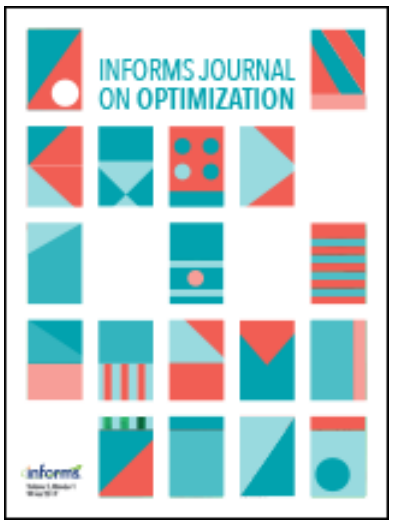


INFORMS Journal on Optimization

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

An Ensemble Learning Framework for Model Fitting and Evaluation in Inverse Linear Optimization

Aaron Babier, Timothy C. Y. Chan, Taewoo Lee, Rafid Mahmood, Daria Terekhov

To cite this article:
Aaron Babier, Timothy C. Y. Chan, Taewoo Lee, Rafid Mahmood, Daria Terekhov (2021) An Ensemble Learning Framework for Model Fitting and Evaluation in Inverse Linear Optimization. INFORMS Journal on Optimization 3(2):119-138. https://doi.org/10.1287/ijoo.2019.0045

http://pubsonline.informs.org/journal/ijoo

INFORMS JOURNAL ON OPTIMIZATION
Vol. 3, No. 2, Spring 2021, pp. 119–138
ISSN 2572-1484 (print), ISSN 2575-1492 (online)

An Ensemble Learning Framework for Model Fitting and Evaluation in Inverse Linear Optimization

Aaron Babier,[a] Timothy C. Y. Chan,[a] Taewoo Lee,[b] Rafid Mahmood,[a] Daria Terekhov[c]

[a] Department of Mechanical and Industrial Engineering, University of Toronto, Toronto, Ontario M5S 3G8, Canada; [b] Department of Industrial Engineering, University of Houston, Houston, Texas 77004; [c] Department of Mechanical, Industrial, and Aerospace Engineering, Concordia University, Montreal, Quebec H3G 1M8, Canada

Contact: ababier@mie.utoronto.ca, https://orcid.org/0000-0002-5949-2500 (AB); tcychan@mie.utoronto.ca, http://orcid.org/0000-0002-4128-1692 (TCYC); tlee6@uh.edu, http://orcid.org/0000-0003-2222-8201 (TL); rafid.mahmood@mail.utoronto.ca, https://orcid.org/0000-0003-1933-066X (RM); daria.terekhov@concordia.ca, https://orcid.org/0000-0003-1585-9806 (DT)

Received: October 18, 2019
Revised: March 15, 2020
Accepted: October 6, 2020
Published Online in Articles in Advance: February 16, 2021

https://doi.org/10.1287/ijoo.2019.0045

Abstract. We develop a generalized inverse optimization framework for fitting the cost vector of a single linear optimization problem given multiple observed decisions. This setting is motivated by ensemble learning, where building consensus from base learners can yield better predictions. We unify several models in the inverse optimization literature under a single framework and derive assumption-free and exact solution methods for each one. We extend a goodness-of-fit metric previously introduced for the problem with a single observed decision to this new setting and demonstrate several important properties. Finally, we demonstrate our framework in a novel inverse optimization-driven procedure for automated radiation therapy treatment planning. Here, the inverse optimization model leverages an ensemble of dose predictions from different machine learning models to construct a consensus treatment plan that outperforms baseline methods. The consensus plan yields better trade-offs between the competing clinical criteria used for plan evaluation.

Funding: This work was supported by Natural Sciences and Engineering Research Council of Canada.
Supplemental Material: The e-companion is available at https://doi.org/10.1287/ijoo.2019.0045.

Keywords: inverse optimization • linear optimization • goodness of fit • model estimation • radiation therapy

1. Introduction

Motivated by the growing availability of data that represent decisions, there is an increasing interest in the use of inverse optimization (IO) to gain insight into decision-generating processes and guide subsequent decision making. Inverse optimization has been used in diverse fields, for example, capturing equilibrium estimates of asset returns for future portfolio optimization (Bertsimas et al. 2012), using past electricity market bids to forecast power consumption (Saez-Gallego et al. 2016), and estimating incentives to design future health insurance subsidies (Aswani et al. 2019).

Inverse optimization determines optimization model parameters to render a data set of observed decisions minimally suboptimal for the model. The literature considers different settings that vary based on data characteristics (e.g., a single feasible decision or multiple points from different instances) or the optimization model (e.g., a linear or convex forward problem). A practitioner also chooses a suboptimality measure to minimize, of which there exist three main variants. The first variant, known as the absolute duality gap, measures the difference between the objective values incurred by data and an imputed optimal value (Bertsimas et al. 2015, Zhao et al. 2015, Saez-Gallego et al. 2016, Esfahani et al. 2018). The second variant, known as the relative duality gap, measures the ratio instead of the absolute difference (Chan et al. 2014, Babier et al. 2018b, Chan et al. 2019). Models using these two measures are referred to as *objective space* models. The third variant is a *decision space* model that minimizes the distance between observed and optimal decisions (Aswani et al. 2018, 2019; Esfahani et al. 2018).

In this paper, we explore an ensemble inverse optimization framework using an arbitrary data set of decisions for a single forward model. Our general motivation is as follows. Consider a single decision-making problem, which we model as a linear program whose cost vector must be estimated. Multiple experts generate decisions for the problem. These experts may be human decision makers with their own parameter estimates or even different heuristics applied to the problem; their proposed decisions may be suboptimal or even infeasible. Using these decisions, we impute a single cost vector that best represents the optimization problem

attempted by the experts (Troutt 1995). We then resolve the problem with the imputed parameter to generate an optimal decision of similar solution quality to the candidate decisions.

Our setting is analogous to ensemble methods in machine learning. Consider the canonical example of a random forest, which averages predictions from a set of decision trees (Breiman 2001). Individual trees train on different subsets of data similar to how individual experts use different experiences to guide their decision making. An ensemble method averages out the biases of the individual models, just as inverse optimization learns an objective that balances the biases of different decision makers (Troutt 1995). Practical evidence from machine learning shows ensemble methods generally outperform base prediction models. We similarly show in our application that ensemble inverse optimization can improve over approaches based on individual decisions.

1.1. Motivating Application

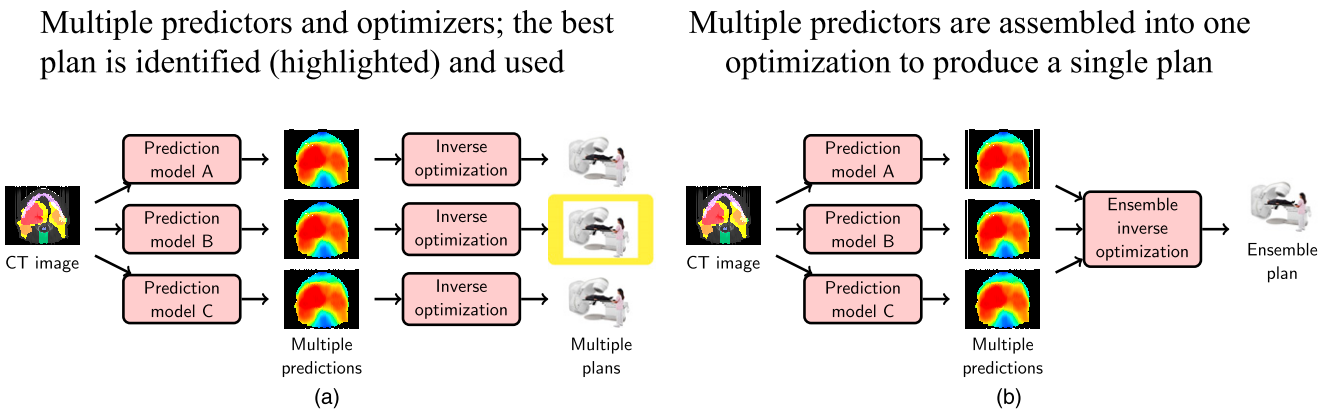
The concrete motivating application in this paper is the automated generation of radiation therapy treatment plans in head-and-neck cancer. Intensity-modulated radiation therapy (IMRT) is one of the most widely-used cancer treatment techniques and is recommended for over 50% of all cancer cases (Delaney et al. 2005). Because there are multiple competing clinical goals in radiation therapy, multiobjective optimization models are used to design clinically acceptable plans. For complex sites such as the head and neck, where there may be multiple targets and critical organs, each patient requires a personalized set of objective function weights, which are typically obtained via manual parameter tuning. This tuning requires going back and forth between a treatment planner and oncologist and may take several days to finalize. This iterative approach, combined with growing patient volumes, leads to strain in the operation of a cancer center and potential delays in treatment for patients (Das et al. 2009).

Knowledge-based planning (KBP) is a machine learning–driven planning procedure that automates the design of personalized treatments for each patient, thereby streamlining planning operations (Sharpe et al. 2014). KBP contains two components: (i) a prediction model that, for a given patient, predicts an appropriate dose distribution (i.e., a function of the planning decision variables) to deliver and (ii) an optimization model that generates a deliverable treatment plan that closely replicates the predicted dose. Although there are various approaches to the optimization stage of KBP, an increasingly popular framework is to use inverse optimization to estimate objective function weights for the original multiobjective planning model by treating the dose predictions as *candidate decisions* obtained from an expert (Chan et al. 2014, Babier et al. 2020). Resolving the planning problem with the imputed parameters then yields an optimal treatment plan.

Modern machine learning has led to a variety of dose prediction techniques that can predict different representations of the dose (McIntosh and Purdie 2016, Kearney et al. 2018, Mahmood et al. 2018). Moreover, treatment plans are clinically evaluated on a set of competing dosimetric criteria, and different prediction models yield plans that overfit to specific criteria. Given a plethora of prediction models where none are strictly dominating, a naive approach may take each prediction, generate a corresponding treatment plan via optimization, and then compare the plans on their dosimetric performance to identify the best plan for a patient (see Figure 1(a)). However, this approach is excessively laborious, and the final plan is still determined from a prediction model that may be overfit to specific clinical criteria.

We propose a natural alternative, which has not been considered previously, to obtain plans that better fit all of the clinical criteria. Analogous to an ensemble learning model combining weak predictors to form a

Figure 1. (Color online) An Existing KBP Pipeline vs. Our Proposed Ensemble Approach



Note. CT, Computerized tomography.

better estimate, we harness a set of prediction models into an ensemble inverse optimization model that yields a single optimal treatment plan (see Figure 1(b)). Differences in the prediction models imitate the biases of different clinical experts that may lead them to suggest different plans for a given patient, even though they all aim to satisfy the same clinical criteria. Our inverse optimization model is a consensus-building treatment planner whose plans compromise between predictions to satisfy aggregate metrics better than any individual model.

1.2. Contributions

Methodologically, we extend the generalized inverse optimization framework of Chan et al. (2019), which considered only a single feasible decision (single point), to the case of multiple observed decisions (ensemble) without any assumption on feasibility. The previous results do not trivially generalize to our assumption-free settings, and we require a suite of new proof techniques to extend the results in these two directions. Our framework is founded on a flexible model template and specializes to several different models via appropriate specification of model hyperparameters. We develop methods to impute the best-fit cost vector for a variety of different loss measures under a general setting (i.e., no assumptions on data), while also introducing efficient techniques under mild application-specific assumptions. Finally, we generalize a previous goodness-of-fit metric for inverse optimization (Chan et al. 2019) to the ensemble case. Together, the model and goodness-of-fit metric form a unified framework for model fitting and evaluation in inverse optimization that is applicable to arbitrary decision data for a single linear optimization problem.

Data-driven inverse optimization has received growing interest, particularly for learning in a class of parametrized convex forward problems (Keshavarz et al. 2011, Bertsimas et al. 2015, Aswani et al. 2018, Esfahani et al. 2018). Contrasting previous papers, which consider a separate feasible set for each decision, our methods are tailored for a single feasible set, given the motivating assumption of different decision makers solving the same forward problem. Our setting admits more efficient solution algorithms that leverage the geometry of linear programming. Furthermore, we develop new bounds relating the performance of different variants that are tighter than bounds adapted from the general convex case to the linear setting (Bertsimas et al. 2015, Esfahani et al. 2018).

The specific contributions of our paper are as follows:

1. We develop an inverse linear optimization approach applicable to decision data sets of arbitrary size and feasibility for a single optimization problem, motivated by ensemble learning methods. This model is expressed in terms of hyperparameters used to derive different model variants.
2. We develop exact and assumption-free solution methods for each of the model variants. Under mild data assumptions, we demonstrate how geometric insights from linear optimization can lead to efficient and even analytic solution approaches.
3. We propose a goodness-of-fit metric measuring the model-data fit between a forward problem and arbitrary decision data. We prove several intuitive properties of the metric, including optimality with respect to the inverse optimization model, boundedness, and monotonicity.
4. We implement the first ensemble-based automated planning pipeline in radiation therapy, using multiple predictions to design a single treatment for head-and-neck cancer patients. Our plans achieve better clinical trade-offs, and our domain-independent goodness-of-fit metric validates our approach.

All proofs can be found in the e-companion.

