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

EC.1. Proofs of Statements

Proof of Proposition 1. For any $\mathbf{c}$, setting each $\boldsymbol{\epsilon}_q = \epsilon_q \mu(\mathbf{c})$ implies $\|\boldsymbol{\epsilon}_q\|_\infty = |\epsilon_q| \, \|\mu(\mathbf{c})\|_\infty = |\epsilon_q|$. Thus, (4a) becomes (5a). Similarly, (4c) becomes (5c), since $\mathbf{c}^\intercal \boldsymbol{\epsilon}_q = \epsilon_q \mathbf{c}^\intercal \mu(\mathbf{c}) = \epsilon_q$. Then, any feasible solution to $\mathbf{GIO}(\hat{\mathcal{X}})$ with the suggested hyperparameters yields a feasible solution to $\mathbf{GIO}_{\mathrm{A}}(\hat{\mathcal{X}})$ and vice versa, with the same objective value. □

Proof of Theorem 2. Let $j^* \in \arg\max_{j \in \mathcal{J}} \left\{ |c_j^*| \right\}$, implying $|c_{j^*}^*| = 1$. Then, $\left(\mathbf{c}^*, \mathbf{y}^*, \epsilon_1^*, \ldots, \epsilon_Q^*\right)$ is feasible to $\mathbf{GIO}_{\mathrm{A}}(\hat{\mathcal{X}}; j^*)$. Conversely, for any $j \in \mathcal{J}$, every feasible solution to $\mathbf{GIO}_{\mathrm{A}}(\hat{\mathcal{X}}; j)$ is feasible to $\mathbf{GIO}_{\mathrm{A}}(\hat{\mathcal{X}})$, so all optimal solutions to each $\mathbf{GIO}_{\mathrm{A}}(\hat{\mathcal{X}}; j)$ lie in the feasible set of $\mathbf{GIO}_{\mathrm{A}}(\hat{\mathcal{X}})$. □

Proof of Proposition 2. If all observations are feasible, then by weak duality $\epsilon_q \geq 0 \; \forall q \in \mathcal{Q}$, and we can simplify the objective function $\sum_{q=1}^{Q} |\epsilon_q| = \sum_{q=1}^{Q} \epsilon_q = \sum_{q=1}^{Q} (\mathbf{c}^\intercal \hat{\mathbf{x}}_q - \mathbf{b}^\intercal \mathbf{y}) = (\mathbf{c}^\intercal \bar{\mathbf{x}} - \mathbf{b}^\intercal \mathbf{y}) \, Q$, where the last equality follows by the definition of the centroid (i.e., $\bar{\mathbf{x}} = \sum_{q=1}^{Q} \hat{\mathbf{x}}_q / Q$). We similarly compress constraint (5c) to a single constraint for $\bar{\mathbf{x}}$, resulting in $\mathbf{GIO}_{\mathrm{A}}(\{\bar{\mathbf{x}}\})$. □

Proof of Proposition 3.

1. Assume without loss of generality that there exist $i, j \in \mathcal{I}$ such that $\mathbf{a}_i^\intercal \hat{\mathbf{x}} > b_i$ and $\mathbf{a}_j^\intercal \hat{\mathbf{x}} < b_j$, respectively. The corresponding $\tilde{y}$ defined in (7) satisfies the strong duality constraint (5c) with $\epsilon = 0$. Furthermore, $(\tilde{\mathbf{c}}, \tilde{\mathbf{y}})$ satisfy the duality feasibility constraints (5b) by construction. We normalize the solution to satisfy constraint (5d). The normalized solution still satisfies all other constraints. This solution is feasible for $\mathbf{GIO}_{\mathrm{A}}(\{\hat{\mathbf{x}}\})$ with zero cost and is thus optimal.
2. Here, the duality gap is non-positive (i.e., $\epsilon \leq 0$). We rewrite the single-point version of (5) with $\delta = -\epsilon$, shown in model (EC.1) below. Now consider the forward problem $\min_{\mathbf{x}} \{-\mathbf{c}^\intercal \mathbf{x} \mid \mathbf{A}\mathbf{x} \leq \mathbf{b}\}$ with the observed solution $\hat{\mathbf{x}}$ and the corresponding inverse optimization model (EC.2).

$$\begin{aligned} \underset{\mathbf{c},\mathbf{y},\delta}{\text{minimize}} \quad & \delta \\ \text{subject to} \quad & \mathbf{A}^\mathsf{T}\mathbf{y} = \mathbf{c},\ \mathbf{y} \geq \mathbf{0} \\ & \mathbf{c}^\mathsf{T}\hat{\mathbf{x}} = \mathbf{b}^\mathsf{T}\mathbf{y} - \delta \\ & \|\mathbf{c}\|_N = 1. \end{aligned} \tag{EC.1}$$

$$\begin{aligned} \underset{\mathbf{c},\mathbf{y},\gamma}{\text{minimize}} \quad & |\gamma| \\ \text{subject to} \quad & \mathbf{A}^\mathsf{T}\mathbf{y} = \mathbf{c},\ \mathbf{y} \geq \mathbf{0} \\ & -\mathbf{c}^\mathsf{T}\hat{\mathbf{x}} = -\mathbf{b}^\mathsf{T}\mathbf{y} + \gamma \\ & \|\mathbf{c}\|_N = 1. \end{aligned} \tag{EC.2}$$

By assumption, $\hat{\mathbf{x}}$ is feasible for the above-defined forward problem and therefore, $\gamma \geq 0$ in (EC.2). Consequently, formulation (EC.1) is equivalent to (EC.2) after removing the absolute value in the objective and rearranging the duality gap constraint. We can solve formulation (EC.2) using Theorem 1, arriving at an optimal solution for the original inverse optimization problem. □

Proof of Corollary 1. Since all observations are infeasible for the initial forward problem, the duality gap terms are all non-positive (i.e., $\epsilon_q \leq 0$ for all $q \in \mathcal{Q}$). As such, we use the same argument as used in Prop. 3 Part 2 to show that the formulation of $\mathbf{GIO}_\mathrm{A}(\hat{\mathcal{X}})$ is equivalent to the formulation of an absolute duality gap inverse optimization problem over the alternative forward problem $\min_{\mathbf{x}}\{-\mathbf{c}^\mathsf{T}\mathbf{x} \mid \mathbf{A}\mathbf{x} \leq \mathbf{b}\}$. As $\hat{\mathcal{X}} \subset \{\mathbf{x} \mid \mathbf{A}\mathbf{x} \leq \mathbf{b}\}$, Proposition 2 reduces the problem to $\mathbf{GIO}_\mathrm{A}(\{\bar{\mathbf{x}}\})$. □

Proof of Proposition 4. For any $\mathbf{c}$, setting $\boldsymbol{\epsilon}_q = \mathbf{b}^\mathsf{T}\mathbf{y}\,(\epsilon_q - 1)\,\mu(\mathbf{c})$ forces $\|\boldsymbol{\epsilon}_q\|_\infty / |\mathbf{b}^\mathsf{T}\mathbf{y}| = |\epsilon_q - 1|$, giving us the objective (8a). The same substitution into (4c) gives the strong duality constraint (8c). Thus, every feasible solution of $\mathbf{GIO}_\mathrm{R}(\hat{\mathcal{X}})$ has a corresponding feasible solution in $\mathbf{GIO}(\hat{\mathcal{X}})$ (after setting the hyperparameters), and vice versa, with the same objective value. □

Remark EC.1. Proposition 4 addresses the case where $\mathbf{b}^\mathsf{T}\mathbf{y}^* \neq 0$ only. However, if $\mathbf{b}^\mathsf{T}\mathbf{y}^* = 0$, $\mathbf{GIO}_\mathrm{R}(\hat{\mathcal{X}})$ and $\mathbf{GIO}(\hat{\mathcal{X}})$ are still equivalent in that they both yield an optimal value of 0. To see this, suppose that an optimal solution to $\mathbf{GIO}_\mathrm{R}(\hat{\mathcal{X}})$ satisfies $\mathbf{b}^\mathsf{T}\mathbf{y}^* = 0$. Then for all $q \in \mathcal{Q}$, $\mathbf{c}^{*\mathsf{T}}\hat{\mathbf{x}}_q = 0$ and since ϵ_q becomes a free variable, we set it to 1 and obtain an optimal value of 0. On the other hand, we can use the same $(\mathbf{c}^*, \mathbf{y}^*, \mathbf{0}, \ldots, \mathbf{0})$ as a feasible solution to $\mathbf{GIO}(\hat{\mathcal{X}})$ and observe that setting $\boldsymbol{\epsilon}_q = \mathbf{0}$ for all $q \in \mathcal{Q}$ satisfies the strong duality constraint, giving an optimal value of 0.

