phi$, $\bar{h}(\lambda)$ is non-decreasing in $\mathbb{R}_+$. So, $\lambda^* = 0$ and

$$\pi^* = \sqrt{\frac{\omega}{4\kappa}} \leq \frac{v}{\sqrt{4\kappa}}.$$

When setting $\kappa = 1$, according to inequality (EC.23), we have

$$h(\mathbf{x}, 1+\Delta_\kappa) - h(\mathbf{x},1) \leq \max\left\{\frac{v}{2}, \frac{v}{2\sin\alpha_\gamma}\right\}\cdot\Delta_\kappa = \frac{v\Delta_\kappa}{2\sin\alpha_\gamma}. \tag{EC.24}$$

Step II: We next derive an upper bound on $h(\mathbf{x},1) - h(\mathbf{x},1-\Delta_\kappa)$ for any $\Delta_\kappa \in [0, \sin^2\alpha_\gamma]$.

We prove this in two sub-steps: (a) we bound the slope ρ of the secant line connecting $h(\mathbf{x},1)$ and $h(\mathbf{x},\cos^2\alpha_\gamma)$, and (b) we leverage the concavity of function h to construct a bound on $h(\mathbf{x},1) - h(\mathbf{x},1-\Delta_\kappa)$ based on the bound on slope ρ.

Step II(a): We first calculate $h(\mathbf{x},\cos^2\alpha_\gamma)$. When setting $\kappa = \cos^2\alpha_\gamma$, the feasible region of the maximization problem defined in Equation (EC.22) for evaluating $h(\mathbf{x},\kappa)$ is a singleton $\{\cos\alpha_\gamma\bar{\boldsymbol{\theta}}\}$. So, $h(\mathbf{x},\cos^2\alpha_\gamma) = (\cos\alpha_\gamma)\phi$.

We next calculate $h(\mathbf{x},1)$ using the first-order conditions identified in Step I and bound β accordingly. When in Case I, it is easy to calculate that $h(\mathbf{x},1) = \sin\alpha_\gamma\sqrt{\omega-\phi^2} + \phi\cos\alpha_\gamma$. So, we have

$$\beta = \frac{h(\mathbf{x},1) - h(\mathbf{x},\cos^2\alpha_\gamma)}{1-\cos^2\alpha_\gamma} = \frac{\sqrt{\omega-\phi^2}}{\sin\alpha_\gamma} \leq \frac{v}{\sin\alpha_\gamma}.$$

When in Case II, we have $h(\mathbf{x},1)=\sqrt{\omega}/2\le\phi\sqrt{\sin^2\alpha_\gamma+1}/2$. So, we have

$$\beta=\frac{h(\mathbf{x},1)-h(\mathbf{x},\cos^2\alpha_\gamma)}{1-\cos^2\alpha_\gamma}\le\frac{\phi\sqrt{\sin^2\alpha_\gamma+1}/2-\phi\cos\alpha_\gamma}{\sin^2\alpha_\gamma}\le\frac{\phi}{\sin\alpha_\gamma}\le\frac{\sqrt{w}}{\sin\alpha_\gamma}\le\frac{v}{\sin\alpha_\gamma}.$$

Step II(b): Since $h(\mathbf{x},\kappa)$ is concave in κ, we have

$$h(\mathbf{x},1)-h(\mathbf{x},1-\Delta_\kappa)\le\beta\Delta_\kappa\le\frac{v\Delta_\kappa}{\sin\alpha_\gamma}.$$

Step III: We next bound the approximation error of $g(\mathbf{x})$ to $h(\mathbf{x},0)$. Recall that $g(\mathbf{x})$ is defined in Equation (19) and is associated with a norm $|||\cdot|||$. Since $|||\boldsymbol{\theta}|||-\epsilon\le\|\boldsymbol{\theta}\|_2\le|||\boldsymbol{\theta}|||+\epsilon$ for any fixed $\boldsymbol{\theta}\in\mathbb{R}^d$, we have,

