

Risk Guarantees for End-to-End Prediction and Optimization Processes

Nam Ho-Nguyen

Discipline of Business Analytics, The University of Sydney. nam.ho-nguyen@sydney.edu.au

Fatma Kılınç-Karzan

Tepper School of Business, Carnegie Mellon University. fkilinc@andrew.cmu.edu

Prediction models are often employed in estimating parameters of optimization models. Despite the fact that in an end-to-end view, the real goal is to achieve good optimization performance, the prediction performance is measured on its own. While it is usually believed that good prediction performance in estimating the parameters will result in good subsequent optimization performance, formal theoretical guarantees on this are notably lacking. In this paper, we explore conditions that allow us to explicitly describe how the prediction performance governs the optimization performance. Our weaker condition allows for an asymptotic convergence result, while our stronger condition allows for exact quantification of the optimization performance in terms of the prediction performance. In general, verification of these conditions is a non-trivial task. Nevertheless, we show that our weaker condition is equivalent to the well-known Fisher consistency concept from the learning theory literature. This then allows us to easily check our weaker condition for several loss functions. We also establish that the squared error loss function satisfies our stronger condition. Consequently, we derive the exact theoretical relationship between prediction performance measured with the squared loss, as well as a class of symmetric loss functions, and the subsequent optimization performance. In a computational study on portfolio optimization, fractional knapsack and multiclass classification problems, we compare the optimization performance of using of several prediction loss functions (some that are Fisher consistent and some that are not) and demonstrate that lack of consistency of the loss function can indeed have a detrimental effect on performance.

Key words: stochastic optimization; prediction; end-to-end view

1. Introduction

The optimum solutions of optimization models crucially depend on the parameters defining these models, but these parameters are hardly ever available directly. In practice, these ‘true’ model parameters are predicted from side information and historical data often using statistical inference or machine learning techniques. There are many techniques that quantify the performance of prediction models. Nevertheless, these techniques almost exclusively focus on achieving a good prediction performance and do not take into account the subsequent optimization task. This is despite the fact that the ultimate goal in this process is to make the best decision in the subsequent optimization problem, not necessarily to have the best generic prediction performance of the parameters. In this paper, we consider a joint *end-to-end* view of the prediction and optimization

processes, and identify the critical properties of prediction models in terms of guaranteeing a low optimality gap in the subsequent optimization performance.

More formally, we consider an optimization problem of the form

$$\min_x \{f(x) + c^\top x : x \in X\}, \quad (1)$$

where $X \subset \mathbb{R}^m$ is a convex compact domain, and $f : X \rightarrow \mathbb{R}$ is a convex function (and hence continuous on the relative interior of X). In our setting, the linear vector c is not known exactly, but instead is governed via covariates w . More precisely, we suppose the covariates w belong to a given set $W \subseteq \mathbb{R}^k$, and the vectors c belong to a given set $C \subseteq \mathbb{R}^m$. We assume that $(w, c) \sim \mathbb{P}$ for some unknown distribution \mathbb{P} on $W \times C$, and we need to solve (1) for c yet we are only given information of w . Note that our setup covers the case when c is still noisy even when given w , since the conditional distribution $\mathbb{P}[c | w]$ may not be a point mass, and indeed this will be the more interesting case that we will study. Note also that previous literature studied the case when the function $f = 0$, but we consider a general convex f function, which is relevant in many applications; see Example 2.

While we do not know the distribution \mathbb{P} , we have access to the historical data $H_n := \{(w_i, c_i) : i \in [n]\}$, where the (w_i, c_i) are realizations of independent and identically distributed (i.i.d.) random variables from the unknown distribution \mathbb{P} . We examine an end-to-end view of the following prediction and optimization processes: first, based on data H_n , a prediction model in the form of a function $g : W \rightarrow \mathbb{R}^m$ is built to capture the dependency of c on w ; then, when given a covariate w , (1) is solved with c replaced by the prediction $g(w)$. This setting is commonly used amongst practitioners in decision-making domains for a variety of problems. Below, we give three particular examples, although many more exist.

