Conformal Inverse Optimization

Bo Lin
University of Toronto
blin@mie.utoronto.ca

Erick Delage
HEC Montréal
Mila – Québec AI Institute
erick.delage@hec.ca

Timothy C. Y. Chan
University of Toronto
Vector Institute
tcychan@mie.utoronto.ca

Abstract

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

1 Introduction

Inverse optimization (IO) is a supervised learning approach that fits parameters in an optimization model to decision data. The fitted optimization model can then be used to prescribe future decisions. Such problems naturally arise in AI applications where human preferences are not explicitly given, and instead need to be inferred from historical decisions. For this pipeline to succeed in practice, the prescribed decision should not only be of high quality but also align with human intuition (i.e., perceived to be of high-quality). The latter encourages algorithm adoption (Chen et al., 2023; Donahue et al., 2023), which is critical in many real-world settings (Liu et al., 2023; Sun et al., 2022).

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

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

New framework. We propose a new IO pipeline that integrates i) a novel method to learn uncertainty sets from decision data and ii) a robust optimization model for decision recommendation.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

https://doi.org/10.52202/079017-2029

Theoretical guarantees. We prove that, with high probability, the learned uncertainty set contains parameters that make future observed decisions optimal. This coverage guarantee leads to bounds on the optimality gap of the decisions from conformal IO, as evaluated using the ground-truth parameters and the decision maker's (DM's) perceived parameters.

Performance. Through experiments, we demonstrate strong empirical performance of conformal IO compared to classic IO and provide insights into modeling choices.

2 Literature Review

Inverse optimization. IO is a method to estimate unknown parameters in an optimization problem based on decision data (Ahuja and Orlin, 2001; Chan et al., 2014; Chan and Kaw, 2020). Early IO papers focus on deterministic settings where the observed decisions are assumed to be optimal to the specified optimization model. Recently, IO has been extended to stochastic settings where the observed decisions are subject to errors and bounded rationality. Progress has been made to provide estimators that are consistent (Aswani et al., 2018; Dong et al., 2018), tractable (Chan et al., 2019; Tan et al., 2020; Zattoni Scroccaro et al., 2024), and robust to data corruption (Mohajerin Esfahani et al., 2018). Our paper is in the stochastic stream. Unlike existing methods that provide a point estimation of the unknown parameters, we learn an uncertainty set that can be used in a robust optimization model.

Data-driven uncertainty set construction. Recently, data have become a critical ingredient to design the structure (Delage and Ye, 2010; Mohajerin Esfahani et al., 2018; Gao and Kleywegt, 2023) and calibrate the size (Chenreddy et al., 2022; Sun et al., 2023) of an uncertainty set. Our paper is related to the work of Sun et al. (2023) who first use an ML model to predict the unknown parameters and then calibrate an uncertainty set around the prediction. However, this approach does not apply in our setting as it requires observations of the unknown parameters, which we do not have access to. Our paper presents a new approach to calibrating uncertainty sets using decision data.

Estimate, then optimize. Conformal IO belongs to a family of data-driven optimization methods called "estimate, then optimize" (Elmachtoub et al., 2023). Recent research suggests that even small estimation errors may be amplified in the optimization step, resulting in significant decision errors. This issue can be mitigated by training the estimation model with decision-aware losses (Wilder et al., 2019; Mandi et al., 2022) and robustifying the optimization model (Chan et al., 2023a). We take a similar approach as the second stream, yet deviate from them by i) utilizing decision data instead of observations of the unknown parameters, and ii) focusing on both the ground-truth and perceived solution quality, the latter of which has not been studied in this stream of literature.

Preference learning. Preference learning has been studied in the context of reinforcement learning and has recently attracted significant attention due to its application in AI alignment (Ji et al., 2023). Existing methods focus on learning a reward function/decision policy that is maximally consistent with expert decision trajectories (Ng and Russell, 2000; Wu et al., 2024) or labeled preferences in the form of pair-wise comparison/ranking (Wirth et al., 2017; Christiano et al., 2017; Rafailov et al., 2024). We enrich this stream of literature in two ways. First, we leverage *unlabeled* decision data that are not state-action trajectories, but solutions to an optimization model. Second, instead of learning a policy that imitates expert behaviors, we aim to extract common wisdom from decision data crowd-sourced from DMs who are not necessarily experts to encourage algorithm adoption.

3 Preliminaries

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

3.1 Problem Setup

Data generation. Consider a *forward optimization* problem

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

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

Objective function. We focus on cases where f is linear in $\boldsymbol{\theta}$ and convex in $\mathbf{x}$, i.e., $f(\boldsymbol{\theta}, \mathbf{x}) = \sum_{i\in[d]} \theta_i f_i(\mathbf{x})$ where $f_i : \mathbb{R}^n \to \mathbb{R}$ are convex basis functions. This generalizes the linear objective $f(\boldsymbol{\theta}, \mathbf{x}) = \boldsymbol{\theta}^\intercal \mathbf{x}$. Moreover, $\mathbf{FO}$ with this objective function can be treated as a multi-objective optimization model, which has been used to model routing preferences (Rönnqvist et al., 2017), radiation therapy planning (Chan et al., 2014), and portfolio optimization (Dong and Zeng, 2021). In this setting, the optimal solution to $\mathbf{FO}$ 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.

Learning task. Given a dataset of N decision-exogenous parameter pairs $\mathcal{D} = \{\hat{\mathbf{x}}_k, \mathbf{u}_k\}_{k=1}^N$, we are interested in finding a decision policy $\bar{\mathbf{x}} : \mathcal{U} \to \mathbb{R}^n$ to suggest decisions for future $\mathbf{u}$. We require $\bar{\mathbf{x}}(\mathbf{u}) \in \mathcal{X}(\mathbf{u})$ for any $\mathbf{u} \in \mathcal{U}$. As discussed later, $\bar{\mathbf{x}}(\mathbf{u})$ is usually generated by solving an optimization model that may have multiple optimal solutions. So we consider randomized policies (e.g., uniformly sample from a set of optimal solutions). This is nonrestrictive because a deterministic policy can be recovered from a randomized policy that samples the deterministic solution with probability one.

As a running example, $\mathbf{FOP}$ may represent a shortest path problem, where $\mathbf{u}_k$ specifies the origin and destination of a trip, $\boldsymbol{\theta}^*$ indicates the ground-truth travel times on each road segment, while $\hat{\boldsymbol{\theta}}_k$ is a driver's perception of $\boldsymbol{\theta}^*$ (i.e., perceived travel times). Decision $\mathbf{x}_k$ corresponds to a path taken by a driver based on her perceived travel times. Given a set of historical trips $\{\mathbf{u}_k, \hat{\mathbf{x}}_k\}_{k\in[N]}$, our goal is to derive a routing policy $\bar{\mathbf{x}}$ that can provide path recommendations for future origin-destination pairs.

3.1.1 Assumptions

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

Assumption 2 (Bounded Inverse Feasible Set)**.** *There exists a constant $\eta \in \mathbb{R}_+$ such that, for any $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \boldsymbol{\Theta}^{\mathrm{OPT}}(\hat{\mathbf{x}}, \mathbf{u})$, for some $\hat{\mathbf{x}} \in \mathcal{X}(\mathbf{u})$ and $\mathbf{u} \in \mathcal{U}$, we have $\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_2 \le \eta$, where*

$$\boldsymbol{\Theta}^{\mathrm{OPT}}(\mathbf{x}, \mathbf{u}) := \left\{\boldsymbol{\theta} \in \mathbb{R}^d \,\middle|\, \mathbf{x} \in \mathcal{X}^{\mathrm{OPT}}(\boldsymbol{\theta}, \mathbf{u}), \|\boldsymbol{\theta}\|_2 = 1\right\}. \tag{2}$$

Assumption 3 (Bounded Divergence)**.** *There exists a constant $\sigma \in \mathbb{R}_+$ such that $\|\mathbb{E}(\hat{\boldsymbol{\theta}}) - \boldsymbol{\theta}^*\|_2 \le \sigma$.*

Assumption 1 is standard in the ML and IO literature. Assumption 2 is mild because $\boldsymbol{\Theta}^{\mathrm{OPT}}(\hat{\mathbf{x}}, \mathbf{u})$ is by definition bounded and is usually much smaller than $\boldsymbol{\Theta}$. Assumption 3 states that the l_2 distance between the expected perceived parameters and the ground-truth parameters is upper bounded. It is reasonable in many real-world settings. For example, a driver's perceived travel cost ($\hat{\boldsymbol{\theta}}$) should not be too different from the travel time ($\boldsymbol{\theta}^*$) as the latter is an important factor that drivers consider.

3.1.2 Evaluation Metrics

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

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

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

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

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

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

https://doi.org/10.52202/079017-2029

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

3.2 An Inverse Optimization Pipeline

Finding $\bar{\mathbf{x}}$ is challenging for three reasons. First, unlike many ML tasks where the prediction target is unconstrained, we require $\bar{\mathbf{x}}(\mathbf{u})$ to be feasible to $\mathbf{FO}$ which may involve a large number of constraints. Supervised learning approaches that predict $\hat{\mathbf{x}}$ based on $\mathbf{u}$ can often fail as they typically do not provide feasibility guarantees. An optimization module is often needed to recover feasibility or produce feasible solutions based on $\mathbf{u}$ and some estimated $\bar{\boldsymbol{\theta}}$. Second, we do not directly observe $\boldsymbol{\theta}^*$ or $\hat{\boldsymbol{\theta}}$, which precludes using classic ML techniques to estimate them. Finally, AOG and POG may not necessarily align with each other, so we are essentially dealing with a bi-objective problem.

