Supplementary Material

EC.1. SPO Loss Functions and Duality

EC.1.1. SPO Loss and Duality for the Newsvendor Problem

As we noted in Section 1, Elmachtoub et al. (2020) develop an SPO approach for optimization problems with side information of the form

$$\min_{\phi \in \widetilde{\Phi}} \mathbb{E}\left[U^\intercal \phi(V)\right], \tag{EC.1}$$

where U and V are an exogenous parameter vector and its associated side information, respectively, and where $\widetilde{\Phi}$ is the set of policies mapping values of side information into feasible region $\widetilde{\mathscr{X}}$, which is assumed to be convex. Given a prediction $\hat{u}$ of the exogenous parameter, the corresponding decision selection problem is

$$\min_{x \in \widetilde{\mathscr{X}}} \hat{u}^\intercal x. \tag{EC.2}$$

In the work we present in this paper, we consider the optimization problem with side information

$$\min_{\phi \in \Phi} \mathbb{E}\left[c^\intercal \phi(V) + \min_{y \geq 0} \{q^\intercal y \mid T\phi(V) + Wy = U\}\right], \tag{EC.3}$$

where $\mathscr{X}$ is a polyhedron $\{x \in \mathbb{R}^{n_1} \mid Ax = b, x \geq 0\}$.

Given a prediction $\hat{u}$ for the value of the exogenous parameter, the corresponding decision selection problem can be written as

$$\min_{x,y \in \mathbb{R}^{n_1+n_2}} \{c^\intercal x + q^\intercal y \mid Ax = b, Tx + Wy = \hat{u}, x \geq 0, y \geq 0\} \tag{EC.4}$$

and its dual is

$$\max_{\mu,\lambda \in \mathbb{R}^{m_1+m_2}} \{b^\intercal \mu + \hat{u}^\intercal \lambda \mid A^\intercal \mu + T^\intercal \lambda \leq c, W^\intercal \lambda \leq q\}. \tag{EC.5}$$

Let $\widetilde{\Phi}$ denote the set of policies that produce an action that is feasible for (EC.5). The dual problem (EC.5) is the decision selection problem for the optimization problem with side information

$$\max_{\phi \in \widetilde{\Phi}} \mathbb{E}\left[\left[\begin{smallmatrix} b \\ U \end{smallmatrix}\right]^\intercal \phi(V)\right], \tag{EC.6}$$

which may be viewed as a special case of (EC.2). For this reason, it may be tempting to apply the Elmachtoub and Grigas (2022) procedure to (EC.6) in order to produce a predictor ψ, and subsequently use ψ as the predictor in a predict-then-optimize procedure for (EC.3). However, we argue that such an approach is not likely to work well. While problems (EC.3) and (EC.6) have corresponding decision problems that are dual with each other, the underlying optimization

problems (EC.3) and (EC.6) are not necessarily dual, and the SPO loss produced by one can be structurally quite different from the other. For this reason, minimizing the empirical SPO loss of (EC.6) does not necessarily minimize the empirical SPO loss of (EC.3). This observation does not take away from the value of the Elmachtoub and Grigas (2022) procedure as it was not intended to be applied in such a way.

We expand on this discussion next. Throughout the section, we use the term *primal decision selection problem* to refer to problem (EC.4) whereas we use the term *dual decision selection problem* to refer to problem (EC.5).

Consider the following simple variant of the newsvendor problem. A newsvendor observes some side information V, which can be used to predict the demand for the newspaper, which is not known in advance. Based on this information, the newsvendor choose the quantity x of newspapers to stock. Then, the newsvendor sells as many newspapers as possible at a unit price of q. The number of newspapers sold cannot be greater than the stocked quantity x, nor can it be greater than the quantity demanded, which is a value that we denote by U. We assume that there is a cost c to stock a newspaper. This leads to the optimization problem with side information

$$\min_{\phi\in\Phi} \mathbb{E}\left[c\phi(V) + \min_{y\geq 0}\{-qy \mid y\leq \phi(V),\, y\leq U\}\right].$$

The objective is to minimize the expected difference between costs and revenues, which is equivalent to maximizing profits. We assume that $q > c$; otherwise the optimal solution is to stock no newspapers regardless of demand. We also assume that $c \geq 0$, since the objective is unbounded when $c < 0$. The constraints, in sequence, state that the quantity of newspapers sold is less than or equal to the quantity stocked and the demand. Quantities stocked and sold are nonnegative. For an estimate u' of the demand, the primal decision selection problem is given by

$$\min_{x\geq 0}\left\{cx + \min_{y\geq 0}\{-qy \mid y\leq x,\, y\leq u'\}\right\}.$$

If u' is nonnegative, this has a unique optimal decision given by $x^* = u'$, which achieves an objective value of $(c-q)u'$. Given this decision and a true observation u of the exogenous parameter, the regret is given by

$$\left(cu' + \min_{y\geq 0}\{-qy \mid y\leq u',\, y\leq u\}\right) - \left(\min_{x\geq 0} cx + \min_{y\geq 0}\{-qy \mid y\leq x,\, y\leq u\}\right).$$

Consider the term $\min_{y\geq 0}\{-qy \mid y\leq u',\, y\leq u\}$. This optimization problem is minimized by $y^* = \min\{u', u\}$, which produces the value $-q\min\{u', u\}$. It follows that the SPO loss function $\ell(\cdot, u)$ derived from the primal decision selection problem is given by

$$\ell(u', u) = \begin{cases} (q-c)(u-u') & \text{if } u > u'; \\ c(u'-u) & \text{if } u \leq u'. \end{cases}$$

This loss function is defined only when u and u' are nonnegative.

Given some prediction u' of the exogenous parameter, the dual decision selection problem is

$$\max_{(\mu,\lambda)\in\mathbb{R}^2} \{u'\lambda \mid \mu \le c,\, \lambda - \mu \le -q,\, \mu \ge 0,\, \lambda \le 0\}. \tag{EC.7}$$

If u' is positive, this problem has a unique optimal solution given by $(\mu^*, \lambda^*) = (c, c-q)$. Given a true observation of the exogenous parameter u, this would incur an objective value of $u(c-q)$. When $u' = 0$, every feasible solution to the decision selection problem is optimal. When u' is negative, the decision selection problem is unbounded. Given a prediction u' and observation u of the exogenous parameter, the associated regret is then

$$\max_{(\mu,\lambda)} \{u\lambda \mid \mu \le c,\, \lambda - \mu \le -q,\, \mu \ge 0,\, \lambda \le 0\} - u(c-q) = u(c-q) - u(c-q) = 0.$$

assuming that u' is positive and u is nonnegative. The regret is not well-defined otherwise.

This implies that the SPO loss $\ell_D(u', u)$ derived from the dual decision selection problem is defined only when u' is positive and u is nonnegative, and is exactly zero when it is well-defined. This is in sharp contrast to the form of $\ell(u', u)$ derived above. Hence, the loss function $\ell_D(\cdot, u)$ is an unsuitable substitute for that derived from the primal decision selection problem $\ell(\cdot, u)$, and it would be ill-advised to use $\ell_D(\cdot, u)$ to fit a function to predict values of the exogenous parameter.

