Model Fitting in Generalized Inverse Linear Optimization: Applications in Radiation Therapy

Aaron Babier, Timothy C. Y. Chan
Mechanical & Industrial Engineering, University of Toronto, Toronto, Canada, {ababier, tcychan}@mie.utoronto.ca

Taewoo Lee
Industrial Engineering, University of Houston, Houston, Texas, USA, tlee6@uh.edu

Rafid Mahmood
Mechanical & Industrial Engineering, University of Toronto, Toronto, Canada, rafid.mahmood@mail.utoronto.ca

Daria Terekhov
Mechanical, Industrial, & Aerospace Engineering, Concordia University, Montréal, Canada, daria.terekhov@concordia.ca

We develop a generalized inverse optimization framework for fitting the cost vector of a single linear optimization problem given an ensemble of observed decisions. We unify multiple variants in the inverse optimization literature under a common template and derive assumption-free and exact solution methods for each variant. We then extend a goodness-of-fit metric previously introduced for the problem with a single observed decision to this new setting, proving and numerically illustrating several important properties. Finally, to illustrate our framework, we develop a novel inverse optimization-driven procedure for automated radiation therapy treatment planning. Here, the inverse optimization model leverages the combined power of an ensemble of dose predictions produced by different machine learning models to construct clinical treatment plans that better trade off between the competing clinical objectives that are used for plan evaluation in practice.

Key words: inverse optimization; linear optimization; goodness of fit; model estimation; radiation therapy

1. Introduction

Motivated by the growing availability of data that represents decisions, there is an increasing interest in the use of inverse optimization to gain insight into decision-generating (e.g., optimization) processes. Inverse optimization determines optimization model parameters that render a given data set of observed decisions minimally sub-optimal for the model. The inverse optimization literature studies a range of model types that can be classified along multiple dimensions, including the characteristics of the data set (e.g., a single data point versus multiple, or whether the points are all strictly feasible), the type of the forward optimization problem (e.g., a linear, convex, or integer program), and the loss function in the inverse problem itself. In this paper, we study inverse optimization given an arbitrary data set of decisions for a single linear program. 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 assumptions.

The setting of multiple observed decisions for a single forward model is motivated methodologically from ensemble methods in machine learning. Consider the canonical example of a random forest that averages the predictions obtained from a set of decision trees that are each trained on subsets of the data (Breiman 2001). Each individual tree, by training on a different subset, specializes to a specific region of the feature space, and the ensemble then averages out the over-fit of the individual models. The analogy in inverse optimization is the need to learn from multiple decision-makers each attempting to solve the same underlying optimization problem; they each propose a different solution and imputing the costs implicitly leads to a consensus (Troutt 1995). Our specific motivating application is the automated construction of radiation therapy treatment plans to treat cancer patients. A common framework for this task is a two-stage approach known as knowledge-based automated planning (KBAP), where a machine learning model predicts a desirable (but typically infeasible) dose distribution and inverse optimization learns model parameters that can generate a similar, physically deliverable treatment through the solution of a forward problem (Sharpe et al. 2014, McIntosh et al. 2017, Babier et al. 2018a). Treatments are evaluated using a set of clinical evaluation criteria and different prediction models are typically biased towards certain criteria (Babier et al. 2018b). In this paper, we apply our inverse optimization methods to develop a novel KBAP pipeline that ensembles multiple predictions to obtain a consensus treatment plan that achieves better trade-offs on clinical criteria than any individual model.

Methodologically, our work extends the single observation (feasible) generalized inverse optimization framework of Chan et al. (2018) to the case of multiple observations (with no assumptions on feasibility). Our framework is founded on a flexible model template that can be specialized to multiple different inverse optimization models via appropriate specification of hyperparameters.

The first variant, known as the absolute duality gap, is well-studied in the multi-point inverse optimization literature, commonly in the context of a more general inverse convex optimization (see Bertsimas et al. (2015), Esfahani et al. (2018) for methodology and Zhao et al. (2015), Saez-Gallego and Morales (2017) for applications). The second, known as the relative duality gap, has only been studied in single-point inverse linear optimization (Chan et al. 2014, 2018). Finally, we generalize a previously developed goodness-of-fit metric for inverse optimization (Chan et al. 2018) to the case of multiple observations. Altogether, we construct a unified framework for model fitting and evaluation in inverse linear optimization for an arbitrary data set.

The specific contributions of our paper are as follows:

1. We develop an inverse linear optimization model that generalizes many approaches in the literature and is applicable to arbitrary data sets of decisions for a single forward optimization problem. This model is expressed in terms of a set of hyperparameters used to derive both general and application-specific model variants.
2. We develop new exact, assumption-free solution methods for each of the different variants of our generalized model. 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 decision data. We prove and illustrate several intuitive properties of the metric, including optimality with respect to the inverse optimization model, boundedeness, and monotonicity.
4. We apply our framework to implement the first ensemble KBAP pipeline, capable of using multiple predictions to construct treatment plans for head-and-neck cancer that achieve better clinical trade-offs compared to traditional single-point KBAP plans. We then show how ρ provides a domain-independent validation of our final model.

Proofs are omitted in this work unless particularly relevant.

2. Background on generalized inverse linear optimization

We first review the formulation and main results from Chan et al. (2018), who introduced an inverse optimization model for linear optimization problems (LPs) that unifies both decision and objective space inverse linear optimization 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 $\mathbf{A} \in \mathbb{R}^{m\times n}, \mathbf{b} \in \mathbb{R}^m$ denote the constraint matrix and right-hand side vector, respectively. Let $\mathcal{I} = \{1,\ldots,m\}$ and $\mathcal{J} = \{1,\ldots,n\}$ index the constraints and decision variables respectively. Let $\mathcal{P} = \{\mathbf{x} \mid \mathbf{A}\mathbf{x} \geq \mathbf{b}\}$ be the feasible set, and suppose there are no redundant constraints. We define $\mathbf{FO}(\mathbf{c}) : \min_{\mathbf{x}}\{\mathbf{c}^\mathsf{T}\mathbf{x} \mid \mathbf{x} \in \mathcal{P}\}$ to be the forward optimization problem. For a *feasible* observed decision $\hat{\mathbf{x}} \in \mathcal{P}$, the (*single-point*) generalized inverse linear optimization problem is

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

In formulation (1), $\mathbf{y} \in \mathbb{R}^m$ represents the dual vector for the constraints of the forward problem. The vector $\boldsymbol{\epsilon} \in \mathbb{R}^n$ represents a perturbation that brings $\hat{\mathbf{x}}$ to a point $\mathbf{x}^* = \hat{\mathbf{x}} - \boldsymbol{\epsilon}$ that satisfies strong duality (1c). The norm in the objective is general and can be chosen based on application-specific considerations. Constraints (1b) ensure dual feasibility. 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 vectors via the sets $\mathcal{E}$ and $\mathcal{C}$, respectively. The tuple $(\|\cdot\|, \|\cdot\|_N, \mathcal{C}, \mathcal{E})$ forms the inverse optimization model *hyperparameters*. By selecting these hyperparameters appropriately, $\mathbf{GIO}(\{\hat{\mathbf{x}}\})$ was shown to specialize into models that minimize various different error measures.

Although formulation (1) is non-convex, it admits a closed-form solution, which can be identified by determining the projection from $\hat{\mathbf{x}}$ to the boundary of $\mathcal{P}$ of minimum distance as measured by $\|\cdot\|$. Specifically, let $\mathcal{H}_i = \left\{\mathbf{x} \,\middle|\, \mathbf{a}_i^\mathsf{T}\mathbf{x} = b_i\right\}$ be the hyperplane corresponding to the i^{th} constraint and

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

be the projection of $\hat{\mathbf{x}}$ to $\mathcal{H}_i$. The hyperplane projection problem has an analytic solution $\pi_i(\hat{\mathbf{x}}) = \hat{\mathbf{x}} - \frac{\mathbf{a}_i^\intercal \hat{\mathbf{x}} - b_i}{\|\mathbf{a}_i\|_D}\nu(\mathbf{a}_i)$, where $\|\cdot\|_D$ is the dual norm of $\|\cdot\|$ and $\nu(\mathbf{a}_i) \in \arg\max_{\|\mathbf{v}\|=1} \left\{\mathbf{v}^\intercal \mathbf{a}_i\right\}$ (Mangasarian 1999) . Theorem 1 (Chan et al. 2018) leverages this to obtain an optimal solution to formulation (1).

Theorem 1 **(Chan et al., 2018).** *Let $\hat{\mathbf{x}} \in \mathcal{P}$, $i^* \in \arg\min_{i\in\mathcal{I}} \left\{\frac{\mathbf{a}_i^\intercal \hat{\mathbf{x}} - b_i}{\|\mathbf{a}_i\|_D}\right\}$, and $\mathbf{e}_i$ be the i^{th} unit vector. There exists an optimal solution to* $\mathbf{GIO}(\{\hat{\mathbf{x}}\})$ *of the form*

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

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

3. Generalized inverse linear optimization with arbitrary data sets

In this section, we extend the model and results of the previous section to the case of multiple observed decisions with no restriction on their feasibility. Let $\hat{\mathcal{X}} = \{\hat{\mathbf{x}}_1, \dots, \hat{\mathbf{x}}_Q\}$ represent a data set indexed by $\mathcal{Q} = \{1, \dots, Q\}$. We will determine a single cost vector $\mathbf{c}^*$ to minimize the aggregate error induced by all points with respect to $\mathbf{FO}(\mathbf{c}^*)$. To measure the error, we introduce a perturbation vector $\boldsymbol{\epsilon}_q$ for $q \in \mathcal{Q}$. The *multi-point* generalized inverse linear optimization problem is

$$\begin{aligned}
\mathbf{GIO}(\hat{\mathcal{X}}): \quad \underset{\mathbf{c},\mathbf{y},\boldsymbol{\epsilon}_1,\dots,\boldsymbol{\epsilon}_Q}{\text{minimize}} \quad & \sum_{q=1}^{Q} \|\boldsymbol{\epsilon}_q\| && \text{(4a)}\\
\text{subject to} \quad & \mathbf{A}^\intercal \mathbf{y} = \mathbf{c}, \;\; \mathbf{y} \geq \mathbf{0} && \text{(4b)}\\
& \mathbf{c}^\intercal \hat{\mathbf{x}}_q = \mathbf{b}^\intercal \mathbf{y} + \mathbf{c}^\intercal \boldsymbol{\epsilon}_q, \quad \forall q \in \mathcal{Q} && \text{(4c)}\\
& \|\mathbf{c}\|_N = 1 && \text{(4d)}\\
& \mathbf{c} \in \mathcal{C}, \boldsymbol{\epsilon}_q \in \mathcal{E}_q, \quad \forall q \in \mathcal{Q}. && \text{(4e)}
\end{aligned}$$

Constraints (4b) and (4d) are carried from the single-point model, while (4c) and (4e) are multi-point 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 $\boldsymbol{\epsilon}_q \in \mathcal{E}_q$. The objective minimizes the sum of the norms of the individual perturbation vectors.

