Contextual Optimization Under Model Misspecification: A Tractable and Generalizable Approach

Omar Bennouna
MIT
omarben@mit.edu

Jiawei Zhang
University of Wisconsin–Madison
jzhang2924@wisc.edu

Saurabh Amin
MIT
amins@mit.edu

Asuman Ozdaglar
MIT
asuman@mit.edu

Abstract

Contextual optimization problems arise in decision-making applications where historical data and contextual features are used to learn predictive models that guide optimal decisions. Practical applications often face model misspecification from incomplete knowledge of the data-generating process, leading to suboptimal decisions. Existing methods mainly address well-specified models, leaving a gap in the literature in handling misspecification. We propose a consistent, tractable and generalizable Integrated Learning and Optimization (ILO) framework that successfully addresses this gap in the literature.

1 Introduction

Many real-world decision-making problems involve optimizing decisions under uncertain parameters estimated from contextual information. A canonical example is traffic routing, where edge traversal costs—initially unknown—depend on observable features like weather, time of day, and construction activity. Using historical data, contextual optimization aims to learn a policy mapping observed contexts to near-optimal decisions Sadana et al. (2024).

A standard framework for contextual optimization assumes that the unknown cost vector $c \in \mathbb{R}^d$ is correlated with an observed context $x \in \mathbb{R}^k$, and the objective is to minimize expected decision cost $\mathbb{E}_{(x,c)\sim P}\left(c^\top \pi(x)\right)$ where P is the joint distribution of (x, c), and π a policy that maps contexts to decisions in a bounded convex set $W \subset \mathbb{R}^d$. Since the true cost function is unknown in practice, it must be approximated using a prediction model from a hypothesis class $\mathcal{H}$.

The widely used predict-then-optimize approach first learns a predictor $\hat{c}(x)$ from $\mathcal{H}$ and then selects a decision by solving:

$$\hat{\pi}(x) \in \arg\min_{w \in W} \hat{c}(x)^\top w. \tag{1}$$

Prior works mainly use two predict-then-optimize approaches: *Sequential Learning and Optimization (SLO)* Bertsimas and Kallus (2020), Hu et al. (2022) and *Integrated Learning and Optimization (ILO)* Donti et al. (2017), Elmachtoub and Grigas (2022), Sun et al. (2023). SLO minimizes prediction error for c without accounting for the downstream optimization task, while ILO prioritizes decision performance over cost prediction accuracy. Recent ILO research refines cost prediction to improve decisions, assuming $\mathcal{H}$ contains the true cost function, i.e., is well-specified. In practice, $\mathcal{H}$ is often misspecified, lacking a predictor that perfectly captures the true cost due to incomplete features, unmodeled dependencies, or distribution shifts. Despite its practical importance, contextual optimization under misspecification is underexplored. Most ILO work assumes well-specified $\mathcal{H}$ and focuses on surrogate losses and algorithms to minimize decision error. Some recent empirical works Donti et al. (2017), McKenzie et al. (2023), Kotary et al. (2023), Huang and Gupta (2024) study decision-aware learning under misspecification but lack theoretical guarantees for global optimality and generalization. A more detailed literature review is in Appendix A.3.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

We argue that the key challenge lies in distinguishing between two forms of misspecification (see Appendix A.7 for further details): prediction misspecification, i.e. The inability of a model $\hat{c}(x)$ to accurately estimate $c(x)$ and decision misspecification, i.e. the inability of a model $\hat{c}(x)$ to produce the same decisions as those derived from the true cost function.

Thus, in contextual optimization, the challenge is to design tractable approaches to tackle decision misspecification. Indeed, a model can yield poor predictions yet still make optimal decisions–and vice versa. Yet, standard methods such as Sequential Learning and Optimization (SLO) and Smart Predict-then-Optimize (SPO+) Elmachtoub and Grigas (2022) fail to guarantee optimal decision-making in the presence of misspecification, leading to suboptimal policies.

1.1 Problem setup

For simplicity, we consider that unknown parameters appear linearly in the objective function of the stochastic optimization problem:

$$\min_{w \in W} \mathbb{E}_{(x,c)\sim P} \left(c^\top w | x\right) = \min_{w \in W} \mathbb{E}_{(x,c)\sim P}(c|x)^\top w. \tag{2}$$

Since c is unknown, we approximate it using a predictor $\hat{c} \in \mathcal{H}$ that is learned using historical data. Importantly, $\mathcal{H}$ can be misspecified, i.e. does not contain the ground truth cost predictor. The decision-making process is as follows: first, observe the context x and predict the cost vector $\hat{c}(x)$. Next, compute a decision $w(\hat{c}(x))$ that minimizes the predicted cost: $w(\hat{c}(x)) \in \arg\min_{w \in W} \hat{c}(x)^\top w$. We consider that the hypothesis set $\mathcal{H}$ is parametric, i.e. it can be written as $\mathcal{H} = \{\hat{c}_\theta,\ \theta \in \mathbb{R}^m\}$. Our goal is to find a predictor $\hat{c}_\theta$ that minimizes the worst case decision cost $\ell_P(\theta) = \mathbb{E}_{(x,c)\sim P}\left(\max_{w \in W^\star(\hat{c}_\theta(x))} c^\top w\right)$, where $W^*(\tilde{c}) := \arg\min_{w\in W} \tilde{c}^\top w$ represents the set of optimal solutions given a predicted cost vector $\tilde{c}$. Here, we consider the worst case decision cost because the solution to $\min_{w\in W} \hat{c}_\theta(x)^\top w$ is not necessarily unique in theory. In practice, however, $\arg\min_{w\in W} \hat{c}_\theta(x)^\top w$ is likely to contain a single element. For example, in the common case where $\hat{c}_\theta(x)$ is a continuous random variable and W is a polyhedron (or more generally when the set of directions $\tilde{c}$ for which the solution to $\min_{w\in W} \tilde{c}^\top w$ is not unique is of Lebesgue measure equal to zero), uniqueness holds with probability 1. Hence, due to practical reasons and theoretical convenience, we make the following assumption henceforth:

Assumption 1. For any $\theta \in \mathbb{R}^m$ such that $\hat{c}_\theta(x) \neq 0$ almost surely, $\min_{w\in W} \hat{c}_\theta(x)^\top w$ has a unique solution with probability 1 when $(x, c) \sim P$.

We denote W_P the set of measurable mappings from the support of the joint probability P of (x, c) to the set of feasible decisions W. Under Assumption 1, we can see that the problem of minimizing target loss can be written as follows:

$$\min_{\theta\in\mathbb{R}^m} \ell_P(\theta) = \min_{\theta\in\mathbb{R}^m} \min_{w_P \in W_P} \mathbb{E}_{(x,c)\sim P}\left(c^\top w_P(x)\right) \text{ s.t. } \forall x \in \mathbb{R}^k,\ w_P(x) \in \arg\min_{w\in W} \hat{c}_\theta(x)^\top w. \tag{3}$$

In Appendix A.9, we formulate a proposition giving a sufficient condition for Assumption 1 to hold when W is a polyhedron. Our task is then to solve problem (3). Since known approaches cannot be used here, we develop a surrogate loss–based method with two properties: consistency, meaning its optimal solutions match those of the target loss, and tractability, meaning it is feasible to optimize.

One way to illustrate why traditional approaches such as SLO and (SPO+) fail to handle misspecified $\mathcal{H}$ is to consider the setting of binary classification suppose the decision space is $W = [-1/2, 1/2]$, and the cost vector c takes values in $\{-1, 1\}$, and the minimization in (2) is replaced by maximization. In this setting, we can build an example where SLO and SPO+ favor a suboptimal decision. Further details are provided in Appendix A.1.

1.2 Our contribution

To address these limitations, we propose a novel decision-aware surrogate loss function that explicitly minimizes decision error regardless of whether or not $\mathcal{H}$ is misspecified. Our approach ensures optimal decision performance under model misspecification and provides strong theoretical guarantees, ensuring both optimality and generalization performance. Importantly, our tractable surrogate loss function matches the optimal solutions of the true objective (3), has favorable optimization properties, avoids bad local minima, and generalizes well to new distributions. We prove that our surrogate loss yields globally optimal decision policies (**Theorem 2**), exhibits strong generalization properties (**Theorem 3**) and is computationally tractable, enabling efficient optimization via first-order methods (**Theorem 4**). We also evaluate our approach against SPO+ and SLO, showing superior performance under varying degrees of misspecification.

The remainder of the paper is structured as follows. Section 2 formalizes the problem and introduces our surrogate loss function. Section 3 establishes theoretical guarantees. Section 4 presents empirical results comparing our approach to existing methods.

2 Our approach

We introduce a new surrogate loss function – referred as Consistent Integrated Learning and Optimization (CILO) loss – for contextual optimization under misspecification.

Definition 1. (CILO loss) For $\beta \in \mathbb{R}$, we define the function ℓ_P^β as

$$\ell_P^\beta(\hat{c}) = \min_{\overline{w}_P^\beta \in \overline{W}_P^\beta} \mathbb{E}_{(x,c)\sim P}\left(\hat{c}(x)^\top \overline{w}_P^\beta(x)\right) - \min_{w_P \in W_P} \mathbb{E}_{(x,c)\sim P}\left(\hat{c}(x)^\top w_P(x)\right)$$

for any measurable function $\hat{c}$, where $\overline{W}_P^\beta = \left\{\overline{w}_P \in W_P,\ \mathbb{E}_{(x,c)\sim P}\left(c^\top \overline{w}_P(x)\right) \leq \beta\right\}$. Furthermore, for any $\theta \in \mathbb{R}^m$, we denote $\ell_P^\beta(\theta) := \ell_P^\beta(\hat{c}_\theta)$.

We denote $\beta^\star_{\mathcal{H},P} := \min_{\hat{c}\in\mathcal{H}} \ell_P(\hat{c})$ and $\beta_{\max,P} = \mathbb{E}_{(x,c)\sim P}(\max_{w\in W} c^\top w)$. In Section 1.1, we have mentioned that we can construct an example in the binary classification setting where SLO and SPO+ favor a suboptimal solution. We can show that for the same example, our surrogate favors an optimal solution (see Appendix A.4). We will now prove that the consistency observed in example 2 holds in general. We now formulate our main consistency theorem.

Theorem 2. *Let $\beta \in \mathbb{R}$ such that $\beta^\star_{\mathcal{H},P} \leq \beta < \beta_{max,P}$. Under Assumption 1, for every $\theta \in \mathbb{R}^m$ with $\hat{c}_\theta(x) \neq 0$ almost surely, $\ell_P(\theta) \leq \beta$ if and only if θ is a minimizer of ℓ_P^β. In particular, when $\beta = \beta^\star_{\mathcal{H},P}$, θ is a minimizer of ℓ_P^β if and only if it is a minimizer of ℓ_P.*

The proof of this theorem is in Appendix A.10. Under mild assumptions, and assuming that θ is bounded away from 0, we can prove a stronger consistency result (Appendix A.6).

The theorem above suggests that a natural way to minimize ℓ_P is to first minimize the CILO loss ℓ_P^β for a suitable β using a tractable method that ensures $\hat{c}_\theta(x) \neq 0$ almost surely. Though $\beta^\star_{\mathcal{H},P}$ is unknown, achieving a small value of ℓ_P^β ensures that ℓ_P has a value close to or smaller than β, and β can be chosen via line search over a suitable interval $[\underline{\beta}, \overline{\beta}]$. In what follows, we develop a tractable (polynomial-time and practical) procedure to optimize ℓ_P^β and formalize the link between near-optimal ℓ_P^β and the target loss ℓ_P. This addresses a key limitation: existing surrogate losses guarantee optimality only under well-specified settings, performing variably under misspecification, whereas our method ensures global optimality across all levels (Section 3).

3 Technical approach and theoretical guarantees

3.1 Generalization performance

We study the generalization performance of the empirical version of CILO loss and show that by optimizing the empirical CILO loss, we can ensure a nearly optimal value for its out of sample counterpart ℓ_P^β. We now present our main generalization result. We denote $\beta_{\min,P} := \mathbb{E}_{(x,c)\sim P}(\min_{w\in W} c^\top w)$.

Theorem 3. *Let $\beta \geq \beta^\star_{\mathcal{H},P}$. Let $\theta^\star \in \mathbb{R}^m$ such that $\ell_{P_n}^\beta(\theta^\star) \leq \varepsilon$ where P_n is the uniform distribution over n fixed samples from P. Assume that there exists $D \geq 0$ such that $\|\theta^\star\| \leq D$. Under assumptions 5, and 6, with probability at least $1-\delta$, we have*

$$\ell_P^\beta(\theta^\star) \leq \varepsilon + O\left(\frac{1}{\beta - \beta_{min,P}}\sqrt{\frac{\log\frac{1}{\delta}}{n}}\right).$$

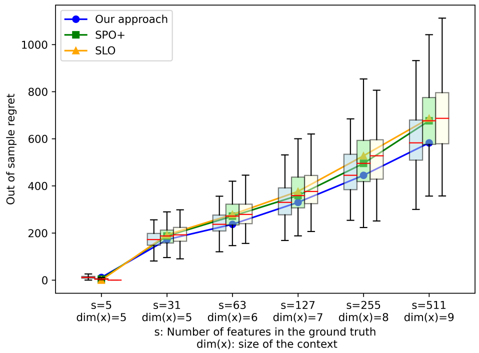

Figure 1: Comparison of SPO+, SLO, and our approach for different levels of misspecification

Notice that when the hypothesis set $\mathcal{H}$ is decision-misspecified, $\beta^{\star}_{\mathcal{H},P} > \beta_{\min,P}$, so the term $\frac{1}{\beta-\beta_{\min,P}}$ is bounded. If $\mathcal{H}$ is (nearly) well-specified, we can pick a larger β to ensure a good generalization bound. This theorem implies our surrogate loss generalizes well when n is large, so minimizing the empirical surrogate loss $\ell^{\beta}_{P_n}$ yields a small optimality gap for its out-of-sample counterpart ℓ^{β}_{P}. Using Theorem 2 and Theorem 3, we can deduce that finding θ such that $\hat{c}_\theta(x) \neq 0$ almost surely and minimizing $\ell^{\beta}_{P_n}$ yields an optimal value for $\ell_P(\theta)$.

3.2 Tractability

We have shown that our surrogate is consistent and generalizable. Hence, in order to find a minimizer of 3, it suffices to minimize the empirical version of our surrogate. The theorem below shows that our surrogate enjoys good landscape properties when the hypothesis set is linear.

Theorem 4. *Under Assumptions 5 and 6, for any $\beta \geq \beta^{\star}_{\mathcal{H},P_n}$, $\ell^{\beta}_{P_n}$ is a difference of convex functions $\ell^{\beta}_{P_n}$. Furthermore, assuming the hypothesis set is linear, i.e. can be written for all $\theta \in \mathbb{R}^m$ as $\hat{c}_\theta : x \longmapsto \Phi(x)\theta$ where $\Phi(x) \in \mathbb{R}^{d\times m}$, all nonzero stationary points of $\ell^{\beta}_{P_n}$ are global minimizers.*

The informal theorem above yields two key conclusions: under mild boundedness assumptions, our surrogate is a difference of convex functions, so we can find a stationary point using state-of-the-art methods for such functions (see Sun and Sun (2021)). Moreover, when the hypothesis set is linear, any such nonzero stationary point is a global optimum for both the surrogate (Definition 1) and the target (3) function. The remaining challenge is ensuring the stationary point is nonzero. Further details on our approach and avoiding zero stationary points are in Appendix A.8.

4 Computational experiments

We make a similar comparison to Hu et al. (2022), where SLO is compared to SPO+, and where they show that when the hypothesis set is well-specified, SLO yields a smaller ℓ_P than SPO+, while under misspecification, SPO+ performs better. Here, we compare our method with SPO+ and SLO in the misspecified setting. The ground truth cost is $c(x) = \theta^{\star}\phi(x)$ for some $\theta^{\star} \in \mathbb{R}^{20\times s}$ and all $x \in \mathbb{R}^k$, with $\phi(x) \in \mathbb{R}^s$. Predictors take the form $\hat{c}_\theta(x) = \theta\tilde{\phi}(x)$ for $\theta \in \mathbb{R}^{20\times 5}$, where $\tilde{\phi}(x) \in \mathbb{R}^5$ is $\phi(x)$ truncated to its first 5 coordinates. When $s = 5$, the hypothesis set is well-specified; as s grows, misspecification increases. The hypothesis set is linear (see Appendix A.20). To optimize $\ell^{\beta}_{P_n}$, we did not need to use our approach to avoid zero (since the iterates did not converge to zero), but only ran gradient descent on our surrogate's Moreau envelope, choosing β by line search. For each s, we ran 96 experiments to get the distribution of testing loss ℓ_P for each method, shown in the box plots of Figure 1. The same plot with relative regret is in Appendix C.

Figure 1 shows the regret of SLO, SPO+, and our method, i.e., $\ell'_P(\theta) = \ell_P(\theta) - \beta_{\min,\mathrm{P}} \geq 0$. In the well-specified setting ($s = 5$), SLO performs best; as misspecification grows ($s > 5$), our method outperforms both, with SLO performing worst. This aligns with Hu et al. (2022) and Elmachtoub et al. (2023), as well as our consistency and generalization results (Theorems 2 and 3).