Elmachtoub and Grigas (2022) also propose a convex approximation to the SPO loss, called the *SPO+ loss function*. Similarly to the exact SPO loss, deriving an SPO+ loss from the dual decision problem does not produce a loss function that is useful for the primal problem. Again, this should not be taken as criticism of the SPO+ loss, as this loss was not intended to be applied in this manner. Let σ be a selection function for the dual decision problem, which is to say that σ is function that takes a value u of the exogenous parameter and returns an optimal solution to the decision selection problem, $\sigma(u) \in \arg\max_{x\in\widetilde{\mathscr{X}}} u^\intercal x$. Then, the SPO+ loss is given as follows:

$$\ell^\sigma_{SPO+}(u', u) = \left(\max_{x\in\widetilde{\mathscr{X}}} u^\intercal x\right) - 2u'^\intercal \sigma(u) - \min_{x\in\widetilde{\mathscr{X}}} \{(u - 2u')^\intercal x\}.$$

For $u > 0$, this can be specialized to the dual decision selection problem (EC.7) as

$$\ell^\sigma_{SPO+}(u', u) = u(c-q) - 2u'(c-q) - \min_{(\mu,\lambda)\in\mathbb{R}^2} \{(u-2u')\lambda \mid \mu \le c, \lambda - \mu \le -q, \mu \ge 0, \lambda \le 0\}. \tag{EC.8}$$

Here, we make use of the previously-discussed fact that when the demand is predicted to be some positive real number u, the unique solution to the dual decision selection problem (EC.7) is given by $(\mu^*, \lambda^*) = (c, c-q)$. By similar arguments as before, if $u < 0$, then the loss is undefined, as the

dual decision selection problem (EC.7) has no optimal solution. In the case where $u = 0$, the SPO+ loss depends on the decision selection function. Consider the last term of (EC.8):

$$\min_{(\mu,\lambda)\in\mathbb{R}^2} \{(u-2u')\lambda \mid \mu \le c, \lambda - \mu \le -q, \mu \ge 0, \lambda \le 0\}. \tag{EC.9}$$

If $u' \le u/2$, then the optimal solution to (EC.9) is again given by $(\mu^*, \lambda^*) = (c, c-q)$ and this term reduces to $(u-2u')(c-q)$. If $u' > u/2$, then the optimization problem (EC.9) is unbounded. This implies that if $u > 0$, the loss $\ell_{SPO+}(u', u)$ is zero for all $u' \le u/2$ and is infinite when $u' > u/2$. This is again not a useful substitute for the primal SPO loss.

EC.1.2. SPO Loss and Duality for the Production Planning and Transportation Problem

A similar phenomenon occurs in the optimization problem used in the computational experiments of Section 5. For this problem, the decision selection problem takes the form

$$\min_{x\ge 0, y\ge 0} \left\{ c^\intercal x + \sum_{\substack{s\in\mathscr{S}\\ t\in\mathscr{T}}} r_{st}^{\text{ship}} y_{st}^{\text{ship}} + \sum_{s\in\mathscr{S}} r_s^{\text{scrap}} y_s^{\text{scrap}} + \sum_{t\in\mathscr{T}} r_t^{\text{unmet}} y_t^{\text{unmet}} \,\middle|\, \begin{array}{ll} x_s = y_s^{\text{scrap}} + \sum_{t\in\mathscr{T}} y_{st}^{\text{ship}}, & \forall s \in \mathscr{S} \\ u_t = y_t^{\text{unmet}} + \sum_{s\in\mathscr{S}} y_{st}^{\text{ship}}, & \forall t \in \mathscr{T} \end{array} \right\}.$$

The dual of this problem, which we again refer to as the dual decision selection problem, is given by

$$\max_{\mu\in\mathbb{R}^{|\mathscr{S}|}, \lambda\in\mathbb{R}^{|\mathscr{T}|}} \left\{ u^\intercal\lambda \,\middle|\, \begin{array}{ll} \mu_s \ge -c_s, & \forall s \in \mathscr{S} \\ \mu_s \le r_s^{\text{scrap}}, & \forall s \in \mathscr{S} \\ \lambda_t \le r_t^{\text{unmet}}, & \forall t \in \mathscr{T} \\ \mu_s + \lambda_t \le r_{st}^{\text{ship}}, & \forall s \in \mathscr{S}\, \forall t \in \mathscr{T} \end{array} \right\}.$$

The primal SPO loss $\ell(u', u)$ for the production planning and transportation planning problem is more complicated than that for the newsvendor problem, and we do not derive an expression for this loss function. However, it is easily observed that the loss $\ell(u', u)$ is typically not equal to zero when $u' \ne u$ and both u' and u are positive vectors. Furthermore given an estimate u' of demand such that $u'_s < 0$ for some $s \in \mathscr{S}$, the primal decision selection problem is infeasible and the dual selection problem is unbounded, so both the primal SPO loss and dual SPO loss are undefined if $u' \ngeq 0$ or $u \ngeq 0$.

Consider the SPO loss $\ell_D(u', u)$ corresponding to the dual decision selection problem. Given a positive estimate $u' > 0$ for the demand, it can be shown that the unique optimal solution to the decision selection problem is given by (μ^*, λ^*) where $\mu^* = -c$ and

$$\lambda_t^* = \min\left\{ r_t^{\text{unmet}}, \min_{s\in\mathscr{S}} \left\{ r_{st}^{\text{ship}} + c_s \right\} \right\}. \tag{EC.10}$$

Equation (EC.10) does not depend on u' beyond the fact that u' must be positive for this expression to be valid. Thus, this is also the unique optimal solution to the decision selection problem where the demand is known to be u, assuming that $u > 0$. This implies that if $u' > 0$ and $u > 0$, then $\ell_D(u', u) = 0$, and we again observe that the dual SPO loss is not useful.

EC.2. Proof of Results from Section 2

For simplicity of notation, we first prove some results concerning the single-stage LP $\mathscr{P} = \min_x \{c^\intercal x | Ax = b, x \geq 0\}$. In this section, we refer to the dimension of x as n.

LEMMA EC.1. *Let B be a basis of A such that the corresponding reduced costs are nonnegative and $A_B^{-1}b \geq 0$. Let ρ be the indices of variables x with positive reduced costs under basis B. Then, the set of optimal solutions to $\mathscr{P}$ is given by*

$$\{x \mid x_\rho = 0, Ax = b, x \geq 0\}.$$

Proof of Lemma EC.1. Let c^* denote the optimal value of $\mathscr{P}$. Since the basis B has nonnegative reduced costs and $A_B^{-1}b \geq 0$, then c^* exists and is given by $c^* = c^\intercal A_B^{-1}b$. Let x be any feasible solution to $\mathscr{P}$. Let ζ (*resp.* ρ) be the indices of non-basic variables with zero (*resp.* positive) reduced costs. Then,

$$x_B = A_B^{-1}b - A_B^{-1}A_\zeta x_\zeta - A_B^{-1}A_\rho x_\rho.$$

Further,

$$c^\intercal x = c_B^\intercal A_B^{-1}b + \left(c_\zeta^\intercal - c_B^\intercal A_B^{-1}A_\zeta\right)x_\zeta + \left(c_\rho^\intercal - c_B^\intercal A_B^{-1}A_\rho\right)x_\rho = c_B^\intercal A_B^{-1}b + \left(c_\rho^\intercal - c_B^\intercal A_B^{-1}A_\rho\right)x_\rho,$$

since $c_\zeta^\intercal - c_B^\intercal A_B^{-1}A_\zeta$ are the reduced costs of the variables x_ζ, which by definition are zero. Similarly, the vector $c_\rho^\intercal - c_B^\intercal A_B^{-1}A_\rho$ contains the reduced costs of the variables x_ρ, which by definition are positive. Since x is feasible, it must be that $x \geq 0$. Thus, $\left(c_\rho^\intercal - c_B^\intercal A_B^{-1}A_\rho\right)x_\rho$ is zero if and only if $x_\rho = 0$. We conclude that $c^\intercal x = c_B^\intercal A_B^{-1}b = c^*$ if and only if $x_\rho = 0$. This is the desired result. □

The following results require an extension of the definition of a lexicographically optimal solution and of a lex-prc basis. Given an index i in $[n]$, we define a partial ordering called the *lexicographical partial ordering up to component i*, denoted by $\prec_i$, so that $x \prec_i x'$ holds if and only if there exists some $j \in [i]$ such that $x_j < x'_j$ and $x_k = x'_k$ for any positive integer k in $[j-1]$. Let $\mathbb{X}^*$ denote the set of optimal solutions to $\mathscr{P}$ and let $\mathbb{X}_i^*$ denote the subset of $\mathbb{X}^*$ that is minimal under $\prec_i$. We will refer to the elements of $\mathbb{X}_i^*$ as *i-lexmin solutions.*

DEFINITION EC.1. Let B be a basis of A. Let $\zeta(0)$ be the set of variables with zero reduced costs under basis B and cost vector c, *i.e.*, $\zeta(0) = \{i \in [n] \mid c_i - c_B^\intercal A_B^{-1}A_i = 0\}$. For each $i \in [n]$, let $\zeta(i)$ be the subset of $\zeta(i-1)$ with zero reduced costs when the cost vector is chosen to be $\boldsymbol{e}(i)$, *i.e.*, $\zeta(i) = \{j \in \zeta(i-1) \mid \boldsymbol{e}(i)_j - \boldsymbol{e}(i)_B^\intercal A_B^{-1}A_j = 0\}$. Let k be in $[n]$. Basis B is *lexicographically positive-reduced-costs up to component k (k-lex-prc)* if:

1. $c^\intercal - c_B^\intercal A_B^{-1}A \geq 0$.
2. $\boldsymbol{e}(i)_j - \boldsymbol{e}(i)_B^\intercal A_B^{-1}A_j \geq 0$ for each $i \in [k]$ and for each $j \in \zeta(i-1)$.