EXAMPLE 1. Suppose we have a collection of service items (e.g., machines, vehicles) which we maintain over a certain time horizon. These items need refurbishment or replacement after a certain number of time periods. The optimal maintenance schedule can be defined as a shortest path problem over an appropriately defined network, where the ‘distances’ are given by the maintenance costs. Note that such future costs are often obtained via forecasts, and thus are not deterministic. In this setting, X is the convex hull of all paths from the starting point to the ending point in the underlying graph (each such path represents a maintenance plan), $f(x) = 0$ for all $x \in X$, and c is the vector of arc distances that represent the maintenance costs. Side information (covariates) w of c can consist of (amongst others) seasonality, demand, supply and other economic factors. ■

EXAMPLE 2. Consider a portfolio optimization problem, where the task is to allocate wealth to m different assets to maximize investment return. In the typical mean-variance formulation,

the goal is to simultaneously minimize the variance of the portfolio return, while maximizing the expected return. Then, X is the set of all possible asset allocations, each $x \in X$ represents an asset allocation (i.e., x_j represents how much wealth to invest into asset j), $f(x) = \gamma x^\top \Sigma x$ is the variance of the portfolio return with Σ being the covariance matrix of the returns between the assets, and $c = -\mu$ is the vector of (negative) returns for each assets. In many settings, Σ is assumed to be stable, and μ is predicted through market factors (e.g., liquidity, value, momentum, volume) which can be considered as the side information w . ■

EXAMPLE 3. Structured prediction is a form of multiclass classification designed to predict structured objects, such as sequences or graphs, from feature data; see e.g., [Goh and Jaillet \(2016\)](#), [Osokin et al. \(2017\)](#) and references therein. In structured prediction, given covariates w , a structured object \tilde{x} from some output space \tilde{X} is chosen as the prediction, often by solving $\min_{\tilde{x} \in \tilde{X}} \tilde{g}(\tilde{x}; w)$. In this setting, \tilde{X} is usually a finite combinatorial set, so $\tilde{g}(\tilde{x}; w)$ is a vector in which each coordinate corresponds to the cost of an object $\tilde{x} \in \tilde{X}$. The ‘true’ structured loss in this setting is measured between the selected \tilde{x} and the ‘correct’ $\tilde{x}^* \in \tilde{X}$, denoted by $L(\tilde{x}, \tilde{x}^*)$. Since \tilde{X} is combinatorial, L can be defined as the Hamming loss, although some other structured losses are possible. Structured prediction fits into our optimization setting by taking X to be a simplex whose vertices correspond to objects in \tilde{X} , $g(w) = \{\tilde{g}(\tilde{x}; w)\}_{\tilde{x} \in \tilde{X}}$, and the Hamming loss can be cast as the optimality gap of a particular constructed cost vector $c_{\tilde{x}^*}$. ■

Given a point $(w, c) \in W \times \mathbb{R}^m$ and a prediction function $g : W \rightarrow \mathbb{R}^m$, in order to assess the quality of using $g(w)$ in place of c in (1), we define the *true loss* as the optimality gap of the solution obtained with $g(w)$ on the true objective vector c , that is, the quality of the prediction $d = g(w)$ with respect to (1) is given by the *true loss function*

$$L(d, c) := f(x^*(d)) + c^\top x^*(d) - \min_{x \in X} \{f(x) + c^\top x\},$$

where $x^*(d) \in \arg \min_{x \in X} \{f(x) + d^\top x\}$. Since (w, c) is randomly drawn from \mathbb{P} , we assess the performance of a function $g : W \rightarrow \mathbb{R}^m$ in terms of the expected true loss, i.e., the *true risk*

$$R(g, \mathbb{P}) := \mathbb{E}[L(g(w), c)].$$

A naïve attempt to minimize $R(g, \mathbb{P})$ is to directly use empirical risk minimization (ERM) with the loss L to train the prediction model g , i.e., given the data $H_n = \{(w_i, c_i)\}_{i \in [n]}$, we obtain a prediction function \hat{g} by solving

$$\inf_g \frac{1}{n} \sum_{i \in [n]} L(g(w_i), c_i).$$

However, $L(d, c)$ is not convex in d , and thus it is not possible in general to obtain a polynomial-time approach with certified performance guarantees from minimizing the empirical risk based on

the true loss function L . The natural remedy is to use a *convex surrogate loss function* ℓ in place of L . The use of surrogate loss functions to ensure algorithmic tractability is very common in machine learning. For example, convex surrogates such as hinge loss are used instead of the non-convex true 0-1 loss in classification problems.