References

Utsav Sadana, Abhilash Chenreddy, Erick Delage, Alexandre Forel, Emma Frejinger, and Thibaut Vidal. A survey of contextual optimization methods for decision-making under uncertainty. *European Journal of Operational Research*, 2024.

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

Yichun Hu, Nathan Kallus, and Xiaojie Mao. Fast rates for contextual linear optimization. *Management Science*, 68(6):4236–4245, 2022.

Priya Donti, Brandon Amos, and J Zico Kolter. Task-based end-to-end model learning in stochastic optimization. *Advances in neural information processing systems*, 30, 2017.

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

Chunlin Sun, Shang Liu, and Xiaocheng Li. Maximum optimality margin: A unified approach for contextual linear programming and inverse linear programming. In *International Conference on Machine Learning*, pages 32886–32912. PMLR, 2023.

Daniel McKenzie, Samy Wu Fung, and Howard Heaton. Faster predict-and-optimize with three-operator splitting. *arXiv preprint arXiv:2301.13395*, 2023.

James Kotary, My H Dinh, and Ferdinando Fioretto. Backpropagation of unrolled solvers with folded optimization. *arXiv preprint arXiv:2301.12047*, 2023.

Michael Huang and Vishal Gupta. Learning best-in-class policies for the predict-then-optimize framework. *arXiv preprint arXiv:2402.03256*, 2024.

Kaizhao Sun and Xu Andy Sun. Algorithms for difference-of-convex (dc) programs based on difference-of-moreau-envelopes smoothing. *arXiv e-prints*, pages arXiv–2104, 2021.

Adam N Elmachtoub, Henry Lam, Haofeng Zhang, and Yunfan Zhao. Estimate-then-optimize versus integrated-estimation-optimization: A stochastic dominance perspective. *arXiv preprint arXiv:2304.06833*, 2023.

Justin Domke. Generic methods for optimization-based modeling. In *Artificial Intelligence and Statistics*, pages 318–326. PMLR, 2012.

Vishal Monga, Yuelong Li, and Yonina C Eldar. Algorithm unrolling: Interpretable, efficient deep learning for signal and image processing. *IEEE Signal Processing Magazine*, 38(2):18–44, 2021.

Brandon Amos and J Zico Kolter. Optnet: Differentiable optimization as a layer in neural networks. In *International Conference on Machine Learning*, pages 136–145. PMLR, 2017.

Akshay Agrawal, Brandon Amos, Shane Barratt, Stephen Boyd, Steven Diamond, and J Zico Kolter. Differentiable convex optimization layers. *Advances in neural information processing systems*, 32, 2019.

Haixiang Sun, Ye Shi, Jingya Wang, Hoang Duong Tuan, H Vincent Poor, and Dacheng Tao. Alternating differentiation for optimization layers. *arXiv preprint arXiv:2210.01802*, 2022.

Mo Liu, Meng Qi, and Zuo-Jun Max Shen. End-to-end deep learning for automatic inventory management with fixed ordering cost. *Available at SSRN 3888897*, 2021.

Gar Goei Loke, Qinshen Tang, and Yangge Xiao. Decision-driven regularization: A blended model for predict-then-optimize. *Available at SSRN 3623006*, 2022.

Jihwan Jeong, Parth Jaggi, Andrew Butler, and Scott Sanner. An exact symbolic reduction of linear smart predict+ optimize to mixed integer linear programming. In *International Conference on Machine Learning*, pages 10053–10067. PMLR, 2022.

Nathan Kallus and Xiaojie Mao. Stochastic optimization forests. *Management Science*, 69(4):1975–1994, 2023.

Bo Tang and Elias B Khalil. Pyepo: A pytorch-based end-to-end predict-then-optimize library for linear and integer programming. *Mathematical Programming Computation*, 16(3):297–335, 2024.

Boris Mityagin. The zero set of a real analytic function. *arXiv preprint arXiv:1512.07276*, 2015.

Dylan J Foster, Claudio Gentile, Mehryar Mohri, and Julian Zimmert. Adapting to misspecification in contextual bandits. *Advances in Neural Information Processing Systems*, 33:11478–11489, 2020.

Sanath Kumar Krishnamurthy, Vitor Hadad, and Susan Athey. Adapting to misspecification in contextual bandits with offline regression oracles. In *International Conference on Machine Learning*, pages 5805–5814. PMLR, 2021.

Jong-Shi Pang, Meisam Razaviyayn, and Alberth Alvarado. Computing b-stationary points of nonsmooth dc programs. *Mathematics of Operations Research*, 42(1):95–118, 2017.

Jorge Nocedal and Stephen J Wright. *Numerical optimization*. Springer, 1999.

Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge university press, 2019.

A Appendix

A.1 Example where SLO and SPO+ favor a suboptimal decision

In binary classification, the optimization problem then reduces to minimizing:

$$\ell_P(\hat{c}) = -\frac{1}{2} + \mathbb{E}_{(x,c)\sim P}\left(\mathbb{1}_{\text{sign}(\hat{c}(x))\neq c}\right). \tag{4}$$

This formulation is closely related to minimizing classification error in binary classification. We can observe that SLO and SPO+ lead to suboptimal solutions in such a setting.

Example 1. Consider the following hypothesis class: $\mathcal{H} = \{\hat{c}_1, \hat{c}_2\}$, and the context x takes only two values 1 and 2 equally likely, and c always takes the value 1. Assume that $\hat{c}_1(1) = \frac{1}{8}$, $\hat{c}_1(2) = \frac{1}{8}$, $\hat{c}_2(1) = 1$, and $\hat{c}_2(2) = -\frac{1}{6}$. Although $\hat{c}_1$ leads to optimal decision-making, both SPO+ and SLO select $\hat{c}_2$, a suboptimal predictor.

We provide below the calculations justifying the example above.

$$\begin{aligned}
\ell_P(\hat{c}) &= \mathbb{E}_{(x,c)\sim P}\left(\max_{w\in W^\star(\hat{c}(x))} c^\top w\right), \\
\ell_{\text{SPO+}}(\hat{c}) &:= \mathbb{E}_{(x,c)\sim P}\left(\max_{w\in W}(c^\top w - 2(\hat{c}^\top w - \hat{c}^\top w(c)))\right) \\
\ell_{\text{SLO}}(\hat{c}) &= \sqrt{\mathbb{E}_{(x,c)\sim P}\left(\|c - \hat{c}(x)\|^2\right)}.
\end{aligned}$$

Notice that whereas $\ell_P(\hat{c}_1) = -\frac{1}{2}$ and $\ell_P(\hat{c}_2) = 0 > \ell_P(\hat{c}_1)$, we have $\ell_{\text{SPO+}}(\hat{c}_1) = \frac{3}{4}$, $\ell_{\text{SPO+}}(\hat{c}_2) = \frac{2}{3} < \ell_{\text{SPO+}}(\hat{c}_1)$, $\ell_{\text{SLO}}(\hat{c}_1) = \frac{7}{8}$, $\ell_{\text{SLO}}(\hat{c}_2) = \frac{7}{12} < \ell_{\text{SLO}}(\hat{c}_1)$. Although $\hat{c}_2$ leads to suboptimal decisions while $\hat{c}_1$ results in optimal decisions, both SLO and SPO+ appear to favor $\hat{c}_2$.

A.2 Equivalence of the target loss in binary classification

Proof. For every $\hat{c} : \mathbb{R}^k \longrightarrow \mathbb{R}^d$, we denote

$$\ell_P(\hat{c}) = \mathbb{E}_{(x,c)\sim P}\left(\min_{w\in W^\star(\hat{c}(x))} c^\top w\right).$$

We have for every $\hat{c} : \mathbb{R}^k \longrightarrow \mathbb{R}^d$, for every $x \in \mathbb{R}^k$, if $\text{sign}(\hat{c}(x)) = c$, then $W^\star(\hat{c}(x)) = W^\star(c) = \{c/2\}$. Consequently, we have $\min_{w\in W^\star(\hat{c}(x))} c \cdot w = \frac{1}{2}$. If $\text{sign}(\hat{c}(x)) \neq c$, then $\min_{w\in W^\star(\hat{c}(x))} c \cdot w = -\frac{1}{2}$. Consequently, we have for any $x \in \mathbb{R}^k$,

$$\frac{1}{2} + \min_{w\in W^\star(\hat{c}(x))} c \cdot w = \begin{cases} 1 & \text{if sign}(\hat{c}(x)) = c \\ 0 & \text{else.} \end{cases}$$

This directly gives

$$\ell_P(\hat{c}) = \mathbb{E}_{(x,c)\sim P}\left(\mathbb{1}_{\text{sign}(\hat{c}(x))=c}\right) + \frac{1}{2}.$$

□

A.3 Previous approaches to Integrated Learning and Optimization

A more comprehensive review of contextual optimization approaches can be found in the survey by Sadana et al. (2024).

Directly optimizing the target loss. One of the first instances of the ILO method can be traced back to Donti et al. (2017) who provide a practical way to differentiate the target loss ℓ_P under some regularity conditions. Similarly, others have attempted to directly minimize ℓ_P using some estimation of its gradient, such as unrolling Domke (2012), Monga et al. (2021), which consists of keeping track of operations while running gradient descent in order to differentiate the final gradient descent iterate as a function of the model parameters, and implicit differentiation Amos and Kolter (2017), Agrawal et al. (2019), Sun et al. (2022), McKenzie et al. (2023). However, ℓ_P is non-differentiable in general, and even if it were, there is no guarantee that it would be convex. This implies that gradient-based algorithms mentioned above do not guarantee convergence to optimum of the target loss. In contrast, our approach consists of minimizing a tractable smooth surrogate that has the same set of minimizers as ℓ_P. This allows us to avoid the challenge coming from the lack of regularity properties for ℓ_P.

Optimizing a surrogate loss. Another increasingly popular approach, which is most relevant to ours, is optimizing the target loss using a smooth convex surrogate loss. In Elmachtoub and Grigas (2022), a new convex surrogate function, named SPO+. Minimizing SPO+ is proven to also minimize the target loss when the hypothesis set is well-specified, but no consistency results are provided when the chosen predictor class is misspecified. Furthermore, SPO+ seems to outperform SLO (i.e. yield better decision performance) when $\mathcal{H}$ is misspecified, since SLO only focuses on the accuracy of the prediction step, but completely disregards the performance in the optimization step, whereas ILO specifically focuses on the decision performance, and when the hypothesis set is misspecified, maximizing the prediction accuracy of the underlying cost will not necessarily result in good decision performance. Moreover, when $\mathcal{H}$ is well-specified, SLO methods outperform ILO Hu et al. (2022), Elmachtoub and Grigas (2022). Hu et al. (2022) give theoretical and experimental evidence showing that classical SLO methods generalize better than SPO+ when the hypothesis set is well-specified. In fact, evidence in Elmachtoub et al. (2023) suggests that in the well-specified case, SLO might likely have better performance than ILO approaches, and such a behavior is inverted in the misspecified case. Most previous works have not theoretically considered the misspecified case. In Huang and Gupta (2024), a new surrogate is introduced based on an approximation of the directional derivative of ℓ_P, and is shown to be theoretically consistent with ℓ_P. However, only local optimality results are provided for this surrogate, whereas in our work, we provide global optimality guarantees.

Another alternative is the surrogate introduced by Sun et al. (2023), which attempts to maximize the nonbasic reduced costs (when the cost is taken equal to the predicted cost) of past realizations of ground truth optimal decisions. Such a method does not require the knowledge of historical costs, but only solutions of previously seen linear programs. In order for this surrogate to be consistent, authors assume that the chosen hypothesis set is well-specified in the sense of our definition, i.e. $\gamma_{\text{dec}}(\mathcal{H}) = 0$. On one hand, this assumption is weaker than the usual definition of well-specification, which requires that the hypothesis set contains the ground truth predictor. On the other hand, this surrogate is only designed for linear objectives with linear constraints. Our surrogate only requires the set of feasible decisions W to be convex and bounded. In a related paper, Liu et al. (2021) use a neural network structure in an inventory management problem to learn a mapping which provides the optimal merchandise order quantity and order time, and give theoretical guarantees in the well-specified setting. Other surrogates have been considered in the literature Loke et al. (2022), Jeong et al. (2022), Kallus and Mao (2023), which despite having good practical performance benefits, do not seem to theoretically tackle the misspecified case as opposed to our work, though practical toolkits such as PyEPO Tang and Khalil (2024) provide implementations of some of these approaches.

A.4 Revisiting Example 1

Example 2. We revisit Example 1. We have seen that SPO+ and SLO do not favor the best element of the hypothesis set $\mathcal{H}$. Since $\mathcal{H}$ is decision-well-specified, by choosing $\beta = \beta^{\star}_{\mathcal{H},P} = -\frac{1}{2}$. We have $\ell^{\beta}_P(\hat{c}_1) = 0$ and $\ell^{\beta}_P(\hat{c}_2) = \frac{1}{12} > \ell^{\beta}_P(\hat{c}_1)$. We see that indeed here, the global minimum of this new loss is optimal for the target loss.

It is important to keep in mind that in the example above, the model is prediction misspecified, but decision well-specified. Hence, even when there exists a cost predictor that yields the same decisions as c, SPO+ and SLO may fail to favor the best cost predictor in term of decisions. Decision-well-specification is not necessary for the CILO loss' consistency to hold in the example. To see that, it suffices to add a third point in the distribution of P which is labeled the same way by $\hat{c}_1$ and $\hat{c}_2$. This example will yield the same results.

A.5 Assumptions for generalizartion bound

Assumption 5. W is closed and bounded and there exists $K \geq 0$ such that for all $x \in \mathbb{R}^k$, $\|c(x)\| \leq K$. We denote $B_W = \sup_{w \in W} \|w\|$.

Assumption 6. When $\theta = 0$, $\hat{c}_\theta(x) = 0$. Furthermore, there exists $B_\Phi \geq 0$ such that the gradient with respect to θ, $\nabla \hat{c}_\theta(x)$, is piecewise continuous and bounded by B_Φ for every $\theta \in \mathbb{R}^m$ and $x \in \mathbb{R}^k$.

The first part of Assumption 6 is natural when we assume that there exists $\theta_0 \in \mathbb{R}^m$ such that $\hat{c}_{\theta_0} = 0$. By reparametrizing the hypothesis set as $\mathcal{H}' = \hat{c}_{\theta+\theta_0}, \theta \in \mathbb{R}^m$, we obtain $\hat{c}_0 = 0$. Furthermore, a direct consequence of this assumption is that for every $x \in \mathbb{R}^k$ and $\theta \in \mathbb{R}^m$,

$$\|\hat{c}_\theta(x)\| = \left\| \int_0^1 \nabla \hat{c}_{t\theta}(x)^\top \theta dt \right\| \leq \int_0^1 \left\| \nabla \hat{c}_{t\theta}(x)^\top \theta \right\| dt \leq B_\Phi \|\theta\| .$$

A.6 Stronger consistency result

For ease of presentation, we make the following assumption from now on:

Assumption 7. For $\theta \neq 0$, $\hat{c}_\theta(x) \neq 0$ almost surely.

Assumption 7 holds when for every $\theta \in \mathbb{R}^m \setminus \{0\}$, $\hat{c}_\theta$ is a nonzero analytic function and x has a continuous distribution. In this case, the set of zeroes of $\hat{c}_\theta$ is of measure equal to zero (see Mityagin (2015)) and consequently x does not belong to this set almost surely. This assumption, coupled with Assumption 1, implies that for any $\theta \in \mathbb{R}^m \setminus \{0\}$, the problem $\min_{w \in W} \hat{c}_\theta(x)^\top w$ has a unique solution almost surely.

We give a more concrete optimality guarantee obtained by minimizing ℓ_P^β, which sharpens Theorem 2. We first make an assumption similar to the assumption in Hu et al. (2022) which enables the authors to get good generalization guarantees for SLO in the well-specified case. However, our assumption add conditions on the hypothesis set rather than the ground truth cost.

Assumption 8. Assume that W is a polyhedron. We denote $W^\angle$ the set of extreme points of W. For a given context $x \in \mathbb{R}^k$, we denote for every $\theta \in \mathbb{R}^m$,

$$\Delta_\theta(x) = \begin{cases} \min\limits_{w \in W^\angle \setminus W^\star(\hat{c}_\theta(x))} \hat{c}_\theta(x)^\top w - \min\limits_{w \in W^\angle} \hat{c}_\theta(x)^\top w \\ \qquad\qquad\qquad\qquad \text{if } W^\star(\hat{c}_\theta(x)) \neq W \\ 0 \qquad\qquad\qquad\qquad \text{else.} \end{cases}$$

Assume that for some $\alpha > 0$ and $\gamma \geq 0$, for every $\theta \in \mathbb{R}^m$

$$\forall t > 0, \;\; \mathbb{P}(0 < \Delta_\theta(x) \leq \|\theta\| t) \leq \left(\frac{\gamma t}{B_W} \right)^\alpha .$$

The assumption above ensures the stability of the target loss ℓ_P. Specifically, it strengthens the uniqueness guarantee provided by Assumption 1. Besides ensuring that $\min_{w \in W} \hat{c}_\theta(x)^\top w$ has a unique solution almost surely, it offers a sharper measure for how sensitive the mapping $w \longmapsto \hat{c}_\theta(x)^\top w$ is to deviations from $w(\hat{c}_\theta(x))$ with high probability. Assumption 8 is reasonable when $\hat{c}_\theta(x)$ has a continuous distribution, since it is equivalent to say that $\hat{c}_\theta(x)$ is likely to have a direction that is not too close to being perpendicular to one of the faces of the polyhedron W–that is, the probability of this direction to be falling within one of the red cones shown in Figures 2 and 4 decays to 0 as the cones get more narrow. More details about this assumption are provided in Appendix A.11.

