

A Missing proofs

A.1 Proof of Theorem 2.4

Our proof follows similar steps to those in [Elmachtoub and Grigas \(2022\)](#). Since the hypothesis class \mathcal{F} is unrestricted, we can optimize the function values $f(\mathbf{x})$ individually for each $\mathbf{x} \in \mathcal{X}$. Therefore, solving the problems

$$\begin{aligned} f_{\text{cost}}^* &\in \arg \min_f \mathbb{E}_{(\mathbf{x}, \mathbf{c}) \sim \mathcal{D}_{\mathbf{x}, \mathbf{c}}} [\text{cost}(f(\mathbf{x}), \mathbf{c}; \mathcal{U}(\mathbf{x}))], \\ f_{\text{cost}_+}^* &\in \arg \min_f \mathbb{E}_{(\mathbf{x}, \mathbf{c}) \sim \mathcal{D}_{\mathbf{x}, \mathbf{c}}} [\text{cost}_+(f(\mathbf{x}), \mathbf{c}; \mathcal{U}(\mathbf{x}))], \end{aligned}$$

is equivalent to optimizing each $f(\mathbf{x})$ separately. Consequently, for the remainder of the proof, we fix \mathbf{x} to \mathbf{x}_0 , and also $\hat{\mathcal{U}}_0 := \mathcal{U}(\mathbf{x}_0)$, and consider only the conditional distribution of \mathbf{c} . We define the risks associated with the cost and cost metrics as:

$$\begin{aligned} R_{\text{cost}}(\hat{\mathbf{c}}) &:= \mathbb{E}_{\mathbf{c}} [\text{cost}(\hat{\mathbf{c}}, \mathbf{c}; \hat{\mathcal{U}}_0)], \\ R_{\text{cost}_+}(\hat{\mathbf{c}}) &:= \mathbb{E}_{\mathbf{c}} [\text{cost}_+(\hat{\mathbf{c}}, \mathbf{c}; \hat{\mathcal{U}}_0)], \end{aligned} \tag{6}$$

where the $\mathbb{E}_{\mathbf{c}}$ denotes the expectation over \mathbf{c} . Let us define $\bar{\mathbf{c}} := \mathbb{E}_{\mathbf{c}}[\mathbf{c}|\mathbf{x}_0]$. We first list the propositions needed to complete the proof of Theorem 2.4.

Proposition A.1 (Proposition 5 of [Elmachtoub and Grigas \(2022\)](#)). *If a cost vector \mathbf{c}^* is a minimizer of $R_{\text{cost}}(\cdot)$, then $W^*(\mathbf{c}^*, \hat{\mathcal{U}}_0) \subseteq W^*(\bar{\mathbf{c}}, \hat{\mathcal{U}}_0)$. On the other hand, if \mathbf{c}^* is a cost vector such that $W^*(\mathbf{c}^*, \hat{\mathcal{U}}_0)$ is a singleton and $W^*(\mathbf{c}^*, \hat{\mathcal{U}}_0) \subseteq W^*(\bar{\mathbf{c}}, \hat{\mathcal{U}}_0)$, then \mathbf{c}^* is a minimizer of $R_{\text{cost}}(\cdot)$.*

Proposition A.2 (Proposition 6 of [Elmachtoub and Grigas \(2022\)](#)). *Under Assumption 2.3, $\bar{\mathbf{c}}$ is the unique minimizer of $R_{\text{cost}_+}(\cdot)$.*

Since we have fixed \mathbf{x} to \mathbf{x}_0 , the uncertainty set $\hat{\mathcal{U}}_0$ is also fixed. Therefore, Propositions [A.1](#) and [A.2](#) reduce to those presented in [Elmachtoub and Grigas \(2022\)](#), and their proofs follow accordingly. Importantly, these propositions hold true when the constructed uncertainty set satisfies Assumption 2.3, regardless of whether the true parameter \mathbf{a} lies within $\hat{\mathcal{U}}_0$ or not. This means we do not need to be concerned about the quality of $\hat{\mathcal{U}}_0$ to guarantee consistency when learning with the cost_+ metric or SPO-RC+ loss function. (Indeed, recall the MSE loss will also guarantee consistency.) However, since we do not know the true distribution \mathcal{D} and certain assumptions such as having a well-defined hypothesis class \mathcal{F} often do not hold, this motivates us to focus on the region where feasibility is guaranteed and subsequently cost_+ and SPO-RC+ yield valid upper bounds, as discussed in our main paper. We complete the proof of Theorem 2.4 using the above propositions.