Similar to formulation (1), $\mathbf{GIO}(\hat{\mathcal{X}})$ is non-convex due to the bilinear terms in (4c) and the normalization constraint (4d). We first show that $\mathbf{GIO}(\hat{\mathcal{X}})$ specializes to two different objective space variants, before developing tailored and tractable solution methods.

3.1. Absolute duality gap

The absolute duality gap method for inverse optimization minimizes the aggregate duality gap between the observed primal objective for each decision and the dual optimal value for the problem:

$$
\begin{aligned}
\mathbf{GIO}_{\mathrm{A}}(\hat{\mathcal{X}}): \quad \underset{\mathbf{c},\mathbf{y},\epsilon_1,\ldots,\epsilon_Q}{\text{minimize}} \quad & \sum_{q=1}^{Q} |\epsilon_q| && \text{(5a)}\\
\text{subject to} \quad & \mathbf{A}^{\mathsf{T}}\mathbf{y} = \mathbf{c}, \;\; \mathbf{y} \geq \mathbf{0} && \text{(5b)}\\
& \mathbf{c}^{\mathsf{T}}\hat{\mathbf{x}}_q = \mathbf{b}^{\mathsf{T}}\mathbf{y} + \epsilon_q, \quad \forall q \in \mathcal{Q} && \text{(5c)}\\
& \|\mathbf{c}\|_N = 1. && \text{(5d)}
\end{aligned}
$$

This model specializes $\mathbf{GIO}(\hat{\mathcal{X}})$ to measure error in terms of scalar duality gap variables $(\epsilon_1,\ldots,\epsilon_Q)$. It can be recovered from the generalized model by appropriately selecting hyperparameters.

Proposition 1. *Let $\mu(\mathbf{c}) \in \mathbb{R}^n$ be a parameter satisfying $\|\mu(\mathbf{c})\|_\infty = 1$ and $\mu(\mathbf{c})^{\mathsf{T}}\mathbf{c} = 1$. A solution $\left(\mathbf{c}^*, \mathbf{y}^*, \epsilon_1^*, \ldots, \epsilon_Q^*\right)$ is optimal to $\mathbf{GIO}_{\mathrm{A}}(\hat{\mathcal{X}})$ if and only if $\left(\mathbf{c}^*, \mathbf{y}^*, \epsilon_1^*\mu(\mathbf{c}^*), \ldots, \epsilon_Q^*\mu(\mathbf{c}^*)\right)$ 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 $\boldsymbol{\epsilon}_q$ being a rescaling of a parameter $\mu(\mathbf{c})$ dependent only on the cost vector. Note that $\mu(\mathbf{c})$ is just a vehicle to aid the specialization of $\mathbf{GIO}(\hat{\mathcal{X}})$, and is useful primarily to interpret the solutions of $\mathbf{GIO}_{\mathrm{A}}(\hat{\mathcal{X}})$ in the broader 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 $\mu(\mathbf{c})^{\mathsf{T}}\mathbf{c} = 1$. Given a specific choice of $\|\cdot\|_N$, we can then propose a structured form for $\mu(\mathbf{c})$. For example, if $\|\cdot\|_N = \|\cdot\|_1$, we can set $\mu(\mathbf{c}) = \operatorname{sgn}(\mathbf{c})$ to be the sign vector of $\mathbf{c}$, ensuring that the two conditions on $\mu(\mathbf{c})$ are satisfied for all $\mathbf{c}$ with $\|\mathbf{c}\|_1 = 1$. Alternatively, if $\|\cdot\|_N = \|\cdot\|_\infty$, we can set $\mu(\mathbf{c}) = \operatorname{sgn}(c_{j^*})\mathbf{e}_{j^*}$ to be a signed j^*-th unit vector, where $j^* \in \arg\max_{j\in\mathcal{J}}\{|c_j|\}$.

General solution method. Since the normalization constraint is the sole non-convexity in $\mathbf{GIO}_{\mathrm{A}}(\hat{\mathcal{X}})$, this model can be solved exactly by polynomial decomposition. Here, effeciency depends on the choice of $\|\cdot\|_N$ (e.g., if $\|\cdot\|_N = \|\cdot\|_\infty$, then $\mathbf{GIO}_{\mathrm{A}}(\hat{\mathcal{X}})$ can be solved by solving $2n$ LPs).

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

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

The disjunctive problem (6), which is written for each j, can be separated into two LPs (one with the constraint $c_j = 1$ and the other with $c_j = -1$), thus totaling $2n$ LPs. In general though, such as when $\|\cdot\|_N = \|\cdot\|_1$, an exponential number of LPs will be required. However, there are special cases where the the 1-norm is an efficient choice for $\|\cdot\|_N$, such as when the cost vector is non-negative. Next, we discuss several special cases that simplify the solution approach for $\mathbf{GIO}_{\mathrm{A}}(\hat{\mathcal{X}})$.

Non-negative cost vectors. In many real-world applications, feasible cost vectors should be non-negative (i.e., $\mathcal{C} \subseteq \mathbb{R}^n_+$). Here, it is advantageous to set $\|\cdot\|_N = \|\cdot\|_1$, because the normalization constraint becomes $\mathbf{c}^{\mathsf{T}}\mathbf{1} = 1$ and $\mathbf{GIO}_{\mathrm{A}}(\hat{\mathcal{X}})$ simplifies to a single LP.

Feasible observed decisions. Most inverse optimization literature focuses 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 first 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 $\mathbf{GIO}_{\mathrm{A}}(\{\bar{\mathbf{x}}\})$.*

Proposition 2 combined with Theorem 1 implies that $\mathbf{GIO}_{\mathrm{A}}(\hat{\mathcal{X}})$ is solved analytically when $\hat{\mathcal{X}} \subset \mathcal{P}$.

Infeasible observed decisions. Finally, we address scenarios where the observed decisions all lie outside of the feasible region. We first consider the case where $\hat{\mathcal{X}}$ consists of a single, infeasible observed decision, $\hat{\mathbf{x}}$, in which case $\mathbf{GIO}_{\mathrm{A}}(\{\hat{\mathbf{x}}\})$ possesses an analytic solution.

Proposition 3. *Assume $\hat{\mathbf{x}} \notin \mathcal{P}$.*

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

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

and $\tilde{\mathbf{c}} = \mathbf{A}^\mathsf{T}\tilde{\mathbf{y}}$. 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. That is, 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 $\mathcal{H} = \left\{\mathbf{x} \mid \mathbf{c}^{*\mathsf{T}}\mathbf{x} = \mathbf{b}^\mathsf{T}\mathbf{y}^*\right\}$ 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 $\mathbf{a}_i^\mathsf{T}\hat{\mathbf{x}} > b_i$ for some i, then there always exists a supporting hyperplane that intersects with $\hat{\mathbf{x}}$ (see Figure 1a for an example). If $\mathbf{A}\hat{\mathbf{x}} \leq \mathbf{b}$, then no such supporting hyperplane exists. Instead, we show that with the alternate forward problem $\mathbf{FOA}(\mathbf{c}) := \min_{\mathbf{x}} \left\{-\mathbf{c}^\mathsf{T}\mathbf{x} \mid \mathbf{A}\mathbf{x} \leq \mathbf{b}\right\}$, 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. Since $\hat{\mathbf{x}}$ is feasible for $\mathbf{FOA}(\mathbf{c})$ by definition, $\mathbf{GIO}_\mathrm{A}(\{\hat{\mathbf{x}}\})$ can be solved via Theorem 1. The geometric insight is that the constraints of $\mathbf{FOA}(\mathbf{c})$ correspond to the nearest supporting hyperplanes of $\mathbf{FO}(\mathbf{c})$. Solving one problem solves the other (see Figure 1b, where $\hat{\mathbf{x}}$ projects to an infeasible point with respect to $\mathbf{FO}(\mathbf{c})$ that nevertheless has no duality gap). This geometric approach can be extended to the case of multiple infeasible decisions.

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 $\mathbf{FOA}(\mathbf{c})$.*

3.2. Relative duality gap

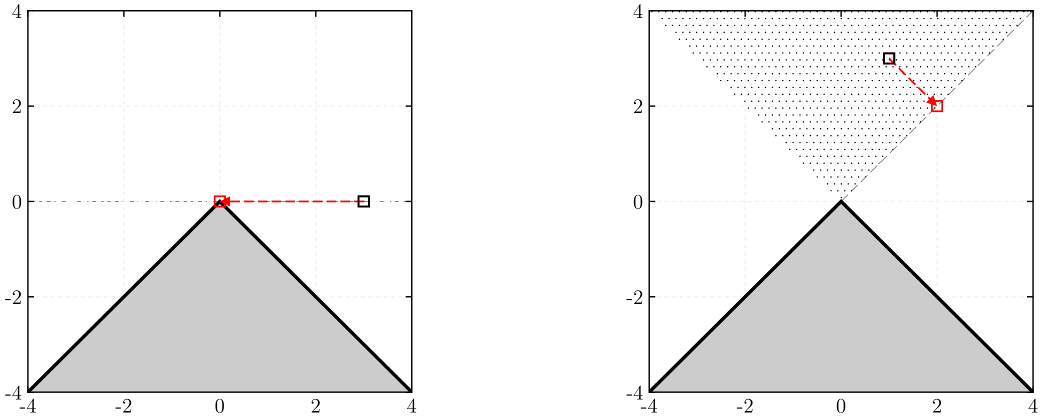
The relative duality gap variant minimizes the sum of the ratios between the duality gap for each decision and the imputed dual optimal value for the forward problem:

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

Figure 1 $\mathbf{GIO}_{\mathrm{A}}(\{\hat{\mathbf{x}}\})$ **with** $\mathbf{FO}(\mathbf{c})$ **shaded and** $\mathbf{FOA}(\mathbf{c})$ **hatched. Illustration of Proposition 3.**

(a) Illustration of Proposition 3 Part 1.

(b) Illustration of Proposition 3 Part 2.



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

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

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

Duality gap ratio variables ϵ_q replace the perturbation vectors used in the general formulation (4). These variables are well-defined except when the imputed forward problem has an optimal value of 0. In this subsection, we assume $\mathbf{b} \neq \mathbf{0}$ and that if $\mathbf{b}^{\mathsf{T}}\mathbf{y} = \mathbf{0}$ for some feasible $\mathbf{y}$, then $\epsilon_q := 1$.