The following theorem provides a relationship between optimality gaps of $\ell_P^\beta(\theta)$ and $\ell_P(\theta)$ when θ is bounded away from 0, i.e. there exists $u > 0$ such that $\|\theta\| \geq u$.

Theorem 9. *Under assumptions 5, 6 and 8, there exists $\alpha > 0$ such that for all $\theta \in \mathbb{R}^m$ bounded away from 0 and $\beta \geq \beta^\star_{\mathcal{H},P}$,*

$$\ell_P(\theta) \leq \beta + O\left(\ell_P^\beta(\theta)^{1 - \frac{1}{1+\alpha}} \right).$$

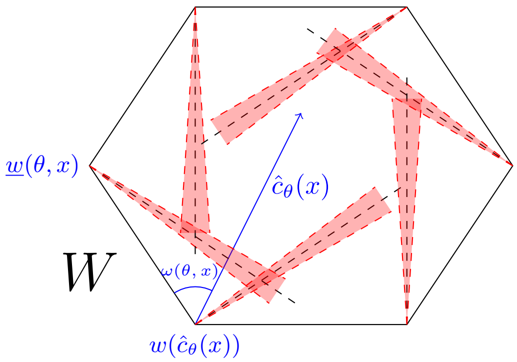

Figure 2: Less likely directions for $\hat{c}_\theta(x)$ relative to W when it is a polyhedron in $\mathbb{R}^2$

The proof of this theorem can be found in Appendix A.14. Assuming that θ is bounded away from 0 is crucial here, because when $\theta = 0$, the uniqueness property in Assumption 1 is not guaranteed, and the consistency in Theorem 2 is not satisfied. Similarly, when θ is close to 0, ℓ_P^β exhibits poor behavior, as it approaches the scenario where consistency breaks down. Although Assumption 8 requires W to be a polyhedron, it is possible to consider a similar yet reasonable assumption that does not impose this requirement. This alternative leads to a stronger consistency inequality (see Appendix A.11).

Remark 1. This theorem suggests that to minimize ℓ_P, one should seek an approximate minimizer θ of ℓ_P^β for a well-chosen $\beta \in \mathbb{R}$ with $\|\theta\|$ kept away from zero. By applying Theorem 3, it follows that it suffices to find some $\theta \in \mathbb{R}^m$ such that $\ell_{P_n}^\beta(\theta)$ is small while ensuring $\|\theta\|$ remains bounded away from zero.

A.7 Predicition misspecification and decision misspecification

In the litterature, a classical metric to quantify the level of prediction misspecification of a given hypothesis set $\mathcal{H}$ is the L^1 or L^2 norm of the distance between the parameter (or the function) one seeks to estimate and the chosen hypothesis set. In our setting, it can be written as

$$\gamma_{\text{pred}}(\mathcal{H}) = \sqrt{\min_{\hat{c} \in \mathcal{H}} \mathbb{E}_{(x,c)\sim P}\left((\hat{c}(x) - c(x))^2\right)}.$$

This metric has been previously adopted in statistics and contextual bandits Foster et al. (2020), Krishnamurthy et al. (2021). We argue that $\gamma_{\text{pred}}(\mathcal{H})$ can be a good metric for evaluating the effect of misspecification for prediction problems, but not for contextual optimization problems in which parameter estimates enter in downstream optimization problem. This can be seen by noting that in our setting, the best achievable cost should not change when multiplying the elements of $\mathcal{H}$ by a positive constant, whereas $\gamma_{\text{pred}}(\mathcal{H})$ is not invariant to such a transformation. To address this, we introduce a new metric to quantify misspecification in contextual optimization.

Definition 2. We define the decision misspecification gap as

$$\gamma_{\text{dec}}(\mathcal{H}) = \min_{\hat{c} \in \mathcal{H}} \mathbb{E}_{(x,c)\sim P}\left(\max_{w \in W^\star(\hat{c}(x))} c(x)^\top w\right) - \mathbb{E}\left(\min_{w \in W} c(x)^\top w\right).$$

The function class $\mathcal{H}$ is decision-well-specified if $\gamma_{\text{dec}}(\mathcal{H}) = 0$, and decision-misspecified otherwise.

First, note that the first term is the smallest possible value of the target loss when choosing a cost predictor from $\mathcal{H}$, and the second term is the smallest possible value of the target loss in the case where $\mathcal{H}$ contains the ground truth cost predictor. Hence, it follows to conclude that $\gamma_{\text{dec}}(\mathcal{H})$ captures the optimality gap (in terms of decision performance) caused by model misspecification. Second, notice that if $\mathcal{H}$ is decision-misspecified, i.e. $\gamma_{\text{dec}}(\mathcal{H}) > 0$, then necessarily it is prediction-misspecified, i.e. $\gamma_{\text{pred}}(\mathcal{H}) > 0$. Third, $\gamma_{\text{dec}}(\mathcal{H})$ is clearly invariant when multiplying the cost predictors by a constant. In Example 1, that we have $\gamma_{\text{pred}}(\mathcal{H}) > 0$ but $\gamma_{\text{dec}}(\mathcal{H}) = 0$. This means that while SPO+ and SLO are consistent when the hypothesis set is prediction-well-specified, their consistency property no longer holds even when the hypothesis set is decision-well-specified in general.

For a given cost predictor $\hat{c}$, denoting $\gamma_{\text{dec}}(\hat{c}) = \mathbb{E}_{(x,c)\sim P}\left(\max_{w\in W^\star(\hat{c}(x))} c^\top w\right) - \mathbb{E}\left(\min_{w\in W} c^\top w\right)$ and $\gamma_{\text{pred}}(\hat{c}) = \sqrt{\mathbb{E}_{(x,c)\sim P}\left((\hat{c}(x) - c(x))^2\right)}$, we can see that SLO focuses on bringing $\gamma_{\text{pred}}(\hat{c})$ as close as possible to $\gamma_{\text{pred}}(\mathcal{H})$, whereas it is more natural to focus on bringing $\gamma_{\text{dec}}(\hat{c})$ as close to $\gamma_{\text{dec}}(\mathcal{H})$ as possible, since small prediction error does not necessarily mean good decision performance in the misspecified case. In the other hand, even if current ILO methods such as SPO+ aim to bring $\gamma_{\text{dec}}(\hat{c})$ as close to $\gamma_{\text{dec}}(\mathcal{H})$, it is unclear whether they can do so in the misspecified case, whereas our surrogate loss's optimality gap dominates $\gamma_{\text{dec}}(\hat{c})$ (see Theorem 9).

A.8 Optimizing our surrogate

Based on theorems 2, 3, and 9, it is sufficient to find a $\theta \in \mathbb{R}^m$ such that $\ell^\beta_{P_n}$ is small and θ is kept bounded away from zero to minimize ℓ_P. The empirical surrogate loss $\ell^\beta_{P_n}$ can be written as:

$$\forall \theta \in \mathbb{R}^m,\ \ell^\beta_{P_n}(\theta) = g_{P_n}(\theta) - \overline{g}^\beta_{P_n}(\theta),$$

where

$$g_{P_n}(\theta) := -\mathbb{E}_{(c,x)\sim P_n}\left(\min_{w\in W} \hat{c}_\theta(x)^\top w\right) \tag{5}$$

$$= -\min_{w_{P_n}\in W_{P_n}} \mathbb{E}_{(c,x)\sim P_n}\left(\hat{c}_\theta(x)^\top w_{P_n}(x)\right) \tag{6}$$

$$\overline{g}^\beta_{P_n}(\theta) := -\min_{w_{P_n}\in \overline{W}^\beta_{P_n}} \mathbb{E}_{(c,x)\sim P_n}\left(\hat{c}_\theta(x)^\top \overline{w}^\beta_{P_n}(x)\right). \tag{7}$$

The equality in (6) is satisfied because choosing a policy w_P minimizing $E_{(c,x)\sim P_n}\left(\hat{c}_\theta(x)^\top \overline{w}^\beta_{P_n}(x)\right)$ is equivalent to choosing $w_P(x)$ minimizing $\hat{c}_\theta(x)^\top w_P(x)$ for every x in $\mathbb{R}^k$.

Landscape properties and smoothing. We take a closer look at the structure of our surrogate. We make the following mild assumption.

Assumption 10. The function $\theta \longmapsto \hat{c}_\theta(x)$ is differentiable for every $x \in \mathbb{R}^k$, and there exists $B_L > 0$ such that for all $\theta, \theta' \in \mathbb{R}^m$ and $x \in \mathbb{R}^k$,

$$\|\nabla_\theta \hat{c}_\theta(x) - \nabla_\theta \hat{c}_{\theta'}(x)\| \leq B_L \|\theta - \theta'\|.$$

This assumption is sufficient for $\ell^\beta_{P_n}$ to be written as a difference of two convex functions. Specifically, we have the following result:

Proposition 1. *Under assumptions 5 and 10, $\ell^\beta_{P_n}$ is a DC (difference of two convex functions) function.*

The proof of the proposition above (see Appendix A.13) relies on the fact that under assumptions 10 and 5, g_{P_n} and $\overline{g}^\beta_{P_n}$ are both weakly convex functions, and consequently their difference is a difference of convex functions.

We aim to find $\theta \in \mathbb{R}^m \setminus \{0\}$ that minimizes $\ell^\beta_{P_n}$. A natural approach is to identify a stationary point of $\ell^\beta_{P_n}$ (i.e., a θ where the subgradient of $\ell^\beta_{P_n}$ contains 0). Despite $\ell^\beta_{P_n}$ having a DC structure, it may still be non-smooth, making the task of finding a stationary point challenging. To address this, we introduce a smoothed version of $\ell^\beta_{P_n}$.

Definition 3. (s-CILO loss) For all $\lambda \in \mathbb{R}^m$ and $\beta \geq \beta_{\mathcal{H},P_n}$, we denote $g^1_{P_n} = g_{P_n}$ and $g^2_{P_n} = \overline{g}^\beta_{P_n}$. For $i \in \{1, 2\}$, let

$$M^i_{P_n}(\lambda) := \min_{\theta\in\mathbb{R}^m} (g^i_{P_n}(\theta) + \frac{1}{2}\|\lambda - \theta\|^2),$$

$$\theta^i_{P_n}(\lambda) = \arg\min_{\theta\in\mathbb{R}^m} (g^i_{P_n}(\theta) + \frac{1}{2}\|\theta - \lambda\|^2).$$

We define a smooth surrogate to the CILO loss, which we call the s-CILO loss $r^\beta_{P_n}(\lambda) := M^1_{P_n}(\lambda) - M^2_{P_n}(\lambda)$. When there is an ambiguity about the value of β, we denote $\theta_{P_n} := \theta^1_{P_n}$, $\overline{\theta}^\beta_{P_n} := \theta^2_{P_n}$, $M_{P_n} := M^1_{P_n}$, $\overline{M}^\beta_{P_n} := M^2_{P_n}$.

Following Sun and Sun (2021) (proposition 1, page 10), the smooth surrogate above has the following property: for every stationary point λ of $r^{\beta}_{P_n}$, $\overline{\theta}^{\beta}_{P_n}(\lambda)$ and $\theta_{P_n}(\lambda)$ are equal and are stationary points of $\ell^{\beta}_{P_n}$. Similar to $\ell^{\beta}_{P_n}$, s-CILO satisfies the positivity property and the fact that its minimum value is equal to 0 from $\ell^{\beta}_{P_n}$.

We now have a way to find a stationary point of $\ell^{\beta}_{P_n}$. A practical way to ensure that the iterates of an optimization algorithm for $\ell^{\beta}_{P_n}$ do not converge to $\theta = 0$ is to add a constraint $\|\theta\| \geq z$ where $z > 0$. Since the function $\theta \longmapsto z - \|\theta\|$ can be written as a difference of two convex functions, we can find a constrained stationary point to $r^{\beta}_{P_n}$ (see Pang et al. (2017) for tractability results for DC constrained minimization).

The DC structure of our surrogate enables us to find a stationary point of $\ell^{\beta}_{P_n}$, thus allowing us to approximately minimize the target loss ℓ_P. Moreover, as we will demonstrate (see Theorem 12 below), for specific choices of the hypothesis set (e.g., a linear hypothesis set), our surrogate has a favorable landscape, meaning that every stationary point of $\ell^{\beta}_{P_n}$ is a global minimum. Consequently, first-order optimization algorithms can effectively find a global minimum of our surrogate.

From now on, we assume that our hypothesis set $\mathcal{H}$ is linear, meaning there exists a mapping $\Phi : \mathbb{R}^k \longrightarrow \mathbb{R}^{m \times d}$ such that for all $\theta \in \mathbb{R}^m$, $\hat{c}_\theta(x) = \Phi(x)\theta$. In this scenario, Assumption 7 is satisfied if $\Phi(x)$ is full rank almost surely. Additionally, Assumption 10 is inherently satisfied, and Assumption 6 can be replaced with the following:

Assumption 11. For any x, the largest singular value of $\Phi(x)$ is bounded above by B_Φ.

With this special choice of hypothesis set, our surrogate loss and its smoothed version enjoy good landscape properties. In particular, every stationary point of $\ell^{\beta}_{P_n}$ is a global minimum. We formulate our main landscape result.

Theorem 12. *$r^{\beta}_{P_n}$ is everywhere differentiable. If λ is a stationary point of $r^{\beta}_{P_n}$ such that $\hat{c}_{\overline{\theta}^{\beta}_{P_n}(\lambda)}(x) \neq 0$ almost surely, then $\overline{\theta}^{\beta}_{P_n}(\lambda)$ is a global minimum of $\ell^{\beta}_{P_n}$. In particular, under assumptions 5 and 6, we have for every $\theta \in \mathbb{R}^m$ and $\varepsilon \geq 0$,*

$$\left\| \nabla r^{\beta}_{P_n} \right\| \leq \varepsilon \Longrightarrow \ell^{\beta}_{P_n}(\theta) \leq 8 B_W B_\Phi \varepsilon.$$

A more complete version of this theorem along with its proof is provided in Appendix A.15. Assuming a priori that running gradient descent to optimize our smooth surrogate, we do not fall into the case $\hat{c}_\theta(x) = 0$ almost surely, our practical procedure to optimize ℓ_P simply consists of finding a suitable β using line search and run gradient descent on $\ell^{\beta}_{P_n}$.

Now, the only remaining problem to address is the case where the algorithm gives $\theta = 0$.

Avoiding zero solutions. Even if we successfully minimize $\ell^{\beta}_{P_n}$, we could encounter the pathological case where $\theta = 0$, which would fail to ensure that $\hat{c}_\theta(x) \neq 0$ almost surely. In this situation, the conditions for Theorems 2 and 9 would not be satisfied. To prevent this, we leverage the linear structure of the hypothesis set. We begin by stating a proposition that highlights key properties of our surrogate when the hypothesis set is linear.

Definition 4. For a set V in $\mathbb{R}^m$ and element $u \in \mathbb{R}^m$, denote $-V = \{-v,\ v \in V\}$ and $d(u, V) = \min_{v \in V} \|u - v\|$ the L^2 distance between u and V.

Proposition 2. *Let $\lambda \in \mathbb{R}^m$ and $\beta \geq \beta^{\star}_{\mathcal{H},P_n}$, we denote*

$$V_{P_n} := \left\{ \mathbb{E}_{(x,c)\sim P} \left(\Phi^\top(x) w_{P_n}(x) \right), w_{P_n} \in W_{P_n} \right\},$$
$$\overline{V}^{\beta}_{P_n} := \left\{ \mathbb{E}_{(x,c)\sim P} \left(\Phi^\top(x) w_{P_n}(x) \right), w_{P_n} \in \overline{W}^{\beta}_{P_n} \right\}.$$

$r^{\beta}_{P_n}$ can be rewritten as

$$r^{\beta}_{P_n}(\lambda) = \frac{1}{2} \left(d(\lambda, -\overline{V}^{\beta}_{P_n})^2 - d(\lambda, -V_{P_n})^2 \right).$$

The proof of this proposition can be found in Appendix A.17.

We observe that $r^{\beta}_{P_n}$ exhibits an interesting property: recovering a candidate solution $\theta \in \mathbb{R}^m$ that is equal to zero is equivalent to finding $\lambda \in \mathbb{R}^m$ such that $d(\lambda, -\overline{V}^{\beta}_{P_n}) = d(\lambda, -V_{P_n}) = 0$, and hence, we would like to find a near stationary point for $r^{\beta}_{P_n}$ such that these distances are non zero. In order to do that, we introduce the following log-barrier surrogate.

Definition 5. (log-CILO loss) Let $\beta \geq \beta^{\star}_{\mathcal{H},P}$, we define for every $\theta \in \mathbb{R}^m$,

$$f^{\beta}_{P_n}(\lambda) = \log\ d(\lambda, -\overline{V}^{\beta}_{P_n}) - \log\ d(\lambda, -V_{P_n}). \tag{8}$$

We call this function the log-CILO loss.