2. Background on Generalized Inverse Linear Optimization

We first review the formulation and main results from Chan et al. (2019), which introduced an inverse optimization model for linear optimization problems (LPs) unifying both decision and objective space models, but only for a data set with a single feasible observed decision. Let $\mathbf{x}, \mathbf{c} \in \mathbb{R}^n$ denote the decision and cost vectors, respectively, and let $\mathbf{A} \in \mathbb{R}^{m \times n}, \mathbf{b} \in \mathbb{R}^m$ denote the constraint matrix and right-hand side vector, respectively. Let $\mathcal{I} = \{1, \ldots, m\}$ and $\mathcal{J} = \{1, \ldots, n\}$. We refer to the following LP as the forward optimization model:

$$\mathbf{FO}(\mathbf{c}): \quad \underset{\mathbf{x}}{\text{minimize}} \quad \mathbf{c}^\top \mathbf{x}$$
$$\text{subject to} \quad \mathbf{x} \in \mathcal{P} := \{\mathbf{x} \mid \mathbf{A}\mathbf{x} \geq \mathbf{b}\}.$$

We assume $\mathcal{P}$ is full-dimensional and $\mathbf{FO}(\mathbf{c})$ has no redundant constraints. Given a *feasible* decision $\hat{\mathbf{x}} \in \mathcal{P}$, the *single-point* generalized inverse linear optimization problem (Chan et al. 2019) is

$$\begin{aligned} \mathbf{GIO}(\{\hat{\mathbf{x}}\}): \quad \underset{\mathbf{c},\mathbf{y},\epsilon}{\text{minimize}} \quad & \|\epsilon\| && \text{(1a)}\\ \text{subject to} \quad & \mathbf{A}^\top \mathbf{y} = \mathbf{c}, \ \ \mathbf{y} \geq \mathbf{0}, && \text{(1b)}\\ & \mathbf{c}^\top \hat{\mathbf{x}} = \mathbf{b}^\top \mathbf{y} + \mathbf{c}^\top \epsilon, && \text{(1c)}\\ & \|\mathbf{c}\|_N = 1, && \text{(1d)}\\ & \mathbf{c} \in \mathcal{C}, \epsilon \in \mathcal{E}. && \text{(1e)} \end{aligned}$$

Above, $\mathbf{y} \in \mathbb{R}^m$ represents the dual vector for the constraints of the forward problem. Constraints (1b) ensure $\mathbf{y}$ is dual feasible with respect to $\mathbf{c}$. Constraint (1c) connects $\mathbf{c}$ and $\mathbf{y}$ with a perturbation vector $\epsilon \in \mathbb{R}^n$ by enforcing that the pair $(\hat{\mathbf{x}} - \epsilon, \mathbf{y})$ satisfy strong duality with respect to $\mathbf{c}$. Note that these constraints do not imply that the pair is primal-dual optimal (as we have not enforced primal feasibility), but rather that $\hat{\mathbf{x}} - \epsilon$ lies on a supporting hyperplane $\{\mathbf{x} \mid \mathbf{c}^\top \mathbf{x} = \mathbf{b}^\top \mathbf{y}\}$ of the feasible set. Constraint (1d) is a normalization constraint to prevent the trivial solution of $\mathbf{c} = \mathbf{0}$, where $\|\cdot\|_N$ denotes an arbitrary norm that may differ from the one in the objective. Finally, constraints (1e) define application-specific perturbation and cost vector restrictions via the sets $\mathcal{E}$ and $\mathcal{C}$, respectively. We leave the choice of the norm in the objective open. The tuple $(\|\cdot\|, \|\cdot\|_N, \mathcal{C}, \mathcal{E})$ forms the inverse optimization model *hyperparameters*. By selecting them appropriately, $\mathbf{GIO}(\{\hat{\mathbf{x}}\})$ specializes into models that minimize error in objective or decision space.

Although $\mathbf{GIO}(\{\hat{\mathbf{x}}\})$ is nonconvex, it admits a closed-form solution if $\hat{\mathbf{x}} \in \mathcal{P}$, by finding a projection of $\hat{\mathbf{x}}$ to the boundary of $\mathcal{P}$ of minimum distance as measured by $\|\cdot\|$. Specifically, let $\mathcal{H}_i = \{\mathbf{x} \mid \mathbf{a}_i^\top \mathbf{x} = b_i\}$ be the hyperplane corresponding to the *ith* constraint, and let

$$\pi_i(\hat{\mathbf{x}}) = \arg\min_{\mathbf{x} \in \mathcal{H}_i} \|\hat{\mathbf{x}} - \mathbf{x}\| \tag{2}$$

be the projection of $\hat{\mathbf{x}}$ to $\mathcal{H}_i$. The hyperplane projection problem has an analytic solution $\pi_i(\hat{\mathbf{x}}) = \hat{\mathbf{x}} - \mathbf{a}_i^\top \hat{\mathbf{x}} - b_i / \|\mathbf{a}_i\|_D \nu(\mathbf{a}_i)$, where $\|\cdot\|_D$ is the dual norm of $\|\cdot\|$ and $\nu(\mathbf{a}_i) \in \arg\max_{\|\mathbf{v}\|=1} \{\mathbf{v}^\top \mathbf{a}_i\}$ (Mangasarian 1999). This result leads to an analytic optimal solution to $\mathbf{GIO}(\hat{\mathbf{x}})$ when $\hat{\mathbf{x}}$ is feasible.

Theorem 1 (Chan et al. 2019). *Let $\hat{\mathbf{x}} \in \mathcal{P}$, $i^* \in \arg\min_{i \in \mathcal{I}} \{\mathbf{a}_i^\top \hat{\mathbf{x}} - b_i / \|\mathbf{a}_i\|_D\}$, and $\mathbf{e}_i$ be the i*th *unit vector. There exists an optimal solution to $\mathbf{GIO}\hat{\mathbf{x}}$ of the form*

$$(\mathbf{c}^*, \mathbf{y}^*, \epsilon^*) = \left(\frac{\mathbf{a}_{i^*}}{\|\mathbf{a}_{i^*}\|_N}, \frac{\mathbf{e}_{i^*}}{\|\mathbf{a}_{i^*}\|_N}, \hat{\mathbf{x}} - \pi_{i^*}(\hat{\mathbf{x}}) \right). \tag{3}$$

If $\hat{\mathbf{x}} \in \mathcal{P}$, then, by Theorem 1, an optimal cost vector describes a supporting hyperplane (i.e., $\{\mathbf{x} \mid \mathbf{c}^{*\top} \mathbf{x} = \mathbf{b}^\top \mathbf{y}^*\}$) that also corresponds to a constraint of the forward problem.

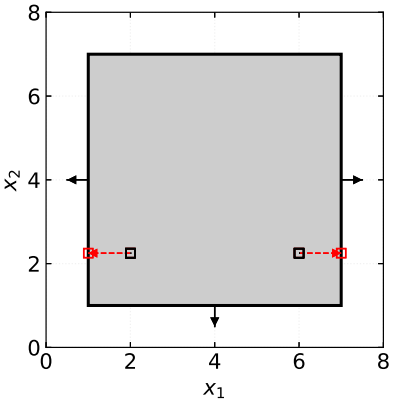
3. Generalized Inverse Linear Optimization with an Ensemble of Decisions

We extend $\mathbf{GIO}\hat{\mathbf{x}}$ to the case of multiple observed decisions with no data assumptions. Let $\hat{\mathcal{X}} = \{\hat{\mathbf{x}}_1, \ldots, \hat{\mathbf{x}}_Q\}$ be a data set, that is, an ensemble of Q observed decisions, indexed by $\mathcal{Q} = \{1, \ldots, Q\}$. We seek to impute a single cost vector $\mathbf{c}^*$ that minimizes the aggregate loss over all decisions.

Given that Theorem 1 admits an analytic solution, one computationally desirable approach may be to solve $\mathbf{GIO}\hat{\mathbf{x}}_q$ for each $\hat{\mathbf{x}}_q$ and impute a set of cost vector estimates. We may then consider classical ensemble methods like a random forest, which average weak predictions (Breiman 2001). However, such a method applied to our setting effectively ignores the geometry of $\mathcal{P}$, which provides useful information in the estimation of a single cost vector. For example, naively averaging the set of cost vectors to obtain a consensus may lead to pathological outcomes.

Example 1. Let $\mathbf{FO}(\mathbf{c}): \min_{\mathbf{x}} \{c_1 x_1 + c_2 x_2 \mid x_1 \leq 7, \ x_2 \leq 7, \ x_1 \geq 1, \ x_2 \geq 1\}$, and consider two points $\hat{\mathbf{x}}_1 = (2, 2.5)$ and $\hat{\mathbf{x}}_2 = (6, 2.25)$. Solving $\mathbf{GIO}(\{\hat{\mathbf{x}}_1\})$ and $\mathbf{GIO}(\{\hat{\mathbf{x}}_2\})$ yields cost vectors $(-1, 0)$ and $(1, 0)$, respectively, with an average cost vector $\bar{\mathbf{c}} = \mathbf{0}$. Note that a more intuitive best-fit cost vector would be $\mathbf{c}^* = (0, -1)$, pointing to the bottom facet of $\mathcal{P}$. Figure 2 illustrates this example.

Figure 2. (Color online) Illustration of Example 1



Notes. The feasible set for **FO**(**c**) is shaded. The squares on the interior and boundary are $\hat{\mathbf{x}}_q$ and $\hat{\mathbf{x}}_q - \epsilon_q^*$ from solving **GIO**($\{\hat{\mathbf{x}}_q\}$) independently. The solid arrows are potential cost vectors.

Instead, we design an ensemble inverse optimization model to minimize the aggregate error induced by all points with respect to a single imputed cost vector. We introduce perturbation vectors ϵ_q for every $q \in \mathcal{Q}$ and form our problem:

$$\mathbf{GIO}(\hat{\mathcal{X}}): \quad \underset{\mathbf{c},\mathbf{y},\epsilon_1,\ldots,\epsilon_Q}{\text{minimize}} \quad \sum_{q=1}^{Q} \|\epsilon_q\| \tag{4a}$$

$$\text{subject to} \quad \mathbf{A}^\top \mathbf{y} = \mathbf{c}, \ \ \mathbf{y} \geq \mathbf{0}, \tag{4b}$$

$$\mathbf{c}^\top \hat{\mathbf{x}}_q = \mathbf{b}^\top \mathbf{y} + \mathbf{c}^\top \epsilon_q, \quad \forall q \in \mathcal{Q}, \tag{4c}$$

$$\|\mathbf{c}\|_N = 1, \tag{4d}$$

$$\mathbf{c} \in \mathcal{C}, \epsilon_q \in \mathcal{E}_q, \quad \forall q \in \mathcal{Q}. \tag{4e}$$

Constraints (4b) and (4d) are carried from the single-point model, whereas (4c) and (4e) are ensemble extensions of (1c) and (1e), respectively, ensuring that for each $q \in \mathcal{Q}$, the data points $\hat{\mathbf{x}}_q$ achieve strong duality with respect to $\mathbf{c}$ after being perturbed by $\epsilon_q \in \mathcal{E}_q$. The objective minimizes the sum of the norms of the individual perturbation vectors. Note that this problem is nonconvex because of the bilinear terms in (4c) and the normalization constraint (4d). We first show that **GIO**($\hat{\mathcal{X}}$) specializes to objective and decision space variants, before developing tailored solution methods.

3.1. Objective Space

Inverse linear optimization in the objective space is based on the premise that suboptimal observed decisions are characterized by suboptimal objective values. Consider the dual problem for **FO**(**c**). For each decision $\hat{\mathbf{x}}_q$, the corresponding duality gap is a distance measure between the objective value of $\hat{\mathbf{x}}_q$ and the optimal value of the dual problem. By choosing the norm in the objective (4a) and the sets $\mathcal{E}_q$ for each $q \in \mathcal{Q}$ appropriately, the problem is transformed to measure a function of the duality gap. We consider two objective space models, the absolute and relative duality gaps.

3.1.1. Absolute Duality Gap. The absolute duality gap method minimizes the aggregate duality gap between the primal objectives of each decision and the imputed dual optimal value:

$$\mathbf{GIO}_{\mathrm{A}}(\hat{\mathcal{X}}): \quad \underset{\mathbf{c},\mathbf{y},\epsilon_1,\ldots,\epsilon_Q}{\text{minimize}} \quad \sum_{q=1}^{Q} |\epsilon_q| \tag{5a}$$

$$\text{subject to} \quad \mathbf{A}^\top \mathbf{y} = \mathbf{c}, \ \ \mathbf{y} \geq \mathbf{0}, \tag{5b}$$

$$\mathbf{c}^\top \hat{\mathbf{x}}_q = \mathbf{b}^\top \mathbf{y} + \epsilon_q, \quad \forall q \in \mathcal{Q}, \tag{5c}$$

$$\|\mathbf{c}\|_N = 1. \tag{5d}$$

This model specializes $\mathbf{GIO}(\hat{\mathcal{X}})$ by measuring error in terms of scalar duality gap variables. We show that it can be recovered from $\mathbf{GIO}(\hat{\mathcal{X}})$ with an appropriate choice of model hyperparameters.

Proposition 1. *Let $\mu(\mathbf{c}) \in \mathbb{R}^n$ be a parameter satisfying $\|\mu(\mathbf{c})\|_\infty = 1$ and $\mu(\mathbf{c})^\top \mathbf{c} = 1$. A solution $(\mathbf{c}^*, \mathbf{y}^*, \epsilon_1^*, \ldots, \epsilon_Q^*)$ is optimal to $\mathbf{GIO}_{\mathrm{A}}(\hat{\mathcal{X}})$ if and only if $(\mathbf{c}^*, \mathbf{y}^*, \epsilon_1^* \mu(\mathbf{c}^*), \ldots, \epsilon_Q^* \mu(\mathbf{c}^*))$ is optimal to $\mathbf{GIO}(\hat{\mathcal{X}})$ with hyperparameters $(\|\cdot\|, \|\cdot\|_N, \mathcal{C}, \mathcal{E}_1, \ldots, \mathcal{E}_Q) = (\|\cdot\|_\infty, \|\cdot\|_N, \mathbb{R}^n, \{\epsilon_1 \mu(\mathbf{c})\}, \ldots, \{\epsilon_Q \mu(\mathbf{c})\})$.*

Proposition 1 shows that the specialization of $\mathbf{GIO}(\hat{\mathcal{X}})$ to $\mathbf{GIO}_{\mathrm{A}}(\hat{\mathcal{X}})$ depends on each ϵ_q being a rescaling of some $\mu(\mathbf{c})$ that is dependent only on the cost vector. Note that $\mu(\mathbf{c})$ is only a vehicle to aid in the specialization of $\mathbf{GIO}(\hat{\mathcal{X}})$, and is useful to interpret solutions of $\mathbf{GIO}_{\mathrm{A}}(\hat{\mathcal{X}})$ in the context of $\mathbf{GIO}(\hat{\mathcal{X}})$. For all $\mathbf{c}$ satisfying $\|\mathbf{c}\|_N = 1$, $\mu(\mathbf{c})$ must satisfy $\|\mu(\mathbf{c})\|_\infty = 1$ and $\mu(\mathbf{c})^\top \mathbf{c} = 1$. Given a specific $\|\cdot\|_N$, we can propose a structured $\mu(\mathbf{c})$. For example, if $\|\cdot\|_N = \|\cdot\|_1$, let $\mu(\mathbf{c}) = \operatorname{sgn}(\mathbf{c})$ be the sign vector of $\mathbf{c}$, ensuring that the conditions on $\mu(\mathbf{c})$ are satisfied for all $\mathbf{c}$ with $\|\mathbf{c}\|_1 = 1$. If $\|\cdot\|_N = \|\cdot\|_\infty$, let $\mu(\mathbf{c}) = \operatorname{sgn}(c_{j^*})\mathbf{e}_{j^*}$ be the j^*th unit vector, where $j^* \in \arg\max_{j \in \mathcal{J}} \{|c_j|\}$.