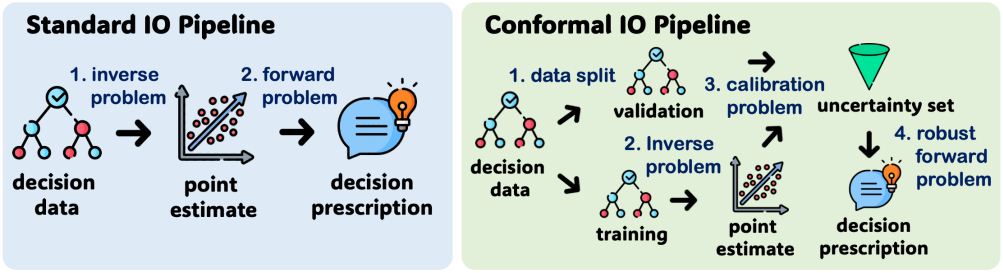

Figure 1: Classic and conformal IO pipelines.

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

$$\mathbf{IO}(\mathcal{D}) : \underset{\boldsymbol{\theta} \in \boldsymbol{\Theta}}{\text{minimize}} \ \frac{1}{N} \sum_{k \in [N]} \ell\left(\hat{\mathbf{x}}_k, \mathcal{X}^{\mathrm{OPT}}(\boldsymbol{\theta}, \mathbf{u}_k)\right), \tag{5}$$

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

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

$$\ell_S\left(\hat{\mathbf{x}}, \mathcal{X}^{OPT}(\boldsymbol{\theta}, \mathbf{u})\right) := \max_{\mathbf{x} \in \mathcal{X}(\mathbf{u})} f(\boldsymbol{\theta}, \hat{\mathbf{x}}) - f(\boldsymbol{\theta}, \mathbf{x}). \tag{6}$$

The sub-optimality loss penalizes the optimality gap achieved by the observed decision under the estimated parameters. As remarked by Mohajerin Esfahani et al. (2018), this loss function has better computational properties than its alternatives as it is convex in the unknown parameters. In fact, when the dataset $\mathcal{D}$ is large in size and the unknown parameters $\boldsymbol{\theta}$ are high-dimensional, the sub-optimality loss is usually the only loss function that leads to a tractable inverse problem, although it does not enjoy properties such as statistical consistency (see Chan et al. (2023b) for detailed discussions). While such a trade-off is acceptable in some applications, we suggest that it is undesirable in our setting because the resulting policy can achieve arbitrarily large AOG and POG. To see this, consider the following example that satisfies Assumptions 1–3. This example is visualized in Figure 2.

Example 1. *Let $\mathbf{FO}(\boldsymbol{\theta}, u)$ be the following problem*

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

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

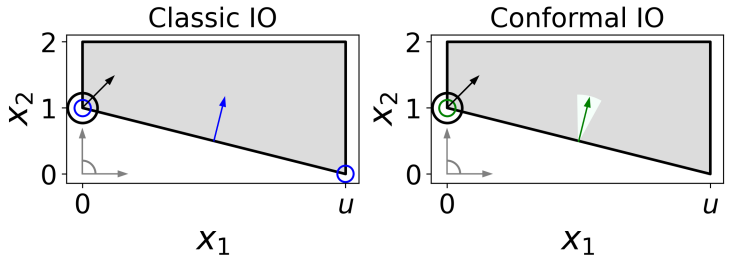

Figure 2: Illustration of Example 1. The gray areas are the feasible region $\mathcal{X}(u)$. The black arrows are the ground-truth parameter $\boldsymbol{\theta}^*$. The gray arrows are the extreme rays of $\boldsymbol{\Theta}$. The blue and green arrows are the point estimation $\bar{\boldsymbol{\theta}}$. The green area is the uncertainty set $\mathcal{C}(\bar{\boldsymbol{\theta}}, \alpha)$. The black circles are the optimal solution to $\mathbf{FO}(\boldsymbol{\theta}^*, u)$. The blue and green circles are the suggested decisions. Note that $\bar{\mathbf{x}}_{\mathrm{IO}}$ may suggest any decisions on the facet of $x_1 + ux_2 \geq u$, which are omitted for clarity.

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

Lemma 1 shows that, when using $\mathbf{IO}(\mathcal{D})$ with the sub-optimality loss to estimate the unknown parameter in Example 1, the probability of the estimated parameter being $\boldsymbol{\theta}_u$ converges to one asymptotically. This implies that asymptotically we are almost certain that $\bar{\mathbf{x}}_{\mathrm{IO}}(u) = \tilde{\mathbf{x}}(\boldsymbol{\theta}_u, u)$, i.e. the policy that samples uniformly from the facet corresponding to the constraint $x_1 + ux_2 \geq u$. As a result, $\bar{\mathbf{x}}_{\mathrm{IO}}$ can achieve arbitrarily large AOG and POG when u is set to a large enough value.

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

3.3 Robustifying the Inverse Optimization Pipeline

A natural idea to improve the AOG and POG of $\bar{\mathbf{x}}$ is to robustify the decision pipeline. Specifically, instead of solving $\mathbf{FO}$ with some estimated parameters $\bar{\boldsymbol{\theta}}$, we solve the following *robust forward optimization problem* with an uncertainty set around $\bar{\boldsymbol{\theta}}$ to prescribe decisions (final steps in Figure 1).

$$\mathbf{RFO}\left(\mathcal{C}(\bar{\boldsymbol{\theta}}, \boldsymbol{\alpha}), \mathbf{u}\right) : \underset{\mathbf{x} \in \mathcal{X}(\mathbf{u})}{\text{minimize}} \ \underset{\boldsymbol{\theta} \in \mathcal{C}(\bar{\boldsymbol{\theta}}, \boldsymbol{\alpha})}{\text{maximize}} \ f(\boldsymbol{\theta}, \mathbf{x}) \tag{8}$$

where $\mathcal{C}$ is an uncertainty set with $\bar{\boldsymbol{\theta}}$ being its center and $\boldsymbol{\alpha}$ representing parameters that control its shape/size. Given the support set $\boldsymbol{\Theta}$ defined in Section 3.1, we focus on the following uncertainty set.

$$\mathcal{C}(\bar{\boldsymbol{\theta}}, \alpha) := \left\{\boldsymbol{\theta} \in \mathbb{R}^d \,|\, \|\boldsymbol{\theta}\|_2 = 1,\ \boldsymbol{\theta}^\intercal \bar{\boldsymbol{\theta}} \geq \cos\alpha\right\} \tag{9}$$

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

Remark 1. *An alternative approach to robustify the IO pipeline is to replace $\mathbf{IO}$ with a distributionally robust IO for parameter estimation (Mohajerin Esfahani et al., 2018). However, this approach would not help as $\boldsymbol{\theta}_u$ is still optimal to the distributionally robust IO in Example 1. So, the AOG and POG can still be arbitrarily large. See Appendix A.3 for complete statement and discussions.*

Now, in Example 1, we analyze the performance of a policy that utilizes $\mathbf{RFO}$ to prescribe decisions.

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

Lemmas 2 shows that, when using $\mathbf{RFO}$ to prescribe new decisions, the probability of achieving upper-bounded AOG and POG converges to one as N goes to infinity, as long as $\alpha \in (0, \pi/2)$. These bounds are independent of u, in contrast to the AOG and POG of classic IO that can be arbitrarily large as u changes (Proposition 1). However, the performance of this approach still depends on the choice of α, which is non-trivial when $\mathbf{FO}$ is more complex than a two-dimensional linear program. We address this problem next.

https://doi.org/10.52202/079017-2029

4 Conformal Inverse Optimization

In this section, we present a principled approach to learn uncertainty sets that lead to provable performance guarantees. As presented later, the learned uncertainty set contains parameters that make the next DM's decision optimal with a specified probability. We call this approach conformal IO due to its connection to conformal prediction (Vovk et al., 2005), which aims to predict a set that contains the next prediction target with a specified probability. Our approach first converts each context-decision observation $(\mathbf{u}_k, \hat{\mathbf{x}}_k)$ to a parameter set $\mathbf{\Theta}^{\text{OPT}}(\mathbf{u}_k, \hat{\mathbf{x}}_k)$ that contains all the parameters that explain $\hat{\mathbf{x}}_k$ under $\mathbf{u}_k$. We then adapt conformal prediction to produce a set that has γ probability of containing at least one member of the next sampled $\mathbf{\Theta}^{\text{OPT}}(\mathbf{u}, \hat{\mathbf{x}})$. Note that if $\mathbf{\Theta}^{\text{OPT}}(\mathbf{u}, \hat{\mathbf{x}})$ is a singleton almost surely, the approach is equivalent to applying conformal prediction to $\boldsymbol{\theta}_k$ directly. As illustrated in Figure 1, there are three training steps in conformal IO: i) data split, ii) point estimation, and iii) uncertainty set calibration. We present these steps in Section 4.1 and analyze the properties of conformal IO in Section 4.2.

4.1 Learning an Uncertainty Set

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

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

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

$$
\begin{aligned}
\mathbf{CP}(\bar{\boldsymbol{\theta}}, \mathcal{D}_{\text{val}}, \gamma): \underset{\alpha, \{\boldsymbol{\theta}_k\}_{k\in\mathcal{K}_{\text{val}}}}{\text{minimize}} \quad & \alpha && \text{(10a)}\\
\text{subject to} \quad & \hat{\mathbf{x}}_k \in \mathcal{X}^{\text{OPT}}(\boldsymbol{\theta}_k, \mathbf{u}_k),\ \forall k \in \mathcal{K}_{\text{val}} && \text{(10b)}\\
& \sum_{k\in\mathcal{K}_{\text{val}}} \mathbb{1}\left[\boldsymbol{\theta}_k \in \mathcal{C}(\bar{\boldsymbol{\theta}}, \alpha)\right] \geq \gamma(N_{\text{val}}+1) && \text{(10c)}\\
& \|\boldsymbol{\theta}_k\|_2 = 1,\ \forall k \in \mathcal{K}_{\text{val}} && \text{(10d)}\\
& 0 \leq \alpha \leq \pi, && \text{(10e)}
\end{aligned}
$$

where decision α controls the size of the uncertainty set, $\boldsymbol{\theta}_k$ represent a possible parameter vector associated with data point $k \in \mathcal{K}_{\text{val}}$, $\gamma \in [0, 1]$ is a DM-specified confidence level. Constraints (10b) ensure that $\boldsymbol{\theta}_k$ can make the decision $\hat{\mathbf{x}}_k$ optimal for $k \in \mathcal{K}_{\text{val}}$. Constraint (10c) ensures that at least $\gamma(N_{\text{val}}+1)$ of the decisions in $\mathcal{D}_{\text{val}}$ can find a vector in $\mathcal{C}$ that makes it optimal. Constraints (10d) ensure that the parameter vectors are on the unit sphere as defined in Equation (9).