The function $f^{\beta}_{P_n}$ inherits the properties we previously observed in $r^{\beta}_{P_n}$. Specifically, since $d(\lambda, -\overline{V}^{\beta}_{P_n}) \geq d(\lambda, -V_{P_n})$ for every $\lambda \in \mathbb{R}^m$ (due to the fact that $\overline{V}^{\beta}_{P_n} \subset V_{P_n}$), and because the logarithm is a non-decreasing function, we have $f^{\beta}_{P_n}(\lambda) \geq 0$ for every $\lambda \in \mathbb{R}^m$. Furthermore, similar to $r^{\beta}_{P_n}$, if λ is a minimizer of $f^{\beta}_{P_n}$, then the distances $d(\lambda, -\overline{V}^{\beta}_{P_n})$ and $d(\lambda, -V_{P_n})$ are equal. Additionally, it can be shown that any stationary point of $f^{\beta}_{P_n}$ is also a stationary point of $r^{\beta}_{P_n}$ (see Theorem 15).

If gradient descent is used to optimize $r^{\beta}_{P_n}$ and a subsequence of the iterates converges outside the set $-V_{P_n}$ (and hence outside $-\overline{V}^{\beta}_{P_n}$ as well), we can recover from the limit of this sequence a $\theta \in \mathbb{R}^m$ that is bounded away from zero and serves as a minimizer of $\ell^{\beta}_{P_n}$. As a result, by Theorems 3 and 9, we successfully obtain a desirable approximate minimizer of ℓ_P. However, if every limit point of the gradient descent iterates converges within the sets $-V_{P_n}$ and $-V^{\beta}_{P_n}$, the function $f^{\beta}_{P_n}$ can be used instead. Gradient descent applied to $f^{\beta}_{P_n}$ will not yield iterates converging inside the sets $-V_{P_n}$ and $-V^{\beta}_{P_n}$ due to the presence of $\log$ barriers. The following proposition outlines the possible outcomes when running gradient descent on $f^{\beta}_{P_n}$.

Proposition 3. *Consider the backtracking line search gradient descent method applied to the function $f^{\beta}_{P_n}$, ensuring sufficient decrease at each iteration (refer to Algorithm 3.1 in Nocedal and Wright (1999)). Suppose the gradient descent sequence has at least one limit point. There are two possible cases:*

1. *The limit point lies strictly outside the interior of $-V_{P_n}$ or $-\overline{V}^{\beta}_{P_n}$;*
2. *The limit point lies on the common boundary of $-V_{P_n}$ and $-\overline{V}^{\beta}_{P_n}$.*

The proof of this proposition is provided in Appendix A.19.

If the gradient sequence $\{\lambda_{i_t}\}$ converges to the common boundary of $-V_{P_n}$ and $-\overline{V}^{\beta}_{P_n}$, the corresponding sequence $\{\overline{\theta}^{\beta}_{P_n}(\lambda_{i_t})\}$ approaches zero, which is not a desirable solution. Fortunately, this issue can be addressed. Since the sequence converges to a limit point on the boundary of $-V_{P_n}$ and $-\overline{V}^{\beta}_{P_n}$ which is also a stationary point of $r^{\beta}_{P_n}$, there must exist some λ_{i_t} such that is an ε-stationary point of $r^{\beta}_{P_n}$. We then move λ_{i_t} orthogonally away from the boundary of $-V_{P_n}$ (see Figure 5). This yields a new approximate stationary point that is sufficiently distant from both $-V_{P_n}$ and $-V^{\beta}_{P_n}$. As a result, the corresponding θ is an approximate global minimizer of $\ell^{\beta}_{P_n}$ that is bounded away from zero.

Theorem 15 in Appendix A.18 provides an overview of this procedure, as well as a relationship between the optimality for $f^{\beta}_{P_n}$ and $r^{\beta}_{P_n}$.

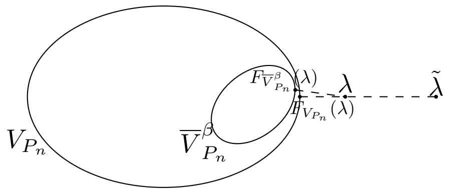
$F_{\overline{V}^{\beta}_{P_n}}(\lambda)$
λ
$\tilde{\lambda}$
$F_{V_{P_n}}(\lambda)$
V_{P_n}
$\overline{V}^{\beta}_{P_n}$

Figure 3: Construction of $\tilde{\lambda}$: when λ is close to the sets $\overline{V}^{\beta}_{P_n}$ and V_{P_n}, we construct a new approximately stationary point by moving away from V_{P_n} orthogonally to its boundary.

A.9 Sufficient condition for uniqueness assumption to hold when W is a polyhedron

Proposition 4. *If the context variable x has a continuous probability distribution and the mapping $x \longmapsto \hat{c}_\theta(x)$ is a nonzero analytic function for any $\theta \neq 0$, then Assumption 1 holds for any polyhedron W that satisfies $\hat{c}_\theta(x)^\top v \neq 0$ for any $\theta \in \mathbb{R}^m \setminus \{0\}$ and any nonzero $v \in W$.*

Proof. Let $\{S_i\}_{i\in I}$ be all the faces of W, where I is a finite index set. Let v_i be an arbitrary tangent direction of S_i and let $\mathcal{V} = \{v_i\}_{i\in I}$ be a finite set. Let U be the set containing all u such that the solution to the linear program $\min_{w\in W} u^\top w, u \in U$ is not unique. Then any $u \in U$ satisfies $u^\top v_i = 0$ for some i. Therefore, $\hat{c}_\theta(x) \in U$ if $\hat{c}_\theta(x)^\top v_i = 0$ for some i. Let $X_0 := \{x \in \mathbb{R}^k \mid \hat{c}_\theta(x) \in U\}$ and $X_i := \{x \in \mathbb{R}^k \mid \hat{c}_\theta(x)^\top v_i = 0\}$. We have $X_0 \subseteq \bigcup_{i\in I} X_i$. By the assumption in the Proposition, $\hat{c}_\theta(x)^\top v_i$ is a nonzero analytic function and then by Mityagin (2015), X_i is a zero-measure set. Therefore, $X_0 = \bigcup_{i\in I} X_i$ is a zero-measure set since I is finite. Finally, since X_0 is of measure zero and x is a continuous random variable, x lies outside of X_0 almost surely, i.e. the solution to $\min_{w\in W} \hat{c}_\theta(x)^\top w$ is unique almost surely. □

A.10 Proof of Theorem 2

Proof. We denote $\beta^\star_{\mathcal{H},P} = \min_{\theta\in\mathbb{R}^m} \ell_P(\theta)$, $\beta_{\max,P} = \mathbb{E}_{(x,c)\sim P}(\max_{w\in W} c^\top w)$, and $\beta_{\min,P} = \mathbb{E}_{(x,c)\sim P}\left(\min_{w\in W} c^\top w\right)$. Let $\beta \in \mathbb{R}$. We first study the property $\ell_P(\theta) \leq \beta$ when β satisfies $\beta^\star_{\mathcal{H},P} \leq \beta < \beta_{\max,P}$ and $\hat{c}_\theta(x) \neq 0$ almost surely. The inequality $\beta^\star_{\mathcal{H},P} \leq \beta$ ensures that the condition $\ell_P(\theta) \leq \beta$ is feasible, and the inequality $\beta < \beta_{\max,P}$ ensures that the condition $\ell_P(\theta) \leq \beta$ is not trivial. Notice that if we take $\beta = \beta^\star_{\mathcal{H},P}$, any $\theta \in \mathbb{R}^m$ satisfying $\ell_P(\theta) \leq \beta$ is a minimizer of $\ell_P(\theta)$. When Assumption 1 holds, and when $\hat{c}_\theta(x) \neq 0$ almost surely, $W^\star(\hat{c}_\theta(x))$ contains only one element for almost every x. Hence, we denote $w(\hat{c}_\theta(x))$ to be the unique solution in $W^\star(\hat{c}_\theta(x))$ when $W^\star(\hat{c}_\theta(x))$ contains only one element. The condition $\ell_P(\theta) \leq \beta$ can be rewritten as $\mathbb{E}_{(x,c)\sim P}(c^\top w(\hat{c}_\theta(x))) \leq \beta$ when $\hat{c}_\theta(x) \neq 0$ almost surely. For a given $\theta \in \mathbb{R}^m$ and $w_P \in W_P$, if w_P satisfies

$$w_P \in \arg\min_{w_P\in W_P} \mathbb{E}_{(x,c)\sim P}(\hat{c}_\theta(x)^\top w_P(x)), \tag{9}$$

then we have, $w_P(x) = w(\hat{c}_\theta(x))$ almost surely, as for each x, the unique minimizer of $\hat{c}_\theta(x)^\top w$ over $w \in W$ is given by $w\left(\hat{c}_\theta(x)\right)$, according to Assumption 1.

If $\theta \in \mathbb{R}^m$ satisfies $\ell_P(\theta) \leq \beta$, and w_P is the measurable mapping satisfying condition (9) (and hence $w_P(x) = w(\hat{c}_\theta(x))$ almost surely), then we have $\mathbb{E}_{(x,c)\sim P}\left(c^\top w_P\left(x\right)\right) \leq \beta$. This suggests that adding the linear constraint $\mathbb{E}_{(x,c)\sim P}\left(c^\top w_P\left(x\right)\right) \leq \beta$ to (9) does not change the set of minimizers of $w_P \longmapsto \mathbb{E}_{(x,c)\sim P}(\hat{c}_\theta(x)^\top w_P(x))$ when $\ell_P(\theta) \leq \beta$. Defining $\overline{W}^{\beta}_P = \left\{\overline{w}_P \in W_P,\ \mathbb{E}_{(x,c)\sim P}\left(c^\top \overline{w}_P(x)\right) \leq \beta\right\}$, our latter statement means that if $\ell_P(\theta) \leq \beta$, the two optimization problems

$$\min_{w_P\in W_P} \mathbb{E}_{(x,c)\sim P}\left(\hat{c}_\theta(x)^\top w_P(x)\right) \tag{10}$$

and

$$\min_{\overline{w}_P\in \overline{W}^{\beta}_P} \mathbb{E}_{(x,c)\sim P}\left(\hat{c}_\theta(x)^\top \overline{w}_P(x)\right) \tag{11}$$

would result in identical values of the objective function.

Let $\beta \in \mathbb{R}$ such that $\beta^{\star}_{\mathcal{H},P} \le \beta < \beta_{\max,P}$ and $\theta \in \mathbb{R}^m$ satisfying $\hat{c}_\theta(x) \neq 0$ almost surely. Assume that θ is optimal for ℓ^{β}_{P}, i.e. $\ell^{\beta}_{P}(\theta) = 0$. Let $\overline{w}_P(x)$ be the solution of (11). Then $\ell^{\beta}_{P}(\theta) = 0$ implies

$$\mathbb{E}_{(x,c)\sim P}\left(\hat{c}_\theta(x)^\top \overline{w}_P(x) - \hat{c}_\theta(x)^\top w\left(\hat{c}_\theta(x)\right)\right) = 0.$$

Since the random variable $\hat{c}_\theta(x)^\top \overline{w}_P(x) - \hat{c}_\theta(x)^\top w\left(\hat{c}_\theta(x)\right)$ is positive almost surely, the equality above implies that $\hat{c}_\theta(x)^\top \overline{w}_P(x) - \hat{c}_\theta(x)^\top w\left(\hat{c}_\theta(x)\right) = 0$ almost surely, i.e. $\overline{w}_P$ is also a minimizer of $w_P \longmapsto E_{(x,c)\sim P}\left(\hat{c}_\theta(x)^\top w_P(x)\right)$, and hence under Assumption 1, we have $\overline{w}_P(x) = w(\hat{c}_\theta(x))$ almost surely. This finally gives

$$\ell_P(\theta) = \mathbb{E}_{(x,c)\sim P}(c^\top w\left(\hat{c}_\theta(x)\right)) = \mathbb{E}_{(x,c)\sim P}(c^\top \overline{w}_P(x)) \le \beta.$$

□

A.11 More details about Assumption 8

We first mention the assumption in Hu et al. (2022) which enables the authors to get good generalization guarantees for SLO in the well-specified case. Note that we will not be making this assumption, but we mention it just for reference.

Assumption 13. Assume that W is a polyhedron. We denote $W^{\angle}$ the set of extreme points of W. For a given context $x \in \mathbb{R}^k$, we denote

$$\Delta(x) = \begin{cases} \min_{w\in W^{\angle}\setminus W^{\star}(c(x))} c(x)^\top w - \min_{w\in W^{\angle}} c(x)^\top w & \text{if } W^{\star}(c(x)) \neq W \\ 0 & \text{else.} \end{cases}$$

Assume that for some $\alpha, \gamma \ge 0$,

$$\forall t > 0,\ \mathbb{P}(0 < \Delta(x) \le t) \le \left(\frac{\gamma t}{B_W}\right)^{\alpha}.$$

We provide further justification of the remarks following Assumption 8. When W is a polyhedron, denoting for $\theta \neq 0$, $\underline{w}(\theta, x) = \arg\min_{w\in W^{\angle}\setminus\{w(\hat{c}_\theta(x))\}} \hat{c}_\theta(x)^\top w$, and

$$\cos\left(\omega(\theta, x)\right) = \frac{\hat{c}_\theta(x)^\top(\underline{w}(\theta,x) - w(\hat{c}_\theta(x)))}{\|\hat{c}_\theta(x)\|\,\|\underline{w}(\theta,x) - w(\hat{c}_\theta(x))\|},$$

Assumption 8 can be rewritten as

$$\mathbb{P}\left(\|\hat{c}_\theta(x)\|\,\|\underline{w}(\theta,x) - w(\hat{c}_\theta(x))\|\cos(\omega(\theta,x)) \le \|\theta\|\, t\right) \le \left(\frac{\gamma t}{B_W}\right)^{\alpha}. \tag{12}$$

We do not need to focus on the term $\|\underline{w}(\theta,x) - w(\hat{c}_\theta(x))\|$, as it is both upper bounded and bounded away from zero. Therefore, inequality (12) is equivalent to $\hat{c}_\theta(x)$ having a norm that is bounded away from zero with high probability, and an angle $\omega(\theta, x)$ that is bounded away from $\frac{\pi}{2}$ when θ is bounded away from 0. In other words, $\hat{c}_\theta(x)$ is likely to have a direction that is not too close to being perpendicular to one of the faces of the polyhedron W. In other words, the probability of this direction to be falling within one of the red cones shown in Figure 2 and 4 decays to 0 as the cones get more narrow. When $\hat{c}_\theta(x)$ follows a continuous distribution, such a property on the direction of $\hat{c}_\theta(x)$ is reasonable.

The key idea of the proof of Theorem 9 relies on the following sensitivity inequality resulting from Assumption 8 which holds for every $t \ge 0$ and mapping $w \in W_P$:

$$\mathbb{E}_{(x,c)\sim P}\left(\left|\hat{c}_\theta(x)^\top w(\hat{c}_\theta(x)) - \hat{c}_\theta(x)^\top w(x)\right|\right) \ge tA_1\|\theta\|\,\mathbb{E}_{(x,c)\sim P}\left(\|w(\hat{c}_\theta(x)) - w(x)\|\right) - A_2\|\theta\|\, t^{\alpha},$$

where A_1, A_2 are positive constants. The left-hand side equals $\ell^{\beta}_{P}(\theta)$ for a well-chosen $w \in W_P$, while the right-hand side is directly related to $\ell_P(\theta) - \beta$. Optimizing over t in the right-hand side yields the desired inequality. By omitting the term $A_2|\theta|t^{\alpha}$ and the dependence of the right-hand side on t, it is possible to obtain a stronger bound that does not require W to be a polyhderon. Specifically, if we assume the existence of $B_s > 0$ such that for every $\theta \in \mathbb{R}^m$ and every mapping $w \in W_P$,

$$\mathbb{E}_{(x,c)\sim P}\left(\left|\hat{c}_\theta(x)^\top w(\hat{c}_\theta(x)) - \hat{c}_\theta(x)^\top w(x)\right|\right) \ge B_s\|\theta\|\,\mathbb{E}_{(x,c)\sim P}\left(\|w(\hat{c}_\theta(x)) - w(x)\|\right), \tag{13}$$

the bound $\ell_P(\theta) \le \beta + O\left(\ell^{\beta}_{P}(\theta)\right)$ holds when $\|\theta\|$ is bounded away from 0. This bound is stronger since $1 - \frac{1}{1+\alpha} < 1$ for every $\alpha > 0$ and does not require W to be a polyhedron. This assumption is also reasonable when P is continuous.

A.12 Proof of Theorem 3

Proof. We proceed in two steps.