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

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

General solution method. Unlike the absolute duality gap problem, which is non-convex only because of the normalization constraint, $\mathbf{GIO}_{\mathrm{R}}(\hat{\mathcal{X}})$ possesses an additional non-convexity due to a bilinear term in the duality gap constraint (8c). We first address the bilinearity by introducing three sub-problems. We then use polyhedral decomposition to address the normalization constraint.

Proposition 5. *Consider the following three problems:*

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

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

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

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 $\left(\mathbf{c}^*, \mathbf{y}^*, \epsilon_1^*, \ldots, \epsilon_Q^*\right)$ be an optimal solution for the corresponding problem. We assume $\epsilon_1^* = \cdots = \epsilon_Q^* = 1$ for $\mathbf{GIO}_{\mathrm{R}}^{0}(\hat{\mathcal{X}};K)$. 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 $\left(\mathbf{c}^*/\left\|\mathbf{c}^*\right\|_N, \mathbf{y}^*/\left\|\mathbf{c}^*\right\|_N, \epsilon_1^*, \ldots, \epsilon_Q^*\right)$.

Proof of Proposition 5. Let $(\hat{\mathbf{c}}, \hat{\mathbf{y}})$ be an optimal solution to $\mathbf{GIO}_{\mathrm{R}}(\hat{\mathcal{X}})$ and let

$$K = \begin{cases} 1/|\mathbf{b}^{\mathsf{T}}\hat{\mathbf{y}}| & \text{if } \mathbf{b}^{\mathsf{T}}\hat{\mathbf{y}} \neq 0 \\ 1/\hat{\mathbf{y}}^{\mathsf{T}}\mathbf{1} & \text{otherwise.} \end{cases} \tag{12}$$

We omit the variables $(\epsilon_1, \ldots, \epsilon_Q)$ when writing optimal solutions for conciseness. First, we show that $(\hat{\mathbf{c}}, \hat{\mathbf{y}})$ maps to a corresponding feasible solution for one of $\mathbf{GIO}_{\mathrm{R}}^{+}(\hat{\mathcal{X}};K)$, $\mathbf{GIO}_{\mathrm{R}}^{-}(\hat{\mathcal{X}};K)$, or $\mathbf{GIO}_{\mathrm{R}}^{0}(\hat{\mathcal{X}};K)$ with the same objective value. Conversely, every feasible solution to formulations (9)–(11) has a corresponding feasible solution in $\mathbf{GIO}_{\mathrm{R}}(\hat{\mathcal{X}})$ with the same objective value.

First, suppose $\mathbf{b}^{\mathsf{T}}\hat{\mathbf{y}} > 0$ and consider $(\tilde{\mathbf{c}}, \tilde{\mathbf{y}}) = (\hat{\mathbf{c}}/\mathbf{b}^{\mathsf{T}}\hat{\mathbf{y}}, \hat{\mathbf{y}}/\mathbf{b}^{\mathsf{T}}\hat{\mathbf{y}})$. This solution is feasible to $\mathbf{GIO}_{\mathrm{R}}^{+}(\hat{\mathcal{X}};K)$ as $\mathbf{b}^{\mathsf{T}}\tilde{\mathbf{y}} = 1$ and $\|\tilde{\mathbf{c}}\|_N = K$. Furthermore, by substituting $\tilde{\mathbf{c}} = \hat{\mathbf{c}}/\mathbf{b}^{\mathsf{T}}\hat{\mathbf{y}}$, we see that the objective value of this solution for $\mathbf{GIO}_{\mathrm{R}}^{+}(\hat{\mathcal{X}};K)$ is equal to the optimal value for $\mathbf{GIO}_{\mathrm{R}}(\hat{\mathcal{X}})$: $\sum_{q=1}^{Q} \left|\tilde{\mathbf{c}}^{\mathsf{T}}\hat{\mathbf{x}}_q - 1\right| = \sum_{q=1}^{Q} \left|(\hat{\mathbf{c}}^{\mathsf{T}}\hat{\mathbf{x}}_q)/(\mathbf{b}^{\mathsf{T}}\hat{\mathbf{y}}) - 1\right|$. Similarly, when $\mathbf{b}^{\mathsf{T}}\hat{\mathbf{y}} < 0$, we construct $(\tilde{\mathbf{c}}, \tilde{\mathbf{y}}) = (\hat{\mathbf{c}}/|\mathbf{b}^{\mathsf{T}}\hat{\mathbf{y}}|, \hat{\mathbf{y}}/|\mathbf{b}^{\mathsf{T}}\hat{\mathbf{y}}|)$, which is feasible to $\mathbf{GIO}_{\mathrm{R}}^{-}(\hat{\mathcal{X}};K)$ and incurs the same objective value as the optimal value of $\mathbf{GIO}_{\mathrm{R}}(\hat{\mathcal{X}})$. Finally, if $\mathbf{b}^{\mathsf{T}}\hat{\mathbf{y}} = 0$, then the optimal value of $\mathbf{GIO}_{\mathrm{R}}(\hat{\mathcal{X}})$ is 0. Let $(\tilde{\mathbf{c}}, \tilde{\mathbf{y}}) = (\hat{\mathbf{c}}/\hat{\mathbf{y}}^{\mathsf{T}}\mathbf{1}, \hat{\mathbf{y}}/\hat{\mathbf{y}}^{\mathsf{T}}\mathbf{1})$. It is straightforward to show that this solution is feasible for $\mathbf{GIO}^{0}_{\mathrm{R}}(\hat{\mathcal{X}}; K)$. Thus, an optimal solution to $\mathbf{GIO}_{\mathrm{R}}(\hat{\mathcal{X}})$ can be scaled to construct a solution that is feasible for exactly one of the formulations (9)–(11).

The converse is proven by showing that every feasible solution of (9)–(11) can be scaled to a feasible solution of $\mathbf{GIO}_{\mathrm{R}}(\hat{\mathcal{X}})$. Let $(\tilde{\mathbf{c}}, \tilde{\mathbf{y}})$ be a feasible solution to one of (9)–(11), and let $(\hat{\mathbf{c}}, \hat{\mathbf{y}}) = (\tilde{\mathbf{c}}/\left\|\tilde{\mathbf{c}}\right\|_N, \tilde{\mathbf{y}}/\left\|\tilde{\mathbf{c}}\right\|_N)$. This solution is feasible for $\mathbf{GIO}_{\mathrm{R}}(\hat{\mathcal{X}})$ with the same objective function value.

In terms of objective value, all feasible solutions of $\mathbf{GIO}^{+}_{\mathrm{R}}(\hat{\mathcal{X}}; K)$, $\mathbf{GIO}^{-}_{\mathrm{R}}(\hat{\mathcal{X}}; K)$, and $\mathbf{GIO}^{0}_{\mathrm{R}}(\hat{\mathcal{X}}; K)$ have a one-to-one correspondence with feasible solutions of $\mathbf{GIO}_{\mathrm{R}}(\hat{\mathcal{X}})$ and the best optimal solution to formulations (9)–(11) can be scaled to an optimal solution for $\mathbf{GIO}_{\mathrm{R}}(\hat{\mathcal{X}})$. □

While $\mathbf{GIO}_{\mathrm{R}}(\hat{\mathcal{X}})$ can be reformulated into three sub-problems, there are two primary concerns. First, the proof of Proposition 5 requires selecting an appropriate value for the parameter K. In practice, an appropriate value is selected by solving the following auxiliary problem:

$$\begin{aligned} \underset{\mathbf{y}}{\text{maximize}} \quad & \max\left\{|\mathbf{b}^{\mathsf{T}}\mathbf{y}|, \mathbf{y}^{\mathsf{T}}\mathbf{1}\right\} \\ \text{subject to} \quad & \left\|\mathbf{A}^{\mathsf{T}}\mathbf{y}\right\|_N = 1, \; \mathbf{y} \geq \mathbf{0}. \end{aligned} \tag{13}$$

We refer to formulation (13) as the auxiliary problem for $\mathbf{GIO}_{\mathrm{R}}(\hat{\mathcal{X}})$. The auxiliary problem can be written as three optimization problems, each with the same constraints as (13) but a different objective: $\mathbf{b}^{\mathsf{T}}\mathbf{y}$, $-\mathbf{b}^{\mathsf{T}}\mathbf{y}$, and $\mathbf{y}^{\mathsf{T}}\mathbf{1}$. Since the auxiliary problem has a normalization constraint similar to the one in $\mathbf{GIO}_{\mathrm{A}}(\hat{\mathcal{X}})$, we can use the same methods to solve it. Let K^* be defined as the reciprocal of the optimal value of the auxiliary problem. Note that K^* is well-defined, because any feasible $\mathbf{y}$ to (13) must satisfy $\mathbf{y}^{\mathsf{T}}\mathbf{1} > 0$. This parameter value is then sufficient to solve $\mathbf{GIO}_{\mathrm{R}}(\hat{\mathcal{X}})$.

Corollary 2. *Let z^* be the optimal value of the auxiliary problem* (13) *and let $K^* = 1/z^*$. Then, Proposition 5 holds for any $K \geq K^*$.*

The second concern is that $\mathbf{GIO}^{+}_{\mathrm{R}}(\hat{\mathcal{X}}; K)$, $\mathbf{GIO}^{-}_{\mathrm{R}}(\hat{\mathcal{X}}; K)$, and $\mathbf{GIO}^{0}_{\mathrm{R}}(\hat{\mathcal{X}}; K)$ are still non-convex due to the normalization constraint $\left\|\mathbf{c}\right\|_N \geq K$. As in $\mathbf{GIO}_{\mathrm{A}}(\hat{\mathcal{X}})$ however, with an appropriate choice of $\|\cdot\|_N$, these problems can be solved via polyhedral decomposition. For example if $\|\cdot\|_N = \|\cdot\|_\infty$, $\mathbf{GIO}^+_{\mathrm{R}}(\hat{\mathcal{X}};K)$ is solved by enumerating over the $2n$ linear programs $\mathbf{GIO}^+_{\mathrm{R}}(\hat{\mathcal{X}};K,j)$, where the norm constraint is replaced by $c_j \geq K$ or $c_j \leq -K$. The process is equivalent to that used in Theorem 2. We repeat this approach to decompose $\mathbf{GIO}^-_{\mathrm{R}}(\hat{\mathcal{X}};K)$ and $\mathbf{GIO}^0_{\mathrm{R}}(\hat{\mathcal{X}};K)$.