The question now becomes: which surrogate loss function should we use? While quite a number of surrogate loss functions have been proposed and used in this context, it is not yet well-understood how using a regression technique performs in terms of the true risk. More precisely, we define the *surrogate risk* as

$$R_\ell(g, \mathbb{P}) := \mathbb{E}[\ell(g(w), c)].$$

This paper aims to understand how a minimization scheme for the surrogate risk $R_\ell(g, \mathbb{P})$, which is well-established and implementable in practice, can impact the true risk $R(g, \mathbb{P})$. In other words, if we employ an established regression technique to obtain a prediction function \hat{g} , what can we say about the true risk $R(\hat{g}, \mathbb{P})$ of \hat{g} ? Consequently, to fill this gap in the literature, in this paper we explore this relationship and identify important properties of surrogate loss functions ℓ that enable us to derive guarantees on the true risk. We make these concepts mathematically rigorous, and describe their relationship to traditional notions of statistical consistency, in Section 3. For a visual summary of our framework see Figure 1.

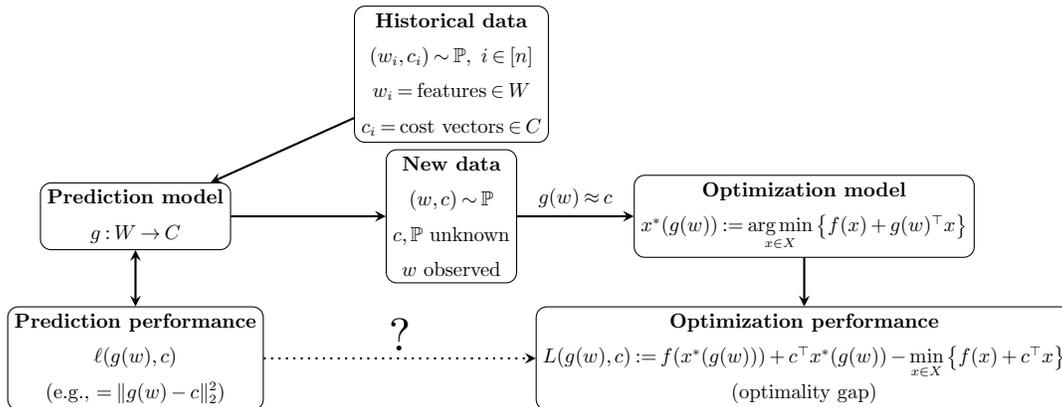


Figure 1 The end-to-end prediction and optimization framework.

1.1. Outline and Contributions

In this paper, we examine, from an end-to-end view, how the performance of the prediction part relates to the performance of the optimization part. In particular, we establish conditions for the existence of explicit relationships between the prediction performance, i.e., the surrogate risk, and the optimization performance, i.e., the true risk.

In Section 2, we review the literature related to this topic. In Section 3, we precisely define the problem we address, and outline the challenges.

In Section 4, we rigorously derive technical sufficient conditions on the prediction loss function that allow us to asymptotically minimize the true risk by minimizing the surrogate risk for a given probability distribution \mathbb{P} . These conditions are based solely on the choice of prediction loss function, rather than the class of prediction models g that we wish to select from. They allow us to compare and contrast the resulting optimization performance when different prediction model training methods are used for the estimation of objective function parameters, and thus are instrumental in terms of selecting among such training methods. In addition, our results in this section make the following contributions:

- We show that, in the prediction and optimization context, the concept of calibration introduced by Steinwart (2007), which allows us to establish performance guarantees, is equivalent to the well-known concept of Fisher consistency. To the best of our knowledge, such a relationship was not described in the previous literature. This result provides a tool for easily checking which loss functions lead to performance guarantees, which we exploit in our examples in Section 4.
- We compare several prediction methods from practice through the lens of our conditions. In Example 4, we show that the commonly used squared loss function $\ell(d, c) = \|d - c\|_2^2$ satisfies Fisher consistency. In Examples 6–9, we examine the SPO+ loss function from Elmachtoub and Grigas (2017), which is particularly relevant since it is the only convex loss function (thus far) that incorporates optimization problem information in the prediction and optimization setting. The Fisher consistency of the SPO+ loss function in certain settings was previously established in Elmachtoub and Grigas (2017) (which we summarize in Examples 6–7). Despite this, our Examples 8–9 show that the SPO+ loss function is not Fisher consistent in other fairly natural settings such as multiclass classification. Furthermore, our numerical study in Section 6.3 highlights the importance of having Fisher consistency of a loss function over simply a property that the loss function is customized to the optimization problem.