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

$$K = \begin{cases} 1/|\mathbf{b}^\mathsf{T}\hat{\mathbf{y}}| & \text{if } \mathbf{b}^\mathsf{T}\hat{\mathbf{y}} \neq 0 \\ 1/\hat{\mathbf{y}}^\mathsf{T}\mathbf{1} & \text{otherwise.} \end{cases} \tag{EC.3}$$

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 $\mathbf{b}^\intercal\hat{\mathbf{y}} > 0$ and consider $(\tilde{\mathbf{c}}, \tilde{\mathbf{y}}) = (\hat{\mathbf{c}}/\mathbf{b}^\intercal\hat{\mathbf{y}}, \hat{\mathbf{y}}/\mathbf{b}^\intercal\hat{\mathbf{y}})$. This solution is feasible to $\mathbf{GIO}_{\mathrm{R}}^{+}(\hat{\mathcal{X}}; K)$ as $\mathbf{b}^\intercal\tilde{\mathbf{y}} = 1$ and $\|\tilde{\mathbf{c}}\|_N = K$. Furthermore, by substituting $\tilde{\mathbf{c}} = \hat{\mathbf{c}}/\mathbf{b}^\intercal\hat{\mathbf{y}}$, 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 $\mathbf{GIO}_{\mathrm{R}}(\hat{\mathcal{X}})$: $\sum_{q=1}^{Q} \left|\tilde{\mathbf{c}}^\intercal\hat{\mathbf{x}}_q - 1\right| = \sum_{q=1}^{Q} \left|(\hat{\mathbf{c}}^\intercal\hat{\mathbf{x}}_q)\,/\,(\mathbf{b}^\intercal\hat{\mathbf{y}}) - 1\right|$. Similarly, when $\mathbf{b}^\intercal\hat{\mathbf{y}} < 0$, we construct $(\tilde{\mathbf{c}}, \tilde{\mathbf{y}}) = (\hat{\mathbf{c}}/|\mathbf{b}^\intercal\hat{\mathbf{y}}|, \hat{\mathbf{y}}/|\mathbf{b}^\intercal\hat{\mathbf{y}}|)$, 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 $\mathbf{b}^\intercal\hat{\mathbf{y}} = 0$, then the optimal value of $\mathbf{GIO}_{\mathrm{R}}(\hat{\mathcal{X}})$ is 0. Let $(\tilde{\mathbf{c}}, \tilde{\mathbf{y}}) = (\hat{\mathbf{c}}/\hat{\mathbf{y}}^\intercal\mathbf{1}, \hat{\mathbf{y}}/\hat{\mathbf{y}}^\intercal\mathbf{1})$. It is straightforward to show that this solution is feasible for $\mathbf{GIO}_{\mathrm{R}}^{0}(\hat{\mathcal{X}}; K)$. 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 $(\hat{\mathbf{c}}, \hat{\mathbf{y}}) = (\tilde{\mathbf{c}}/\|\tilde{\mathbf{c}}\|_N, \tilde{\mathbf{y}}/\|\tilde{\mathbf{c}}\|_N)$. 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 $\mathbf{GIO}_{\mathrm{R}}^{0}(\hat{\mathcal{X}}; K)$ 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 $\mathbf{GIO}_{\mathrm{R}}(\hat{\mathcal{X}})$. ☐

Proof of Proposition 6. When all of the observed points are feasible, $\mathbf{c}^\intercal\hat{\mathbf{x}}_q - \mathbf{b}^\intercal\mathbf{y} \geq 0$, $\forall q \in \mathcal{Q}$. Thus, objective (8a) becomes $\sum_{q=1}^{Q} |\epsilon_q - 1| = \sum_{q=1}^{Q} \frac{\mathbf{c}^\intercal\hat{\mathbf{x}}_q - \mathbf{b}^\intercal\mathbf{y}}{|\mathbf{b}^\intercal\mathbf{y}|} = Q\left(\frac{\mathbf{c}^\intercal\bar{\mathbf{x}} - \mathbf{b}^\intercal\mathbf{y}}{|\mathbf{b}^\intercal\mathbf{y}|}\right)$. Noting that $\bar{\mathbf{x}}$ must also be feasible, the last term equals the objective for $\mathbf{GIO}_{\mathrm{R}}(\{\bar{\mathbf{x}}\})$. ☐

Proof of Lemma 1. Without loss of generality, assume that $\|\mathbf{a}_i\|_N = 1$ for all $i \in \mathcal{I}$. Solution (15) is feasible to $\mathbf{GIO}_p(\hat{\mathcal{X}})$ for all $i \in \mathcal{I}$. We show that for any feasible solution that is not of the form (15), there exists a feasible solution of that form whose objective value is at least as good.

Consider a feasible solution $(\tilde{\mathbf{c}}, \tilde{\mathbf{y}}, \tilde{\boldsymbol{\epsilon}}_1, \ldots, \tilde{\boldsymbol{\epsilon}}_Q)$ to $\mathbf{GIO}_p(\hat{\mathcal{X}})$, where $\tilde{\mathbf{y}} \neq \mathbf{e}_i$ for any $i \in \mathcal{I}$. Without loss of generality, assume $\tilde{y}_1, \ldots, \tilde{y}_k > 0$ for some $1 < k \leq m$ and let $\mathcal{K} = \{1, \ldots, k\}$ denote the corresponding index set. Let $\tilde{\mathbf{x}}_q = \hat{\mathbf{x}}_q - \tilde{\boldsymbol{\epsilon}}_q$ denote the perturbed decision for all $q \in \mathcal{Q}$. The primal feasibility constraint (13d) implies that $\mathbf{A}\tilde{\mathbf{x}}_q \geq \mathbf{b}$ for all $q \in \mathcal{Q}$. The strong duality constraint (13c) implies that for all $q \in \mathcal{Q}$, $0 = \mathbf{c}^\mathsf{T}\tilde{\mathbf{x}}_q - \mathbf{b}^\mathsf{T}\tilde{\mathbf{y}} = \sum_{i=1}^k \tilde{y}_i (\mathbf{a}_i^\mathsf{T}\tilde{\mathbf{x}}_q - b_i)$, which follows from substituting $\tilde{\mathbf{c}} = \sum_{i=1}^k \tilde{y}_i \mathbf{a}_i$. Using the non-negativity of $\tilde{\mathbf{y}}$ and primal feasibility (i.e., $\mathbf{a}_i^\mathsf{T}\tilde{\mathbf{x}}_q \geq b_i$ for all $i \in \mathcal{I}$), we see that $\tilde{\mathbf{x}}_q$ for all $q \in \mathcal{Q}$ are feasible solutions to the feasible projection problem (14) for each $i \in \mathcal{K}$.

Let $(\hat{\mathbf{c}}, \hat{\mathbf{y}}, \hat{\boldsymbol{\epsilon}}_1, \ldots, \hat{\boldsymbol{\epsilon}}_Q) = (\mathbf{a}_{i^*}, \mathbf{e}_{i^*}, \hat{\mathbf{x}}_1 - \psi_{i^*}(\hat{\mathbf{x}}_1), \ldots, \hat{\mathbf{x}}_Q - \psi_{i^*}(\hat{\mathbf{x}}_Q))$ for an arbitrary index $i^* \in \mathcal{K}$. For all $q \in \mathcal{Q}$, $\psi_{i^*}(\hat{\mathbf{x}}_q)$ is, by definition, an optimal solution to (14). Therefore, we have $\sum_{q=1}^Q \|\hat{\boldsymbol{\epsilon}}_q\|_p = \sum_{q=1}^Q \|\hat{\mathbf{x}}_q - \psi_{i^*}(\hat{\mathbf{x}}_q)\|_p \leq \sum_{q=1}^Q \|\tilde{\boldsymbol{\epsilon}}_q\|_p$, with the inequality following from the optimality of (14). Thus, given any feasible solution to $\mathbf{GIO}_p(\hat{\mathcal{X}})$ not of the form defined in (15), we can construct a feasible solution of the form (15) with the objective value at least as good as the original. □

Proof of Theorem 3. For each i, the inner optimization problem produces solutions with the structure in (15). Thus, the inner optimization problems, along with the corresponding $(\mathbf{c}, \mathbf{y})$ enumerate all possible solutions to $\mathbf{GIO}_p(\hat{\mathcal{X}})$ with the structure in (15). By Lemma 1, we select the one yielding the lowest objective value. □

Proof of Theorem 4. First note that due to the dominance between p-norms, (i.e., $\|\boldsymbol{\epsilon}\|_p \geq \|\boldsymbol{\epsilon}\|_\infty$) we have $z_p^* \geq z_\infty^*$, since the choice of p only affects the objective and the two problems share the same feasible set. We then lower bound the optimal value of $\mathbf{GIO}_\infty(\hat{\mathcal{X}})$ using Theorem 3:

$$\left.\begin{aligned} \min_{i \in \mathcal{I}} \min_{\boldsymbol{\epsilon}_{1,i}, \ldots, \boldsymbol{\epsilon}_{Q,i}} \quad & \sum_{q=1}^Q \|\boldsymbol{\epsilon}_{q,i}\|_\infty \\ \text{s.t.} \quad & \mathbf{A}(\hat{\mathbf{x}}_q - \boldsymbol{\epsilon}_{q,i}) \geq \mathbf{b}_i, \forall q \in \mathcal{Q} \\ & \mathbf{a}_i^\mathsf{T}(\hat{\mathbf{x}}_q - \boldsymbol{\epsilon}_{q,i}) = b_i, \forall q \in \mathcal{Q} \end{aligned}\right\} = \min_{i \in \mathcal{I}} \left\{ \sum_{q=1}^Q \|\hat{\mathbf{x}}_q - \psi_i(\hat{\mathbf{x}}_q)\|_\infty \right\} \tag{EC.4}$$

$$\geq \min_{i \in \mathcal{I}} \left\{ \sum_{q=1}^Q \|\hat{\mathbf{x}}_q - \pi_i(\hat{\mathbf{x}}_q)\|_\infty \right\} \tag{EC.5}$$

$$= \min_{i \in \mathcal{I}} \left\{ \sum_{q=1}^Q \frac{|\mathbf{a}_i^\mathsf{T}\hat{\mathbf{x}}_q - b_i|}{\|\mathbf{a}_i\|_1} \right\} \tag{EC.6}$$

$$= \min_{i \in \mathcal{I}} \left\{ Q \left(\frac{\mathbf{a}_i^\mathsf{T}\bar{\mathbf{x}} - b_i}{\|\mathbf{a}_i\|_1} \right) \right\}. \tag{EC.7}$$

The inequality in (EC.5) comes from the fact that the projection problem (2) is a relaxation of the feasible projection problem (14), by removing the feasibility constraint. The equality of (EC.6) comes from Mangasarian (1999) (e.g., see Theorem 1), which provides the analytic optimal value of the projection problem. Because $\hat{\mathbf{x}}_q \in \mathcal{P}$ for all $\hat{\mathbf{x}}_q \in \hat{\mathcal{X}}$, we bypass the absolute values to average. Note that (EC.7) is equal to the optimal value of $\mathbf{GIO}(\{\bar{\mathbf{x}}\})$.