Proof. Let $\mathbf{x}_0 \in \mathcal{X}$ be given and let $\hat{\mathcal{U}}_0 = \mathcal{U}(\mathbf{x}_0)$. By Proposition [A.2](#), the expected cost vector $\mathbb{E}[\mathbf{c}|\mathbf{x}_0]$ is the unique minimizer of $R_{\text{cost}_+}(\cdot)$. That is, $f_{\text{cost}_+}^*(\mathbf{x}_0)$ is unique and $f_{\text{cost}_+}^*(\mathbf{x}_0) = \mathbb{E}[\mathbf{c}|\mathbf{x}_0]$. Under Assumption 2.3, the optimal solution set $W^*(\mathbb{E}[\mathbf{c}|\mathbf{x}_0], \hat{\mathcal{U}}_0)$ is a singleton. Applying Proposition [A.1](#), we conclude that $\mathbb{E}[\mathbf{c}|\mathbf{x}_0]$ is a minimizer of $R_{\text{cost}}(\cdot)$. Since this holds for every $\mathbf{x} \in \mathcal{X}$, we have that almost surely $f_{\text{cost}_+}^*$ is unique, $f_{\text{cost}_+}^* = \mathbb{E}[\mathbf{c}|\mathbf{x}]$, and $f_{\text{cost}_+}^*$ also minimizes $\mathbb{E}_{(\mathbf{x}, \mathbf{c}) \sim \mathcal{D}_{\mathbf{x}, \mathbf{c}}} [\text{cost}(f(\mathbf{x}), \mathbf{c}; \mathcal{U}(\mathbf{x}))]$. This shows the Fisher consistency between the cost and cost_+ metrics. Moreover, because the f_{cost}^* and $f_{\text{cost}_+}^*$ remain the same when using SPO-RC and SPO-RC+ loss functions, respectively, this implies Fisher consistency between the SPO-RC and SPO-RC+ loss functions as well. \square

A.2 Proof of Lemma 3.3

Proof. The truncated distribution $\tilde{\mathcal{D}}$ satisfies $\mathbb{P}_{\tilde{\mathcal{D}}}(\mathbf{x}, \mathbf{a}, \mathbf{c}) \propto \mathbb{P}_{\mathcal{D}}(\mathbf{x}, \mathbf{a}, \mathbf{c}) \mathbb{1}(\mathbf{a} \in \mathcal{U}(\mathbf{x}))$, where $\mathbb{1}(\cdot)$ is the indicator function. Therefore, we have

$$\begin{aligned} \mathbb{P}_{\tilde{\mathcal{D}}}(\mathbf{x}, \mathbf{a}, \mathbf{c}) &\propto \mathbb{P}_{\mathcal{D}}(\mathbf{x}, \mathbf{a}, \mathbf{c}) \mathbb{1}(\mathbf{a} \in \mathcal{U}(\mathbf{x})) \\ &= \mathbb{P}_{\mathcal{D}}(\mathbf{c}|\mathbf{x}, \mathbf{a}) \mathbb{P}_{\mathcal{D}}(\mathbf{x}, \mathbf{a}) \mathbb{1}(\mathbf{a} \in \mathcal{U}(\mathbf{x})) \\ &= \mathbb{P}_{\mathcal{D}}(\mathbf{c}|\mathbf{x}, \mathbf{a}) \mathbb{P}_{\tilde{\mathcal{D}}}(\mathbf{x}, \mathbf{a}) \\ &= \mathbb{P}_{\mathcal{D}}(\mathbf{c}|\mathbf{x}) \mathbb{P}_{\tilde{\mathcal{D}}}(\mathbf{x}, \mathbf{a}) \text{ (by Assumption 3.2)}. \end{aligned}$$