Step 1. We first prove the Lipschitz continuity of the generalization, i.e. that there exists a constant $A \geq 0$ such that for every $\theta, \theta' \in \mathbb{R}^m$,

$$\left| \ell^\beta_{P_n}(\theta) - \ell^\beta_P(\theta) - \left(\ell^\beta_{P_n}(\theta') - \ell^\beta_P(\theta') \right) \right| \leq A \left\| \theta - \theta' \right\| . \tag{14}$$

For every $\theta \in \mathbb{R}^m$, using Lagrangian duality,

$$\begin{aligned}
\ell^\beta_{P_n}(\theta) &= \min_{\overline{w}^\beta_{P_n} \in \overline{W}^\beta_{P_n}} \mathbb{E}_{(x,c)\sim P_n} \left(\hat{c}_\theta(x)^\top \overline{w}^\beta_{P_n}(x) \right) - \min_{w_{P_n} \in W_{P_n}} \mathbb{E}_{(x,c)\sim P_n} \left(\hat{c}_\theta(x)^\top w_{P_n}(x) \right) && (15)\\
&= \min_{\overline{w}^\beta_{P_n} \in W_{P_n}} \max_{y \geq 0} \mathbb{E}_{(x,c)\sim P_n} \left(\hat{c}_\theta(x)^\top \overline{w}^\beta_{P_n}(x) \right) + y \left(\mathbb{E}_{(x,c)\sim P_n} \left(c(x)^\top w^\beta_{P_n}(x) \right) - \beta \right) && (16)\\
&- \min_{w_{P_n} \in W_{P_n}} \mathbb{E}_{(x,c)\sim P_n} \left(\hat{c}_\theta(x)^\top w_{P_n}(x) \right) && (17)\\
&= \max_{y \geq 0} \min_{\overline{w}^\beta_{P_n} \in W_{P_n}} \mathbb{E}_{(x,c)\sim P_n} \left(\left(\hat{c}_\theta(x) + y c(x) \right)^\top \overline{w}^\beta_{P_n}(x) \right) - y\beta - \min_{w_{P_n} \in W_{P_n}} \mathbb{E}_{(x,c)\sim P_n} \left(\hat{c}_\theta(x)^\top w_{P_n}(x) \right) . && (18)
\end{aligned}$$

In (18), we have switched the min and the max because of strong duality, since the objective function we are optimizing is linear. For a given $\theta \in \mathbb{R}^m$, we denote $y^\star(\theta)$ to be the optimal dual variable corresponding to θ in the minimization problem above. Let $\theta, \theta' \in \mathbb{R}^m$ and $D \geq \max(\left\|\theta\right\|, \left\|\theta'\right\|)$. We have

$$\left| \ell^\beta_{P_n}(\theta) - \ell^\beta_P(\theta) - \left(\ell^\beta_{P_n}(\theta') - \ell^\beta_P(\theta') \right) \right| \leq \left| \ell^\beta_{P_n}(\theta) - \ell^\beta_{P_n}(\theta') \right| + \left| \ell^\beta_P(\theta) - \ell^\beta_P(\theta') \right| .$$

We now bound the two terms above on the right side of the inequality. We have

$$\begin{aligned}
\left| \ell^\beta_{P_n}(\theta) - \ell^\beta_{P_n}(\theta') \right| = &\Bigg| \max_{y \geq 0} \min_{\overline{w}^\beta_{P_n} \in W_{P_n}} \mathbb{E}_{(x,c)\sim P_n} \left(\left(\hat{c}_\theta(x) + y c(x) \right)^\top \overline{w}^\beta_{P_n}(x) \right) - y\beta - \min_{w_{P_n} \in W_{P_n}} \mathbb{E}_{(x,c)\sim P_n} \left(\hat{c}_\theta(x)^\top w_{P_n}(x) \right) \\
&- \max_{y \geq 0} \min_{\overline{w}^\beta_{P_n} \in W_{P_n}} \mathbb{E}_{(x,c)\sim P_n} \left(\left(\hat{c}_{\theta'}(x) + y c(x) \right)^\top \overline{w}^\beta_{P_n}(x) \right) - y\beta - \min_{w_{P_n} \in W_{P_n}} \mathbb{E}_{(x,c)\sim P_n} \left(\hat{c}_{\theta'}(x)^\top w_{P_n}(x) \right) \Bigg| \\
= &\Bigg| \min_{\overline{w}^\beta_{P_n} \in W_{P_n}} \mathbb{E}_{(x,c)\sim P_n} \left(\left(\hat{c}_\theta(x) + y^\star(\theta) c(x) \right)^\top \overline{w}^\beta_{P_n}(x) \right) - y^\star(\theta)\beta \\
&- \min_{w_{P_n} \in W_{P_n}} \mathbb{E}_{(x,c)\sim P_n} \left(\hat{c}_\theta(x)^\top w_{P_n}(x) \right) \\
&- \min_{\overline{w}^\beta_{P_n} \in W_{P_n}} \mathbb{E}_{(x,c)\sim P_n} \left(\left(\hat{c}_{\theta'}(x) + y^\star(\theta') c(x) \right)^\top \overline{w}^\beta_{P_n}(x) \right) \\
&- y^\star(\theta')\beta - \min_{w_{P_n} \in W_{P_n}} \mathbb{E}_{(x,c)\sim P_n} \left(\hat{c}_{\theta'}(x)^\top w_{P_n}(x) \right) \Bigg| .
\end{aligned}$$

In the left hand side of the inequality above, we can see that if $\ell^\beta_{P_n}(\theta) \geq \ell^\beta_{P_n}(\theta')$, then replacing $y^\star(\theta')$ by $y^\star(\theta)$ gives a larger term than the one above, and when $\ell^\beta_{P_n}(\theta) \leq \ell^\beta_{P_n}(\theta')$, replacing $y^\star(\theta)$ by $y^\star(\theta')$ gives a larger term than the one above as well. Let $\tilde{y} \in \{y^\star(\theta), y^\star(\theta')\}$ such that replacing $y^\star(\theta)$ and $y^\star(\theta')$ by $\tilde{y}$ makes the expression above larger. We have, by replacing $y^\star(\theta)$ and $y^\star(\theta')$ by

$\tilde{y}$ in the inequality above, we get

$$\begin{aligned}\left|\ell^{\beta}_{P_n}(\theta) - \ell^{\beta}_{P_n}(\theta')\right| \leq \Bigg| &\min_{\overline{w}^{\beta}_{P_n} \in W_{P_n}} \mathbb{E}_{(x,c)\sim P_n} \left(\left(\hat{c}_{\theta}(x) + \tilde{y}c(x)\right)^{\top} \overline{w}^{\beta}_{P_n}(x)\right) - \min_{w_{P_n} \in W_{P_n}} \mathbb{E}_{(x,c)\sim P_n} \left(\hat{c}_{\theta}(x)^{\top} w_{P_n}(x)\right) \\ &- \min_{\overline{w}^{\beta}_{P_n} \in W_{P_n}} \mathbb{E}_{(x,c)\sim P_n} \left(\left(\hat{c}_{\theta'}(x) + \tilde{y}c(x)\right)^{\top} \overline{w}^{\beta}_{P_n}(x)\right) + \min_{w_{P_n} \in W_{P_n}} \mathbb{E}_{(x,c)\sim P_n} \left(\hat{c}_{\theta'}(x)^{\top} w_{P_n}(x)\right)\Bigg| \\ = \Bigg| &\mathbb{E}_{(x,c)\sim P_n} \left(\min_{w\in W} \left(\hat{c}_{\theta}(x) + \tilde{y}c(x)\right)^{\top} w\right) - \mathbb{E}_{(x,c)\sim P_n} \left(\min_{w\in W} \hat{c}_{\theta}(x)^{\top} w\right) \\ &- \mathbb{E}_{(x,c)\sim P_n} \left(\min_{w\in W} \left(\hat{c}_{\theta'}(x) + \tilde{y}c(x)\right)^{\top} w\right) + \mathbb{E}_{(x,c)\sim P_n} \left(\min_{w\in W} \hat{c}_{\theta'}(x)^{\top} w\right)\Bigg| \\ = &\left|f_{\tilde{y},P_n}(\theta) - f_{\tilde{y},P_n}(\theta')\right|,\end{aligned}$$

where for every $\theta \in \mathbb{R}^m$,

$$f_{\tilde{y},P_n}(\theta) = \mathbb{E}_{(x,c)\sim P_n} \left(\min_{w\in W} \overbrace{\left(\hat{c}_{\theta}(x) + \tilde{y}c(x)\right)^{\top} w}^{h_1(\theta,w,x)}\right) - \mathbb{E}_{(x,c)\sim P_n} \left(\min_{w\in W} \overbrace{\hat{c}_{\theta}(x)^{\top} w}^{h_2(\theta,w,x)}\right)$$

W is a convex bounded set, and for all $x \in \mathbb{R}^k$ and the two functions $w \in \mathbb{R}^d$, $\theta \longmapsto h_1(\theta, w, x)$, $\theta \longmapsto h_2(\theta, w, x)$ are both differentiable with respect to θ. Moreover, $\frac{\partial h_1}{\partial \theta} \longmapsto \nabla \hat{c}_{\theta}(x)^{\top} w$ and $\frac{\partial h_2}{\partial \theta} \longmapsto \nabla \hat{c}_{\theta}(x)^{\top} w$ (where $\nabla \hat{c}_{\theta}(x)$ is the jacobian of $\theta \longmapsto \hat{c}_{\theta}(x)$ at θ) are both continuous with respect to w for all $\theta \in \mathbb{R}^m$ and $x \in \mathbb{R}^k$. Hence, using Danskin's theorem, we can say that for all $x \in \mathbb{R}^k$, $\overline{h}_1(\theta, x) = \min_{w\in W} h_1(\theta, w, x)$ and $\overline{h}_2(\theta, x) := \min_{w\in W} h_2(\theta, w, x)$ are both subdifferentiable, and that for all $\theta \in \mathbb{R}^m$, $\partial \overline{h}_1(\theta, x) = \mathrm{conv}\{\nabla \hat{c}_{\theta}(x)^{\top} w,\ w \in \arg\min_{w\in W} h_1(\theta, w, x)\}$, $\partial \overline{h}_2(\theta, x) = \mathrm{conv}\{\nabla \hat{c}_{\theta}(x)^{\top} w,\ w \in \arg\min_{w\in W} h_2(\theta, w, x)\}$. These two sets are bounded by $B_W B_{\Phi}$ because of Assumption 6. This means that the two functions in the equality above are both $B_W B_{\Phi}$ lipschitz, which makes $f_{\tilde{y},P_n}$ $2B_W B_{\Phi}$ lipschitz.

We can also easily prove that the inequality above is also true when we replace P_n by P. Hence, we can deduce that

$$\begin{aligned}\left|\ell^{\beta}_{P_n}(\theta) - \ell^{\beta}_{P}(\theta) - \left(\ell^{\beta}_{P_n}(\theta') - \ell^{\beta}_{P}(\theta')\right)\right| &\leq \left|\ell^{\beta}_{P_n}(\theta) - \ell^{\beta}_{P_n}(\theta')\right| + \left|\ell^{\beta}_{P}(\theta) - \ell^{\beta}_{P}(\theta')\right| \\ &\leq \left|f_{\tilde{y},P_n}(\theta) - f_{\tilde{y},P_n}(\theta')\right| + \left|f_{\tilde{y},P}(\theta) - f_{\tilde{y},P}(\theta')\right| \\ &\leq 4B_W B_{\Phi} \left\|\theta - \theta'\right\|,\end{aligned}$$

which yields the desired result of step 1.

Step 2. We will now prove the desired generalization bound by bounding the generalization gap locally in balls covering the set which represents the possible values of $\theta^{\star}$. Let $\mathcal{B}$ be a set of balls of radius $\gamma > 0$ (for the norm $\|.\|$) such that $\bigcup_{B\in\mathcal{B}} B = B_{\|.\|}(0, D)$. According to Wainwright (2019), it is possible to choose $\mathcal{B}$ such that $\log|\mathcal{B}| \leq d \log\left(1 + \frac{2D}{\gamma}\right)$. For every $B \in \mathcal{B}$, let θ_B be an element of B. For a given $\varepsilon > 0$ we would like to bound the probability $\mathbb{P}\left(\left|\ell^{\beta}_{P_n}(\theta) - \ell^{\beta}_{P}(\theta)\right| > \varepsilon\right)$ using Hoeffding's inequality. In order to do that, we first notice that using Lagrangian duality, we get

$$\begin{aligned}\ell^{\beta}_{P}(\theta) - \ell^{\beta}_{P_n}(\theta) = &\max_{y\geq 0} \min_{\overline{w}^{\beta}_{P} \in W_{P}} \mathbb{E}_{(x,c)\sim P} \left(\left(\hat{c}_{\theta}(x) + yc(x)\right)^{\top} \overline{w}^{\beta}_{P}(x)\right) - y\beta - \min_{w_{P} \in W_{P}} \mathbb{E}_{(x,c)\sim P} \left(\hat{c}_{\theta}(x)^{\top} w_{P}(x)\right) \\ &- \max_{y\geq 0} \min_{\overline{w}^{\beta}_{P_n} \in W_{P_n}} \mathbb{E}_{(x,c)\sim P_n} \left(\left(\hat{c}_{\theta}(x) + yc(x)\right)^{\top} \overline{w}^{\beta}_{P_n}(x)\right) - y\beta - \min_{w_{P_n} \in W_{P_n}} \mathbb{E}_{(x,c)\sim P_n} \left(\hat{c}_{\theta}(x)^{\top} w_{P_n}(x)\right) \\ = &\min_{\overline{w}^{\beta}_{P} \in W_{P}} \mathbb{E}_{(x,c)\sim P} \left(\left(\hat{c}_{\theta}(x) + y^{\star}_{P}(\theta)c(x)\right)^{\top} \overline{w}^{\beta}_{P}(x)\right) - y^{\star}_{P}(\theta)\beta - \min_{w_{P} \in W_{P}} \mathbb{E}_{(x,c)\sim P} \left(\hat{c}_{\theta}(x)^{\top} w_{P}(x)\right) \\ &- \Bigg(\min_{\overline{w}^{\beta}_{P_n} \in W_{P_n}} \mathbb{E}_{(x,c)\sim P_n} \left(\left(\hat{c}_{\theta}(x) + y^{\star}_{P_n}(\theta)c(x)\right)^{\top} \overline{w}^{\beta}_{P_n}(x)\right) - y^{\star}_{P_n}(\theta)\beta \\ &- \min_{w_{P_n} \in W_{P_n}} \mathbb{E}_{(x,c)\sim P_n} \left(\hat{c}_{\theta}(x)^{\top} w_{P_n}(x)\right)\Bigg),\end{aligned}$$

where $y^\star_P(\theta)$ and $y^*_{P_n}(\theta)$ are respectively optimal dual variables in the bottom and top line above. As we have seen before, replacing $y^*_{P_n}(\theta)$ by $y^\star_P(\theta)$ makes the above term larger. Hence, we have

$$\begin{aligned}
\ell^\beta_P(\theta) - \ell^\beta_{P_n}(\theta) &\leq \min_{\overline{w}^\beta_P \in W_P} \mathbb{E}_{(x,c)\sim P}\left(\left(\hat{c}_\theta(x) + y^\star_P(\theta)c(x)\right)^\top \overline{w}^\beta_P(x)\right) - \min_{w_P \in W_P} \mathbb{E}_{(x,c)\sim P}\left(\hat{c}_\theta(x)^\top w_P(x)\right) \\
&\quad - \left(\min_{\overline{w}^\beta_{P_n} \in W_{P_n}} \mathbb{E}_{(x,c)\sim P_n}\left(\left(\hat{c}_\theta(x) + y^\star_P(\theta)c(x)\right)^\top \overline{w}^\beta_{P_n}(x)\right) - \min_{w_{P_n} \in W_{P_n}} \mathbb{E}_{(x,c)\sim P_n}\left(\hat{c}_\theta(x)^\top w_{P_n}(x)\right)\right) \\
&= \underbrace{\mathbb{E}_{(x,c)\sim P}\left(\min_{w\in W}\left(\hat{c}_\theta(x) + y^\star_P(\theta)c(x)\right)^\top w - \min_{w\in W}\hat{c}_\theta(x)^\top w\right)}_{\tilde{\ell}^\beta_P(\theta)} \\
&\quad - \underbrace{\mathbb{E}_{(x,c)\sim P_n}\left(\min_{w\in W}\left(\hat{c}_\theta(x) + y^\star_P(\theta)c(x)\right)^\top w - \min_{w\in W}\hat{c}_\theta(x)^\top w\right)}_{\tilde{\ell}^\beta_{P_n}(\theta)}
\end{aligned}$$

Hence, we have $\mathbb{P}\left(\ell^\beta_P(\theta) - \ell^\beta_{P_n}(\theta) > \varepsilon\right) \leq \mathbb{P}\left(\tilde{\ell}^\beta_P(\theta) - \tilde{\ell}^\beta_{P_n}(\theta) > \varepsilon\right)$. This enables us to apply Hoeffding's inequality to the term on the right given that we bound the random variable

$$m(x,\theta) := \min_{w\in W}\left(\hat{c}_\theta(x) + y^\star_P(\theta)c(x)\right)^\top w - \min_{w\in W}\hat{c}_\theta(x)^\top w.$$

It is easy to see that using Cauchy-Schwartz inequality, we have

$$|m(x,\theta)| \leq \left(\|\hat{c}_\theta(x)\| + |y^\star_P(\theta)|\,\|c(x)\|\right) B_W + \|\hat{c}_\theta(x)\|\, B_W$$

