Generalization Bounds in the Predict-then-Optimize Framework

Othman El Balghiti

Department of Industrial Engineering and Operations Research, Columbia University, New York, NY, 10027,
oe2161@columbia.edu

Adam N. Elmachtoub

Department of Industrial Engineering and Operations Research and Data Science Institute, Columbia University, New York,
NY, 10027, adam@ieor.columbia.edu

Paul Grigas

Department of Industrial Engineering and Operations Research, UC Berkeley, Berkeley, CA, 94720, pgrigas@berkeley.edu

Ambuj Tewari

Department of Statistics, University of Michigan, Ann Arbor, MI, 48109, tewaria@umich.edu

The predict-then-optimize framework is fundamental in many practical settings: predict the unknown parameters of an optimization problem, and then solve the problem using the predicted values of the parameters. A natural loss function in this environment is to consider the cost of the decisions induced by the predicted parameters, in contrast to the prediction error of the parameters. This loss function was recently introduced in Elmachtoub and Grigas (2022) and referred to as the Smart Predict-then-Optimize (SPO) loss. In this work, we seek to provide bounds on how well the performance of a prediction model fit on training data generalizes out-of-sample, in the context of the SPO loss. Since the SPO loss is non-convex and non-Lipschitz, standard results for deriving generalization bounds do not apply.

We first derive bounds based on the Natarajan dimension that, in the case of a polyhedral feasible region, scale at most logarithmically in the number of extreme points, but, in the case of a general convex feasible region, have linear dependence on the decision dimension. By exploiting the structure of the SPO loss function and a key property of the feasible region, which we denote as the strength property, we can dramatically improve the dependence on the decision and feature dimensions. Our approach and analysis rely on placing a margin around problematic predictions that do not yield unique optimal solutions, and then providing generalization bounds in the context of a modified margin SPO loss function that is Lipschitz continuous. Finally, we characterize the strength property and show that the modified SPO loss can be computed efficiently for both strongly convex bodies and polytopes with an explicit extreme point representation.

Key words: generalization bounds; prescriptive analytics; regression; predict-then-optimize

1. Introduction

A common application of machine learning is to *predict-then-optimize*, i.e., predict unknown parameters of an optimization problem and then solve the optimization problem using the predictions.

For instance, consider a navigation task that requires solving a shortest path problem. The key inputs into this problem are the travel times on each edge, typically called edge costs. Although the exact costs are not known at the time the problem is solved, the edge costs are predicted using a machine learning model trained on training data consisting of features (time of day, weather, etc.) and edge costs (collected from app data). Fundamentally, a good prediction model induces the optimization problem to find good shortest paths, as measured by the true edge costs. In fact, recent work has been developed to consider how to solve problems in similar environments (Bertsimas and Kallus 2020, Kao et al. 2009, Donti et al. 2017). In particular, Elmachtoub and Grigas (2022) developed the Smart Predict-then-Optimize (SPO) loss function for the case of optimization with an unknown linear objective subject to a convex feasible region, also known as contextual linear optimization. The SPO loss exactly measures the quality of a prediction by the decision error, in contrast to the prediction error as measured by standard loss functions such as mean squared error. A key property of a loss function is to show that the generalization error is small, i.e., that the loss observed in-sample is close to what we would observe out-of-sample. A generalization bound is an upper bound on the generalization error that holds with high probability, and should be relatively small to justify the use of minimizing the empirical risk on the training data. A generalization bound also implies a convergence rate on the number of samples required to find the best prediction model within a given hypothesis class. In this work, we seek to provide an assortment of novel and strong generalization bounds for the SPO loss function.

Specifically, we shall assume that our optimization task is to minimize a linear objective over a feasible region, and that this task can be completed efficiently in the practical sense. This captures a wide range of optimization problems including linear programming, mixed-integer linear programming, and a large class of nonlinear programs. In the shortest path example, the feasible region is a polyhedron. We assume the objective cost vector is not known at the time that the optimization problem is solved, but rather predicted from a prediction model that maps from features to cost vectors. A decision is made with respect to the predicted cost vector, and the SPO loss is computed by evaluating the decision on the true cost vector and then subtracting the optimal cost assuming knowledge of the true cost vector. Unfortunately, the SPO loss is non-convex and non-Lipschitz, and therefore proving generalization bounds is not immediate.

Our results consider two approaches, and in all cases achieve bounds that are $\tilde{O}(\frac{1}{\sqrt{n}})$, where n is the number of samples. Each approach uses the classic result of Bartlett and Mendelson (2002) (Theorem 1), which allows us to bound the generalization error by using the Rademacher complexity of the hypothesis class composed with predict-then-optimize models. In the first and more straightforward approach, we treat the optimization task as a generalization of multiclass classification (see Example 3). When the feasible region is a polyhedron, this resorts to treating every

extreme point as a class. The final bound depends on the square root of the Natarajan dimension of the hypothesis class, and the square root of the log of the number of extreme points (Theorem 2). The Natarajan dimension essentially measures how rich the hypothesis class of prediction models is, and in the case of linear models it is exactly the number of parameters pd, where p is the feature dimension and d is the decision dimension (Corollary 2). For a general convex body, we transform the optimization task into a multiclass classification problem by discretizing the feasible region with a ball covering that only results in a loss of $O(\frac{1}{n})$ (Corollary 3). The ball covering can be exponential in d, and the final generalization bound can thus depend on $d\sqrt{p}$ in the case of a linear hypothesis class.

Fortunately, we show that for a large class of feasible regions, tighter generalization bounds can be obtained with improved dependence on p and d, using margin-based methods. The key idea of our approach is to reduce the analysis of the Rademacher complexity of the SPO framework to a multivariate Rademacher complexity of the vector-valued hypothesis class, for which we can obtain improved bounds in many cases. This reduction relies on a vector contraction inequality that only works for Lipschitz loss functions. Since SPO loss is not Lipschitz, we transform it to a Lipschitz loss function by using a margin-based analysis that bounds away predictions that yield multiple optimal solutions (and cause discontinuities in the SPO loss). We define a property of the feasible region which we call the strength property, and show that our margin-based loss function is Lipschitz if the strength property is satisfied (Theorem 3). In turn, this allows to provide generalization bounds with respect to the margin-based loss, the strength parameter, and with the more amenable Rademacher complexity of the hypothesis class (Theorem 4). We also show the margin parameter can be cross-validated by showing the generalization bound holds uniformly over an interval of margin parameters (Theorem 5). We show that for suitably constrained linear hypothesis classes, we get a much improved dependence on problem dimensions (Theorem 6). Finally, we show the strength property and how to compute the margin-based loss for strongly convex sets (Theorem 7) and polyhedra with an explicit extreme point representation (Theorem 8).

1.1. Relation to Previous Work

We first summarize a few additional approaches proposed in the literature which for solving predict-then-optimize problems. Kao et al. (2009) propose a loss function for training linear regression models which minimizes a convex combination between the prediction error and decision error of an unconstrained quadratic optimization. Donti et al. (2017) provide a more general methodology that allows for constraints which relies on differentiating the optimization problem using a technique from Amos and Kolter (2017). Wilder et al. (2019a), Mandi et al. (2020), Ferber et al. (2020), Wilder et al. (2019b), Pogančić et al. (2019), Mandi and Guns (2020), Berthet et al. (2020) consider

training ML models using "decision-focused" loss functions for various linear and combinatorial optimization problems; their methods do not attempt to minimize SPO loss directly but rather employ differentiable surrogate loss functions. Elmachtoub and Grigas (2022) propose a convex surrogate loss funcion called SPO+ loss which is shown to be Fisher consistent with SPO loss. Loke et al. (2021) consider a form of regularization that generalizes the SPO+ loss function. Demirovic et al. (2019, 2020) propose algorithms for training linear regression models to directly minimize SPO loss, but their methodology is specialized for ranking optimization problems and dynamic programming problems, respectively. Ho-Nguyen and Kılınç-Karzan (2022) study the relationship between the mean squared error loss and the SPO loss. See Kotary et al. (2021) for a recent survey of more approaches. The SPO loss has been used in applications such as last mile delivery (Chu et al. 2021) and ship inspection (Yan et al. 2020). Elmachtoub et al. (2020) and Kallus and Mao (2020) train decision trees and random forests to directly minimize decision error in contextual linear and nonlinear optimization problems, respectively.

A preliminary version of this paper appeared in El Balghiti et al. (2019) that did not include our general framework for margin-based bounds using the strength property. Our results have found use in Hu et al. (2022) which slightly improved Theorem 2 by a factor of $\sqrt{\log n}$. They also show - in settings where there is no model misspecification (realizability) - how to obtain regret rates that can be faster than $O(\frac{1}{\sqrt{n}})$ when using SPO loss and least squares loss. Wang et al. (2020) uses our Theorem 2 to derive generalization bounds on compact surrogate loss functions. Generalization bounds for related types of decision-making problems have been derived in Ban and Rudin (2019), Ho and Hanasusanto (2019), Bertsimas and Kallus (2020).

As previously mentioned, the SPO framework is a generalization of multi-class classification, and thus our results have some connections to results in this literature stream. In fact, our results can be seen as extending the classic margin-based generalization bounds for classification (Koltchinskii and Panchenko 2002, Koltchinskii et al. 2001, Mohri et al. 2018) to the predict-then-optimize framework. Note that the reduction of a predict-then-optimize problem to multi-class classification throws away potentially important information, namely the numerical values of the cost vectors. Therefore a naive application of results in the classification literature would not lead to bounds on our loss function of interest, the SPO loss. Furthermore, our approaches remove a strong explicit dependency on the number of extreme points (classes) by exploiting the structure of the SPO loss. Our data-independent bounds grow only logarithmically in the number of extreme points. In contrast, for data-independent worst-case bounds in multiclass classification, the dependency is at best square root in the number of classes (Guermeur 2007, Daniely et al. 2015). Our data-dependent margin based bounds remove an explicit dependency on the number of extreme points entirely. Since the seminal work of Koltchinskii and Panchenko (2002), much work in the classification

literature has focused on improving the dependency on the number of classes for nonparameteric hypothesis classes such as multi-class kernels. For example, by using data-dependent (margin-based) approaches and local Rademacher complexity analysis, Lei et al. (2015) and Li et al. (2018) obtain logarithmic complexity in the number of classes for such hypothesis classes.

2. Predict-then-Optimize Framework

We now formally describe the predict-then-optimize framework, which is central to many applications of machine learning for use in operations research problems. We are given a nominal optimization problem of interest which models a downstream decision-making task. We assume that the nominal problem has a linear objective and that the decision variable $w \in \mathbb{R}^d$ and feasible region $S \subseteq \mathbb{R}^d$ are well-defined and known with certainty. We assume S is a nonempty, compact, and convex set. However, the cost vector of the objective, $c \in \mathbb{R}^d$, is not observed directly, and rather an associated feature vector $x \in \mathbb{R}^p$ is observed. Let $\mathcal{D}$ be the underlying joint distribution of (x, c) and let $\mathcal{D}_x$ be the conditional distribution of c given x. Then, the goal for the decision maker is to solve the contextual stochastic optimization problem

$$\min_{w \in S} \mathbb{E}_{c \sim \mathcal{D}_x}[c^T w | x] \; = \; \min_{w \in S} \mathbb{E}_{c \sim \mathcal{D}_x}[c | x]^T w \; . \tag{1}$$

The predict-then-optimize framework relies on using a prediction (estimate) for $\mathbb{E}_{c \sim \mathcal{D}_x}[c | x]$, which we denote by $\hat{c}$, and solving the deterministic version of the optimization problem based on $\hat{c}$. From (1), it is clear that the linearity of the objective makes the predict-then-optimize approach viable, where as a nonlinear objective may require an alternate approach such as that developed by Bertsimas and Kallus (2020). Note that knowledge of the distribution $\mathcal{D}_x$ is typically unavailable, thus necessitating a data-driven approach to solving (1).

We define $P(\hat{c})$ to be the optimization task with objective cost vector $\hat{c}$, namely

$$P(\hat{c}): \quad \begin{array}{l} \min_{w} \hat{c}^T w \\ \text{s.t. } w \in S. \end{array} \tag{2}$$

Since we are focusing on linear optimization problems, the assumption that S is convex is without loss of generality as we can take S to be the convex hull of any potentially non-convex feasible region. We let $w^*(\cdot) : \mathbb{R}^d \to S$ denote any oracle for solving $P(\cdot)$. That is, $w^*(\cdot)$ is a fixed deterministic mapping such that $w^*(c) \in \arg\min_{w \in S} \{c^T w\}$ for all $c \in \mathbb{R}^d$. For instance, if (2) corresponds to a linear, conic, or mixed-integer optimization problem (in which case S can be implicitly described as a convex set), then a commercial optimization solver or a specialized algorithm suffices for $w^*(\cdot)$.

In this framework, we assume that predictions are made from a model that is learned on a training data set. Specifically, we are provided sample training data $(x_1, c_1), \ldots, (x_n, c_n)$ drawn i.i.d. from the joint distribution $\mathcal{D}$, where $x_i \in \mathcal{X} \subseteq \mathbb{R}^p$ is a feature vector representing auxiliary

information associated with the cost vector $c_i \in \mathcal{C} \subseteq \mathbb{R}^d$. We denote by $\mathcal{H}$ our hypothesis class of cost vector prediction models. Thus, for a function $f \in \mathcal{H}$, we have that $f : \mathcal{X} \to \mathbb{R}^d$. Most approaches for learning a model $f \in \mathcal{H}$ from the training data are based on specifying a loss function that quantifies the error in making prediction $\hat{c}$ when the realized (true) cost vector is actually c. Following prior work of Elmachtoub and Grigas (2022), our primary loss function of interest is the Smart Predict-then-Optimize (SPO) loss function that directly takes the nominal optimization problem $P(\cdot)$ into account when measuring errors in predictions. Namely, we consider the SPO loss function defined by:

$$\ell_{\mathrm{SPO}}(\hat{c}, c) := c^T w^*(\hat{c}) - c^T w^*(c) \ ,$$

where $\hat{c}$ is the predicted cost vector and c is the true realized cost vector. The SPO loss can be described as the decision error of making the decision based on $\hat{c}$ when the actual cost vector was c, evaluated on the actual cost vector c. The SPO loss is always non-negative and is exactly zero when the induced decisions from $\hat{c}$ and c are equal, even if $\hat{c}$ and c themselves are not necessarily close (which is traditionally the goal in prediction). Note that, primarily for simplicity of presentation, we consider the version of the SPO loss defined relative to the specific and fixed optimization oracle $w^*(\cdot)$ (Definition 1 of Elmachtoub and Grigas (2022)) rather than the unambiguous variant that does not depend on a specific optimization oracle (Definition 2 of Elmachtoub and Grigas (2022)).

Given a fixed sample $(x_1, c_1), \ldots, (x_n, c_n)$, we define the empirical risk with respect to the SPO loss of a function $f \in \mathcal{H}$ as

$$\hat{R}_{\mathrm{SPO}}(f) := \frac{1}{n} \sum_{i=1}^{n} \ell_{\mathrm{SPO}}(f(x_i), c_i) \ ,$$

and the expected (or Bayes) risk as

$$R_{\mathrm{SPO}}(f) := \mathbb{E}_{(x,c) \sim \mathcal{D}}[\ell_{\mathrm{SPO}}(f(x), c)].$$

One can interpret the expected risk as the out-of-sample performance of f, and the empirical risk as the in-sample performance. The empirical risk minimization principle states that we should determine a prediction model $\hat{f}_n \in \mathcal{H}$ by solving the optimization problem

$$\min_{f \in \mathcal{H}} \ \hat{R}_{\mathrm{SPO}}(f) \ . \tag{3}$$

Empirical risk minimization with respect to the SPO loss is not the only way to train a prediction model. Indeed, as pointed out in Elmachtoub and Grigas (2022), the SPO loss function is generally non-convex, may even be discontinuous, and is in fact a strict generalization of the 0-1 loss function in binary classification. Thus, solving (3) to optimally is intractable even when $\mathcal{H}$ is a linear hypothesis class. To circumvent these difficulties, alternative approaches include optimizing

a convex surrogate loss such as the SPO+ loss presented in (Elmachtoub and Grigas 2022) or the squared ℓ_2 loss, using methods exploiting structural properties of $\mathcal{H}$ (Elmachtoub et al. 2020), or modifying $P(\cdot)$ and finding local optimal solutions (Wilder et al. 2019a). Most generally, one may use any training procedure that determines a model in the hypothesis class $\mathcal{H}$ and therefore it is important that generalization bounds apply *uniformly* for all functions in $\mathcal{H}$. Our focus herein is on deriving generalization bounds that hold uniformly over a given hypothesis class $\mathcal{H}$ and thus are valid for *any* training approach, including the use of a surrogate or alternate loss function within the framework of empirical risk minimization. We also note that a generalization bound for the SPO loss directly translates to an upper bound guarantee for problem (1) that holds "on average" over the distribution.