Remark 2 (Optimality Conditions)**.** *The specific form of Constraints* (10b) *depends on the structure of* **FO**. *For example, when* **FO** *is a linear program, Constraints* (10b) *can be replaced with the dual feasibility and strong duality constraints. When the* **FO** *is a general convex optimization problem, we can use the KKT conditions. For non-convex forward problems, we can replace Constraints* (10b) *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 in a cutting-plane fashion.*

Remark 3 (Feasibility)**.** *For* **CP** *to be feasible, we require, for each observed decision, there exists a* $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ *that make it optimal. This condition holds for a range of problems, e.g., routing problems and the knapsack problem, even if the DM is subject to bounded rationality, i.e., the DM settles for suboptimal solutions due to cognitive/computational limitations. For problems where this condition is violated, we may pre-process* $\mathcal{D}_{val}$ *to project* $\hat{\mathbf{x}}$ *to a point in* $\mathcal{X}(\mathbf{u})$ *such that the condition is satisfied.*

Solving **CP** is hard. First, **CP** is non-convex due to Constraints (10d). Second, Constraints (10b) involve the optimality conditions of N_{val} problems, so the size of **CP** grows quickly as N_{val} in-

creases. Nevertheless, considering a large $\mathcal{D}_{\text{val}}$ is critical to ensure desirable properties of the learned uncertainty set (Section 4.2). Below we introduce a decomposition method to solve **CP** efficiently.

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 first solving N_{val} optimization problems whose size is independent of N_{val} and then find a quantile in a set of N_{val} elements. The first step is parallelizable and the second step can be done in $O\left(N_{\text{val}} \log(\tau)\right)$ time. Since the problem required for evaluating c_k is a maximization problem, we can replace the constraint $\|\boldsymbol{\theta}_k\|_2 = 1$ with $\|\boldsymbol{\theta}_k\|_2 \leq 1$ if $\bar{\boldsymbol{\theta}} \in \mathbb{R}^d_+$, so this problem is convex when the forward problem is a linear program.

4.2 Properties of Conformal IO

Theorem 2 (Uncertainty 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 $\mathbb{P}_{(\boldsymbol{\theta},\mathbf{u})}$, $\hat{\mathbf{x}} = \tilde{\mathbf{x}}(\hat{\boldsymbol{\theta}}, \mathbf{u})$, $\hat{\boldsymbol{\Theta}} := \boldsymbol{\Theta}^{\text{OPT}}(\hat{\mathbf{x}}, \mathbf{u})$, and α_γ be an optimal solution to $\mathbf{CP}(\bar{\boldsymbol{\theta}}, \mathcal{D}_{\text{val}}, \gamma)$ where $\bar{\boldsymbol{\theta}} \in \mathbb{R}^d$. We have, for any $\gamma \in [0, N_{val}/(N_{val}+1)]$, that*

$$\mathbb{P}\left(\hat{\boldsymbol{\Theta}} \cap \mathcal{C}(\bar{\boldsymbol{\theta}}, \alpha_\gamma) \neq \varnothing\right) \geq \gamma. \tag{11}$$

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

$$\left|\mathbb{P}\left(\hat{\boldsymbol{\Theta}} \cap \mathcal{C}(\bar{\boldsymbol{\theta}}, \alpha_\gamma) \neq \varnothing\right) - \gamma\right| \leq \epsilon(N_{val}) := \sqrt{\frac{8\log(N_{val}+1) + 2\log N_{val}}{N_{val}}} + \frac{2}{N_{val}}. \tag{12}$$

Theorem 2 states that our learned uncertainty set is conservatively valid and asymptotically exact (Vovk et al., 2005). More specifically, first, our method will produce a set that contains a $\boldsymbol{\theta}$ that makes the next DM's decision optimal no less than γ of the time that it is used (conservatively valid). The probability in Inequality (11) is with respect to the joint distribution over $\mathcal{D}_{\text{val}}$ and the new sample. Second, once the set is given, we have high confidence that, the probability of the next DM's decision being covered is within $\epsilon(N_{\text{val}})$ from γ. The probability in Inequality (12) 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 $\mathbb{P}(\hat{\boldsymbol{\Theta}} \cap \mathcal{C}(\bar{\boldsymbol{\theta}}, \alpha_\gamma) \neq \varnothing)$ to γ as N goes to infinity. Finally, we note that in many practical applications, the number of decision observations N can be quite large. For example, in our motivating example, rideshare platforms observe millions of trips on a daily basis, providing ample data for both point estimation and uncertainty set calibration in our pipeline.

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

Theorem 3 (POG Bound)**.** *Let $\bar{\mathbf{x}}_{\text{CIO}}(\mathbf{u})$ be an optimal solution to $\mathbf{RFO}\left(\mathcal{C}(\bar{\boldsymbol{\theta}}, \alpha_1), \mathbf{u}\right)$ for any $\mathbf{u} \in \mathcal{U}$, where $\bar{\boldsymbol{\theta}} \in \mathbb{R}^d$ and α_1 are chosen such that, for a new sample $(\boldsymbol{\theta}', \mathbf{u}')$ from $\mathbb{P}_{(\boldsymbol{\theta},\mathbf{u})}$ and $\mathbf{x}' = \tilde{\mathbf{x}}(\boldsymbol{\theta}', \mathbf{u}')$, $\mathbb{P}\left(\mathcal{C}(\bar{\boldsymbol{\theta}}, \alpha_1) \cap \boldsymbol{\Theta}^{\text{OPT}}(\mathbf{u}', \mathbf{x}') \neq \varnothing\right) = 1$. If Assumptions 2–3 hold, then*

$$\text{POG}(\bar{\mathbf{x}}_{\text{CIO}}) \leq (\eta - 2\cos 2\alpha_1 + 2)\mu + \eta\mu_{\text{CIO}}, \tag{13}$$

and

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

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 states that, when $\mathcal{C}(\bar{\boldsymbol{\theta}}, \alpha)$ contains a $\boldsymbol{\theta}$ that makes the next DM's decision optimal almost surely, conformal IO achieves upper-bounded POG and AOG. While we can meet this condition by using a large α, the bounds would be large as they increase as α increases, reflecting that the decisions may be overly conservative. Instead, we can use **CP** to obtain a α that achieves close-to-100% coverage and possibly add a small $\Delta_\alpha \in \mathbb{R}_+$ as extra protection. Moreover, we show in Section 5

https://doi.org/10.52202/079017-2029

that, when using $\gamma < 100\%$, conformal IO still demonstrates favorable performance compared to classic IO. Our bounds have problem-specific constants. To demonstrate tightness, we present their numerical values in Example 1 with $\alpha = \pi/4$ (Table 1). They closely follow the performance of conformal IO, which outperforms classic IO by a large margin.

Table 1: Performance profile of classic and conformal IO in Example 1.

	AOG				POG			
u	2	10	50	100	2	10	50	100
Classic IO	0.35	3.18	17.32	35	0.74	4.55	24.51	49.50
Conformal IO	0.00	0.00	0.00	0.00	0.70	0.16	0.03	0.02
Conformal IO bound	0.70	0.05	0.002	0.001	1.58	0.67	0.15	0.08

5 Numerical Studies

Data generation. We consider two forward problems: The shortest path problem (linear program, $d = 120$) and knapsack problem (integer program, $d = 10$). See Appendix C.2 for their formulations. We use synthetic instances, which is a common practice as there is no well established IO benchmark (Tan et al., 2020; Dong et al., 2018). For both problems, we randomly generate a ground-truth parameters $\boldsymbol{\theta}^*$ and a dataset of $N = 1000$ DMs. For each DM $k \in [N]$, we generate her perceived parameters as $\hat{\theta}^i_k = (\theta^{i*}_k * p^i_k + \epsilon^i_k)^+ + \epsilon_0$ for $i \in [d]$ where p^i_k is uniformly drawn from $[1/2, 2]$, ϵ^i_k is drawn from a normal distribution with mean 0 and standard deviation 1, and $\epsilon_0 = 0.1$. For the shortest path problem, parameters $\mathbf{u}_k$ represent a random origin-destination pair on the network. For the knapsack problem, $\mathbf{u}_k$ correspond to the weights of different items and the DM's budget. The item weight w^i for $i \in [d]$ are uniformly drawn from $[1, 10]$ and are shared among DMs. For each DM $k \in [N]$, we generate a budget $u_k = q_k \sum_i w^i$ where q_k is uniformly drawn from $[1/5, 5]$.

Experiment design. Conformal IO is compatible with any approach that can provide a point estimation of unknown parameters using *decision data* (Step 2 in Section 4.1). To the best of our knowledge, i) **IO** 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 these two methods. They also serve as our baselines. We call both i) and ii) classic IO to emphasize that they rely on a point estimation for decision prescription, although PFYL is not an IO approach. See Appendix C for implementation details. In all experiments, conformal IO uses the training set for point estimation and the validation set for calibration, while classic IO uses the union of the training and validation sets for point estimation. So, they have access to the same amount of data and are evaluated on the same test set. Unless otherwise noted, experiments are based on a 60/20/20 train-validation-test split and are repeated 10 times with different random seeds.