Now consider $\mathbf{GIO}_{\mathrm{A}}(\hat{\mathcal{X}})$. Because, $\hat{\mathcal{X}} \subset \mathcal{P}$, Proposition 2 yields $z_{\mathrm{A}}^* = z^*\big(\mathbf{GIO}(\{\bar{\mathbf{x}}\})\big)$, i.e., the optimal solution to $\mathbf{GIO}(\{\bar{\mathbf{x}}\})$ where $\bar{\mathbf{x}} = \sum_{q=1}^{Q} \hat{\mathbf{x}}_q / Q$ is the centroid of $\hat{\mathcal{X}}$. In conjunction with (EC.7), we conclude that $z_p^* \geq z_\infty^* \geq z_{\mathrm{A}}^*$. □

Proof of Corollary 3. We remark that Corollary 3 is in fact a special case of a more general statement regarding error measures in the absolute versus relative space. Below, we prove a more general statement and specialize the result to the case of inverse optimization.

Let $f(\mathbf{x})$ and $g(\mathbf{x})$ be two functions and $f(\mathbf{x}) \neq \mathbf{0}$ for all $\mathbf{x}$. Consider two optimization problems:

$$\begin{aligned} \underset{\mathbf{x}}{\text{minimize}} \quad & \sum_{q=1}^{Q} |g_q(\mathbf{x}) - f(\mathbf{x})| \\ \text{subject to} \quad & \mathbf{x} \in \mathcal{X} \end{aligned} \qquad \text{(EC.8)} \qquad\qquad \begin{aligned} \underset{\mathbf{x}}{\text{minimize}} \quad & \sum_{q=1}^{Q} \left| \frac{g_q(\mathbf{x}) - f(\mathbf{x})}{f(\mathbf{x})} \right| \\ \text{subject to} \quad & \mathbf{x} \in \mathcal{X} \end{aligned} \qquad \text{(EC.9)}$$

Let $\mathbf{x}_{\mathrm{A}}^*$ and z_{A}^* be an optimal solution and value, respectively for (EC.8). Similarly, let $\mathbf{x}_{\mathrm{R}}^*$ and z_{R}^* be an optimal solution and value, respectively for (EC.9). We will prove that $|f(\mathbf{x}_{\mathrm{R}}^*)| z_{\mathrm{R}}^* \geq z_{\mathrm{A}}^* \geq z_{\mathrm{R}}^* |f(\mathbf{x}_{\mathrm{A}}^*)|$.

First note that $\mathbf{x}_{\mathrm{A}}^*$ is feasible for (EC.9) and $\mathbf{x}_{\mathrm{R}}^*$ is feasible for (EC.8). Then,

$$z_{\mathrm{A}}^* = \sum_{q=1}^{Q} |g_q(\mathbf{x}_{\mathrm{A}}^*) - f(\mathbf{x}_{\mathrm{A}}^*)| \leq \sum_{q=1}^{Q} |g_q(\mathbf{x}_{\mathrm{R}}^*) - f(\mathbf{x}_{\mathrm{R}}^*)| = \sum_{q=1}^{Q} \left| \frac{g_q(\mathbf{x}_{\mathrm{R}}^*) - f(\mathbf{x}_{\mathrm{R}}^*)}{f(\mathbf{x}_{\mathrm{R}}^*)} \right| |f(\mathbf{x}_{\mathrm{R}}^*)| = z_{\mathrm{R}}^* |f(\mathbf{x}_{\mathrm{R}})|.$$

The inequality comes from the feasibility of $\mathbf{x}_{\mathrm{R}}^*$ for (EC.8) and the second equality comes from multiplying by $|f(\mathbf{x}_{\mathrm{R}}^*)|/|f(\mathbf{x}_{\mathrm{R}}^*)|$. This proves the left inequality.

We next show

$$z_{\mathrm{R}}^* = \sum_{q=1}^{Q} \left| \frac{g_q(\mathbf{x}_{\mathrm{R}}^*) - f(\mathbf{x}_{\mathrm{R}}^*)}{f(\mathbf{x}_{\mathrm{R}}^*)} \right| \leq \sum_{q=1}^{Q} \left| \frac{g_q(\mathbf{x}_{\mathrm{A}}^*) - f(\mathbf{x}_{\mathrm{A}}^*)}{f(\mathbf{x}_{\mathrm{A}}^*)} \right| = \sum_{q=1}^{Q} |g_q(\mathbf{x}_{\mathrm{A}}^*) - f(\mathbf{x}_{\mathrm{A}}^*)| \frac{1}{|f(\mathbf{x}_{\mathrm{A}}^*)|} = z_{\mathrm{A}}^* \frac{1}{|f(\mathbf{x}_{\mathrm{A}}^*)|}.$$

The inequality comes from the feasibility of $\mathbf{x}_{\mathrm{A}}^*$ for (EC.9). This proves the right inequality.

Finally, we observe that letting $\mathbf{x} = (\mathbf{c}, \mathbf{y})$, $\mathcal{X} = \{(\mathbf{c}, \mathbf{y}) \mid \mathbf{A}^\intercal \mathbf{y} = \mathbf{c}, \mathbf{y} \geq \mathbf{0}, \|\mathbf{c}\|_N = 1\}$, $g_q(\mathbf{x}) = \mathbf{c}^\intercal \hat{\mathbf{x}}_q$, and $f(\mathbf{x}) = \mathbf{b}^\intercal \mathbf{y}$ converts (EC.8) into $\mathbf{GIO}_{\mathrm{A}}(\hat{\mathcal{X}})$ and (EC.9) into $\mathbf{GIO}_{\mathrm{R}}(\hat{\mathcal{X}})$. Finally note that for any feasible pair $(\mathbf{c}, \mathbf{y})$, $\mathbf{b}^\mathsf{T}\mathbf{y}$ is equal to the optimal value of the forward problem $\mathbf{FO}(\mathbf{c})$. Substituting the terms for the absolute and relative duality gap problems respectively completes the inequality. $\square$

Proof of Theorem 5.

1. Given $\hat{\mathcal{X}}$, $\mathbf{A}$, and $\mathbf{b}$, the denominator term in ρ is fixed. An optimal solution to $\mathbf{GIO}(\hat{\mathcal{X}})$ minimizes the numerator of $1-\rho$, thus maximizing ρ.
2. We prove $1-\rho \in [0,1]$. It is easy to see that $1-\rho \geq 0$, because it is the ratio of sums of norms, which are nonnegative. To show $1-\rho \leq 1$, note that $\sum_{q=1}^{Q} \left\|\boldsymbol{\epsilon}_q^*\right\| \leq \sum_{q=1}^{Q} \left\|\boldsymbol{\epsilon}_{q,i}\right\|$ for all i, as setting $\mathbf{c} = \mathbf{a}_i / \left\|\mathbf{a}_i\right\|_N$ will yield a feasible but not necessarily optimal solution to $\mathbf{GIO}(\hat{\mathcal{X}})$.
3. An optimal solution to $\mathbf{GIO}^{(k)}(\hat{\mathcal{X}})$ is feasible for $\mathbf{GIO}^{(k+1)}(\hat{\mathcal{X}})$, since the latter problem is a relaxation of the former. Invoking the first statement in this theorem, $\rho^{(k)} \leq \rho^{(k+1)}$. $\square$

EC.2. A general solution method for $\mathbf{GIO}_{\mathrm{R}}(\hat{\mathcal{X}})$

Although Proposition 5 reformulates $\mathbf{GIO}_{\mathrm{R}}(\hat{\mathcal{X}})$ into three sub-problems, the norm constraint $\left\|\cdot\right\|_N \geq K$ in the sub-problems adds two challenges: first, the constraint itself is non-convex, and second, an appropriate value for K must be chosen in order for Proposition 5 to hold. As the non-convex constraint can be handled by polyhedral decomposition, we first discuss how to choose a valid K. We then consider a relaxed reformulation of $\mathbf{GIO}_{\mathrm{R}}(\hat{\mathcal{X}})$ that often works well in practice. Finally, we summarize all of these results into a general solution algorithm for inverse optimization minimizing the relative duality gap. These steps are summarized in Algorithm 1.

The proof of Proposition 5 shows that for any $K > 0$, every feasible solution of $\mathbf{GIO}_{\mathrm{R}}^{+}(\hat{\mathcal{X}}; K)$, $\mathbf{GIO}_{\mathrm{R}}^{-}(\hat{\mathcal{X}}; K)$, and $\mathbf{GIO}_{\mathrm{R}}^{0}(\hat{\mathcal{X}}; K)$ can be mapped to a feasible solution of $\mathbf{GIO}_{\mathrm{R}}(\hat{\mathcal{X}})$. The normalization constraint $\left\|\mathbf{c}\right\|_N \geq K$ implies that the feasible region for each sub-problem grows as K decreases. The proof then shows that for some sufficiently small $K > 0$, an optimal solution to $\mathbf{GIO}_{\mathrm{R}}(\hat{\mathcal{X}})$ can be mapped to a feasible (and therefore, also optimal) solution of one of (9)–(11).