PROPOSITION EC.1. *If B is a k-lex-prc basis for A such that $A_B^{-1}b \geq 0$, then the solution x given by $(x_B, x_{[n]\setminus B}) = (A_B^{-1}b, 0)$ is an element of $\mathbb{X}_k^*$.*

Proof of Proposition EC.1. Let B be a k-lex-prc basis of A such that $A_B^{-1}b \geq 0$. We define the 0^{th} level problem $\mathscr{P}_0$ to be $\mathscr{P}$, and we let $\mathbb{X}_0^* = \mathbb{X}^*$. Further, for $j \geq 1$, we define the j^{th} level problem, $\mathscr{P}_j$, to be

$$\min_{x \in \mathbb{R}^n} \left\{ x_j \mid x \in \mathbb{X}_{j-1}^* \right\},$$

The set of optimal solutions to $\mathscr{P}_j$ is exactly $\mathbb{X}_j^*$.

Let $\zeta(j)$ be as in the definition of a k-lex-prc basis, *i.e.*,

$$\begin{aligned}
\zeta(0) &= \left\{ i \in [n] \mid c_i - c_B^\intercal A_B^{-1} A_i = 0 \right\} \\
\zeta(j) &= \left\{ i \in \zeta(j-1) \mid \boldsymbol{e}(j)_i - \boldsymbol{e}(j)_B^\intercal A_B^{-1} A_i = 0 \right\},
\end{aligned}$$

for j in $[k-1]$. Let $\rho(0)$ be the set of indices of variables with positive reduced costs under cost vector c, and let $\rho(j)$ be the subset of $\zeta(j-1)$ with positive reduced costs under cost function $\boldsymbol{e}(j)$:

$$\begin{aligned}
\rho(0) &= \left\{ i \in [n] \mid c_i - c_B^\intercal A_B^{-1} A_i > 0 \right\} \\
\rho(j) &= \left\{ i \in \zeta(j-1) \mid \boldsymbol{e}(j)_i - \boldsymbol{e}(j)_B^\intercal A_B^{-1} A_i > 0 \right\}.
\end{aligned}$$

Lemma EC.1 implies that $\mathscr{P}_k$ can be formulated as

$$\min_x \left\{ x_k \mid Ax = b,\, x \geq 0,\, x_{\rho(i)} = 0, \forall i \in [k-1] \right\}.$$

Clearly, we can omit the variables whose indices fall in the set $\rho(i)$ for some i in $[k-1]$; these are precisely the variables whose indices are not in the set $\zeta(k-1)$. This leads to the equivalent problem $\mathscr{P}_k'$ defined as

$$\min_{x_{\zeta(k-1)}} \left\{ x_k \mid A_{\zeta(k-1)} x_{\zeta(k-1)} = b,\, x_{\zeta(k-1)} \geq 0 \right\}.$$

Any optimal solution for $\mathscr{P}_k'$ can be extended into an optimal solution to $\mathscr{P}_k$ by assigning a value of zero to each variable whose index is not in $\zeta(k-1)$. For the given k-lex-prc basis B, we define a solution χ' for $\mathscr{P}_k'(b)$ by $\chi_B' = A_B^{-1}b, \chi_i' = 0,\ \forall i \in \zeta(k-1) \setminus B$. By definition, B must be a subset of $\zeta(k-1)$, since basic variables always have zero reduced costs. This ensures that χ' is well-defined. By assumption, $A_B^{-1}b$ is nonnegative, so χ' is feasible. Further, it is an optimal solution for $\mathscr{P}_k'$, since it is a basic solution associated with basis B, which is a basis for $\mathscr{P}_k'$ corresponding to a basic feasible solution whose associated reduced costs are nonnegative. As previously noted, the solution χ' can be extended into an optimal solution χ for $\mathscr{P}_k$ by assigning a value of zero to each variable whose index is not in $\zeta(k-1)$. This results in the solution $(\chi_B, \chi_{[n]\setminus B}) = (A_B^{-1}b, 0)$. Since χ is an optimal solution for $\mathscr{P}_k$, it must be in $\mathbb{X}_k^*$. The form of this solution matches that given in the statement of the proposition, proving the result. $\square$

Proposition EC.1 immediately implies Proposition 1.

A procedure for identifying lexicographically minimal solutions is given in Akgul (1984). We further show that it can be used to identify a corresponding k-lex-prc basis. The procedure is summarized in Algorithm 3, where we use the notation $\mathscr{L}(b, \mathcal{I}, i)$ to refer to the optimization problem

$$\min_x \{x_i \mid A_{\mathcal{I}} x_{\mathcal{I}} = b,\, x_{\mathcal{I}} \geq 0\}.$$

where $\mathcal{I}$ is a subset of $[n]$ and i in $\mathcal{I}$.

Algorithm 3 Let A be an $m \times n$ constraint matrix, b a right-hand side vector, c an objective vector, and k a variable index in $[n]$. This procedure finds a k-lexmin basic solution and k-lex-prc basis

procedure LexMin(k, b, A, c)
  Let x^* be an optimal solution to $\mathscr{P}$ with corresponding basis $B(0)$.
  Let $\mathcal{I}(0) = [n]$.
  for j in $[n] \setminus B(0)$ **do**
    Calculate $r_j = c_j - c_{B(0)}^{\intercal} A_{B(0)}^{-1} A_j$.
    If $r_j > 0$, remove j from $\mathcal{I}(0)$.
  end for
  for i in $[k]$ **do**
    if $i \in \mathcal{I}(i-1)$ **then**
      Solve $\mathscr{L}(b, \mathcal{I}(i-1), i)$; let x^* be an optimal solution with corresponding basis $B(i)$.
      Let $\mathcal{I}(i) = \mathcal{I}(i-1)$.
      for j in $\mathcal{I}(i-1) \setminus B(i)$ **do**
        Calculate $r_j = \boldsymbol{e}(i)_j - \boldsymbol{e}(i)_{B(i)}^{\intercal} A_{B(i)}^{-1} A_j$.
        If $r_j > 0$, remove j from $\mathcal{I}(i)$.
      end for
    else
      Let $B(i) = B(i-1)$.
      Let $\mathcal{I}(i) = \mathcal{I}(i-1)$.
    end if
  end for
  return k-lexmin solution x^* and corresponding k-lex-prc basis $B(k)$.
end procedure

PROPOSITION EC.2. *Suppose that $\mathscr{P}$ has an optimal solution and the rows of A are linearly independent. Algorithm 3 returns an element x^* of $\mathbb{X}_k^*$ and a corresponding k-lex-prc basis B.* □

Proof of Proposition EC.2. Throughout this proof, whenever we use $\mathcal{I}(i)$ for $i \in [k]$, we refer to the value at the termination of the algorithm. We also use the notation $\mathscr{T}(0, B)$ to denote the simplex tableau of $\mathscr{P}$ under basis B, whereas $\mathscr{T}(i, B)$ denotes the simplex tableau of $\mathscr{L}(b, \mathcal{I}(i-1), i)$ under basis B.

First, we claim that the algorithm is well-defined, which is to say that the optimal basic solution produced in steps 2 and 10 exists. By assumption, $\mathscr{P}$ has an optimal solution and the constraint A matrix has linearly independent rows; the simplex algorithm will then return an optimal basic solution. Thus, step 2 is well-defined. The problem $\mathscr{L}(b, \mathcal{I}(0), 1)$ is feasible because the optimal solution x^* produced in step 2 is feasible to this problem; likewise, for $i > 1$ the problem $\mathscr{L}(b, \mathcal{I}(i-1), i)$ is feasible for each i, because the optimal solution x^* produced in the previous iteration is a feasible solution if $i > 1$. Also, it is clear that the objective of this problem is bounded below by zero, so the problem is not unbounded. Thus, the optimal basic solutions referred to in steps 2 and 10 exist.