Uncertainty validity. To verify Theorem 2, we empirically evaluate the out-of-sample coverage achieved by our uncertainty set under different target levels γ and sample sizes N_{val}. The point estimation is generated by **IO** with the sub-optimality loss. As shown in Figure 3, when the validation set is small ($N_{\text{val}} = 10$), we always achieve the specified target but $\mathcal{C}(\hat{\boldsymbol{\theta}}, \alpha_\gamma)$ tends to over-cover (conservatively valid). When using larger validation sets ($N_{\text{val}} \in \{100, 200\}$), our coverage level gets closer to the specified γ (asymptotic exact). These empirical findings echo our theoretical analysis.

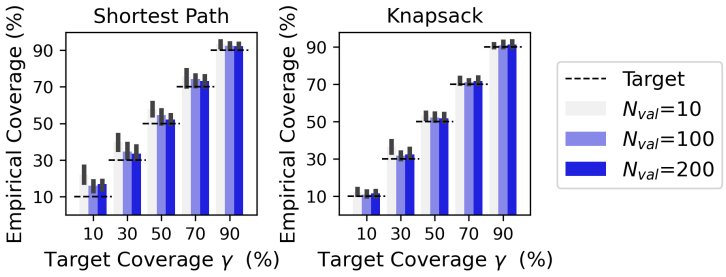

Figure 3: Empirical coverage achieved by the learned uncertainty set (error bar = range).

The value of robustness. As shown in Figure 4, solving **RFO** with an uncertainty set learned by conformal IO leads to decisions of lower AOG and POG, compared to solving **FO** with a point estimation from classic IO. On average, when varying γ, our approach improves AOG by 20.1–30.4% and POG by 15.0–23.2% for the shortest path problem, and improves AOG by 40.3–57.0% and POG by 13.5–20.1% for the knapsack problem. The improvement is orthogonal to the point estimation method. Our decisions are of higher quality and better align with human intuition than classic IO. Finally, the performance of conformal IO generally improves as the quality of the point estimate increases (see Appendix D for details). Therefore, even though the conformal IO pipeline is robust to estimation errors, using the best available point estimation method is still recommended.

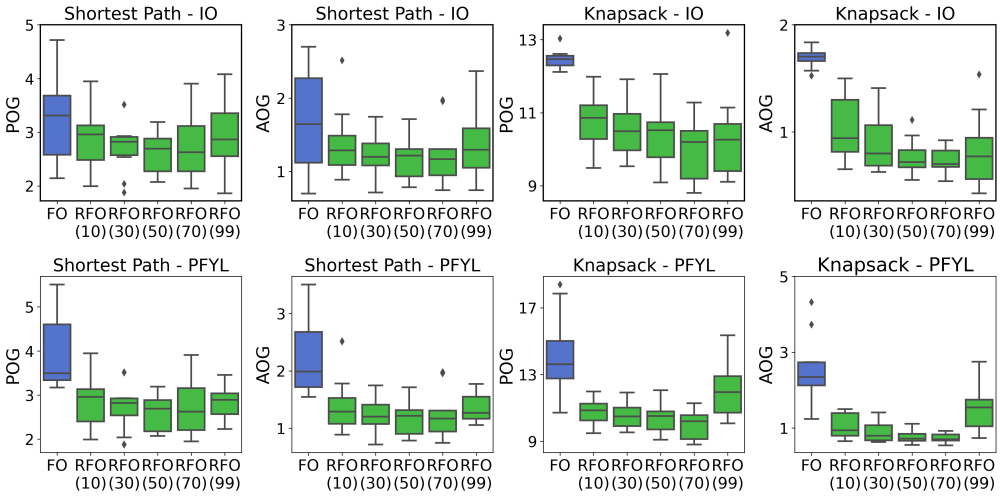

Figure 4: Performance profile of classic (blue) and conformal IO (green).

Computational efficiency. As shown in Table 2, conformal IO and classic IO require similar training times. When **FO** is an integer program (knapsack), the training of conformal IO is even faster 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 that are parallelizable (Theorem 1). At the prediction time, our method achieves lower AOG and POG at the cost of solving a more challenging **RFO**. Nevertheless, the solution time of **RFO** is within one second in our instances.

Table 2: Average (std) computational time of classic and conformal IO in seconds.

	Training		Prediction (per decision)	
Problem	Classic IO	Conformal IO	**FO**	**RFO**
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)

Important hyper-parameters. Finally, we provide empirical evidence that sheds light on the choice of two important hyper-parameters in conformal IO: 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 using cross-validation. However, this requires an additional validation dataset. If such a dataset is unavailable, setting γ to a relatively large value (e.g., 0.99) usually yields decent performance, which aligns with our theoretical analysis. Regarding *train-validation split ratio*, intuitively, both the estimation and calibration can benefit from more data. However, when the dataset is small, we need to strike a balance between these two steps aiming to achieve lower AOG and POG. 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, there is no benefit of using conformal IO because we do not have enough data to obtain good point estimation and uncertainty set simultaneously. However,

https://doi.org/10.52202/079017-2029

the performance of classic IO quickly plateaus as the dataset grows. When given a mid- or large-sized dataset, we can generally benefit from putting more data in $\mathcal{D}_{\text{val}}$, echoing our theoretical analysis.

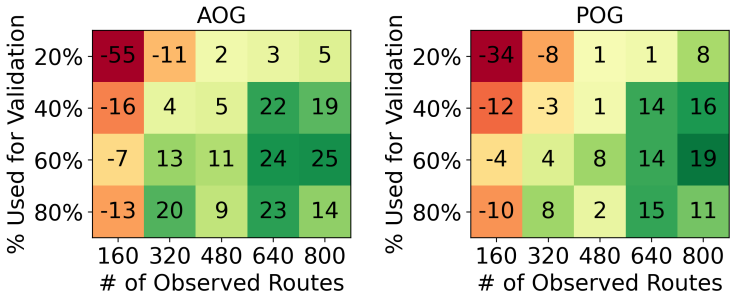

Figure 5: Percentage reduction in AOG and POG when using the conformal IO vs classic IO.

6 Conclusion and Future Work

In this paper, we propose conformal IO, a novel IO pipeline to recommend high-quality decisions that align with human intuition. We present a new approach to learning uncertainty sets from decision data, which is then utilized in a robust optimization model to prescribe new decisions. We prove that conformal IO achieves bounded optimality gaps, as measured by the ground-truth parameters and the DM's perceived parameters. We demonstrate the strong empirical performance of conformal IO via extensive numerical studies. Finally, we highlight several challenges that underscore future research directions. First, we focus on objectives that are linear in the unknown parameters. While such objectives are ubiquitous in practice, it is of interest to extend our results for general convex objectives. Second, the robust forward problem, while leading to better decisions, is computationally more costly than the forward problem. Future research can be done to accelerate its solution process. Finally, while Conformal IO consistently outperforms classic IO when the point estimation methods are fixed, our computational experiments suggest that its performance hinges on the quality of the point estimate. Future research could explore point estimation methods that directly optimize the performance of the downstream robust optimization model.

Acknowledgments

The authors are grateful to the anonymous reviewers for their valuable feedback and insightful comments. Erick Delage was partially supported by the Canadian Natural Sciences and Engineering Research Council [Grant RGPIN-2022-05261] and by the Canada Research Chair program [950-230057].

References

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

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

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

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

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

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

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

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

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

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

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

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

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

Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., and Amodei, D. (2017). Deep reinforcement learning from human preferences. In Advances in neural information processing systems, volume 30.

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

Donahue, K., Kollias, K., and 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., and Zeng, B. (2018). Generalized inverse optimization through online learning. In Advances in Neural Information Processing Systems, volume 31.

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

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

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

Hu, X., Cirit, O., Binaykiya, T., and 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.

Ji, J., Qiu, T., Chen, B., Zhang, B., Lou, H., Wang, K., Duan, Y., He, Z., Zhou, J., Zhang, Z., et al. (2023). AI alignment: A comprehensive survey. arXiv preprint arXiv:2310.19852.

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

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

 https://doi.org/10.52202/079017-2029

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

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

Mohri, M., Rostamizadeh, A., and Talwalkar, A. (2018). Foundations of Machine Learning. MIT press.

Ng, A. Y. and Russell, S. J. (2000). Algorithms for inverse reinforcement learning. In Proceedings of the Seventeenth International Conference on Machine Learning, page 663–670.

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

Rafailov, R., Sharma, A., Mitchell, E., Manning, C. D., Ermon, S., and Finn, C. (2024). Direct preference optimization: Your language model is secretly a reward model. Advances in Neural Information Processing Systems, 36.

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

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

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

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

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

Tang, B. and Khalil, E. B. (2024). PyEPO: A pytorch-based end-to-end predict-then-optimize library for linear and integer programming. Mathematical Programming Computation, 16(3):297–335.

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

Wainwright, M. J. (2019). High-dimensional Statistics: A non-Asymptotic Viewpoint, volume 48. Cambridge University Press.

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

Wirth, C., Akrour, R., Neumann, G., and Fürnkranz, J. (2017). A survey of preference-based reinforcement learning methods. Journal of Machine Learning Research, 18(136):1–46.

Wu, F., Ke, J., and Wu, A. (2024). Inverse reinforcement learning with the average reward criterion. In Advances in Neural Information Processing Systems, volume 36.

Zattoni Scroccaro, P., Atasoy, B., and Mohajerin Esfahani, P. (2024). Learning in inverse optimization: Incenter cost, augmented suboptimality loss, and algorithms. Operations Research, 0(0).