To determine a sufficiently small K, note that the mapping of a solution of $\mathbf{GIO}_{\mathrm{R}}(\hat{\mathcal{X}})$ to solutions of one of (9)–(11) involves scaling the solution by $\mathbf{b}^\mathsf{T}\mathbf{y}$, $-\mathbf{b}^\mathsf{T}\mathbf{y}$, or $\mathbf{y}^\mathsf{T}\mathbf{1}$, respectively. Bounding these terms allows us to determine a sufficiently small K. Formally, consider the following problem:

$$\begin{aligned} \underset{\mathbf{y}}{\text{maximize}} \quad & \max\left\{|\mathbf{b}^\mathsf{T}\mathbf{y}|, \mathbf{y}^\mathsf{T}\mathbf{1}\right\} \\ \text{subject to} \quad & \left\|\mathbf{A}^\mathsf{T}\mathbf{y}\right\|_N = 1,\ \mathbf{y} \geq \mathbf{0}. \end{aligned} \tag{EC.10}$$

We refer to formulation (EC.10) 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 (EC.10) 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. That is, the auxiliary problem always has a non-zero solution, because any feasible $\mathbf{y}$ to (EC.10) must have $\mathbf{y} \geq \mathbf{0}$ and at least one non-zero $y_i > 0$, meaning $\mathbf{y}^\mathsf{T}\mathbf{1} > 0$ must always hold. We use K^* to reformulate $\mathbf{GIO}_{\mathrm{R}}(\hat{\mathcal{X}})$ to $\mathbf{GIO}_{\mathrm{R}}^{+}(\hat{\mathcal{X}}; K^*)$, $\mathbf{GIO}_{\mathrm{R}}^{-}(\hat{\mathcal{X}}; K^*)$, and $\mathbf{GIO}_{\mathrm{R}}^{0}(\hat{\mathcal{X}}; K^*)$.

THEOREM EC.1. *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 $\left(\mathbf{c}^*, \mathbf{y}^*, \epsilon_1^*, \dots, \epsilon_Q^*\right)$ be the corresponding optimal solution. Then, $\left(\mathbf{c}^*/\left\|\mathbf{c}^*\right\|_N, \mathbf{y}^*/\left\|\mathbf{c}^*\right\|_N, \epsilon_1^*, \dots, \epsilon_Q^*\right)$ is optimal to $\mathbf{GIO}_{\mathrm{R}}(\hat{\mathcal{X}})$.*

Proof of Theorem EC.1. Let $(\hat{\mathbf{c}}, \hat{\mathbf{y}})$ be optimal to $\mathbf{GIO}_{\mathrm{R}}(\hat{\mathcal{X}})$ and K be defined as in (EC.3). Since $\hat{\mathbf{y}}$ is feasible for the auxiliary problem (EC.10), $1/K^* \geq \max\left\{|\mathbf{b}^\mathsf{T}\hat{\mathbf{y}}|, \hat{\mathbf{y}}^\mathsf{T}\mathbf{1}\right\}$, implying $K^* \leq K$.

The proof of Proposition 5 showed that scaling $(\hat{\mathbf{c}}, \hat{\mathbf{y}})$ appropriately yielded a corresponding feasible solution to one of $\mathbf{GIO}_{\mathrm{R}}^{+}(\hat{\mathcal{X}}; K)$, $\mathbf{GIO}_{\mathrm{R}}^{-}(\hat{\mathcal{X}}; K)$, or $\mathbf{GIO}_{\mathrm{R}}^{0}(\hat{\mathcal{X}}; K)$. Because $K^* \leq K$, the scaled solution must also be feasible for the respective $\mathbf{GIO}_{\mathrm{R}}^{+}(\hat{\mathcal{X}}; K^*)$, $\mathbf{GIO}_{\mathrm{R}}^{-}(\hat{\mathcal{X}}; K^*)$, or $\mathbf{GIO}_{\mathrm{R}}^{0}(\hat{\mathcal{X}}; K^*)$. Moreover, every solution of $\mathbf{GIO}_{\mathrm{R}}^{+}(\hat{\mathcal{X}}; K^*)$, $\mathbf{GIO}_{\mathrm{R}}^{-}(\hat{\mathcal{X}}; K^*)$, or $\mathbf{GIO}_{\mathrm{R}}^{0}(\hat{\mathcal{X}}; K^*)$ can be scaled to a feasible solution of $\mathbf{GIO}_{\mathrm{R}}(\hat{\mathcal{X}})$, completing the proof. □

Algorithm 1 General solution method for $\mathbf{GIO}_{\mathrm{R}}(\hat{\mathcal{X}})$

Input: Data set $\hat{\mathcal{X}}$

Output: Imputed model parameters $\left(\mathbf{c}^*, \mathbf{y}^*, \epsilon_1^*, \ldots, \epsilon_Q^*\right)$

1: Let $z_{\mathrm{LP}}^+ \leftarrow \mathbf{GIO}_{\mathrm{R,LP}}^+(\hat{\mathcal{X}})$, $z_{\mathrm{LP}}^- \leftarrow \mathbf{GIO}_{\mathrm{R,LP}}^-(\hat{\mathcal{X}})$, $z_{\mathrm{LP}}^0 \leftarrow \mathbf{GIO}_{\mathrm{R,LP}}^0(\hat{\mathcal{X}})$ be the optimal values.

2: Let $z_{\mathrm{LP}}^* \leftarrow \min\left\{z_{\mathrm{LP}}^+, z_{\mathrm{LP}}^-, z_{\mathrm{LP}}^0\right\}$ and $\left(\mathbf{c}^*, \mathbf{y}^*, \epsilon_1^*, \ldots, \epsilon_Q^*\right)$ be the corresponding optimal solution.

3: **if** $\mathbf{c}^* \neq \mathbf{0}$ **then**

4:     **return** $\left(\mathbf{c}^*, \mathbf{y}^*, \epsilon_1^*, \ldots, \epsilon_Q^*\right)$

5: **else**

6:     Solve the auxiliary problem (EC.10). Let K^* be the reciprocal of the optimal value.

7:     Let $z^+ \leftarrow \mathbf{GIO}_{\mathrm{R}}^+(\hat{\mathcal{X}}; K^*)$, $z^- \leftarrow \mathbf{GIO}_{\mathrm{R}}^-(\hat{\mathcal{X}}; K^*)$, $z^0 \leftarrow \mathbf{GIO}_{\mathrm{R}}^0(\hat{\mathcal{X}}; K^*)$ be the optimal values.

8:     Let $z^* \leftarrow \min\{z^+, z^-, z^0\}$ and $\left(\mathbf{c}^*, \mathbf{y}^*, \epsilon_1^*, \ldots, \epsilon_Q^*\right)$ be the corresponding optimal solution.

9:     **return** $\left(\mathbf{c}^*, \mathbf{y}^*, \epsilon_1^*, \ldots, \epsilon_Q^*\right)$

10: **end if**

In the most general case, solving $\mathbf{GIO}_{\mathrm{R}}(\hat{\mathcal{X}})$ is more computationally intensive than solving $\mathbf{GIO}_{\mathrm{A}}(\hat{\mathcal{X}})$. We must first identify K^*, which we can use to reformulate $\mathbf{GIO}_{\mathrm{R}}(\hat{\mathcal{X}})$ into three norm-constrained optimization problems. Subsequently, given an appropriate choice of $\|\cdot\|_N$, each problem is decomposed into a series of LPs. For instance, doing so leads to $2n$ LPs if $\|\cdot\|_N = \|\cdot\|_\infty$ and 2^n LPs if $\|\cdot\|_N = \|\cdot\|_1$. These steps coupled with the auxiliary problem (EC.10) used to determine K^* require the solution of $12n$ LPs when $\|\cdot\|_N = \|\cdot\|_\infty$, or $6(2^n)$ when $\|\cdot\|_N = \|\cdot\|_1$. In some cases, however, it may be possible to find an optimal solution to $\mathbf{GIO}_{\mathrm{R}}(\hat{\mathcal{X}})$ by solving exactly three LPs.

Corollary EC.1. *Let $\mathbf{GIO}_{\mathrm{R,LP}}^+(\hat{\mathcal{X}})$, $\mathbf{GIO}_{\mathrm{R,LP}}^-(\hat{\mathcal{X}})$, and $\mathbf{GIO}_{\mathrm{R,LP}}^0(\hat{\mathcal{X}})$ be the LP relaxations of $\mathbf{GIO}_{\mathrm{R}}^+(\hat{\mathcal{X}}; K)$, $\mathbf{GIO}_{\mathrm{R}}^-(\hat{\mathcal{X}}; K)$, and $\mathbf{GIO}_{\mathrm{R}}^0(\hat{\mathcal{X}}; K)$, respectively, obtained by removing the normalization constraint $\|\mathbf{c}\|_N \geq K$. Let z_{LP}^+ be the optimal value of $\mathbf{GIO}_{\mathrm{R,LP}}^+(\hat{\mathcal{X}})$ if it is feasible, otherwise $z_{\mathrm{LP}}^+ = \infty$. Let z_{LP}^- and z_{LP}^0 be defined similarly for $\mathbf{GIO}_{\mathrm{R,LP}}^-(\hat{\mathcal{X}})$ and $\mathbf{GIO}_{\mathrm{R,LP}}^0(\hat{\mathcal{X}})$, respectively. Let $z_{\mathrm{LP}}^* = \min\left\{z_{\mathrm{LP}}^+, z_{\mathrm{LP}}^-, z_{\mathrm{LP}}^0\right\}$ and let $\left(\mathbf{c}^*, \mathbf{y}^*, \epsilon_1^*, \ldots, \epsilon_Q^*\right)$ be an optimal solution of* *the corresponding problem. If* $\mathbf{c}^* \neq \mathbf{0}$*, then* z^*_{LP} *is equal to the optimal value of* $\mathbf{GIO}_{\text{R}}(\hat{\mathcal{X}})$ *and* $\left(\mathbf{c}^*/\left\|\mathbf{c}^*\right\|_N, \mathbf{y}^*/\left\|\mathbf{c}^*\right\|_N, \epsilon^*_1, \ldots, \epsilon^*_Q\right)$ *is an optimal solution to* $\mathbf{GIO}_{\text{R}}(\hat{\mathcal{X}})$*.*