$$\begin{aligned}
&\left\{\boldsymbol{\theta}\in\mathbb{R}^d_+\,\middle|\,\bar{\boldsymbol{\theta}}^\intercal\boldsymbol{\theta}\ge\cos\alpha_\gamma,\,\|\boldsymbol{\theta}\|_2^2\le(1-\epsilon)^2\right\}\\
=&\left\{\boldsymbol{\theta}\in\mathbb{R}^d_+\,\middle|\,\bar{\boldsymbol{\theta}}^\intercal\boldsymbol{\theta}\ge\cos\alpha_\gamma,\,\|\boldsymbol{\theta}\|_2\le1-\epsilon\right\}\\
\subseteq&\left\{\boldsymbol{\theta}\in\mathbb{R}^d_+\,\middle|\,\bar{\boldsymbol{\theta}}^\intercal\boldsymbol{\theta}\ge\cos\alpha_\gamma,\,|||\boldsymbol{\theta}|||\le1\right\}\\
\subseteq&\left\{\boldsymbol{\theta}\in\mathbb{R}^d_+\,\middle|\,\bar{\boldsymbol{\theta}}^\intercal\boldsymbol{\theta}\ge\cos\alpha_\gamma,\,\|\boldsymbol{\theta}\|_2\le1+\epsilon\right\}\\
=&\left\{\boldsymbol{\theta}\in\mathbb{R}^d_+\,\middle|\,\bar{\boldsymbol{\theta}}^\intercal\boldsymbol{\theta}\ge\cos\alpha_\gamma,\,\|\boldsymbol{\theta}\|_2^2\le(1+\epsilon)^2\right\}.
\end{aligned}$$

So, the approximation error is bounded as

$$\begin{aligned}
|g(\mathbf{x})-h(\mathbf{x},1)|\le&\max\left\{h(\mathbf{x},\epsilon^2+2\epsilon+1)-h(\mathbf{x},1),h(\mathbf{x},1)-h(\mathbf{x},\epsilon^2-2\epsilon+1)\right\}\\
\le&\max\left\{\frac{v(\epsilon^2+2\epsilon)}{2\sin\alpha_\gamma},\frac{v(2\epsilon-\epsilon^2)}{\sin\alpha_\gamma}\right\}\\
\le&\frac{v(\epsilon^2+2\epsilon)}{2\sin\alpha_\gamma}.
\end{aligned}$$

Step IV: Finally, we relate the approximation error to the decision error. Let $\mathbf{x}^*$ and $\mathbf{x}'$ be optimal solutions to Problems (17) and (18), respectively. We have

$$\begin{aligned}
h(\mathbf{x}',0)-h(\mathbf{x}^*,0)=&h(\mathbf{x}',0)-g(\mathbf{x}')+g(\mathbf{x}')-h(\mathbf{x}^*,0)+g(\mathbf{x}^*)-g(\mathbf{x}^*)\\
\le&\frac{v(\epsilon^2+2\epsilon)}{\sin\alpha_\gamma}+g(\mathbf{x}')-g(\mathbf{x}^*)\\
\le&\frac{v(\epsilon^2+2\epsilon)}{\sin\alpha_\gamma}.
\end{aligned}$$

The first inequality holds due to the bound on the approximation error of $g(\mathbf{x})$ to $h(\mathbf{x},0)$. The second inequality holds due to the definition of $\mathbf{x}'$. □

EC.2.9. Proof of Corollary 1

Poof. We first define notations. Let $\mathbf{\Theta}(\kappa) = \left\{\boldsymbol{\theta} \in \mathbb{R}^d \,\middle|\, \bar{\boldsymbol{\theta}}^\intercal \boldsymbol{\theta} \geq \cos\alpha_\gamma,\ \|\boldsymbol{\theta}\|_2^2 \leq \kappa\right\}$ be the feasible region associated with the maximization problem required for evaluating $h(\mathbf{x}, \kappa)$ as defined in the proof of Proposition 2, $\tilde{\mathbf{\Theta}} = \left\{\boldsymbol{\theta} \in \mathbb{R}^d \,\middle|\, \bar{\boldsymbol{\theta}}^\intercal \boldsymbol{\theta} \geq \cos\alpha_\gamma,\ |||\cdot|||_{\beta_1,\beta_2} \leq 1\right\}$ be the feasible region of the maximization problem required for evaluating $g(\mathbf{x})$. Note that function $g(\mathbf{x})$ is associated with the data-driven norm $|||\cdot|||_{\beta_1,\beta_2}$ in this section.