Solving $\mathbf{GIO}_{\mathrm{R}}(\hat{\mathcal{X}})$ is often computationally taxing, due to the fact that we must first determine an appropriate K before considering the reformulations into $\mathbf{GIO}^+_{\mathrm{R}}(\hat{\mathcal{X}};K)$, $\mathbf{GIO}^-_{\mathrm{R}}(\hat{\mathcal{X}};K)$, and $\mathbf{GIO}^0_{\mathrm{R}}(\hat{\mathcal{X}};K)$. In practice, it is often easier to try solving an LP relaxation of these sub-problems (by removing the normalization constraint). If an optimal solution of the LP relaxation can be found for which the optimal $\mathbf{c}^* \neq \mathbf{0}$, then this is often sufficient for an application. When the LP relaxation fails to do so, we must consider the more complex general method. In the rest of this section, we examine several cases leading to more efficient solution approaches for $\mathbf{GIO}_{\mathrm{R}}(\hat{\mathcal{X}})$.

Feasible observed decisions. Consider the scenario where the observed decisions are all feasible for the forward problem. As in the absolute duality gap case, the relative duality gap model reduces to a single-point problem, which has an analytic solution using 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 $\mathbf{GIO}_{\mathrm{R}}(\{\bar{\mathbf{x}}\})$.*

Infeasible observed decisions. For a single infeasible point, the solution to $\mathbf{GIO}_{\mathrm{R}}(\{\hat{\mathbf{x}}\})$ is broken into two cases. The proofs (omitted) are similar to the absolute duality gap case.

Proposition 7. *Assume $\hat{\mathbf{x}} \notin \mathcal{P}$.*

1. *If $\hat{\mathbf{x}}$ satisfies $\mathbf{a}_i^\intercal \hat{\mathbf{x}} > b_i$ 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 $\mathbf{GIO}_{\mathrm{R}}(\{\hat{\mathbf{x}}\})$.*

Proposition 7 is the relative duality gap analogue of Proposition 3, as it provides an analytic solution for $\mathbf{GIO}_{\mathrm{R}}(\{\hat{\mathbf{x}}\})$ when $\hat{\mathbf{x}}$ is infeasible. Thus, geometric insights similar to those of the absolute duality gap case are applicable here. Furthermore, Part 2 of Proposition 7 can be extended to multiple observations. The proof (omitted) is derived by extending the proof of Proposition 7 in the same way that Corollary 1 extends the proof of Proposition 3.

Table 1 Summary of the two variants of GIO($\hat{\mathcal{X}}$).

	$\|\cdot\|$	$\|\cdot\|_N$	$\mathcal{C}$	$\mathcal{E}_q, \forall q \in \mathcal{Q}$	Solution approach
$\mathbf{GIO}_{\mathrm{A}}(\hat{\mathcal{X}})$	$\|\cdot\|_\infty$	$\|\cdot\|_N$	$\mathbb{R}^n$	$\{\boldsymbol{\epsilon}_q \mid \boldsymbol{\epsilon}_q = \epsilon_q \mu(\mathbf{c})\}$	Polyhedral decomposition
$\mathbf{GIO}_{\mathrm{R}}(\hat{\mathcal{X}})$	$\|\cdot\|_\infty / \|\mathbf{b}^\intercal \mathbf{y}\|$	$\|\cdot\|_N$	$\mathbb{R}^n$	$\left\{\boldsymbol{\epsilon}_q \,\middle|\, \boldsymbol{\epsilon}_q = \mathbf{b}^\intercal \mathbf{y}\,(\epsilon_q - 1)\,\mu(\mathbf{c})\right\}$	Three sub-problems

Corollary 3. *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* $\mathbf{FOA}(\mathbf{c})$.

3.3. Model discussion and comparisons

Table 1 summarizes how the three variants specialize from $\mathbf{GIO}(\hat{\mathcal{X}})$ using the model hyperparameters $(\|\cdot\|, \|\cdot\|_N, \mathcal{C}, \mathcal{E}_1, \ldots, \mathcal{E}_Q)$. In this subsection, we compare the variants against the literature.

$\mathbf{GIO}_{\mathrm{A}}(\hat{\mathcal{X}})$ can be seen as special cases of previous inverse convex optimization models in the literature (Bertsimas et al. 2015, 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 instance-specific parameter given to the inverse optimizer. Thus, the data set is now $\hat{\mathcal{X}} = \{(\hat{\mathbf{x}}_1, \hat{\mathbf{u}}_1), \ldots, (\hat{\mathbf{x}}_Q, \hat{\mathbf{u}}_Q)\}$. To translate inverse convex optimization to our setting, we remove $\mathbf{u}$ and define $f(\mathbf{x};\mathbf{c}) = \mathbf{c}^\intercal \mathbf{x}$ and $g(\mathbf{x};\mathbf{c}) = \mathbf{b} - \mathbf{A}\mathbf{x}$ and obtain a linear forward problem with a fixed feasible set. Bertsimas et al. (2015) study inverse optimization by minimizing a first-order variational inequality (i.e., the absolute duality gap in LPs) and construct a convex inverse problem without any normalization constraint (e.g., $\|\mathbf{c}\|_N = 1$). Although normalization can be avoided if an application-specific $\mathcal{C}$ is convex and excludes $\mathbf{0}$, setting $f(\mathbf{x};\mathbf{u},\mathbf{c}) = \mathbf{c}^\intercal \mathbf{x}$ with a general $\mathcal{C} = \mathbb{R}^n$ implies that $(\mathbf{c}, \mathbf{y}, \epsilon_1, \ldots, \epsilon_Q) = (\mathbf{0}, \mathbf{0}, 0, \ldots, 0)$ is a trivially optimal solution (Chan et al. 2018). Esfahani et al. (2018) also study absolute duality gap inverse linear optimization with a normalization constraint. However, they study this in the context of a distributionally robust inverse convex optimization problem; their formulation decomposes to a finite set of conic optimization problems after polyhedral decomposition. On the other hand, we pursue a geometric exploration that yields several special cases with efficient solutions (e.g., Propositions 2 and 3, as well as Corollary 1).

We remark that the relative duality gap variant has not been studied in inverse convex optimization. However, $\mathbf{GIO}_{\mathrm{R}}(\hat{\mathcal{X}})$ is often competitive with $\mathbf{GIO}_{\mathrm{A}}(\hat{\mathcal{X}})$ in terms of solution efficiency. Moreover, our case study on radiation therapy treatment planning in Section 5 compares several models and shows that $\mathbf{GIO}_{\mathrm{R}}(\hat{\mathcal{X}})$ outperforms $\mathbf{GIO}_{\mathrm{A}}(\hat{\mathcal{X}})$ on the downstream clinical task.

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 1, Figure 2 (below), simply assessing the aggregate error from the inverse optimization model may not provide a complete picture of model fitness. Moreover, a context-free goodness of fit metric is useful when comparing different forward models for a given data set, or when faced with an unfamiliar application.

There exist previously proposed fitness measures for inverse optimization, but they are less general compared to the one developed in this section. For example, the measures were either context-specific (e.g., Troutt et al. (2006) for production planning and Chow and Recker (2012) for traffic assignment) or applicable only for a single feasible observed decision (Chan et al. 2018). Our metric expands on this latter metric, referred to as the *coefficient of complementarity* and denoted $\rho(\{\hat{\mathbf{x}}\})$, which provides a scale-free, unitless measure of goodness of fit, analogous to the *coefficient of determination* R^2 in linear regression. It was defined as

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

The numerator of the ratio is the residual error from the estimated cost vector, equivalently 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 (i.e., $\boldsymbol{\epsilon}_i = \hat{\mathbf{x}} - \pi_i(\hat{\mathbf{x}})$ 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 to a baseline corresponding to the average error induced by m candidate optimal cost vectors (recall Theorem 1).

In this section, we generalize $\rho(\{\hat{\mathbf{x}}\})$ for use with $\mathbf{GIO}(\hat{\mathcal{X}})$. When it is clear, we omit the data set from notation and denote the absolute and relative duality gap variants as ρ_{A} and ρ_{R}, respectively.

4.1. Multi-point coefficient of complementarity

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

$$\rho(\hat{\mathcal{X}}) = 1 - \frac{\sum_{q=1}^{Q} \left\| \boldsymbol{\epsilon}_q^* \right\|}{\frac{1}{m}\sum_{i=1}^{m}\left(\sum_{q=1}^{Q} \left\| \boldsymbol{\epsilon}_{q,i} \right\|\right)}. \tag{14}$$

The numerator is the optimal value of $\mathbf{GIO}(\hat{\mathcal{X}})$, i.e., the residual error from an optimal solution to the inverse optimization problem. The denominator terms $\sum_{q=1}^{Q} \|\boldsymbol{\epsilon}_{q,i}\|$ 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 absolute duality gap, $\mathbf{GIO}_{\mathrm{A}}(\hat{\mathcal{X}})$,

$$\sum_{q=1}^{Q} \|\boldsymbol{\epsilon}_{q,i}\| = \sum_{q=1}^{Q} \frac{\left|\mathbf{a}_i^{\intercal}\hat{\mathbf{x}}_q - b_i\right|}{\|\mathbf{a}_i\|_1}. \tag{15}$$

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

$$\sum_{q=1}^{Q} \|\boldsymbol{\epsilon}_{q,i}\| = \sum_{q=1}^{Q} \left|\frac{\mathbf{a}_i^{\intercal}\hat{\mathbf{x}}_q}{b_i} - 1\right|. \tag{16}$$

The denominator in $\rho(\hat{\mathcal{X}})$ represents a baseline against which the inverse solution is measured. Our choice of baseline is a direct extension from the single-point case, where an optimal cost vector can be found by selecting amongst 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 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., (15) and (16)). Third, this definition provides a direct generalization of the single-point metric, inheriting several attractive mathematical properties that we present in Section 4.2. In addition, for objective space models, following the same development as Propositions 2 and 6, the multi-point coefficient of complementarity is equal to the single-point version when all data points are feasible.

Proposition 8. *Let $\bar{\mathbf{x}}$ be the centroid of $\hat{\mathcal{X}} \subset \mathcal{P}$. Then, $\rho_{\mathrm{A}}(\hat{\mathcal{X}}) = \rho_{\mathrm{A}}(\{\bar{\mathbf{x}}\})$ and $\rho_{\mathrm{R}}(\hat{\mathcal{X}}) = \rho_{\mathrm{R}}(\{\bar{\mathbf{x}}\})$.*

4.2. Properties of ρ

Chan et al. (2018) demonstrated that $\rho(\{\hat{\mathbf{x}}\})$ possesses several mathematical properties analogous to the properties of R^2 from linear regression. The properties also hold for the more general $\rho(\hat{\mathcal{X}})$.

Theorem 3. *The following properties hold for ρ defined in* (14)*:*

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

The first property underlines how ρ fits into the generalized inverse optimization framework. One can select any cost vector and calculate the induced error and ρ value with respect to the data $\hat{\mathcal{X}}$. However, a solution obtained via $\mathbf{GIO}(\hat{\mathcal{X}})$ 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, mathematically rigorous framework for model fitting and evaluation in inverse optimization.