Often in statistical learning, we are given minimal knowledge of the distribution \mathbb{P} . Therefore, the distribution dependent nature of our results from Section 4 is not so desirable. In Section 5, building on the results from Steinwart (2007), we establish conditions for distribution-independent relationships between the true risk and the surrogate risk. In this section, our main contributions are as follows:

- Since checking these conditions is difficult for loss functions in general, we first focus on a tractable special case of using the squared loss function to measure prediction performance (i.e., the least squares method to train a prediction model). For the first time in the literature of joint prediction and optimization setting, using our conditions, we prove an explicit relationship between

the surrogate squared risk and the true optimization risk. This then allows us to relate our true risk to a class of symmetric loss functions by exploiting existing results on regression from [Steinwart \(2007\)](#); see Section 5.3.

- We also study distribution-independent risk relationships for the SPO+ loss function proposed in [Elmachtoub and Grigas \(2017\)](#) in Section 5.4. The SPO+ loss has not been formally studied in this context before, and due to the importance of the SPO+ loss to the prediction and optimization setting, we believe that such a study is warranted.

In Section 6, we carry out a computational study on three problem classes on real and simulated data: portfolio optimization, fractional knapsack, and multiclass classification. Our study on portfolio optimization is based on real-world data, where consistency is not known a priori. Our study with the fractional knapsack problem on simulated data allows us to choose some parameters to control the degree of non-linearity of the underlying data model, and thereby the model misspecification of certain loss functions. Lastly, we examine multiclass classification on simulated data, where the SPO+ loss is provably Fisher inconsistent (see Example 9), but the squared loss is consistent.

Our numerical findings support our theoretical results by indicating that the conditions we identified for the loss function have a non-trivial effect on practical performance. On the real world instances of portfolio optimization, we observe that there is little difference between using squared loss and SPO+ loss, where it is likely that both of these loss functions are consistent and there is no model misspecification. Indeed, we show in Section 6.1 that for a specific convex quadratic program with uncertain linear term and a single linear constraint and no non-negativity constraints arising in the mean-variance portfolio optimization, the SPO+ loss is equivalent to the true loss L . Furthermore, for this problem, we show that for the class of linear predictors, the optimal least squares risk predictor is also optimal for the true risk. Because of this, we carry out our experiments on more interesting case of the portfolio optimization instances with nonnegativity constraints. In contrast, our experiments on multiclass classification highlight an important insight: consistency of a loss function matters as much as (if not more) whether the loss function takes into account optimization problem information. In particular, despite the fact that the SPO+ loss takes into account information from the optimization problem, its inconsistency for multiclass classification problem resulted in poor performance. On the fractional knapsack instances, we re-affirm the observation of [Elmachtoub and Grigas \(2017\)](#) that the degree of model misspecification plays a role favoring SPO+ loss over squared loss when there is no significant difference between consistency and calibration properties of the loss functions.

We relegate all of the proofs to the appendices of the corresponding sections.

Notation. We use of the following notation. Given a positive integer N , $[N] := \{1, \dots, N\}$. Throughout, $k, m \in \mathbb{N}$ are the dimensions of the Euclidean spaces where W, C live respectively, $j \in [m]$ always denotes an index for the component of a vector in \mathbb{R}^m , and $i \in [n]$ denotes an index for a data point $(w_i, c_i) \in H_n$. Given a vector $d \in \mathbb{R}^m$, we let $X^*(d) := \arg \min_{x \in X} \{f(x) + d^\top x\}$ to be the argmin mapping, and $x^*(d)$ denotes some selection from $X^*(d)$ selected in a deterministic manner. More precisely, $x^* : \mathbb{R}^m \rightarrow X$ is a function such that for any $d \in \mathbb{R}^m$, $x^*(d) \in X^*(d)$. Our results are agnostic to the specific choice of algorithm picking $x^*(d) \in X^*(d)$.

2. Related Literature

Both prediction and optimization have been studied extensively on their own. In particular, the selection of the prediction function $g : W \rightarrow \mathbb{R}^m$ to minimize some measure of prediction error on the given data H_n is studied extensively in statistics and machine learning, see e.g., [Bousquet et al. \(2004\)](#). Moreover, a classical machine learning application, that is the classification problem, where a prediction model is built first from training data based on a loss function, presents a setup close to our end-to-end joint prediction and optimization view. The benchmark loss function in the context of the classification problem is the 0-1 loss, but it is nonconvex. Thus, in order to get polynomial-time algorithms for training, 0-1 loss is often replaced with a convex surrogate loss function. Consequently, this necessitates the study of the relationship between the surrogate loss functions and the true 0-1 loss within this context. This is a topic well-studied and understood; for example, [Steinwart \(2002a,b\)](#), [Lin \(2004\)](#), [Zhang \(2004\)](#), [Steinwart \(2005\)](#), [Bartlett et al. \(2006\)](#) have developed a general theory for the minimization of the true 0-1 risk via a surrogate risk which satisfies certain criteria. This was extended to robust regression and density estimation problems by [Steinwart \(2007\)](#), who builds a theory for the relationship between true and surrogate risk. Our work can be seen as a generalization of these results to optimization problems involving prediction parameters. In this context, our optimality gap is analogous to the 0-1 loss in classification, but is much more complicated.