In order to upper bound the right-hand side, we need to upper bound $y^\star_P(\theta)$. In order to do this, we prove that the Lagrange multiplier $y^\star_P(\theta)$ is bounded for any $\theta \in \mathbb{R}^m$. Indeed, we have

$$\min_{w_P\in W_P} \mathbb{E}_{(x,c)\sim P}(\hat{c}_\theta(x)^\top w_P(x)) \leq \min_{w_P\in W_P}\max_{y\geq 0} \mathbb{E}_{(x,c)\sim P}(\hat{c}_\theta(x)^\top w_P(x)) + y(\mathbb{E}_{(x,c)\sim P}(c(x)^\top w_P(x)) - \beta) \tag{19}$$

$$= \min_{w_P\in W_P} \mathbb{E}_{(x,c)\sim P}(\hat{c}_\theta(x)^\top w_P(x)) + y^\star_P(\theta)(\mathbb{E}_{(x,c)\sim P}(c(x)^\top w_P(x)) - \beta) \tag{20}$$

$$\leq \mathbb{E}_{(x,c)\sim P}(\hat{c}_\theta(x)^\top w(c(x))) + y^\star_P(\theta)(\mathbb{E}_{(x,c)\sim P}(c(x)^\top w(c(x))) - \beta). \tag{21}$$

Inequality (19) holds because the left-hand side is the evaluation of the right-hand side at $y = 0$. Hence,

$$\begin{aligned}
\min_{w_P\in W_P} \mathbb{E}_{(x,c)\sim P}(\hat{c}_\theta(x)^\top w_P(x)) &\leq \mathbb{E}_{(x,c)\sim P}(\hat{c}_\theta(x)^\top w(c(x))) + y^\star_P(\theta)(\mathbb{E}_{(x,c)\sim P}(c(x)^\top w(c(x))) - \beta) \\
&= \mathbb{E}_{(x,c)\sim P}(\hat{c}_\theta(x)^\top w(c(x))) + y^\star_P(\theta)(\beta_{\min,P} - \beta).
\end{aligned}$$

This yields

$$y^\star_P(\theta) \leq \frac{\mathbb{E}_{(x,c)\sim P}(\hat{c}_\theta(x)^\top w(c(x))) - \mathbb{E}_{(x,c)\sim P}(\hat{c}_\theta(x)^\top w(\hat{c}_\theta(x)))}{\beta - \beta_{\min,P}} \leq \frac{2DB_W B_\Phi}{\beta - \beta_{\min,P}}.$$

Hence, we get the following upper bound for the random variables we are working with

$$|m(x,\theta)| \leq 2DB_W B_\Phi + KB_W \frac{2DB_W B_\Phi}{\beta - \beta_{\min,P}} := U$$

we can apply Hoeffding's inequality, and say that

$$\mathbb{P}\left(\tilde{\ell}^\beta_P(\theta) - \tilde{\ell}^\beta_{P_n}(\theta) > \varepsilon\right) \leq 2e^{-\frac{n\varepsilon^2}{2U^2}}.$$

This yields

$$\mathbb{P}\left(\sup_{B\in\mathcal{B}}\left(\tilde{\ell}^\beta_P(\theta_B) - \tilde{\ell}^\beta_{P_n}(\theta_B)\right) > \varepsilon\right) \leq |\mathcal{B}|\, 2e^{-\frac{n\varepsilon^2}{2U^2}}.$$

By denoting $\delta := |\mathcal{B}|\, 2e^{-\frac{n\varepsilon^2}{2U^2}}$, we can say that with probability at least $1-\delta$, we have

$$\forall B \in \mathcal{B},\ \tilde{\ell}^{\beta}_{P}(\theta_B) - \tilde{\ell}^{\beta}_{P_n}(\theta_B) \le U\sqrt{2}\sqrt{\frac{\log|\mathcal{B}| + \log\frac{2}{\delta}}{n}},$$

which gives

$$\forall B \in \mathcal{B},\ \ell^{\beta}_{P}(\theta_B) - \ell^{\beta}_{P_n}(\theta_B) \le U\sqrt{2}\sqrt{\frac{\log|\mathcal{B}| + \log\frac{2}{\delta}}{n}} \le U\sqrt{2}\sqrt{\frac{d\log\left(1+\frac{2D}{\gamma}\right) + \log\frac{2}{\delta}}{n}}.$$

Now we have for every $\theta \in B_{\|.\|}(0, D)$, denoting $B \in \mathcal{B}$ such that $\theta \in B$,

$$\begin{aligned}
\ell^{\beta}_{P}(\theta) - \ell^{\beta}_{P_n}(\theta) &\le \ell^{\beta}_{P}(\theta_B) - \ell^{\beta}_{P_n}(\theta_B) + \left|\ell^{\beta}_{P_n}(\theta) - \ell^{\beta}_{P}(\theta) - \left(\ell^{\beta}_{P_n}(\theta_B) - \ell^{\beta}_{P}(\theta_B)\right)\right| && (22)\\
&\le U\sqrt{2}\sqrt{\frac{d\log\left(1+\frac{2D}{\gamma}\right) + \log\frac{2}{\delta}}{n}} + 4B_W B_\Phi \|\theta - \theta_B\| && (23)\\
&\le U\sqrt{2}\sqrt{\frac{d\log\left(1+\frac{2D}{\gamma}\right) + \log\frac{2}{\delta}}{n}} + 4B_W B_\Phi \gamma. && (24)
\end{aligned}$$

Inequality (24) is resulting from the Lipschitz inequality (14). Taking $\gamma = O\left(\sqrt{\frac{\log\frac{1}{\delta}}{n}}\right)$, we get

$$\ell^{\beta}_{P}(\theta) - \ell^{\beta}_{P_n}(\theta) \le O\left((U + 4B_W B_\Phi)\sqrt{\frac{\log\frac{1}{\delta}}{n}}\right).$$

Since we have $\ell^{\beta}_{P_n}(\theta^\star) \le \varepsilon$, we can replace in the inequality above θ by $\theta^\star$ and get

$$\ell^{\beta}_{P}(\theta^\star) \le \ell^{\beta}_{P_n}(\theta^\star) + O\left((U + 4B_W B_\Phi)\sqrt{\frac{\log\frac{1}{\delta}}{n}}\right) \le \varepsilon + O\left(\frac{1}{\beta - \beta_{\min,P}}\sqrt{\frac{\log\frac{1}{\delta}}{n}}\right).$$

□

A.13 Proof of Proposition 1

Proof. We denote for every $w \in W$ and $\theta \in \mathbb{R}^m$, $f(\theta; w) = -\hat{c}_\theta(x)^\top w$. We want to prove that for every $w \in W$, $\theta \longrightarrow f(\theta; w)$ is weakly convex. More precisely, we want to prove that there exists $\alpha \ge 0$ such that for every $w \in W$, $\theta \longmapsto f(\theta; w) + \alpha\|\theta\|^2$ is convex. We have for every $\theta_1, \theta_2 \in \mathbb{R}^m$, using assumptions 10 and 5,

$$\begin{aligned}
\|\nabla f(\theta_1; w) - \nabla f(\theta_2; w)\| &= \left\|\left(\nabla\hat{c}_{\theta_1}(x) - \nabla\hat{c}_{\theta_2}(x)\right)^\top w\right\|\\
&\le B_W \|\nabla\hat{c}_{\theta_1}(x) - \nabla\hat{c}_{\theta_2}(x)\|\\
&\le B_W B_L \|\theta_1 - \theta_2\|.
\end{aligned}$$

We denote for every $\theta \in \mathbb{R}^m$ and $w \in W$, $h(\theta; w) = f(\theta; w) + 2B_W B_L \|\theta\|^2$. We have for every $\theta_1, \theta_2 \in \mathbb{R}^m$,

$$\begin{aligned}
(\nabla h(\theta_1; w) - \nabla h(\theta_2; w))^\top(\theta_1 - \theta_2) &= (\nabla f(\theta_1; w) - \nabla f(\theta_2; w))^\top(\theta_1 - \theta_2) + 2B_W B_L \|x - y\|^2\\
&\ge -\|\nabla f(\theta_1; w) - \nabla f(\theta_2; w)\|\,\|\theta_1 - \theta_2\| + 2B_W B_L \|\theta_1 - \theta_2\|^2\\
&\ge -B_W B_L \|\theta_1 - \theta_2\|^2 + 2B_W B_L \|\theta_1 - \theta_2\|^2\\
&= B_W B_L \|\theta_1 - \theta_2\|^2 \ge 0.
\end{aligned}$$

Hence, for every $w \in W$, $\theta \longmapsto h(\theta; w)$ is convex. This implies that for any measurable mapping $w : \mathbb{R}^k \longrightarrow W$, $\theta \longmapsto \mathbb{E}_{(x,c)\sim P_n}\left(h(\theta; w(x))\right)$ is convex. Consequently, $\theta \longmapsto$

$\max_{w_{P_n} \in W_{P_n}} \mathbb{E}_{(x,c)\sim P_n}(h(\theta; w(x)))$ and $\theta \longmapsto \max_{w_{P_n} \in \overline{W}^{\beta}_{P_n}} \mathbb{E}_{(x,c)\sim P_n}(h(\theta; w(x)))$ are both maximums of a family of convex functions, and hence are also convex functions. Finally we have for every $\theta \in \mathbb{R}^m$,

$$\max_{w_{P_n} \in W_{P_n}} \mathbb{E}_{(x,c)\sim P_n}(h(\theta; w(x))) = g_{P_n}(\theta) + 2B_L \left\|\theta\right\|^2$$
$$\max_{w_{P_n} \in \overline{W}^{\beta}_{P_n}} \mathbb{E}_{(x,c)\sim P_n}(h(\theta; w(x))) = \overline{g}^{\beta}_{P_n}(\theta) + 2B_L \left\|\theta\right\|^2 .$$

Finally, we can write for every $\theta \in \mathbb{R}^m$,

$$\ell^{\beta}_{P_n}(\theta) = g_{P_n}(\theta) + 2B_L \left\|\theta\right\|^2 - \left(\overline{g}^{\beta}_{P_n}(\theta) + 2B_L \left\|\theta\right\|^2\right)$$

which is indeed a difference of convex functions. □

A.14 Proof of Theorem 9

Proof. Let $\theta \in \mathbb{R}^m$ and $t > 0$. Assumption 8 gives

$$\mathbb{P}(0 < \Delta_\theta(x) \leq \left\|\theta\right\| t) \leq \left(\frac{\gamma t}{B}\right)^{\alpha} .$$

where

$$\Delta_\theta(x) = \begin{cases} \min_{w \in W^{\angle} \setminus W^{\star}(\hat{c}_\theta(x))} \hat{c}_\theta(x)^\top w - \min_{w \in W^{\angle}} \hat{c}_\theta(x)^\top w & \text{if } W^{\star}(\hat{c}_\theta(x)) \neq W \\ 0 & \text{else.} \end{cases}$$

We have for $\theta \neq 0$, denoting $W^{\angle} = \{w_1, \ldots, w_N\}$, $N \in \mathbb{N}$,

$$\begin{aligned}
&\Delta_\theta(x) > \left\|\theta\right\| t \Longrightarrow \forall w_{\angle} \in W^{\angle} \setminus \{w(\hat{c}_\theta)\},\ \hat{c}_\theta(x)^\top w_{\angle} - \hat{c}_\theta(x)^\top w(\hat{c}_\theta(x)) > \left\|\theta\right\| t \\
&\Longrightarrow \forall w_{\angle} \in W^{\angle} \setminus \{w(\hat{c}_\theta)\},\ \hat{c}_\theta(x)^\top w_{\angle} - \hat{c}_\theta(x)^\top w(\hat{c}_\theta(x)) > \left\|\theta\right\| \frac{\left\|w(\hat{c}_\theta(x)) - w_{\angle}\right\|}{2B_W} t \\
&\Longrightarrow \forall \lambda_1, \ldots, \lambda_n \geq 0 \text{ s.t. } \lambda_1 + \cdots + \lambda_n = 1, \\
&\sum_{i=1}^{N} \lambda_i \left(\hat{c}_\theta(x)^\top w_i - \hat{c}_\theta(x)^\top w(\hat{c}_\theta(x))\right) > \frac{1}{2B_W} \left\|\theta\right\| t \sum_{i=1}^{N} \lambda_i \left\|w(\hat{c}_\theta(x)) - w_i\right\| \\
&\Longrightarrow \forall \lambda_1, \ldots, \lambda_n \geq 0 \text{ s.t. } \lambda_1 + \cdots + \lambda_n = 1, \\
&\hat{c}_\theta(x)^\top \sum_{i=1}^{N} \lambda_i w_i - \hat{c}_\theta(x)^\top w(\hat{c}_\theta(x)) > \frac{1}{2B_W} \left\|\theta\right\| t \left\|w(\hat{c}_\theta(x)) - \sum_{i=1}^{N} \lambda_i w_i\right\| \\
&\Longrightarrow \forall w \in W,\ \hat{c}_\theta(x)^\top w - \hat{c}_\theta(x)^\top w(\hat{c}_\theta(x)) > \frac{1}{2B_W} \left\|\theta\right\| \left\|w(\hat{c}_\theta(x)) - w\right\| t.
\end{aligned}$$

Hence, we have

$$\mathbb{P}\left(\forall w \in W,\ \hat{c}_\theta(x)^\top w - \hat{c}_\theta(x)^\top w(\hat{c}_\theta(x)) > \frac{1}{2B_W} \left\|\theta\right\| \left\|w(\hat{c}_\theta(x)) - w\right\| t\right) \geq \mathbb{P}\left(\Delta_\theta(x) > \left\|\theta\right\| t\right) > 1 - \left(\frac{\gamma t}{B}\right)^{\alpha} \tag{25}$$

We denote for $w \in W$, $t \geq 0$, and $\theta \in \mathbb{R}^m \setminus \{0\}$,

$$G(x, \theta, w, t) = \hat{c}_\theta(x)^\top w - \hat{c}_\theta(x)^\top w(\hat{c}_\theta(x)) - \frac{1}{2B_W} \left\|\theta\right\| \left\|w(\hat{c}_\theta(x)) - w\right\| t$$

and the event $A(x, \theta, t)$ defined by

$$A(x, \theta, t) = \{\forall w \in W,\ G(x, \theta, w(x), t) > 0\} .$$

Hence, we have for every mapping $w : \mathbb{R}^k \longrightarrow w$ and $\theta \in \mathbb{R}^m \setminus \{0\}$,

$$\begin{aligned}
\mathbb{E}_{(x,c)\sim P}(G(x, \theta, w(x), t)) &= \mathbb{E}_{(x,c)\sim P}(G(x, \theta, w(x), t) 1_{A(x,\theta,t)}) + \mathbb{E}_{(x,c)\sim P}(G(x, \theta, w(x), t) 1_{A(x,\theta,t)^c}) \\
&\geq \mathbb{E}_{(x,c)\sim P}(G(x, \theta, w(x), t) 1_{A(x,\theta,t)^c}).
\end{aligned}$$

Furthermore, we have for every $w \in W$

$$\begin{aligned}|G(x,\theta,w,t)| &= \left|\hat{c}_\theta(x)^\top w - \hat{c}_\theta(x)^\top w(\hat{c}_\theta(x)) - \frac{1}{2B_W}\|\theta\|\,\|w(\hat{c}_\theta(x)) - w\|\, t\right| \\ &\leq B_\Phi B_W \|\theta\| + B_\Phi B_W \|\theta\| + \|\theta\|\, t = \|\theta\|\,(2B_W B_\Phi + t)\,.\end{aligned}$$

Notice that (25) can be rewritten as $\mathbb{P}\left(A(x,\theta,t)\right) > 1 - \left(\frac{\gamma t}{B}\right)^\alpha$. Hence, we have

$$\begin{aligned}\mathbb{E}_{(x,c)\sim P}(G(x,\theta,w(x),t)) &\geq -(2B_W B_\Phi + t)\,\|\theta\|\,\mathbb{P}(A(x,\theta,t)^c) \\ &\geq -(2B_W B_\Phi + t)\,\|\theta\| \left(\frac{\gamma t}{B}\right)^\alpha.\end{aligned}$$

In conclusion, we have for every $t \geq 0$ and every $\theta \in \mathbb{R}^m$, and mapping $w : \mathbb{R}^k \longrightarrow W$,

$$\mathbb{E}_{(x,c)\sim P}\left(\hat{c}_\theta(x)^\top w(x) - \hat{c}_\theta(x)^\top w(\hat{c}_\theta(x))\right) \geq \frac{1}{2B_W}\|\theta\|\,\mathbb{E}_{(x,c)\sim P}\left(\|w(\hat{c}_\theta(x)) - w(x)\|\right) t - (2B_W B_\Phi + t)\,\|\theta\|\left(\frac{\gamma t}{B}\right)^\alpha.$$

the inequality above is trivially verified when $\theta = 0$. Also, when taking $w \in \arg\min_{w \in \overline{W}_P^\beta} \mathbb{E}_{(x,c)\sim P}\left(\hat{c}_\theta(x)^\top w(x)\right)$, we get

$$\begin{aligned}\ell_P^\beta(\theta) &\geq \frac{1}{2B_W}\|\theta\|\,\mathbb{E}_{(x,c)\sim P}\left(\|w(\hat{c}_\theta(x)) - w(x)\|\right) t - (2B_W B_\Phi + t)\,\|\theta\|\left(\frac{\gamma t}{B}\right)^\alpha \\ &\geq \frac{\|\theta\|\, t}{2KB_W}(\ell_P(\theta) - \beta) - (2B_W B_\Phi + t)\,\|\theta\|\left(\frac{\gamma t}{B}\right)^\alpha\end{aligned}$$

Let $\varepsilon > 0$. Assume that θ is bounded away from 0. We have

$$\ell_P(\theta) - \beta \leq \frac{2KB_W}{t}\left(\frac{\ell_P^\beta(\theta)}{\|\theta\|} + (2B_W B_\Phi + t)\left(\frac{\gamma t}{B}\right)^\alpha\right) \leq O\left(\frac{1}{t}\left(\ell_P^\beta\,(\theta) + t^\alpha\right)\right)$$

To optimize the right hand side and ignoring the constant, we take $t = \ell_P^\beta\,(\theta)^{\frac{1}{1+\alpha}}$. This gives

$$\ell_P(\theta) \leq \beta + O\left(\ell_P^\beta\,(\theta)^{1 - \frac{1}{1+\alpha}}\right).$$

□

A.15 A more complete version of Theorem 12

Theorem 14.