A Omitted Statements and Proofs in Section 3

A.1 Poof of Lemma 1

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

$$\hat{\mathbf{x}}_k = \begin{cases} (0,1), \text{ w.p. } 2\delta_u/\pi \\ (u,0), \text{ w.p. } (\pi-2\delta_u)/\pi. \end{cases} \tag{15}$$

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

$$\bar{\boldsymbol{\theta}}_N := \operatorname*{argmin}_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \quad \frac{N_1}{N} l_1(\boldsymbol{\theta}) + \frac{N_2}{N} l_2(\boldsymbol{\theta}) \tag{16}$$

where

$$l_1(\boldsymbol{\theta}) = \begin{cases} 0, & \text{if } \boldsymbol{\theta}\in\boldsymbol{\Theta}_1, \\ \theta_2 - u\theta_1, & \text{if } \boldsymbol{\theta}\in\boldsymbol{\Theta}_2, \end{cases} \tag{17}$$

and

$$l_2(\boldsymbol{\theta}) = \begin{cases} u\theta_1 - \theta_2, & \text{if } \boldsymbol{\theta}\in\boldsymbol{\Theta}_1, \\ 0, & \text{if } \boldsymbol{\theta}\in\boldsymbol{\Theta}_2. \end{cases} \tag{18}$$

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

$$\mathbb{P}(N_1N_2>0) \le \mathbb{P}\left(\bar{\boldsymbol{\theta}}_N = (\cos\delta_u,\sin\delta_u)\right) \le 1. \tag{19}$$

Given the probability distribution given in Equation (15) and that $\mathcal{D}$ is generated using i.i.d. samples from $\mathbb{P}_\theta$, we have

$$\mathbb{P}(N_1N_2>0) = 1 - \left(\frac{2\delta_u}{\pi}\right)^N - \left(1-\frac{2\delta_u}{\pi}\right)^N, \tag{20}$$

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

A.2 Proof of Proposition 1

Proof. We first consider the AOG and POG of the decision policy $\bar{\mathbf{x}}_{\text{IO}}(u) := \tilde{\mathbf{x}}(\boldsymbol{\theta}_u,u)$ separately in the following two lemmas.

Lemma 4. *In Example 1, let $\bar{\mathbf{x}}_{\text{IO}}(u) = \tilde{\mathbf{x}}(\boldsymbol{\theta}_u,u)$. For any $v\in\mathbb{R}_+$ there exists some $\bar{u}>1$ such that $\text{AOG}(\bar{\mathbf{x}}_{\text{IO}}) > v$ for any $u>\bar{u}_{AOG}$.*

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

$$\mathcal{X}^{\text{OPT}}(\boldsymbol{\theta}_u,u) = \left\{ \left(\frac{ut}{\sqrt{u^2+1}}, 1-\frac{t}{\sqrt{u^2+1}}\right) \,\middle|\, t\in\left[0,\sqrt{u^2+1}\right] \right\}. \tag{21}$$

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

$$\text{AOG}(\bar{\mathbf{x}}_{\text{IO}}) = \int_0^{\sqrt{u^2+1}} \frac{\sqrt{2}}{2\sqrt{u^2+1}} \left(1 - \frac{t}{\sqrt{u^2+1}} + \frac{ut}{\sqrt{u^2+1}}\right) dt - \frac{\sqrt{2}}{2} = \frac{\sqrt{2}(u-1)}{4} \tag{22}$$

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

https://doi.org/10.52202/079017-2029

Lemma 5. *In Example 1, let* $\bar{\mathbf{x}}_{\mathrm{IO}}(u) = \tilde{\mathbf{x}}(\boldsymbol{\theta}_u, u)$. *for any* $v \in \mathbb{R}_+$ *there exists some* $\bar{u}_{POG} > 1$ *such that* $\mathrm{POG}(\bar{\mathbf{x}}_{\mathrm{IO}}) > v$ *for any* $u > \bar{u}_{POG}$.

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

$$\mathcal{X}^{\mathrm{OPT}}(\boldsymbol{\theta}_u, u) = \left\{ \left(\frac{ut}{\sqrt{u^2+1}}, 1 - \frac{t}{\sqrt{u^2+1}} \right) \middle| \, t \in \left[0, \sqrt{u^2+1}\right] \right\}. \tag{23}$$

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

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

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

Based on Lemmas 4 and 5, we conclude that for any $v \in \mathbb{R}_+$, there exists $\bar{u} = \max\{\bar{u}_{\mathrm{AOG}}, \bar{u}_{\mathrm{POG}}\} = 2\sqrt{2}v + 1$ such that $\mathrm{AOG}(\bar{\mathbf{x}}_{\mathrm{IO}}) > v$ and $\mathrm{POG}(\bar{\mathbf{x}}_{\mathrm{IO}}) > v$ for any $u > \bar{u}$. □

A.3 Robustifying the Inverse Problem

An alternative approach to robustify the classic IO pipeline is to solve a distributionally robust inverse optimization problem. Specifically, consider the following loss function proposed by Mohajerin Esfahani et al. (2018).

Definition 4. *The distributionally robust sub-optimality loss of* $\boldsymbol{\theta}$ *is given by*

$$\ell_{DR\text{-}S}(\boldsymbol{\theta}) := \sup_{\mathbb{Q} \in \mathfrak{B}_r^p(\hat{\mathbb{P}}_{\mathbf{u},\hat{\mathbf{x}}})} \rho^{\mathbb{Q}} \left[\ell_S \left(\hat{\mathbf{x}}, \mathcal{X}^{OPT}(\boldsymbol{\theta}, \mathbf{u}) \right) \right] \tag{24}$$

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

The *distributionally robust inverse optimization* problem is

$$\mathbf{DRIO}(\mathcal{D}) : \underset{\boldsymbol{\theta} \in \boldsymbol{\Theta}}{\text{minimize}} \quad \ell_{\text{DR-S}}(\boldsymbol{\theta}). \tag{25}$$

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

Lemma 6. *In Example 1,* $\boldsymbol{\theta}_u$ *is an optimal solution to* $\mathbf{DRIO}(\mathcal{D})$.

Proof. As shown in the proof of Lemma 7, when $\alpha \in (0, \pi/2)$, the $\mathbf{RFO}\left(\mathcal{C}(\boldsymbol{\theta}_u, \alpha), u\right)$ has a unique optimal solution $(0, 1)$. So $\bar{\mathbf{x}}_{\mathrm{CIO}}(u) = (0, 1)$ almost surely, when $\bar{\boldsymbol{\theta}}_N = \boldsymbol{\theta}_u$ and $\alpha \in (0, \pi/2)$. It is easy to verify that, when $\hat{\boldsymbol{\theta}} \in \boldsymbol{\Theta}_1 := \{(\cos\delta, \sin\delta) \,|\, \delta \in (0, \delta_u]\}$, we have $\hat{\mathbf{x}}_k = \tilde{\mathbf{x}}(\hat{\boldsymbol{\theta}}, u) = (0, 1)$

almost surely; When $\hat{\boldsymbol{\theta}} \in \boldsymbol{\Theta}_2 := \{(\cos\delta, \sin\delta) \,|\, \delta \in (\delta_u, \pi/2)\}$, we have $\hat{\mathbf{x}}_k = \tilde{\mathbf{x}}(\hat{\boldsymbol{\theta}}, u) = (u, 0)$ almost surely. Hence, we have

$$\mathrm{POG}(\bar{\mathbf{x}}_{\mathrm{CIO}}) = \int_0^{\delta_u} \frac{\pi}{2} \times 0 \; d\delta + \int_{\delta_u}^{\pi/2} \frac{\pi}{2} \times \sin\delta \; d\delta = -\frac{\pi}{2}\cos\delta \Big|_{\delta_u}^{\pi/2} = \frac{\pi}{2\sqrt{1+u^2}} < \frac{\pi}{2\sqrt{2}}. \quad (26)$$

The inequality holds because $u > 1$.

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

Lemma 6 shows that, in Example 1, the estimated parameter from $\mathbf{DRIO}(\mathcal{D})$ may still be $\boldsymbol{\theta}_u$. Hence, the decision policy is identical to $\bar{\mathbf{x}}_{\mathrm{IO}}$ whose AOG and POG can be unbounded. The fundamental reason behind these negative results is the misalignment between the sub-optimality loss and the evaluation metrics. Achieving a low sub-optimality loss means that the suggested and observed decisions are of similar quality as evaluated using the estimated parameters. However, this does not speak to the similarity between these two decisions with respect to the DM's perceived parameters (POG) or the ground-truth parameters (AOG). Therefore, the out-of-sample guarantees on the sub-optimality loss do not translate into bounded AOG or POG.

A.4 Proof of Lemma 2

We consider the AOG and POG of conformal IO separately in the following two lemmas.