815 To find the marginal distribution $\mathbb{P}_{\tilde{\mathcal{D}}}(\mathbf{x}, \mathbf{c})$, we sum over all possible \mathbf{a} :

$$\begin{aligned}\mathbb{P}_{\tilde{\mathcal{D}}}(\mathbf{x}, \mathbf{c}) &= \mathbb{P}_{\tilde{\mathcal{D}}}(\mathbf{c}|\mathbf{x})\mathbb{P}_{\tilde{\mathcal{D}}}(\mathbf{x}) \\ &\propto \mathbb{P}_{\mathcal{D}}(\mathbf{c}|\mathbf{x})\mathbb{P}_{\tilde{\mathcal{D}}}(\mathbf{x}).\end{aligned}$$

816 Dividing both sides by $\mathbb{P}_{\tilde{\mathcal{D}}}(\mathbf{x})$ gives $\mathbb{P}_{\tilde{\mathcal{D}}}(\mathbf{c}|\mathbf{x}) = \mathbb{P}_{\mathcal{D}}(\mathbf{c}|\mathbf{x})$.

817

□

818 B Extended theoretical results

819 B.1 Generalization bound

820 In this section, we present additional theoretical results, specifically generalization bounds for the cost
821 metric. This extends the generalization bounds presented in [El Balghiti et al. \(2023\)](#) to accommodate
822 context-dependent feasibility sets. We define the population risk of a function f with respect to the
823 cost metric as

$$\mathcal{R}_{\mathcal{D}}(f) := \mathbb{E}_{(\mathbf{x}, \mathbf{c}) \sim \mathcal{D}_{\mathbf{x}, \mathbf{c}}} [\text{cost}(f(\mathbf{x}), \mathbf{c}; \mathcal{U}(\mathbf{x}))],$$

824 and denote its empirical risk over n samples collected in a dataset \mathcal{D}^n as

$$\hat{\mathcal{R}}_{\mathcal{D}}^n(f) := \frac{1}{n} \sum_{i=1}^n \text{cost}(f(\mathbf{x}_i), \mathbf{c}_i; \mathcal{U}(\mathbf{x}_i)).$$

825 The multivariate Rademacher complexity $\mathfrak{R}_{\text{cost}}^n(\mathcal{F})$ ([Bertsimas and Kallus \(2020\)](#)) of the hypothesis
826 class \mathcal{F} with respect to the cost metric is defined as:

$$\mathfrak{R}_{\text{cost}}^n(\mathcal{F}) := \mathbb{E}_{\mathbf{x}, \mathbf{c}} \mathbb{E}_{\sigma} \left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \sigma_i \text{cost}(f(\mathbf{x}_i), \mathbf{c}_i; \mathcal{U}(\mathbf{x}_i)) \right],$$

827 where σ_i are i.i.d Rademacher random variables for $i = 1, \dots, n$. We denote the quantity $\Omega_S(\mathcal{C})$ as
828 an upper bound on the maximum possible objective value over all solutions in S across the entire
829 space \mathcal{C} . Specifically, it is given by:

$$\Omega_S(\mathcal{C}) := \sup_{\mathbf{c} \in \mathcal{C}} \left(\max_{\mathbf{w} \in S} \mathbf{c}^\top \mathbf{w} \right).$$

830 By applying the cost metric to the renowned result from [Bartlett and Mendelson \(2002\)](#), we obtain
831 the following theorem.

832 **Theorem B.1** (Theorem from [Bartlett and Mendelson \(2002\)](#)). *Let \mathcal{F} be a hypothesis class and let
833 $\delta > 0$. The following inequality holds with probability at least $1 - \delta$ over an i.i.d. dataset \mathcal{D}^n for all
834 $f \in \mathcal{F}$:*

$$\mathcal{R}_{\mathcal{D}}(f) \leq \hat{\mathcal{R}}_{\mathcal{D}}^n(f) + 2\mathfrak{R}_{\text{cost}}^n(\mathcal{F}) + \Omega_S(\mathcal{C}) \sqrt{\frac{\log(1/\delta)}{2n}}. \quad (7)$$

835 In particular, when the hypothesis class \mathcal{F} consist of linear functions, we can further bound the
836 Rademacher complexity $\mathfrak{R}_{\text{cost}}^n(\mathcal{F})$ in terms of the sample size n and other relevant quantities. We
837 define $\mathbb{S} := \{\mathcal{S}(\mathbf{x}) : \mathbf{x} \in \mathcal{X}\}$ as the collection of all possible feasible sets and introduce the upper
838 bound on their radius $\rho(\mathbb{S}) := \max_{S \in \mathbb{S}} \max_{\mathbf{w} \in S} \|\mathbf{w}\|_2$ to characterize the size of these sets.