3.1.1.1. General Solution Method. Because the normalization constraint is the sole nonconvexity in $\mathbf{GIO}_{\mathrm{A}}(\hat{\mathcal{X}})$, this model can be solved exactly by polyhedral decomposition. The efficiency of this approach depends on the choice of the norm. For example, $2n$ LPs are needed if $\|\cdot\|_N = \|\cdot\|_\infty$.

Theorem 2. *Let $(\mathbf{c}^*, \mathbf{y}^*, \epsilon_1^*, \ldots, \epsilon_Q^*)$ be optimal to $\mathbf{GIO}_{\mathrm{A}}(\hat{\mathcal{X}})$ under $\|\cdot\|_N = \|\cdot\|_\infty$. There exists $j \in \mathcal{J}$ such that $(\mathbf{c}^*, \mathbf{y}^*, \epsilon_1^*, \ldots, \epsilon_Q^*)$ is also optimal to $\mathbf{GIO}_{\mathrm{A}}(\hat{\mathcal{X}}; j)$, defined as*

$$
\begin{aligned}
\mathbf{GIO}_{\mathrm{A}}(\hat{\mathcal{X}}; j): \quad & \underset{\mathbf{c}, \mathbf{y}, \epsilon_1, \ldots, \epsilon_Q}{\text{minimize}} && \sum_{q=1}^{Q} |\epsilon_q| \\
& \text{subject to} && \mathbf{A}^\top \mathbf{y} = \mathbf{c}, \ \ \mathbf{y} \geq \mathbf{0}, \\
& && \mathbf{c}^\top \hat{\mathbf{x}}_q = \mathbf{b}^\top \mathbf{y} + \epsilon_q, \quad \forall q \in \mathcal{Q}, \\
& && (c_j = 1) \vee (c_j = -1), \\
& && |c_k| \leq 1, \quad \forall k \in \mathcal{J} / \{j\}.
\end{aligned} \tag{6}
$$

For each j, the problem $\mathbf{GIO}_{\mathrm{A}}(\hat{\mathcal{X}}; j)$ separates into two LPs (one with the constraint $c_j = 1$ and the other with $c_j = -1$), thus totaling $2n$ LPs. When $\|\cdot\|_N \neq \|\cdot\|_\infty$ in general, an exponential number of LPs may be required. We next discuss special cases where this approach simplifies.

3.1.1.2. Nonnegative Cost Vectors. In many real-world applications, feasible cost vectors should be nonnegative (i.e., $\mathcal{C} \subseteq \mathbb{R}^n_+$). Here, it is advantageous to set $\|\cdot\|_N = \|\cdot\|_1$, because the normalization constraint becomes $\mathbf{c}^\top \mathbf{1} = 1$, and $\mathbf{GIO}_{\mathrm{A}}(\hat{\mathcal{X}})$ simplifies to a single linear optimization problem.

3.1.1.3. Feasible Observed Decisions. Most inverse optimization works focus on the situation where all observed decisions are feasible for the forward model (i.e., $\hat{\mathcal{X}} \subset \mathcal{P}$). In this case, $\hat{\mathcal{X}}$ can be replaced by the singleton $\{\bar{\mathbf{x}}\}$, where $\bar{\mathbf{x}}$ is the centroid of the points in $\hat{\mathcal{X}}$. A similar result was presented in Goli (2015, chapter 4), but for a model with a different normalization constraint that did not prevent trivial solutions. We present the analogous result in the context of our model (5).

Proposition 2. *If $\hat{\mathcal{X}} \subset \mathcal{P}$ and $\bar{\mathbf{x}}$ is the centroid of $\hat{\mathcal{X}}$, $\mathbf{GIO}_{\mathrm{A}}(\hat{\mathcal{X}})$ is equivalent to $\mathbf{GIO}_{\mathrm{A}}(\{\bar{\mathbf{x}}\})$.*

Together, Proposition 2 and Theorem 1 imply that $\mathbf{GIO}_{\mathrm{A}}(\hat{\mathcal{X}})$ is analytically solvable when $\hat{\mathcal{X}} \subset \mathcal{P}$.

3.1.1.4. Infeasible Observed Decisions. Finally, we address scenarios where the observed decisions are all infeasible. We first consider the case where $\hat{\mathcal{X}}$ is a single, infeasible observed decision $\hat{\mathbf{x}}$.

Proposition 3. *Assume $\hat{\mathbf{x}} \notin \mathcal{P}$. Then we have the following:*

1. *If $\hat{\mathbf{x}}$ satisfies $\mathbf{a}_i^\top \hat{\mathbf{x}} > b_i$ for some $i \in \mathcal{I}$, then there also exists $i^* \in \mathcal{I}$ such that $\tilde{\mathbf{y}}$ is*

$$
\tilde{y}_i = \frac{1}{\mathbf{a}_i^\top \hat{\mathbf{x}} - b_i}, \quad \tilde{y}_{i^*} = \frac{1}{b_{i^*} - \mathbf{a}_{i^*}^\top \hat{\mathbf{x}}}, \quad \tilde{y}_k = 0 \quad \forall k \in \mathcal{I} \setminus \{i, i^*\} \tag{7}
$$

and $\tilde{\mathbf{c}} = \mathbf{A}^\top \tilde{\mathbf{y}}$. The corresponding normalized solution $(\mathbf{c}^*, \mathbf{y}^*, \epsilon^*) = (\tilde{\mathbf{c}}/\|\tilde{\mathbf{c}}\|_N, \tilde{\mathbf{y}}/\|\tilde{\mathbf{c}}\|_N, 0)$ is an optimal solution to $\mathbf{GIO}_{\mathrm{A}}(\{\hat{\mathbf{x}}\})$ and the optimal value is 0.

2. *If* $\mathbf{A}\hat{\mathbf{x}} \leq \mathbf{b}$, *there exists* $i^* \in \mathcal{I}$ *such that* (3) *is an optimal solution to* $\mathbf{GIO}_{\mathrm{A}}(\{\hat{\mathbf{x}}\})$.

Proposition 3 provides geometric insights regarding the structure of optimal solutions. In objective space inverse optimization, all points that lie on a level set of a cost vector yield the same duality gap. Recall that the hyperplane $\mathcal{H} = \{\mathbf{x} \mid \mathbf{c}^{*\top}\mathbf{x} = \mathbf{b}^\top \mathbf{y}^*\}$ is a supporting hyperplane of $\mathcal{P}$, or, in other words, a level set of the cost vector with zero duality gap. If $\hat{\mathbf{x}} \notin \mathcal{P}$ but satisfies $\mathbf{a}_i^\top \hat{\mathbf{x}} > b_i$ for some i, then there always exists a supporting hyperplane that intersects with $\hat{\mathbf{x}}$ (e.g., Figure 3(a)). If $\mathbf{A}\hat{\mathbf{x}} \leq \mathbf{b}$, then no such supporting hyperplane exists. However, consider the alternate forward problem $\mathbf{FOA}(\mathbf{c}) := \min_{\mathbf{x}}\{-\mathbf{c}^\top\mathbf{x} \mid \mathbf{A}\mathbf{x} \leq \mathbf{b}\}$ obtained by reversing the signs of all constraints and the cost vector. The single-point inverse problem for $\hat{\mathbf{x}}$ and $\mathbf{FOA}(\mathbf{c})$ is equivalent to the original problem. Because $\hat{\mathbf{x}}$ is feasible for $\mathbf{FOA}(\mathbf{c})$, Theorem 1 applies for $\mathbf{GIO}_{\mathrm{A}}(\{\hat{\mathbf{x}}\})$. Geometrically, the constraints of $\mathbf{FOA}(\mathbf{c})$ describe the nearest supporting hyperplanes of $\mathbf{FO}(\mathbf{c})$. Solving one problem solves the other (e.g., Figure 3(b), where $\hat{\mathbf{x}}$ projects to an infeasible point for $\mathbf{FO}(\mathbf{c})$ with no duality gap). We use this insight to extend Proposition 3, Statement 2, for multiple infeasible decisions. If all data points violate all constraints, then the multipoint problem reduces to a single point.

Corollary 1. *Suppose that* $\mathbf{A}\hat{\mathbf{x}}_q \leq \mathbf{b}$ *for all* $q \in \mathcal{Q}$, *and* $\hat{\mathcal{X}} \subset \mathbb{R}^n \setminus \mathcal{P}$. *Let* $\bar{\mathbf{x}}$ *be the centroid of* $\hat{\mathcal{X}}$. *Then,* $\mathbf{GIO}_{\mathrm{A}}(\hat{\mathcal{X}})$ *for the forward problem* $\mathbf{FO}(\mathbf{c})$ *is equivalent to* $\mathbf{GIO}_{\mathrm{A}}(\{\bar{\mathbf{x}}\})$ *for* $\mathbf{FOA}(\mathbf{c})$.

3.1.2. Relative Duality Gap. The relative duality gap variant minimizes the sum of the ratios between the duality gap for each decision and the imputed dual optimal value for the forward problem:

$$\mathbf{GIO}_{\mathrm{R}}(\hat{\mathcal{X}}): \quad \underset{\mathbf{c},\mathbf{y},\epsilon_1,\ldots,\epsilon_Q}{\text{minimize}} \quad \sum_{q=1}^{Q} |\epsilon_q - 1| \tag{8a}$$

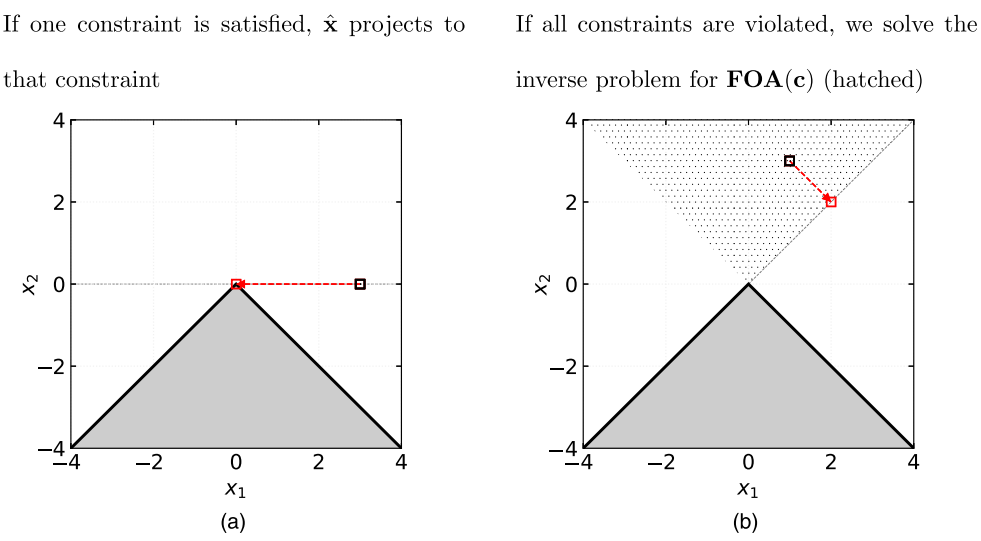
$$\text{subject to} \quad \mathbf{A}^\top\mathbf{y} = \mathbf{c}, \ \ \mathbf{y} \geq \mathbf{0}, \tag{8b}$$

$$\mathbf{c}^\top\hat{\mathbf{x}}_q = \epsilon_q \mathbf{b}^\top\mathbf{y}, \quad \forall q \in \mathcal{Q}, \tag{8c}$$

$$\|\mathbf{c}\|_N = 1. \tag{8d}$$

Duality gap ratio variables ϵ_q replace the perturbation vectors used in the general formulation $\mathbf{GIO}(\hat{\mathcal{X}})$. These variables are well defined except when the imputed forward problem has an optimal value of 0. In this subsection, we assume $\mathbf{b} \neq \mathbf{0}$. Furthermore, note that if $\mathbf{b}^\top\mathbf{y} = \mathbf{0}$ for feasible $\mathbf{y}$, then ϵ_q are free variables; in this

Figure 3. (Color online) Illustration of Proposition 3



Notes. The feasible set for $\mathbf{FO}(\mathbf{c})$ is shaded. The squares on the exterior and boundary are $\hat{\mathbf{x}}$ and $\hat{\mathbf{x}} - \epsilon^*$, respectively. The dashed line is the supporting hyperplane yielding an optimal value of 0.

case, we assume $\epsilon_q := 1$ for all q. First, we show $\mathbf{GIO}_{\mathrm{R}}(\hat{\mathcal{X}})$ can be recovered from $\mathbf{GIO}(\hat{\mathcal{X}})$ with appropriate hyperparameters.

Proposition 4. *Let $\mu(\mathbf{c})$ be a function that satisfies $\|\mu(\mathbf{c})\|_\infty = 1$ and $\mu(\mathbf{c})^\top \mathbf{c} = 1$ for all $\mathbf{c}$. A solution $(\mathbf{c}^*, \mathbf{y}^*, \epsilon_1^*, \ldots, \epsilon_Q^*)$ for which $\mathbf{b}^\top \mathbf{y}^* \neq 0$ is optimal to $\mathbf{GIO}_{\mathrm{R}}(\hat{\mathcal{X}})$ if and only if $(\mathbf{c}^*, \mathbf{y}^*, \mathbf{b}^\top \mathbf{y}^*(\epsilon_1^* - 1)\mu(\mathbf{c}^*), \ldots, \mathbf{b}^\top \mathbf{y}^*(\epsilon_Q^* - 1)\mu(\mathbf{c}^*))$ is optimal to $\mathbf{GIO}(\hat{\mathcal{X}})$ with hyperparameters*

$$\big(\|\cdot\|, \|\cdot\|_N, \mathcal{C}, \mathcal{E}_1, \ldots, \mathcal{E}_Q\big) = \big(\|\cdot\|_\infty / |\mathbf{b}^\top \mathbf{y}^*|, \|\cdot\|_N, \mathbb{R}^n, \{\mathbf{b}^\top \mathbf{y}^*(\epsilon_1 - 1)\mu(\mathbf{c}^*)\}, \ldots, \{\mathbf{b}^\top \mathbf{y}^*(\epsilon_Q - 1)\mu(\mathbf{c}^*)\}\big).$$

3.1.2.1. General Solution Method. Unlike the absolute duality gap problem, which is nonconvex only because of the normalization constraint, $\mathbf{GIO}_{\mathrm{R}}(\hat{\mathcal{X}})$ possesses an additional nonconvexity due to a bilinear term in the duality gap constraint (8c). We first address the bilinearity by introducing three subproblems. We then use polyhedral decomposition to address the normalization constraint.