Proof of Corollary EC.1. Let $(\hat{\mathbf{c}}, \hat{\mathbf{y}}, \hat{\epsilon}_1, \ldots, \hat{\epsilon}_Q)$ be an optimal solution to $\mathbf{GIO}_{\text{R}}(\hat{\mathcal{X}})$. From Proposition 5, this solution can be rescaled to construct a feasible solution for one of $\mathbf{GIO}^+_{\text{R,LP}}(\hat{\mathcal{X}})$, $\mathbf{GIO}^-_{\text{R,LP}}(\hat{\mathcal{X}})$, and $\mathbf{GIO}^0_{\text{R,LP}}(\hat{\mathcal{X}})$ with the same objective value. Conversely, for each of the relaxed problems, let $(\tilde{\mathbf{c}}, \tilde{\mathbf{y}}, \tilde{\epsilon}_1, \ldots, \tilde{\epsilon}_Q)$ be a feasible solution. Assuming that $\tilde{\mathbf{c}} \neq \mathbf{0}$, this solution can be rescaled to construct $(\hat{\mathbf{c}}, \hat{\mathbf{y}}, \hat{\epsilon}_1, \ldots, \hat{\epsilon}_Q) = (\tilde{\mathbf{c}}/\left\|\tilde{\mathbf{c}}\right\|_N, \tilde{\mathbf{y}}/\left\|\tilde{\mathbf{c}}\right\|_N, \tilde{\epsilon}_1, \ldots, \tilde{\epsilon}_Q)$, which is a feasible solution to $\mathbf{GIO}_{\text{R}}(\hat{\mathcal{X}})$ with the same objective value. Thus, if the minimum of $\mathbf{GIO}^+_{\text{R,LP}}(\hat{\mathcal{X}})$, $\mathbf{GIO}^-_{\text{R,LP}}(\hat{\mathcal{X}})$, and $\mathbf{GIO}^0_{\text{R,LP}}(\hat{\mathcal{X}})$ yields an optimal solution with a non-zero imputed cost vector, the two problems share the same optimal solution. □

The key difference between Proposition 5 and Corollary EC.1 is the non-zero assumption (i.e., $\mathbf{c}^* \neq \mathbf{0}$). By relaxing the normalization constraint, we permit potential solutions for which $\mathbf{c}^* = \mathbf{A}^\intercal \mathbf{y}^* = \mathbf{0}$ is a linearly dependent combination of the rows of $\mathbf{A}$. However, if $\mathbf{c}^* \neq \mathbf{0}$ is an optimal solution to the relaxed problem, it is also an optimal solution to $\mathbf{GIO}_{\text{R}}(\hat{\mathcal{X}})$. Therefore, to solve $\mathbf{GIO}_{\text{R}}(\hat{\mathcal{X}})$, we suggest first solving the three relaxed problems, which are LPs, from Corollary EC.1. If $\mathbf{c}^* = \mathbf{0}$, then we use the more general approach. Section 5 (with details on the formulations in EC.4) shows a case where the LP relaxations via Corollary EC.1 are sufficient.

EC.3. Related work in inverse convex optimization

Multi-point inverse optimization has recently received significant interest under the setting of convex forward problems, with several notable inverse optimization models having been proposed for arbitrary convex forward problems (i.e., Bertsimas et al. (2015), Aswani et al. (2018), Esfahani et al. (2018)). The methods proposed in this prior work specialize to linear forward problems and overlap in formulation with the absolute duality and the decision space models proposed in this paper. However, the geometric nature of LPs poses new challenges, but also allows for some efficient solutions, that are not present in the strictly convex domain. In this section, we highlight the previous formulations and discuss several differences in the solution methods.

The inverse convex models in prior work assume that the data set consists of points corresponding to different forward problem instances. As we focus on inverse optimization for a fixed forward feasible region, we illustrate the results in the previous work by fixing $\mathcal{P}$.

EC.3.1. Inverse variational inequality

Let $f(\mathbf{x};\mathbf{c}):\mathbb{R}^n \to \mathbb{R}$ be a convex function in $\mathbf{x}$ parametrized by $\mathbf{c}$ and $\mathcal{K}$ be a convex cone. Bertsimas et al. (2015) considered the forward problem $\min_{\mathbf{x}} \{f(\mathbf{x};\mathbf{c}) \mid \mathbf{A}\mathbf{x} = \mathbf{b}, \mathbf{x} \in \mathcal{K}\}$ and proposed an inverse optimization model that minimized the residuals from failing to satisfy the variational inequality of the first-order optimality condition. The inverse variational inequality problem is

$$\begin{aligned}
\underset{\mathbf{c},\mathbf{y}_1,\ldots,\mathbf{y}_Q,\epsilon_1,\ldots,\epsilon_Q}{\text{minimize}} \quad & \sum_{q=1}^{Q} |\epsilon_q| \\
\text{subject to} \quad & \mathbf{A}^\mathsf{T}\mathbf{y}_q \leq_{\mathcal{K}} \nabla f(\hat{\mathbf{x}}_q;\mathbf{c}), \quad \forall q \in \mathcal{Q} \\
& \nabla f(\hat{\mathbf{x}}_q;\mathbf{c})^\mathsf{T}\hat{\mathbf{x}}_q - \mathbf{b}^\mathsf{T}\mathbf{y}_q \leq \epsilon_q, \quad \forall q \in \mathcal{Q} \\
& \mathbf{c} \in \mathcal{C}.
\end{aligned} \tag{EC.11}$$

Setting $\mathcal{K} = \mathbb{R}^n_+$, $f(\mathbf{x};\mathbf{c}) = \mathbf{c}^\mathsf{T}\mathbf{x}$, and $\mathcal{C} = \{\mathbf{c} \in \mathbb{R}^n \mid \|\mathbf{c}\|_N = 1\}$ makes formulation (EC.11) equivalent to $\mathbf{GIO}_\text{A}(\hat{\mathcal{X}})$, i.e., formulation (6).

In the original work, Bertsimas et al. (2015) focused mostly on strictly convex forward problems and on ensuring a convex inverse optimization formulation. While the non-convex normalization constraint is not always necessary when the forward problem is strictly convex, setting $f(\mathbf{x};\mathbf{c}) = \mathbf{c}^\mathsf{T}\mathbf{x}$ 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. 2019, Esfahani et al. 2018). Note furthermore that convex normalization constraints exist in the literature, e.g., Keshavarz et al. (2011) proposed setting $c_0 = 1$. However, these convex normalization constraints often bias the parameter space. For instance, setting $c_0 = 1$ prevents imputing non-trivial cost vectors where $c_0 = 0$. We enforce the non-convex constraint within all of the inverse optimization models in the current paper and propose polyhedral decomposition-based solution methods in the general setting for $\mathbf{GIO}_\text{A}(\hat{\mathcal{X}})$. Furthermore, we find it important to explore special cases where the non-convexity can be bypassed, leading to simpler, sometimes analytic results (see Proposition 2 and 3, as well as Corollary 1).

Finally, Bertsimas et al. (2015) discussed a decision space alternative to formulation (EC.11) where instead of minimizing the variational inequality residual, they minimized $\|\hat{\mathbf{x}}_q - \mathbf{x}_q\|$, where $\mathbf{x}_q$ is a variable that satisfies $f(\mathbf{x}_q;\mathbf{c}) = \mathbf{b}^\top\mathbf{y}$. Furthermore, they assumed that the gradient of the objective function is strongly monotone, i.e., there exists $\gamma > 0$ such that

$$(\nabla f(\mathbf{x};\mathbf{c}) - \nabla f(\mathbf{y};\mathbf{c}))^\top(\mathbf{x}-\mathbf{y}) \geq \gamma \|\mathbf{x}-\mathbf{y}\|_2, \quad \forall \mathbf{x},\mathbf{y} \in \mathcal{P}.$$

By focusing on the variational inequality nature of objective space inverse optimization, Bertsimas et al. (2015, Theorem 1) translated the variational inequality error bound of Pang (1987) to show that if there exists an solution $(\mathbf{c}^*, \mathbf{y}^*, \epsilon_1^*, \dots, \epsilon_Q^*)$ to formulation (EC.11), then there exists $\mathbf{x}_1^*, \dots, \mathbf{x}_Q^*$ that are optimal solutions to the forward problem and satisfy $\left\|\hat{\mathbf{x}}_q - \mathbf{x}_q^*\right\|_2 \leq \sqrt{\epsilon_q/\gamma}$ for all q. That is, given the feasible solution to an objective space inverse optimization problem, we can obtain a corresponding feasible solution to a decision space problem where the error is bounded. Note, however, that in the linear case, $\nabla f(\mathbf{x};\mathbf{c}) = \mathbf{c}$ does not satisfy the strong monotone property, i.e., $\gamma = 0$. As a result, the previous bound does not hold for inverse linear optimization.

EC.3.2. Inverse empirical risk minimization

Let $f(\mathbf{x};\mathbf{u},\mathbf{c}) : \mathbb{R}^n \to \mathbb{R}$ and $g(\mathbf{x};\mathbf{u},\mathbf{c}) : \mathbb{R}^n \to \mathbb{R}^m$ be convex functions in $\mathbf{x}$ that are both parametrized by $\mathbf{u}$ and $\mathbf{c}$. Aswani et al. (2018) considered the general convex forward problem $\min_{\mathbf{x}} \{f(\mathbf{x};\mathbf{u},\mathbf{c}) \mid g(\mathbf{x};\mathbf{u},\mathbf{c}) \leq \mathbf{0}\}$ and proposed a bilevel inverse optimization model that minimized the empirical distance between a data set $\hat{\mathcal{X}} = \{(\hat{\mathbf{x}}_1, \hat{\mathbf{u}}_1), \dots, (\hat{\mathbf{x}}_Q, \hat{\mathbf{u}}_Q)\}$ of Q points sampled i.i.d. from some joint probability distribution $\mathbb{P}_{\mathbf{x},\mathbf{u}}$ and the optimal solution sets. The corresponding inverse risk minimization problem is

$$\begin{aligned} \underset{\mathbf{c},\boldsymbol{\epsilon}_1,\dots,\boldsymbol{\epsilon}_Q}{\text{minimize}} \quad & \sum_{q=1}^{Q} \|\boldsymbol{\epsilon}_q\|_p \\ \text{subject to} \quad & \hat{\mathbf{x}}_q - \boldsymbol{\epsilon}_q \in \arg\min_{\mathbf{x}} \{f(\mathbf{x};\hat{\mathbf{u}}_q,\mathbf{c}) \mid g(\mathbf{x};\hat{\mathbf{u}}_q,\mathbf{c}) \leq \mathbf{0}\}, \quad \forall q \in \mathcal{Q} \\ & \mathbf{c} \in \mathcal{C}. \end{aligned} \tag{EC.12}$$

Setting $f(\mathbf{x};\mathbf{u},\mathbf{c}) = \mathbf{c}^\top\mathbf{x}$, $g(\mathbf{x};\mathbf{u},\mathbf{c}) = \mathbf{b} - \mathbf{A}\mathbf{x}$, and $\mathcal{C} = \{\mathbf{c} \in \mathbb{R}^n \mid \|\mathbf{c}\|_N = 1\}$ specializes formulation (EC.12) to an equivalent form as $\mathbf{GIO}_p(\hat{\mathcal{X}})$.