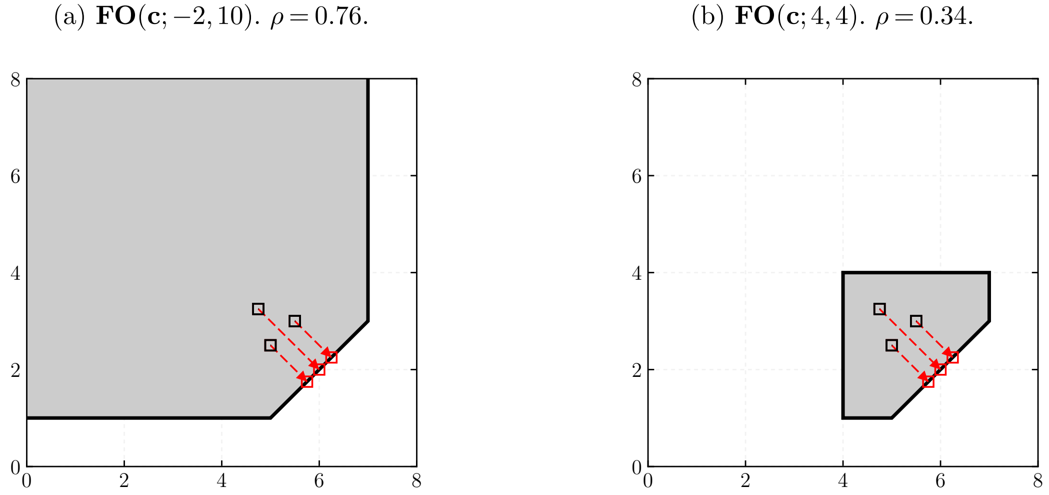
The second property makes ρ easily interpretable as a measure of goodness of fit, with higher values indicating better fit. Note that $\rho = 1$ if and only if $\sum_{q=1}^{Q} \left\|\boldsymbol{\epsilon}_q^*\right\| = 0$ (i.e., every point in $\hat{\mathcal{X}}$ is an optimal solution to $\mathbf{FO}(\mathbf{c}^*)$). 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 $\sum_{q=1}^{Q} \left\|\boldsymbol{\epsilon}_q^*\right\| = \sum_{q=1}^{Q} \left\|\boldsymbol{\epsilon}_{q,i}\right\|$ 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 practitioner. This property is 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 one of the weaknesses of R^2, namely the potential of overfitting. Thus, when using ρ to compare the goodness of fit of several inverse optimization models, a user should ensure that higher values of ρ represent true improvements in fit, rather than artificial increases that lack generalizability.

4.3. Numerical examples

Next, we present two examples highlighting behavioral properties and usefulness of the coefficient of complementarity. Example 1 illustrates the value of using ρ instead of an unnormalized measure

Figure 2 $\mathbf{GIO}_{\mathrm{A}}(\hat{\mathcal{X}})$ **with two** $\mathbf{FO}(\mathbf{c};u,v)$**, same** $\mathbf{c}^*$ **and** $\boldsymbol{\epsilon}^*$**, but different** ρ**. Illustration of Example 1.**


(a) $\mathbf{FO}(\mathbf{c};-2,10)$. $\rho = 0.76$.

(b) $\mathbf{FO}(\mathbf{c};4,4)$. $\rho = 0.34$.

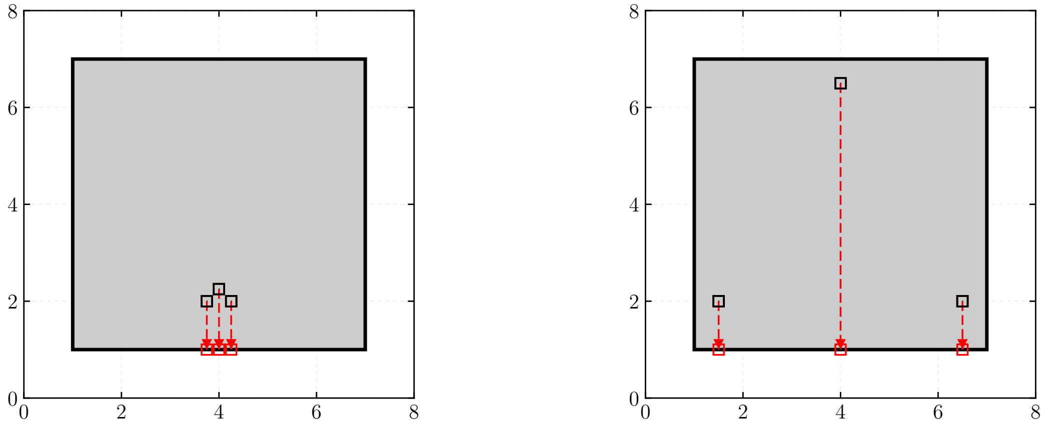
of error when comparing different models. Intuitively, a given error within a larger feasible region indicates better fit than the same error within a smaller feasible region: ρ captures this intuition by measuring error in the context of the geometry of the feasible set of the forward model.

Example 1 *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* $\hat{\mathcal{X}} = \{(5,2.5),(4.75,3.75),(5.5,3)\}$. $\mathbf{GIO}_{\mathrm{A}}(\hat{\mathcal{X}})$ *in both cases yields* $\mathbf{c}^* = (-0.5,0.5)$ *and* $\sum_{q=1}^{3}|\epsilon_q^*| = 2.75$. *However, for* $\mathbf{FO}(\mathbf{c};-2,10)$, $\rho = 0.76$, *while for* $\mathbf{FO}(\mathbf{c};4,4)$, $\rho = 0.34$. *In Fig. 2a, the data points are closer to the bottom facet, relative to the other facets, while in Fig. 2b, the data points are near the "center" of the polyhedron rather than a specific facet.*

Next, we present an example that demonstrates the behaviour of ρ when the forward problem remains the same but the data set changes. In Example 2 case (a), the points are close together and all project to the same facet, resulting in the same optimal cost vector. In (b), the points are further apart, each with a different preferred cost vector, but the aggregate error is minimized by selecting a cost vector that is not preferred by any of them. In the latter case, the inverse solution is a compromise between the preferences of the individual data points, resulting in poorer fit.

Example 2 *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* $\mathbf{GIO}_{\mathrm{A}}(\hat{\mathcal{X}}_2)$

Figure 3 $\mathbf{GIO}_{\mathrm{A}}(\hat{\mathcal{X}}_1)$ **and** $\mathbf{GIO}_{\mathrm{A}}(\hat{\mathcal{X}}_2)$ **for the same** $\mathbf{FO}(\mathbf{c})$**. Illustration of Example 2.**



(a) $\hat{\mathcal{X}}_1 = \{(3.75, 2), (4, 2.25), (4.25, 2)\}$. $\rho = 0.64$.

(b) $\hat{\mathcal{X}}_2 = \{(1.5, 2), (4, 6.25), (6.5, 2)\}$. $\rho = 0.17$.

impute $\mathbf{c}^* = (0, 1)$*. In Fig. 3a, the data points are closer together and all clearly prefer the bottom facet, while in Fig. 3b, the points are further apart, with each point biased towards a different facet. We find* $\rho = 0.64$ *and* $\rho = 0.17$ *for the two problems, respectively.*

5. Knowledge-based automated planning in radiation therapy

In this section, we implement several $\mathbf{GIO}(\hat{\mathcal{X}})$ variants and demonstrate the use of ρ in the context of intensity-modulated radiation therapy (IMRT) treatment planning in head-and-neck cancer. IMRT, where a linear accelerator (LINAC) delivers beamlets of radiation to a tumor, is one of the most widely-used cancer treatment techniques. A treatment is typically designed using a multi-objective optimization model, but model parameters for a given patient (e.g., objective weights) are not known *a priori*. Consequently, standard clinical practice is to iterate between treatment planners (for design) and oncologists (for approval) via ad-hoc parameter tuning.

As an alternative to trial-and-error, knowledge-based automated planning (KBAP) has been proposed to streamline the treatment design process (Sharpe et al. 2014). KBAP consists of two components: (1) a prediction model that, for a given patient, predicts an appropriate dose distribution; and (2) an optimization model that generates a deliverable treatment plan that closely replicates the predicted dose distribution. Early approaches predicted summary statistics of the dose distribution using techniques that include look-up tables, linear regression, and principle component analysis (Wu et al. 2009, Zhu et al. 2011, Appenzoller et al. 2012). Recent approaches predict the entire 3D dose distribution using random forests (McIntosh and Purdie 2016) and neural networks (Mahmood et al. 2018, Kearney et al. 2018, Babier et al. 2018c).

There are two main approaches to the optimization component in KBAP. The first is "dose mimicking", which amounts to minimizing a 2-norm to determine a deliverable plan that closely reproduces the original predicted dose (McIntosh et al. 2017). The second approach uses inverse optimization: first estimating the parameters of a treatment (i.e., forward) optimization model assuming that the predicted dose is the "observed decision", and then solving the forward optimization model with these parameters to generate the actual treatment plan (Lee et al. 2013, Chan et al. 2014, Boutilier et al. 2015, Babier et al. 2018b).

To date, the prediction stage in KBAP have exclusively output a single prediction, which is then converted into a treatment plan via optimization. However, different prediction models lead to plans with different trade-offs between the clinical evaluation criteria. Consequently, rather than using a single prediction in KBAP, we harness an ensemble of different predictions to generate a single treatment plan that balances the trade-offs and improves upon the single-point models. To be clear, instead of averaging many predictions (like in a random forest), we keep each prediction in the ensemble separate, and use inverse optimization over all of them simultaneously. Until now, KBAP optimization has never been used to generate a single treatment plan from multiple predictions.

Our numerical results suggest that this is a promising avenue of research. Given a set of high-quality predictions, our inverse optimization model generates treatments that better achieve clinical trade-offs than when given a single prediction. In particular, we first experiment over the two different objective space models to show that the relative duality gap model is the appropriate choice for this application. We then show that our multi-point (ensemble) approach produces new treatment plans with better overall performance compared to traditional single-point KBAP models. However, as our final multi-point model requires clinically-driven model engineering, we show that ρ offers a useful domain-independent alternative when choosing models.

Using a set of different predictions from different learning models also has an intuitive clinical interpretation. In practice, different clinicians will typically generate different treatment plans for the same patient. With machine learning models that produce different predictions, we imitate the situation where clinicians have different decision-making criteria. Multi-point inverse optimization can compromise between different predictions to construct better plans than any single-point model.