Proposition 5. *Consider the following three problems:*

$$\begin{aligned}
\mathbf{GIO}_{\mathrm{R}}^{+}(\hat{\mathcal{X}}; K): \quad \min_{\substack{\mathbf{c},\mathbf{y},\\ \epsilon_1,\ldots,\epsilon_Q}} \quad & \sum_{q=1}^{Q} |\epsilon_q - 1| \\
\text{s.t.} \quad & \mathbf{A}^\top \mathbf{y} = \mathbf{c}, \ \ \mathbf{y} \geq \mathbf{0} \\
& \mathbf{c}^\top \hat{\mathbf{x}}_q = \epsilon_q, \quad \forall q \in \mathcal{Q} \\
& \mathbf{b}^\top \mathbf{y} = 1 \\
& \|\mathbf{c}\|_N \geq K,
\end{aligned} \tag{9}$$

$$\begin{aligned}
\mathbf{GIO}_{\mathrm{R}}^{-}(\hat{\mathcal{X}}; K): \quad \min_{\substack{\mathbf{c},\mathbf{y},\\ \epsilon_1,\ldots,\epsilon_Q}} \quad & \sum_{q=1}^{Q} |\epsilon_q - 1| \\
\text{s.t.} \quad & \mathbf{A}^\top \mathbf{y} = \mathbf{c}, \ \ \mathbf{y} \geq \mathbf{0} \\
& \mathbf{c}^\top \hat{\mathbf{x}}_q = -\epsilon_q, \quad \forall q \in \mathcal{Q} \\
& \mathbf{b}^\top \mathbf{y} = -1 \\
& \|\mathbf{c}\|_N \geq K,
\end{aligned} \tag{10}$$

$$\begin{aligned}
\mathbf{GIO}_{\mathrm{R}}^{0}(\hat{\mathcal{X}}; K): \quad \min_{\mathbf{c},\mathbf{y}} \quad & 0 \\
\text{s.t.} \quad & \mathbf{A}^\top \mathbf{y} = \mathbf{c}, \ \ \mathbf{y} \geq \mathbf{0} \\
& \mathbf{c}^\top \hat{\mathbf{x}}_q = 0, \quad \forall q \in \mathcal{Q} \\
& \mathbf{b}^\top \mathbf{y} = 0, \quad \mathbf{y}^\top \mathbf{1} = 1 \\
& \|\mathbf{c}\|_N \geq K.
\end{aligned} \tag{11}$$

Let z^+ be the optimal value of $\mathbf{GIO}_{\mathrm{R}}^{+}(\hat{\mathcal{X}}; K)$ if it is feasible; otherwise, $z^+ = \infty$. Let z^- and z^0 be defined similarly for $\mathbf{GIO}_{\mathrm{R}}^{-}(\hat{\mathcal{X}}; K)$ and $\mathbf{GIO}_{\mathrm{R}}^{0}(\hat{\mathcal{X}}; K)$, respectively. Let $z^* = \min\{z^+, z^-, z^0\}$, and let $(\mathbf{c}^*, \mathbf{y}^*, \epsilon_1^*, \ldots, \epsilon_Q^*)$ be an optimal solution for the corresponding problem. We assume $\epsilon_1^* = \cdots = \epsilon_Q^* = 1$ for $\mathbf{GIO}_{\mathrm{R}}^{0}(\hat{\mathcal{X}}; K)$. Then there exists K such that the optimal value of $\mathbf{GIO}_{\mathrm{R}}(\hat{\mathcal{X}})$ is equal to z^* and an optimal solution to $\mathbf{GIO}_{\mathrm{R}}(\hat{\mathcal{X}})$ is $(\mathbf{c}^*/\|\mathbf{c}^*\|_N, \mathbf{y}^*/\|\mathbf{c}^*\|_N, \epsilon_1^*, \ldots, \epsilon_Q^*)$.

Proposition 5 breaks $\mathbf{GIO}_{\mathrm{R}}(\hat{\mathcal{X}})$ into three cases: $\mathbf{b}^\top \mathbf{y} > 0$, $\mathbf{b}^\top \mathbf{y} < 0$, and $\mathbf{b}^\top \mathbf{y} = 0$. We then normalize $\mathbf{b}^\top \mathbf{y}$ and relax the cost vector normalization constraint (8d). When $\mathbf{b}^\top \mathbf{y} = \mathbf{0}$ for $\mathbf{GIO}_{\mathrm{R}}^{0}(\hat{\mathcal{X}})$, the ϵ_q terms become free variables, motivating $\epsilon_q = 1$ for all q in order to obtain an objective function 0.

There are two issues to note. First, Proposition 5 requires the selection of an appropriate value for the parameter K, which can be accomplished by solving an auxiliary problem (see Section EC.2 of the e-companion for details). Second, formulations (9), (10), and (11) are still nonconvex because of the normalization constraint

$\|\mathbf{c}\|_N \geq K$. As in $\mathbf{GIO}_{\mathrm{A}}(\hat{\mathcal{X}})$, this can be addressed via polyhedral decomposition. For example, if $\|\cdot\|_N = \|\cdot\|_\infty$, $\mathbf{GIO}^{+}_{\mathrm{R}}(\hat{\mathcal{X}}; K)$ decomposes to $2n$ linear programs $\mathbf{GIO}^{+}_{\mathrm{R}}(\hat{\mathcal{X}}; K, j)$ (see Theorem 2):

$$\begin{aligned}
\mathbf{GIO}^{+}_{\mathrm{R}}(\hat{\mathcal{X}}; K, j): \quad \underset{\mathbf{c},\mathbf{y},\epsilon_1,\ldots,\epsilon_Q}{\text{minimize}} \quad & \sum_{q=1}^{Q} |\epsilon_q - 1| \\
\text{subject to} \quad & \mathbf{A}^\top \mathbf{y} = \mathbf{c}, \ \ \mathbf{y} \geq \mathbf{0}, \\
& \mathbf{c}^\top \hat{\mathbf{x}}_q = \epsilon_q, \quad \forall q \in \mathcal{Q}, \\
& \mathbf{b}^\top \mathbf{y} = 1, \\
& (c_j \geq K) \vee (c_j \leq -K). \qquad (12)
\end{aligned}$$

A complete algorithm for solving $\mathbf{GIO}_{\mathrm{R}}(\hat{\mathcal{X}})$ exactly in this assumption-free setting is provided in Section EC.2 of the e-companion. We briefly remark here that an alternative approach is to relax the normalization constraints in formulations (9), (10), and (11). If solving the relaxations yields an optimal $\mathbf{c}^* \neq \mathbf{0}$, then this cost vector can be rescaled as in Proposition 5 (see Corollary EC.1 in the e-companion).

3.1.2.2. Feasible Observed Decisions. As in the absolute duality gap case, the relative duality gap model reduces to a single-point problem, which has an analytic solution according to Theorem 1.

Proposition 6. *If $\hat{\mathcal{X}} \subset \mathcal{P}$ and $\bar{\mathbf{x}}$ is the centroid of $\hat{\mathcal{X}}$, $\mathbf{GIO}_{\mathrm{R}}(\hat{\mathcal{X}})$ is equivalent to $\mathbf{GIO}_{\mathrm{R}}(\{\bar{\mathbf{x}}\})$.*

3.1.2.3. Infeasible Observed Decisions. Proposition 7 is analogous to Proposition 3 and yields an analytic solution for $\mathbf{GIO}_{\mathrm{R}}(\{\hat{\mathbf{x}}\})$ if $\hat{\mathbf{x}} \notin \mathcal{P}$. Corollary 2 extends Proposition 7, Statement 2, to multiple decisions similar to Corollary 1 with Proposition 3. The proofs (omitted) are similar to Proposition 3 and Corollary 1.

Proposition 7. *Assume $\hat{\mathbf{x}} \notin \mathcal{P}$. Then we have the following:*

1. *If $\hat{\mathbf{x}}$ satisfies $\mathbf{a}_i^\top \hat{\mathbf{x}} > b_i$ for some $i \in \mathcal{I}$, then there exists $i^* \in \mathcal{I}$ such that (7) is an optimal solution to $\mathbf{GIO}_{\mathrm{R}}(\{\hat{\mathbf{x}}\})$ and the optimal value is 0.*
2. *If $\mathbf{A}\hat{\mathbf{x}} \leq \mathbf{b}$, there exists $i^* \in \mathcal{I}$ such that (3) is an optimal solution to $\mathbf{GIO}_{\mathrm{R}}(\{\hat{\mathbf{x}}\})$.*

Corollary 2. *Suppose that $\mathbf{A}\hat{\mathbf{x}}_q \leq \mathbf{b}$ for all $q \in \mathcal{Q}$, and let $\bar{\mathbf{x}}$ be the centroid of $\hat{\mathcal{X}}$. Then, $\mathbf{GIO}_{\mathrm{R}}(\hat{\mathcal{X}})$ for the forward problem $\mathbf{FO}(\mathbf{c})$ is equivalent to $\mathbf{GIO}_{\mathrm{R}}(\{\bar{\mathbf{x}}\})$ for $\mathbf{FOA}(\mathbf{c})$.*

3.2. Decision Space

Inverse optimization in the decision space measures error by distance from optimal decisions, rather than objective values. The model identifies a cost vector that produces optimal decisions for the forward problem that are of minimum aggregate distance to the corresponding observed decisions:

$$\begin{aligned}
\mathbf{GIO}_p(\hat{\mathcal{X}}): \quad \underset{\mathbf{c},\mathbf{y},\epsilon_1,\ldots,\epsilon_Q}{\text{minimize}} \quad & \sum_{q=1}^{Q} \|\epsilon_q\|_p && \text{(13a)} \\
\text{subject to} \quad & \mathbf{A}^\top \mathbf{y} = \mathbf{c}, \ \ \mathbf{y} \geq \mathbf{0}, && \text{(13b)} \\
& \mathbf{c}^\top \hat{\mathbf{x}}_q = \mathbf{b}^\top \mathbf{y} + \mathbf{c}^\top \epsilon_q, \quad \forall q \in \mathcal{Q}, && \text{(13c)} \\
& \mathbf{A}(\hat{\mathbf{x}}_q - \epsilon_q) \geq \mathbf{b}, \quad \forall q \in \mathcal{Q}, && \text{(13d)} \\
& \|\mathbf{c}\|_N = 1. && \text{(13e)}
\end{aligned}$$

Model $\mathbf{GIO}_p(\hat{\mathcal{X}})$ resembles $\mathbf{GIO}(\hat{\mathcal{X}})$, except that the objective function is the sum of p-norms ($p \geq 1$), and constraint (13d) is added to enforce primal feasibility of the perturbed decisions $\hat{\mathbf{x}}_q - \epsilon_q$. Unlike in the objective space models, we require primal feasibility because the ϵ_q perturbation vectors have a physical meaning as the distance from observed $\hat{\mathbf{x}}_q$ to optimal $\mathbf{x}_q^*$ decisions. It is straightforward to show $\mathbf{GIO}_p(\hat{\mathcal{X}})$ is a specialization of $\mathbf{GIO}(\hat{\mathcal{X}})$ (proof omitted).

Proposition 8. *A solution $(\mathbf{c}^*, \mathbf{y}^*, \epsilon_1^*, \ldots, \epsilon_Q^*)$ is optimal to $\mathbf{GIO}_p(\hat{\mathcal{X}})$ if and only if it is optimal to $\mathbf{GIO}(\hat{\mathcal{X}})$ with the following hyperparameters: $(\|\cdot\|, \|\cdot\|_N, \mathcal{C}, \mathcal{E}_1, \ldots, \mathcal{E}_Q) = (\|\cdot\|_p, \|\cdot\|_N, \mathbb{R}^n, \{\epsilon_1 \mid \mathbf{A}(\hat{\mathbf{x}}_1 - \epsilon_1) \geq \mathbf{b}\}, \ldots, \{\epsilon_Q \mid \mathbf{A}(\hat{\mathbf{x}}_Q - \epsilon_Q) \geq \mathbf{b}\})$.*

Although $\mathbf{GIO}_p(\hat{\mathcal{X}})$ is nonconvex, we show that an optimal cost vector coincides with one of the constraints (e.g., Theorem 1). However, directly projecting all $\hat{\mathbf{x}}_q$ to a hyperplane may result in projections being infeasible, violating (13d). Thus, we define the *feasible projection problem*:

$$\begin{aligned} \underset{\mathbf{x}}{\text{minimize}} \quad & \|\hat{\mathbf{x}}_q - \mathbf{x}\|_p \\ \text{subject to} \quad & \mathbf{A}\mathbf{x} \geq \mathbf{b}, \\ & \mathbf{a}_i^\top \mathbf{x} = b_i. \end{aligned} \tag{14}$$

Let $\psi_i(\hat{\mathbf{x}}_q)$ be an optimal solution to problem (14), which identifies the closest point in $\mathcal{P}$ to $\hat{\mathbf{x}}_q$ on the hyperplane $\mathcal{H}_i = \{\mathbf{x} \mid \mathbf{a}_i^\top \mathbf{x} = b_i\}$. We first derive a structured optimal solution to $\mathbf{GIO}_p(\hat{\mathcal{X}})$.

Lemma 1. *There exists $i \in \mathcal{I}$ such that an optimal solution to $\mathbf{GIO}_p(\hat{\mathcal{X}})$ is given by*

$$\left(\mathbf{c}^*, \mathbf{y}^*, \epsilon_1^*, \ldots, \epsilon_Q^*\right) = \left(\frac{\mathbf{a}_i}{\|\mathbf{a}_i\|_N}, \frac{\mathbf{e}_i}{\|\mathbf{a}_i\|_N}, \hat{\mathbf{x}}_1 - \psi_i(\hat{\mathbf{x}}_1), \ldots, \hat{\mathbf{x}}_Q - \psi_i(\hat{\mathbf{x}}_Q)\right). \tag{15}$$

The intuition behind Lemma 1 is as follows. Given a feasible set of vectors $\epsilon_1, \ldots, \epsilon_Q$, every observed decision $\hat{\mathbf{x}}_q$ is perturbed by ϵ_q to a point that satisfies both strong duality and primal feasibility. Strong duality implies that $\mathcal{H} = \{\mathbf{x} \mid \mathbf{c}^{*\top}\mathbf{x} = \mathbf{b}^\top \mathbf{y}^*\}$ is a supporting hyperplane, and so $\hat{\mathbf{x}}_q - \epsilon_q$ lies on that supporting hyperplane for all $q \in \mathcal{Q}$. Every feasible solution not of the form (15) must satisfy multiple constraints with equality, and is dominated by solutions that involve the feasible projection to just one of those constraints. Because Lemma 1 holds regardless of the chosen norm and feasibility of the observed decisions, we can show $\mathbf{GIO}_p(\hat{\mathcal{X}})$ can be solved via m convex optimization problems (which become linear with appropriate p-norms).