Step (i): We first show that $\mathbf{\Theta}(\cos^2\alpha_\gamma) \subset \tilde{\mathbf{\Theta}}$. It is easy to verify that $\mathbf{\Theta}(\cos^2\alpha_\gamma)$ is a singleton $\left\{\cos\alpha_\gamma \bar{\boldsymbol{\theta}}\right\}$. Moreover, we have

$$\left|\left|\left|\cos\alpha_\gamma \bar{\boldsymbol{\theta}}\right|\right|\right|_{\beta_1,\beta_2} = (\cos\alpha_\gamma)\left(\beta_1\|\bar{\boldsymbol{\theta}}\|_1 + \beta_2\|\bar{\boldsymbol{\theta}}\|_\infty\right) \leq 1.$$

So, we have $\mathbf{\Theta}(\cos^2\alpha_\gamma) \subset \tilde{\mathbf{\Theta}}$.

Step (iii): We then show that $\tilde{\mathbf{\Theta}} \subset \mathbf{\Theta}(2-\cos^2\alpha_\gamma)$. It suffices to show that $\|\boldsymbol{\theta}\|_2^2 \leq 2 - \cos^2\alpha_\gamma$, for any $\boldsymbol{\theta} \in \tilde{\mathbf{\Theta}}$. To derive an upper bound on $\|\boldsymbol{\theta}\|_2^2$, we solve the following optimization problem.

$$\begin{aligned}
\underset{\boldsymbol{\theta},\tau}{\text{maximize}} \quad & \sum_{i\in[d]} \theta_i^2 && \text{(EC.25a)}\\
\text{subject to} \quad & \sum_{i\in[d]} \bar{\theta}_i\theta_i \geq \cos\alpha_\gamma && \text{(EC.25b)}\\
& \beta_1 \sum_{i\in[d]} \theta_i + \beta_2\tau = 1 && \text{(EC.25c)}\\
& \tau \geq \theta_i,\ \forall i \in [d] && \text{(EC.25d)}\\
& \theta_i \geq 0,\ \forall i \in [d]. && \text{(EC.25e)}
\end{aligned}$$

Since we are maximizing a convex objective over a polytope, the solution must be an extreme point, which requires at least $d+1$ of the $2d+2$ constraints to be binding. We consider the following two possible cases.

Case I: Constraint (EC.25c), k of Constraints (EC.25d) and $d-k$ of Constraints (EC.25e) are binding. In this case, it is easy to verify that the optimal $\boldsymbol{\theta}$ has k entries being $1/(k\beta_1 + \beta_2$ and all other entries being zero. So, the objective value is

$$\frac{k}{(k\beta_1+\beta_2)^2} \leq \frac{1}{4\beta_1\beta_2} \leq (2-\cos^2\alpha_\gamma).$$

The first inequality becomes equality when $k = \beta_2/\beta_1$ because, by examining the first-order condition, we can see that the function $k/(k\beta_1+\beta_2)^2$ is monotonically increasing in $[0, \beta_2/\beta_1]$ and monotonically decreasing in $[\beta_2/\beta_1, \infty)$.

Case II: Constraint (EC.25b), Constraint (EC.25c), k of Constraints (EC.25d) and $d-k-1$ of Constraints (EC.25e) are binding. Since the objective function is convex, by Taylor expanding it at $\bar{\boldsymbol{\theta}}$, we have

$$\|\boldsymbol{\theta}\|_2^2 \le 1 + 2\bar{\boldsymbol{\theta}}(\boldsymbol{\theta}-\bar{\boldsymbol{\theta}}) = 2\bar{\boldsymbol{\theta}}^\intercal\boldsymbol{\theta} - 1 = 2\cos\alpha_\gamma - 1 \le (2-\cos^2\alpha_\gamma).$$

The second equality holds because Constraint (EC.25b) is binding.

So, we have $\tilde{\boldsymbol{\Theta}} \subset \boldsymbol{\Theta}(2-\cos^2\alpha_\gamma)$.

Step (iv): Now, we revisit steps III and IV in the proof of Proposition 2. Based on step III, we have

$$|g(\mathbf{x}) - h(\mathbf{x},1)| \le \frac{v(1-\cos^2\alpha_\gamma)}{2\sin\alpha_\gamma}.$$

Let $\mathbf{x}^*$ and $\mathbf{x}'$ be optimal solutions to Problems (17) and (18), respectively. Following Step IV, we have

$$\begin{aligned} h(\mathbf{x}',0) - h(\mathbf{x}^*,0) =& h(\mathbf{x}',0) - g(\mathbf{x}') + g(\mathbf{x}') - h(\mathbf{x}^*,0) + g(\mathbf{x}^*) - g(\mathbf{x}^*) \\ \le& \frac{v(1-\cos^2\alpha_\gamma)}{\sin\alpha_\gamma} + g(\mathbf{x}') - g(\mathbf{x}^*) \\ \le& v\sin\alpha_\gamma. \end{aligned}$$