839 **Proposition B.2** (Corollary 3 of [El Balghiti et al. \(2023\)](#)). *If $\mathcal{F}_{\text{lin}} := \{\mathbf{x} \rightarrow \mathbf{B}\mathbf{x} | \mathbf{B} \in \mathbb{R}^{d \times p}\}$ is the
840 linear hypothesis class, then we have*

$$\mathfrak{R}_{\text{cost}}^n(\mathcal{F}_{\text{lin}}) \leq 2d\Omega_S(\mathcal{C}) \sqrt{\frac{2p \log(2n\rho(\mathbb{S})d)}{n}} + O\left(\frac{1}{n}\right).$$

841 Notice that the extension can be easily made by adjusting the definition of $\rho(\mathbb{S})$, which characterizes
842 the size of the feasible sets. By incorporating the Rademacher complexity bound from Proposition
843 [B.2](#) into [\(7\)](#), we obtain a generalization bound for the linear hypothesis class in our framework.

844 In addition to ensuring consistency when truncating, importance reweighting allows us to extend
845 the generalization bounds presented in Theorem [B.1](#). The following lemma provides a bound on
846 the difference between the empirical risks calculated under the true distribution and the importance-
847 reweighted truncated distribution.

Table 2: Comparison of NormSPORCTest and optimal decision boundary across different datasets

Data	R_A	R_B	R_C	Opt Boundary
\mathcal{D}_O	8.4%	12%	11%	$x = \bar{x}_2$
\mathcal{D}_T	12.2%	9.2%	15.8%	Not Intersect
\mathcal{D}_{IR}	8.4%	12.2%	11.2%	$x = \bar{x}_2$

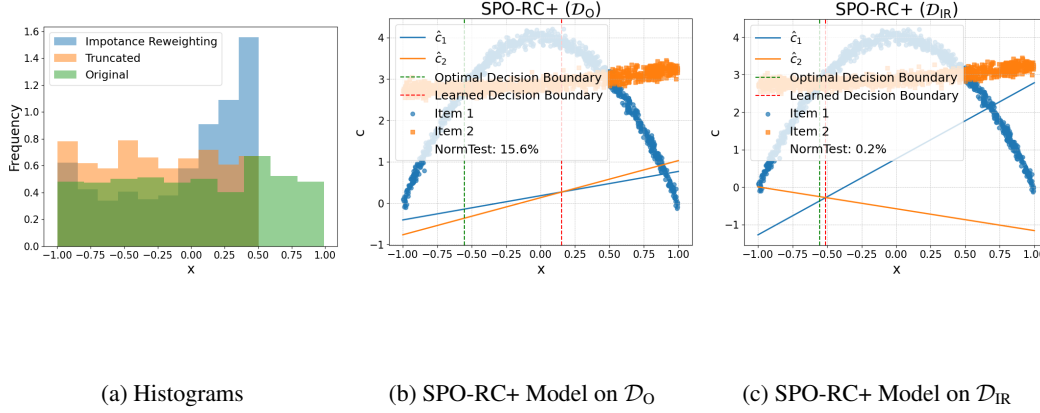


Figure 4: Visualization of the truncation toy example

848 **Lemma B.3** (Lemma 4 from [Huang et al. \(2006\)](#)). Suppose we have knowledge of $\beta(\mathbf{x}) \in [0, B]$
849 and let $\delta > 0$. With probability at least $1 - \delta$ over n i.i.d. samples from the true distribution \mathcal{D} , and
850 their corresponding truncations drawn from the truncated distribution $\tilde{\mathcal{D}}$, we have for all $f \in \mathcal{F}$:

$$|\hat{\mathcal{R}}_{\mathcal{D}}^n(f) - \hat{\mathcal{R}}_{\beta\tilde{\mathcal{D}}}^n(f)| \leq (1 + \sqrt{2\log(2/\delta)})\Omega_S(\mathcal{C})\sqrt{\frac{B^2 + 1}{n}},$$

851 where $\beta\tilde{\mathcal{D}}$ represents the truncated distribution adjusted for the importance weight β . Using Lemma
852 [B.3](#), we can replace $\hat{\mathcal{R}}_{\mathcal{D}}^n(f)$ in the generalization bound [\(7\)](#) with $\hat{\mathcal{R}}_{\beta\tilde{\mathcal{D}}}^n(f)$, thus ensuring that our
853 generalization analysis remains valid when using importance-reweighted truncated data.