Theorem 3. *Consider the following optimization problem:*

$$\min_{i \in \mathcal{I}} \quad \min_{\epsilon_{1,i}, \ldots, \epsilon_{Q,i}} \quad \sum_{q=1}^{Q} \|\epsilon_{q,i}\|_p \tag{16a}$$

$$\text{s.t.} \quad \mathbf{A}\left(\hat{\mathbf{x}}_q - \epsilon_{q,i}\right) \geq \mathbf{b}, \quad \forall q \in \mathcal{Q}, \tag{16b}$$

$$\mathbf{a}_i^\top \left(\hat{\mathbf{x}}_q - \epsilon_{q,i}\right) = b_i, \quad \forall q \in \mathcal{Q}. \tag{16c}$$

For each $i \in \mathcal{I}$, let $(\epsilon_{1,i}^*, \ldots, \epsilon_{Q,i}^*)$ denote an optimal solution to the inner optimization problem, and let $i^* \in \arg\min_{i \in \mathcal{I}} \sum_{q=1}^{Q} \|\epsilon_{q,i}^*\|$ denote an optimal index determined by the outer optimization problem. Then, $(\mathbf{a}_{i^*}/\|\mathbf{a}_{i^*}\|_N, \mathbf{e}_{i^*}/\|\mathbf{a}_{i^*}\|_N, \epsilon_{1,i^*}^*, \ldots, \epsilon_{Q,i^*}^*)$ is an optimal solution to $\mathbf{GIO}_p(\hat{\mathcal{X}})$.

3.3. Summary of Models and Comparison with Literature

Table 1 summarizes the model variants. Next, we relate the optimal values of the three variants.

Theorem 4. *Assume $\hat{\mathcal{X}} \subset \mathcal{P}$, and let z_{A}^* and z_p^* denote the optimal values of $\mathbf{GIO}_{\mathrm{A}}(\hat{\mathcal{X}})$ and $\mathbf{GIO}_p(\hat{\mathcal{X}})$, respectively. Then $z_p^* \geq z_{\mathrm{A}}^*$.*

Theorem 4 implies that if the decision space model returns a low error, so does the absolute duality gap model. Note that although bounds between objective and decision space inverse convex optimization models exist (Bertsimas et al. 2015, theorem 1; Esfahani et al. 2018, proposition 2.5), the previous bounds were developed using constants based on the nonlinearity of the objective function of the forward problem (e.g., Bertsimas et al. (2015) assumes the gradient of the objective is strongly monotone), which are not applicable in our linear setting. Furthermore, because of the nature of relative versus absolute measures, we can also bound the performance of the absolute and relative duality gap models, and consequently connect all three variants.

Corollary 3. *Let z_{A}^* and z_{R}^* denote the optimal values of $\mathbf{GIO}_{\mathrm{A}}(\hat{\mathcal{X}})$ and $\mathbf{GIO}_{\mathrm{R}}(\hat{\mathcal{X}})$, respectively. Let f_{A}^* and f_{R}^* be the optimal values of the forward problem $\mathbf{FO}(\mathbf{c})$ using cost vectors obtained by $\mathbf{GIO}_{\mathrm{A}}(\hat{\mathcal{X}})$ and $\mathbf{GIO}_{\mathrm{R}}(\hat{\mathcal{X}})$, respectively. Then, $|f_{\mathrm{R}}^*| z_{\mathrm{R}}^* \geq z_{\mathrm{A}}^* \geq |f_{\mathrm{A}}^*| z_{\mathrm{R}}^*$.*

Table 1. Summary of the Different Variants of $\mathbf{GIO}(\hat{\mathcal{X}})$

Model	$\|\cdot\|$	$\|\cdot\|_N$	$\mathcal{C}$	$\mathcal{E}_q, \forall q \in \mathcal{Q}$	Solution approach
$\mathbf{GIO}_{\mathrm{A}}(\hat{\mathcal{X}})$	$\|\cdot\|_\infty$	$\|\cdot\|_N$	$\mathbb{R}^n$	$\{\epsilon_q \mid \epsilon_q = \epsilon_q \mu(\mathbf{c})\}$	Polyhedral decomposition
$\mathbf{GIO}_{\mathrm{R}}(\hat{\mathcal{X}})$	$\|\cdot\|_\infty/\|\mathbf{b}^\top \mathbf{y}\|$	$\|\cdot\|_N$	$\mathbb{R}^n$	$\{\epsilon_q \mid \epsilon_q = \mathbf{b}^\top \mathbf{y}(\epsilon_q - 1)\mu(\mathbf{c})\}$	Three subproblems
$\mathbf{GIO}_p(\hat{\mathcal{X}})$	$\|\cdot\|_p$	$\|\cdot\|_N$	$\mathbb{R}^n$	$\{\epsilon_q \mid \mathbf{A}(\mathbf{x}_q - \epsilon_q) \geq \mathbf{b}\}$	Formulation (16)

Next, we briefly compare our models with similar models from the literature. In-depth technical comparisons are provided in Section EC.3 of the e-companion. Models $\mathbf{GIO}_{\mathrm{A}}(\hat{\mathcal{X}})$ and $\mathbf{GIO}_p(\hat{\mathcal{X}})$ can be seen as special cases of previous inverse convex optimization models; see Bertsimas et al. (2015), Aswani et al. (2018), and Esfahani et al. 2018. There, the forward problem is $\min_{\mathbf{x}}\{f(\mathbf{x};\mathbf{u},\mathbf{c}) \mid g(\mathbf{x};\mathbf{u},\mathbf{c}) \leq \mathbf{0}\}$, where $f(\mathbf{x};\mathbf{u},\mathbf{c})$ and $g(\mathbf{x};\mathbf{u},\mathbf{c})$ are convex differentiable functions and $\mathbf{u}$ is an exogenous instance-specific parameter. Thus, the data set in each of their settings is $\hat{\mathcal{X}} = \{(\hat{\mathbf{x}}_1, \hat{\mathbf{u}}_1), \ldots, (\hat{\mathbf{x}}_Q, \hat{\mathbf{u}}_Q)\}$. In our setting, we remove $\mathbf{u}$ and define $f(\mathbf{x};\mathbf{c}) = \mathbf{c}^\top\mathbf{x}$ and $g(\mathbf{x};\mathbf{c}) = \mathbf{b} - \mathbf{A}\mathbf{x}$ to obtain a linear forward problem with a fixed feasible set.

Although the assumption of instance-specific parameters generalizes our setting, we observe that the consequent formulations and methods are, on the whole, less efficient than those presented in our paper. Incorporating different forward models requires additional dual variables and dual feasibility constraints for each feasible set. For a large-scale forward optimization problem, the number of additional variables and constraints required to formulate the inverse problem grows both in the number of feasible sets and the size of $\hat{\mathcal{X}}$. For example, in our application in Section 5, n (dimension of the decision vector) and m (number of constraints) for the forward problem are on the order of 10^5. Inverse optimization frameworks from the literature (which impute instance-specific parameters) lead to inverse problems that grow significantly with every data point. In contrast, our ensemble approach using a single forward model does not suffer from this curse.

Bertsimas et al. (2015) study inverse optimization by minimizing a first-order variational inequality (which reduces to the absolute duality gap in LPs) and construct a convex inverse problem with an alternate normalization constraint (e.g., setting $\mathbf{c}^\top\mathbf{1} = 1$, rather than $\|\mathbf{c}\|_N = 1$), which may lead to trivial solutions unless the forward model is carefully chosen. For instance, setting $\mathcal{C} = \mathbb{R}^n$ and $f(\mathbf{x};\mathbf{u},\mathbf{c}) = \mathbf{c}^\top\mathbf{x}$ implies $(\mathbf{c}, \mathbf{y}, \epsilon_1, \ldots, \epsilon_Q) = (\mathbf{0}, \mathbf{0}, 0, \ldots, 0)$ is a trivially optimal solution.

Esfahani et al. (2018) study a distributionally robust inverse convex optimization problem, which can specialize to absolute duality gap inverse linear optimization with a normalization constraint. Their formulation decomposes to a finite set of conic optimization problems after polyhedral decomposition. Whereas their approach specializes to ours in the nonrobust case, we further analyze several other special cases that yield efficient solution methods (e.g., Propositions 2 and 3, and Corollary 1).

Aswani et al. (2018) propose a decision space inverse convex optimization model that satisfies a statistical consistency property given several identifiability conditions that assume the data set of decisions are noisy perturbations of optimal solutions to different forward problems. However, these assumptions may not hold in general, for example, if they arrive from an ensemble of independent prediction models, as in our application (see Section EC.3.2 of the e-companion for details). Furthermore, our solution method reformulates $\mathbf{GIO}_p(\hat{\mathcal{X}})$ to m convex problems. In contrast, Aswani et al. (2018) enumeratively solve the inverse problem using fixed $\mathbf{c}$ from a quantized subset of $\mathcal{C}$. They state that their algorithm is practical only when the parameter space $\mathcal{C}$ is modest (i.e., at most four or five parameters). However, for $\mathbf{GIO}_p(\hat{\mathcal{X}})$, we assume $\mathcal{C} = \mathbb{R}^n$, and our algorithm for $\mathbf{GIO}_p(\hat{\mathcal{X}})$ is insensitive to n.

Finally, we remark that the relative duality gap variant has not been studied in inverse convex optimization. It has been studied in inverse linear optimization but only when given a single feasible decision (Chan et al. 2019). Our case study in Section 5 demonstrates the value of $\mathbf{GIO}_{\mathrm{R}}(\hat{\mathcal{X}})$.

4. Measuring Goodness of Fit

In this section, we present a unified view of measuring model-data fitness by developing a metric that is easily and consistently interpretable across different inverse linear optimization methods, forward models, and applications. As shown in Example 2 below, assessing the aggregate error may not provide a complete picture of model fitness, necessitating a context-free fitness metric.

Previously proposed fitness measures for inverse optimization exist but are less general (e.g., for application-specific measures, Troutt et al. 2006, Chow and Recker 2012; for a metric applicable to only a single feasible decision, Chan et al. 2019). Our new metric builds off the latter metric, referred to as the *coefficient of complementarity* and denoted by $\rho(\{\hat{\mathbf{x}}\})$:

$$\rho(\{\hat{\mathbf{x}}\}) = 1 - \frac{\|\epsilon^*\|}{\frac{1}{m}\sum_{i=1}^{m}\|\epsilon_i\|}.$$

Analogous to the *coefficient of determination* R^2 in linear regression, $\rho(\{\hat{\mathbf{x}}\})$ provides a scale-free, unitless measure of goodness of fit. The numerator is the residual error from the estimated cost vector, equal to the optimal value of $\mathbf{GIO}(\{\hat{\mathbf{x}}\})$. The denominator is the average of the errors corresponding to the projections of $\hat{\mathbf{x}}$

to each of the m constraints defining the forward feasible region (i.e., $\epsilon_i = \hat{\mathbf{x}} - \pi_i(\hat{\mathbf{x}})$ for $i \in \mathcal{I}$). Just as R^2 calculates the ratio of error of a linear regression model over a baseline mean-only model, $\rho(\{\hat{\mathbf{x}}\})$ measures the relative improvement in error from using $\mathbf{FO}(\mathbf{c}^*)$ compared with a baseline of the average error induced by m candidate cost vectors.

We now generalize $\rho(\{\hat{\mathbf{x}}\})$ for $\mathbf{GIO}(\hat{\mathcal{X}})$. For convenience, we omit the data set and denote the absolute duality gap, relative duality gap, and p-norm variants of ρ as ρ_{A}, ρ_{R}, and ρ_p, respectively.

4.1. Ensemble Coefficient of Complementarity

We define the (*ensemble*) *coefficient of complementarity*, $\rho(\hat{\mathcal{X}})$, as

$$\rho(\hat{\mathcal{X}}) = 1 - \frac{\sum_{q=1}^{Q} \|\epsilon_q^*\|}{\frac{1}{m}\sum_{i=1}^{m}\left(\sum_{q=1}^{Q} \|\epsilon_{q,i}\|\right)}. \tag{17}$$

The numerator is the optimal value of $\mathbf{GIO}(\hat{\mathcal{X}})$, that is, the residual error from an optimal solution to the inverse optimization problem. The denominator terms $\sum_{q=1}^{Q} \|\epsilon_{q,i}\|$ represent the aggregate error induced by choosing baseline feasible solutions $(\mathbf{c}, \mathbf{y}) = (\mathbf{a}_i/\|\mathbf{a}_i\|_N, \mathbf{e}_i/\|\mathbf{a}_i\|_N)$:

- For the absolute duality gap, $\mathbf{GIO}_{\mathrm{A}}(\hat{\mathcal{X}})$,

$$\sum_{q=1}^{Q} \|\epsilon_{q,i}\| = \sum_{q=1}^{Q} \frac{\left|\mathbf{a}_i^\top \hat{\mathbf{x}}_q - b_i\right|}{\|\mathbf{a}_i\|_1}. \tag{18}$$

- For the relative duality gap, $\mathbf{GIO}_{\mathrm{R}}(\hat{\mathcal{X}})$, under the assumption that $b_i \neq 0$ for all $i \in \mathcal{I}$,

$$\sum_{q=1}^{Q} \|\epsilon_{q,i}\| = \sum_{q=1}^{Q} \left|\frac{\mathbf{a}_i^\top \hat{\mathbf{x}}_q}{b_i} - 1\right|. \tag{19}$$

- For the decision space, $\mathbf{GIO}_p(\hat{\mathcal{X}})$, $\sum_{q=1}^{Q} \|\epsilon_{q,i}\|$ are the optimal values of the inner problems in (16).