Provided with any prediction model $\hat{f}$ trained on the given data, we can now simply express the predict-then-optimize approach to solve (1); when presented with a feature vector x, select the optimal solution with respect to the predicted cost vector, i.e., the decision made is $w^*(\hat{f}(x))$. Ideally, in order to select see if $\hat{f}$ is a good model, we would evaluate the expected future real-world performance of $\hat{f}$, i.e., the decision-maker would evaluate the expected risk $R_{\mathrm{SPO}}(\hat{f})$. Unfortunately, as the true distribution is unknown, one can only estimate this quantity using data and the most natural estimate is the empirical risk $\hat{R}_{\mathrm{SPO}}(\hat{f})$. Generalization bounds, which bound the gap between the empirical and expected risk with high probability, are thus critical as they validate that $\hat{R}_{\mathrm{SPO}}(\hat{f})$ is a reasonable estimate of $R_{\mathrm{SPO}}(\hat{f})$. Strong validation bounds validate the usage of empirical risk minimization and other model selection ideas that estimate the empirical risk on the training data. Indeed, one may use a (two-sided) generalization bound to construct a confidence interval for $R_{\mathrm{SPO}}(\hat{f})$, for example. At a more cursory level, generalization bounds reveal how the difference between the expected and empirical risks scale with key parameters such as the sample size and dimensions of the feature and decisions spaces.

To conclude this subsection, we present several important examples to illustrate the applicability and generality of the SPO loss function and framework.

EXAMPLE 1 (SHORTEST PATH). In the shortest path problem, the feature vector x may include features such as weather, speed limit, and time data that may be used to predict the cost vector c representing the travel times along each edge of the network. In this case, the network is assumed to be given (e.g., the road network of a city) and the feasible region S is a network flow polytope that represents flow conservation and capacity constraints on the underlying network. More generally, the class of vehicle routing problems with unknown edge costs falls in our framework, although the feasible region is more complex as integer constraints are typically required. □

EXAMPLE 2 (PORTFOLIO OPTIMIZATION). In portfolio optimization, the returns of potential investments can depend on many features which typically include historical returns, news, economic

factors, social media, and others. We presume that these features may be used to predict the vector of returns different assets, but that the covariance matrix of the asset returns does not depend on the features. Maximizing the expected return is a linear objective, and with convex constraints that the total amount invested is in the simplex and the overall variance of the portfolio is bounded. □

EXAMPLE 3 (MULTICLASS CLASSIFICATION). Our setting also captures multi-class (and binary) classification by the following characterization: S is the d-dimensional unit simplex, namely $S := \{w \in \mathbb{R}^d : \sum_{j=1}^d w_j = 1, w \geq 0\}$, where d is the number of classes, and $\mathcal{C} = \{-e_i | i = 1, \ldots, d\}$ where e_i is the ith unit vector in $\mathbb{R}^d$. It is easy to see that each vertex of the simplex corresponds to a class, and correct/incorrect classification has a loss of 0/1. Specifically, $\ell_{\text{SPO}}(\hat{c}, c) = 0$ if $w^*(\hat{c}) = -c$ and $\ell_{\text{SPO}}(\hat{c}, c) = 1$ if $w^*(\hat{c}) \neq -c$. □

2.1. Other Notation

We make use of a generic given norm $\| \cdot \|$ on $w \in \mathbb{R}^d$, as well as the ℓ_q-norm denoted by $\| \cdot \|_q$ for $q \in [1, \infty]$. For the given norm $\| \cdot \|$ on $\mathbb{R}^d$, $\| \cdot \|_*$ denotes the dual norm defined by $\|c\|_* := \max_{w:\|w\| \leq 1} c^T w$. Let $B(\bar{w}, r) := \{w : \|w - \bar{w}\| \leq r\}$ denote the ball of radius r centered at $\bar{w}$, and we analogously define $B_q(\bar{w}, r)$ for the ℓ_q-norm and $B_*(c, r)$ for the dual norm. For a set $W \subseteq \mathbb{R}^d$, let $\text{dist}_W(\bar{w}) := \inf_{w \in W} \{\|w - \bar{w}\|\}$ be the distance from the point $\bar{w} \in \mathbb{R}^d$ to W measured in the norm $\| \cdot \|$. Likewise, for a set $C \subseteq \mathbb{R}^d$, define $\text{dist}_C^*(\bar{c}) := \inf_{c \in C} \{\|c - \bar{c}\|_*\}$ which is measured in the dual norm. For a set $S \subseteq \mathbb{R}^d$, we define the size of S in the norm $\| \cdot \|$ by $\rho(S) := \sup_{w \in S} \|w\|$, and the diameter of S by $\Delta(S) := \sup_{w_1, w_2 \in S} \|w_1 - w_2\|$. We analogously define $\rho_q(\cdot)$ for the ℓ_q-norm and $\rho_*(\cdot)$ for the dual norm. We define the "linear optimization gap" of S with respect to c by $\omega_S(c) := \max_{w \in S} \{c^T w\} - \min_{w \in S} \{c^T w\}$, and for a set $\mathcal{C} \subseteq \mathbb{R}^d$ we slightly abuse notation by defining $\omega_S(\mathcal{C}) := \sup_{c \in \mathcal{C}} \omega_S(c)$. Note that the $\ell_{\text{SPO}}(\hat{c}, c)$ is non-negative and bounded above by $\omega_S(\mathcal{C})$ for all $\hat{c} \in \mathbb{R}^d$ and $c \in \mathcal{C}$. Finally, we define the class of predict-then-optimize functions by $w^*(\mathcal{H}) := \{x \mapsto w^*(f(x)) : f \in \mathcal{H}\}$.

2.2. Rademacher Complexity and Generalization Bounds

Let us now briefly review the notion of Rademacher complexity and its application in our framework. We define the empirical Rademacher complexity of $\mathcal{H}$ with respect to the SPO loss, i.e., the empirical Rademacher complexity of the function class obtained by composing ℓ_{SPO} with $\mathcal{H}$, by

$$\hat{\mathfrak{R}}_{\text{SPO}}^n(\mathcal{H}) := \mathbb{E}_\sigma \left[\sup_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^n \sigma_i \ell_{\text{SPO}}(f(x_i), c_i) \right],$$

where $\sigma_1, \ldots, \sigma_n$ are i.i.d. Rademacher random variables, i.e., $\mathbb{P}(\sigma_i = 1) = \mathbb{P}(\sigma_i = -1) = \frac{1}{2}$. The expected version of the Rademacher complexity is defined as $\mathfrak{R}_{\text{SPO}}^n(\mathcal{H}) := \mathbb{E}\left[\hat{\mathfrak{R}}_{\text{SPO}}^n(\mathcal{H})\right]$ where the

expectation is w.r.t. an i.i.d. sample drawn from the underlying distribution $\mathcal{D}$. This Rademacher complexity intuitively measures how expressive $\mathcal{H}$ is in terms of correlating $\ell_{\mathrm{SPO}}(f(x_i), c_i)$ with random noise σ_i. The following theorem is an adaptation of the classical generalization bounds based on Rademacher complexity due to Bartlett and Mendelson (2002) to our setting.

THEOREM 1 (**Bartlett and Mendelson (2002), Theorem 3.3 in Mohri et al. (2018)**). *Let $\mathcal{H}$ be a family of functions mapping from $\mathcal{X}$ to $\mathbb{R}^d$. Then, for any $\delta > 0$, with probability at least $1 - \delta$ over an i.i.d. sample drawn from the distribution $\mathcal{D}$, each of the following holds for all $f \in \mathcal{H}$:*

$$R_{\mathrm{SPO}}(f) \le \hat{R}_{\mathrm{SPO}}(f) + 2\mathfrak{R}_{\mathrm{SPO}}^n(\mathcal{H}) + \omega_S(\mathcal{C})\sqrt{\frac{\log(1/\delta)}{2n}} \ .$$

REMARK 1. Note that the above Theorem provides a one-sided upper generalization bound and that its proof relies on McDiarmid's inequality. A two-sided generalization bound can be obtained with only additional absolute constants by applying McDiarmid's inequality twice. Such a result yields upper and lower bounds on $R_{\mathrm{SPO}}(f)$, enabling the construction of a two-sided confidence interval. Since we are ultimately interested in minimizing $R_{\mathrm{SPO}}(f)$, our focus herein is on providing one-sided upper generalization bounds; however many of our results (with the notable exception of the margin-based bounds in Section 4) can readily be extended to two-sided bounds for $R_{\mathrm{SPO}}(f)$. □

As previously noted, we are interested in guarantees that hold across a wide range of training procedures, beyond simply the empirical risk minimizer of the SPO loss. The fact that Theorem 1 holds *uniformly* over the hypothesis class $\mathcal{H}$ critically provides us with a tool for obtaining such guarantees: as long as the training procedure returns a predictor in $f \in \mathcal{H}$ and $\mathcal{H}$ has a bounded Rademacher complexity $\mathfrak{R}_{\mathrm{SPO}}^n(\mathcal{H})$, then Theorem 1 yields a useful upper bound on the out-of-sample SPO risk $R_{\mathrm{SPO}}(f)$. We also note that, equipped with Theorem 1, we obtain "for free" an additional guarantee on the *excess risk* of the empirical risk minimizer of the SPO loss $\hat{f}_n$. Indeed, Corollary 1 below is a standard combination of Theorem 1 with Hoeffding's inequality. The full proof is in Appendix A.1.

COROLLARY 1. *Let $\mathcal{H}$ be a family of functions mapping from $\mathcal{X}$ to $\mathbb{R}^d$. Then, for any $\delta > 0$, with probability at least $1 - \delta$ over an i.i.d. sample drawn from the distribution $\mathcal{D}$, any empirical risk minimizer $\hat{f}_n$ solving (3) satisfies:*

$$R_{\mathrm{SPO}}(\hat{f}_n) - \min_{f \in \mathcal{H}} R_{\mathrm{SPO}}(f) \le 2\mathfrak{R}_{\mathrm{SPO}}^n(\mathcal{H}) + 2\omega_S(\mathcal{C})\sqrt{\frac{\log(2/\delta)}{2n}} \ .$$

Finally, we review an extension of Rademacher complexity to the case of vector-valued function classes. Following Bertsimas and Kallus (2020) and Maurer (2016), given a fixed sample $(x_1, c_1), \ldots, (x_n, c_n)$, we define the *multivariate empirical Rademacher complexity* of $\mathcal{H}$ as

$$\hat{\mathfrak{R}}^n(\mathcal{H}) := \mathbb{E}_\sigma \left[\sup_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^d \sigma_{ij} f_j(x_i) \right] = \mathbb{E}_{\boldsymbol{\sigma}} \left[\sup_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^n \boldsymbol{\sigma}_i^T f(x_i) \right] , \tag{4}$$

where σ_{ij} are i.i.d. Rademacher random variables for $i = 1, \ldots, n$ and $j = 1, \ldots, d$, and $\boldsymbol{\sigma}_i := (\sigma_{i1}, \ldots, \sigma_{id})^T$. The expected version of the multivariate Rademacher complexity is defined as $\mathfrak{R}^n(\mathcal{H}) := \mathbb{E}\left[\hat{\mathfrak{R}}^n(\mathcal{H}) \right]$ where the expectation is taken with respect to the i.i.d. sample drawn from the underlying distribution $\mathcal{D}$.

3. Generalization Bounds Based on Multiclass Classification

In this section, we primarily consider the case where S is a polyhedron and derive generalization bounds based on bounding the Rademacher complexity of $\mathcal{H}$ with respect to the SPO loss and applying Theorem 1. The idea is similar to the bounds show for multiclass classification based on Natarajan dimension (Shalev-Shwartz and Ben-David 2014, Ch. 29), where we now treat each extreme point as a class label. Since S is polyhedral, the optimal solution of $P(\cdot)$ can always be found by considering only the finite set of extreme points of S, which we denote by the set $\mathfrak{S}$. Since the number of extreme points may be exponential in d, our goal is to provide bounds that are logarithmic in $|\mathfrak{S}|$. Note that, in the case of ties, we do not require that $w^*(\hat{c})$ necessarily returns an extreme point of S. At the end of the section, we extend our analysis to any compact and convex feasible region S by extending the polyhedral analysis with a covering number argument.

In order to derive a bound on the Rademacher complexity, we critically rely on the notion of Natarajan dimension (Natarajan 1989), which is an extension of the VC-dimension to the multiclass classification setting and is defined in our setting as follows.

DEFINITION 1 (NATARAJAN DIMENSION). Let $\mathcal{F} \subseteq S^{\mathcal{X}}$ be a hypothesis space of functions mapping from $\mathcal{X}$ to S, and let $\mathbb{X} \subseteq \mathcal{X}$ be given. We say that $\mathcal{F}$ N-shatters $\mathbb{X}$ if there exists $g_1, g_2 : \mathbb{X} \to S$ such that

- $g_1(x) \neq g_2(x)$ for all $x \in \mathbb{X}$
- For all $T \subseteq \mathbb{X}$, there exists $g \in \mathcal{F}$ such that (i) for all $x \in T$, $g(x) = g_1(x)$ and (ii) for all $x \in \mathbb{X} \backslash T$, $g(x) = g_2(x)$.

The Natarajan dimension of $\mathcal{F}$, denoted $d_N(\mathcal{F})$, is the maximal cardinality of a set N-shattered by $\mathcal{F}$. □

Notice that our definition of the Natarajan dimension above is a slight generalization of the standard definition, as we allow for functions that map to an infinite set S instead of a finite set of "labels." In our context, the labels correspond to the optimal decisions induced by the optimization

oracle $w^*(\cdot)$. The slight generalization we consider allows us to accommodate cases when $w^*(\hat{c})$ is not necessarily an extreme point of S (in the case of a tie). The Natarajan dimension is a measure for the richness of a hypothesis class. In Theorem 2, we show that the Rademacher complexity for the SPO loss can be bounded as a function of the Natarajan dimension of $w^*(\mathcal{H}) := \{x \mapsto w^*(f(x)) : f \in \mathcal{H}\}$ and the number of extreme points of S. The proof, in Appendix B.1, follows a classical argument and makes strong use of Massart's lemma and the Natarajan lemma.

THEOREM 2. *Suppose that S is a polyhedron and $\mathfrak{S}$ is the set of its extreme points. Let $\mathcal{H}$ be a family of functions mapping from $\mathcal{X}$ to $\mathbb{R}^d$. Then we have that*

$$\mathfrak{R}^n_{\mathrm{SPO}}(\mathcal{H}) \leq \omega_S(\mathcal{C})\sqrt{\frac{2d_N(w^*(\mathcal{H}))\log(n|\mathfrak{S}|^2)}{n}}.$$

Furthermore, for any $\delta > 0$, with probability at least $1 - \delta$ over an i.i.d. sample $(x_1, c_1), \ldots, (x_n, c_n)$ drawn from the distribution $\mathcal{D}$, for all $f \in \mathcal{H}$ we have

$$R_{\mathrm{SPO}}(f) \leq \hat{R}_{\mathrm{SPO}}(f) + 2\omega_S(\mathcal{C})\sqrt{\frac{2d_N(w^*(\mathcal{H}))\log(n|\mathfrak{S}|^2)}{n}} + \omega_S(\mathcal{C})\sqrt{\frac{\log(1/\delta)}{2n}}.$$

Next, we show that when $\mathcal{H}$ is restricted to the linear hypothesis class $\mathcal{H}_{\mathrm{lin}} = \{x \mapsto Bx : B \in \mathbb{R}^{d \times p}\}$, then the Natarajan dimension of $w^*(\mathcal{H}_{lin})$ can be bounded by dp. The proof, in Appendix B.2, relies on translating our problem to an instance of a linear multiclass prediction problem and using a result of Daniely and Shalev-Shwartz (2014).