Formulation (EC.11) satisfies statistical consistency (i.e., given sufficient points, the imputed $\mathbf{c}$ converges to a true data-generating $\mathbf{c}$) under several assumptions on the data set and the forward model (Aswani et al. 2018):

1. **Assumption 2:** The parameter space $\mathcal{C}$ is convex.
2. **Regularity 1:** The feasible set $\mathcal{P}$ is closed and bounded.
3. **Identifiability condition:** There exists a unique $\mathbf{c}^*$ such that:
   (a) The data set corresponds to noisy perturbations of optimal solutions, i.e., $\hat{\mathbf{x}}_q = \mathbf{x}_q^* + \mathbf{w}_q$, where $\mathbf{x}_q^* \in \arg\min_{\mathbf{x}} \{f(\mathbf{x};\mathbf{u},\mathbf{c}) \mid g(\mathbf{x};\mathbf{u},\mathbf{c}) \leq \mathbf{0}\}$, and $\mathbf{w}_q$ is a random variable with mean 0 and finite variance.

   (b) For any $\mathbf{c} \neq \mathbf{c}^*$, there exists $\mathcal{U}_{\mathbf{c}}$ such that the marginal distribution $\mathbb{P}_{\mathbf{u}}(\mathbf{u} \in \mathcal{U}_{\mathbf{c}}) > 0$ and the optimal value for

$$
\begin{aligned}
\inf_{\mathbf{x},\mathbf{x}^*} \quad & \|\mathbf{x} - \mathbf{x}^*\| \\
\text{s.t.} \quad & \mathbf{x} \in \arg\min_{\mathbf{w}} \{f(\mathbf{w};\mathbf{u},\mathbf{c}) \mid g(\mathbf{w};\mathbf{u},\mathbf{c}) \leq \mathbf{0}\} \\
& \mathbf{x}^* \in \arg\min_{\mathbf{w}} \{f(\mathbf{w};\mathbf{u},\mathbf{c}^*) \mid g(\mathbf{w};\mathbf{u},\mathbf{c}^*) \leq \mathbf{0}\}
\end{aligned}
$$

   is equal to 0 for all $\mathbf{u} \in \mathcal{U}_{\mathbf{c}}$.

   (c) For all $\mathbf{c}$,

$$
\mathbb{P}_{\mathbf{u}}\left(\left\{\mathbf{u} \,\middle|\, \left|\arg\min_{\mathbf{x}} \{f(\mathbf{x};\mathbf{u},\mathbf{c}) \mid g(\mathbf{x};\mathbf{u},\mathbf{c}) \leq \mathbf{0}\}\right| > 1\right\}\right) = 0
$$

These assumptions do not hold in this work where we focus on a fixed linear forward problem for all data points. Particularly, setting $f(\mathbf{x};\mathbf{u},\mathbf{c}) = \mathbf{c}^\intercal \mathbf{x}$ and $g(\mathbf{x};\mathbf{u},\mathbf{c}) = \mathbf{b} - \mathbf{A}\mathbf{x}$ implies that the forward and inverse optimization problem do not depend on $\mathbf{u}$. Consequently, the second Identifiability condition does not hold in many settings. A trivial example is $\mathcal{P} = \{(x_1, x_2) \mid 0 \leq x_1, x_2 \leq 1\}$. Here, for any cost vector $\mathbf{c}^*$, there exists another cost vector $\mathbf{c}_i = \mathbf{a}_i / \|\mathbf{a}_i\|_N$ such that the facet described by $\mathbf{c}_i$ contains an optimal vertex of $\mathbf{FO}(\mathbf{c}^*)$. Furthermore, the third condition is also trivially violated when $\mathbf{c} = \mathbf{a}_i$ for any $i \in \mathcal{I}$. Finally, our application in Section 5 is an example where the dataset does not correspond to noisy perturbations, but is obtained via several prediction models; we therefore cannot guarantee a well-behaved $\mathbf{w}_q$. We also remark that our problem setting permits the feasible set $\mathcal{P}$ to be unbounded. A last consequence of $\mathbf{u}$ not existing in our setting is that the parameter space becomes non-convex due to the norm constraint. Overall, we find our problem setting to be incompatible with the statistical consistency guarantees in Aswani et al. (2018).

Aswani et al. (2018) propose an efficient semi-parametric algorithm to solve formulation (EC.12) under the assumption that the forward problem is strictly convex in $\mathbf{x}$. For when $f(\mathbf{x};\mathbf{u},\mathbf{c})$ is linear however, Aswani et al. (2018) introduce an enumerative algorithm for solving formulation (EC.12) that relies on quantizing the set $\mathcal{C}$ to a finite set $\hat{\mathcal{C}}$, and solving the corresponding formulation with fixed $\mathbf{c}\in\hat{\mathcal{C}}$. This algorithm is effective primarily because, for fixed $\mathbf{c}$, formulation (EC.12) (and incidentally, $\mathbf{GIO}_p(\hat{\mathcal{X}})$) are convex. However, the authors state that due to the enumerative nature, the algorithm is generally only applicable when the parameter space is modest (e.g., $n \leq 5$ is recommended). We find that the algorithm of Aswani et al. (2018) is complementary to ours. That is, their algorithm is inefficient for large n, while our decision space algorithm is relatively insensitive to the increase in n, but is inefficient for large m.

EC.3.3. Distributionally robust inverse optimization

Esfahani et al. (2018) study distributionally robust generalized inverse optimization for convex forward problems. Let $\varrho(\cdot)$ denote a risk measure such as the Value-at-Risk (VaR) or Conditional Value-at-Risk (CVaR). The *non-robust* version of their formulation is

$$\begin{aligned} \underset{\mathbf{c},\boldsymbol{\epsilon}_1,\dots,\boldsymbol{\epsilon}_Q}{\text{minimize}} \quad & \varrho(\|\boldsymbol{\epsilon}_1\|,\dots,\|\boldsymbol{\epsilon}_Q\|) \\ \text{subject to} \quad & \text{Constraints in (EC.11) or (EC.12)} \end{aligned} \tag{EC.13}$$

Esfahani et al. (2018) consider several different variants of inverse convex optimization to encapsulate previous methods; the variants are referred to as predictability (i.e., inverse risk minimization), sub-optimality, first-order (i.e., inverse variational inequality), and bounded rationality. When the forward problem is an LP, the sub-optimality loss model is in fact equivalent to the first-order loss model, and therefore also equivalent to $\mathbf{GIO}_{\mathrm{A}}(\hat{\mathcal{X}})$ proposed here.

A consequence of the general formulation (EC.13) is that it leads to a new dominance relationship to bound the optimal values between predictability and sub-optimality losses. Similar to Bertsimas et al. (2015), Esfahani et al. (2018) define the parameter $\gamma \geq 0$ to be the largest parameter satisfying

$$f(\mathbf{x};\mathbf{u},\mathbf{c}) - f(\mathbf{y};\mathbf{u},\mathbf{c}) \geq \nabla f(\mathbf{x};\mathbf{u},\mathbf{c})^{\mathsf{T}}(\mathbf{x}-\mathbf{y}) + \frac{\gamma}{2}\|\mathbf{x}-\mathbf{y}\|_2^2, \quad \forall \mathbf{x},\mathbf{y} \in \mathcal{P}, \mathbf{u} \in \mathcal{U}.$$

Under this definition, Esfahani et al. (2018) show that their sub-optimality (i.e., objective space) loss upper bounds their predictability (i.e., decision space) loss by a multiplicative factor $\gamma/2$. However, similar to the scenario in the previous bound, $\gamma = 0$ when $f(\mathbf{x};\mathbf{u},\mathbf{c}) = \mathbf{c}^{\mathsf{T}}\mathbf{x}$. Consequently, this bound also does not hold for LP forward problems.

Esfahani et al. (2018) focus on solving a distributionally robust version of formulation (EC.13), where the robustness is over the worst-case distribution of data. As they primarily address the sub-optimality loss model, which specializes to the absolute duality gap model in this work, the comparison between their solution methods and ours is similar to the comparison between Bertsimas et al. (2015) and ours. That is, we focus on developing efficient algorithms based on LP geometry, and as a consequence, yield several new efficiencies in the absolute duality gap setting.

EC.4. Automated radiation therapy treatment planning

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 design of an IMRT treatment plan is typically done by mathematical optimization where the decision variable $\mathbf{x} = (\mathbf{w}, \mathbf{d})$ is composed of two components that represent the beamlets and the dose delivered (in Gy) as a result of 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 $\mathcal{B}$ denote the index set of beamlets and w_b be the radiation intensity of beamlet $b \in \mathcal{B}$. Similarly, let $\mathcal{V}$ denote the index set of voxels within a patient and d_v be the dose of radiation delivered to voxel $v \in \mathcal{V}$. Dose is calculated via a weighted linear combination of all beamlet intensities, 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.