Our choice of baseline (denominator) is a direct extension from the single-point case, where an optimal cost vector can be found by selecting among one of the vectors $\mathbf{a}_i$ defining the m constraints. We maintain this choice of baseline for several reasons. First, an optimal solution will be exactly one of the $\mathbf{a}_i$ in the general decision space problem (see Lemma 1) and in several special cases of the objective space problem (see Propositions 2 and 6). Second, calculation of the denominator is straightforward either directly from the data (e.g., (18) and (19)) or via the solution of m convex optimization problems (16). Third, this definition directly generalizes the single-point metric, inheriting several attractive mathematical properties that we present in Section 4.2. Finally, given Propositions 2 and 6, the ensemble coefficient of complementarity is equal to the single-point version for objective space models when all data points are feasible (proof omitted).

Proposition 9. *Let $\overline{\mathbf{x}}$ be the centroid of $\hat{\mathcal{X}} \subset \mathcal{P}$. Then, $\rho_{\mathrm{A}}(\hat{\mathcal{X}}) = \rho_{\mathrm{A}}(\{\overline{\mathbf{x}}\})$ and $\rho_{\mathrm{R}}(\hat{\mathcal{X}}) = \rho_{\mathrm{R}}(\{\overline{\mathbf{x}}\})$.*

4.2. Properties of ρ

Theorem 5. *The following properties hold for ρ defined in (17):*

1. *Optimality*: *ρ is maximized by an optimal solution to $\mathbf{GIO}(\hat{\mathcal{X}})$.*
2. *Boundedness*: *$\rho \in [0, 1]$.*
3. *Monotonicity*: *For $1 \leq k < n$, let $\mathbf{GIO}^{(k)}(\hat{\mathcal{X}})$ be $\mathbf{GIO}(\hat{\mathcal{X}})$ with additional constraints $c_i = 0$, for $k+1 \leq i \leq n$, and let $\rho^{(k)}$ be the coefficient of complementarity. Then, $\rho^{(k)} \leq \rho^{(k+1)}$.*

These properties are analogous to the properties of R^2. The first property underlines how ρ integrates into $\mathbf{GIO}(\hat{\mathcal{X}})$. Although one can select any cost vector and calculate the ρ value with respect to the data $\hat{\mathcal{X}}$, an optimal cost vector obtained by solving $\mathbf{GIO}(\hat{\mathcal{X}})$ is guaranteed to attain the maximum value for ρ. Like least squares regression and R^2, our inverse optimization model and this ρ metric form a unified framework for model fitting and evaluation in inverse linear optimization.

The second property makes ρ easily interpretable as a measure of goodness of fit, with higher values indicating better fit. Note that $\rho = 1$ if and only if $\sum_{q=1}^{Q} \|\epsilon_q^*\| = 0$ (i.e., every point in $\hat{\mathcal{X}}$ lies on a supporting hyperplane of $\mathcal{P}$). In this case, the model perfectly describes all of the data points, analogous to the best fit line passing through all data points in a linear regression. Conversely, $\rho = 0$ if and only if $\sum_{q=1}^{Q} \|\epsilon_q^*\| = \sum_{q=1}^{Q} \|\epsilon_{q,i}\|$ for

all $i \in \mathcal{I}$. This scenario occurs when an optimal solution to the inverse optimization problem does not reduce the model-data fit error with respect to any of the baseline solutions, akin to when a linear regression returns an intercept-only model.

The third property states that goodness of fit is nondecreasing as additional degrees of freedom are provided to the modeler, analogous to the property that R^2 is nondecreasing in the number of features in a linear regression model. Because of this similarity, ρ also shares a weakness of R^2 related to overfitting. When using ρ to compare several models, one should ensure that higher values of ρ represent true improvements in fit, rather than artificial increases that lack generalizability.

4.3. Numerical Examples

Examples 2 and 3 illustrate the value of using ρ instead of an unnormalized error measure such as the aggregate error. Intuitively, a given error with a larger feasible set indicates better fit than the same error in a smaller set. Furthermore, ρ degrades when individual data points are forced to deviate from their preferred cost vector to minimize aggregate error. Example 4 showcases ρ for a problem where three points in the data set are fixed and the fourth is varied. Because of primal feasibility in $\mathbf{GIO}_p(\hat{\mathcal{X}})$, the decision and objective space models yield different values of ρ.

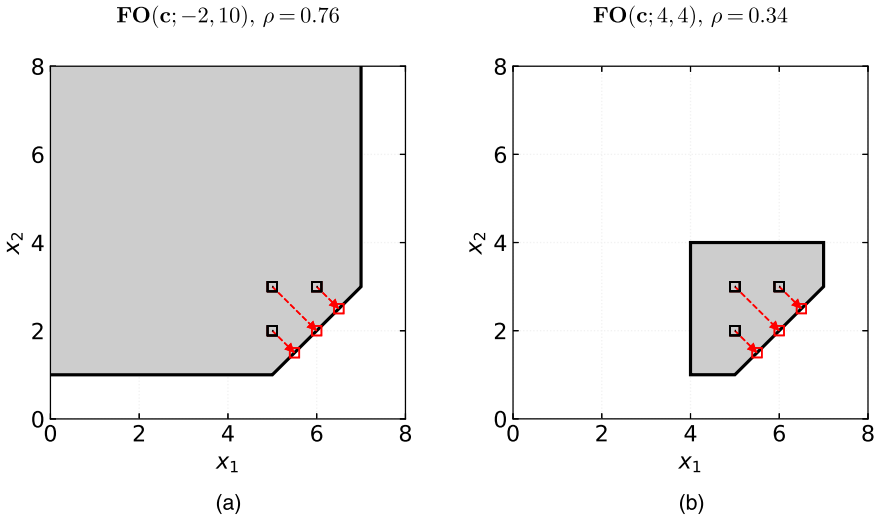
Example 2. Let $\mathbf{FO}(\mathbf{c}; u, v) : \min_{\mathbf{x}}\{c_1x_1 + c_2x_2 \mid -0.71x_1 + 0.71x_2 \geq -2.83,\ x_1 \leq 7,\ x_2 \leq v,\ x_1 \geq u;\ x_2 \geq 1\}$, and let $\hat{\mathcal{X}} = \{(5, 2.5), (4.75, 3.75), (5.5, 3)\}$. Consider two cases: $\mathbf{FO}(\mathbf{c}; -2, 10)$ and $\mathbf{FO}(\mathbf{c}; 4, 4)$. Model $\mathbf{GIO}_{\mathrm{A}}(\hat{\mathcal{X}})$ yields $\mathbf{c}^* = (-0.5, 0.5)$ and $\sum_{q=1}^{3} |\epsilon_q^*| = 2.75$ for both, but $\rho = 0.76$ for $\mathbf{FO}(\mathbf{c}; -2, 10)$ and $\rho = 0.34$ for $\mathbf{FO}(\mathbf{c}; 4, 4)$. In Figure 4(a), $\hat{\mathcal{X}}$ is closer to the bottom facet, relative to the other facets, whereas in Figure 4(b), $\hat{\mathcal{X}}$ is near the "center" of the polyhedron, rather than one facet.

Example 3. Let $\mathbf{FO}(\mathbf{c}) : \min_{\mathbf{x}}\{c_1x_1 + c_2x_2 \mid x_1 \leq 7,\ x_2 \leq 7,\ x_1 \geq 1,\ x_2 \geq 1\}$, $\hat{\mathcal{X}}_1 = \{(3.75, 2), (4, 2.25), (4.25, 2)\}$, and $\hat{\mathcal{X}}_2 = \{(1.5, 2), (4, 6.25), (6.5, 2)\}$. Both $\mathbf{GIO}_{\mathrm{A}}(\hat{\mathcal{X}}_1)$ and $\mathbf{GIO}_{\mathrm{A}}(\hat{\mathcal{X}}_2)$ impute $\mathbf{c}^* = (0, 1)$. In Figure 5(a), the points are close together and prefer the bottom facet ($\rho = 0.64$). In Figure 5(b), the points are further apart, each with a different preferred cost vector, but aggregate error is minimized by selecting a new different cost vector, resulting in poorer model fit ($\rho = 0.17$).

Examples 2 and 3 show that the aggregate error and the imputed cost vector from inverse optimization can hide poor model-data fitness. However, poor fitness arises from poor models or poor data. In Example 2, the forward model $\mathbf{FO}(\mathbf{c}; 4, 4)$ uses constraints that are potentially too tight given the data. On the other hand, in Example 3, the data set $\hat{\mathcal{X}}_2$ is spread out and unlikely to be all generated with respect to a single objective.

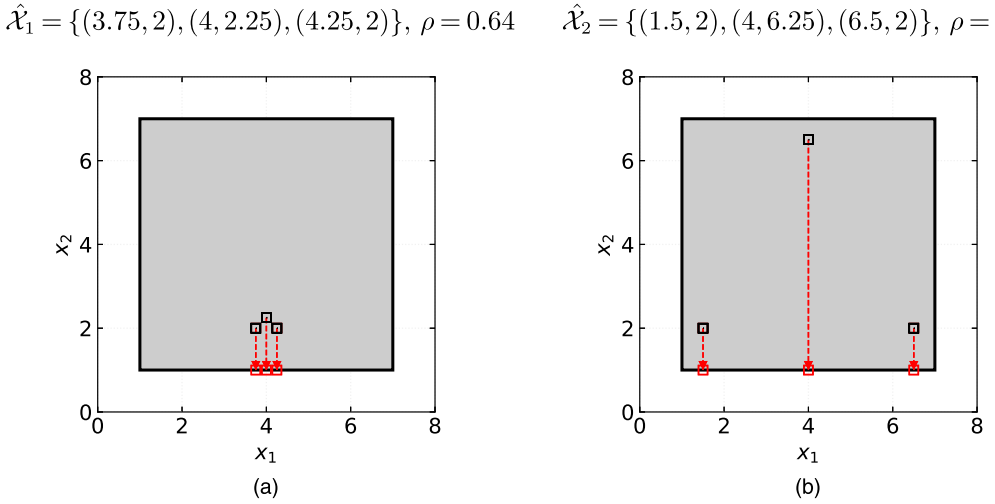
Example 4. Let $\mathbf{FO}(\mathbf{c}) : \min_{\mathbf{x}}\{c_1x_1 + c_2x_2 \mid 0.71x_1 + 0.71x_2 \geq 4.24,\ 0.71x_1 - 0.71x_2 \geq -2.83,\ x_1 \leq 7,\ x_2 \leq 7,\ x_2 \geq 1\}$, and consider all data sets of the form $\hat{\mathcal{X}} = \{(2, 5), (3, 6), (5, 4), (\gamma_1, \gamma_2)\}$, where $-2 \leq \gamma_1, \gamma_2 \leq 10$. Figure 6 shows

Figure 4. (Color online) Illustration of Example 2



Notes. The figure shows $\mathbf{GIO}_{\mathrm{A}}(\hat{\mathcal{X}})$ with two different $\mathbf{FO}(\mathbf{c}; u, v)$. The feasible sets are shaded. The squares on the interior and boundary are $\hat{\mathbf{x}}_q$ and $\hat{\mathbf{x}}_q - \epsilon_q^*$, respectively. Both problems yield the same $\mathbf{c}^*$ and ϵ_q^*, but have different model fitness ρ.

Figure 5. (Color online) Illustration of Example 3



Notes. The figure shows $\mathbf{GIO}_{\mathrm{A}}(\hat{\mathcal{X}})$ with two different data sets for the same $\mathbf{FO}(\mathbf{c})$. The feasible set is shaded. The squares on the interior and boundary are $\hat{\mathbf{x}}_q$ and $\hat{\mathbf{x}}_q - \epsilon_q^*$, respectively. Both problems yield the same $\mathbf{c}^*$ but have different errors ϵ_q^* and model fitness ρ.

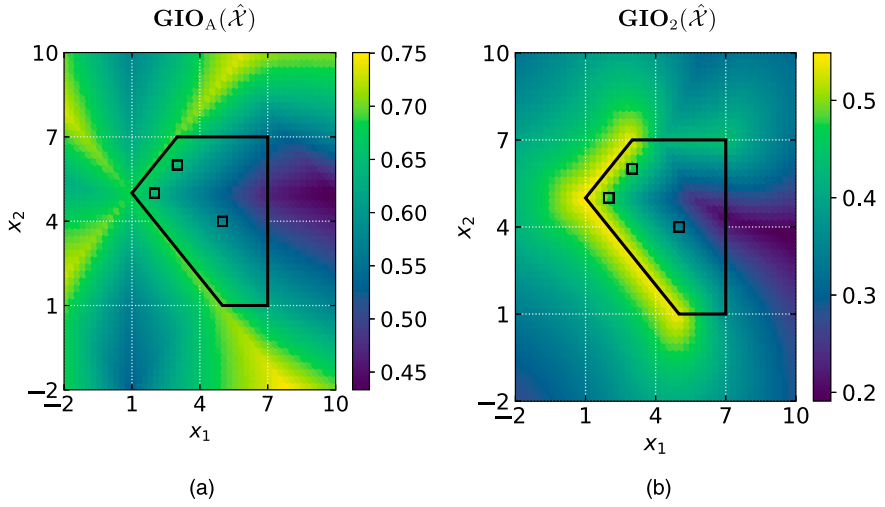
heat maps of ρ for $\mathbf{GIO}_{\mathrm{A}}(\hat{\mathcal{X}})$ and $\mathbf{GIO}_2(\hat{\mathcal{X}})$. For $\mathbf{GIO}_{\mathrm{A}}(\hat{\mathcal{X}})$, ρ is maximized when the fourth point lies on $\mathcal{H}_1 = \{(x_1, x_2) \mid 0.71x_1 - 0.71x_2 = -2.83\}$. If we solve $\mathbf{GIO}_{\mathrm{A}}(\hat{\mathcal{X}})$ with the three fixed points, then $\mathbf{c}^* = (0.5, -0.5)$. Thus, when the fourth point lies on $\mathcal{H}_1$, there is zero additional loss. The value of ρ is also high when the fourth point lies on $\mathcal{H}_2 = \{(x_1, x_2) \mid 0.71x_1 + 0.71x_2 = 4.24\}$, and degrades as it moves away from these two hyperplanes.

We observe different behavior for ρ in $\mathbf{GIO}_2(\hat{\mathcal{X}})$: maximum model fitness occurs when the fourth point lies along the facets of $\mathcal{P}$ defined by $\mathcal{H}_1$ and $\mathcal{H}_2$. Because of primal feasibility, if the fourth point is infeasible, it must project to $\mathcal{P}$ and thus incur some positive loss.