COROLLARY 2. *Suppose that S is a polyhedron and $\mathfrak{S}$ is the set of its extreme points. Let $\mathcal{H}_{lin}$ be the hypothesis class of all linear functions, i.e., $\mathcal{H}_{\mathrm{lin}} = \{x \mapsto Bx : B \in \mathbb{R}^{d \times p}\}$. Then we have*

$$d_N(w^*(\mathcal{H}_{\mathrm{lin}})) \leq dp.$$

Furthermore, for any $\delta > 0$, with probability at least $1 - \delta$ over an i.i.d. sample $(x_1, c_1), \ldots, (x_n, c_n)$ drawn from the distribution $\mathcal{D}$, for all $f \in \mathcal{H}_{\mathrm{lin}}$ we have

$$R_{\mathrm{SPO}}(f) \leq \hat{R}_{\mathrm{SPO}}(f) + 2\omega_S(\mathcal{C})\sqrt{\frac{2dp\log(n|\mathfrak{S}|^2)}{n}} + \omega_S(\mathcal{C})\sqrt{\frac{\log(1/\delta)}{2n}}.$$

Next, we will build off the previous results to prove a generalization bound in the case where S is a general compact convex set. The arguments we made earlier made extensive use of the extreme points of the polyhedron. Nevertheless, this combinatorial argument can be modified in order to derive similar results for general S. The approach is to approximate S by a grid of points corresponding to the smallest cardinality ϵ-covering of S. To optimize over these grid of points, we first find the optimal solution in S and then round to the nearest point in the grid. Both the grid representation and the rounding procedure can fortunately both be handled by similar arguments made in Theorems 2 and Corollary 2, yielding a generalization bound below. The full proof is in Appendix B.3.

COROLLARY 3. *Let S be any compact and convex set, and let $\mathcal{H}_{\mathrm{lin}}$ be the hypothesis class of all linear functions. Then, for any $\delta > 0$, with probability at least $1 - \delta$ over an i.i.d. sample $(x_1, c_1), \ldots, (x_n, c_n)$ drawn from the distribution $\mathcal{D}$, for all $f \in \mathcal{H}_{\mathrm{lin}}$ we have*

$$R_{\mathrm{SPO}}(f) \leq \hat{R}_{\mathrm{SPO}}(f) + 4d\omega_S(\mathcal{C})\sqrt{\frac{2p\log(2n\rho_2(S)d)}{n}} + \omega_S(\mathcal{C})\sqrt{\frac{\log(1/\delta)}{2n}} + O\left(\frac{1}{n}\right).$$

Although the dependence on the sample size n in the above bound is favorable, the dependence on the dimension of the feature space p and the dimension of the feasible region d is relatively weak. Given that the proofs of Corollary 3 and Theorem 2 are purely combinatorial and hold for worst-case distributions, this is not surprising. In the next section, we demonstrate how to exploit the structure of the SPO loss function and additional properties of the distribution $\mathcal{D}$ and the feasible region S in order to develop improved bounds.

4.   Margin-based Generalization Bounds

In this section, we seek to obtain generalization bounds that have better dependency on the feature dimension p and the decision dimension d than those in Corollaries 2 and 3. The key idea in Section 3 is to essentially treat S as a collection of extreme points and treat the predict-then-optimize problem as a multiclass classification problem. This approach invariably leads to a $\sqrt{pd}$ dependence on the number of parameters pd, with an additional factor of $\sqrt{d}$ in the general convex case or a square root logarithmic dependence on the number of extreme points in the polyhedral case. This is not surprising, as the Natarajan dimension measures the maximum cardinality set (from the collection of extreme points) that can be shattered by a linear hypothesis class with pd parameters.

However, given some constraints on the linear hypothesis class (i.e., ridge or lasso regularization) one would hope to improve the dependence on the number of parameters leveraging relevant results on the regular Rademacher complexity of the constrained version of $\mathcal{H}_{\mathrm{lin}}$. One would also hope that the dependency on the number of extreme points is also not relevant to the generalization bound. Interestingly, both the dependence on the number of parameters and the dependence on the number of extreme points may be improved if the SPO loss were L-Lipschitz (w.r.t. the ℓ_2-norm) due to the vector concentration inequality of Maurer (2016), which would imply that $\hat{\mathfrak{R}}^n_{\mathrm{SPO}}(\mathcal{H})$ is at most $\sqrt{2}L\hat{\mathfrak{R}}^n(\mathcal{H})$. Although the Rademacher complexity of a multivariate hypothesis class is less commonly used in machine learning, we show in Section 4.4 that many of the results that bring down the dependency on the number of parameters in the univariate case still apply in the multivariate case.

Unfortunately, the SPO loss is *not* Lipschitz. It is not so difficult to see that there can be sets S with predictions $\hat{c}$ that yield wildly different solutions if $\hat{c}$ is perturbed, i.e., $w^*(\hat{c})$ is not Lipschitz. Furthermore, the SPO loss is actually discontinuous when $\hat{c}$ is a vector that has multiple optimal

solutions (e.g., when $\hat{c}$ is the zero vector or a normal vector to a face of a polyhedron). On the other hand, the SPO loss may have desirable continuity properties when $\hat{c}$ is bounded away from the "degenerate" cost vectors that lead to multiple optimal solutions. This intuition motivates us to use a modified version of the SPO loss called the *margin SPO loss*. Our developments are akin to and in fact are a strict generalization of the margin analysis for classification developed in Koltchinskii and Panchenko (2002). In particular, we extend the notion of margin from classification problems to the predict-then-optimize setting by associating the "margin" of a cost vector prediction $\hat{c}$ to its "distance to degeneracy," where here degeneracy refers to when problem (2) has multiple optimal solutions. The margin SPO loss imposes an additional penalty on cost vector predictions when they are close to being degenerate. We prove that the margin SPO loss is a Lipschitz function when S satisfies what we call the *strength property*, which measures the sensitivity of the optimal solution around $\hat{c}$ with respect to the distance to degeneracy. We leverage the Lipschitz property of the margin SPO loss in order to develop improved generalization bounds that use the multivariate Rademacher complexity of $\mathcal{H}$ as discussed above, at the small expense of using the empirical margin SPO loss rather than the empirical SPO loss as in Theorem 1.

The analysis and results in this section hold generally for any convex and compact feasible region S that satisfies the strength property, but the ability to compute the distance to degeneracy and subsequently evaluate the empirical margin SPO loss depends on the structural properties of S. We note also that the Lipschitz parameter of the margin SPO loss depends on this strength property and the size of the margin. A smaller margin implies a larger Lipschitz parameter yet a closer imitation of the margin SPO loss to the original SPO loss. In Section 5, we discuss two special cases, strongly convex sets and polyhedral sets with known convex hull representations, where the strength property is satisfied and where one can efficiently compute the distance to degeneracy, enabling the application of the generalization bounds developed herein.

4.1. Margin SPO Loss
Let us now present the definition of the set of degenerate cost vectors and the distance to degeneracy. Recall that $\|\cdot\|$ is a generic given norm on $\mathbb{R}^d$ with dual norm denoted by $\|\cdot\|_*$.

DEFINITION 2 (DISTANCE TO DEGENERACY). The set of *degenerate* cost vector predictions is $\mathcal{C}^\circ := \{\hat{c} \in \mathbb{R}^d : P(\hat{c}) \text{ has multiple optimal solutions}\}$, and the distance function (measured in the dual norm) to this set is denoted by $\nu_S(\hat{c}) := \mathrm{dist}^*_{\mathcal{C}^\circ}(\hat{c}) = \inf_{c \in \mathcal{C}^\circ}\{\|c - \hat{c}\|_*\}$. □

Notice that the set $\mathcal{C}^\circ$ and hence the function $\nu_S(\cdot)$ is determined completely by the set S, see Figure 1 for an example. In this section, we assume without loss of generality that S is not a singleton. Indeed, if S is a singleton then the SPO loss $\ell_{\mathrm{SPO}}(\cdot, c)$ is identically equal to 0 for all $c \in \mathbb{R}^d$ and generalization bounds are trivial. Since S is not a singleton, we have that $0 \in \mathcal{C}^\circ$ and

Figure 1   A geometric example of distance to degeneracy.
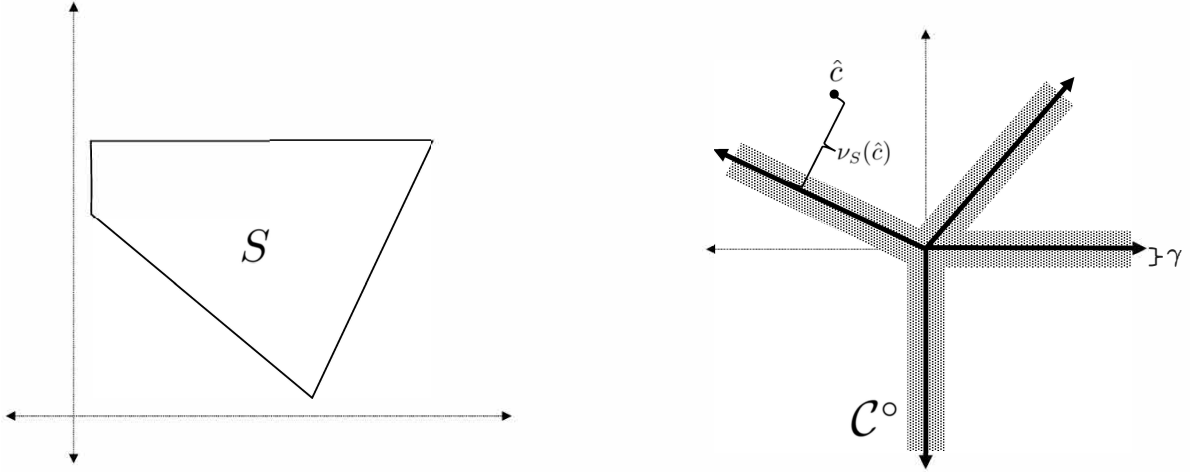
Note. The left figure shows an example of a two-dimensional feasible region S that corresponds to a polyhedron. The right figure shows the set of degenerate cost vectors $\mathcal{C}^\circ$ in the black arrows. The shaded area around $\mathcal{C}^0$ represents a γ-margin around the degenerate set of cost vectors. For some cost vector $\hat{c}$, we show the distance to the closest degenerate point which is exactly $\nu_S(\hat{c})$ when using the 2-norm (whose dual norm is also the 2-norm).

therefore $\nu_S(\hat{c}) \in [0, \infty)$ for all $\hat{c} \in \mathbb{R}^d$. (In fact, $0 \in \mathcal{C}^\circ$ implies that $\nu_S(\hat{c}) \leq \|\hat{c}\|_*$.) Therefore, since $\nu_S(\cdot)$ is a distance function to a non-empty set, it is also a 1-Lipschitz function (see, e.g., Rockafellar and Wets (2009) for a reference on this standard fact). A key property of the feasible region S and the function $\nu_S(\cdot)$ that will enable us to prove a "Lipschitz-like" property of the optimization oracle $w^*(\cdot)$ is the following strength property formalized in Definition 3 below.

DEFINITION 3 (STRENGTH PROPERTY). The feasible region S satisfies the *strength property* if for some constant $\mu > 0$,

$$\hat{c}^T(w - w^*(\hat{c})) \geq \left(\frac{\mu \cdot \nu_S(\hat{c})}{2}\right) \|w - w^*(\hat{c})\|^2 \quad \text{for all } w \in S \text{ and } \hat{c} \in \mathbb{R}^d , \tag{5}$$

where $\nu_S(\cdot)$ is the distance to degeneracy function. We refer to μ as the *strength parameter*. $\square$

The strength property above may be thought of as a type of "quantitative optimality condition" that justifies the use of the term "strength." Indeed, (5) indicates that larger values of $\nu_S(\hat{c})$ provide stronger bounds on the optimality gap $\hat{c}^T(w - w^*(\hat{c}))$ relative to the squared distance $\|w - w^*(\hat{c})\|^2$. The distance to degeneracy function is chosen to make the strength property as tight as possible, and thus ensuring better generalization bounds since we want to avoid points in the γ-margin of $\nu_S(\hat{c})$ as much as possible. The name is also inspired by results in the case when S is strongly convex, although in Section 5 it it shown that the strength property holds in the polyhedral case as well. An example of a convex set that does not satisfy the strength property for any $\mu > 0$ is an ℓ_p ball where $2 < p < \infty$.

REMARK 2 (NORMAL CONE INTERPRETATION). Notice that the strength property in Definition 3 is independent of the optimization oracle $w^*(\cdot)$ since, for any $\hat{c}$ where there are multiple choices for $w^*(\hat{c})$, by definition it holds that $\nu_S(\hat{c}) = 0$ and thus (5) is satisfied. In other words, we can equivalently restate (5) as follows. Let the normal cone of S at a point $\bar{w} \in S$ be denoted by $N_S(\bar{w}) := \{c \in \mathbb{R}^d : c^T(w - \bar{w}) \leq 0 \text{ for all } w \in S\}$. Then, (5) is equivalent to requiring that

$$N_S(\bar{w}) = \left\{c \in \mathbb{R}^d : c^T(w - \bar{w}) \leq -\left(\frac{\mu \cdot \nu_S(-c)}{2}\right) \|w - \bar{w}\|^2 \text{ for all } w \in S\right\} \quad \text{for all } \bar{w} \in S.$$

In Section 5, we show that strongly convex and polyhedral sets satisfy the strength property and thus the equivalent property above. The above is a stronger characterization of the normal cones of S and may be of interest in other contexts as well. For example, for the differentiable nonlinear minimization problem $\min_{w \in S} F(w)$, the above gives a tighter characterization of the optimality condition $-\nabla F(\bar{w}) \in N_S(\bar{w})$.   □

As discussed, the distance to degeneracy $\nu_S(\hat{c})$ provides a measure of "confidence" regarding the cost vector prediction $\hat{c}$ and its implied decision $w^*(\hat{c})$. This intuition motivates us to define a margin-based version of the SPO loss, which places a greater penalty on cost vector predictions that are nearly degenerate. In other words, the "margin SPO loss" encourages predictions that have small SPO loss values as well as predictions that are far enough away from the set of degenerate cost vector predictions $\mathcal{C}^\circ$. Definition 4 presents the definition of the "γ-margin SPO loss," which requires a fixed parameter $\gamma > 0$ that controls the size of the "margin" around $\mathcal{C}^\circ$ where cost vector predictions are more heavily penalized.

DEFINITION 4 (γ-MARGIN SPO LOSS). For a fixed parameter $\gamma > 0$, given a cost vector prediction $\hat{c}$ and a realized cost vector c, the *γ-margin SPO loss* $\ell_{\mathrm{mSPO}}^\gamma(\hat{c}, c)$ is defined as:

$$\ell_{\mathrm{mSPO}}^\gamma(\hat{c}, c) := \begin{cases} \ell_{\mathrm{SPO}}(\hat{c}, c) & \text{if } \nu_S(\hat{c}) > \gamma \\ \left(\frac{\nu_S(\hat{c})}{\gamma}\right) \ell_{\mathrm{SPO}}(\hat{c}, c) + \left(1 - \frac{\nu_S(\hat{c})}{\gamma}\right) \omega_S(c) & \text{if } \nu_S(\hat{c}) \leq \gamma. \end{cases} \quad □$$

Recall that, for any $\hat{c}, c \in \mathbb{R}^d$, it holds that $\ell_{\mathrm{SPO}}(\hat{c}, c) \leq \omega_S(c)$. Hence, we also have that $\ell_{\mathrm{SPO}}(\hat{c}, c) \leq \ell_{\mathrm{mSPO}}^\gamma(\hat{c}, c)$, that is the γ-margin SPO loss provides an upper bound on the SPO loss. Notice that the γ-margin SPO loss linearly interpolates between the SPO loss $\ell_{\mathrm{SPO}}(\hat{c}, c)$ and the upper bound $\omega_S(c)$ whenever $\nu_S(\hat{c}) \leq \gamma$. The γ-margin SPO loss also satisfies a simple monotonicity property whereby $\ell_{\mathrm{SPO}}^\gamma(\hat{c}, c) \leq \ell_{\mathrm{SPO}}^{\bar{\gamma}}(\hat{c}, c)$ for any $\hat{c}, c \in \mathbb{R}^d$ and $\bar{\gamma} \geq \gamma > 0$.

We are now ready to state a theorem concerning the Lipschitz properties of the optimization oracle $w^*(\cdot)$ and the γ-margin SPO loss, which will then be used to derive margin-based generalization bounds. Theorem 3 below first demonstrates that the optimization oracle $w^*(\cdot)$ and SPO loss both satisfy a "Lipschitz-like" property when the distance to degeneracy function $\nu_S(\cdot)$ is bounded away from zero. Subsequently, these Lipschitz-like properties are a key ingredient in demonstrating that the γ-margin SPO loss is a Lipschitz function.