To prove the result, we show by induction that the following hypotheses hold for each $i \in \{0, \ldots, k\}$:

1. $c^\intercal - c_{B(i)}^\intercal A_{B(i)}^{-1} A = c^\intercal - c_{B(0)}^\intercal A_{B(0)}^{-1} A$,
2. $\boldsymbol{e}(j)_{\mathcal{I}(j-1)}^\intercal - c_{B(i)}^\intercal A_{B(i)}^{-1} A_{\mathcal{I}(j-1)} = \boldsymbol{e}(j)_{\mathcal{I}(j-1)}^\intercal - c_{B(j)}^\intercal A_{B(j)}^{-1} A_{\mathcal{I}(j-1)}$, for $i \geq 1$ and $j \in [i]$.

The case in which $i = 0$ holds trivially. Suppose that the hypotheses hold when $i = \iota$ for some ι in $\{0, \ldots, k-1\}$. By definition, all variables in $\mathcal{I}(0)$ have zero reduced costs in the tableau $\mathscr{T}(0, B(0))$. By the first induction hypothesis, the reduced costs of the variables $\mathcal{I}(0)$ must also be zero in the tableau $\mathscr{T}(0, B(\iota))$. The set $\mathcal{I}(\iota)$ is a subset of $\mathcal{I}(0)$, so all variables in $\mathcal{I}(\iota)$ must have zero reduced costs in the tableau $\mathscr{T}(0, B(\iota))$. By definition, $B(\iota)$ is a subset of $\mathcal{I}(\iota)$, and it can easily be verified that $B(\iota)$ corresponds to a feasible basic solution of $\mathscr{L}(b, \mathcal{I}(\iota), \iota+1)$. By definition, $B(\iota+1)$ also corresponds to a basic feasible solution of $\mathscr{L}(b, \mathcal{I}(\iota), \iota+1)$. This implies that there is a series of simplex pivots that maintain a feasible basic solution and that transform $\mathscr{T}(\iota+1, B(\iota))$ into $\mathscr{T}(\iota+1, B(\iota+1))$. The same sequence of pivots could be applied to $\mathscr{T}(0, B(\iota))$ to transform it into $\mathscr{T}(0, B(\iota+1))$. These pivots only use variables in $\mathcal{I}(\iota)$, since these are the variables of $\mathscr{L}(b, \mathcal{I}(\iota), \iota+1)$. Since all of these variables have zero reduced costs in the tableau $\mathscr{T}(0, B(\iota))$, then pivoting any of these variables does not affect reduced costs. This implies that the reduced costs in the tableau $\mathscr{T}(0, B(\iota+1))$ must be the same as those of $\mathscr{T}(0, B(\iota))$. The first induction hypothesis then gives that $c^\intercal - c_{B(\iota+1)}^\intercal A_{B(\iota+1)}^{-1} A = c^\intercal - c_{B(0)}^\intercal A_{B(0)}^{-1} A$. This shows that the first induction hypothesis holds.

The proof of the second induction hypothesis is nearly identical. Let $j \in [\iota+1]$. If $j = \iota+1$, then the second induction hypothesis holds trivially, so we can assume that $j \leq \iota$. By definition, all of the variables in $\mathcal{I}(j)$ have zero reduced costs in the tableau $\mathscr{T}(j, B(j))$. By the second induction hypothesis, the reduced costs of the variables $\mathcal{I}(j)$ must also be zero in the tableau $\mathscr{T}(j, B(\iota))$. The set $\mathcal{I}(\iota)$ is a subset of $\mathcal{I}(j)$, so all variables in $\mathcal{I}(\iota)$ must have zero reduced costs in the tableau $\mathscr{T}(j, B(\iota))$. By definition, $B(\iota)$ is a subset of $\mathcal{I}(\iota)$, and it can easily be verified that $B(\iota)$ corresponds to a feasible basic solution of $\mathscr{L}(b, \mathcal{I}(\iota), \iota+1)$. By definition, $B(\iota+1)$ also corresponds to a basic feasible solution of $\mathscr{L}(b, \mathcal{I}(\iota), \iota+1)$. This implies that there is a series of simplex pivots that maintain a feasible basic solution and that transform $\mathscr{T}(\iota+1, B(\iota))$ into $\mathscr{T}(\iota+1, B(\iota+1))$. The same sequence of pivots could be applied to $\mathscr{T}(j, B(\iota))$ to transform it into $\mathscr{T}(j, B(\iota+1))$. These pivots only use variables in $\mathcal{I}(\iota)$, since these are the variables of $\mathscr{L}(b, \mathcal{I}(\iota), \iota+1)$. All of these variables have zero reduced costs in the tableau $\mathscr{T}(j, B(\iota))$, so entering any of these variables does not affect the reduced costs. This implies that the reduced costs in the tableau $\mathscr{T}(j, B(\iota+1))$ must be the same as those of $\mathscr{T}(j, B(\iota))$. Applying the second induction hypothesis again, we obtain that

$$\boldsymbol{e}(j)^\intercal_{\mathcal{I}(j-1)} - c^\intercal_{B(\iota+1)} A^{-1}_{B(\iota+1)} A_{\mathcal{I}(j-1)} = \boldsymbol{e}(j)^\intercal_{\mathcal{I}(j-1)} - c^\intercal_{B(j)} A^{-1}_{B(j)} A_{\mathcal{I}(j-1)}.$$

This completes the proof of the second induction hypothesis.

We are now ready to prove that $B(k)$ is a k-lex-prc basis. Following the definition of a k-lex-prc basis, we let $\zeta(0) = \{i \in [n] \mid c_i^\intercal - c^\intercal_{B(k)} A^{-1}_{B(k)} A_i = 0\}$ and $\zeta(i) = \{j \in \zeta(i-1) \mid \boldsymbol{e}(i)^\intercal_j - \boldsymbol{e}(i)^\intercal_{B(k)} A^{-1}_{B(k)} A_j = 0\}$. It follows immediately from the two hypotheses and from the algorithm that $\zeta(0) = \mathcal{I}(0)$ and that $\zeta(i) = \mathcal{I}(i)$ for any $i \in [k]$. Furthermore, it also follows immediately from the definition of the algorithm that $\boldsymbol{c}^\intercal - c^\intercal_{B(0)} A^{-1}_{B(0)} A \geq 0$ and $\boldsymbol{e}(i)^\intercal_{\mathcal{I}(i-1)} - c^\intercal_{B(i)} A^{-1}_{B(i)} A_{\mathcal{I}(i-1)} \geq 0$, for i in $[k]$. Then, using the first hypothesis,

$$\boldsymbol{c}^\intercal - c^\intercal_{B(k)} A^{-1}_{B(k)} A = \boldsymbol{c}^\intercal - c^\intercal_{B(0)} A^{-1}_{B(0)} A \geq 0$$

and, using the second hypothesis,

$$\boldsymbol{e}(i)^\intercal_{\zeta(i-1)} - c^\intercal_{B(k)} A^{-1}_{B(k)} A_{\zeta(i-1)} = \boldsymbol{e}(i)^\intercal_{\mathcal{I}(i-1)} - c^\intercal_{B(i)} A^{-1}_{B(i)} A_{\mathcal{I}(i-1)} \geq 0.$$

This completes the proof that $B(k)$ is a k-lex-prc basis. □

Proposition EC.2 implies that LexMin$(n_1, \left[\begin{smallmatrix} b \\ u \end{smallmatrix}\right], \Lambda, \left[\begin{smallmatrix} c \\ q \end{smallmatrix}\right])$ provides a solution (x^*, y^*) and a n_1-lex-prc basis B of Λ such that x^* is the lexmin solution. This provides a proof of Proposition 2.