5. Automated Knowledge-Based Planning in Radiation Therapy

We now implement $\mathbf{GIO}(\hat{\mathcal{X}})$ and demonstrate the use of ρ in the context of IMRT treatment planning. In IMRT, a linear accelerator fires beamlets of different intensities that deliver a radiation dose to a tumor. A personalized treatment (consisting of beamlet intensity and dose variables) can be designed using a multi-objective optimization model, where the objective weights for a given patient are not known a priori. Knowledge-based planning offers an automated treatment design process. We consider a KBP pipeline

Figure 6. (Color online) Illustration of Example 4



Notes. The figure shows heat maps of ρ for different $\mathbf{GIO}(\hat{\mathcal{X}})$, where $\hat{\mathcal{X}}$ consists of three fixed points and the fourth variable point. The feasible set is highlighted, and the squares are the fixed $\hat{\mathbf{x}}_q$ of $\hat{\mathcal{X}}$. The value of ρ is high for $\mathbf{GIO}_{\mathrm{A}}(\hat{\mathcal{X}})$ along the relevant supporting hyperplanes, but is high for $\mathbf{GIO}_2(\hat{\mathcal{X}})$ only along the facets.

where (i) a machine learning model first predicts an appropriate dose distribution for a given patient, (ii) an inverse optimization model treats the dose as an observed decision to impute candidate objective weights, and (iii) the objective weights are input to the multiobjective planning problem to reconstruct a final plan (Babier et al. 2018b).

Different prediction models lead to plans that find different trade-offs between clinical evaluation criteria. Rather than a single prediction in KBP, we harness an ensemble of predictions to generate a single treatment plan. However, instead of averaging predictions (like in a random forest), we keep each prediction separate and feed them all into one inverse optimization model (see Figure 1(b)). Until now, KBP has never been used to generate a single plan from multiple predictions.

We develop an ensemble KBP approach using eight different predictions and show that the relative duality gap model dominates the absolute duality gap model for this application. Plans from the relative duality gap model outperform most single-point models on our overall clinical metric. Finally, by removing certain low-quality predictions, we design a final model that outperforms the single-point as well as conventional ensemble baselines. Although the final model requires clinically driven model engineering, we use ρ as domain-independent validation of the clinical intuition.

5.1. Data and Methods

We use a data set of 217 clinical treatment plans for patients with oropharyngeal (a subset of head-and-neck) cancer, randomly split into 130 plans for training and 87 plans for testing. The training set is used only to pretrain eight prediction models; the test set is used to implement our KBP pipeline. With each patient k, we associate parameters $(\mathbf{C}_k, \mathbf{A}_k, \mathbf{b}_k)$ to a multiobjective radiation therapy linear forward optimization problem $\mathbf{RT}\text{–}\mathbf{FO}(\boldsymbol{\alpha}_k) : \min_{\mathbf{x}}\{\boldsymbol{\alpha}_k^\top \mathbf{C}_k\mathbf{x} \mid \mathbf{A}_k\mathbf{x} \geq \mathbf{b}_k, \mathbf{x} \geq \mathbf{0}\}$, where $\mathbf{C}_k$ is a matrix whose rows represent different cost vectors, and $\boldsymbol{\alpha}_k$ is a vector of objective weights. The decision vector contains two subvectors, $\mathbf{x} = (\mathbf{w}, \mathbf{d})$, where $\mathbf{w}$ is the intensity of each radiation beamlet and $\mathbf{d}$ is the dose delivered to each voxel (4 mm × 4 mm × 2 mm volumetric pixel) of the patient's body, computed by a linear transformation of $\mathbf{w}$. This multiobjective model fits into $\mathbf{GIO}(\hat{\mathcal{X}}_k)$ by specifying the set of feasible cost vectors for patient k as $\mathcal{C}_k = \{\mathbf{C}_k^\top \boldsymbol{\alpha} \mid \boldsymbol{\alpha} \geq \mathbf{0}\}$. Note that the optimization problem for each patient is distinct. For a specific patient, the feasible set is fixed and a single treatment optimization problem is solved. The ensemble arises from the multiple dose predictions for the patient. Furthermore, our dose predictions are function outputs of the decisions and sufficient for use in inverse optimization. The prediction and optimization models are detailed in Section EC.4 of the e-companion and summarized below.

We first train four different dose prediction models from the literature, labeled Random Forest (RF), 2-Dimensional Red Green Blue Generative Adversarial Network (2-D RGB GAN), 2-D Generative Adversarial Network for Computationally Efficient Radiotherapy (GANCER), 3-D GANCER (McIntosh et al. 2017; Babier et al. 2018b, 2020; Mahmood et al. 2018). For each model, we also implement versions with scaled predictions (suffixed with "-sc."), which are known to produce plans that better satisfy target (tumor) criteria (Babier et al. 2020). Thus, we have eight predictions per patient, which vary in their dose trade-offs between the targets and healthy organs. We predict the dose $\hat{\mathbf{d}}_{k,q}$ for each test patient $k \in \{1, \ldots, 87\}$ with prediction model $q \in \{1, \ldots, 8\}$ and let $\hat{\mathcal{X}}_k = \{\hat{\mathbf{d}}_{k,1}, \ldots, \hat{\mathbf{d}}_{k,8}\}$ be data for each patient-specific problem. We then use inverse optimization to construct an optimal treatment plan given these predictions.

For each patient k in the test set, we implement the absolute and relative duality gap radiation therapy inverse optimization models, referred to as $\mathbf{RT}\text{–}\mathbf{IO}_{\mathrm{A}}(\hat{\mathcal{X}}_k)$ and $\mathbf{RT}\text{–}\mathbf{IO}_{\mathrm{R}}(\hat{\mathcal{X}}_k)$, respectively. They are derived from $\mathbf{GIO}_{\mathrm{A}}(\hat{\mathcal{X}}_k)$ and $\mathbf{GIO}_{\mathrm{R}}(\hat{\mathcal{X}}_k)$ by setting $\mathcal{C}_k$ as defined above, along with the template hyperparameters of Proposition 1 and Proposition 4, respectively. Once an objective weight vector $\boldsymbol{\alpha}_k^*$ is imputed from one of the inverse models, we solve $\mathbf{RT}\text{–}\mathbf{FO}(\boldsymbol{\alpha}_k^*)$ to determine the beamlets $\mathbf{w}_k^*$ and dose $\mathbf{d}_k^*$. The dose $\mathbf{d}_k^*$ is then evaluated using different clinical criteria. Note that we are not attempting to reconstruct beamlets or a dose distribution that is similar in p-norm to the predictions, but rather learn the objective function weights that the predictions appear to prioritize in order to construct a plan that best reflects clinical preferences in the ground truth $\hat{\mathbf{d}}_k$. Because plan quality is evaluated on dosimetric values in practice, we focus only on the objective space model variants.

5.2. The Value of Ensemble Inverse Optimization

In practice, a suite of quantitative metrics are evaluated to assess whether a sufficient dose is delivered to the tumor and the surrounding healthy tissue is sufficiently spared. In line with clinical practice, we use 10 binary criteria for plan evaluation (see the first two columns of Table 2; see also Babier et al. 2018a). These criteria

Table 2. The Percentage of Final Plans of Each KBP Population That Satisfy the Same Clinical Criteria as the Corresponding Clinical Plans

		$\mathbf{RT}-\mathbf{IO}_{\mathrm{A}}(\hat{\mathcal{X}})$ (%)	$\mathbf{RT}-\mathbf{IO}_{\mathrm{R}}(\hat{\mathcal{X}})$ (%)			
Structure	Criteria (grays)	8 pts.	8 pts.	6 pts.	4 pts.	2 pts.
Brainstem	Max ≤ 54	100	100	100	100	100
Spinal cord	Max ≤ 48	100	100	98.9	98.9	100
Right parotid	Mean ≤ 26	58.8	88.2	88.2	82.4	94.1
Left parotid	Mean ≤ 26	63.6	81.8	81.8	81.8	81.8
Larynx	Mean ≤ 45	59.2	95.9	95.9	93.9	95.9
Mandible	Mean ≤ 45	74.4	100	100	100	100
Esophagus	Max ≤ 73.5	51.5	100	98.5	95.5	97.0
PTV70	99th percentile ≥ 66.5	51.7	91.4	94.8	96.6	86.2
PTV63	99th percentile ≥ 59.9	50.0	98.0	98.0	98.0	98.0
PTV56	99th percentile ≥ 53.2	30.4	45.7	80.4	100	69.6
All		26.4	60.9	75.9	83.9	70.1

Notes. OARs are assigned a mean or maximum dose criterion depending on relevance. PTVs are assigned criteria to the 99th percentile. pts., points.

cover seven organs at risk (OARs) and three planning target volumes (PTVs). OARs are healthy structures whose dose should remain below a specific threshold (e.g., the maximum dose delivered to any voxel in the brainstem should be less than 54 Gy). The PTVs are regions that encompass the tumor sites and are each assigned a criterion specifying the minimum dose that at least 99% of its volume should receive. To evaluate our plans on these criteria, we first check whether the corresponding clinical (ground truth) plan satisfied given criteria. If the clinical plan satisfied the criteria, we evaluate whether the generated plan also satisfied that criteria.

The columns of Table 2 list the proportions of plans generated by $\mathbf{RT}-\mathbf{IO}_{\mathrm{A}}(\hat{\mathcal{X}})$ and $\mathbf{RT}-\mathbf{IO}_{\mathrm{R}}(\hat{\mathcal{X}})$ that satisfied the corresponding clinical criteria. The row "All" reflects the percentage of plans that satisfied all of the criteria that were also met by the corresponding clinical plans and is an aggregate measure of plan quality. We first use all eight predictions to solve $\mathbf{RT}-\mathbf{IO}_{\mathrm{A}}(\hat{\mathcal{X}})$ (third column) and $\mathbf{RT}-\mathbf{IO}_{\mathrm{R}}(\hat{\mathcal{X}})$ (fourth column). Model $\mathbf{RT}-\mathbf{IO}_{\mathrm{R}}(\hat{\mathcal{X}})$ substantially outperforms model $\mathbf{RT}-\mathbf{IO}_{\mathrm{A}}(\hat{\mathcal{X}})$ over every criterion, suggesting that the absolute duality gap model is not well suited to this application. This result is consistent with results observed for single-point inverse optimization in IMRT (Chan et al. 2014, 2019; Goli et al. 2018), and we conjecture that it is due to the wide range of objective function magnitudes in the forward problem. The absolute duality gap model adjusts each objective value by the same absolute amount, causing relatively large adjustments to objectives with low values and small adjustments to those with high values; thus, it has difficulty balancing different criteria.

Although $\mathbf{RT}-\mathbf{IO}_{\mathrm{R}}(\hat{\mathcal{X}})$ with eight predictions is generally effective at satisfying the OAR criteria, these plans sacrifice the PTV criteria, especially PTV56. We hypothesize that this performance for PTV criteria is due to the large variability in the quality of predictions. For example, the 2-D RGB GAN, 2-D GANCER, and 3-D GANCER models are known to produce plans that emphasize OAR criteria at the expense of the PTV. Criteria satisfaction for single-point $\mathbf{RT}-\mathbf{IO}_{\mathrm{R}}(\{\hat{\mathbf{x}}\})$ using each of the individual predictions is shown in Table 3. Depending on which prediction is used, the single-point KBP population varies from 10.9% to 95.7% in terms of satisfying the PTV56 criteria. The ability of the single-point models to satisfy all clinical criteria ranges between 44.8% and 80.5%, suggesting that some single-point KBP models make poorer trade-offs in criteria satisfaction than others. Regardless of the variability among predictions, the ensemble model outperforms all but the top three single-point models in satisfying all criteria. In cases where the cost of determining model performance is expensive (e.g., having to solve inverse and forward models over multiple predictions and patients), ensemble inverse optimization can reliably provide high-quality plans.

Using multiple points of varying quality as input to the ensemble model may lead to poor model-data fit (see Example 3). We experiment with IO models based on subsets of the eight predictions to determine which subset of KBP prediction models best fit $\mathbf{RT}-\mathbf{FO}(\boldsymbol{\alpha})$. The clinical KBP literature shows that some of the prediction models generally perform better than others: scaled GANCER models typically predict better than RF models, which themselves predict better than RGB GAN and unscaled GANCER models (Mahmood et al. 2018, Babier et al. 2020). Using the prior literature and qualitative assessment from a clinical collaborator, we propose an ordering of the models from weak to strong: 3-D GANCER, 2-D RGB GAN, 2-D GANCER, 2-D

Table 3. The Percentage of Single-Point Inverse Optimization Plans of Each KBP Population That Satisfy the Same Clinical Criteria as the Clinical Plans

		$\mathbf{RT}-\mathbf{IO}_{\mathrm{R}}(\{\hat{\mathbf{x}}_q\})$ (%)							
Structure	Criteria (grays)	3-D GANCER	2-D RGB GAN	2-D GANCER	2-D RGB GAN-sc.	RF-sc.	RF	2-D GANCER-sc.	3-D GANCER-sc.
Brainstem	Max ≤ 54	100	100	100	100	98.9	100	100	100
Spinal cord	Max ≤ 48	100	98.9	100	98.9	98.9	100	98.9	98.9
Right parotid	Mean ≤ 26	94.1	94.1	82.4	88.2	94.1	88.2	88.2	94.1
Left parotid	Mean ≤ 26	100	90.9	81.8	63.6	72.8	63.6	81.8	81.8
Larynx	Mean ≤ 45	98.0	89.8	89.8	87.8	95.9	91.8	85.7	93.9
Mandible	Mean ≤ 45	100	100	100	100	98.7	100	100	100
Esophagus	Max ≤ 73.5	100	100	100	98.5	100	100	89.4	84.8
PTV70	99th percentile ≥ 66.5	81.0	36.2	81.0	69.0	63.8	91.4	98.3	100
PTV63	99th percentile ≥ 59.9	92.0	100	100	100	98.0	98.0	100	100
PTV56	99th percentile ≥ 53.2	10.9	58.7	19.6	82.6	47.8	65.2	95.7	95.7
All		44.8	47.1	47.1	59.8	55.2	67.8	77.0	80.5

RGB GAN-sc., RF-sc., RF, 2-D GANCER-sc., and 3-D GANCER-sc. Note the general pattern is more important than the exact ordering; that is, we rate the scaled GANCER models as strongest, followed by RF models, followed by RGB GAN and unscaled GANCER. We implement $\mathbf{RT}-\mathbf{IO}_{\mathrm{R}}(\hat{\mathcal{X}})$ with data sets of decreasing size by sequentially removing the two weakest predictors. For example, the six-point IO model uses the six strongest predictions, whereas the four-point model uses only the scaled GANCER and RF. The fifth, sixth, and seventh columns of Table 2 show the performance of the three subset IO models. The six-point model markedly improves over the eight-point model on PTV criteria, while satisfying almost all OAR criteria, resulting in an additional 15% of the final plans being able to satisfy all criteria. Similarly, the four-point model improves over the six-point model by achieving near perfect PTV criterion satisfaction while mostly preserving OAR performance. In fact, this model now outperforms the best single-point model, 3-D GANCER-sc. (see Table 3). Interestingly, performance does not improve in the two-point model. This model uses two predictions (2-D GANCER-sc. and 3-D GANCER-sc.) that individually achieve high PTV satisfaction in their single-point models, but fail to do so when combined in an ensemble. We conjecture that the two-point model reaches a local minimum in PTV satisfaction because the forward objectives do not directly target PTV criteria (see Section EC.4 of the e-companion).