854 **Proposition B.4.** Suppose we have knowledge of $\beta(\mathbf{x}) \in [0, B]$, let \mathcal{F} be a hypothesis class, and let
855 $\delta > 0$. The following inequality holds with probability at least $1 - \delta$ over an i.i.d. dataset \mathcal{D}^n , and
856 its corresponding truncated dataset $\tilde{\mathcal{D}}^n$, for all $f \in \mathcal{F}$:

$$\mathcal{R}_{\mathcal{D}}(f) \leq \hat{\mathcal{R}}_{\beta\tilde{\mathcal{D}}}^n(f) + 2\mathfrak{R}_{\text{cost}}^n(\mathcal{F}) + \Omega_S(\mathcal{C}) \left(\sqrt{\frac{\log(1/\delta)}{2n}} + (1 + \sqrt{2\log(2/\delta)})\sqrt{\frac{B^2 + 1}{n}} \right).$$

857 C Additional experimental details and results

858 In this section, we provide additional details on the numerical experiments presented in Section [4](#) and
859 some additional results. The experiments were conducted on a MacBook Pro equipped with an Intel
860 chip and 16 GB of RAM.

861 C.1 Additional details and results on toy examples

862 This section provides a more detailed explanation and additional toy examples supplementing the
863 toy example presented in Section [4](#). To illustrate the effectiveness of importance reweighting and
864 truncation, we consider a simplified version of the fractional knapsack problem. Specifically, we
865 consider a case where our goal is to predict which of two items has a higher value based on a
866 uniformly distributed context $x \in [-1, 1]$. The true relationships between the item values c_1 and c_2
867 and the context x are given by:

$$c_1 = -4x^2 + 4 + \zeta_1, \quad c_2 = \frac{1}{8}(x+1)^2 + 2.75 + \zeta_2,$$

where ζ_1 and ζ_2 are i.i.d. normal random variables with mean 0 and standard deviation 0.1. We generate 1,000 samples of (x, c_1, c_2) . Figure 1a shows the scatter plot of these samples along with their true conditional expectations $\mathbb{E}[c_1|x]$ and $\mathbb{E}[c_2|x]$. The curves intersect at two points, denoted \bar{x}_1 and \bar{x}_2 , partitioning the interval $[-1, 1]$ into three distinct regions: $A = [-1, \bar{x}_1]$, $B = [\bar{x}_1, \bar{x}_2]$, and $C = [\bar{x}_2, 1]$. Our goal is to use linear regression to predict c_1 and c_2 and make a decision based on which predicted value is higher. This is a special case of the fractional knapsack problem where the uncertain weight constraint is replaced with $w_1 + w_2 = 1$ without any uncertainty.

C.1.1 Importance reweighting example

In regions A and C , we observe that $c_2 < c_1$, whereas in region B , $c_1 < c_2$. Since the optimal decision changes across these regions and linear models may not capture the non-linear relationships perfectly, the decision made by the models will be incorrect in at least one of these regions. For instance, Figure 1c shows the learned decision boundary (red vertical line) of a linear model, which incorrectly predicts item 1 in region A , even though the true optimal choice is item 2. Notably, the optimal decision boundary (green vertical line in Figure 1c) makes incorrect decisions only in region A , where the difference between the two true curves is minimal. In fact, the area between the curves can be quantified using the NormSPORCTest metric on samples from each region, denoted as R_A , R_B , and R_C , corresponding to regions A , B , and C , respectively.

Suppose we randomly remove 30% of the data where $|x| < 0.5$, reducing samples in region B . This shifts the smallest region from A to B (as shown in Table 2) thereby changing the optimal decision boundary. However, by applying importance reweighting, we correct for this and realign the decision boundary with the true distribution, as shown in Table 2. Figure 1b and 1c illustrate the learned SPO-RC+ models on \mathcal{D}_T and \mathcal{D}_{IR} . Notice that the learned decision boundary of SPO-RC+ on \mathcal{D}_{IR} , the model with importance reweighting, is much closer to the true one than the model without it (SPO-RC+ on \mathcal{D}_T). Additionally, the NormSPORCTest measured with 500 test samples was lower in SPO-RC+ on \mathcal{D}_{IR} , indicating better performance.