Recall the statement of Theorem 1:

THEOREM 1. *Under Assumptions 1, 2, 3, and 4, the lex-SPO loss function $\ell(\cdot, u)$ is continuous and piecewise-linear for each u in $\mathscr{F}_0$ with a finite number of pieces.* □

Proof of Theorem 1. First, we construct a suitable polyhedral subdivision. Let l be the dimension of $\mathscr{F}_0$. For any basis B of Λ, let P_B be the polyhedron given by $P_B = \{\upsilon \mid \Lambda_B^{-1} \left[\begin{smallmatrix} b \\ \upsilon \end{smallmatrix}\right] \geq 0\}$. Let $\mathcal{P}$ be the set $\{P_B \mid B \text{ is a lex-prc basis}\}$. Let $\bar{\mathcal{P}}$ and $\underline{\mathcal{P}}$ be the subsets of $\mathcal{P}$ consisting of those polyhedra whose dimension is equal to l and those less than l respectively. We claim that $\bar{\mathcal{P}}$ is a polyhedral subdivision of $\mathscr{F}_0$. By definition, each polyhedron $P \in \bar{\mathcal{P}}$ is l-dimensional. Also, since there is a finite number of bases, the set $\bar{\mathcal{P}}$ is finite. We show next that $\bar{\mathcal{P}}$ covers $\mathscr{F}_0$. Each element of $\bar{\mathcal{P}}$ is closed and this set has a finite number of elements. Thus, the union of the polyhedra in $\bar{\mathcal{P}}$ is a closed set and therefore $\bar{\mathcal{P}}$ must contain all of its limit points. Hence, it is sufficient to show that every element of $\mathscr{F}_0$ is a limit point of $\bar{\mathcal{P}}$, which is to say that for any set Υ containing υ that is open relative to $\mathscr{F}_0$, there exists some element υ' in Υ and some P in $\bar{\mathcal{P}}$ such that υ' is in P. Let $\tilde{\mathscr{F}}_0$ be the affine hull of $\mathscr{F}_0$. There is an invertible affine transformation ω between $\tilde{\mathscr{F}}_0$ and $\mathbb{R}^l$. The image $\omega(\Upsilon)$ must contain an open set of $\mathbb{R}^l$ and therefore must have non-zero Lebesgue measure. For any polyhedron P in $\underline{\mathcal{P}}$, the image $\omega(P)$ has dimension less than l, which implies that each of these images must have Lebesgue measure zero. Since $\underline{\mathcal{P}}$ is finite, the union of images $\bigcup_{P \in \underline{\mathcal{P}}} \omega(P)$ must also have Lebesgue measure zero. Thus, there exists some point ι in $\omega(\Upsilon)$ that is not in $\bigcup_{P \in \underline{\mathcal{P}}} \omega(P)$. Let υ' be equal to $\omega^{-1}(\iota)$. By Proposition EC.2, there exists a lex-prc basis B such that $\Lambda_B^{-1} \left[\begin{smallmatrix} b \\ \upsilon' \end{smallmatrix}\right] \geq 0$. Furthermore, polyhedron P_B has dimension l; if it had dimension less than l, then ι would be contained in $\bigcup_{P \in \underline{\mathcal{P}}} \omega(P)$. This completes the proof that $\bar{\mathcal{P}}$ covers $\mathscr{F}_0$.

Now, we show that $\ell(\cdot, u)$ is piecewise-linear on $\mathscr{F}_0$ with a finite number of pieces. Consider polyhedron P_B in $\bar{\mathcal{P}}$ corresponding to some lex-prc basis B. For any $u' \in P_B$, it follows from Proposition 1 that the lexmin solution $x^{\text{lex}}(u')$ is given by

$$x^{\text{lex}}(u') = \Pi(B)\Lambda_B^{-1} \left[\begin{smallmatrix} b \\ u' \end{smallmatrix}\right].$$

Then,

$$\ell(u', u) = c^{\intercal}\Pi(B)\Lambda_B^{-1} \left[\begin{smallmatrix} b \\ u' \end{smallmatrix}\right] + Q\left(\Pi(B)\Lambda_B^{-1} \left[\begin{smallmatrix} b \\ u' \end{smallmatrix}\right], u\right) - z^*(u).$$

It follows from well-known linear programming results (see, for example, the discussions in Sections 5.2 and 5.3 of Bertsimas and Tsitsiklis (1997)) that, when $Q(\cdot, u)$ is well-defined, it is continuous, convex, and piecewise-linear with a finite number of pieces. This proves that the restriction of $\ell(\cdot, u)$ on each P in $\bar{\mathcal{P}}$ is continuous, convex, and piecewise-linear with a finite number of pieces.

Finally, we show that $\ell(\cdot, u)$ is continuous. Select $\epsilon > 0$ and u' in $\mathscr{F}_0$ that will remain fixed in this paragraph. Let $\dot{\mathcal{P}}$ be the subset of $\bar{\mathcal{P}}$ consisting of those polyhedra P that contain u'. Since each polyhedron in $\bar{\mathcal{P}}$ is closed, then for each P' in the set difference $\bar{\mathcal{P}} \setminus \dot{\mathcal{P}}$ there exists some $\eta(P') > 0$ such that $\|u'' - u'\| > \eta(P')$ for any u'' in P'. Let

$$\eta^* = \begin{cases} \min\limits_{P' \in \bar{\mathcal{P}} \setminus \dot{\mathcal{P}}} \eta(P') & \text{if } \bar{\mathcal{P}} \setminus \dot{\mathcal{P}} \text{ is non-empty,} \\ 1 & \text{otherwise.} \end{cases}$$

Since $\bar{\mathcal{P}}$ is finite, η^* is guaranteed to be a positive number. Since, as we have already demonstrated, the restriction of $\ell(\cdot, u)$ to any polyhedron P of $\bar{\mathcal{P}}$ is continuous, then for any P in $\dot{\mathcal{P}}$ there exists some $\Delta(P) > 0$ such that $|\ell(u'', u) - \ell(u', u)| < \epsilon$ for any u'' in P where $\|u'' - u'\| \leq \Delta(P)$. Let $\Delta^* = \min_{P \in \dot{\mathcal{P}}} \Delta(P)$. The set $\dot{\mathcal{P}}$ is finite and must be non-empty since $\bar{\mathcal{P}}$ covers $\mathscr{F}_0$. Hence, Δ^* exists and is positive. Finally, let u'' be any vector in $\mathscr{F}_0$ such that $\|u'' - u'\| < \min\{\eta^*, \Delta^*\}$. Then, by definition of η^*, the vector u'' must be in some polyhedra P of the set $\dot{\mathcal{P}}$. By definition of Δ^*, $\|u'' - u'\| \leq \Delta(P)$, so $|\ell(u'', u) - \ell(u', u)| < \epsilon$. This proves that $\ell(\cdot, u)$ is continuous. □

Recall the statement of Theorem 2.

THEOREM 2. *Suppose that Assumptions 1, 2, 3, and 4 hold. Further suppose that $\mathscr{X}$ is a bounded polyhedron and that U has finite second moments. For any $\epsilon > 0$, there exists a predictor function $\psi_\epsilon : \mathbb{R}^\nu \to \mathbb{R}^{m_2}$ such that for any v in the support of the side information, $\mathbb{E}[f(x^{\text{lex}}(\psi_\epsilon(v)), U) \mid V = v] \leq \min_{x \in \mathscr{X}} \mathbb{E}[f(x, U) \mid V = v] + \epsilon$.* □

In order to prove Theorem 2, we first prove some properties about a related stochastic optimization problem without side information. Given some probability distribution $\mathbb{P}$ on $\mathbb{R}^{m_1}$, let $S^*(\mathbb{P})$ denote the set of optimal solutions to the optimization problem $\min_{x \in \mathscr{X}} \mathbb{E}_{U \sim \mathbb{P}}[f(x, U)]$. Let $\chi^{\text{lex}}(\mathbb{P})$ be the lexicographically minimal element of $S^*(\mathbb{P})$.

