Supplementary Materials for “Smart Surrogate Losses for Contextual Stochastic Linear Optimization with Robust Constraints”

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

$$f_{\text{cost}} = \arg\min_{f} \mathbb{E}_{(\mathbf{x},\mathbf{c})\sim\mathcal{D}_{\mathbf{x},\mathbf{c}}} \left[\text{cost}\left(f(\mathbf{x}), \mathbf{c}; \mathcal{U}(\mathbf{x})\right)\right],$$
$$f_{\text{cost}_+} = \arg\min_{f} \mathbb{E}_{(\mathbf{x},\mathbf{c})\sim\mathcal{D}_{\mathbf{x},\mathbf{c}}} \left[\text{cost}_+\left(f(\mathbf{x}), \mathbf{c}; \mathcal{U}(\mathbf{x})\right)\right],$$

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:

$$\begin{aligned} R_{\text{cost}}(\hat{\mathbf{c}}) &:= \mathbb{E}_{\mathbf{c}}\left[\text{cost}\left(\hat{\mathbf{c}}, \mathbf{c}; \hat{\mathcal{U}}_0\right)\right], \\ R_{\text{cost}_+}(\hat{\mathbf{c}}) &:= \mathbb{E}_{\mathbf{c}}\left[\text{cost}_+\left(\hat{\mathbf{c}}, \mathbf{c}; \hat{\mathcal{U}}_0\right)\right], \end{aligned} \tag{1}$$

where the $\mathbb{E}_{\mathbf{c}}$ denotes the expectation over $\mathbf{c}$. Let us define $\bar{\mathbf{c}} := \mathbb{E}_{\mathbf{c}}\left[\mathbf{c}|\mathbf{x}_0\right]$. 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* $R_{cost}(\cdot)$*, 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* $R_{cost}(\cdot)$*.*

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

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 $\hat{\mathcal{U}}_0$ when learning with cost metric or SPO-RC+ loss function. 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, as shown 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_{cost_+}. Therefore $f^*_{\text{cost}_+}(\mathbf{x}_0) = \mathbb{E}[\mathbf{c}|\mathbf{x}_0]$. 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 the minimizer of R_{cost}, which implies $f^*_{\text{cost}}(\mathbf{x}_0) = \mathbb{E}[\mathbf{c}|\mathbf{x}_0]$. Since this holds for every $\mathbf{x} \in \mathcal{X}$ and we have

$$f^*_{\text{cost}} = f^*_{\text{cost}_+} = \mathbb{E}[\mathbf{c}|\mathbf{x}].$$

This equality 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. □

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

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

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

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

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

□

B Extended Theoretical Results

B.1 Generalization Bound

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

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

and denote its empirical risk over n samples as

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

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

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

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

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

By applying the cost metric to the renowned result from Bartlett and Mendelson (2002), we obtain the following theorem.

Theorem B.1 (Theorem from (Bartlett and Mendelson, 2002))**.** *Let $\mathcal{F}$ be a hypothesis class and let $\delta > 0$. Given a dataset $\mathcal{D}^n$, the following holds for all $f \in \mathcal{F}$ with probability with at least $1 - \delta$:*

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

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

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

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

Notice that the extension can be easily made by adjusting the definition of $\rho(\mathbb{S})$, which characterizes the size of the feasible sets. By incorporating the Rademacher complexity bound from Proposition B.2 into (3), we obtain a generalization bound for the linear hypothesis class in our framework.

In addition to ensuring Fisher consistency, importance reweighting allows us to extend the generalization bounds presented in Theorem B.1. The following lemma provides a bound on the difference between the empirical risks calculated under the true distribution and the importance-reweighted truncated distribution.

Lemma B.3 (Lemma 4 from Huang et al. (2006))**.** *Assuming perfect knowledge of $\beta(\mathbf{x}) \in [0, B]$ and given n samples from both the trunacted distribution $\tilde{\mathcal{D}}$ and the true distribution $\mathcal{D}$, we have for all $f \in \mathcal{F}$:*

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

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

Proposition B.4. *Let $\mathcal{F}$ be a hypothesis class and let $\delta > 0$. Given datasets $\mathcal{D}^n$ and (subsequently) $\tilde{\mathcal{D}}^n$, the following holds for all $f \in \mathcal{F}$ with probability with at least $1-\delta$:*

$$\mathcal{R}_{\mathcal{D}}(f) \le \hat{\mathcal{R}}^n_{\beta\tilde{\mathcal{D}}}(f) + 2\mathfrak{R}^n_{\text{cost}}(\mathcal{F}) + \Omega_S(\mathcal{C})\left(\sqrt{\frac{\log(1/\delta)}{2n}} + (1+\sqrt{2\log(2/\delta)})\sqrt{\frac{B^2+1}{n}}\right). \tag{3}$$

Remark B.5. *Break up LHS to $1-\alpha$ and α.*

References

Bartlett, P. L. and Mendelson, S. (2002). Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482.

Bertsimas, D. and Kallus, N. (2020). From predictive to prescriptive analytics. *Management Science*, 66(3):1025–1044.

El Balghiti, O., Elmachtoub, A. N., Grigas, P., and Tewari, A. (2023). Generalization bounds in the predict-then-optimize framework. *Mathematics of Operations Research*, 48(4):2043–2065.

Elmachtoub, A. N. and Grigas, P. (2022). Smart "predict, then optimize". *Management Science*, 68(1):9–26.

Huang, J., Gretton, A., Borgwardt, K., Schölkopf, B., and Smola, A. (2006). Correcting sample selection bias by unlabeled data. *Advances in neural information processing systems*, 19.