□

EC.2.10. Proof of Proposition 3

We first present the full formulation of Problem (18) with the data-driven norm $|||\cdot|||_{\beta_1,\beta_2}$. We introduce an auxiliary decision variable $\theta_{\max} \in \mathbb{R}_+$ to indicate the l_∞ norm of $\boldsymbol{\theta}$. We present the following formulation.

$$\begin{aligned} \underset{\mathbf{x}\in\mathcal{X}(\mathbf{u})}{\text{minimize}} \ \underset{\boldsymbol{\theta},\theta_{\max}}{\text{maximize}} \quad & \boldsymbol{\theta}^\intercal\mathbf{f}(\mathbf{x}) && \text{(EC.26a)} \\ \text{subject to} \quad & \bar{\boldsymbol{\theta}}^\intercal\boldsymbol{\theta} \ge \cos\alpha_\gamma && \text{(EC.26b)} \\ & \beta_1\sum_{i\in[d]}\theta_i + \beta_2\theta_{\max} \le 1 && \text{(EC.26c)} \\ & \theta_{\max} - \theta_i \ge 0, \quad \forall i\in[d] && \text{(EC.26d)} \\ & \theta_i \ge 0, \quad \forall i\in[d] && \text{(EC.26e)} \end{aligned}$$

Let $\lambda \in \mathbb{R}_+$, $\zeta \in \mathbb{R}_+$, and $\boldsymbol{\phi} \in \mathbb{R}^d_+$ be the dual decision variables associated with Constraints (EC.26b)–(EC.26e), respectively. We next dualize the inner maximization problem and obtain

$$\underset{\mathbf{x},\boldsymbol{\phi},\lambda,\zeta}{\text{minimize}} \quad \zeta - \lambda \cos \alpha_\gamma \tag{EC.27a}$$

$$\text{subject to} \quad \beta_1 \zeta + \phi_i - \bar{\theta}_i \lambda \geq f_i(\mathbf{x}), \quad \forall i \in [d] \tag{EC.27b}$$

$$\beta_2 \zeta + \mathbf{1}^\intercal \boldsymbol{\phi} \geq 0 \tag{EC.27c}$$

$$\mathbf{x} \in \mathcal{X}(\mathbf{u}) \tag{EC.27d}$$

$$\boldsymbol{\phi} \in \mathbb{R}^d_+ \tag{EC.27e}$$

$$\lambda, \zeta \geq 0. \tag{EC.27f}$$

EC.3. Numerical Experiment Details

EC.3.1. 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.

EC.3.2. Forward Problems

EC.3.2.1. Shortest-path We consider the shortest path problem on a 5×5 grid network $G(\mathcal{N}, \mathcal{E})$ where $\mathcal{N}$ and $\mathcal{E}$ indicate the node and edge sets, respectively. 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.

$$\underset{\mathbf{x}}{\text{minimize}} \quad \sum_{(i,j)\in\mathcal{E}} \theta_{ij} x_{ij} \tag{EC.28a}$$

$$\text{subject to} \quad \sum_{(j,i)\in\mathcal{E}^+(i)} x_{ji} - \sum_{(i,j)\in\mathcal{E}^-(i)} x_{ij} = \begin{cases} 1, & \text{if } i = u^d \\ -1, & \text{if } i = o^d \\ 0, & \text{otherwise} \end{cases}, \quad \forall i \in \mathcal{N} \tag{EC.28b}$$

$$x_{ij} \in \{0,1\}, \quad (i,j) \in \mathcal{E}. \tag{EC.28c}$$

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 $0 \le x_{ij} \le 1$ for any $(i,j) \in \mathcal{E}$ when implementing this model.

EC.3.2.2. Knapsack We consider a knapsack problem of $d = 10$ items. 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.

$$\underset{\mathbf{x}}{\text{maximize}} \quad \sum_{i\in[d]} \theta_i x_i \tag{EC.29a}$$

$$\text{subject to} \quad \sum_{i\in[d]} w_i x_i \le u \tag{EC.29b}$$

$$x_i \in \{0,1\}, \forall i \in [d]. \tag{EC.29c}$$

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.