C.1.2 Truncation example

In this example, we compare SPO-RC+ on \mathcal{D}_O with SPO-RC+ on \mathcal{D}_{IR} , to show the effectiveness of truncation. Building on the setting of the previous toy example, we introduce an additional capacity constraint $w_2 \leq a_2$, where a_2 equals 100 if $x < 0.8$ and a_2 equals 0 otherwise. This makes the decision $(w_1, w_2) = (0, 1)$ infeasible when $x > 0.8$. We use a linear regression model to predict a_2 and construct the uncertainty set $\hat{\mathcal{U}}$ using conformal prediction with $\alpha = 0.25$. In the region where $x < 0.8$, we set a_2 large so that the constructed uncertainty set $\hat{\mathcal{U}}$ is large enough to include the original feasibility set defined by $w_1 + w_2 = 1$, for all $x \in [-1, 1]$. Thus, this example is equivalent to the previous one, except that the induced optimal solution can now be infeasible when $x > 0.8$.

Figure 4a shows the histograms of each dataset, indicating that all samples in region C (including the possible infeasible range $[0.8, 1]$) are truncated. This means that for x in the region C , $a_2 \notin \hat{\mathcal{U}}$. Figures 4b and 4c show the learned linear models from SPO-RC+ on \mathcal{D}_O and \mathcal{D}_{IR} , respectively. The decision boundary learned from \mathcal{D}_{IR} closely matches the true optimal boundary, whereas the model trained on \mathcal{D}_O produces a less accurate solution. Evaluated on a dataset sampled from the feasibility-guaranteed region ($x < 0.5$), SPO-RC+ on \mathcal{D}_O performs poorly, achieving a NormSPORCTest of 16.7%. In contrast, SPO-RC+ on \mathcal{D}_{IR} performs almost perfectly, with a NormSPORCTest value of 0.2%. Furthermore, in the region $x > 0.8$, the induced solution of SPO-RC+ on \mathcal{D}_O is $(w_1^*, w_2^*) = (1, 0)$, which is infeasible. These results highlight the effectiveness of truncation when model complexity is limited and suggest focusing our learning capacity on the feasibility-guaranteed region.

C.2 Additional details on the fractional knapsack instances

Data generation process: We consider a fractional knapsack problem with $d = 5$ items and $p = 10$ features. We adapt a popular data generation process in the literature, starting with Elmachoub and Grigas (2022). We sample each element of the context vector $\mathbf{x} \in \mathbb{R}^p$ from the uniform distribution Uniform $(-1, 1)$, and generate the cost vector \mathbf{c} as:

$$c_{ij} = \frac{5}{3.5^{\deg_c}} \left[\left(\frac{(\mathbf{B}_c \mathbf{x}_i)_j}{\sqrt{p}} + 3 \right)^{\deg_c} + 10 \right] + \epsilon_{ij}^c,$$

where $i = 1, \dots, n$ indexes the instances, $j = 1, \dots, d$ indexes the items, and $\epsilon_{ij}^c \sim \mathcal{N}(0, 1)$. The matrix $\mathbf{B}_c \in [0, 1]^{d \times p}$ has elements sampled from a binomial distribution with probability 0.5. The parameter deg_c is a fixed integer that determines the complexity of the cost function. Similarly, the weight vector \mathbf{a} is generated as:

$$a_{ij} = \frac{5}{3.5^{\text{deg}_a}} \left(\frac{(\mathbf{B}_a \mathbf{x}_i)_j}{\sqrt{p}} + 3 \right)^{\text{deg}_a} + \frac{(p - \|\mathbf{x}_i\|_1) \epsilon_{ij}^a}{p},$$

where \mathbf{B}_a and ϵ_{ij}^a are generated in the same manner as \mathbf{B}_c and ϵ_{ij}^c , respectively. Notice that the noise in a_{ij} increases with $\|\mathbf{x}_i\|_1$. We fix $\text{deg}_a = 4$ for the weight function, while varying $\text{deg}_c = 2, 4, 6, 8$ to increase the complexity of the cost function, with $b = 20$.