THEOREM 3. *Suppose that the feasible region S satisfies the strength property with parameter μ. Then, we have the following:*

(a) *The optimization oracle $w^*(\cdot)$ satisfies the following Lipschitz-like property:*

$$\|w^*(\hat{c}_1) - w^*(\hat{c}_2)\| \ \leq \ \left(\frac{1}{\mu \cdot \min\{\nu_S(\hat{c}_1), \nu_S(\hat{c}_2)\}}\right) \|\hat{c}_1 - \hat{c}_2\|_* \ \ for \ all \ \hat{c}_1, \hat{c}_2 \in \mathbb{R}^d.$$

(b) *For any fixed $c \in \mathbb{R}^d$, the SPO loss satisfies the following Lipschitz-like property:*

$$|\ell_{\mathrm{SPO}}(\hat{c}_1, c) - \ell_{\mathrm{SPO}}(\hat{c}_2, c)| \ \leq \ \left(\frac{\|c\|_*}{\mu \cdot \min\{\nu_S(\hat{c}_1), \nu_S(\hat{c}_2)\}}\right) \|\hat{c}_1 - \hat{c}_2\|_* \ \ for \ all \ \hat{c}_1, \hat{c}_2 \in \mathbb{R}^d.$$

(c) *For any fixed $c \in \mathbb{R}^d$ and $\gamma > 0$, the γ-margin SPO loss is $\frac{1}{\gamma\mu}(\|c\|_* + \mu \cdot \omega_S(c))$-Lipschitz with respect to the dual norm $\|\cdot\|_*$, i.e., it holds that:*

$$|\ell^\gamma_{\mathrm{mSPO}}(\hat{c}_1, c) - \ell^\gamma_{\mathrm{mSPO}}(\hat{c}_2, c)| \ \leq \ \left(\frac{\|c\|_* + \mu \cdot \omega_S(c)}{\gamma\mu}\right) \|\hat{c}_1 - \hat{c}_2\|_* \ \ for \ all \ \hat{c}_1, \hat{c}_2 \in \mathbb{R}^d \ .$$

Proof. If S is a singleton, then all of these results are trivial. Otherwise, as mentioned previously, $\nu_S(\cdot)$ is finite valued and 1-Lipschitz on $\mathbb{R}^d$. Let $\hat{c}_1, \hat{c}_2 \in \mathbb{R}^d$ be given and let $\tau := \min\{\nu_S(\hat{c}_1), \nu_S(\hat{c}_2)\}$.

Part (a): If $\tau = 0$ then (a) holds since the right-hand side of (a) is equal to $+\infty$ by convention. Otherwise $\tau > 0$ and applying (5) twice yields

$$\hat{c}_1^T(w^*(\hat{c}_2) - w^*(\hat{c}_1)) \ \geq \ \left(\tfrac{\mu \cdot \nu_S(\hat{c}_1)}{2}\right) \|w^*(\hat{c}_2) - w^*(\hat{c}_1)\|^2 \ \geq \ \left(\tfrac{\mu\tau}{2}\right) \|w^*(\hat{c}_1) - w^*(\hat{c}_2)\|^2 \ ,$$

and

$$\hat{c}_2^T(w^*(\hat{c}_1) - w^*(\hat{c}_2)) \ \geq \ \left(\tfrac{\mu \cdot \nu_S(\hat{c}_2)}{2}\right) \|w^*(\hat{c}_1) - w^*(\hat{c}_2)\|^2 \ \geq \ \left(\tfrac{\mu\tau}{2}\right) \|w^*(\hat{c}_1) - w^*(\hat{c}_2)\|^2 \ .$$

Adding the above two inequalities together yields

$$\mu\tau\|w^*(\hat{c}_1) - w^*(\hat{c}_2)\|^2 \ \leq \ (\hat{c}_2 - \hat{c}_1)^T(w^*(\hat{c}_1) - w^*(\hat{c}_2)) \ \leq \ \|\hat{c}_1 - \hat{c}_2\|_*\|w^*(\hat{c}_1) - w^*(\hat{c}_2)\| \ ,$$

where the second inequality is Hölder's inequality. Dividing both sides of the above by $\mu\tau\|w^*(\hat{c}_1) - w^*(\hat{c}_2)\|$ yields item (a).

Part (b): Now let $c \in \mathbb{R}^d$ be given and note that

$$|\ell_{\mathrm{SPO}}(\hat{c}_1, c) - \ell_{\mathrm{SPO}}(\hat{c}_2, c)| \ = \ |c^T(w^*(\hat{c}_1) - w^*(\hat{c}_2))| \ \leq \ \|c\|_*\|w^*(\hat{c}_1) - w^*(\hat{c}_2)\| \ ,$$

where the inequality follows from Hölder's inequality. Hence, (b) follows by combining the above with (a).

Part (c): Without loss of generality, we consider three cases: *(i)* $\nu_S(\hat{c}_1) \leq \gamma$ and $\nu_S(\hat{c}_2) \leq \gamma$, *(ii)* $\nu_S(\hat{c}_1) \geq \gamma$ and $\nu_S(\hat{c}_2) \geq \gamma$, and *(iii)* $\nu_S(\hat{c}_1) \leq \gamma$ and $\nu_S(\hat{c}_2) > \gamma$.

Let us first consider case *(i)* and let us suppose without loss of generality that $0 < \nu_S(\hat{c}_1) \leq \nu_S(\hat{c}_2) \leq \gamma$. By definition, we know that $|\ell_{\mathrm{mSPO}}^\gamma(\hat{c}_1, c) - \ell_{\mathrm{mSPO}}^\gamma(\hat{c}_2, c)|$ is equal to

$$
\begin{aligned}
&\left| \left(\frac{\nu_S(\hat{c}_1)}{\gamma} \right) \ell_{\mathrm{SPO}}(\hat{c}_1, c) + \left(1 - \frac{\nu_S(\hat{c}_1)}{\gamma} \right) \omega_S(c) - \left(\frac{\nu_S(\hat{c}_2)}{\gamma} \right) \ell_{\mathrm{SPO}}(\hat{c}_2, c) - \left(1 - \frac{\nu_S(\hat{c}_2)}{\gamma} \right) \omega_S(c) \right| \\
&= \left| \left(\frac{\nu_S(\hat{c}_1)}{\gamma} \right) [\ell_{\mathrm{SPO}}(\hat{c}_1, c) - \ell_{\mathrm{SPO}}(\hat{c}_2, c)] + \left(\frac{\ell_{\mathrm{SPO}}(\hat{c}_2, c) - \omega_S(c)}{\gamma} \right) [\nu_S(\hat{c}_1) - \nu_S(\hat{c}_2)] \right| \\
&\leq \left(\frac{\nu_S(\hat{c}_1)}{\gamma} \right) |\ell_{\mathrm{SPO}}(\hat{c}_1, c) - \ell_{\mathrm{SPO}}(\hat{c}_2, c)| + \left(\frac{\omega_S(c)}{\gamma} \right) |\nu_S(\hat{c}_1) - \nu_S(\hat{c}_2)| \\
&\leq \left(\frac{\nu_S(\hat{c}_1)}{\gamma} \right) \left(\frac{\|c\|_*}{\mu \cdot \nu_S(\hat{c}_1)} \right) \|\hat{c}_1 - \hat{c}_2\|_* + \left(\frac{\omega_S(c)}{\gamma} \right) \|\hat{c}_1 - \hat{c}_2\|_* \\
&= \frac{1}{\gamma} \left(\frac{\|c\|_*}{\mu} + \omega_S(c) \right) \|\hat{c}_1 - \hat{c}_2\|_* \,,
\end{aligned}
$$

where the first inequality follows from the triangle inequality and the definition of $\omega_S(c)$ and the second inequality follows from part (b) and the facts that $0 < \nu_S(\hat{c}_1) = \min\{\nu_S(\hat{c}_1), \nu_S(\hat{c}_2)\}$ and $\nu_S(\cdot)$ is 1-Lipschitz.

Now, in case *(ii)*, we have that $\ell_{\mathrm{mSPO}}^\gamma(\hat{c}_1, c) = \ell_{\mathrm{SPO}}(\hat{c}_1, c)$ and $\ell_{\mathrm{mSPO}}^\gamma(\hat{c}_2, c) = \ell_{\mathrm{SPO}}(\hat{c}_2, c)$. Therefore, applying part (b) yields

$$
|\ell_{\mathrm{mSPO}}^\gamma(\hat{c}_1, c) - \ell_{\mathrm{mSPO}}^\gamma(\hat{c}_2, c)| \leq \frac{\|c\|_*}{\mu \cdot \min\{\nu_S(\hat{c}_1), \nu_S(\hat{c}_2)\}} \|\hat{c}_1 - \hat{c}_2\|_* \leq \frac{\|c\|_*}{\gamma\mu} \|\hat{c}_1 - \hat{c}_2\|_* \leq \frac{\|c\|_* + \mu \cdot \omega_S(c)}{\gamma\mu} \|\hat{c}_1 - \hat{c}_2\|_* \,,
$$

where the second inequality follows since $\gamma \leq \min\{\nu_S(\hat{c}_1), \nu_S(\hat{c}_2)\}$.

Finally, in case *(iii)*, by the intermediate value theorem we can define $\bar{c} := \lambda \hat{c}_1 + (1 - \lambda)\hat{c}_2$ where $\lambda \in (0, 1]$ such that $\nu_S(\bar{c}) = \gamma$. Then, we have that $|\ell_{\mathrm{mSPO}}^\gamma(\hat{c}_1, c) - \ell_{\mathrm{mSPO}}^\gamma(\hat{c}_2, c)|$ is equal to

$$
\begin{aligned}
&|(\ell_{\mathrm{mSPO}}^\gamma(\hat{c}_1, c) - \ell_{\mathrm{mSPO}}^\gamma(\bar{c}, c)) + (\ell_{\mathrm{mSPO}}^\gamma(\bar{c}, c) - \ell_{\mathrm{mSPO}}^\gamma(\hat{c}_2, c))| \\
&\leq |\ell_{\mathrm{mSPO}}^\gamma(\hat{c}_1, c) - \ell_{\mathrm{mSPO}}^\gamma(\bar{c}, c)| + |\ell_{\mathrm{mSPO}}^\gamma(\bar{c}, c) - \ell_{\mathrm{mSPO}}^\gamma(\hat{c}_2, c)| \\
&\leq \left(\frac{\|c\|_* + \mu \cdot \omega_S(c)}{\gamma\mu} \right) \|\hat{c}_1 - \bar{c}\| + \left(\frac{\|c\|_* + \mu \cdot \omega_S(c)}{\gamma\mu} \right) \|\bar{c} - \hat{c}_2\| \\
&= \left(\frac{\|c\|_* + \mu \cdot \omega_S(c)}{\gamma\mu} \right) (\|\hat{c}_1 - \bar{c}\| + \|\bar{c} - \hat{c}_2\|),
\end{aligned}
$$

where the first inequality follows from the triangle inequality, the second inequality follows from cases *(i)* and *(ii)*, and the final equality follows the definition of $\bar{c}$ □

4.2.   Connecting to Multivariate Rademacher Complexity

As we are extending standard margin-based generalization theory from binary classification to the predict-then-optimize setting, it is instructive to recall the steps involved in the binary classification setting (e.g., see (Mohri et al. 2018, Section 5.4)). First, the 0-1 loss function is upper bounded by

the γ-margin loss function. Second, the Rademacher complexity of the γ-margin loss function class is upper bounded by the Rademacher complexity of the underlying function class. This second step uses the Lipschitz property of the γ-margin loss function and a result known as the Ledoux-Talagrand contraction inequality (e.g., see (Mohri et al. 2018, Lemma 5.7))

We have already accomplished step one above by constructing the γ-margin SPO loss which is Lipschitz and also upper bounds the SPO loss. Note that this step was significantly more challenging than in the binary classification case. Step two also poses its own set of challenges. In the binary classification case, the function class is a class of scalar-valued functions. However, in our case, the class $\mathcal{H}$ consists of *vector-valued functions*. This raises two important questions. First, how do we define the Rademacher complexity of a vector-valued function class? Second, what is the analogue of the Ledoux-Talagrand contraction inequality for the vector case?

Recall that the multivariate empirical Rademacher complexity of $\mathcal{H}$ is defined as

$$\hat{\mathfrak{R}}^n(\mathcal{H}) := \mathbb{E}_\sigma \left[\sup_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^d \sigma_{ij} f_j(x_i) \right] = \mathbb{E}_{\boldsymbol{\sigma}} \left[\sup_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^n \boldsymbol{\sigma}_i^T f(x_i) \right] , \qquad (6)$$

where σ_{ij} are i.i.d. Rademacher random variables for $i = 1, \ldots, n$ and $j = 1, \ldots, d$, and $\boldsymbol{\sigma}_i := (\sigma_{i1}, \ldots, \sigma_{id})^T$. The expected version of the multivariate Rademacher complexity is defined as $\mathfrak{R}^n(\mathcal{H}) := \mathbb{E}\left[\hat{\mathfrak{R}}^n(\mathcal{H}) \right]$ where the expectation is taken with respect to the i.i.d. sample drawn from the underlying distribution $\mathcal{D}$. It is often the case that the structure of the hypothesis class $\mathcal{H}$ naturally leads to a bound on $\mathfrak{R}^n(\mathcal{H})$ that can have mild, even logarithmic, dependence on dimensions p and d. We give examples of such bounds in Section 4.4, after first presenting our margin-based generalization bounds.

Let us also introduce the empirical γ-margin SPO loss:

$$\hat{R}_{\mathrm{SPO}}^\gamma(f) := \frac{1}{n} \sum_{i=1}^n \ell_{\mathrm{SPO}}^\gamma(f(x_i), c_i) ,$$

and the empirical Rademacher complexity of $\mathcal{H}$ with respect to the γ-margin SPO loss:

$$\hat{\mathfrak{R}}_{\gamma\mathrm{SPO}}^n(\mathcal{H}) := \mathbb{E}_\sigma \left[\sup_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^n \sigma_i \ell_{\mathrm{SPO}}^\gamma(f(x_i), c_i) \right] ,$$

where $f \in \mathcal{H}$ on the left side above and σ_i are i.i.d. (scalar) Rademacher random variables for $i = 1, \ldots, n$. So we need tools that allow us to transition from the standard Rademacher complexity of $\mathcal{H}$ with respect to the γ-margin SPO loss to the multivariate Rademacher complexity of $\mathcal{H}$.

Let us now briefly review the concept of vector contraction inequalities for multivariate Rademacher complexities, which turns out to provide precisely the right set of tools. Let $\Phi_i : \mathbb{R}^d \to \mathbb{R}$ for $i \in \{1, \ldots, n\}$ be a collection of L-Lipschitz functions with respect to the given norm $\|\cdot\|$ defined on $\mathbb{R}^d$:

$$|\Phi_i(u) - \Phi_i(v)| \le L \cdot \|u - v\| \quad \text{for all } u, v \in \mathbb{R}^d.$$

A *vector contraction inequality* takes the form:

$$\mathbb{E}_{\sigma}\left[\sup_{f\in\mathcal{H}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i\Phi_i(f(x_i))\right] \le CL\cdot\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{f\in\mathcal{H}}\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\sigma}_i^T f(x_i)\right] = CL\cdot\hat{\mathfrak{R}}^n(\mathcal{H})\ , \tag{7}$$

where C is a constant. As we have already demonstrated that the γ-margin SPO loss is Lipschitz, our next step is to apply a vector contraction inequality in order to obtain improved generalization bounds.

4.3. Generalization Bounds in the ℓ_2 Case

We now focus on the case of the ℓ_2-norm set-up, i.e., the norm on the space of w variables as well as the norm on the space of cost vectors c are both the ℓ_2-norm. Our primary reason for focusing on the ℓ_2 case is that we can apply an elegant vector contraction inequality due to Maurer (2016), which exactly takes the form of (7) with $C = \sqrt{2}$. Vector contraction inequalities have also been developed in other norm setups, but the result of Maurer (2016) in the ℓ_2 case appears to be the most elegant and the most applicable result for our purposes.