Lemma 7. *In Example 1, let $\bar{\mathbf{x}}_{\mathrm{CIO}}(u)$ be an optimal solution to $\mathbf{RFO}\left(\mathcal{C}(\bar{\boldsymbol{\theta}}_N, \alpha), u\right)$ where $\bar{\boldsymbol{\theta}}_N$ is an optimal solution to $\mathbf{IO}(\mathcal{D})$ with the sub-optimality loss (6). When $\alpha \in (0, \pi/2)$, we have $\mathbb{P}\left[\mathrm{AOG}(\mathbf{x}_{\mathrm{CIO}}) = 0\right] \to 1$ as $N \to \infty$.*

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

$$R(\mathbf{x}) := \max_{\boldsymbol{\theta} \in \mathcal{C}(\boldsymbol{\theta}_u, \alpha)} \theta_1 x_1 + \theta_2 x_2. \quad (27)$$

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

It is easy to verify that

$$R(\mathbf{x}_1) = \begin{cases} \sin(\delta_u + \alpha), & \text{if } \alpha \in (0, \pi/2 - \delta_u], \\ 1, & \text{if } \alpha \in (\pi/2 - \delta_u, \pi/2). \end{cases} \quad (28)$$

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

$$R(\mathbf{x}_2) = \begin{cases} 2\sin(\delta_u + \alpha), & \text{if } \alpha \in (0, \pi/2 - \delta_u], \\ 2, & \text{if } \alpha \in (\pi/2 - \delta_u, \pi/2). \end{cases} \quad (29)$$

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

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

$$R(\mathbf{x}_3) = \begin{cases} u\cos(\delta_u - \alpha), & \text{if } \alpha \in (0, \delta_u], \\ u, & \text{if } \alpha \in (\delta_u, \pi/2). \end{cases} \quad (30)$$

https://doi.org/10.52202/079017-2029

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

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

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

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

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

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

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

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

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

$$R(\mathbf{x}_4) = \max_{\delta \in \mathcal{C}(\delta_u, \alpha)} u\cos\delta + 2\sin\delta. \tag{31}$$

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

$$R(\mathbf{x}_4) = \max_{\delta \in \mathcal{C}(\delta_u, \alpha)} u\cos\delta + 2\sin\delta \geq u\cos\delta_1^* + 2\sin\delta_1^* > \sin\delta_1^* = R(\mathbf{x}_1). \tag{32}$$

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

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

□

Lemma 8. *In Example 1, let $\bar{\mathbf{x}}_{\mathrm{CIO}}(u)$ be an optimal solution to* **RFO** $\left(\mathcal{C}(\bar{\boldsymbol{\theta}}_N, \alpha), u\right)$ *where $\bar{\boldsymbol{\theta}}_N$ is an optimal solution to* $\mathbf{IO}(\mathcal{D})$ *with the sub-optimality loss (6). When $\alpha \in (0, \pi/2)$, we have $\mathbb{P}\left[\mathrm{POG}(\bar{\mathbf{x}}_{\mathrm{CIO}}) < \pi/2\sqrt{2}\right] \to 1$ as $N \to \infty$.*

Proof. As shown in the proof of Lemma 7, when $\alpha \in (0, \pi/2)$, the **RFO** $\left(\mathcal{C}(\boldsymbol{\theta}_u, \alpha), u\right)$ has a unique optimal solution $(0, 1)$. So $\bar{\mathbf{x}}_{\mathrm{CIO}}(u) = (0, 1)$ almost surely, when $\bar{\boldsymbol{\theta}}_N = \boldsymbol{\theta}_u$ and $\alpha \in (0, \pi/2)$. It is easy to verify that, when $\hat{\boldsymbol{\theta}} \in \boldsymbol{\Theta}_1 := \{(\cos\delta, \sin\delta) \,|\, \delta \in (0, \delta_u]\}$, we have $\hat{\mathbf{x}}_k = \tilde{\mathbf{x}}(\hat{\boldsymbol{\theta}}, u) = (0, 1)$ almost surely; When $\hat{\boldsymbol{\theta}} \in \boldsymbol{\Theta}_2 := \{(\cos\delta, \sin\delta) \,|\, \delta \in (\delta_u, \pi/2)\}$, we have $\hat{\mathbf{x}}_k = \tilde{\mathbf{x}}(\hat{\boldsymbol{\theta}}, u) = (u, 0)$ almost surely. Hence, we have

$$\mathrm{POG}(\bar{\mathbf{x}}_{\mathrm{CIO}}) = \int_0^{\delta_u} \frac{\pi}{2} \times 0 \, d\delta + \int_{\delta_u}^{\pi/2} \frac{\pi}{2} \times \sin\delta \, d\delta = -\frac{\pi}{2} \cos\delta \Big|_{\delta_u}^{\pi/2} = \frac{\pi}{2\sqrt{1+u^2}} < \frac{\pi}{2\sqrt{2}}. \quad (33)$$

The inequality holds because $u > 1$.

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

B Proof of Statements in Section 4

B.1 Definitions

Definition 5 (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*

$$\hat{\mathfrak{R}}_{\mathcal{D}}(\mathcal{F}) := \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{f \in \mathcal{F}} \frac{1}{N} \sum_{i \in [N]} \sigma_i f(Z_i)\right] \quad (34)$$

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 6 (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 $\mathbb{P}$:*

$$\mathfrak{R}_N(\mathcal{F}) := \mathbb{E}_{\mathcal{D} \sim \mathbb{P}^N}\left[\hat{\mathfrak{R}}_{\mathcal{D}}(\mathcal{F})\right] \quad (35)$$

Definition 7 (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*

$$\Pi_{\mathcal{H}}(N) := \max_{(Z_1, Z_2, \ldots, Z_N) \in \mathcal{Z}^N} |\{(h(Z_1), h(Z_2), \ldots, h(Z_N)) \,|\, h \in \mathcal{H}\}| \quad (36)$$

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

B.2 Useful Lemmas

Lemma 9 (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*

$$\mathfrak{R}_N(\mathcal{H}) \leq \sqrt{\frac{2 \log \Pi_{\mathcal{H}}(N)}{N}}. \quad (37)$$

Lemma 10 (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*

$$\sup_{f \in \mathcal{F}} \left| \frac{1}{N} \sum_{i \in [N]} f(X_i) - \mathbb{E}\left[f(X_i)\right] \right| \leq 2\mathfrak{R}_N(\mathcal{F}) + \delta \quad (38)$$

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

https://doi.org/10.52202/079017-2029

B.3 Proof of Theorem 1

Proof. We first present the extensive formulation of Problem (10). For convenience, we define $\hat{\boldsymbol{\Theta}}_k := \boldsymbol{\Theta}^{\text{OPT}}(\mathbf{u}_k, \hat{\mathbf{x}}_k)$ for any $k \in [N]$. When $\alpha \in [0, \pi]$, $\cos\alpha$ is a strictly decreasing in α. Therefore, minimizing α is equivalent to maximizing the value of $\cos\alpha$. We can replace the decision variable α in Problem (10) with a new decision variable $c := \cos\alpha$ with an additional constraint t with $-1 \leq c \leq 1$. In addition, we introduce a new set of decision variables $y_k \in \{0, 1\}$ that indicate if $\hat{\boldsymbol{\Theta}}_k$ intersects with the learned uncertainty set ($= 1$) or not ($= 0$) for any $k \in \mathcal{K}_{\text{val}}$. Problem (10) can be presented as follows.

$$\underset{c, \{\boldsymbol{\theta}_k\}_{k\in\mathcal{K}_{\text{val}}}, \{y_k\}_{k\in\mathcal{K}_{\text{val}}}}{\text{maximize}} \quad c \tag{39a}$$

$$\text{subject to} \quad \hat{\mathbf{x}}_k \in \mathcal{X}^{\text{OPT}}(\boldsymbol{\theta}_k, \mathbf{u}_k), \quad \forall k \in \mathcal{K}_{\text{val}} \tag{39b}$$

$$\boldsymbol{\theta}_k^\intercal \bar{\boldsymbol{\theta}} \geq c + 2(y_k - 1), \quad \forall k \in \mathcal{K}_{\text{val}} \tag{39c}$$

$$\sum_{k\in\mathcal{K}_{\text{val}}} y_k \geq \lceil \gamma(N_{\text{val}} + 1) \rceil \tag{39d}$$

$$\|\boldsymbol{\theta}_k\|_2 = 1, \quad \forall k \in \mathcal{K}_{\text{val}} \tag{39e}$$

$$-1 \leq c \leq 1 \tag{39f}$$

$$y_k \in \{0, 1\}, \quad \forall k \in \mathcal{K}_{\text{val}}. \tag{39g}$$

Constraints (39b) ensure that $\boldsymbol{\theta}_k$ is a member of $\hat{\boldsymbol{\Theta}}_k$ for any $k \in \mathcal{K}_{\text{val}}$. Constraints (39c) decide if $\boldsymbol{\theta}_k$ should be taken into account when calculating the maximal cosine value c based on if $\hat{\boldsymbol{\Theta}}_k$ intersects with $\mathcal{C}$. Constraint (39d) ensures that $\mathcal{C}$ intersects with at least $\lceil \gamma(N_{\text{val}} + 1) \rceil$ inverse feasible sets. Constraint (39e) enforces $\boldsymbol{\theta}_k$ to be on the unit sphere as defined in Equation (9). Constraints (39f)–(39g) specify the ranges of the decision variables.

Observing that the objective of Problem (39) is to maximize c and that decision variables $\boldsymbol{\theta}_k$ of different data points only interact in Constraints (39c). We can re-write Problem (39) as

$$\text{maximize} \quad c \tag{40a}$$

$$\text{subject to} \quad c \leq c_k - 2(y_k - 1), \quad \forall k \in \mathcal{K}_{\text{val}} \tag{40b}$$

$$\sum_{k\in\mathcal{K}_{\text{val}}} y_k \geq \lceil \gamma(N_{\text{val}} + 1) \rceil \tag{40c}$$

$$-1 \leq c \leq 1 \tag{40d}$$

$$y_k \in \{0, 1\}, \quad \forall k \in \mathcal{K}_{\text{val}}, \tag{40e}$$

where

$$c_k := \underset{\boldsymbol{\theta}_k}{\text{maximize}} \quad \boldsymbol{\theta}_k^\intercal \bar{\boldsymbol{\theta}} \tag{41a}$$

$$\text{subject to} \quad \hat{\mathbf{x}}_k \in \mathcal{X}^{\text{OPT}}(\boldsymbol{\theta}_k, \mathbf{u}_k) \tag{41b}$$

$$\|\boldsymbol{\theta}_k\|_2 \leq 1. \tag{41c}$$

Note that we replace Constraints (39e) with Constraints (41c) because the objective of Problem (41) 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 (40a) is to set $y_k = 1$ for all k such that $c_k \geq \Gamma_\tau\left(\{c_k\}_{k\in\mathcal{K}_{\text{val}}}\right)$ and $y_k = 0$ otherwise. Therefore, the optimal objective value of Problem (40a) is $c = \Gamma_\tau\left(\{c_k\}_{k\in\mathcal{K}_{\text{val}}}\right)$ corresponding to $\alpha_\gamma = \arccos\Gamma_\tau\left(\{c_k\}_{k\in\mathcal{K}_{\text{val}}}\right)$. □