PROPOSITION EC.3. *Suppose that Assumptions 1, 2, 3 and 4 hold. If $\mathbb{P}$ is supported on a finite subset of $\mathscr{F}_0$, then $x^{\text{lex}}(T\chi^{\text{lex}}(\mathbb{P})) = \chi^{\text{lex}}(\mathbb{P})$.*

Proof of Proposition EC.3. Let $u^1, \ldots, u^k$ be the elements in the support of $\mathbb{P}$ with corresponding probabilities $p_1, \ldots, p_k$. In this case, the optimization problem $\min_{x \in \mathscr{X}} \mathbb{E}_{U \sim \mathbb{P}}[f(x, U)]$ can be written in extensive form as the linear program:

$$\min \quad c^\intercal x + \sum_{i=1}^{k} q^\intercal y^i p_i \tag{EC.11a}$$

$$s.t. \quad Ax = b \tag{EC.11b}$$

$$Tx + Wy^i = u^i, \qquad \forall i \in [k] \tag{EC.11c}$$

$$y^i \geq 0, \qquad \forall i \in [k] \tag{EC.11d}$$

$$x \geq 0. \tag{EC.11e}$$

Let LP^1 denote this LP. It follows from Assumptions 1 and 2 that LP^1 has an optimal solution. Let Γ denote the constraint matrix of LP^1. Assumptions 3 and 4 imply that the rows of Γ are linearly independent. Let B be a n_1-lex-prc basis for LP^1 that has a corresponding feasible basic solution $(\tilde{x}, \tilde{y}^1, \ldots, \tilde{y}^k)$; such a basis is guaranteed to exist by Proposition EC.2. Then, it follows from

Proposition EC.1 that $\chi^{\text{lex}}(\mathbb{P}) = \tilde{x}$. Consider now $x^{\text{lex}}(T\tilde{x})$. This is the lexicographically minimal solution to the optimization problem $\min_{x \in \mathscr{X}} f(x, T\tilde{x})$, which can equivalently be written as

$$\min \quad c^\intercal x + \sum_{i=1}^{k} q^\intercal y^i p_i \tag{EC.12a}$$

$$s.t. \quad Ax = b \tag{EC.12b}$$

$$Tx + Wy^i = T\tilde{x}, \qquad \forall i \in [k] \tag{EC.12c}$$

$$y^i \geq 0, \qquad \forall i \in [k] \tag{EC.12d}$$

$$x \geq 0. \tag{EC.12e}$$

Let LP^2 denote this LP. Since LP^2 has the same cost vector and constraint matrix as LP^1, and since the n_1-lex-prc property only depends on the constraint matrix and cost vector, then B is also a n_1-lex-prc basis for LP^2. Let $\bar{x}, \bar{y}^1, \ldots, \bar{y}^k$ be the basic solution to LP^2 corresponding to the basis B. If this solution vector can be shown to be non-negative, then Proposition EC.1 implies that $\bar{x} = x^{\text{lex}}(T\tilde{x})$. Let $B(0)$ denote the subset of B corresponding to the vector of variables x and let $B(i)$ denote the subset of B corresponding to the vector of variable y^i. Then, by definition, the basic solution $(\bar{x}, \bar{y}^1, \ldots, \bar{y}^k)$ is the unique solution to the system of equations:

$$A_{B(0)}\bar{x}_{B(0)} = b \tag{EC.13a}$$

$$T_{B(0)}\bar{x}_{B(0)} + W\bar{y}^i_{B(i)} = T\tilde{x}, \qquad \forall i \in [k]. \tag{EC.13b}$$

We now note that by definition, it must be true that $A_{B(0)}\tilde{x}_{B(0)} = b$. Furthermore, $T\tilde{x}_i = 0$ for all $i \notin B(0)$, so $T_{B(0)}\tilde{x}_{B(0)} = T\tilde{x}$. So, the vector given by $\bar{x} = \tilde{x}$ and $\bar{y}^i = 0$ for all $i \in [k]$ is the solution to the system (EC.13a) and (EC.13b). Thus, $x^{\text{lex}}(T\chi^{\text{lex}}(\mathbb{P})) = \chi^{\text{lex}}(\mathbb{P})$ $\square$

Given a subset $S \subseteq \mathbb{R}^{n_1}$ and an element $\dot{x} \in \mathbb{R}^{n_1}$, define the *distance* from $\dot{x}$ to S, denoted $\text{dist}(\dot{x}, S)$, as $\text{dist}(\dot{x}, S) = \inf_{x \in S} \|x - \dot{x}\|$ where $\|\cdot\|$ denotes the usual Euclidean norm.

Given two sets $S^1, S^2 \subseteq \mathbb{R}^{n_1}$, define the *deviation* from S^1 to S^2, denoted $\mathbb{D}(S^1, S^2)$, as $\mathbb{D}(S^1, S^2) = \sup_{x \in S^1} \text{dist}(x, S^2)$. Let $\mathbb{P}_v$ denote the probability distribution of the exogenous parameter given that the side information takes the value v, so that for any Borel set $S \subseteq \mathbb{R}^{m_2}$,

$$\mathbb{P}_v(S) = P(U \in S | V = v).$$

To simplify notation, let $\mathscr{Q}_v(x) = \mathbb{E}[f(x, U)|V = v]$.

We can now prove the desired result.

Proof of Theorem 2. We construct ψ_ϵ in a pointwise fashion. Consider $v \in \mathbb{R}^\nu$ in the support of the side information and $\epsilon > 0$. Since U has bounded second moments and the objective function f takes the assumed form $f(x, u) = c^\intercal x + \mathbb{E}[Q(x, u)]$, it follows from known results concerning

stochastic programs that $\mathscr{Q}_v$ is continuous; for example, see Theorem 6 of Birge and Louveaux (2011). Then, the assumption that $\mathscr{X}$ is bounded implies that $\mathscr{Q}_v$ is uniformly continuous on $\mathscr{X}$, so that there exists some $\delta > 0$ such that if x and x' are elements of $\mathscr{X}$ with $\|x - x'\| \leq \delta$, then $|\mathscr{Q}_v(x) - \mathscr{Q}_v(x')| \leq \epsilon$. It follows from known results that there exists some probability distribution $\widehat{\mathbb{P}}_v$ with finite support such that $\mathbb{D}\left(S^*\left(\widehat{\mathbb{P}}_v\right), S^*(\mathbb{P}_v)\right) \leq \delta$; for example, see Theorem 5.3 of Shapiro et al. (2014). Then, we claim that if we let

$$\psi_\epsilon(v) = T\chi^{\text{lex}}\left(\widehat{\mathbb{P}}_v\right)$$

the desired property is satisfied. From Proposition EC.3, we have that $x^{\text{lex}}(\psi_\epsilon(v)) = \chi^{\text{lex}}\left(\widehat{\mathbb{P}}_v\right)$. Furthermore, by definition, $\chi^{\text{lex}}\left(\widehat{\mathbb{P}}_v\right) \in S^*\left(\widehat{\mathbb{P}}_v\right)$. By assumption, the deviation $\mathbb{D}\left(S^*\left(\widehat{\mathbb{P}}_v\right), S^*(\mathbb{P}_v)\right) \leq \delta$. Thus, there must exist some solution $x^* \in S^*(\mathbb{P}_v)$ such that $\|x^{\text{lex}}(\psi_\epsilon(v)) - x^*\| \leq \delta$. Then, by definition of δ, it must be true that $\left|\mathscr{Q}_v(x^{\text{lex}}(\psi_\epsilon(v))) - \mathscr{Q}_v(x^*)\right| \leq \epsilon$. This provides the desired result.

EC.3. Proof of Theorem 3

Proof of Theorem 3. Suppose by contradiction that the algorithm does not terminate after a finite number of iterations. Under this assumption, we first make the following simple observations:

1. For any i, j with $j > i \geq 1$, $\theta_i^{\text{UB}} > \theta_j > \theta_i^{\text{LB}}$.
2. There exists $\theta^* \in [0,1]$ such that $\lim_{i\to\infty} \theta_i = \lim_{i\to\infty} \theta_i^{\text{UB}} = \lim_{i\to\infty} \theta_i^{\text{LB}} = \theta^*$.
3. Let $t^* = \lim_{i\to\infty} \dot{t}(i)$. Then t^* exists and $t^* = \theta^* u^\circ + (1-\theta^*)\dot{u}$.
4. For any $i \geq 1$, $g\left(\theta_i^{\text{UB}}\dot{u} + (1-\theta_i^{\text{UB}})u^\circ\right) \geq \theta_i^{\text{UB}} g(\dot{u}) + (1-\theta_i^{\text{UB}})g(u^\circ)$.
5. For any $i \geq 1$, $g\left(\theta_i^{\text{LB}}\dot{u} + (1-\theta_i^{\text{LB}})\, u^\circ\right) \leq \theta_i^{\text{LB}} g(\dot{u}) + (1-\theta_i^{\text{LB}})\, g(u^\circ)$.