EC.4.1. 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 of predictions; we consider $\theta \in \Theta := \{0.25, 0.5, 0.75, 0.9, 0.975\}$. Such objectives for each OAR can be computed as follows:

$$z_o^{\text{mean}} = \frac{1}{|\mathcal{V}_o|} \sum_{v \in \mathcal{V}_o} d_v, \quad \forall o \in \mathcal{O} \tag{EC.14}$$

$$z_o^{\text{max}} = \max_{v \in \mathcal{V}_o} \{d_v\}, \quad \forall o \in \mathcal{O} \tag{EC.15}$$

$$z_o^{\text{thresh},\theta} = \frac{1}{|\mathcal{V}_o|} \sum_{v \in \mathcal{V}_o} \max\left\{0, d_v - \phi_o^\theta\right\}, \quad \forall \theta \in \Theta, \forall o \in \mathcal{O}. \tag{EC.16}$$

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:

$$z_t^{\text{over}} = \frac{1}{|\mathcal{V}_t|} \sum_{v \in \mathcal{V}_t} \max\{0, d_v - \phi_t\}, \quad \forall t \in \mathcal{T} \tag{EC.17}$$

$$z_t^{\text{under}} = \frac{1}{|\mathcal{V}_t|} \sum_{v \in \mathcal{V}_t} \max\{0, \phi_t - d_v\}, \quad \forall t \in \mathcal{T} \tag{EC.18}$$

$$z_t^{\text{max}} = \max_{v \in \mathcal{V}_t} \{d_v\}, \quad \forall t \in \mathcal{T}. \tag{EC.19}$$

EC.4.2. 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. Every OAR is assigned a constraint to ensure that the mean dose and maximum dose do not exceed baseline safety limits. Similarly, every PTV is assigned a constraint to ensure that it receives a baseline dose on average.

The safety constraints are relaxations of the clinical criteria used to evaluate plans. Recall that clinical plans typically do not satisfy all of the clinical criteria. In fact, satisfying all of the criteria is infeasible for most patients. Consequently, we set these safety constraints so that all plans can satisfy at least these baseline doses for each of the OARs and PTVs; we then use the objectives to push the doses to achieving the clinical criteria. The baseline values, i.e., right-hand-side of the constraints, are obtained from the average and maximum dose delivered by the 130 clinical plans in our training set. We list the constraints below:

$$\text{Brain stem:} \quad z_o^{\text{mean}} \leq 30, \quad z_o^{\max} \leq 53 \tag{EC.20}$$

$$\text{Spinal cord:} \quad z_o^{\text{mean}} \leq 30, \quad z_o^{\max} \leq 46 \tag{EC.21}$$

$$\text{Left parotid:} \quad z_o^{\text{mean}} \leq 68, \quad z_o^{\max} \leq 77 \tag{EC.22}$$

$$\text{Right parotid:} \quad z_o^{\text{mean}} \leq 68, \quad z_o^{\max} \leq 78 \tag{EC.23}$$

$$\text{Larynx:} \quad z_o^{\text{mean}} \leq 68, \quad z_o^{\max} \leq 77 \tag{EC.24}$$

$$\text{Esophagus:} \quad z_o^{\text{mean}} \leq 52, \quad z_o^{\max} \leq 75 \tag{EC.25}$$

$$\text{Mandible:} \quad z_o^{\text{mean}} \leq 63, \quad z_o^{\max} \leq 76 \tag{EC.26}$$

$$\text{limPostNeck:} \quad z_o^{\text{mean}} \leq 21, \quad z_o^{\max} \leq 46 \tag{EC.27}$$

$$\text{PTV56:} \quad z_t^{\text{mean}} \geq 58 \tag{EC.28}$$

$$\text{PTV63:} \quad z_t^{\text{mean}} \geq 63 \tag{EC.29}$$

$$\text{PTV70:} \quad z_t^{\text{mean}} \geq 69 \tag{EC.30}$$

Note that we introduce a z_t^{mean} variable for the targets, analogous to z_o^{mean} in (EC.14).

Finally, we include a constraint on the "complexity" or physical deliverability of the treatment plan. This constraint, known as the sum-of-positive-gradients (SPG), restricts the variation of radiation doses from neighboring beamlets so that the resulting dose shape is deliverable by the LINAC (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 $\mathcal{B}_r$ be the index set of beamlets along that row. Then, we add the following constraint to restrict the variation of doses to be delivered from different beamlets:

$$\sum_{a\in\mathcal{A}} \max_{r\in\mathcal{R}_a} \left\{ \sum_{b\in\mathcal{B}_r} \max\{0, w_b - w_{b+1}\} \right\} \leq 55, \tag{EC.31}$$

where we set $w_{b+1} = 0$ for the last beamlet in each row. The right-hand-side, i.e., the SPG, is set to 55 Gy, following the convention from previous literature (Babier et al. 2020).

EC.4.3. Forward optimization problem

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

$$\begin{aligned} \mathbf{RT\text{–}FO}(\boldsymbol{\alpha}): \quad \underset{\mathbf{z},\mathbf{w},\mathbf{d}}{\text{minimize}} \quad & \sum_{o\in\mathcal{O}} \left(\alpha_o^{\text{mean}} z_o^{\text{mean}} + \alpha_o^{\max} z_o^{\max} + \sum_{\theta\in\Theta} \alpha_o^{\text{thresh},\theta} z_o^{\text{thresh},\theta} \right) + \\ & \sum_{t\in\mathcal{T}} \left(\alpha_t^{\text{over}} z_t^{\text{over}} + \alpha_t^{\text{under}} z_t^{\text{under}} + \alpha_t^{\max} z_t^{\max} \right) \\ \text{subject to} \quad & (\text{EC.14}) - (\text{EC.31}) \\ & \sum_{b\in\mathcal{B}} D_{v,b} w_b = d_v, \quad \forall v \in \mathcal{V} \\ & w_b, d_v \geq 0, \quad \forall b \in \mathcal{B}, \forall v \in \mathcal{V}. \end{aligned} \tag{EC.32}$$

We compress the notation of the above forward problem to $\mathbf{FO}(\boldsymbol{\alpha}): \min_{\mathbf{x}} \left\{ \boldsymbol{\alpha}^\intercal \mathbf{C}\mathbf{x} \;\middle|\; \mathbf{A}\mathbf{x} \geq \mathbf{b}, \mathbf{x} \geq \mathbf{0} \right\}$. 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. We discuss these in Section EC.4.5.

EC.4.4. Generating a data set of predicted treatments

We use the training set of 130 patients to implement several machine learning models from the KBP literature. Each model is trained via supervised learning using a data set of paired patient

CT images (i.e., features) and clinically delivered dose distributions (i.e., target). There are some variations in how each model approaches the task. We use the same training techniques for each model as described in their original papers and summarize the results below:

1. **Random Forest:** A random forest that uses 10 hand-crafted geometric features derived from the CT images (e.g., distance to nearest tumor structure) to predict the dose for each voxel $\hat{d}_v$ of the patient individually (McIntosh et al. 2017, Mahmood et al. 2018). We apply the random forest to predict each voxel for a given patient independently and concatenate the predictions to construct a dose distribution.
2. **2-D RGB GAN:** A generative adversarial network that uses 2-D axial slices of the patient's CT as an RGB image to predict corresponding 2-D axial slices of the patient's dose also as an RGB image (Mahmood et al. 2018). We convert the images to grayscale and run 2-D RGB GAN over all 128 axial slices of the patient and concatenate the predictiosn to produce a 3-D dose distribution.
3. **2-D GANCER:** A generative adversarial network that uses 2-D axial slices of the patient's CT as an RGB image to predict 2-D axial slices of the patient's dose in grayscale directly (Babier et al. 2020). This model is a variant of the 2-D RGB GAN. We run this model over all 128 axial slices of a patient and concatenate the predictions to produce a 3-D dose distributions.
4. **3-D GANCER:** A generative adversarial network that uses the full 3-D patient's CT image as input to predict the full 3-D dose distribution $\hat{\mathbf{d}}$ in one shot (Babier et al. 2020).

Babier et al. (2020) noted that plans predicted using the above models often sought to deliver low dose (i.e., significantly spare healthy tissue) at the cost of not satisfying the prescription criteria for the PTVs, and implemented a rescaling method to create a modified prediction to address this issue. In their experiments, they showed that treatment plans constructed using inverse optimization-based KBP and the normalized dose distributions would better satisfy the prescription criteria while performing slightly poorer on sparing healthy tissue. Consequently, we implement

Table EC.1 The percentage of predictions that are feasible with respect to their forward problems.

Predictive model	%-age of feasible predictions
3-D GANCER	95.3
2-D RGB GAN	90.1
2-D GANCER	82.3
2-D RGB GAN-sc.	83.9
RF-sc.	82.3
RF	86.2
2-D GANCER-sc.	87.7
3-D GANCER-sc.	86.9

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., $\hat{\mathcal{X}} = \{\hat{\mathbf{z}}_1, \ldots, \hat{\mathbf{z}}_8\}$. Note that we do not require $\hat{\mathbf{x}}_q = (\hat{\mathbf{w}}_q, \hat{\mathbf{d}}_q)$, but only the objective function values. Inverse optimization using this data set then yields a weight vector $\boldsymbol{\alpha}_k$, with which we solve $\mathbf{FO}(\boldsymbol{\alpha}_k)$ to obtain a reconstructed personalized treatment.

Dose predictions may be feasible and sub-optimal or infeasible. Recall from Proposition 6 and 6 that if all decisions in the data set, then solving the ensemble absolute or relative duality gap inverse optimization is equivalent to solving a single-point model using the centroid. Table EC.1 highlights the percentage of the patients for which the predictions are feasible dose distributions with respect to $\mathbf{RT}$–$\mathbf{FO}(\boldsymbol{\alpha})$. Typically we observe that about 85% of predictions are feasible, suggesting that there is usually at least one prediction for every patient which is infeasible.