B.4 Proof of Lemma 3

Proof. 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. □

B.5 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

$$A_{\bar{\boldsymbol{\theta}}}(\hat{\mathbf{x}}, \mathbf{u}) := \underset{\boldsymbol{\theta}}{\text{maximize}} \quad \boldsymbol{\theta}^\intercal \bar{\boldsymbol{\theta}} \tag{42a}$$

$$\text{subject to} \quad \boldsymbol{\theta} \in \boldsymbol{\Theta}^{\text{OPT}}(\hat{\mathbf{x}}, \mathbf{u}) \tag{42b}$$

$$\|\boldsymbol{\theta}\|_2 \leq 1. \tag{42c}$$

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

$$\mathbb{P}\left(\boldsymbol{\Theta}^{\text{OPT}}(\hat{\mathbf{x}}, \mathbf{u}) \cap \mathcal{C}(\bar{\boldsymbol{\theta}}, \alpha) \neq \varnothing\right) = \mathbb{P}\left(A_{\bar{\boldsymbol{\theta}}}(\hat{\mathbf{x}}, \mathbf{u}) \geq \Gamma^\tau(\mathcal{A})\right). \tag{43}$$

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 (43). We note that all the probabilities are over the joint distribution of $\mathcal{D}_{\text{val}}$ and the new sample, i.e. $\mathcal{D}'$.

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

$$\mathcal{H} = \left\{h(z, \alpha) = \mathbb{1}\left[\boldsymbol{\Theta}^{\text{OPT}}(\hat{\mathbf{x}}, \mathbf{u}) \cap \mathcal{C}(\bar{\boldsymbol{\theta}}, \alpha)\right] \,\middle|\, \alpha \in (0, \pi)\right\}. \tag{44}$$

Let $\Pi_{\mathcal{H}}$ denote the growth function of $\mathcal{H}$ as defined in Definition 7. It is easy to verify that

$$\Pi_{\mathcal{H}}(N_{\text{val}}) = N_{\text{val}} + 1 \tag{45}$$

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 9, we have

$$\mathfrak{R}_{N_{\text{val}}}(\mathcal{H}) \leq \sqrt{\frac{2\log(N_{\text{val}} + 1)}{N_{\text{val}}}} \tag{46}$$

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 6.

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

$$\frac{1}{N_{\text{val}}} \sum_{k\in\mathcal{K}_{\text{val}}} h(z_k, \alpha) = \frac{1}{N_{\text{val}}} \sum_{k\in\mathcal{K}_{\text{val}}} \mathbb{1}\left[\boldsymbol{\Theta}^{\text{OPT}}(\hat{\mathbf{x}}_k, \mathbf{u}_k) \cap \mathcal{C}(\bar{\boldsymbol{\theta}}, \alpha)\right] = \frac{\lceil \gamma(N_{\text{val}} + 1) \rceil}{N_{\text{val}}}, \tag{47}$$

https://doi.org/10.52202/079017-2029

so we have

$$\gamma \leq \frac{1}{N_{\text{val}}} \sum_{k \in \mathcal{K}_{\text{val}}} h(z_k, \alpha) \leq \gamma + \frac{2}{N_{\text{val}}}. \tag{48}$$

The second inequality holds because

$$\frac{\lceil \gamma(N_{\text{val}}+1) \rceil}{N_{\text{val}}} = \frac{\lfloor \gamma N_{\text{val}} \rfloor + \lceil \gamma N_{\text{val}} - \lfloor \gamma N_{\text{val}} \rfloor + \gamma \rceil}{N_{\text{val}}} \leq \frac{\gamma N_{\text{val}} + \lceil \gamma N_{\text{val}} - \lfloor \gamma N_{\text{val}} \rfloor + \gamma \rceil}{N_{\text{val}}} \leq \gamma + \frac{2}{N_{\text{val}}}. \tag{49}$$

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 $\mathbb{E}[h(z, \alpha)]$, which can be interpreted as $\mathbb{P}\left(\mathbf{\Theta}^{\text{OPT}}(\hat{\mathbf{x}}, \mathbf{u}) \cap \mathcal{C}(\bar{\boldsymbol{\theta}}, \alpha)\right)$ 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 10, we have, with probability at least $\delta = 1 - 1/N_{\text{val}}$,

$$\left| \frac{1}{N_{\text{val}}} \sum_{k \in \mathcal{K}_{\text{val}}} h(z_k, \alpha) - \mathbb{E}[h(z, \alpha)] \right| \leq 2\mathfrak{R}_{N_{\text{val}}}(\mathcal{H}) + \sqrt{\frac{2 \log N_{\text{val}}}{N_{\text{val}}}}. \tag{50}$$

By combing (46)–(50), we have, with probability at least $1 - 1/N_{\text{val}}$,

$$\left| \mathbb{P}\left(\mathbf{\Theta}^{\text{OPT}}(\hat{\mathbf{x}}, \mathbf{u}) \cap \mathcal{C}(\bar{\boldsymbol{\theta}}, \alpha)\right) - \gamma \right| \leq \sqrt{\frac{8 \log(N_{\text{val}}+1) + 2 \log N_{\text{val}}}{N_{\text{val}}}} + \frac{2}{N_{\text{val}}}. \tag{51}$$

□

B.6 Proof of Theorem 3

We bound the AOG and POG of conformal IO separately in the following two propositions.

Proposition 2 (Conformal IO Achieves Bounded POG)**.** *Let $\bar{\mathbf{x}}_{\text{CIO}}(\mathbf{u})$ be an optimal solution to* $\mathbf{RFO}\left(\mathcal{C}(\bar{\boldsymbol{\theta}}, \alpha_1), \mathbf{u}\right)$ *for any $\mathbf{u} \in \mathcal{U}$, where $\bar{\boldsymbol{\theta}} \in \mathbb{R}^d$ and α_1 are chosen such that, for a new sample $(\boldsymbol{\theta}', \mathbf{u}')$ from $\mathbb{P}_{(\boldsymbol{\theta}, \mathbf{u})}$ and $\mathbf{x}' = \tilde{\mathbf{x}}(\boldsymbol{\theta}', \mathbf{u}')$, $\mathbb{P}\left(\mathcal{C}(\bar{\boldsymbol{\theta}}, \alpha_1) \cap \mathbf{\Theta}^{\text{OPT}}(\mathbf{u}', \mathbf{x}') \neq \varnothing\right) = 1$. If Assumptions 2–3 hold, then*

$$\text{POG}(\bar{\mathbf{x}}_{\text{CIO}}) \leq (\eta - 2\cos 2\alpha_1 + 2)\mu + \eta \mu_{\text{CIO}} \tag{52}$$

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

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

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

The first line holds due to Lemma 3. The second line holds due to assumption 2. The third line holds due to the definition of $\bar{\boldsymbol{\theta}}_{\text{CIO}}(u)$. The fourth line holds because $\left(\bar{\mathbf{x}}_{\text{CIO}}(\mathbf{u}), \bar{\boldsymbol{\theta}}_{\text{CIO}}(\mathbf{u})\right)$ is an optimal solution to $\mathbf{RFO}\left(\mathcal{C}(\bar{\boldsymbol{\theta}}, \alpha_1), \mathbf{u}\right)$. The fifth line holds due to Lemma 3. The sixth line holds because both $\hat{\boldsymbol{\theta}}_{\text{CIO}}(\mathbf{u})$ and $\tilde{\boldsymbol{\theta}}$ are in $\mathcal{C}(\bar{\boldsymbol{\theta}}, \alpha_1)$ so the angle between them is no larger than $2\alpha_1$. Since both $\hat{\boldsymbol{\theta}}_{\text{CIO}}(\mathbf{u})$ and $\tilde{\boldsymbol{\theta}}$ are on the unit sphere, the L_2 distance between them are bounded by $2(1-\cos 2\alpha_1)$.

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

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

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

Proposition 3 (Conformal IO Achieves Bounded AOG)**.** *Let $\bar{\mathbf{x}}_{\text{CIO}}(\mathbf{u})$ be an optimal solution to $\mathbf{RFO}\left(\mathcal{C}(\bar{\boldsymbol{\theta}}, \alpha_1), \mathbf{u}\right)$ for any $\mathbf{u} \in \mathcal{U}$, where $\bar{\boldsymbol{\theta}} \in \mathbb{R}^d$ and α_1 are chosen such that, for a new sample $(\boldsymbol{\theta}', \mathbf{u}')$ from $\mathbb{P}_{(\boldsymbol{\theta},\mathbf{u})}$ and $\mathbf{x}' = \tilde{\mathbf{x}}(\boldsymbol{\theta}', \mathbf{u}')$, $\mathbb{P}\left(\mathcal{C}(\bar{\boldsymbol{\theta}}, \alpha_1) \cap \boldsymbol{\Theta}^{\text{OPT}}(\mathbf{u}', \mathbf{x}') \neq \varnothing\right) = 1$. If Assumptions 2–3 hold, then*

$$\text{AOG}(\bar{\mathbf{x}}_{\text{CIO}}) \leq (2 - 2\cos 2\alpha_1 + \eta + \sigma)\mu^* + (\eta + \sigma)\mu_{\text{CIO}} \tag{53}$$

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

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

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

The first line holds because of Lemma 3. The second line holds due to Assumption 3. 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

https://doi.org/10.52202/079017-2029

3. The fifth line holds due to Assumption 2. 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 3. 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

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

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

C Numerical Experiment Details

C.1 Computational Setup

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

C.2 Forward Problems

C.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.