REMARK 3 (OTHER VECTOR CONTRACTION INEQUALITIES). Indeed, for the case when Φ_i are Lipschitz with respect to the ℓ_∞-norm, the first relevant inequality is that of Bertsimas and Kallus (2020) (see Lemma EC.1 in their supplementary material). Note that this inequality is inferior to the result of Maurer (2016) since the ℓ_∞ Lipschitz constant of a function is always larger than, or equal to, the ℓ_2 Lipschitz constant. An alternative contraction inequality for the ℓ_∞ case is due to Foster and Rakhlin (2019), which involves a worst case alternative to the multivariate Rademacher complexity and may be favorable in some situations. Finally, Zatarain-Vera (2019) has recently considered the case of an ℓ_p-norm for $p \in (1,2)$, which involves replacing Rademacher random variables with p-stable random variables. The result of Zatarain-Vera (2019) is not particularly useful in the case of strongly convex sets S since all common examples of strongly convex sets with strong convexity constant independent of dimension are strongly convex with respect to an ℓ_p norm for some $p \le 2$. However, this result may have potential uses for some polyhedral sets.   □

With Theorem 3 and the vector contraction inequality of Maurer (2016) in place, we are now ready to present our margin-based generalization bound in the predict-then-optimize setting in Theorem 4 below. Recall that $\mathcal{C}$ denotes the domain of the true cost vectors c, $\rho_2(\mathcal{C}) := \sup_{c\in\mathcal{C}}\|c\|_2$, and $\omega_S(\mathcal{C}) := \sup_{c\in\mathcal{C}}\omega_S(c)$.

THEOREM 4. *Suppose that the feasible region S satisfies the strength property with parameter $\mu > 0$ and with respect to the ℓ_2-norm. Let $\gamma > 0$ be fixed, and let $\mathcal{H}$ be a family of functions mapping from $\mathcal{X}$ to $\mathbb{R}^d$. Then, for any fixed sample $((x_1, c_1)...(x_n, c_n))$ we have that*

$$\hat{\mathfrak{R}}^n_{\gamma\mathrm{SPO}}(\mathcal{H}) \ \le\ \left(\frac{\sqrt{2}\rho_2(\mathcal{C}) + \sqrt{2}\mu\cdot\omega_S(\mathcal{C})}{\gamma\mu}\right)\hat{\mathfrak{R}}^n(\mathcal{H})\ .$$

Furthermore, for any $\delta > 0$, with probability at least $1 - \delta$ over an i.i.d. sample drawn from the distribution $\mathcal{D}$, the following holds for all $f \in \mathcal{H}$

$$R_{\mathrm{SPO}}(f) \ \leq \ \hat{R}^{\gamma}_{\mathrm{SPO}}(f) + \left(\frac{2\sqrt{2}\rho_2(\mathcal{C}) + 2\sqrt{2}\mu \cdot \omega_S(\mathcal{C})}{\gamma\mu} \right) \mathfrak{R}^n(\mathcal{H}) + \omega_S(\mathcal{C})\sqrt{\frac{\log(1/\delta)}{2n}} \ .$$

Proof. The bound on $\hat{\mathfrak{R}}^n_{\gamma\mathrm{SPO}}(\mathcal{H})$ follows simply by combining Theorem 3(c) with the vector contraction inequality (7) due to Maurer (2016). The subsequent generalization bound then simply follows since $R_{\mathrm{SPO}}(f) \leq R^{\gamma}_{\mathrm{SPO}}(f)$ for all $f \in \mathcal{H}$ and by applying the version of Theorem 1 for the γ-margin SPO loss. $\square$

Theorem 4 may also be extended to a bound that holds uniformly over all values of $\gamma \in (0, \bar{\gamma}]$, where $\bar{\gamma} > 0$ is a fixed parameter. This extension is presented below in Theorem 5 and proved in Appendix C.1. The advantage of the bound presented in Theorem 5 versus the bound presented in Theorem 4 is that γ in Theorem 5 can be chosen in a data-driven way, e.g., through cross-validation.

THEOREM 5. *Suppose that the feasible region S satisfies the strength property with parameter $\mu > 0$ and with respect to the ℓ_2-norm. Let $\bar{\gamma} > 0$ be fixed, and let $\mathcal{H}$ be a family of functions mapping from $\mathcal{X}$ to $\mathbb{R}^d$. Then, for any $\delta > 0$, with probability at least $1 - \delta$ over an i.i.d. sample drawn from the distribution $\mathcal{D}$, the following holds for all $f \in \mathcal{H}$ and for all $\gamma \in (0, \bar{\gamma}]$*

$$R_{\mathrm{SPO}}(f) \leq \hat{R}^{\gamma}_{\mathrm{SPO}}(f) + \left(\frac{4\sqrt{2}\rho_2(\mathcal{C}) + 4\sqrt{2}\mu \cdot \omega_S(\mathcal{C})}{\gamma\mu} \right) \mathfrak{R}^n(\mathcal{H}) + \omega_S(\mathcal{C}) \left(\sqrt{\frac{\log(\log_2(2\bar{\gamma}/\gamma))}{n}} + \sqrt{\frac{\log(2/\delta)}{2n}} \right) \ .$$

The proof of Theorem 5 is presented in Appendix C. Note that a natural choice for $\bar{\gamma}$ in Theorem 5 is $\bar{\gamma} \leftarrow \sup_{f \in \mathcal{H}, x \in \mathcal{X}} \nu_S(f(x))$ or an upper bound thereof, presuming that one can obtain such a bound based on the properties of S, $\mathcal{H}$, and $\mathcal{X}$. In addition, since $\nu_S(\hat{c}) \leq \|\hat{c}\|_2$ for all $\hat{c} \in \mathbb{R}^d$, one can also take $\bar{\gamma} \leftarrow \sup_{f \in \mathcal{H}, x \in \mathcal{X}} \|f(x)\|_2$ or an upper bound thereof. As mentioned, the value of γ in Theorem 5 can be chosen in a data-driven way so that, given a prediction function $\hat{f}$ trained on the observed data, the upper bound on $\hat{R}_{\mathrm{SPO}}(\hat{f})$ given by Theorem 5 is minimized. Since Theorem 5 is a uniform result over $\gamma \in (0, \bar{\gamma}]$, this data-driven procedure for choosing γ is indeed valid.

It is important to note that the utility of Theorems 4 and 5 is strengthened when the underlying distribution $\mathcal{D}$ has a "favorable margin property," whereby $\hat{R}^{\gamma}_{\mathrm{SPO}}(f) \approx \hat{R}_{\mathrm{SPO}}(f)$ for some value of $\gamma > 0$ that is reasonably large. A sufficient condition to ensure such a favorable margin property is if $\nu_S(f(x)) \geq \gamma$ with high probability over x, i.e., most predictions of the model f are far from degenerate. When such a favorable margin property occurs, the bounds in Theorems 4 and 5 can be much stronger than those of Corollaries 2 and 3. Indeed, we demonstrate in Section 4.4 and 5 examples of hypothesis classes and feasible regions, respectively, such that the dependence on the number of parameters pd in the bounds of Theorems 4 and 5 is significantly stronger than that of Corollaries 2 and 3.

4.4. Bounding the Multivariate Rademacher Complexity for Linear Classes

As mentioned, it is often the case that the structure of the hypothesis class $\mathcal{H}$ naturally leads to a bound on $\mathfrak{R}^n(\mathcal{H})$ that has favorable dependence on dimensions p and d. For example, let us consider the general setting of a constrained linear function class, namely $\mathcal{H} = \mathcal{H}_{\mathcal{B}} := \{f : f(x) = Bx \text{ for some } B \in \mathbb{R}^{d \times p}, B \in \mathcal{B}\}$, where $\mathcal{B} \subseteq \mathbb{R}^{d \times p}$. Theorems 4 and 5 apply more broadly than just linear function classes, but these examples are illustrative and comparable to Corollaries 2 and 3. Theorem 6 below is an extension of Theorem 3 of Kakade et al. (2008) to multivariate Rademacher complexity and provides a convenient way to bound $\mathfrak{R}^n(\mathcal{H}_{\mathcal{B}})$ in the case when $\mathcal{B}$ lies in a level set of a strongly convex function. In a slight abuse of notation, in Theorem 6 only we use $\|\cdot\|$ and $\|\cdot\|_*$ to refer to a norm and corresponding dual norm, respectively, on the space of *matrices* $\mathbb{R}^{d \times p}$. The full proof is in Appendix C.2.

THEOREM 6. *Let $\mathcal{S} \subseteq \mathbb{R}^{d \times p}$ be a closed convex set, and let $F : \mathcal{S} \to \mathbb{R}$ be α-strongly convex with respect to $\|\cdot\|_*$ and satisfy $\inf_{B \in \mathcal{S}} F(B) = 0$. Suppose that $\mathcal{B} \subseteq \{B \in \mathcal{S} : F(B) \leq \beta^2\}$ and let Ω be such that*

$$\sup_{\boldsymbol{\sigma} \in \{\pm 1\}^d} \sup_{x \in \mathcal{X}} \|\boldsymbol{\sigma} x^T\| \leq \Omega.$$

Then, it holds that

$$\mathfrak{R}^n(\mathcal{H}_{\mathcal{B}}) \leq \Omega\beta\sqrt{\frac{2}{\alpha n}}.$$

This theorem can be applied with many different strongly convex functions of matrices (Kakade et al. 2012, Section 2.4). We give some interesting examples below.

EXAMPLE 4 (BOUNDED FROBENIUS NORM, I.E., RIDGE REGULARIZATION). The most basic case is $\mathcal{B} = \{B \in \mathbb{R}^{d \times p} : \frac{1}{2}\|B\|_F^2 \leq \beta^2\}$, in which case $F(B) = \frac{1}{2}\|B\|_F^2$ is 1-strongly convex on $\mathbb{R}^{d \times p}$ w.r.t. $\|\cdot\|_F$. Note that

$$\sup_{\boldsymbol{\sigma} \in \{\pm 1\}^d} \sup_{x \in \mathcal{X}} \|\boldsymbol{\sigma} x^T\|_F = \sup_{\boldsymbol{\sigma} \in \{\pm 1\}^d} \|\boldsymbol{\sigma}\|_2 \cdot \sup_{x \in \mathcal{X}} \|x\|_2 = \sqrt{d} \sup_{x \in \mathcal{X}} \|x\|_2.$$

Therefore, if $\frac{1}{2}\|B\|_F^2 \leq \beta^2$ and $\rho_2(\mathcal{X}) := \sup_{x \in \mathcal{X}} \|x\|_2$ we have

$$\mathfrak{R}^n(\mathcal{H}_{\mathcal{B}}) \leq \rho_2(\mathcal{X})\beta\sqrt{\frac{2d}{n}}.$$

□

EXAMPLE 5 (BOUNDED ℓ_1 NORM OF VECTORIZED MATRIX, I.E., LASSO REGULARIZATION). Another case is when $\mathcal{B} = \{B \in \mathbb{R}^{d \times p} : \frac{1}{2}\|B\|_1^2 \leq \beta^2\}$, where $\|B\|_q$ is ℓ_q norm of the vectorized matrix B. We set $F(B) = \frac{1}{2}\|B\|_q^2$ for $q = \frac{\log(pd)}{\log(pd)-1}$ which is $1/(3\log(pd))$-strongly convex w.r.t. $\|\cdot\|_1$ (Kakade et al. 2012, Corollary 10). Since $\|B\|_q \leq \|B\|_1$, we clearly have $F(B) \leq \beta^2$. Note that

$$\sup_{\boldsymbol{\sigma} \in \{\pm 1\}^d} \sup_{x \in \mathcal{X}} \|\boldsymbol{\sigma} x^T\|_\infty = \sup_{\boldsymbol{\sigma} \in \{\pm 1\}^d} \|\boldsymbol{\sigma}\|_\infty \cdot \sup_{x \in \mathcal{X}} \|x\|_\infty = \sup_{x \in \mathcal{X}} \|x\|_\infty.$$

Therefore, if $\frac{1}{2}\|B\|_1^2 \leq \beta^2$ and $\rho_\infty(\mathcal{X}) := \sup_{x\in\mathcal{X}} \|x\|_\infty$ we have

$$\mathfrak{R}^n(\mathcal{H}_\mathcal{B}) \ \leq \ \rho_\infty(\mathcal{X})\beta\sqrt{\frac{6\log(pd)}{n}}.$$

□

EXAMPLE 6 (BOUNDED GROUP-LASSO NORM). In cases where the feature dimension p is large, we might want to encode prior knowledge that only a subset of the p input variables are relevant for making predictions. The vectorized ℓ_1 norm considered in the previous example encourages sparsity but does not result in shared sparsity structure over the rows of B. That is, it does not cause entire columns to be set to zero. In multivariate regression, the group-lasso norm (Tibshirani et al. 2015, Section 4.3) is used to enforce such a structured from of sparsity. Define the norm

$$\|B\|_{2,q} = \left(\sum_{j=1}^p \|B_{\cdot j}\|_2^q\right)^{1/q}.$$

The subscripts above remind us that we first take the ℓ_2 norms of columns $B_{\cdot j}$ and then take the ℓ_q norm of the p resulting values. The group-lasso norm is simply $\|\cdot\|_{2,1}$. Let us consider the case when the matrices B are constrained to have low group-lasso norm, i.e., $\mathcal{B} = \{B \in \mathbb{R}^{d\times p} : \frac{1}{2}\|B\|_{2,1}^2 \leq \beta^2\}$. We set $F(B) = \frac{1}{2}\|B\|_{2,q}^2$ for $q = \frac{\log(p)}{\log(p)-1}$ which is $1/(3\log(p))$-strongly convex w.r.t. $\|\cdot\|_{2,1}$ (Kakade et al. 2012, Corollary 14). Since $\|B\|_{2,q} \leq \|B\|_{2,1}$, we clearly have $F(B) \leq \beta^2$. Note that the dual norm of the $\|\cdot\|_{2,1}$ norm is the $\|\cdot\|_{2,\infty}$ norm and we have

$$\sup_{\boldsymbol{\sigma}\in\{\pm1\}^d}\sup_{x\in\mathcal{X}} \|\boldsymbol{\sigma}x^T\|_{2,\infty} = \sup_{\boldsymbol{\sigma}\in\{\pm1\}^d} \|\boldsymbol{\sigma}\|_2 \cdot \sup_{x\in\mathcal{X}} \|x\|_\infty \leq \sqrt{d}\sup_{x\in\mathcal{X}} \|x\|_\infty.$$

Therefore, if $\frac{1}{2}\|B\|_{2,1}^2 \leq \beta^2$ and $\rho_\infty(\mathcal{X}) := \sup_{x\in\mathcal{X}} \|x\|_\infty$ we have

$$\mathfrak{R}^n(\mathcal{H}_\mathcal{B}) \ \leq \ \rho_\infty(\mathcal{X})\beta\sqrt{\frac{6d\log(p)}{n}}.$$

□

REMARK 4 (DIMENSION DEPENDENCE). The reader will notice that in Examples 4 and 6, the dependence on the output/decision dimension d is $O(\sqrt{d})$ in contrast to the dependence on input/feature dimension p which is either absent (Example 4) or logarithmic (Examples 5 and 6). This is a direct consequence of the fact that multivariate Rademacher complexity involves a d-dimensional *vector* of Radamacher variables *per training example*. In standard Rademacher complexity, the output dimension is 1 and therefore the dependence on d is not manifest. However, it shows up in our more general setting.   □

5. Applications to Strongly Convex and Polyhedral Sets

In this section, we consider two special cases for the feasible region S: *(i)* S is a strongly convex set, and *(ii)* S is polyhedral. In both cases, we show that the distance to degeneracy function $\nu_S(\cdot)$ satisfies the strength property and specify the strength parameter μ. In the strongly convex case, $\nu_S(\hat{c}) = \|\hat{c}\|_*$ and is easily computable whenever the dual norm is. In the polyhedral case, we show that a sufficient condition for computing $\nu_S(\cdot)$ is that S has a known convex hull representation, i.e., $S = \mathrm{conv}\{v_1, \ldots, v_K\}$ where $v_1, \ldots, v_K \in \mathbb{R}^d$ are given vectors. Thus, in these two cases, we can readily compute the empirical margin SPO loss and apply Theorems 4 and 5.

5.1. Strongly Convex Sets

We adopt the classical definition of strongly convex sets (see, e.g., Vial (1983), Journée et al. (2010), Garber and Hazan (2015)), which is reviewed in Definition 5 below. Recall that $\|\cdot\|$ is a generic given norm on $\mathbb{R}^d$ with dual norm $\|\cdot\|_*$, and $B(\bar{w}, r) := \{w : \|w - \bar{w}\| \le r\}$ denotes the ball of radius r centered at $\bar{w}$. Recall also that the normal cone of S at a point $\bar{w} \in S$ is defined by $N_S(\bar{w}) := \{c \in \mathbb{R}^d : c^T(w - \bar{w}) \le 0 \text{ for all } w \in S\}$.