Observations 4 and 5 hold trivially when $i = 1$ as they reduce to $g(\dot{u}) \geq g(\dot{u})$ and $g(u^\circ) \leq g(u^\circ)$. Further, one can establish that they continue to hold for $i > 1$ by induction, because of lines 14-18 of Algorithm 1. In fact, the "if" condition of these lines changes exactly one value of θ_{i+1}^{LB} and θ_{i+1}^{UB} (from their previous values θ_i^{LB} and θ_i^{UB}) so that the conditions are met, while the other remains unchanged, and hence verified by induction. It follows from these claims and the fact that g is continuous that

$$\lim_{i\to\infty} g(\dot{t}(i)) = g(t^*) = g(\theta^*\dot{u} + (1-\theta^*)u^\circ) = \theta^* g(\dot{u}) + (1-\theta^*)g(u^\circ). \tag{EC.14}$$

The set $Z = \{\theta\dot{u} + (1-\theta)u^\circ \mid \theta \in [0,1]\}$ is bounded. Let $\mathcal{P}$ be the polyhedral subdivision of $\mathbb{R}^n$ on which g is defined. By the definition of polyhedral subdivision, there exists $P \in \mathcal{P}$ and an increasing function $\eta : \mathbb{N} \mapsto \mathbb{N}$ such that $\dot{t}(\eta(i)) \in P$ for all $i \in \mathbb{N}$ and $\dot{d}(\eta(i)) \in \delta g_P(\dot{t}(\eta(i)))$, *i.e.*, there is a subsequence of the pairs of points and piece-subgradients $(\dot{t}(i), \dot{d}(i))$ such that all the points in the subsequence fall in P and each piece-subgradient is a subgradient of the piece function defined on P. Since P is closed, t^* must belong to P. We now consider four cases. In discussing them, we use $\dot{t}_i$, $\dot{\theta}_i$, and $\dot{d}_i$ as shorthand notations for $\dot{t}(\eta(i))$, $\theta_{\eta(i)}$, and $\dot{d}(\eta(i))$, respectively.

Case 1: There exists a positive integer i such that

$$g(\dot{t}_i) > \dot{\theta}_i g(\dot{u}) + (1-\dot{\theta}_i) g(u^\circ). \tag{EC.15}$$

Then, $\theta^* \le \theta^{\text{UB}}_{\eta(i)+1} = \dot{\theta}_i \le 1$. Define function $\pi : \mathbb{R} \mapsto \mathbb{R}$ as $\pi(\theta) = \dot{d}_i^\intercal(\theta\dot{u} + (1-\theta)u^\circ - \dot{t}_i) + g(\dot{t}_i)$ and function $\omega : \mathbb{R} \mapsto \mathbb{R}$ as $\omega(\theta) = \theta g(\dot{u}) + (1-\theta)g(u^\circ)$. Observe that

$$\pi(\dot{\theta}_i) = g(\dot{t}_i) > \dot{\theta}_i g(\dot{u}) + (1-\dot{\theta}_i) g(u^\circ) = \omega(\dot{\theta}_i),$$

where the inequality holds because of (EC.15). Observe also that

$$\pi(\theta^*) = \dot{d}_i^\intercal \left(\theta^*\dot{u} + (1-\theta^*)u^\circ - \dot{t}_i\right) + g(\dot{t}_i) = \dot{d}_i^\intercal \left(t^* - \dot{t}_i\right) + g\left(\dot{t}_i\right) \le g\left(t^*\right) = \omega(\theta^*),$$

where the inequality holds because $\dot{d}_i$ is a subgradient of g_P at $\dot{t}_i$ and the last equality because of (EC.14). As (i) $\theta^* \le \dot{\theta}_i$, (ii) $\pi(\theta^*) \le \omega(\theta^*)$, and (iii) $\pi(\dot{\theta}_i) > \omega(\dot{\theta}_i)$, there exists $\bar{\theta} \in [\theta^*, \dot{\theta}_i)$ such that $\pi(\theta) \le \omega(\theta)$ for $\theta \le \bar{\theta}$ and $\pi(\theta) \ge \omega(\theta)$ for $\theta \ge \bar{\theta}$ because π and ω are affine functions. Then,

$$\dot{d}_i^\intercal(\dot{u} - \dot{t}_i) + g(\dot{t}_i) = \pi(1) \ge \omega(1) = g(\dot{u}) > g(\dot{u}) - \epsilon$$

and

$$\dot{d}_i^\intercal(u^\circ - \dot{t}_i) + g(\dot{t}_i) = \pi(0) \le \omega(0) = g(u^\circ).$$

This establishes that element $(\dot{t}_i, \dot{d}_i)$ of $D(g)$ cuts $(\dot{u}, g(\dot{u}) - \epsilon)$ but does not cut $(u^\circ, g(u^\circ))$. This contradicts the assumption that the algorithm did not terminate at step $\eta(i)$.

Case 2: There exists a positive integer i that satisfies

$$g(\dot{t}_i) < \dot{\theta}_i g(\dot{u}) + (1-\dot{\theta}_i) g(u^\circ) \text{ and } g(\dot{t}_i) \le \dot{\theta}_i (g(\dot{u}) - \epsilon) + (1-\dot{\theta}_i) g(u^\circ). \tag{EC.16}$$

In this case, $1 \ge \theta^* \ge \theta^{\text{LB}}_{\eta(i)+1} = \dot{\theta}_i > 0$. Define functions $\pi_1, \omega : \mathbb{R} \mapsto \mathbb{R}$ where $\pi_1(\theta) = g(\theta\dot{u} + (1-\theta)u^\circ)$ and $\omega(\theta) = \theta(g(\dot{u}) - \epsilon) + (1-\theta)g(u^\circ)$. Now,

$$\begin{aligned}\lim_{j\to\infty} \pi_1(\dot{\theta}_j) - \omega(\dot{\theta}_j) &= \lim_{j\to\infty} g\left(\dot{t}(j)\right) - \lim_{j\to\infty} \omega(\dot{\theta}_j) = \theta^* g(\dot{u}) + (1-\theta^*)g(u^\circ) - \omega(\theta^*) \\ &= \theta^* g(\dot{u}) + (1-\theta^*)g(u^\circ) - \theta^*(g(\dot{u}) - \epsilon) - (1-\theta^*)g(u^\circ) = \theta^*\epsilon > 0.\end{aligned}$$

This implies that there must be $j > i$ such that $\pi_1(\dot{\theta}_j) > \omega(\dot{\theta}_j)$. Now, define function $\pi_2 : \mathbb{R} \mapsto \mathbb{R}$ as $\pi_2(\theta) = \dot{d}_j^\intercal(\theta\dot{u} + (1-\theta)u^\circ - \dot{t}_j) + g\left(\dot{t}_j\right)$. Since $j > i$, $\dot{\theta}_j > \theta^{LB}_{\eta(i)+1} = \dot{\theta}_i$. Now, since $\dot{d}_j$ is a subgradient of g_P at $\dot{\theta}_j$,

$$\pi_2(\dot{\theta}_i) = \dot{d}_j^\intercal(\dot{t}_i - \dot{t}_j) + g\left(\dot{t}_j\right) \le g\left(\dot{t}_i\right) \le \dot{\theta}_i(g(\dot{u}) - \epsilon) + (1-\dot{\theta}_i)g(u^\circ) = \omega(\dot{\theta}_i),$$

where the last inequality holds because of (EC.16). Further, $\pi_2(\dot{\theta}_j) = \pi_1(\dot{\theta}_j) > \omega(\dot{\theta}_j)$. As (i) $\dot{\theta}_j > \dot{\theta}_i$, (ii) $\pi_2(\dot{\theta}_j) > \omega(\dot{\theta}_j)$, and (iii) $\pi_2(\dot{\theta}_i) \leq \omega(\dot{\theta}_i)$, there exists $\bar{\theta} \in [\dot{\theta}_i, \dot{\theta}_j)$ such that $\pi_2(\theta) \leq \omega(\theta)$ for $\theta \leq \bar{\theta}$ and $\pi_2(\theta) > \omega(\theta)$ for $\theta > \bar{\theta}$ because π_2 and ω are affine functions. Then,

$$\dot{d}_j^{\intercal}(u^\circ - \dot{t}_j) + g\left(\dot{t}_j\right) = \pi_2(0) \leq \omega(0) = g(u^\circ)$$

and

$$\dot{d}_j^{\intercal}(\dot{u} - \dot{t}_j) + g\left(\dot{t}_j\right) = \pi_2(1) > \omega(1) = g(\dot{u}) - \epsilon.$$

This establishes that the element $(\dot{t}_j, \dot{d}_j)$ of $D(g)$ cuts $(\dot{u}, g(\dot{u}) - \epsilon)$ but does not cut $(u^\circ, g(u^\circ))$. This contradicts the assumption that the algorithm did not terminate at step $\eta(j)$.