$$\underset{\mathbf{x}}{\text{minimize}} \quad \sum_{(i,j)\in\mathcal{E}} \theta_{ij} x_{ij} \tag{54a}$$

$$\text{subject to} \quad \sum_{(j,i)\in\mathcal{E}^+(i)} x_{ji} - \sum_{(i,j)\in\mathcal{E}^-(i)} x_{ij} = \begin{cases} 1, & \text{if } i = u^d \\ -1, & \text{if } i = o^d \\ 0, & \text{otherwise} \end{cases}, \quad \forall i \in \mathcal{N} \tag{54b}$$

$$x_{ij} \in \{0,1\}, \quad (i,j) \in \mathcal{E}. \tag{54c}$$

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

C.2.2 Knapsack

We 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.

$$\underset{\mathbf{x}}{\text{maximize}} \quad \sum_{i\in[d]} \theta_i x_i \tag{55a}$$

$$\text{subject to} \quad \sum_{i\in[d]} w_i x_i \leq u \tag{55b}$$

$$x_i \in \{0,1\}, \forall i \in [d]. \tag{55c}$$

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

C.3 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 (2024). 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

$$\begin{aligned}
\underset{\boldsymbol{\theta}\in\mathbb{R}^{|\mathcal{E}|},\boldsymbol{\epsilon}\in\mathbb{R}_+^{n_{\text{train}}}}{\text{minimize}} \quad & \frac{1}{N_{\text{train}}}\sum_{k\in\mathcal{K}_{\text{train}}} l_k && \text{(56a)}\\
\text{subject to} \quad & l_k \geq \boldsymbol{\theta}^\intercal\hat{\mathbf{x}}_k - \boldsymbol{\theta}^\intercal\mathbf{x}, \quad \forall \mathbf{x}\in\mathcal{X}_k,\ k\in\mathcal{K}_{\text{train}} && \text{(56b)}\\
& \|\boldsymbol{\theta}-\mathbf{1}\|_1 \leq \frac{|\mathcal{E}|}{4}. && \text{(56c)}
\end{aligned}$$

This problem is initialized without Constraints (56b), which were added iteratively using a cutting-plane method. Specifically, in each iteration, after solving Problem (56), let $\boldsymbol{\theta}'$ and $\{l'_k\}_{k\in\mathcal{K}_{\text{train}}}$ be the optimal solution. For each data point $k\in\mathcal{K}_{\text{train}}$, we solve the following sub-problem

$$\underset{\mathbf{x}_k\in\mathcal{X}(\mathbf{u}_k)}{\text{minimize}} \quad \boldsymbol{\theta}'^\intercal\mathbf{x}_k. \tag{57}$$

Let $\mathbf{x}'_k$ be the optimal solution to the sub-problem. If $l'_k < \boldsymbol{\theta}'^\intercal\hat{\mathbf{x}}_k - \boldsymbol{\theta}'^\intercal\mathbf{x}'_k$, we add the following cut to Problem (56)

$$l_k \geq \boldsymbol{\theta}^\intercal\hat{\mathbf{x}}_k - \boldsymbol{\theta}^\intercal\mathbf{x}'_k. \tag{58}$$

We keep running this procedure until no cut is added to the master Problem (56).

C.4 Solving the Calibration Problem

C.4.1 Shortest Path

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

$$\begin{aligned}
\underset{\boldsymbol{\theta}\in\mathbb{R}^{|\mathcal{E}|},\mathbf{w}\in\mathbb{R}^{\mathcal{N}},\mathbf{v}\in\mathbb{R}_+^{|\mathcal{E}|}}{\text{maximize}} \quad & \bar{\boldsymbol{\theta}}^\intercal\boldsymbol{\theta} && \text{(59a)}\\
\text{subject to} \quad & w_{d_k} - w_{o_k} - \sum_{(i,j)\in\mathcal{E}} v_{ij} = \boldsymbol{\theta}^\intercal\hat{\mathbf{x}}_k && \text{(59b)}\\
& w_j - w_i - v_{ij} \leq c_{ij}, \quad \forall (i,j)\in\mathcal{E} && \text{(59c)}\\
& \|\boldsymbol{\theta}\|_2 \leq 1. && \text{(59d)}
\end{aligned}$$

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

C.4.2 Knapsack

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

$$\begin{aligned}
\underset{\boldsymbol{\theta}\in\mathbb{R}^d}{\text{maximize}} \quad & \bar{\boldsymbol{\theta}}^\intercal\boldsymbol{\theta} && \text{(60a)}\\
\text{subject to} \quad & \boldsymbol{\theta}^\intercal\hat{\mathbf{x}}_k \geq \boldsymbol{\theta}^\intercal\mathbf{x}, \quad \forall \mathbf{x}\in\mathcal{X}(\mathbf{u}_k) && \text{(60b)}\\
& \|\boldsymbol{\theta}\|_2 \leq 1. && \text{(60c)}
\end{aligned}$$

https://doi.org/10.52202/079017-2029

We initialize this problem without Constraints (60b). In each iteration, after solving the calibration problem, let $\boldsymbol{\theta}'$ denote the optimal solution. We solve $\mathbf{FO}(\boldsymbol{\theta}', \mathbf{u}_k)$ and let $\mathbf{x}'$ denote the optimal solution. If $\boldsymbol{\theta}'^{\intercal}\mathbf{x}' > \boldsymbol{\theta}'^{\intercal}\hat{\mathbf{x}}_k$, we then add the corresponding cut to the model. We keep running this process until no cut is added.

C.5 Solving the Robust Forward Problem

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

$$\underset{\mathbf{x}\in\mathcal{X}(\mathbf{u})}{\text{minimize}} \ \underset{\boldsymbol{\theta}\in\mathbb{R}^{|\mathcal{E}|}}{\text{maximize}} \quad \boldsymbol{\theta}^{\intercal}\mathbf{x} \tag{61a}$$

$$\text{subject to} \quad \bar{\boldsymbol{\theta}}^{\intercal}\boldsymbol{\theta} \geq \cos(\alpha) \tag{61b}$$

$$\|\boldsymbol{\theta}\|_2 \leq 1. \tag{61c}$$

We initialize this problem as follows.

$$\underset{\mathbf{x}\in\mathcal{X}(\mathbf{u}),\Omega\in\mathbb{R}_+}{\text{minimize}} \quad \Omega \tag{62a}$$

$$\text{subject to} \quad \boldsymbol{\theta}^{\intercal}\mathbf{x} \leq \Omega, \quad \forall \boldsymbol{\theta} \in \tilde{\boldsymbol{\Theta}}. \tag{62b}$$

We initialize $\tilde{\boldsymbol{\Theta}} = \varnothing$. We first solve Problem (62), let $\mathbf{x}'$ and Ω' denote the optimal solution. Then we solve the following sub-problem

$$\underset{\boldsymbol{\theta}\in\mathbb{R}^{|\mathcal{E}|}}{\text{maximize}} \quad \boldsymbol{\theta}^{\intercal}\mathbf{x}' \tag{63a}$$

$$\text{subject to} \quad \bar{\boldsymbol{\theta}}^{\intercal}\boldsymbol{\theta} \geq \cos(\alpha) \tag{63b}$$

$$\|\boldsymbol{\theta}\|_2 \leq 1. \tag{63c}$$

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

D Additional Computational Results

In this section, we conduct additional computational experiments to assess how point estimate quality affects the performance of the Conformal IO pipeline. Specifically, we focus on the knapsack problem, where we randomly generate point estimates with an angular deviation δ from the ground-truth parameter $\boldsymbol{\theta}^*$, with δ serving as a proxy for point estimate quality. We then apply our calibration method to the point estimate and subsequently use the robust forward problem to generate decision recommendations. We vary the value of δ and report the average (std) out-of-sample AOG and POG achieved by our pipeline in Table 3.

Table 3: Mean (std) AOG and POG by conformal IO when varying point estimate quality.

δ	0	$\pi/20$	$\pi/10$	$3\pi/20$	$\pi/5$
POG	8.45 (0.67)	8.29 (0.66)	8.94 (0.79)	9.91 (0.95)	11.58 (1.40)
AOG	0.00 (0.00)	0.16 (0.07)	0.30 (0.19)	0.74 (0.31)	1.38 (0.53)

In general, conformal IO benefits from more accurate point estimates (i.e., when δ is small). However, the trend is not always strictly monotonic. For instance, as δ increases from 0 to $\pi/10$, the POG initially decreases before rising again. While beyond the scope of this work, exploring point estimation methods that optimize for the performance of the downstream robust optimization could be a valuable direction for future research.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Theoretical results are proved in Sections 3 and Section 4 (complete proofs are in the appendix). These results are verified numerically in Section 5.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Limitations and future research directions are discussed in Section 6.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

https://doi.org/10.52202/079017-2029

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: Assumptions are stated and justified in Section 3.1.1. Proofs are in the appendix.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Experiment details are well documented in the paper. We also open source our implementation at `https://anonymous.4open.science/r/ConformalIO-B776`.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Data and source code available at https://anonymous.4open.science/r/ConformalIO-B776.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Yes, this information is documented in Section 5.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

https://doi.org/10.52202/079017-2029

Justification: We report error bars/standard errors for all experimental results in Section 4.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: This information is disclosed in Appendix C.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have reviewed the NeurIPS Code and Ethics and confirm that our paper conform with it.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.

- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: This is a theoretical/methodological paper that presents foundational research in optimization. We do not see a direct path to any negative societal impact.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: This is a theoretical/methodological paper that presents foundational research in optimization. It does not involve any high-risk data/model.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

 https://doi.org/10.52202/079017-2029

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We properly cite code/papers on which our research is built on.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: Our code and data along with documentations are available at `https://anonymous.4open.science/r/ConformalIO-B776`.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: Our paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: Our paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

https://doi.org/10.52202/079017-2029