DEFINITION 5 (STRONGLY CONVEX SET). For a constant $\bar{\mu} \ge 0$, we say that a convex set $S \subseteq \mathbb{R}^d$ is $\bar{\mu}$-strongly convex with respect to the norm $\|\cdot\|$ if, for any $w_1, w_2 \in S$ and for any $\lambda \in [0, 1]$, it holds that:

$$B\left(\lambda w_1 + (1 - \lambda) w_2, \left(\tfrac{\bar{\mu}}{2}\right)\lambda(1-\lambda)\|w_1 - w_2\|^2\right) \subseteq S . \quad \square$$

Informally, Definition 5 says that, for every convex combination of points in S, a ball of appropriate radius also lies in S. Several examples of strongly convex sets are presented by Journée et al. (2010) and Garber and Hazan (2015), including ℓ_q and Schatten ℓ_q balls for $q \in (1, 2]$, certain group norm balls, and generally any level set of a smooth and strongly convex function.

Our analysis herein relies on the following proposition, which gives a tighter characterization of the normal cones of strongly convex sets. Proposition 1 can be derived as a consequence of a result of Vial (1983), but we include its proof in Appendix D.1 in our notation for completeness.

PROPOSITION 1 (**Vial (1983), Proposition 2.9**). *Let $S \subseteq \mathbb{R}^d$ be a $\bar{\mu}$-strongly convex set for some $\bar{\mu} \ge 0$. Then, for any $\bar{w} \in S$ it holds that:*

$$N_S(\bar{w}) = \left\{c \in \mathbb{R}^d : c^T(w - \bar{w}) \le -\left(\tfrac{\bar{\mu}}{2}\right)\|c\|_*\|w - \bar{w}\|^2 \text{ for all } w \in S\right\} . \tag{8}$$

We are now ready to state our main theorem in the strongly convex case, which shows that the distance to degeneracy is exactly the dual norm and that (5) is satisfied.

THEOREM 7. *Suppose that S is is not a singleton and is $\bar{\mu}$-strongly convex for some $\bar{\mu} > 0$. Then the distance to degeneracy function satisfies $\nu_S(\hat{c}) = \|\hat{c}\|_*$ for all $\hat{c} \in \mathbb{R}^d$, and the strength property is satisfied with $\mu \leftarrow \bar{\mu}$.*

Proof. To show that $\nu_S(\hat{c}) = \|\hat{c}\|_*$, it suffices to show that $\mathcal{C}^\circ = \{0\}$. Since S is not a singleton, we have that $0 \in \mathcal{C}^\circ$. Conversely, suppose that $\hat{c} \neq 0$. Then, since $\|\hat{c}\|_* > 0$ and $-\hat{c} \in N_S(w^*(\hat{c}))$, Proposition 1 implies that $\hat{c}^T w^*(\hat{c}) < \hat{c}^T w$ for all $w \in S$. Thus $w^*(\hat{c})$ is the unique optimal solution of $P(\hat{c})$ and $\hat{c} \notin \mathcal{C}^\circ$. Then the strength property with parameter $\bar{\mu}$ follows immediately from Proposition 1 and the observation that $-\hat{c} \in N_S(w^*(\hat{c}))$. $\square$

With Theorem 7 in place, we can now apply the margin-based generalization bounds of Theorems 4 and 5 to any strongly convex set. When combined with bounds on the multivariate Rademacher complexity, such as those developed in Section 4.4, we obtain generalization bounds with significantly better dependence on the number of parameters than Corollary 3. As a basic example, we also recover known results in binary classification, which is described in Example 7 below.

EXAMPLE 7. In Elmachtoub and Grigas (2022), it is shown that the SPO loss corresponds exactly to the 0-1 loss in binary classification when $d = 1$, $S = [-1/2, +1/2]$, and $\mathcal{C} = \{-1, +1\}$. In this case, using our notation, the margin value of a prediction $\hat{c}$ is $c\hat{c}$. It is also easily seen that $\omega_S(\mathcal{C}) = \rho_2(\mathcal{C}) = 1$, the γ-margin SPO loss corresponds exactly to the margin loss (or ramp loss) that interpolates between 1 and 0 when $c\hat{c} \in [0, \gamma]$ Furthermore, note that the interval $S = [-\frac{1}{2}, +\frac{1}{2}]$ is 2-strongly convex Garber and Hazan (2015). In this case, the right hand side of the generalization bound of Theorem 4, for example, can be upper bounded by $O(\mathfrak{R}^n(\mathcal{H})/\gamma + \sqrt{\log(1/\delta)/n})$, where the $O(\cdot)$ notation hides absolute constants. This recovers a well known result in binary classification (Koltchinskii and Panchenko 2002) and, altogether, Theorems 4, 5, and 7 generalize the well-known results on margin guarantees based on Rademacher complexity for binary classification. $\square$

5.2. Polyhedral Sets

Let us now assume that S is polyhedral with a known convex hull representation, i.e., $S = \text{conv}\{v_1, \ldots, v_K\}$ where $v_1, \ldots, v_K \in \mathbb{R}^d$ are given vectors. We allow redundancies in the convex hull representation (i.e., it may be possible to express v_j as a convex combination of other points in the basis), but for simplicity and ease of exposition we do require that $v_i \neq v_j$ for $i \neq j$.

Let $\mathcal{K}_j := -N_S(v_j)$ be the negative normal cone at v_j for $j = 1, \ldots, K$. Equivalently, we have that

$$\mathcal{K}_j = \{\hat{c} \in \mathbb{R}^d : \hat{c}^T(w - v_j) \geq 0 \text{ for all } w \in S\} = \{\hat{c} \in \mathbb{R}^d : \hat{c}^T(v_i - v_j) \geq 0 \text{ for all } i = 1, \ldots, K\}, \quad (9)$$

and hence $\mathcal{K}_j$ is a polyhedral cone. Note also that, by the fundamental theorem of linear optimization, we have that $\cup_{j=1}^K \mathcal{K}_j = \mathbb{R}^d$. Moreover, $\mathcal{K}_1, \ldots, \mathcal{K}_K$ form part of a polyhedral complex called the normal fan of a polytope (Ziegler 2012). In our context, we can use $\mathcal{K}_1, \ldots, \mathcal{K}_K$ to characterize the set $\mathcal{C}^\circ$ and subsequently compute the distance to degeneracy function $\nu_S(\cdot)$. The following proposition exactly characterizes $\mathcal{C}^\circ$ in terms of $\mathcal{K}_1, \ldots, \mathcal{K}_K$.

PROPOSITION 2. *Suppose that S is polyhedral with $S = conv\{v_1, \ldots, v_K\}$. Then, for any $\hat{c} \in \mathbb{R}^d$, $P(\hat{c})$ has a unique optimal solution if and only if there exists $j \in \{1, \ldots, K\}$ such that $\hat{c} \in \mathrm{int}(\mathcal{K}_j)$. Subsequently, it holds that $\mathcal{C}^\circ = \mathbb{R}^d \setminus \cup_{j=1}^{K} \mathrm{int}(\mathcal{K}_j)$.*

Proof. If $P(\hat{c})$ has a unique optimal solution, then $w^*(\hat{c}) = v_j$ is the unique optimal solution for some $j \in \{1, \ldots, K\}$. Furthermore, we have that $\hat{c}^T v_j < \hat{c}^T v_i$ for all $i \in \{1, \ldots, K\}$, $i \neq j$, which implies that $\hat{c} \in \mathrm{int}(\mathcal{K}_j)$. Conversely, if $\hat{c} \in \mathrm{int}(\mathcal{K}_j)$ then (9) implies that $\hat{c}^T(v_i - v_j) > 0$ for all $i \neq j$. Therefore, since any $w \in S$ is a convex combination of $v_1, \ldots, v_K$, it must be the case that v_j is the unique optimal solution of $P(\hat{c})$. Note that we have shown that $\hat{c}$ is in the complement of $\mathcal{C}^\circ$ if and only if $\hat{c} \in \cup_{j=1}^{K} \mathrm{int}(\mathcal{K}_j)$, hence $\mathcal{C}^\circ = \mathbb{R}^d \setminus \cup_{j=1}^{K} \mathrm{int}(\mathcal{K}_j)$. □

We are now ready to state our main theorem in the polyhedral case, which gives an exact formula for the distance to degeneracy function $\nu_S(\cdot)$ in terms of $v_1, \ldots, v_K$ and shows that the strength property is satisfied. Recall that the diameter of S in the given norm is defined by $\Delta(S) := \sup_{w_1, w_2 \in S} \|w_1 - w_2\|$, which in the polyhedral case satisfies $\Delta(S) = \max_{i,j \in \{1, \ldots, K\}} \|v_i - v_j\|$.

THEOREM 8. *Suppose that S is not a singleton and is polyhedral with $S = conv\{v_1, \ldots, v_K\}$. Then the distance to degeneracy function satisfies*

$$\nu_S(\hat{c}) = \min_{j:v_j \neq w^*(\hat{c})} \left\{ \frac{\hat{c}^T(v_j - w^*(\hat{c}))}{\|v_j - w^*(\hat{c})\|} \right\} \quad \text{for all } \hat{c} \in \mathbb{R}^d , \tag{10}$$

and the strength property in (5) is satisfied with $\mu \leftarrow \frac{2}{\Delta(S)}$.

Proof. Let us first consider the case where $\hat{c} \in \mathcal{C}^\circ$, and hence $\nu_S(\hat{c}) = 0$. Then, since there are multiple optimal solutions of $P(\hat{c})$, there exists v_j that is an optimal solution of $P(\hat{c})$ with $v_j \neq w^*(\hat{c})$. Thus, the right side of (10) is equal to zero, which matches $\nu_S(\hat{c})$.

Let us now consider the case where $\hat{c} \notin \mathcal{C}^\circ$. Since $\nu_S(\cdot)$ is a distance function in $\mathbb{R}^d$, we have for any $\hat{c} \notin \mathcal{C}^\circ$ that $\nu_S(\hat{c}) = \mathrm{dist}^*_{\mathcal{C}^\circ}(\hat{c}) = \sup_r \{r : B_*(\hat{c}, r) \subseteq \mathbb{R}^d \setminus \mathcal{C}^\circ\}$. Thus, by Proposition 2, we have that $\nu_S(\hat{c}) = \sup_r \{r : B_*(\hat{c}, r) \subseteq \cup_{j=1}^{K} \mathrm{int}(\mathcal{K}_j)\}$ for $\hat{c} \notin \mathcal{C}^\circ$.

In the remainder of the proof, we let $j^*(\cdot) : \mathbb{R}^d \to \{1, \ldots, K\}$ be any mapping such that $v_{j^*(\hat{c})}$ is an optimal solution of $P(\hat{c})$ for all $\hat{c} \in \mathbb{R}^d$. Now, for any $r > 0$ and $\hat{c} \notin \mathcal{C}^\circ$, using $\nu_S(\hat{c}) = \sup_r \{r : B_*(\hat{c}, r) \subseteq \cup_{j=1}^{K} \mathrm{int}(\mathcal{K}_j)\}$ we have the following chain of equivalences:

$$\nu_S(\hat{c}) \geq r \iff B_*(\hat{c}, r - \epsilon) \subseteq \cup_{j=1}^{K} \mathrm{int}(\mathcal{K}_j) \text{ for all } \epsilon \in (0, r]$$

$$\iff B_*(\hat{c}, r - \epsilon) \subseteq \mathrm{int}(\mathcal{K}_{j^*(\hat{c})}) \text{ for all } \epsilon \in (0, r]$$

$$\iff B_*(\hat{c}, r) \subseteq \mathcal{K}_{j^*(\hat{c})}$$

$$\iff (\hat{c} + \Delta)^T(v_j - v_{j^*(\hat{c})}) \geq 0 \text{ for all } j = 1, \ldots, K \text{ and } \Delta \text{ s.t. } \|\Delta\|_* \leq r$$

$$\iff \hat{c}^T(v_j - v_{j^*(\hat{c})}) - r\|v_j - v_{j^*(\hat{c})}\| \geq 0 \text{ for all } j = 1, \ldots, K$$

$$\iff r \leq \min_{j \neq j^*(\hat{c})} \left\{ \frac{\hat{c}^T(v_j - v_{j^*(\hat{c})})}{\|v_j - v_{j^*(\hat{c})}\|} \right\} = \min_{j:v_j \neq w^*(\hat{c})} \left\{ \frac{\hat{c}^T(v_j - w^*(\hat{c}))}{\|v_j - w^*(\hat{c})\|} \right\} .$$

where the second equivalence holds since $\cup_{j=1}^{K} \text{int}(\mathcal{K}_j)$ is a disjoint union of open sets, the third equivalence holds since $\mathcal{K}_{j*(\hat{c})}$ is a closed set, the fifth equivalence uses the fact that $\|\cdot\|$ is the dual norm of $\|\cdot\|_*$, and the equality at the end uses the fact that $v_{j*(\hat{c})} = w^*(\hat{c})$ since $\hat{c} \notin \mathcal{C}^\circ$. Since $\mathcal{C}^\circ$ is a closed set by Proposition 2, we have that $\nu_S(\hat{c}) > 0$. Furthermore, since $w^*(\hat{c})$ is the unique optimal solution of $P(\hat{c})$, we have that $\min_{j:v_j \neq w^*(\hat{c})} \left\{ \frac{\hat{c}^T(v_j - w^*(\hat{c}))}{\|v_j - w^*(\hat{c})\|} \right\} > 0$. Therefore, applying the above chain of equivalences twice with both $r \leftarrow \nu_S(\hat{c})$ and $r \leftarrow \min_{j:v_j \neq w^*(\hat{c})} \left\{ \frac{\hat{c}^T(v_j - w^*(\hat{c}))}{\|v_j - w^*(\hat{c})\|} \right\}$ yields (10).

Now, to see that (5) (strength property) holds, let $\hat{c} \in \mathbb{R}^d$ and $w \in S$ be arbitrary. Then, there exist nonnegative scalars $\alpha_1, \ldots, \alpha_K$ such that $w = \sum_{j=1}^{K} \alpha_j v_j$ and $\sum_{j=1}^{K} \alpha_j = 1$. By (10), it holds that

$$\hat{c}^T(v_j - w^*(\hat{c})) \geq \nu_S(\hat{c})\|v_j - w^*(\hat{c})\| \quad \text{for all } j = 1, \ldots, K \; . \tag{11}$$

Now we have

$$
\begin{aligned}
\hat{c}^T(w - w^*(\hat{c})) &= \sum_{j=1}^{K} \alpha_j \hat{c}^T(v_j - w^*(\hat{c})) \\
&\geq \nu_S(\hat{c}) \sum_{j=1}^{K} \alpha_j \|v_j - w^*(\hat{c})\| \\
&\geq \nu_S(\hat{c}) \|w - w^*(\hat{c})\| \\
&\geq \left(\frac{\nu_S(\hat{c})}{\Delta(S)} \right) \|w - w^*(\hat{c})\|^2 \; ,
\end{aligned}
$$

where the first inequality uses (11), the second uses convexity, and the third uses the definition of the diameter $\Delta(S)$. $\quad\square$

With Theorem 8 in place, we can now apply the margin-based generalization bounds of Theorems 4 and 5 to any polyhedral set with a known convex hull representation. The most basic example of such a set is the unit simplex, which corresponds to the case of multiclass-classification as described in Examples 3 and 8. More generally, when combined with bounds on the multivariate Rademacher complexity, such as those developed in Section 4.4, we obtain generalization bounds in the polyhedral case that may have significantly better dependence on the number of parameters than Corollary 2.