EC.3.3. Obtaining a Point Estimation

We consider two methods to obtain point estimations of the unknown parameters. They are i) data-driven inverse optimization with the sub-optimality loss and ii) the gradient-based method proposed by Berthet et al. (2020). We implement the method from Berthet et al. (2020) with the package provided by Tang and Khalil (2022). Hyper-parameters are tuned using a separate validation set of 200 decision data points. Batch size is set to 64. We use the Adam optimizer with an initial learning rate of 0.1. We train the model for 20 epochs.

We present the implementation details of the data-driven inverse optimization method next. We consider solving

$$\underset{\boldsymbol{\theta}\in\mathbb{R}^{|\mathcal{E}|},\boldsymbol{\epsilon}\in\mathbb{R}_+^{n_{\text{train}}}}{\text{minimize}} \quad \frac{1}{N_{\text{train}}} \sum_{k\in\mathcal{K}_{\text{train}}} l_k \tag{EC.30a}$$

$$\text{subject to} \quad l_k \ge \boldsymbol{\theta}^\intercal \hat{\mathbf{x}}_k - \boldsymbol{\theta}^\intercal \mathbf{x}, \quad \forall \mathbf{x} \in \mathcal{X}_k,\ k \in \mathcal{K}_{\text{train}} \tag{EC.30b}$$

$$\|\boldsymbol{\theta} - \mathbf{1}\|_1 \le \frac{|\mathcal{E}|}{4}. \tag{EC.30c}$$

This problem is initialized without Constraints (EC.30b), which were added iteratively using a cutting-plane method. Specifically, in each iteration, after solving Problem (EC.30), 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

$$\underset{\mathbf{x}_k\in\mathcal{X}(\mathbf{u}_k)}{\text{minimize}} \quad \boldsymbol{\theta}'^{\mathsf{T}}\mathbf{x}_k. \tag{EC.31}$$

Let $\mathbf{x}'_k$ be the optimal solution to the sub-problem. If $l'_k < \boldsymbol{\theta}'^{\mathsf{T}}\hat{\mathbf{x}}_k - \boldsymbol{\theta}'^{\mathsf{T}}\mathbf{x}'_k$, we add the following cut to Problem (EC.30)

$$l_k \geq \boldsymbol{\theta}^{\mathsf{T}}\hat{\mathbf{x}}_k - \boldsymbol{\theta}^{\mathsf{T}}\mathbf{x}'_k. \tag{EC.32}$$

We keep running this procedure until no cut is added to the master Problem (EC.30).

EC.3.4. Solving the Calibration Problem

EC.3.4.1. Shortest Path For each data point in the validation set, we calculate the value of c_k by solving the following problem

$$\begin{aligned}
\underset{\boldsymbol{\theta}\in\mathbb{R}^{|\mathcal{E}|},\mathbf{w}\in\mathbb{R}^{\mathcal{N}},\mathbf{v}\in\mathbb{R}_+^{|\mathcal{E}|}}{\text{maximize}} \quad & \bar{\boldsymbol{\theta}}^{\mathsf{T}}\boldsymbol{\theta} && \text{(EC.33a)}\\
\text{subject to} \quad & w_{d_k} - w_{o_k} - \sum_{(i,j)\in\mathcal{E}} v_{ij} = \boldsymbol{\theta}^{\mathsf{T}}\hat{\mathbf{x}}_k && \text{(EC.33b)}\\
& w_j - w_i - v_{ij} \leq c_{ij}, \quad \forall (i,j)\in\mathcal{E} && \text{(EC.33c)}\\
& \|\boldsymbol{\theta}\|_2 \leq 1. && \text{(EC.33d)}
\end{aligned}$$

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.

EC.3.4.2. Knapsack For each data point in the validation set, we calculate the value of c_k by solving the following calibration problem

$$\begin{aligned}
\underset{\boldsymbol{\theta}\in\mathbb{R}^d}{\text{maximize}} \quad & \bar{\boldsymbol{\theta}}^{\mathsf{T}}\boldsymbol{\theta} && \text{(EC.34a)}\\
\text{subject to} \quad & \boldsymbol{\theta}^{\mathsf{T}}\hat{\mathbf{x}}_k \geq \boldsymbol{\theta}^{\mathsf{T}}\mathbf{x}, \quad \forall \mathbf{x}\in\mathcal{X}(\mathbf{u}_k) && \text{(EC.34b)}\\
& \|\boldsymbol{\theta}\|_2 \leq 1. && \text{(EC.34c)}
\end{aligned}$$

We initialize this problem without Constraints (EC.34b). 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.