Model description for ℓ_2 -norm conformal prediction score: When the ℓ_2 -norm score is used for conformal prediction, the resulting robust reformulation takes the form of an SOCP:

$$\max_{\mathbf{w} \in \mathbb{R}^d} \mathbf{c}^\top \mathbf{w} \quad \text{s.t.} \quad Q^{1-\alpha} \|\mathbf{w}\|_2 \leq b - \hat{\mathbf{a}}^\top \mathbf{w}, \quad \mathbb{1}_d^\top \mathbf{w} = 1, \quad \mathbf{w} \in [0, 1]^d.$$

To predict the weight vector $\hat{\mathbf{a}}$, we use a one-layer neural network with ReLU activation functions. We set $\alpha = 0.2$ to compute the conformal prediction quantile. For the cost vector $\hat{\mathbf{c}}$, we evaluate three models: Linear Regression (MSE), Random Forest (RF), and SPO-RC+ using a linear model. All models are trained using the Adam optimizer. The learning rate is set to 1×10^{-3} for the MSE and RF models, and 4×10^{-3} for the SPO-RC+ model. Training is conducted for 50 epochs with early stopping based on validation loss to prevent overfitting.

Model description for ℓ_1 -norm conformal prediction score: When the ℓ_1 -norm score is used for conformal prediction, the resulting robust reformulation can be formulated as an LP:

$$\max_{\mathbf{w} \in \mathbb{R}^d} \mathbf{c}^\top \mathbf{w} \quad \text{s.t.} \quad Q^{1-\alpha} \|\mathbf{w}\|_\infty \leq b - \hat{\mathbf{a}}^\top \mathbf{w}, \quad \mathbb{1}_d^\top \mathbf{w} = 1, \quad \mathbf{w} \in [0, 1]^d.$$

To predict the weight vector $\hat{\mathbf{a}}$, we use a one-layer neural network with ReLU activation functions, and use $\alpha = 0.2$ to compute the conformal prediction quantile. For the cost vector $\hat{\mathbf{c}}$, we evaluate both Linear Regression (MSE) and SPO-RC+ using a linear model. For the SPO-RC+ model, we warm start our method with MSE for faster training. All models are trained using the Adam optimizer. The learning rate is set to 1×10^{-2} for the MSE and RF models, and 5×10^{-2} for the SPO-RC+ model. Training is conducted for 50 epochs with early stopping based on validation loss to prevent overfitting.

C.3 Additional details on the alloy production instances

Data generation process: For the brass production scenario, we sample each element of the context vector and the cost vectors \mathbf{c} using the same strategy used for the fractional knapsack problem. For each ore, we generate the concentration according to a Gamma distribution

$$G_i = \max \{ \text{Gamma}(P_i) + \epsilon_k, 0 \} \quad \text{for each supplier } i = 1, \dots, d$$

where $P \in \mathbb{R}^{m \times d}$ is a base preference matrix independently sampled using a uniform distribution $[0, 1]$. Since brass production requires around 30% of Zinc and 70% Copper to produce the alloy, we set the requirements to $\mathbf{h} = [2.9, 7.1]$.

Model description: When the ℓ_2 -norm score is used for conformal prediction, the resulting robust reformulation can be formulated as an SOCP:

$$\min_{\mathbf{w} \in \mathbb{R}^d} \mathbf{c}^\top \mathbf{w} \quad \text{s.t.} \quad Q^{1-\alpha} \|\mathbf{w}\|_2 \leq \hat{\mathbf{a}}_j^\top \mathbf{w} - h_j \quad \forall j = 1, \dots, m, \quad \mathbf{w} \geq 0,$$

To predict the weight vector $\hat{\mathbf{a}}$, we use a one-layer neural network with ReLU activation functions. The conformal prediction quantile is computed using $\alpha = 0.2$. For the cost vector $\hat{\mathbf{c}}$, we evaluate two models: Linear Regression (MSE), and SPO-RC+, again using a linear model class and with the MSE solution as a warm start. All models are trained using the Adam optimizer. The learning rate is set to 1×10^{-2} for MSE and 4×10^{-2} for SPO-RC+. Training is run for 50 epochs with an adaptive learning rate scheduler and early stopping based on validation loss to prevent overfitting.