Case 3: There exists a positive integer i that satisfies

$$\begin{aligned} &(a) \quad g(\dot{t}_i) < \dot{\theta}_i g(\dot{u}) + (1 - \dot{\theta}_i) g(u^\circ), \text{ and} \\ &(b) \quad g(\dot{t}_i) > \dot{\theta}_i (g(\dot{u}) - \epsilon) + (1 - \dot{\theta}_i) g(u^\circ). \end{aligned} \tag{EC.17}$$

In this case, $1 \geq \theta^* \geq \theta^{\text{LB}}_{\eta(i)+1} = \dot{\theta}_i > 0$. Define real number $v^* = \frac{g\left(\dot{t}_i\right) - (1-\dot{\theta}_i) g(u^\circ)}{\dot{\theta}_i}$. Next, define functions $\pi_1,\, \omega : \mathbb{R} \mapsto \mathbb{R}$ as $\pi_1(\theta) = g(\theta \dot{u} + (1-\theta) u^\circ)$ and $\omega(\theta) = \theta v^* + (1-\theta) g(u^\circ)$. Now,

$$\begin{aligned} \lim_{j \to \infty} \pi_1(\dot{\theta}_j) - \omega(\dot{\theta}_j) &= \lim_{j \to \infty} g\left(\dot{t}(j)\right) - \lim_{j \to \infty} \omega(\dot{\theta}_j) = \theta^* g(\dot{u}) + (1-\theta^*) g(u^\circ) - \omega(\theta^*) \\ &= \theta^* g(\dot{u}) + (1-\theta^*) g(u^\circ) - \theta^* v^* - (1-\theta^*) g(u^\circ) = \theta^* \left(g(\dot{u}) - v^*\right) \\ &= \theta^* \left(g(\dot{u}) - \frac{g\left(\dot{t}_i\right) - (1-\dot{\theta}_i) g(u^\circ)}{\dot{\theta}_i} \right) > 0, \end{aligned}$$

where the last inequality holds because of (EC.17a). This implies that there must be $j > i$ such that $\pi_1(\dot{\theta}_j) > \omega(\dot{\theta}_j)$. Now, define function $\pi_2 : \mathbb{R} \mapsto \mathbb{R}$ where $\pi_2(\theta) = \dot{d}_j^{\intercal}(\theta \dot{u} + (1-\theta) u^\circ - \dot{t}_j)) + g\left(\dot{t}_j\right)$. Clearly, $\pi_2(\dot{\theta}_j) = \pi_1(\dot{\theta}_j) > \omega(\dot{\theta}_j)$. Since $j > i$, $\dot{\theta}_j > \theta^{LB}_{\eta(i)+1} = \dot{\theta}_i$. Now, since $\dot{d}_j$ is a subgradient of g_P at $\dot{t}_j$,

$$\pi_2(\dot{\theta}_i) = \dot{d}_j^{\intercal}(\dot{t}_i - \dot{t}_j) + g\left(\dot{t}_j\right) \leq g\left(\dot{t}_i\right) = \omega(\dot{\theta}_i).$$

As (i) $\dot{\theta}_i < \dot{\theta}_j$, (ii) $\pi_2(\dot{\theta}_i) \leq \omega(\dot{\theta}_i)$, and (iii) $\pi_2(\dot{\theta}_j) > \omega(\dot{\theta}_j)$, there exists $\bar{\theta} \in [\dot{\theta}_i, \dot{\theta}_j)$ such that $\pi_2(\theta) \leq \omega(\theta)$ for $\theta \leq \bar{\theta}$ and $\pi_2(\theta) \geq \omega(\theta)$ for $\theta \geq \bar{\theta}$ because π_2 and ω are affine functions. Then,

$$\dot{d}_j^{\intercal}(u^\circ - \dot{t}_j) + g\left(\dot{t}_j\right) = \pi_2(0) \leq \omega(0) = g(u^\circ)$$

and

$$\begin{aligned} \dot{d}_j^{\intercal}(\dot{u} - \dot{t}_j) + g\left(\dot{t}_j\right) = \pi_2(1) > \omega(1) = v^* &= \frac{g\left(\dot{t}_i\right) - (1-\dot{\theta}_i) g(u^\circ)}{\dot{\theta}_i} \\ &> \frac{\dot{\theta}_i (g(\dot{u}) - \epsilon) + (1-\dot{\theta}_i) g(u^\circ) - (1-\dot{\theta}_i) g(u^\circ)}{\dot{\theta}_i} = g(\dot{u}) - \epsilon, \end{aligned}$$

where the last inequality holds because of (EC.17b). This establishes that the element $(\dot{t}_j, \dot{d}_j)$ of $D(g)$ cuts $(\dot{u}, g(\dot{u}) - \epsilon)$ but does not cut $(u^\circ, g(u^\circ))$. This contradicts the assumption that the algorithm did not terminate at step $\eta(j)$.

Case 4: For all i,

$$g(\dot{t}_i) = \dot{\theta}_i g(\dot{u}) + (1-\dot{\theta}_i) g(u^\circ). \tag{EC.18}$$

Note that $0 = \theta^* \leq \dot{\theta}_2 < \theta^{\text{UB}}_{\eta(1)+1} = \dot{\theta}_1 < 1$. Define function $\pi : \mathbb{R} \mapsto \mathbb{R}$ as $\pi(\theta) = \dot{d}_1^\intercal(\theta\dot{u} + (1-\theta)u^\circ - \dot{t}_1) + g(\dot{t}_1)$. Define function $\omega : \mathbb{R} \mapsto \mathbb{R}$ as $\omega(\theta) = \theta g(\dot{u}) + (1-\theta) g(u^\circ)$. Now,

$$\pi(\dot{\theta}_1) = g(\dot{t}_1) = \dot{\theta}_1 g(\dot{u}) + (1-\dot{\theta}_1) g(u^\circ) = \omega(\dot{\theta}_1),$$

where the second equality holds because of (EC.18) and

$$\pi(\theta^*) = \dot{d}_1^\intercal \left(\theta^*\dot{u} + (1-\theta^*)u^\circ - \dot{t}_1\right) + g(\dot{t}_1) = \dot{d}_1^\intercal \left(t^* - \dot{t}_1\right) + g\left(\dot{t}_1\right) \leq g\left(t^*\right) = \omega(\theta^*),$$

because $\dot{d}_1$ is a subgradient of g_P at $\dot{t}_1$.

As (i) $\theta^* < \dot{\theta}_1$, (ii) $\pi(\theta^*) \leq \omega(\theta^*)$, and (iii) $\pi(\dot{\theta}_1) = \omega(\dot{\theta}_1)$, there exists $\bar{\theta} \in [\theta^*, \dot{\theta}_1]$ such that $\pi(\theta) \leq \omega(\theta)$ for all $\theta \leq \bar{\theta}$ and $\pi(\theta) \geq \omega(\theta)$ for all $\theta \geq \bar{\theta}$ because π and ω are affine functions. Then,

$$\dot{d}_1^\intercal(\dot{u} - \dot{t}_1) + g(\dot{t}_1) = \pi(1) \geq \omega(1) = g(\dot{u}) > g(\dot{u}) - \epsilon$$

and

$$\dot{d}_1^\intercal(u^\circ - \dot{t}_1) + g(\dot{t}_1) = \pi(0) \leq \omega(0) = g(u^\circ).$$

This establishes that element $(\dot{t}_1, \dot{d}_1)$ of $D(g)$ cuts $(\dot{u}, g(\dot{u}) - \epsilon)$ but does not cut $(u^\circ, g(u^\circ))$. This contradicts the assumption that the algorithm did not terminate at step $\eta(1)$.