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. With each patient k, we associate parameters $(\mathbf{C}_k, \mathbf{A}_k, \mathbf{b}_k)$ and the corresponding multi-objective linear optimization problem $\mathbf{RT\text{–}FO}(\boldsymbol{\alpha}_k) : \min_{\mathbf{x}} \left\{ \boldsymbol{\alpha}_k^\top \mathbf{C}_k \mathbf{x} \;\middle|\; \mathbf{A}_k \mathbf{x} \geq \mathbf{b}_k, \mathbf{x} \geq \mathbf{0} \right\}$, where $\mathbf{C}_k$ is the matrix whose rows represent different linear cost vectors and $\boldsymbol{\alpha}_k$ is the vector of objective weights. The decision vector is composed of two subvectors, $\mathbf{x} = (\mathbf{w}, \mathbf{d})$, where $\mathbf{w}$ represents the intensity of each beamlet of radiation and $\mathbf{d}$ is the dose to be delivered to every voxel (4 mm × 4 mm × 2 mm volumetric pixel) of the patient's body, computed as a linear transformation of $\mathbf{w}$. Note this multi-objective model fits into the $\mathbf{GIO}(\hat{\mathcal{X}}_k)$ framework by specifying the set of feasible cost vectors for patient k as $\mathcal{C}_k = \left\{ \mathbf{C}_k^\top \boldsymbol{\alpha} \;\middle|\; \boldsymbol{\alpha} \geq \mathbf{0} \right\}$. Furthermore, the optimization problem for each patient is distinct. Given a specific patient, the feasible set is fixed and a single treatment optimization problem is solved. The multi-point nature arises from the multiple dose predictions comprising each patient's data set $\mathbf{GIO}(\hat{\mathcal{X}}_k)$. Note that the use of predictions, not actual observed decisions, also constitutes an innovative application of inverse optimization. We treat the predictions as functions of the decisions, which are used to learn parameters of the underlying decision-generating process.

We first train four different 3D dose prediction models from the literature, labeled Random Forest (RF), 2-D RGB GAN, 2-D GANCER, and 3-D GANCER (Babier et al. 2018b,c, 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. 2018c). 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 $\hat{\mathcal{X}}_k = \left\{ \hat{\mathbf{d}}_{k,1}, \ldots, \hat{\mathbf{d}}_{k,8} \right\}$ be data for each patient-specific problem.

For each patient k in the test set, we implement absolute and relative duality gap models, referred to as $\mathbf{RT–IO}_{\mathrm{A}}(\hat{\mathcal{X}}_k)$ and $\mathbf{RT–IO}_{\mathrm{R}}(\hat{\mathcal{X}}_k)$, 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 $\mathbf{RT–FO}(\boldsymbol{\alpha}_k^*)$ to determine the beamlets $\mathbf{w}_k^*$ and dose $\mathbf{d}_k^*$. The dose $\mathbf{d}_k^*$ is then evaluated using different clinical criteria. Detailed descriptions of the prediction models and the formulation of the inverse optimization models are provided in Appendix A.

5.2. The value of multi-point inverse optimization

In practice, a suite of quantitative metrics are evaluated to assess whether 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; also Babier et al. (2018a)). These criteria 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, each of which is assigned a criterion specifying the minimum dose that at least 99% of its volume should receive. For every criteria, we evaluate whether the generated plan satisfied the clinical trade-off (i.e., if the ground truth clinical plan satisfied a criteria, whether the generated plan also satisfy it.) We say that a generated plan replicated the clinical practice if the generated plan satisfied all of the same criteria as the clinical plan.

The columns of Table 2 list the proportion of plans generated by $\mathbf{RT–IO}_{\mathrm{A}}(\hat{\mathcal{X}})$ and $\mathbf{RT–IO}_{\mathrm{R}}(\hat{\mathcal{X}})$ that satisfied clinical trade-offs. The 'All' row reflects the percentage of plans that perfectly replicated all of the trade-offs as the clinical practice. We first use all eight predictions to solve $\mathbf{RT–IO}_{\mathrm{A}}(\hat{\mathcal{X}})$ (column 3) and $\mathbf{RT–IO}_{\mathrm{R}}(\hat{\mathcal{X}})$ (column 4). $\mathbf{RT–IO}_{\mathrm{R}}(\hat{\mathcal{X}})$ substantially outperforms the $\mathbf{RT–IO}_{\mathrm{A}}(\hat{\mathcal{X}})$ over every criterion, suggesting that the absolute duality gap model is not well-suited to this specific application. This result is consistent with results observed for single-point inverse optimization in IMRT (Chan et al. 2014, 2018, Goli et al. 2018) and we conjecture that it is due

Table 2 The percentage of final plans of each KBAP population that satisfy the same clinical criteria as the corresponding clinical plans. OARs are assigned a mean or maximum dose criteria depending on relevance. PTVs are assigned criteria to the 99%-ile.

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

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 relatively small adjustments to objectives with high values; thus, it has a hard time balancing different clinical criteria.

Although $\mathbf{RT\text{–}IO}_{\mathrm{R}}(\hat{\mathcal{X}})$ with all eight predictions is generally effective at satisfying the OAR criteria, these plans sacrifice the PTV criteria, especially PTV56. We hypothesize that this 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 $\mathbf{RT\text{–}IO}_{\mathrm{R}}(\{\hat{\mathbf{x}}\})$ using each of the individual predictions is shown in Table 3. Depending on which prediction to use, the single-point KBAP population varies from 10.9% to 80.5% in terms of satisfying the PTV56 criteria. When evaluating on whether they can replicate all of the clinical criteria, this variance lies between 44.8% to 80.5%, suggesting that some KBAP models are definitively making poorer trade-offs than others. More importantly, each prediction has different imputed cost vectors, which as discussed in Section 4, may lead to poor model fitness if the points are used together in multi-point $\mathbf{RT}\text{–}\mathbf{IO}_{\mathrm{R}}(\hat{\mathcal{X}}_k)$.

We next vary the set of predictions given to the multi-point inverse optimization model and examine how performance changes. Using the 'All' criteria and previous work (see Mahmood et al. (2018), Babier et al. (2018c)), we rank the eight predictions from worst to best as follows: 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. Note that we rank RF-sc. above 2-D RGB GAN-sc.; this is to prevent over-emphasizing GAN-based models in the top predictions. Using our ranking, we implement $\mathbf{RT}\text{–}\mathbf{IO}_{\mathrm{R}}(\hat{\mathcal{X}})$, but with data sets of decreasing size; we sequentially remove the two worst predictions and re-solve $\mathbf{RT}\text{–}\mathbf{IO}_{\mathrm{R}}(\hat{\mathcal{X}})$. The results using the best six, four, and two predictions are provided in columns 5–7 of Table 2. By removing the worst two predictions, the 6 Pts. model markedly improves on PTV criteria, while satisfying almost all OAR criteria at the same rate as the original model. This alone results in an additional 15% of the final plans being able to replicate clinical trade-offs. Similarly, the 4 Pts. model improves over the 6 Pts. model, by achieving near perfect PTV criteria satisfaction while mostly preserving OAR performance. In fact, this model now outperforms the best single-point model, 3-D GANCER-sc. (see cf. Table 3). Interestingly, performance does not continue to improve in the 2 Pts. model. This model uses two predictions (2-D GANCER-sc. and 3-D GANCER-sc.) that individually achieve high PTV satisfaction, but fail to do so when combined in the multi-point model. We conjecture that the 2 Pts. model reaches a local minimum in criteria satisfaction because the two predictions lie in different regions of the feasible set.

Overall, we demonstrate that the multi-point model creates significant value as a conduit for turning an ensemble of predictions into a single treatment plan. An interesting second point is that the data must be carefully chosen to truly maximize performance in KBAP.

Table 3 **The percentage of single-point inverse optimization plans of each KBAP population that satisfy the same clinical criteria as the clinical plans.**

Structure	Criteria (Gy)	$\mathbf{RT\text{–}IO}_{\mathrm{R}}(\{\hat{\mathbf{x}}_q\})$							
		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	99%-ile ≥ 66.5	81.0	36.2	81.0	69.0	63.8	91.4	98.3	100
PTV63	99%-ile ≥ 59.9	92.0	100	100	100	98.0	98.0	100	100
PTV56	99%-ile ≥ 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

Table 4 ρ **and criteria satisfaction for the best, average, and worst subsets of 6, 4, and 2 Pts. Highest performing models are bolded.**

(a) ρ.

	Best	Average	Worst
6 Pts.	**0.666**	0.626	0.637
4 Pts.	**0.734**	0.672	0.555
2 Pts.	**0.851**	0.664	0.631

(b) All structures criteria satisfaction (%).

	Best	Average	Worst
6 Pts.	**74.9**	57.5	51.7
4 Pts.	**83.9**	62.1	30.1
2 Pts.	**70.1**	60.9	42.5

5.3. Using ρ to validate the best subset of the data

Given a ranking of the eight predictions over their performance in a single-point KBAP pipeline we showed in the previous subsection how reducing the data set to include only the four best predictions resulted in treatment plans that on average achieved better trade-offs than the best single-point model (i.e., $\mathbf{RT}\text{–}\mathbf{IO}_{\mathrm{R}}(\{\hat{\mathbf{x}}\})$ using only 3-D GANCER-sc. predictions). When the data contains points from different regions of the feasible set, an inverse optimization model is unable to determine a good cost vector to describe all points. However, the coefficient of complementarity can determine model fitness in terms of the points (e.g., see Example 2). Here, we demonstrate a clinical analogue by validating the data selection in the Best 6 Pts., 4 Pts., and 2 Pts. models.

We consider three different variants for each of the 6 Pts., 4 Pts., and 2 Pts. models by selecting the best, average, and worst subsets, according to the pre-defined ranking. Note that we are not studying the effect of data set size, but rather the effect of data set quality. Table 4a compares ρ across models with varying quality of input predictions. For a fixed data set size (i.e., along the rows), the Best model always yields the highest ρ, which suggests that the Best predictions are the best fit for the clinical forward model. The high ρ values correlate with the predictions that perform best in isolation (i.e., the single-point scenario). Furthermore in Table 4b, we show that the clinical criteria satisfaction rates for each of the models also reflect the same trends as ρ. Since ρ is a general metric, we can evaluate the model quality for a set number of points without domain specific knowledge, and come to the same conclusion as we do using the clinical criteria, which are domain-specific evaluation metrics. An interesting result is that the data set with the best fit (2 Pts.) is not necessarily the one that results in the best final treatment plan (4 Pts.). This result is due to the fact that some predictions may be overly optimistic, which is tempered after conversion to a treatment in the optimization step of KBAP. Overall, these experiments demonstrate that in the absence of domain-specific knowledge, practitioners can use ρ to evaluate model-data fit.