EC.4.5. 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, as there is no clinical interpretation for negative objective function weights.

A specific inverse optimization problem is formulated by appropriately selecting the model hyperparameters $(\|\cdot\|, \mathcal{C}, \mathcal{E}_1, \ldots, \mathcal{E}_Q)$ from $\mathbf{GIO}(\hat{\mathcal{X}})$. In our experiments, we use the default parameters, except with the custom $\mathcal{C}$ to ensure the objective function is a weighted combination of the different objectives. Moreover, we set $\|\cdot\|_N = \|\cdot\|_1$.

EC.4.5.1. Absolute duality gap. Using Proposition 1 and our specific choice of $\mathcal{C}$, we formulate an absolute duality gap inverse optimization problem:

$$\begin{aligned} \mathbf{RT\text{–}IO}_{\mathrm{A}}(\hat{\mathcal{X}}): \quad & \min_{\boldsymbol{\alpha}, \mathbf{y}, \epsilon_1, \ldots, \epsilon_Q} \sum_{q=1}^{Q} |\epsilon_q| \\ & \text{s.t.} \quad \mathbf{C}^\intercal \boldsymbol{\alpha} \geq \mathbf{A}^\intercal \mathbf{y}, \quad \mathbf{y} \geq \mathbf{0} \\ & \qquad \boldsymbol{\alpha}^\intercal \hat{\mathbf{z}}_q = \mathbf{b}^\intercal \mathbf{y} + \epsilon_q, \quad \forall q \in \mathcal{Q} \\ & \qquad (\mathbf{C}^\intercal \boldsymbol{\alpha})^\intercal \mathbf{1} = 1 \\ & \qquad \boldsymbol{\alpha} \geq \mathbf{0}. \end{aligned} \tag{EC.33}$$

$\mathbf{RT\text{–}IO}_{\mathrm{A}}(\hat{\mathcal{X}})$ is obtained by substituting $\mathbf{c} = \mathbf{C}^\intercal \boldsymbol{\alpha}$ into formulation (4), and noting that $\left\|\mathbf{C}^\intercal \boldsymbol{\alpha}\right\|_1 = (\mathbf{C}^\intercal \boldsymbol{\alpha})^\intercal \mathbf{1}$ when both $\boldsymbol{\alpha} \geq \mathbf{0}$ and $\mathbf{C} \geq \mathbf{0}$.

EC.4.5.2. Relative duality gap. Using Proposition 4 and our specific choice of $\mathcal{C}$, we formulate a relative duality gap inverse optimization problem. We then use Corollary EC.1 to obtain the LP relaxation of the relative duality gap problem. The two relevant formulations are given below.

$$\begin{aligned} & \mathbf{RT\text{–}IO}_{\mathrm{R}}(\hat{\mathcal{X}}): \\ & \min_{\boldsymbol{\alpha}, \mathbf{y}, \epsilon_1, \ldots, \epsilon_Q} \sum_{q=1}^{Q} |\epsilon_q - 1| \\ & \text{s.t.} \quad \mathbf{C}^\intercal \boldsymbol{\alpha} \geq \mathbf{A}^\intercal \mathbf{y}, \quad \mathbf{y} \geq \mathbf{0} \\ & \qquad \boldsymbol{\alpha}^\intercal \hat{\mathbf{z}}_q = \epsilon_q \mathbf{b}^\intercal \mathbf{y}, \quad \forall q \in \mathcal{Q} \\ & \qquad (\mathbf{C}^\intercal \boldsymbol{\alpha})^\intercal \mathbf{1} = 1 \\ & \qquad \boldsymbol{\alpha} \geq \mathbf{0}. \end{aligned} \tag{EC.34}$$

$$\begin{aligned} & \mathbf{RT\text{–}IO}_{\mathrm{R,LP}}(\hat{\mathcal{X}}): \\ & \min_{\boldsymbol{\alpha}, \mathbf{y}, \epsilon_1, \ldots, \epsilon_Q} \sum_{q=1}^{Q} |\epsilon_q - 1| \\ & \text{s.t.} \quad \mathbf{C}^\intercal \boldsymbol{\alpha} \geq \mathbf{A}^\intercal \mathbf{y}, \quad \mathbf{y} \geq \mathbf{0} \\ & \qquad \boldsymbol{\alpha}^\intercal \hat{\mathbf{z}}_q = \epsilon_q, \quad \forall q \in \mathcal{Q} \\ & \qquad \mathbf{b}^\intercal \mathbf{y} = 1 \\ & \qquad \boldsymbol{\alpha} \geq \mathbf{0}. \end{aligned} \tag{EC.35}$$

Using Algorithm 1, we first solve the LP relaxation of $\mathbf{RT\text{–}IO}_{\mathrm{R}}(\hat{\mathcal{X}})$, stated above as $\mathbf{RT\text{–}IO}_{\mathrm{R,LP}}(\hat{\mathcal{X}})$. Note that this relaxation is the application-specific analogue of $\mathbf{GIO}^{+}_{\mathrm{R,LP}}(\hat{\mathcal{X}})$, which is only one of the three reformulations of the relative duality gap problem. We do not construct or solve relaxations of the other two (e.g., $\mathbf{GIO}^{-}_{\mathrm{R,LP}}(\hat{\mathcal{X}})$ and $\mathbf{GIO}^{0}_{\mathrm{R,LP}}(\hat{\mathcal{X}})$) due to the following reasons. First, the analogue to $\mathbf{GIO}^{-}_{\mathrm{R,LP}}(\hat{\mathcal{X}})$ is infeasible; in our application, $\mathbf{b} \geq \mathbf{0}$ implying $\mathbf{b}^{\mathsf{T}}\mathbf{y} \geq \mathbf{0}$ for all $\mathbf{y} \geq \mathbf{0}$. Second, the application-specific analogue of $\mathbf{GIO}^{0}_{\mathrm{R}}(\hat{\mathcal{X}})$ in practice is often infeasible or generates plans that perform poorly on the clinical criteria satisfaction metrics compared to $\mathbf{RT\text{–}IO}_{\mathrm{R,LP}}(\hat{\mathcal{X}})$. Recall that $\mathbf{GIO}^{0}_{\mathrm{R}}(\hat{\mathcal{X}})$ 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. The only objectives where this situation could occur are the threshold objectives (EC.16)–(EC.18). Thus, the application-specific analogue of $\mathbf{GIO}^{0}_{\mathrm{R}}(\hat{\mathcal{X}})$ is either infeasible or distributes all of the objective weights to these three objectives. By strictly focusing on the threshold objectives however, the inverse problem then generally fails to meet a large number of the clinical criteria. Consequently, we advocate in this application to strictly use $\mathbf{RT\text{–}IO}_{\mathrm{R,LP}}(\hat{\mathcal{X}})$ to solve the relative duality gap inverse optimization problem.

EC.4.6. Baseline implementations

In Section 5.3, we implement two conventional ensemble learning baselines to compare with ensemble inverse optimization. The first baseline is an ensemble-then-inverse optimization model. Here, we first compute the average of the individual decisions and then solve a single-point inverse optimization problem to obtain a cost vector. The second baseline is a Multiplicative Weights Algorithm (MWA). In our experiments, we implemented both models using all eight predictions as well as for the 4 Pts. predictions (RF-sc., RF, 2-D GANCER-sc., 3-D GANCER-sc.). We also use grid search with the training set patients to identify the best learning rate for the MWA.

Algorithm 2 summarizes the steps for the MWA. We use an offline learning variant of the Weighted Majority update rule of Arora et al. (2012). Each of the prediction models $F_q(\cdot)$ in the

Algorithm 2 Multiplicative Weights Algorithm Baseline

Input: Data set of CT images for training patients $\mathcal{C}$, Data set of CT images for testing patients $\tilde{\mathcal{C}}$, Dose prediction models $F_1(\cdot), \ldots, F_Q(\cdot)$, Learning rate $\eta \leq 0.5$.

Output: Treatment plans for each patient

Initialize weights $w_q = 1$ for $q \in \mathcal{Q}$.
for Each patient in the training data set $\hat{\mathbf{c}}_k \in \tilde{\mathcal{C}}$ **do**
  for $q \in \mathcal{Q}$ **do**
    Let $\hat{\mathbf{d}}_{q,k} \leftarrow F_q(\hat{\mathbf{c}}_k)$.
    Let $z_{q,k} \leftarrow \mathbf{RT\text{–}IO}_{\mathrm{R}}(\{\hat{\mathbf{d}}_{q,k}\})$.
    Let $w_q \leftarrow w_q(1 - \eta z_{q,k})$.
  end for
end for
Normalize weights $w_q \leftarrow w_q / (\sum_{q'=1}^{Q} w_{q'})$.
for Each patient in the testing data set $\hat{\mathbf{c}}_k \in \mathcal{C}$ **do**
  Select prediction model $F_q(\cdot)$ with probability w_q.
  Let $\hat{\mathbf{d}}_{q,k} \leftarrow F_q(\hat{\mathbf{c}}_k)$.
  Let $\boldsymbol{\alpha}_k^*$ be the optimal solution to $\mathbf{RT\text{–}IO}_{\mathrm{R}}(\{\hat{\mathbf{d}}_{q,k}\})$.
  Let $\mathbf{x}_k^* \leftarrow \mathbf{RT\text{–}FO}(\boldsymbol{\alpha}_k^*)$ and evaluate the corresponding treatment plan.
end for

ensemble KBP pipeline is treated as an expert and we initialize a weight $w_q = 1$ for each model. For each of the 130 training set patients k and each prediction model, we predict a dose $\hat{d}_{q,k}$, solve a single-point inverse optimization problem and update the weight w_q by a penalty factor corresponding to the aggregate error of the inverse optimization problem. After repeating this process for all of the training set patients, we normalize the weights to a probability distribution and freeze them. Then for each of the patients in the test set, we randomly select an 'expert' KBP pipeline to generate a treatment plan.