EXAMPLE 8 (CONTINUATION OF EXAMPLE 3). Recall that $S := \{w \in \mathbb{R}^d : \sum_{j=1}^{d} w_j = 1, w \geq 0\}$ and, using the ℓ_2-norm, we have that $\Delta(S) = 2$. Additionally, $\mathcal{C} = \{-e_i | i = 1, \ldots, d\}$ and therefore $\rho_2(\mathcal{C}) = \omega_S(\mathcal{C}) = 1$. In this case, the right hand side of the generalization bound of Theorem 4, for example, can be upper bounded by $O(\mathfrak{R}^n(\mathcal{H})/\gamma + \sqrt{\log(1/\delta)/n})$, which recovers a well known result in multi-class classification (see, e.g., Li et al. (2018) and the references therein). Altogether, Theorems 4, 5, and 7 generalize known results on margin guarantees based on Rademacher complexity for multi-class classification. $\quad\square$

6. Conclusions and Future Directions

Our work extends learning theory, as developed for binary and multiclass classification, to predict-then-optimize problems in two very significant directions: (i) obtaining worst-case generalization bounds using combinatorial parameters that measure the capacity of function classes, and (ii) exploiting special structure in data by deriving margin-based generalization bounds that scale more gracefully w.r.t. problem dimensions. It also motivates several interesting avenues for future work. Beyond the margin theory, other aspects of the problem that lead to improvements over worst case rates should be studied. In this respect, developing a theory of local Rademacher complexity for predict-then-optimize problems would be a promising approach. Developing a theory of surrogate losses, especially convex ones, that are calibrated w.r.t. the non-convex SPO loss will also be extremely important. Finally, the assumption that the optimization objective is linear could be relaxed to include non-linear objectives, in which case a new framework and methodology would be required.

References

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

Ban, Gah-Yi, Cynthia Rudin. 2019. The big data newsvendor: Practical insights from machine learning. *Operations Research* **67**(1) 90–108.

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

Berthet, Quentin, Mathieu Blondel, Olivier Teboul, Marco Cuturi, Jean-Philippe Vert, Francis Bach. 2020. Learning with differentiable pertubed optimizers. *Advances in neural information processing systems* **33** 9508–9519.

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

Chu, Hongrui, Wensi Zhang, Pengfei Bai, Yahong Chen. 2021. Data-driven optimization for last-mile delivery. *Complex & Intelligent Systems* 1–14.

Daniely, Amit, Sivan Sabato, Shai Ben-David, Shai Shalev-Shwartz. 2015. Multiclass learnability and the erm principle. *The Journal of Machine Learning Research* **16**(1) 2377–2404.

Daniely, Amit, Shai Shalev-Shwartz. 2014. Optimal learners for multiclass problems. *Conference on Learning Theory*. 287–316.

Demirovic, Emir, Peter Stuckey, James Bailey, Jeffrey Chan, Christopher Leckie, Kotagiri Ramamohanarao, Tias Guns. 2020. Dynamic programming for predict+ optimise. *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI 2020)*. 1444–1451.

Demirovic, Emir, Peter J Stuckey, James Bailey, Jeffrey Chan, Christopher Leckie, Kotagiri Ramamohanarao, Tias Guns. 2019. Predict+ optimise with ranking objectives: Exhaustively learning linear functions. *IJCAI-19* 1078–1085.

Donti, Priya, Brandon Amos, J Zico Kolter. 2017. Task-based end-to-end model learning in stochastic optimization. *Advances in Neural Information Processing Systems*. 5484–5494.

El Balghiti, Othman, Adam N Elmachtoub, Paul Grigas, Ambuj Tewari. 2019. Generalization bounds in the predict-then-optimize framework. *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. 14412–14421.

Elmachtoub, Adam, Jason Cheuk Nam Liang, Ryan McNellis. 2020. Decision trees for decision-making under the predict-then-optimize framework. *International Conference on Machine Learning*. PMLR, 2858–2867.

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

Ferber, Aaron, Bryan Wilder, Bistra Dilkina, Milind Tambe. 2020. Mipaal: Mixed integer program as a layer. *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34. 1504–1511.

Foster, Dylan J, Alexander Rakhlin. 2019. ℓ_∞ vector contraction for rademacher complexity. *arXiv preprint arXiv:1911.06468* .

Garber, Dan, Elad Hazan. 2015. Faster rates for the frank-wolfe method over strongly-convex sets. *International Conference on Machine Learning*. PMLR, 541–549.

Guermeur, Yann. 2007. Vc theory of large margin multi-category classifiers. *Journal of Machine Learning Research* **8**(Nov) 2551–2594.

Ho, Chin Pang, Grani A Hanasusanto. 2019. On data-driven prescriptive analytics with side information: A regularized nadaraya-watson approach. *URL: http://www. optimization-online. org/DB FILE/2019/01/7043. pdf* .

Ho-Nguyen, Nam, Fatma Kılınç-Karzan. 2022. Risk guarantees for end-to-end prediction and optimization processes. *Management Science* .

Hu, Yichun, Nathan Kallus, Xiaojie Mao. 2022. Fast rates for contextual linear optimization. *Management Science* .

Journée, Michel, Yurii Nesterov, Peter Richtárik, Rodolphe Sepulchre. 2010. Generalized power method for sparse principal component analysis. *Journal of Machine Learning Research* **11**(Feb) 517–553.

Kakade, Sham M., Shai Shalev-Shwartz, Ambuj Tewari. 2012. Regularization techniques for learning with matrices. *Journal of Machine Learning Research* **13** 1865–1890.

Kakade, Sham M, Karthik Sridharan, Ambuj Tewari. 2008. On the complexity of linear prediction: Risk bounds, margin bounds, and regularization. *Advances in neural information processing systems*. 793–800.

Kallus, Nathan, Xiaojie Mao. 2020. Stochastic optimization forests. *arXiv preprint arXiv:2008.07473* .

Kao, Yi-hao, Benjamin V Roy, Xiang Yan. 2009. Directed regression. *Advances in Neural Information Processing Systems*. 889–897.

Koltchinskii, Vladimir, Dmitriy Panchenko, Fernando Lozano. 2001. Some new bounds on the generalization error of combined classifiers. *Advances in neural information processing systems* 245–251.

Koltchinskii, Vladimir, Dmitry Panchenko. 2002. Empirical margin distributions and bounding the generalization error of combined classifiers. *The Annals of Statistics* **30**(1) 1–50.

Kotary, James, Ferdinando Fioretto, Pascal Van Hentenryck, Bryan Wilder. 2021. End-to-end constrained optimization learning: A survey. *arXiv preprint arXiv:2103.16378* .

Lei, Yunwen, Urun Dogan, Alexander Binder, Marius Kloft. 2015. Multi-class svms: From tighter data-dependent generalization bounds to novel algorithms. *Advances in Neural Information Processing Systems*. 2035–2043.

Li, Jian, Yong Liu, Rong Yin, Hua Zhang, Lizhong Ding, Weiping Wang. 2018. Multi-class learning: From theory to algorithm. *Advances in Neural Information Processing Systems 31*. 1586–1595.

Liu, Heyuan, Paul Grigas. 2021. Risk bounds and calibration for a smart predict-then-optimize method. *Advances in Neural Information Processing Systems* **34**.

Loke, Gar Goei, Qinshen Tang, Yangge Xiao. 2021. Decision-driven regularization: A blended model for predict-then-optimize. *Available at SSRN 3623006* .

Mandi, Jayanta, Tias Guns. 2020. Interior point solving for lp-based prediction+ optimisation. *Advances in Neural Information Processing Systems* **33** 7272–7282.

Mandi, Jayanta, Peter J Stuckey, Tias Guns, et al. 2020. Smart predict-and-optimize for hard combinatorial optimization problems. *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34. 1603–1610.

Maurer, Andreas. 2016. A vector-contraction inequality for rademacher complexities. *International Conference on Algorithmic Learning Theory*. Springer, 3–17.

Mohri, Mehryar, Afshin Rostamizadeh, Ameet Talwalkar. 2018. *Foundations of machine learning*. 2nd ed. MIT press.

Natarajan, Balas K. 1989. On learning sets and functions. *Machine Learning* **4**(1) 67–97.

Pogančić, Marin Vlastelica, Anselm Paulus, Vit Musil, Georg Martius, Michal Rolinek. 2019. Differentiation of blackbox combinatorial solvers. *International Conference on Learning Representations*.

Rockafellar, R Tyrrell, Roger J-B Wets. 2009. *Variational analysis*, vol. 317. Springer Science & Business Media.

Shalev-Shwartz, Shai, Shai Ben-David. 2014. *Understanding machine learning: From theory to algorithms*. Cambridge university press.

Tibshirani, Robert, Martin Wainwright, Trevor Hastie. 2015. *Statistical learning with sparsity: the lasso and generalizations*. Chapman and Hall/CRC.

Vial, Jean-Philippe. 1983. Strong and weak convexity of sets and functions. *Mathematics of Operations Research* **8**(2) 231–259.

Wang, Kai, Bryan Wilder, Andrew Perrault, Milind Tambe. 2020. Automatically learning compact quality-aware surrogates for optimization problems. *Advances in Neural Information Processing Systems* **33** 9586–9596.

Wilder, Bryan, Bistra Dilkina, Milind Tambe. 2019a. Melding the data-decisions pipeline: Decision-focused learning for combinatorial optimization. *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33. 1658–1665.

Wilder, Bryan, Eric Ewing, Bistra Dilkina, Milind Tambe. 2019b. End to end learning and optimization on graphs. *Advances in Neural Information Processing Systems*. 4674–4685.

Yan, Ran, Shuaian Wang, Kjetil Fagerholt. 2020. A semi-"smart predict then optimize"(semi-spo) method for efficient ship inspection. *Transportation Research Part B: Methodological* **142** 100–125.

Zatarain-Vera, Oscar. 2019. A vector-contraction inequality for rademacher complexities using p-stable variables. *arXiv preprint arXiv:1912.10136* .

Ziegler, Günter M. 2012. *Lectures on polytopes*, vol. 152. Springer Science & Business Media.

Appendix

A.   Proofs for Section 2

A.1.   Proof of Corollary 1

Proof.   Let $f^* \in \arg\min_{f \in \mathcal{H}} R_{\mathrm{SPO}}(f)$. Then,

$$R_{\mathrm{SPO}}(\hat{f}_n) - R_{\mathrm{SPO}}(f^*) = R_{\mathrm{SPO}}(\hat{f}_n) - \hat{R}_{\mathrm{SPO}}(\hat{f}_n) + \hat{R}_{\mathrm{SPO}}(\hat{f}_n) - \hat{R}_{\mathrm{SPO}}(f^*) + \hat{R}_{\mathrm{SPO}}(f^*) - R_{\mathrm{SPO}}(f^*)$$

$$\leq R_{\mathrm{SPO}}(\hat{f}_n) - \hat{R}_{\mathrm{SPO}}(\hat{f}_n) + \hat{R}_{\mathrm{SPO}}(f^*) - R_{\mathrm{SPO}}(f^*)$$

The inequality follows from the fact that $\hat{f}_n$ minimizes the empirical risk, i.e., $\hat{f}_n \in \arg\min_{f \in \mathcal{H}} \hat{R}_{\mathrm{SPO}}(f)$. From Theorem 1, we know that

$$R_{\mathrm{SPO}}(\hat{f}_n) - \hat{R}_{\mathrm{SPO}}(\hat{f}_n) \leq 2\mathfrak{R}_{\mathrm{SPO}}^n(\mathcal{H}) + \omega_S(\mathcal{C})\sqrt{\frac{\log(2/\delta)}{2n}}$$

with probability at least $1 - \frac{\delta}{2}$. From From Hoeffding's inequality, we have that

$$\hat{R}_{\mathrm{SPO}}(f^*) - R_{\mathrm{SPO}}(f^*) \le \omega_S(\mathcal{C}) \sqrt{\frac{\log(2/\delta)}{2n}}$$

with probability at least $1 - \frac{\delta}{2}$. Thus, with probability at least $1 - \delta$ we can combine the three previous inequalities and obtain the desired result. $\square$

B. Proofs for Section 3

B.1. Proof of Theorem 2

Proof. The proof is along the lines of Corollary 3.8 in Mohri et al. (2018). Fix a sample of data $\mathcal{S}_n = (\mathbb{X}, \mathbb{C}) \in (\mathcal{X}, \mathcal{C})^n$, where $\mathbb{X} = (x_1, \ldots, x_n)$ and $\mathbb{C} = (c_1, \ldots, c_n)$. Let $\mathfrak{F}_{|\mathbb{X}} := \{(w^*(f(x_1)), \ldots, w^*(f(x_n))) : f \in \mathcal{H}\}$. From the definition of empirical Rademacher complexity, we have that

$$\hat{\mathfrak{R}}_{\mathrm{SPO}}^n(\mathcal{H}) = \mathbb{E}_\sigma \left[\sup_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^n \sigma_i \ell_{\mathrm{SPO}}(f(x_i), c_i) \right]$$

$$= \mathbb{E}_\sigma \left[\sup_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^n \sigma_i c_i^T (w^*(f(x_i)) - w^*(c_i)) \right]$$

$$= \mathbb{E}_\sigma \left[\sup_{(w_1, \ldots, w_n) \in \mathfrak{F}_{|\mathbb{X}}} \frac{1}{n} \sum_{i=1}^n \sigma_i c_i^T (w_i - w^*(c_i)) \right]$$

$$\le \omega_S(\mathcal{C}) \sqrt{\frac{2 \log |\mathfrak{F}_{|\mathbb{X}}|}{n}}$$

$$\le \omega_S(\mathcal{C}) \sqrt{\frac{2 d_N(w^*(\mathcal{H})) \log(n|\mathfrak{S}|^2)}{n}}$$

where the first inequality is directly due to Massart's lemma and the definition of $\omega_S(\mathcal{C})$ and the second inequality follows from the Natarajan Lemma (see Lemma 29.4 in Shalev-Shwartz and Ben-David (2014)). When using the Natarajan Lemma, we observe that there $w^*(\cdot)$ has at most $|\mathfrak{S}|$ possible values w.l.o.g. by the equivalence of optimizing over S to optimizing over $\mathfrak{S}$. The bound for the expected version of the Rademacher complexity follows immediately from the bound on the empirical Rademacher complexity. Applying this bound with Theorem 1 concludes the proof. $\square$

B.2. Proof of Corollary 2

Proof. We will prove that $w^*(\mathcal{H}_{\mathrm{lin}})$ is an instance of a linear multiclass predictor for a particular class-sensitive feature mapping Ψ. In our application of linear multiclass predictors, let $\Psi : \mathcal{X} \times S \mapsto \mathbb{R}^{d \times p}$ be a function that takes as input a feature vector and an extreme point and outputs a matrix and let

$$\mathcal{H}_\Psi = \{x \mapsto \underset{w \in S}{\mathrm{argmax}} \langle B, \Psi(x, w) \rangle : B \in \mathbb{R}^{d \times p}\}.$$

We will show that, for $\Psi(x, w) = wx^T$, we have that $w^*(\mathcal{H}_{\mathrm{lin}}) \subseteq \mathcal{H}_\Psi$. Consider any $f \in \mathcal{H}_{\mathrm{lin}}$ and the associated matrix B_f. Then

$$w^*(B_f x) \in \underset{w \in S}{\mathrm{argmin}} (B_f x)^T w$$

$$= \underset{w \in S}{\mathrm{argmax}} - \mathrm{Tr}\left((B_f x)^T w\right)$$

$$= \underset{w \in S}{\mathrm{argmax}} - \mathrm{Tr}\left(B_f^T w x^T\right)$$

$$= \underset{w \in S}{\mathrm{argmax}} \langle -B_f, wx^T \rangle.$$

Thus, it is clear that for $\Psi(x,w) = wx^T$, choosing the function in $\mathcal{H}_\Psi$ corresponding to $-B_f$ yields exactly the function f. Therefore $w^*(\mathcal{H}_{\text{lin}}) \subseteq \mathcal{H}_\Psi$. Theorem 7 in Daniely and Shalev-Shwartz (2014) shows that $d_N(\mathcal{H}_\Psi) \leq dp$. Since $w^*(\mathcal{H}_{\text{lin}}) \subseteq \mathcal{H}_\Psi$, then $d_N(w^*(\mathcal{H}_{\text{lin}})) \leq dp$. Combining this bound on the Natarajan dimension with Theorem 2 concludes the proof. □

B.3. Proof of Corollary 3

Proof. Consider the smallest cardinality ϵ-covering of the feasible region S by Euclidean balls of radius ϵ. From Example 27.1 in Shalev-Shwartz and Ben-David (2014), the number of balls needed is at most $\left(\frac{2\rho_2(S)\sqrt{d}}{\epsilon}\right)^d$. Let the set $\tilde{\mathfrak{S}}$ denote the centers of the balls from the smallest cardinality covering. Then it immediately follows that

$$|\tilde{\mathfrak{S}}| \leq \left(\frac{2\rho_2(S)\sqrt{d}}{\epsilon}\right)^d. \tag{12}$$

Finally, let the function $\tilde{w}: S \mapsto \{1,\ldots,|\tilde{\mathfrak{S}}|\}$ be the function that takes a feasible solution in S and maps it to the closest point in in $\tilde{\mathfrak{S}}$.