1. *For all $\theta \in \mathbb{R}^m$, $\overline{g}_{P_n}^\beta$ and g_{P_n} have a non-empty subgradient at θ, and $\ell_{P_n}^\beta(\theta) = \theta^\top v(\theta)$ for some $v(\theta) \in \partial\ell_{P_n}^\beta := \partial\overline{g}_{P_n}^\beta(\theta) - \partial g_{P_n}^\beta(\theta) := \{u_\beta - u,\ (u_\beta, u) \in \partial\overline{g}_{P_n}^\beta(\theta) \times \partial g_{P_n}(\theta)\}$, where $\partial\overline{g}_{P_n}^\beta(\theta)$ and $\partial g_{P_n}(\theta)$ are respectively the subgradients of $\overline{g}_{P_n}^\beta$ and g_{P_n} at θ;*

2. *For all $\theta \in \mathbb{R}^m$, if $v(\theta) = 0$ for some $v(\theta) \in \partial\ell_{P_n}^\beta(\theta)$, then $\ell_{P_n}^\beta(\theta) = 0$, hence θ is a minimizer.*

3. *$r_{P_n}^\beta$ is everywhere differentiable. If λ is a stationary point of $r_{P_n}^\beta$ such that $\hat{c}_{\overline{\theta}_{P_n}^\beta(\lambda)}(x) \neq 0$ almost surely, then $\overline{\theta}_{P_n}^\beta(\lambda)$ is a global minimum of $\ell_{P_n}^\beta$. In particular, under assumptions 5 and 6, we have for every $\theta \in \mathbb{R}^m$ and $\varepsilon \geq 0$,*

$$\left\|\nabla r_{P_n}^\beta\right\| \leq \varepsilon \Longrightarrow \ell_{P_n}^\beta(\theta) \leq 8B_W B_\Phi \varepsilon.$$

Proof. 1. We denote for all $(\theta, v) \in \mathbb{R}^m \times V_P$, $\phi(\theta, v) = \theta^\top v$, v_β and for all $(\theta, v_\beta) \in \mathbb{R}^m \times V_P$, $\phi(\theta, v_\beta) = \theta^\top v_\beta$, In this case, for all $\theta \in \mathbb{R}^m$, the loss writes as

$$\ell_P^\beta(\theta) = g_P(\theta) - \overline{g}_P^\beta(\theta) = \min_{v_\beta \in V_P^\beta}\phi(\theta, v_\beta) - \min_{v \in V_P}\phi(\theta, v).$$

Let $\mathcal{V}^\beta_P(\theta)$ and $\mathcal{V}_P(\theta)$ be the set of minimizers of respectively $v_\beta \longmapsto \phi(\theta, v_\beta)$ over V^β_P and $v \longmapsto \phi(\theta, v)$ over V_P. Given that V_P and V^β_P are compact sets, and $\theta \longmapsto \phi(\theta, v)$ for all $v \in V_P \cup V^\beta_P$ is differentiable, and ϕ is continuous, we can say by Danskin's theorem that for all $\theta \in \mathbb{R}^m$,

$$\partial g_P(\theta) = \text{conv}\left\{\frac{\partial\phi(\theta, v)}{\partial\theta},\ v \in \mathcal{V}_P(\theta)\right\} = \text{conv}\ \mathcal{V}_P(\theta) \underset{(*)}{=} \mathcal{V}_P(\theta),$$

$$\partial\overline{g}^\beta_P(\theta) = \text{conv}\left\{\frac{\partial\phi(\theta, v_\beta)}{\partial\theta},\ v_\beta \in \mathcal{V}^\beta_P(\theta)\right\} = \text{conv}\ \mathcal{V}^\beta_P(\theta) \underset{(**)}{=} \mathcal{V}^\beta_P(\theta).$$

The two inequalities $(*)$ and $(**)$ are due to the fact that $\mathcal{V}_P(\theta)$ and $\mathcal{V}^\beta_P(\theta)$ are convex sets. Furthermore, we have for all $\theta \in \mathbb{R}^m$,

$$\ell^\beta_P(\theta) = \min_{v_\beta \in V^\beta_P} \phi(\theta, v_\beta) - \min_{v \in V_P} \phi(\theta, v) = \theta^\top v^\star_\beta - \theta^\top v^\star = \theta^\top \underbrace{\left(v^\star_\beta - v^\star\right)}_{:=v(\theta)},$$

where $(v^\star_\beta, v^\star) \in \mathcal{V}^\beta_P(\theta) \times \mathcal{V}_P(\theta)$. All of the above clearly yields

$$v(\theta) \in \mathcal{V}^\beta_P(\theta) - \mathcal{V}_P(\theta) = \partial\overline{g}^\beta_P(\theta) - \partial g_P(\theta),$$

which is the result we were seeking to prove.

2. From the above, it is clear that for a given $\theta \in \mathbb{R}^m$ if $v(\theta) = 0$, then $\ell^\beta_P(\theta) = 0$. Furthermore, using the very first definition of ℓ^β_P, since $\overline{W}^\beta_P \subset W_P$, we have for all $\theta \in \mathbb{R}^m$,

$$\ell^\beta_P(\theta) = \min_{w_P \in \overline{W}^\beta_P} \mathbb{E}_{(c,x)\sim P}\left((\hat{c}_\theta(x)^\top w_P(x))\right) - \min_{w_P \in W_P} \mathbb{E}_{(c,x)\sim P}\left(\hat{c}_\theta(x)^\top w_P(x)\right) \geq 0.$$

In conclusion, 0 is a lower bound of ℓ^β_P, and if for a given θ, $v(\theta) = 0$, then $\ell^\beta_P(\theta) = 0$, i.e. θ is a minimizer of ℓ^β_P.

3. r^β_P is the difference of the Moreau envelope of two convex functions, hence it is the difference between two differentiable convex functions. this yields that r^β_P is differentiable. Furthermore, if λ is a stationary point of r^β_P then using proposition 1 from Sun and Sun (2021), $\theta^\beta_P(\lambda)$ is indeed a stationary point of ℓ^β_P. Furthermore, we have for every $\varepsilon \geq 0$, We consider $\lambda \in \mathbb{R}^m$ an ε-stationary point of $r^\beta_{P_n}$, i.e. $\left\|\nabla r^\beta_{P_n}(\lambda)\right\| \leq \varepsilon$. Using Danskin's theorem, we get

$$\begin{aligned}\nabla r^\beta_{P_n}(\lambda) &= \nabla M_{P_n}(\lambda) - \nabla\overline{M}^\beta_{P_n}(\lambda)\\ &= \lambda - \theta_{P_n}(\lambda) - (\lambda - \overline{\theta}^\beta_{P_n}(\lambda))\\ &= \overline{\theta}^\beta_{P_n}(\lambda) - \theta_{P_n}(\lambda),\end{aligned}$$

Where $\overline{\theta}^\beta_{P_n}(\lambda) = \arg\min_{\theta\in\mathbb{R}^m} g^\beta_{P_n}(\theta) + \frac{1}{2}\|\lambda - \theta\|^2$ and $\theta_{P_n}(\lambda) = \arg\min_{\theta\in\mathbb{R}^m} g_{P_n}(\theta) + \frac{1}{2}\|\lambda - \theta\|^2$. By looking at the proof of Lemma 1, we can easily see that we also have $\theta_{P_n}(\lambda) = \lambda + v_{P_n}(\theta_{P_n}(\lambda))$ where $v_{P_n}(\theta_{P_n}(\lambda)) \in \arg\min_{v\in V_{P_n}} \theta_{P_n}(\lambda)^\top v = \partial g_{P_n}(\theta_{P_n}(\lambda))$ and $\overline{\theta}^\beta_{P_n}(\lambda) = \lambda + v^\beta_{P_n}(\overline{\theta}^\beta_{P_n}(\lambda))$ where $v^\beta_{P_n}(\overline{\theta}^\beta_{P_n}(\lambda)) \in \arg\min_{v\in\overline{V}^\beta_{P_n}} \overline{\theta}^\beta_{P_n}(\lambda)^\top v = \partial g^\beta_{P_n}(\overline{\theta}^\beta_{P_n}(\lambda))$. Let

$v_{P_n}(\overline{\theta}^{\beta}_{P_n}(\lambda)) \in \arg\min_{v \in V_{P_n}} \overline{\theta}^{\beta}_{P_n}(\lambda)^\top v = \partial g_{P_n}(\overline{\theta}^{\beta}_{P_n}(\lambda))$. We have

$$\begin{aligned}
\ell^{\beta}_{P_n}(\overline{\theta}^{\beta}_{P_n}(\lambda)) &= \overline{\theta}^{\beta}_{P_n}(\lambda)^\top v^{\beta}_{P_n}(\overline{\theta}^{\beta}_{P_n}(\lambda)) - \overline{\theta}^{\beta}_{P_n}(\lambda) v_{P_n}(\overline{\theta}^{\beta}_{P_n}(\lambda)) && (26)\\
&\leq \left|\overline{\theta}^{\beta}_{P_n}(\lambda)^\top v^{\beta}_{P_n}(\overline{\theta}^{\beta}_{P_n}(\lambda)) - \overline{\theta}^{\beta}_{P_n}(\lambda)^\top v_{P_n}(\theta_{P_n}(\lambda))\right| && (27)\\
&+ \left|\theta_{P_n}(\lambda)^\top v_{P_n}(\theta_{P_n}(\lambda)) - \overline{\theta}^{\beta}_{P_n}(\lambda)^\top v_{P_n}(\overline{\theta}^{\beta}_{P_n}(\lambda))\right| && (28)\\
&+ \left|\theta_{P_n}(\lambda)^\top v_{P_n}(\theta_{P_n}(\lambda)) - \overline{\theta}^{\beta}_{P_n}(\lambda)^\top v_{P_n}(\theta_{P_n}(\lambda))\right| && (29)\\
&\leq \left\|\overline{\theta}^{\beta}_{P_n}(\lambda)\right\| \left\|v^{\beta}_{P_n}(\overline{\theta}^{\beta}_{P_n}(\lambda)) - v_{P_n}(\theta_{P_n}(\lambda))\right\| + \left|\kappa(\theta^{\beta}_{P_n}(\lambda)) - \kappa(\theta_{P_n}(\lambda))\right| && (30)\\
&+ \left\|v_{P_n}(\theta_{P_n}(\lambda))\right\| \left\|\theta_{P_n}(\lambda) - \overline{\theta}^{\beta}_{P_n}(\lambda)\right\| && (31)\\
&\leq (2B_W B_\Phi + 5B_W B_\Phi)\varepsilon + \left|\kappa(\theta^{\beta}_{P_n}(\lambda)) - \kappa(\theta_{P_n}(\lambda))\right| && (32)\\
&\leq (3B_W B_\Phi + 5B_W B_\Phi)\varepsilon = 8B_W B_\Phi \varepsilon && (33)
\end{aligned}$$

Here, κ is defined for any $\theta \in \mathbb{R}^m$ as $\kappa(\theta) = \min_{v \in V_{P_n}} \theta^\top v$. The last equality 33 is due to the fact that the subgradient of κ at some $\theta \in \mathbb{R}^m$ is in V_{P_n} (which can be proven thanks to Danskin's theorem) and is hence bounded by $B_W B_\Phi$. We can see that inequality 33 is indeed the inequality we were seeking to obtain. Taking $\varepsilon = 0$ gives that stationarity for $r^{\beta}_{P_n}$ implies global optimality for $\ell^{\beta}_{P_n}$.

□

A.16 Key lemma to prove Proposition 2

Lemma 1. *For every $\lambda \in \mathbb{R}^m$ and $\beta \geq \beta^{\star}_{\mathcal{H},P}$, we denote $V^1_{P_n} = V_{P_n}$ and $V^2_{P_n} = \overline{V}^{\beta}_{P_n}$. We have for $i \in \{1, 2\}$, $\theta^i_{P_n}(\lambda) = \lambda + v_i$ with $v_i = \arg\min\limits_{v \in V^i_{P_n}} \dfrac{1}{2}\|\lambda + v\|^2$ and $M^i_{P_n}(\lambda) = \frac{1}{2}\|\lambda\|^2 - \min\limits_{v \in V^i_{P_n}} \dfrac{1}{2}\|\lambda + v\|^2$.*

Proof. The main idea of the proof is to switch min and max in the definitions of M^{β}_{P} and $\overline{M}^{\beta}_{P}$.

1. We have for all $\lambda \in \mathbb{R}^m$,

$$\begin{aligned}
M_P(\lambda) &= \min_{\theta \in \mathbb{R}^m} g^{\beta}_{P}(\theta) + \frac{1}{2}\|\lambda - \theta\|^2 && (34)\\
&= \min_{\theta \in \mathbb{R}^m} \left(- \min_{w_P \in W_P} \mathbb{E}(\hat{c}_\theta(x)^\top w_P(x))\right) + \frac{1}{2}\|\lambda - \theta\|^2 && (35)\\
&= \min_{\theta \in \mathbb{R}^m} \max_{w_P \in W_P} -\mathbb{E}(\hat{c}_\theta(x)^\top w_P(x)) + \frac{1}{2}\|\lambda - \theta\|^2 && (36)\\
&= \min_{\theta \in \mathbb{R}^m} \max_{w_P \in W_P} -\theta^\top \underbrace{\mathbb{E}(\Phi(x)^\top w_P(x))}_{v} + \frac{1}{2}\|\lambda - \theta\|^2 && (37)\\
&= \min_{\theta \in \mathbb{R}^m} \max_{v \in V_P} -\theta^\top v + \frac{1}{2}\|\lambda - \theta\|^2 && (38)\\
&= \max_{v \in V_P} \min_{\theta \in \mathbb{R}^m} -\theta^\top v + \frac{1}{2}\|\lambda - \theta\|^2 && (39)
\end{aligned}$$

The equality between 39 and 38 holds because of Sion's minimax theorem: since $\mathbb{R}^m$ and V_P are convex sets, $v \longmapsto -\theta^\top v$ is upper semi-continuous and concave for any $\theta \in \mathbb{R}^m$, and $\theta \longmapsto -\theta^\top v + \frac{1}{2}\|\lambda - \theta\|^2$ is lower semicontinuous and convex for any $v \in V_P$, we can switch min and max. In 39, the minimum is reached when the gradient with respect to θ is zero, i.e. $-v+\theta-\lambda=0$, which gives $\theta=v+\lambda$. In this case, we get

$$\begin{aligned} M_P(\lambda) &= \max_{v\in V_P} -(v+\lambda)^\top v + \frac{1}{2}\|v\|^2 \\ &= \max_{v\in V_P} -\lambda^\top v - \frac{1}{2}\|v\|^2 \\ &= \frac{1}{2}\|\lambda\|^2 + \max_{v\in V_P} -\frac{1}{2}\|\lambda+v\|^2 \\ &= \frac{1}{2}\|\lambda\|^2 - \min_{v\in V_P} \frac{1}{2}\|\lambda+v\|^2 . \end{aligned}$$

The calculations above also give us immediately that $\theta_P(\lambda) = \lambda + v$ where $v = \arg\min_{v\in V_P} \frac{1}{2}\|\lambda+v\|^2$.

2. The proof of the second property is almost identical to the previous one. It suffices to replace V_P by $\overline{V}_P^\beta$ and W_P by $\overline{W}_P^\beta$.

□

A.17 Proof of Proposition 2

This proposition immediately follows from Lemma 1.

A.18 More details on avoiding $\theta = 0$

Theorem 15. *We define $F_{V_{P_n}}(\lambda) := \arg\min_{\lambda'\in -V_{P_n}} \|\lambda-\lambda'\|$ and $F_{\overline{V}^\beta_{P_n}}(\lambda) := \arg\min_{\lambda'\in -\overline{V}^\beta_{P_n}} \|\lambda-\lambda'\|$ to be respectively the projection of λ on $-V_{P_n}$ and on $-\overline{V}^\beta_{P_n}$. Suppose that λ is an ε-solution of $f^\beta_{P_n}$, i.e. $\left\|\nabla f^\beta_{P_n}(\lambda)\right\| \le \varepsilon$ for some $\varepsilon > 0$, and $\lambda \notin V^\beta_{P_n}$. We have $\left\|\nabla r^\beta_{P_n}\right\| \le 27B_W^2B_\Phi^2\varepsilon$. Furthermore, if $d(\lambda, -V_{P_n}) \le B_W B_\Phi$, letting $\tilde\lambda = \lambda + r\frac{\lambda - F_{V_{P_n}}(\lambda)}{\|\lambda - F_{V_{P_n}}(\lambda)\|}$ with $r \in [3B_WB_\Phi, 8B_WB_\Phi]$ and $\varepsilon' = 27B_W^2B_\Phi^2\varepsilon$, we have*

- *The value of $r^\beta_{P_n}(\tilde\lambda)$ is dominated by ε. In particular, we have $r^\beta_{P_n}(\tilde\lambda) \le 11B_WB_\Phi\varepsilon'$ and $\left\|\nabla r^\beta_{P_n}(\tilde\lambda)\right\| \le \sqrt{11B_WB_\Phi\varepsilon'}$.*
- *The norm of the resulting candidate solution for $\ell^\beta_{P_n}$ is bounded from above and below. In particular, $B_WB_\Phi \le \|\bar\theta^\beta_{P_n}(\tilde\lambda)\| \le 11B_WB_\Phi$.*

Figure 5 summarizes the construction of $\tilde\lambda$.