Overall, these experiments demonstrate that ensemble inverse optimization is valuable for turning an ensemble of predictions into a single treatment plan. Although an off-the-shelf ensemble model immediately outperforms most single-point constituents, our results show that careful selection of data is required to maximize performance and beat all single-point KBP models.

5.3. Comparison with Existing Ensemble Learning Techniques

We next compare $\mathbf{RT}-\mathbf{IO}_{\mathrm{R}}(\hat{\mathcal{X}})$ with two conventional ensemble learning baselines that do not account for linear programming geometry. The first baseline is an "ensemble-then-inverse optimization" approach where for each patient k, the centroid $\bar{\mathbf{d}}_k$ of the individual predictions is input into a single-point inverse optimization problem $\mathbf{RT}-\mathbf{IO}_{\mathrm{R}}(\{\bar{\mathbf{d}}_k\})$. The second baseline is a multiplicative weights algorithm (MWA), commonly used in "learning from experts" settings (Arora et al. 2012). Here, we first solve the single-point problem $\mathbf{RT}-\mathbf{IO}_{\mathrm{R}}(\{\hat{\mathbf{d}}_{k,q}\})$ with each prediction model for the training set patients. We treat each prediction model as a different expert and learn a probability distribution over the set of prediction models using the aggregate error as a loss function. Then, for each patient in the test set, we use this distribution to randomly sample a prediction model and solve a single-point problem. Baseline implementation details are given in Section EC.4.6 of the e-companion.

We implement the centroid and the MWA model using all eight predictions per patient (i.e., eight points), as well as the four-point predictions (RF-sc., RF, 2-D GANCER-sc., and 3-D GANCER-sc.). Table 4 compares our incumbent, the four-point $\mathbf{RT}-\mathbf{IO}_{\mathrm{R}}(\hat{\mathcal{X}})$, with the best-performing centroid and MWA models. If all dose predictions were feasible with respect to $\mathbf{RT}-\mathbf{FO}(\boldsymbol{\alpha})$, then, by Proposition 6, our ensemble model and the centroid model would be equivalent. Each prediction model outputs feasible doses for approximately 85% of the patients (see Table EC.1 in the e-companion). Consequently, $\mathbf{RT}-\mathbf{IO}_{\mathrm{R}}(\hat{\mathcal{X}})$ yields different plans from

Table 4. The Percentages of Plans from Different Ensemble Models That Satisfy the Same Clinical Criteria as the Corresponding Clinical Plans

Structure	Criteria (grays)	$\mathbf{RT}-\mathbf{IO}_{\mathrm{R}}(\hat{\mathcal{X}})$ (%)	Centroid $\mathbf{RT}-\mathbf{IO}_{\mathrm{R}}(\{\bar{\mathbf{d}}_k\})$ (%)	MWA $\mathbf{RT}-\mathbf{IO}_{\mathrm{R}}(\{\hat{\mathbf{d}}_{k,q}\})$ (%)
Brainstem	Max ≤ 54	100	100	100
Spinal cord	Max ≤ 48	98.9	100	100
Right parotid	Mean ≤ 26	82.4	88.2	88.2
Left parotid	Mean ≤ 26	81.8	81.8	63.6
Larynx	Mean ≤ 45	93.9	87.8	91.8
Mandible	Mean ≤ 45	100	98.5	100
Esophagus	Max ≤ 73.5	95.5	100	100
PTV70	99th percentile ≥ 66.5	96.6	96.6	93.1
PTV63	99th percentile ≥ 59.9	98.0	100	98.0
PTV56	99th percentile ≥ 53.2	100	80.4	67.4
All		83.9	77.0	69.0

Notes. $\mathbf{RT}-\mathbf{IO}_{\mathrm{R}}(\hat{\mathcal{X}})$ refers to the four-point model from Table 2. We present the best-performing setting for each baseline, that is, the eight-point centroid model and the four-point MWA.

$\mathbf{RT}-\mathbf{IO}_{\mathrm{R}}(\bar{\mathbf{d}}_k)$. Our incumbent outperforms the baseline on all criteria by 6.9%. Nonetheless, the centroid model is similar to $\mathbf{RT}-\mathbf{IO}_{\mathrm{R}}(\hat{\mathcal{X}})$ for each individual criterion. We intuit that if only a small fraction of points in $\hat{\mathcal{X}}$ are infeasible, then centroid inverse optimization is an efficient approximation of ensemble inverse optimization.

The MWA baseline randomly selects a single-point inverse optimization model for each patient according to a learned probability distribution. This approach is a tractable alternative to solving eight inverse optimization problems and selecting the best plan for each patient (see Figure 1(a)). As shown in Table 3, some single-point models are significantly better than others. Thus, most of the test set patients will receive plans from RF, 2-D GANCER-sc., or 3-D GANCER-sc. However, $\mathbf{RT}-\mathbf{IO}_{\mathrm{R}}(\hat{\mathcal{X}})$ already outperforms each of these single-point models on most of the criteria. Consequently, $\mathbf{RT}-\mathbf{IO}_{\mathrm{R}}(\hat{\mathcal{X}})$ outperforms the MWA baseline on all criteria by 14.9%.

5.4. Using ρ to Validate the Best Subset of the Data

We showed previously that using a targeted subset of the predictions yielded a better model. The intuition follows Example 3, where points that are individually far from each other induce poor fit. Although our ranking scheme was domain specific, here we demonstrate a domain-independent validation of the selection of the data sets in the six-, four-, and two-point models using ρ.

We consider three variants for each of the six-, four-, and two-point models by selecting subsets of strong, medium, and weak predictions according to our clinical ordering. Strong subsets correspond to the models developed in Section 5.2, weak subsets use the lowest ordered predictions and sequentially remove the best, and medium subsets use the predictions from 2-D RGB GAN to 2-D GANCER-sc. and sequentially remove one strong and one weak prediction. Table 5 compares ρ across models with varying quality of predictions. Note that we are not studying the effect of data set size Q (along columns of Table 5), but rather the effect of quality (along rows of Table 5). For fixed Q, the strong model always yields the highest ρ, which suggests that the strong predictions are the best fit for the clinical forward model. Furthermore, in parentheses in Table 5, we show that the clinical criteria satisfaction rates for each of the ensemble models also reflect trends similar to those of ρ. Because ρ is a general metric, we can evaluate the model quality for a given number of points without domain-specific knowledge and come to nearly the same conclusion as via the clinical criteria, which are domain specific and require additional computation due to resolving the forward model.

Table 5. ρ for the Weak, Medium, and Strong Subsets of Two, Four, and Six Predictions

Model	Weak	Medium	Strong
Two points	0.63 (42.5)	0.65 (60.9)	**0.90 (70.1)**
Four points	0.56 (30.1)	0.68 (62.1)	**0.73 (83.9)**
Six points	0.64 (51.7)	0.63 (57.5)	**0.67 (75.9)**

Notes. The all-criteria percentage satisfaction for each model is in parentheses. The "Strong" column reflects the predictions used in Table 2. Highest-performing models are bolded.

However, ρ is not a perfect surrogate for criteria satisfaction. For example, the weak six-point model has a slightly higher ρ than the medium six-point model. Note that the two data sets share four of six points, and the relatively similar ρ reflects a similar criteria satisfaction rate. We also observe that the data set with the best fit from an inverse optimization perspective (strong two points) is not the one resulting in the best clinical criteria evaluation (strong four points). This result is because that ρ is calculated via the average distance of the predictions to the constraints, but the constraints only approximate the criteria (see Section EC.4.1 of the e-companion). Because the predictions are close to the constraints but not criteria, ρ is overly optimistic for this model. Using diverse predictions of high clinical quality allows us to obtain ρ values that are more representative of the clinical problem.

6. Conclusion

Inverse optimization is an increasingly popular model-fitting paradigm for estimating the cost vector of an optimization problem given decision data. Motivated by ensemble methods in machine learning, we develop a framework that uses a collection of decisions for a single problem to estimate a cost vector. The data are drawn from different decision makers attempting to solve a single problem or, as in our application, a family of machine learning–generated predictions of an optimal solution. We propose a generalized inverse linear optimization framework that unifies several common variants of inverse optimization from the literature and derive assumption-free exact solution methods for each. Comparing with the inverse convex optimization literature shows that by focusing on our specialized context, we can leverage the geometry of linear optimization to produce tighter performance bounds and more efficient solution methods. To complete our framework, we develop a general goodness of fit metric to measure model-data fit in any inverse linear optimization application. We demonstrate that this metric, by virtue of possessing properties analogous to R^2 in linear regression, is easy to calculate and interpret.

We propose a novel application of ensemble inverse optimization in the automated construction of radiation therapy treatment plans. In contrast to traditional approaches, which generate plans from individual predictions, we use a family of predictions, each with different characteristics and trade-offs, to form treatment plans that better imitate clinically delivered treatments. Finally, although constructing the best inverse optimization model requires careful clinical expertise, we show how our goodness-of-fit metric provides domain-independent validation of our model engineering. Beyond the specific context and application presented in this paper, we believe there will be new applications of predict-and-ensemble inverse optimize frameworks that can build on our foundation.

Acknowledgment

The authors thank the anonymous reviewers, associate editor, and editor-in-chief for their helpful comments which greatly improved the paper.

References

Arora S, Hazan E, Kale S (2012) The multiplicative weights update method: A meta-algorithm and applications. *Theory Comput.* 8(1):121–164.

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

Aswani A, Shen ZJ, Siddiq A (2019) Data-driven incentive design in the Medicare Shared Savings Program. *Oper. Res.* 67(4):1002–1026.

Babier A, Boutilier JJ, McNiven AL, Chan TCY (2018a) Knowledge-based automated planning for oropharyngeal cancer. *Medical Phys.* 45(7):2875–2883.

Babier A, Boutilier JJ, Sharpe MB, McNiven AL, Chan TCY (2018b) Inverse optimization of objective function weights for treatment planning using clinical dose-volume histograms. *Phys. Medical Biol.* 63(10):105004.

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

Bertsimas D, Gupta V, Paschalidis IC (2012) Inverse optimization: A new perspective on the Black-Litterman model. *Oper. Res.* 60(6):1389–1403.

Bertsimas D, Gupta V, Paschalidis IC (2015) Data-driven estimation in equilibrium using inverse optimization. *Math. Programming* 153(2): 595–633.

Breiman L (2001) Random forests.*Machine Learn.* 45(1):5–32.

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

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

Chow JYJ, Recker WW (2012) Inverse optimization with endogenous arrival time constraints to calibrate the household activity pattern problem. *Transportation Res. Part B: Methodological* 46(3):463–479.

Das IJ, Moskvin V, Johnstone PA (2009) Analysis of treatment planning time among systems and planners for intensity-modulated radiation therapy. *J. Amer. College Radiology* 6(7):514–517.

Delaney G, Jacob S, Featherstone C, Barton M (2005) The role of radiotherapy in cancer treatment. *Cancer* 104(6):1129–1137.

Esfahani PM, Shafieezadeh-Abadeh S, Hanasusanto GA, Kuhn D (2018) Data-driven inverse optimization with imperfect information. *Math. Programming* 167(1):191–234.

Goli A (2015) Sensitivity and stability analysis for inverse optimization with applications in intensity-modulated radiation therapy. Unpublished master's thesis, University of Toronto, Toronto, Canada.

Goli A, Boutilier JJ, Craig T, Sharpe MB, Chan TCY (2018) A small number of objective function weight vectors is sufficient for automated treatment planning in prostate cancer. *Phys. Medicine Biol.* 63(19):195004.

Kearney V, Chan JW, Haaf S, Descovich M, Solberg TD (2018) DoseNet: A volumetric dose prediction algorithm using 3D fully-convolutional neural networks. *Phys. Medicine Biol.* 63(23):235022.

Keshavarz A, Wang Y, Boyd S (2011) Imputing a convex objective function. *Proc. IEEE Internat. Sympos. Intelligent Control* (Institute of Electrical and Electronics Engineers, Piscataway, NJ), 613–619.

Mahmood R, Babier A, McNiven A, Diamant A, Chan TCY (2018) Automated treatment planning in radiation therapy using generative adversarial networks. Doshi-Velez F, Fackler J, Jung K, Kale D, Ranganath R, Wallace B, Wiens J, eds. *Proc. Third Machine Learn. Healthcare Conf.* Proceedings of Machine Learning Research, vol. 85, 484–499.

Mangasarian OL (1999) Arbitrary-norm separating plane. *Oper. Res. Lett.* 24(1):15–23.

McIntosh C, Purdie TG (2016) Voxel-based dose prediction with multi-patient atlas selection for automated radiotherapy treatment planning. *Phys. Medicine Biol.* 62(2):415.

McIntosh C, Welch M, McNiven A, Jaffray DA, Purdie TG (2017) Fully automated treatment planning for head and neck radiotherapy using a voxel-based dose prediction and dose mimicking method. *Phys. Medicine Biol.* 62(15):5926.

Saez-Gallego J, Morales JM, Zugno M, Madsen H (2016) A data-driven bidding model for a cluster of price-responsive consumers of electricity. *IEEE Trans. Power Systems* 31(6):5001–5011.

Sharpe MB, Moore KL, Orton CG (2014) Within the next ten years treatment planning will become fully automated without the need for human intervention. *Medical Phys.* 41(12):120601.

Troutt MD (1995) A maximum decisional efficiency estimation principle. *Management Sci.* 41(1):76–82.

Troutt MD, Pang WK, Hou SH (2006) Behavioral estimation of mathematical programming objective function coefficients. *Management Sci.* 53(3):422–434.

Zhao Q, Stettner A, Reznik E, Segrè D, Paschalidis IC (2015) Learning cellular objectives from fluxes by inverse optimization. *Proc.* 54*th IEEE Conf. Decision Control* (Institute of Electrical and Electronics Engineers, Piscataway, NJ), 1271–1276.