6. Conclusion

Inverse linear optimization is an increasingly popular model-fitting paradigm for estimating the cost vector of an optimization problem when given observed decisions. Motivated by ensemble methods in machine learning, we develop a framework that uses a collection of decisions for a single, fixed problem to estimate a consensus cost vector. Here, the data can be noisy observations 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. 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 multi-point inverse optimization in the automated construction of radiation therapy treatment plans. In contrast to the current state-of-the-art, which generates treatment plans from individual dose predictions, we use a family of different dose predictions, each with different characteristics, to form superior treatment plans with better clinical trade-offs. Finally, while constructing the best inverse optimization model requires careful clinical expertise, we show how our goodness-of-fit metric can provide domain-independent validation of our model engineering process. Beyond the specific application presented in this paper, we believe there will be new applications of predict-then-inversely optimize frameworks that can build on our foundation.

References

Appenzoller LM, Michalski JM, Thorstad WL, Mutic S, Moore KL (2012) Predicting dose-volume histograms for organs-at-risk in imrt planning. *Medical physics* 39(12):7446–7461.

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

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 Med Biol* 63(10):105004.

Babier A, Mahmood R, McNiven AL, Diamant A, Chan TCY (2018c) Knowledge-based automated planning with three-dimensional generative adversarial networks. *arXiv preprint arXiv:1812.09309* .

Bertsimas D, Gupta V, Paschalidis IC (2015) Data-driven Estimation In Equilibrium Using Inverse Optimization. *Mathematical Programming* 153(2):595–633.

Boutilier JJ, Lee T, Craig T, Sharpe MB, Chan TCY (2015) Models for predicting objective function weights in prostate cancer imrt. *Medical Physics* 42(4):1586–1595.

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

Chan TCY, Craig T, Lee T, Sharpe MB (2014) Generalized Inverse Multiobjective Optimization with Application to Cancer Therapy. *Operations Research* 62(3):680–695.

Chan TCY, Lee T, Terekhov D (2018) Inverse optimization: Closed-form solutions, geometry, and goodness of fit. *Management Science* .

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

Craft D, Suss P, Bortfeld T (2007) The tradeoff between treatment plan quality and required number of monitor units in intensity-modulated radiotherapy. *International Journal of Radiation Oncology, Biology, Physics* 67:1596–1605.

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

Goli A (2015) *Sensitivity and Stability Analysis for Inverse Optimization with Applications in Intensity-Modulated Radiation Therapy*. Master's thesis, University of Toronto.

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 Med 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 Med Biol* 63(23):235022.

Lee T, Hammad M, Chan TCY, Craig T, Sharpe MB (2013) Predicting objective function weights from patient anatomy in prostate imrt treatment planning. *Medical Physics* 40(12):121706.

Mahmood R, Babier A, McNiven A, Diamant A, Chan TCY (2018) Automated treatment planning in radiation therapy using generative adversarial networks. *Proceedings of the 3rd Machine Learning for Healthcare Conference*, volume 85 of *Proceedings of Machine Learning Research*, 484–499 (PMLR).

Mangasarian OL (1999) Arbitrary-norm separating plane. *Operations Research Letters* 24(1):15–23.

McIntosh C, Purdie TG (2016) Voxel-based dose prediction with multi-patient atlas selection for automated radiotherapy treatment planning. *Physics in Medicine & Biology* 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. *Physics in Medicine & Biology* 62(15):5926.

Saez-Gallego J, Morales JM (2017) Short-term forecasting of price-responsive loads using inverse optimization. *IEEE Transactions on Smart Grid* PP(99):1–1, ISSN 1949-3053.

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 physics* 41(12).

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

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

Wu B, Ricchetti F, Sanguineti G, Kazhdan M, Simari P, Chuang M, Taylor R, Jacques R, McNutt T (2009) Patient geometry-driven information retrieval for IMRT treatment plan quality control. *Med Phys* 36(12):5497–505.

Zhao Q, Stettner A, Reznik E, Segrè D, Paschalidis IC (2015) Learning cellular objectives from fluxes by inverse optimization. *2015 54th IEEE Conference on Decision and Control (CDC)*, 1271–1276.

Zhu X, Ge Y, Li T, Thongphiew D, Yin F, Wu QJ (2011) A planning quality evaluation tool for prostate adaptive IMRT based on machine learning. *Med Phys* 38(2):719–26.

Appendix A: Knowledge based treatment planning in radiation therapy

IMRT treatment is delivered by a linear accelerator (LINAC) that delivers high-energy X-rays from different angles to a patient's tumor. The patient's body is discretized into tiny voxels in order to calculate the dose delivered to each voxel. The optimization problem for designing an IMRT treatment plan is over $\mathbf{x} = (\mathbf{w}, \mathbf{d})$, composed of the beamlets and the dose delivered (in Gy) from the intensities of the beamlets, respectively.

The forward model in our experiments is a modified version of the one used by Babier et al. (2018b). Let $\mathcal{B}$ be the index set of beamlets and w_b be the radiation intensity of beamlet $b \in \mathcal{B}$. Similarly, let $\mathcal{V}$ be the index set of voxels in a patient and d_v be the dose of radiation delivered to voxel $v \in \mathcal{V}$. Dose is calculated linearly, i.e., $d_v = \sum_{b \in \mathcal{B}} D_{v,b} w_b$, where $D_{v,b}$ is the dose influence of beamlet b on voxel v.

For each patient, let $\mathcal{T}$ denote the index set of the three planning target volumes (PTVs) with different prescription doses (i.e., PTV56, PTV63, and PTV70 with 56 Gy, 63 Gy, and 70Gy as prescription doses, respectively) and let $\mathcal{O}$ denote the index set of the eight surrounding OARs (i.e., brain stem, spinal cord, right parotid, left parotid, larynx, esophagus, mandible, and limPostNeck). Note that the limPostNeck is an artificially defined region used solely in optimization; it does not possess a clinical criteria. For each $t \in \mathcal{T}$ and $o \in \mathcal{O}$, let $\mathcal{V}_t$ and $\mathcal{V}_o$ denote the set of voxels corresponding to the given target or OARs, respectively.

A.1. Forward objectives

The IMRT forward problem includes 65 different objectives each minimizing some feature of the dose delivered to an OAR or PTV. For each OAR, we minimize the mean dose delivered, the maximum dose delivered, and the average dose above a threshold ϕ_o^θ. Here, ϕ_o^θ is a fraction θ of the average maximum dose to OAR o over the data set; we consider $\theta \in \Theta := \{0.25, 0.5, 0.75, 0.9, 0.975\}$. These objectives are computed as follows:

$$z_o^{\text{mean}} = \frac{1}{|\mathcal{V}_o|} \sum_{v \in \mathcal{V}_o} d_v, \quad \forall o \in \mathcal{O} \tag{17}$$

$$z_o^{\max} = \max_{v \in \mathcal{V}_o} \{d_v\}, \quad \forall o \in \mathcal{O} \tag{18}$$

$$z_o^{\text{thresh},\theta} = \frac{1}{|\mathcal{V}_o|} \sum_{v \in \mathcal{V}_o} \max\left\{0, d_v - \phi_o^\theta\right\}, \quad \forall \theta \in \Theta, \forall o \in \mathcal{O}. \tag{19}$$

Each PTV is assigned a prescribed dose ϕ_t, i.e., 56 Gy for PTV56, 63 Gy for PTV63, and 70 Gy for PTV70. For each PTV, we minimize the dose over the prescription, under the prescription, and the maximum dose delivered to the target, which can be computed as follows:

$$z_t^{\text{over}} = \frac{1}{|\mathcal{V}_t|} \sum_{v \in \mathcal{V}_t} \max\{0, d_v - \phi_t\}, \quad \forall t \in \mathcal{T} \tag{20}$$

$$z_t^{\text{under}} = \frac{1}{|\mathcal{V}_t|} \sum_{v \in \mathcal{V}_t} \max\{0, \phi_t - d_v\}, \quad \forall t \in \mathcal{T} \tag{21}$$

$$z_t^{\max} = \max_{v \in \mathcal{V}_t} \{d_v\}, \quad \forall t \in \mathcal{T}. \tag{22}$$

A.2. Forward constraints

In order to ensure that no OAR or PTV is prioritized by the objectives at a cost to the other organs, we assign a set of hard constraints for each structure. OARs are assigned constraints ensuring the mean and maximum doses do not exceed safety limits, whereas PTVs are constrained to ensure they receive baselines.

$$\text{Brain stem:} \qquad z_o^{\text{mean}} \le 30, \quad z_o^{\max} \le 53 \tag{23}$$

$$\text{Spinal cord:} \quad z_o^{\text{mean}} \leq 30, \quad z_o^{\text{max}} \leq 46 \tag{24}$$

$$\text{Left parotid:} \quad z_o^{\text{mean}} \leq 68, \quad z_o^{\text{max}} \leq 77 \tag{25}$$

$$\text{Right parotid:} \quad z_o^{\text{mean}} \leq 68, \quad z_o^{\text{max}} \leq 78 \tag{26}$$

$$\text{Larynx:} \quad z_o^{\text{mean}} \leq 68, \quad z_o^{\text{max}} \leq 77 \tag{27}$$

$$\text{Esophagus:} \quad z_o^{\text{mean}} \leq 52, \quad z_o^{\text{max}} \leq 75 \tag{28}$$

$$\text{Mandible:} \quad z_o^{\text{mean}} \leq 63, \quad z_o^{\text{max}} \leq 76 \tag{29}$$

$$\text{limPostNeck:} \quad z_o^{\text{mean}} \leq 21, \quad z_o^{\text{max}} \leq 46 \tag{30}$$

$$\text{PTV56:} \quad z_t^{\text{mean}} \geq 58 \tag{31}$$

$$\text{PTV63:} \quad z_t^{\text{mean}} \geq 63 \tag{32}$$

$$\text{PTV70:} \quad z_t^{\text{mean}} \geq 69 \tag{33}$$

Note that we introduce a z_t^{mean} variable for the targets, analogous to z_o^{mean} in (17).

Finally, we include a constraint on the physical deliverability of the treatment plan. This constraint, known as the sum-of-positive-gradients (SPG), restricts the variation of doses from neighboring beamlets so that the resulting dose shape is deliverable (Craft et al. 2007). Let $a \in \mathcal{A}$ index each angle of the LINAC, $r \in \mathcal{R}_a$ index each row of the LINAC at that angle, and $\mathcal{B}_r$ index beamlets along that row. We add the constraint

$$\sum_{a \in \mathcal{A}} \max_{r \in \mathcal{R}_a} \left\{ \sum_{b \in \mathcal{B}_r} \max\{0, w_b - w_{b+1}\} \right\} \leq 55, \tag{34}$$

where $w_{b+1} = 0$ for the last beamlet in a row. The right-hand-side is 55 Gy as in Babier et al. (2018c).