We can bound the empirical Rademacher complexity by

$$\begin{aligned}
\hat{\mathfrak{R}}^n_{\text{SPO}}(\mathcal{H}) &= \mathbb{E}_\sigma\left[\sup_{f\in\mathcal{H}}\frac{1}{n}\sum_{i=1}^n \sigma_i\ell_{\text{SPO}}(f(x_i),c_i)\right]\\
&= \mathbb{E}_\sigma\left[\sup_{f\in\mathcal{H}}\frac{1}{n}\sum_{i=1}^n \sigma_i c_i^T(w^*(f(x_i))-w^*(c_i))\right]\\
&= \mathbb{E}_\sigma\left[\sup_{f\in\mathcal{H}}\frac{1}{n}\sum_{i=1}^n \sigma_i c_i^T\left[w^*(f(x_i))-\tilde{w}(w^*(f(x_i)))+\tilde{w}(w^*(f(x_i)))-w^*(c_i)\right]\right]\\
&\leq \mathbb{E}_\sigma\left[\sup_{f\in\mathcal{H}}\frac{1}{n}\sum_{i=1}^n \sigma_i c_i^T\left[w^*(f(x_i))-\tilde{w}(w^*(f(x_i)))\right]\right]+\mathbb{E}_\sigma\left[\sup_{f\in\mathcal{H}}\frac{1}{n}\sum_{i=1}^n \sigma_i c_i^T\left[\tilde{w}(w^*(f(x_i)))-w^*(c_i)\right]\right]\\
&\leq 2\epsilon\rho_2(\mathcal{C})+\mathbb{E}_\sigma\left[\sup_{f\in\mathcal{H}}\frac{1}{n}\sum_{i=1}^n \sigma_i c_i^T\left[\tilde{w}(w^*(f(x_i)))-w^*(c_i)\right]\right]\\
&\leq 2\epsilon\rho_2(\mathcal{C})+(\omega_S(\mathcal{C})+2\epsilon\rho_2(\mathcal{C}))\sqrt{\frac{2d_N(\tilde{w}(w^*(\mathcal{H})))\log(n|\tilde{\mathfrak{S}}|^2)}{n}}
\end{aligned}\tag{13}$$

The first inequality follows from the triangle inequality. The second inequality follows from the fact that $w^*(f(x_i))$ and $\tilde{w}(w^*(f(x_i)))$ are at most 2ϵ away by the definition of $\tilde{w}$. In the worst case, the difference is in the direction of c_i, and $c_i^T[w^*(f(x_i))-\tilde{w}(w^*(f(x_i)))] \leq 2\epsilon\|c_i\| \leq 2\epsilon\rho_2(\mathcal{C})$. The third inequality follows from the same exact argument as that in Theorem 2 with two additional observations. First, the maximum value of $c_i^T[\tilde{w}(w^*(f(x_i)))-w^*(c_i)]$ is $\omega_S(\mathcal{C})+2\epsilon\rho_2(\mathcal{C})$ using a similar reasoning as in the second inequality. Second, the number of possible values for $\tilde{w}(\cdot)$ is at most $|\tilde{\mathfrak{S}}|$ by definition.

Thus, all that remains is to bound $d_N(\tilde{w}(w^*(\mathcal{H})))$. To do this, we first observe that $d_N(w^*(\mathcal{H})) \leq dp$, where the proof follows exactly that of Corollary 2. Finally, we observe that

$$d_N(\tilde{w}(w^*(\mathcal{H}))) \leq d_N(w^*(\mathcal{H})) \leq dp \tag{14}$$

since $\tilde{w}$ is simply a deterministic function, and thus the number of dichotomies (labelings) that can be generated by $\tilde{w}(w^*(\mathcal{H}))$ is at most that of $w^*(\mathcal{H})$.

Now setting $\epsilon = \frac{1}{n}$, and combining Eq. (12), Eq. (13), and Eq. (14) yields

$$\hat{\mathfrak{R}}_{\mathrm{SPO}}^n(\mathcal{H}) \leq \frac{2\rho_2(\mathcal{C})}{n}\left(1 + \sqrt{\frac{2dp\log(n(2n\rho_2(S)\sqrt{d})^{2d})}{n}}\right) + \omega_S(\mathcal{C})\sqrt{\frac{2dp\log(n(2n\rho_2(S)\sqrt{d})^{2d})}{n}}$$

$$\leq \frac{2\rho_2(\mathcal{C})}{n}\left(1 + 2d\sqrt{\frac{2p\log(2n\rho_2(S)d)}{n}}\right) + 2d\omega_S(\mathcal{C})\sqrt{\frac{2p\log(2n\rho_2(S)d)}{n}}. \qquad (15)$$

Finally, combining Eq. (15) with Theorem 1 yields

$$R_{\mathrm{SPO}}(f) \leq \hat{R}_{\mathrm{SPO}}(f) + 4d\omega_S(\mathcal{C})\sqrt{\frac{2p\log(2n\rho_2(S)d)}{n}} + \omega_S(\mathcal{C})\sqrt{\frac{\log(1/\delta)}{2n}} + O\left(\frac{1}{n}\right).$$

$\square$

C.   Proofs for Section 4

C.1.   Proof of Theorem 5

Proof.   The argument here follows closely the proof of Theorem 5.9 of Mohri et al. (2018). Define $\epsilon := \omega_S(\mathcal{C})\sqrt{\frac{\log(2/\delta)}{2n}}$ and two sequences $\{\gamma_k\}_{k=1}^\infty$ and $\{\epsilon_k\}_{k=1}^\infty$ by

$$\epsilon_k := \epsilon + \omega_S(\mathcal{C})\sqrt{\frac{\log(k)}{n}} \ , \ \text{ and } \ \gamma_k := \frac{\bar{\gamma}}{2^k} \ , \ \text{ for } k \geq 1 \ .$$

Define the following events:

$$A_k := \left\{\sup_{f \in \mathcal{H}}\left\{R_{\mathrm{SPO}}(f) - \hat{R}_{\mathrm{SPO}}^{\gamma_k}(f) - \frac{2\sqrt{2}(\rho_2(\mathcal{C}) + \mu \cdot \omega_S(\mathcal{C}))\mathfrak{R}^n(\mathcal{H})}{\gamma_k\mu} - \epsilon_k\right\} > 0\right\} \text{ for } k \geq 1 \ ,$$

$$\tilde{A} := \bigcup_{k=1}^\infty A_k \ , \ \text{ and }$$

$$\check{A} := \left\{\sup_{f \in \mathcal{H}, \gamma \in (0,\bar{\gamma}]}\left\{R_{\mathrm{SPO}}(f) - \hat{R}_{\mathrm{SPO}}^{\gamma}(f) - \frac{4\sqrt{2}(\rho_2(\mathcal{C}) + \mu \cdot \omega_S(\mathcal{C}))\mathfrak{R}^n(\mathcal{H})}{\gamma\mu} - \omega_S(\mathcal{C})\sqrt{\frac{\log(\log_2(2\bar{\gamma}/\gamma))}{n}} - \epsilon\right\} > 0\right\} \ .$$

Let us first argue that $\check{A} \subseteq \tilde{A}$. Indeed, suppose that $\check{A}$ occurs. Then, there exists some $f \in \mathcal{H}$ and some $\gamma \in (0, \bar{\gamma}]$ such that:

$$R_{\mathrm{SPO}}(f) - \hat{R}_{\mathrm{SPO}}^{\gamma}(f) - \frac{4\sqrt{2}(\rho_2(\mathcal{C}) + \mu \cdot \omega_S(\mathcal{C}))\mathfrak{R}^n(\mathcal{H})}{\gamma\mu} - \omega_S(\mathcal{C})\sqrt{\frac{\log(\log_2(2\bar{\gamma}/\gamma))}{n}} - \epsilon > 0 \ . \qquad (16)$$

By definition of the sequence $\{\gamma_k\}$, there exists $k \geq 1$ such that $\gamma_k \leq \gamma \leq 2\gamma_k$. Thus, $\gamma_k \leq \gamma$ implies that $\hat{R}_{\mathrm{SPO}}^{\gamma_k}(f) \leq \hat{R}_{\mathrm{SPO}}^{\gamma}(f)$. Moreover, $\gamma \leq 2\gamma_k$ implies that $-1/\gamma_k \geq -2/\gamma$, $k \leq \log_2(2\bar{\gamma}/\gamma)$, and thus

$$\epsilon_k = \epsilon + \omega_S(\mathcal{C})\sqrt{\frac{\log(k)}{n}} \leq \epsilon + \omega_S(\mathcal{C})\sqrt{\frac{\log(\log_2(2\bar{\gamma}/\gamma))}{n}} \ .$$

Now, combining the previous inequalities together with (16) yields:

$$R_{\mathrm{SPO}}(f) - \hat{R}_{\mathrm{SPO}}^{\gamma_k}(f) - \frac{2\sqrt{2}(\rho_2(\mathcal{C}) + \mu \cdot \omega_S(\mathcal{C}))\mathfrak{R}^n(\mathcal{H})}{\gamma_k\mu} - \epsilon_k > 0 \ ,$$

which means that the event A_k and correspondingly the event $\tilde{A}$ have occurred.

Now, for each $k \geq 1$, we apply Theorem 4 using $\gamma \leftarrow \gamma_k$ and $\delta \leftarrow \exp((-2n\epsilon_k^2)/\omega_S(\mathcal{C})^2)$, which yields $\mathbb{P}(A_k) \leq \exp((-2n\epsilon_k^2)/\omega_S(\mathcal{C})^2)$. We now apply $\mathbb{P}(\check{A}) \leq \mathbb{P}(\tilde{A})$ and the union bound to obtain:

$$
\begin{aligned}
\mathbb{P}(\check{A}) &\leq \sum_{k=1}^{\infty} \exp\left(-\frac{2n\epsilon_k^2}{\omega_S(\mathcal{C})^2}\right) \\
&= \sum_{k=1}^{\infty} \exp\left(-2n\left(\sqrt{\frac{\log(2/\delta)}{2n}} + \sqrt{\frac{\log(k)}{n}}\right)^2\right) \\
&< \sum_{k=1}^{\infty} \exp\left(-(\log(2/\delta) + 2\log(k))\right) \\
&= \frac{\delta}{2}\sum_{k=1}^{\infty}\frac{1}{k^2} = \frac{\delta}{2}\cdot\frac{\pi^2}{6} < \delta \ .
\end{aligned}
$$

Thus, we have completed the proof. $\square$

C.2. Proof of Theorem 6

Proof.   Define $\boldsymbol{\sigma}_i = (\sigma_{i1}, \ldots, \sigma_{id})^T$ for $i = 1, \ldots, n$ whose components are i.i.d. Rademacher random variables. Then, we have

$$
\begin{aligned}
\mathfrak{R}^n(\mathcal{H}_\mathcal{B}) &= \mathbb{E}\left[\sup_{B\in\mathcal{B}}\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\sigma}_i^T B x_i\right] \\
&= \mathbb{E}\left[\sup_{B\in\mathcal{B}}\frac{1}{n}\sum_{i=1}^{n}\mathrm{Tr}\left(\boldsymbol{\sigma}_i^T B x_i\right)\right] \\
&= \mathbb{E}\left[\sup_{B\in\mathcal{B}}\frac{1}{n}\sum_{i=1}^{n}\mathrm{Tr}\left(B x_i \boldsymbol{\sigma}_i^T\right)\right] \\
&= \mathbb{E}\left[\sup_{B\in\mathcal{B}}\mathrm{Tr}\left(B\left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\sigma}_i x_i^T\right)^T\right)\right] \\
&= \mathbb{E}\left[\sup_{B\in\mathcal{B}}\langle B, \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\sigma}_i x_i^T\rangle\right].
\end{aligned}
$$

Note that the inner product between matrices $B, A \in \mathbb{R}^{d\times p}$ is defined as $\langle B, A\rangle := \sum_{i,j} B_{ij} A_{ij} = \mathrm{Tr}\left(BA^T\right)$. Now fix $x_1, \ldots, x_n$ and note that, by our assumption, we have, for all i,

$$
\sup_{\boldsymbol{\sigma}\in\{\pm 1\}^d}\|\boldsymbol{\sigma}x_i^T\| \leq \Omega.
$$

Let Θ be the random matrix $\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\sigma}_i x_i^T$. Choose arbitrary $\lambda > 0$. By Fenchel's inequality,

$$
\langle B, \Theta\rangle \leq \frac{F(B)}{\lambda} + \frac{F^*(\lambda\Theta)}{\lambda}.
$$

Since $F(B) \leq \beta^2$ for all $B \in \mathcal{B}$, we have

$$
\sup_{B\in\mathcal{B}}\langle B, \Theta\rangle \leq \frac{\beta^2}{\lambda} + \frac{F^*(\lambda\Theta)}{\lambda}.
$$

Taking expectations (w.r.t. σ_{ij}) gives

$$
\mathbb{E}[\sup_{B\in\mathcal{B}}\langle B, \Theta\rangle] \leq \frac{\beta^2}{\lambda} + \frac{\mathbb{E}[F^*(\lambda\Theta)]}{\lambda}
$$

Now let $Z_i = \frac{\lambda}{n}\boldsymbol{\sigma}_i x_i^T$ so that $S_n := \sum_{i=1}^n Z_i = \lambda\Theta$. Note that $\|Z_i\| \leq \frac{\lambda}{n}\Omega$. So the conditions of Lemma 4 in [Kakade et al.](2008) are satisfied with $V^2 = \lambda^2\Omega^2/n^2$. That lemma gives us $\mathbb{E}[F^*(\lambda\Theta)] \leq \lambda^2\Omega^2/(2\alpha n)$. Plugging this above, we have

$$\mathbb{E}[\sup_{B \in \mathcal{B}}\langle B, \Theta\rangle] \leq \frac{\beta^2}{\lambda} + \frac{\lambda\Omega^2}{2\alpha n}.$$

Setting $\lambda = (\beta/\Omega)\sqrt{2\alpha n}$ gives

$$\mathbb{E}[\sup_{B \in \mathcal{B}}\langle B, \Theta\rangle] \leq \Omega\beta\sqrt{\frac{2}{\alpha n}}$$

which completes the proof. $\quad\square$

D.   Proofs for Section 5

D.1.   Proof of Proposition 1

Proof. It is clear that the set on the right-hand side of (8) is a subset of $N_S(\bar{w})$. To see the reverse, let $c \in N_S(\bar{w})$ and $w \in S$ be given. Define $\hat{w}(\lambda) := \lambda w + (1-\lambda)\bar{w}$ and $r(\lambda) := \left(\frac{\bar{\mu}}{2}\right)\lambda(1-\lambda)\|w - \bar{w}\|^2$ for $\lambda \in [0,1]$. By the $\bar{\mu}$-strong convexity of S, we have that $B(\hat{w}(\lambda), r(\lambda)) \subseteq S$. Now notice that

$$c^T(\tilde{w} - \bar{w}) \leq 0 \ \ \text{for all } \tilde{w} \in B(\hat{w}(\lambda), r(\lambda)) \ ,$$

since $c \in N_S(\bar{w})$ and $B(\hat{w}(\lambda), r(\lambda)) \subseteq S$. Clearly the above condition is equivalent to:

$$c^T\bar{w} \ \geq \ \max_{\tilde{w} \in B(\hat{w}(\lambda), r(\lambda))} \left\{c^T\tilde{w}\right\} \ = \ c^T\hat{w}(\lambda) + r(\lambda)\|c\|_* \ ,$$

where the equality above follows from the definition of the dual norm $\|\cdot\|_*$. Rearranging the above and using $\hat{w}(\lambda) - \bar{w} = \lambda(w - \bar{w})$ as well as the definition of $r(\lambda)$ yields:

$$\lambda c^T(w - \bar{w}) \ \leq \ -\left(\frac{\bar{\mu}}{2}\right)\lambda(1-\lambda)\|c\|_*\|w - \bar{w}\|^2 \ \ \text{for all } \lambda \in [0,1] \ .$$

Now suppose that $\lambda > 0$. Dividing the above by λ yields:

$$c^T(w - \bar{w}) \ \leq \ -\left(\frac{\bar{\mu}}{2}\right)(1-\lambda)\|c\|_*\|w - \bar{w}\|^2 \ \ \text{for all } \lambda \in (0,1] \ .$$

Taking the limit as $\lambda \to 0$ yields

$$c^T(w - \bar{w}) \ \leq \ -\left(\frac{\bar{\mu}}{2}\right)\|c\|_*\|w - \bar{w}\|^2 \ ,$$

which completes the proof. $\quad\square$