EC.3.5. Solving the Robust Forward Problem

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

$$\underset{\mathbf{x}\in\mathcal{X}(\mathbf{u})}{\text{minimize}} \ \underset{\boldsymbol{\theta}\in\mathbb{R}^{|\mathcal{E}|}}{\text{maximize}} \quad \boldsymbol{\theta}^{\intercal}\mathbf{x} \tag{EC.35a}$$

$$\text{subject to} \quad \bar{\boldsymbol{\theta}}^{\intercal}\boldsymbol{\theta} \geq \cos(\alpha) \tag{EC.35b}$$

$$\|\boldsymbol{\theta}\|_2 \leq 1. \tag{EC.35c}$$

We initialize this problem as follows.

$$\underset{\mathbf{x}\in\mathcal{X}(\mathbf{u}),\Omega\in\mathbb{R}_+}{\text{minimize}} \quad \Omega \tag{EC.36a}$$

$$\text{subject to} \quad \boldsymbol{\theta}^{\intercal}\mathbf{x} \leq \Omega, \quad \forall \boldsymbol{\theta} \in \tilde{\boldsymbol{\Theta}}. \tag{EC.36b}$$

We initialize $\tilde{\boldsymbol{\Theta}} = \varnothing$. We first solve Problem (EC.36), let $\mathbf{x}'$ and Ω' denote the optimal solution. Then we solve the following sub-problem

$$\underset{\boldsymbol{\theta}\in\mathbb{R}^{|\mathcal{E}|}}{\text{maximize}} \quad \boldsymbol{\theta}^{\intercal}\mathbf{x}' \tag{EC.37a}$$

$$\text{subject to} \quad \bar{\boldsymbol{\theta}}^{\intercal}\boldsymbol{\theta} \geq \cos(\alpha) \tag{EC.37b}$$

$$\|\boldsymbol{\theta}\|_2 \leq 1. \tag{EC.37c}$$

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 (EC.36). We keep running this procedure until no new solution is added to $\tilde{\boldsymbol{\Theta}}$.

EC.3.6. Training the Data-Driven Norm

In this section, we present an approach to building a training data set for fitting the data-driven norm. This procedure is summarized in Algorithm 1. Specifically, we initialize the dataset as the cost vectors generated in the uncertainty set calibration step, i.e., in the

solution of **CP** (see Section 4.1). We then augment this dataset by randomly sample cost vector pairs and create a linear combination of them. We introduce a hyper-parameter $\rho_{\max}$ to control the extrapolation ratio for generating the linear combination. This is parameter can be tuned using standard cross-validation techniques. We set $\rho_{\max} = 2$ in our computational experiments.

Algorithm 1 A data generation procedure for training the data-driven norm.

Input: Estimated cost parameters $\left\{\hat{\boldsymbol{\theta}}_k\right\}_{k\in\mathcal{K}_{\text{val}}}$; Number of new data points to generate n_{new}; Maximal Extrapolation ratio $\rho_{\max}$.

Output: A dataset $\mathcal{S}$ for training the data-driven norm.

1: Initialize the dataset $\mathcal{S} \leftarrow \left\{\left(l_2(\hat{\boldsymbol{\theta}}_k), l_1(\hat{\boldsymbol{\theta}}_k), l_\infty(\hat{\boldsymbol{\theta}}_k)\right)\right\}_{k\in\mathcal{K}_{\text{val}}}$.
2: **for** $i \in [n_{\text{new}}]$ **do**
3: Randomly draw two distinct vectors $\boldsymbol{\theta}'$ and $\boldsymbol{\theta}''$ from $\left\{\hat{\boldsymbol{\theta}}_k\right\}_{k\in\mathcal{K}_{\text{val}}}$.
4: Draw a random ratio ρ uniformly from $[-\rho_{\max}, \rho_{\max}+1]$.
5: Generate a new vector $\boldsymbol{\theta}''' \leftarrow \rho\boldsymbol{\theta}' + (1-\rho)\boldsymbol{\theta}''$.
6: Update the dataset $\mathcal{S} \leftarrow \mathcal{S} \cup \{(l_2(\boldsymbol{\theta}'''), l_1(\boldsymbol{\theta}'''), l_\infty(\boldsymbol{\theta}'''))\}$.
7: **return** $\mathcal{S}$