A.3. Forward optimization problem

The final forward problem is then to minimize a weighted combination of the objectives:

$$\begin{aligned} \textbf{RT–FO}(\boldsymbol{\alpha}): \quad \underset{\mathbf{z},\mathbf{w},\mathbf{d}}{\text{minimize}} \quad & \sum_{o \in \mathcal{O}} \left(\alpha_o^{\text{mean}} z_o^{\text{mean}} + \alpha_o^{\text{max}} z_o^{\text{max}} + \sum_{\theta \in \Theta} \alpha_o^{\text{thresh},\theta} z_o^{\text{thresh},\theta} \right) + \\ & \sum_{t \in \mathcal{T}} \left(\alpha_t^{\text{over}} z_t^{\text{over}} + \alpha_t^{\text{under}} z_t^{\text{under}} + \alpha_t^{\text{max}} z_t^{\text{max}} \right) \\ \text{subject to} \quad & (17) - (34) \\ & \sum_{b \in \mathcal{B}} D_{v,b} w_b = d_v, \quad \forall v \in \mathcal{V} \\ & w_b, d_v \geq 0, \quad \forall b \in \mathcal{B}, \forall v \in \mathcal{V}. \end{aligned} \tag{35}$$

We compress the notation of the above forward problem to $\mathbf{FO}(\boldsymbol{\alpha}) : \min_{\mathbf{x}} \left\{ \boldsymbol{\alpha}^{\mathsf{T}}\mathbf{C}\mathbf{x} \,\middle|\, \mathbf{A}\mathbf{x} \geq \mathbf{b}, \mathbf{x} \geq \mathbf{0} \right\}$. This problem has several useful properties. Firstly under this notation, the matrix of objective functions $\mathbf{C}$ is non-negative. Furthermore, the constraint vector $\mathbf{b}$ is also non-negative. These properties are useful specifically as they allow for constructing almost entirely linear inverse optimization problems (see Section A.5 for details).

A.4. Generating a data set of predicted treatments

We use the training set of 130 patients to implement a series of machine learning models that were previously proposed in the KBAP literature. Each machine learning model takes as input a segmented CT image of the patients tumour site and predicts the dose distribution $\hat{\mathbf{d}}$. We briefly describe the models below:

1. **Random Forest:** A random forest that predicts the dose to be delivered to each voxel of the dose distribution $\hat{d}_v$ individually (Mahmood et al. 2018)

2. **2-D RGB GAN:** A generative adversarial network (GAN) that predicts an RGB image representation of each axial slice of the dose distribution individually (Mahmood et al. 2018)

3. **2-D GANCER:** A GAN that predicts the dose distribution of each axial slice (Babier et al. 2018c).

4. **3-D GANCER:** A GAN that predicts the dose distribution vector $\hat{\mathbf{d}}$ in one shot (Babier et al. 2018c).

Babier et al. (2018c) noted that plans predicted using the above models often delivered low dose (i.e., significantly spare healthy tissue) at the cost of failing prescription criteria for the PTVs, and implemented a rescaling method to create an improved prediction. They showed that treatment plans constructed using inverse optimization-based KBAP and the scaled dose would better satisfy prescription criteria while performing poorer on sparing healthy tissue. We implement the rescaling method on all predictions from the models, and use both the non-scaled and scaled predictions as input for the inverse optimization model. Thus, for each patient there is a data set of 8 dose distributions, i.e., $\hat{\mathscr{X}} = \{\hat{\mathbf{z}}_1, \ldots, \hat{\mathbf{z}}_8\}$. Note that we do not require $\hat{\mathbf{x}}_q = (\hat{\mathbf{w}}_q, \hat{\mathbf{d}}_q)$. Inverse optimization yields a weight vector $\boldsymbol{\alpha}_k$, with which we then re-solve $\mathbf{FO}(\boldsymbol{\alpha}_k)$.

A.5. Inverse optimization problems

In order to frame $\mathbf{FO}(\boldsymbol{\alpha})$ for generalized inverse optimization, we restrict imputed cost vectors to be in the image of $\mathbf{C}$, i.e., $\mathcal{C} = \left\{ \mathbf{C}^{\mathsf{T}}\boldsymbol{\alpha} \,\middle|\, \boldsymbol{\alpha} \geq \mathbf{0} \right\}$. Note that $\boldsymbol{\alpha} \geq \mathbf{0}$ is an application-specific constraint.A specific inverse optimization problem is then formulated by appropriately selecting the model hyperparameters $(\|\cdot\|, \|\cdot\|_N, \mathcal{C}, \mathcal{E}_1, \ldots, \mathcal{E}_Q)$ from $\mathbf{GIO}(\hat{\mathscr{X}})$. We use the default parameters, except with the custom $\mathcal{C}$ to ensure the objective function is a weighted combination of the different objectives. We set $\|\cdot\|_N = \|\cdot\|_1$.

A.5.1. Absolute duality gap Using Proposition 1 and our choice of $\mathcal{C}$, we formulate the following:

$$\begin{aligned}
\textbf{RT–IO}_{\text{A}}(\hat{\mathcal{X}}): \quad \min_{\boldsymbol{\alpha},\mathbf{y},\epsilon_1,\dots,\epsilon_Q} \quad & \sum_{q=1}^{Q} |\epsilon_q| \\
\text{s.t.} \quad & \mathbf{C}^\mathsf{T}\boldsymbol{\alpha} \geq \mathbf{A}^\mathsf{T}\mathbf{y}, \quad \mathbf{y} \geq \mathbf{0} \\
& \boldsymbol{\alpha}^\mathsf{T}\hat{\mathbf{z}}_q = \mathbf{b}^\mathsf{T}\mathbf{y} + \epsilon_q, \quad \forall q \in \mathcal{Q} \\
& (\mathbf{C}^\mathsf{T}\boldsymbol{\alpha})^\mathsf{T}\mathbf{1} = 1 \\
& \boldsymbol{\alpha} \geq \mathbf{0}.
\end{aligned} \tag{36}$$

$\textbf{RT–IO}_{\text{A}}(\hat{\mathcal{X}})$ is obtained by substituting $\mathbf{c} = \mathbf{C}^\mathsf{T}\boldsymbol{\alpha}$ and noting $\left\|\mathbf{C}^\mathsf{T}\boldsymbol{\alpha}\right\|_1 = \boldsymbol{\alpha}^\mathsf{T}\mathbf{C}\mathbf{1}$ when $\boldsymbol{\alpha} \geq \mathbf{0}$ and $\mathbf{C} \geq \mathbf{0}$.

A.5.2. Relative duality gap Using Proposition 4 and our specific choice of $\mathcal{C}$, we formulate a relative duality gap inverse optimization problem, for which we then form an LP relaxation.

$$\begin{aligned}
&\textbf{RT–IO}_{\text{R}}(\hat{\mathcal{X}}): \\
\min_{\boldsymbol{\alpha},\mathbf{y},\epsilon_1,\dots,\epsilon_Q} \quad & \sum_{q=1}^{Q} |\epsilon_q - 1| \\
\text{s.t.} \quad & \mathbf{C}^\mathsf{T}\boldsymbol{\alpha} \geq \mathbf{A}^\mathsf{T}\mathbf{y}, \quad \mathbf{y} \geq \mathbf{0} \\
& \boldsymbol{\alpha}^\mathsf{T}\hat{\mathbf{z}}_q = \epsilon_q\mathbf{b}^\mathsf{T}\mathbf{y}, \quad \forall q \in \mathcal{Q} \\
& (\mathbf{C}^\mathsf{T}\boldsymbol{\alpha})^\mathsf{T}\mathbf{1} = 1 \\
& \boldsymbol{\alpha} \geq \mathbf{0}.
\end{aligned} \tag{37}$$

$$\begin{aligned}
&\textbf{RT–IO}_{\text{R,LP}}(\hat{\mathcal{X}}): \\
\min_{\boldsymbol{\alpha},\mathbf{y},\epsilon_1,\dots,\epsilon_Q} \quad & \sum_{q=1}^{Q} |\epsilon_q - 1| \\
\text{s.t.} \quad & \mathbf{C}^\mathsf{T}\boldsymbol{\alpha} \geq \mathbf{A}^\mathsf{T}\mathbf{y}, \quad \mathbf{y} \geq \mathbf{0} \\
& \boldsymbol{\alpha}^\mathsf{T}\hat{\mathbf{z}}_q = \epsilon_q, \quad \forall q \in \mathcal{Q} \\
& \mathbf{b}^\mathsf{T}\mathbf{y} = 1 \\
& \boldsymbol{\alpha} \geq \mathbf{0}.
\end{aligned} \tag{38}$$

We first solve the LP relaxation of $\textbf{RT–IO}_{\text{R}}(\hat{\mathcal{X}})$, stated above as $\textbf{RT–IO}_{\text{R,LP}}(\hat{\mathcal{X}})$. Note that this relaxation is application-specific analogue of relaxing $\textbf{GIO}^{+}_{\text{R}}(\hat{\mathcal{X}})$, which is only one of the three reformulations of the relative duality gap problem. Note that the analogue to $\textbf{GIO}^{-}_{\text{R}}(\hat{\mathcal{X}})$ is infeasible; in our application, $\mathbf{b} \geq \mathbf{0}$ implying $\mathbf{b}^\mathsf{T}\mathbf{y} \geq \mathbf{0}$ for all $\mathbf{y} \geq \mathbf{0}$. Similarly, the application-specific analogue of $\textbf{GIO}^{0}_{\text{R}}(\hat{\mathcal{X}})$ in practice is often infeasible or generates plans that perform poorly on the clinical criteria satisfaction metrics compared to $\textbf{RT–IO}_{\text{R,LP}}(\hat{\mathcal{X}})$. Recall that $\textbf{GIO}^{0}_{\text{R}}(\hat{\mathcal{X}})$ requires $\mathbf{c}^\mathsf{T}\hat{\mathbf{x}}_q = 0$ for all $q \in \mathcal{Q}$. In the application-specific analogue (where the constraint is $\boldsymbol{\alpha}^\mathsf{T}\hat{\mathbf{z}}_q = 0$), both $\boldsymbol{\alpha} \geq 0$ and $\hat{\mathbf{z}}_q \geq 0$, which means that the problem is feasible only when there exists an element of $\hat{\mathbf{z}}_q$ that is equal to 0 for all of the predictions. This situation could only occur for the threshold objectives (19)–(21). In our case, $\textbf{GIO}^{0}_{\text{R}}(\hat{\mathcal{X}})$ is infeasible for every patient in the data set.