The theorem above provides us with an algorithm to optimize $\ell^\beta_{P_n}$:

1. Run gradient descent on $r^\beta_{P_n}$.
2. If we obtain a solution reasonably far from V_{P_n}, we simply return that iterate as an output. Else, we run gradient descent on $f^\beta_{P_n}$.
3. If the gradient descent iterates do not converge to a common boundary to $-V_{P_n}$ and $-\overline{V}^\beta_{P_n}$, we return the final iterate as an output. Else, we use the procedure in Theorem 15 to obtain a near stationary point which is far from the boundary, and return it as an output.

Proof. We will sequentially prove the two points. We assume that $\lambda \in \mathbb{R}^m$ is a stationary point of $f^{\beta}_{P_n}$. In this case, we have

$$0 = \nabla f^{\beta}_{P_n}(\lambda) = \frac{\lambda - F_{\overline{V}^{\beta}_{P_n}}(\lambda)}{\left\|\lambda - F_{\overline{V}^{\beta}_{P_n}}(\lambda)\right\|^2} - \frac{\lambda - F_{V_{P_n}}(\lambda)}{\left\|\lambda - F_{V_{P_n}}(\lambda)\right\|^2}.$$

This yields

$$\frac{\lambda - F_{\overline{V}^{\beta}_{P_n}}(\lambda)}{\left\|\lambda - F_{\overline{V}^{\beta}_{P_n}}(\lambda)\right\|^2} = \frac{\lambda - F_{V_{P_n}}(\lambda)}{\left\|\lambda - F_{V_{P_n}}(\lambda)\right\|^2}.$$

Taking the norm in both sides, we get $\left\|\lambda - F_{\overline{V}^{\beta}_{P_n}}(\lambda)\right\| = \left\|\lambda - F_{V_{P_n}}(\lambda)\right\|$, and finally, replugging this in the equality above, we directly get $F_{\overline{V}^{\beta}_{P_n}}(\lambda) = F_{V_{P_n}}(\lambda)$, i.e. $\nabla r^{\beta}_{P_n}(\lambda) = 0$. Let us now prove the second result. We denote $D^{\beta}_n = \lambda - F_{\overline{V}^{\beta}_{P_n}}(\lambda)$ and $D_n = \lambda - F_{V_{P_n}}(\lambda)$. We assume now that $\left\|f^{\beta}_{P_n}(\lambda)\right\| \leq \varepsilon$. This inequality gives us two inequalities

$$\left|\frac{1}{\|D_n\|} - \frac{1}{\left\|D^{\beta}_n\right\|}\right| \leq \varepsilon \text{ and } \left\|\frac{D_n}{\|D_n\|^2} - \frac{D^{\beta}_n}{\left\|D^{\beta}_n\right\|^2}\right\| \leq \varepsilon.$$

Hence, we have

$$\begin{aligned}
\left\|\nabla r^{\beta}_{P_n}\right\| &= \left\|F_{\overline{V}^{\beta}_{P_n}}(\lambda) - F_{V_{P_n}}(\lambda)\right\| \\
&= \left\|D^{\beta}_n - D_n\right\| \\
&= \|D_n\|^2 \left\|\frac{D_n}{\|D_n\|^2} - \frac{D^{\beta}_n}{\|D_n\|^2}\right\| \\
&\leq \|D_n\|^2 \left(\left\|\frac{D_n}{\|D_n\|^2} - \frac{D^{\beta}_n}{\left\|D^{\beta}_n\right\|^2}\right\| + \left\|D^{\beta}_n\right\| \left|\frac{1}{\|D_n\|^2} - \frac{1}{\left\|D^{\beta}_n\right\|^2}\right|\right) \\
&\leq 9B_W^2 B_\Phi^2 \varepsilon + 9B_W^2 B_\Phi^2 \left(1 + \frac{\|D_n\|}{\left\|D^{\beta}_n\right\|}\right) \left|\frac{1}{\|D_n\|} - \frac{1}{\left\|D^{\beta}_n\right\|}\right| \\
&\leq 27B_W^2 B_\Phi^2 \varepsilon.
\end{aligned}$$

Let $\lambda \in \mathbb{R}^m$ such that $\left\|\nabla r^{\beta}_{P_n}(\lambda)\right\| \leq \varepsilon$, i.e.

$$\left\|F_{\overline{V}^{\beta}_{P_n}}(\lambda) - F_{V_{P_n}}(\lambda)\right\| \leq \varepsilon. \tag{40}$$

We have

$$\begin{aligned}
\left\|\tilde{\lambda} - F_{\overline{V}^{\beta}_{P_n}}(\tilde{\lambda})\right\| &\leq \left\|\tilde{\lambda} - F_{\overline{V}^{\beta}_{P_n}}(\lambda)\right\| && (41) \\
&\leq \left\|\tilde{\lambda} - F_{V_{P_n}}(\lambda)\right\| + \left\|F_{V_{P_n}}(\lambda) - F_{\overline{V}^{\beta}_{P_n}}(\lambda)\right\| && (42) \\
&\leq \left\|\tilde{\lambda} - F_{V_{P_n}}(\tilde{\lambda})\right\| + \varepsilon. && (43)
\end{aligned}$$

The last inequality is due to inequality 40 and $F_{V_{P_n}}(\tilde{\lambda}) = F_{\overline{V}^{\beta}_{P_n}}(\lambda)$. We finish by proving this last equality. We have for every $v \in -V_{P_n}$,

$$\left\|\tilde{\lambda} - v\right\|^2 = \left\|\lambda + r\frac{\lambda - F_{V_{P_n}}(\lambda)}{\|\lambda - F_{V_{P_n}}(\lambda)\|} - v\right\|^2 \tag{44}$$

$$= \|\lambda - v\|^2 + r^2 + 2r\left\langle \lambda - v,\ \frac{\lambda - F_{V_{P_n}}(\lambda)}{\|\lambda - F_{V_{P_n}}(\lambda)\|}\right\rangle \tag{45}$$

$$\geq \|\lambda - v\|^2 + r^2. \tag{46}$$

Notice that when the left-hand side in 46 is minimized (i.e. $v = F_{V_{P_n}}(\lambda)$, the right-hand side takes the same value. Hence, $v = F_{V_{P_n}}(\lambda)$ is also a minimizer of the left-hand side, i.e. $F_{V_{P_n}}(\tilde{\lambda}) = F_{\overline{V}^{\beta}_{P_n}}(\lambda)$. Finally, inequality 43 yields

$$\left\|\tilde{\lambda} - F_{\overline{V}^{\beta}_{P_n}}(\tilde{\lambda})\right\| - \left\|\tilde{\lambda} - F_{V_{P_n}}(\tilde{\lambda})\right\| \leq \varepsilon,$$

Which gives

$$\begin{aligned}
r^{\beta}_{P_n}(\tilde{\lambda}) &= \frac{1}{2}\left\|\tilde{\lambda} - F_{\overline{V}^{\beta}_{P_n}}(\tilde{\lambda})\right\|^2 - \frac{1}{2}\left\|\tilde{\lambda} - F_{V_{P_n}}(\tilde{\lambda})\right\|^2 \\
&= \frac{1}{2}\left(\left\|\tilde{\lambda} - F_{\overline{V}^{\beta}_{P_n}}(\tilde{\lambda})\right\| - \left\|\tilde{\lambda} - F_{V_{P_n}}(\tilde{\lambda})\right\|\right)\left(\left\|\tilde{\lambda} - F_{\overline{V}^{\beta}_{P_n}}(\tilde{\lambda})\right\| + \left\|\tilde{\lambda} - F_{V_{P_n}}(\tilde{\lambda})\right\|\right) \\
&\leq \frac{1}{2}\left(\left\|\tilde{\lambda} - F_{\overline{V}^{\beta}_{P_n}}(\tilde{\lambda})\right\| + \left\|\tilde{\lambda} - F_{V_{P_n}}(\tilde{\lambda})\right\|\right)\varepsilon \\
&\leq \frac{1}{2}(6B_W B_\Phi + 2r)\varepsilon \leq 11 B_W B_\Phi \varepsilon.
\end{aligned}$$

Using this, we also prove the bound on the gradient of $r^{\beta}_{P_n}$. Using the properties of the projection $F_{V_{P_n}}(\tilde{\lambda})$, we have

$$\begin{aligned}
\left\|\nabla r^{\beta}_{P_n}(\tilde{\lambda})\right\| &= \sqrt{\left\|F_{\overline{V}^{\beta}_{P_n}}(\tilde{\lambda}) - F_{V_{P_n}}(\tilde{\lambda})\right\|^2} \\
&= \sqrt{\left\|F_{\overline{V}^{\beta}_{P_n}}(\tilde{\lambda}) - \tilde{\lambda}\right\|^2 - \left\|F_{V_{P_n}}(\tilde{\lambda}) - \tilde{\lambda}\right\|^2 - 2\underbrace{\left\langle F_{\overline{V}^{\beta}_{P_n}}(\tilde{\lambda}) - F_{V_{P_n}}(\tilde{\lambda}),\ F_{V_{P_n}}(\tilde{\lambda}) - \tilde{\lambda}\right\rangle}_{\geq 0}} \\
&\leq \sqrt{r^{\beta}_{P_n}(\tilde{\lambda})} \leq \sqrt{11 B_W B_\Phi \varepsilon}.
\end{aligned}$$

We finish by proving the bound on $\overline{\theta}^{\beta}_{P_n}(\tilde{\lambda})$. Using Lemma 2, we have $\overline{\theta}^{\beta}_{P_n}(\tilde{\lambda}) = \tilde{\lambda} + F_{\overline{V}^{\beta}_{P_n}}(\tilde{\lambda})$, hence,

$$B_W B_\Phi \leq \left\|\overline{\theta}^{\beta}_{P_n}(\tilde{\lambda})\right\| \leq 11 B_W B_\Phi.$$

□

A.19 Proof of Proposition 3

Proof. Let $\{\lambda_i\}$ be the sequence of iterates generated by applying gradient descent to $f^{\beta}_{P_n}$. Suppose a subsequence $\{\lambda_{i_t}\}$ of this sequence converges to a limit point $\lambda^\star$. If $\lambda^\star$ is not on the boundary of $-V_{P_n}$ or $-\overline{V}^{\beta}_{P_n}$, then by Nocedal and Wright (1999), $\lambda^\star$ is a stationary point of $f^{\beta}_{P_n}$. We now show that if the limit point $\lambda^\star$ lies on the boundary of $-V_{P_n}$, then it must also lie on the boundary of $-\overline{V}^{\beta}_{P_n}$. Indeed, on the one hand, the objective function $f^{\beta}_{P_n}(\lambda_{i_t})$ is non-increasing, so $f^{\beta}_{P_n}(\lambda_{i_t}) \leq f^{\beta}_{P_n}(\lambda_0)$ for all t. On the other hand, $-\log(d(\lambda^{i_t}, -V_{P_n})) \to \infty$, which implies that $-\log(d(\lambda^{i_t}, -\overline{V}^{\beta}_{P_n})) \to \infty$. Therefore, $d(\lambda^\star, -\overline{V}^{\beta}_{P_n}) \to 0$, yielding the desired result. □

A.20 Experimental setting

To do line search for β, we used $\beta_{\text{min},P} = \mathbb{E}_{(x,c)\sim P_n}\left(c^\top w(c)\right)$ as a lower bound to β, and $\beta_{\text{SPO+}} = \mathbb{E}_{(x,c)\sim P_n}\left(c^\top w\left(\hat{c}_{\theta^\star_{\text{SPO+}}}(x)\right)\right)$ where $\theta^\star_{\text{SPO+}}$ is the solution obtained by optimizing the SPO+ loss. For every value of s, we tested 96 evenly spaced values of β in the interval $[\beta_{\text{min},P}, \beta_{\text{SPO+}}]$, and picked β yielding the solution with the best decision performance. The optimal β is likely to fall into (or near) the interval $[\beta_{\text{min},P}, \beta_{\text{SPO+}}]$, and hence for at least one of the values of β that we test, ℓ_P^β enjoys good optimality guarantees, which ensures that the solution we obtain yields a small value for ℓ_P.

We set $(d, j) = (20, 5)$ and W to be a polyhedron and written as $W = \{w \in \mathbb{R}^d,\ Aw = b,\ 10 \geq w \geq 0\}$ where $A \in \mathbb{R}^{j\times d}$ ($j \leq d$) and $b \in \mathbb{R}^j$. W is closed and bounded, so Assumption 5 holds. In every experiment, we sample $x \sim \mathcal{N}(0, I)$ while all of its coordinates are conditioned to be between 0 and 10. and the coefficients of A from a standard normal Gaussian distribution, and b to be equal to $A\,|w|$, where w has standard normal random coefficients and $|w|$ is the vector whose coordinates are equal to the absolute value of the coordinates of w.

To create a framework where we can enforce misspecification and progressively analyze how the performance of different methods varies with the level of misspecification $\gamma_{\text{dec}}(\mathcal{H})$, we need to define a model that allows for easy adjustment of the misspecification level while still yielding meaningful results. In order to do so, we define for $x \in \mathbb{R}^k$, $c(x)$ to be a linear combination of polynomial functions of x. This could be written in a compact way as $c(x) = M\phi(x)$ where M is a real valued matrix, and $\phi(x)$ is a vector whose coordinates are polynomial in x, defined by

$$\forall x \in \mathbb{R}^k,\ \phi(x) = \begin{pmatrix} x_1 & \dots & x_k & x_1x_2 & \dots & x_1\dots x_k \end{pmatrix}^\top = \left(\prod_{i=1}^{k} x_i^{y_i}\right)_{y\in\{0,1\}^5, y\neq 0}.$$

Furthermore, we define $\hat{c}_\theta(x)$ for $x \in \mathbb{R}^k$ and $\theta \in \mathbb{R}^m$ as $\hat{c}_\theta(x) = M'(\theta)\tilde{\phi}(x)$ where $\tilde{\phi}(x)$ is equal to $\phi(x)$ truncated to its 5 first coordinates, and $M'(\theta)$ is a matrix representation of θ. To increase the level of misspecification, increase the size of x and consequently the gap between $\phi(x)$ and $\tilde{\phi}(x)$ in terms of number of features. To show that this experimental setting indeed corresponds to the theoretical setting studied in this paper, we can see that both the elements of the hypothesis set and the ground truth can be written for all $x \in \mathbb{R}^k$ as $\hat{c}_\theta(x) = \tilde{\Phi}(x)\theta$ and $c(x) = \Phi(x)\theta^\star$, $\theta^\star \in \mathbb{R}^m$, where

$$\tilde{\Phi}(x) = \begin{pmatrix} \tilde{\phi}(x)^\top & & 0 \\ & \ddots & \\ 0 & & \tilde{\phi}(x)^\top \end{pmatrix} \in \mathbb{R}^{20\times 100},\quad \Phi(x) = \begin{pmatrix} \phi(x)^\top & & 0 \\ & \ddots & \\ 0 & & \phi(x)^\top \end{pmatrix} \in \mathbb{R}^{20\times 20s}.$$

Furthermore, for any matrix M, we denote $L_r(M)$ the $r-$th row of M. In this case, we have $\theta^\star = \begin{pmatrix} L_1(M) & \dots & L_{20}(M) \end{pmatrix}^\top \in \mathbb{R}^{20s}$ and $\theta = \begin{pmatrix} L_1(M'(\theta)) & \dots & L_{20}(M'(\theta)) \end{pmatrix}^\top \in \mathbb{R}^{100}$.

B Additional Figures

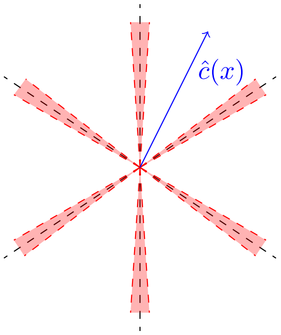

Figure 4: Less likely directions for $\hat{c}_\theta(x)$: $\hat{c}_\theta(x)$ is less likely to have a direction included in the red cones, which represent nearly perpendicular directions to one of the faces of the polyhedron (see Assumption 8).

C Plot with relative regret

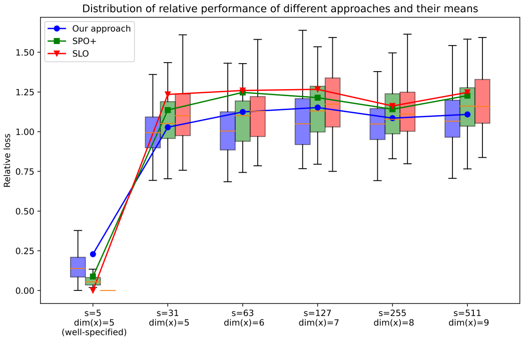

Figure 5: Experiment with relative regret results