From an end-to-end point of view, the relationship between the prediction models used to obtain model parameters and the subsequent optimization performance has, to our knowledge, only been examined by a few papers. This line of work was initiated by [Bengio \(1997\)](#) who explored the use of a financial training criterion in neural networks rather than a prediction criterion. In the context of newsvendor inventory control problem, [Liyanage and Shanthikumar \(2005\)](#) show that, rather than analyzing the optimal order quantity derived for the distribution that is estimated from the data, it is better to propose a broader class of order policies and choose the optimal policy that maximizes the expected profit on the data. More recently, [Kao et al. \(2009\)](#), [Elmachtoub and Grigas \(2017\)](#) and [Donti et al. \(2017\)](#) contributed to this line of research. These papers examined

designing or using alternative loss functions in training the prediction model so as to improve the final optimization performance. [Kao et al. \(2009\)](#) study the specialized setting where $X = \mathbb{R}^m$, f is a strongly convex quadratic, and the prediction model g is restricted to be linear, and present theoretical guarantees under a particular data distribution. [Donti et al. \(2017\)](#) propose a scheme to directly differentiate the optimality gap, which gives rise to a stochastic gradient descent scheme for directly training the prediction model via the optimality gap. While superior numerical performance of this algorithmic scheme was demonstrated in [Donti et al. \(2017\)](#), they provide no theoretical guarantees for the convergence of the risk quantities in their approach. In a setting closest to ours, [Elmachtoub and Grigas \(2017\)](#) examine the true optimality gap loss, and propose a convex surrogate loss from a quantity upper bounding the true optimality gap, and suggest that this convex surrogate loss function, referred to as the SPO+ loss, should be used in prediction model training. They show Fisher consistency (see Definition 4) of their surrogate loss under certain distributional assumptions, but do not give explicit relationships on how the performance of the prediction part governs the optimization performance. In contrast to their work on designing a new surrogate loss function, the main goal of our paper is essentially to close this theoretical gap in the literature by identifying *properties* of loss functions that ensure good performance and providing explicit relationships between the performance of the prediction loss and the optimization loss for *general* classes of loss functions satisfying these properties; see Sections 4 and 5. As such the focus and the results presented in our paper are very different than the ones from [Elmachtoub and Grigas \(2017\)](#). Note that in certain parts of our paper, we use specific loss functions, such as the squared loss or SPO+ loss to demonstrate that they possess or lack certain properties that we have identified. For this purpose, the squared loss is rather classical, and the main purpose in designing the SPO+ loss in [Elmachtoub and Grigas \(2017\)](#) was to keep the end-to-end framework in view, and therefore it is a very natural candidate to examine.

As we discussed in Example 3, this paper is also related to structured prediction. [Osokin et al. \(2017\)](#) provides risk relationships between surrogate methods to predict the vector $\{\tilde{g}(\tilde{x}; w)\}_{\tilde{x} \in \tilde{X}}$ (see Example 3) and the true structured loss. Our goal in this paper is to provide results for the more general optimization setting, where there are a potentially infinite number of ‘objects’, and when the true loss is the optimality gap.

As an alternative approach to this end-to-end view of the predict-then-optimize framework, one may wish to avoid appealing to an explicit prediction model completely, and instead use density estimation as a compelling method to incorporate the covariates w . Specifically, given w and historical data H_n , a density estimation of the conditional distribution $\mathbb{P}[c | w]$ can be built using a kernel: $\mathbb{P}[\cdot | w] \approx \sum_{i \in [n]} k_{w_i}(w) \delta_{c_i}(\cdot)$ where $k_{w_i}(w)$ are convex combination weights which increase as the covariates w become closer to w_i (often obtained via a kernel), and $\delta_{c_i}(\cdot)$ is point mass at c_i .

Then, a stochastic optimization problem with the estimated conditional distribution can be solved. This approach was studied by [Hannah et al. \(2010\)](#), [Hanasusanto and Kuhn \(2013\)](#), [Bertsimas and Kallus \(2014\)](#), [Ban and Rudin \(2019\)](#), [Bertsimas and Van Parys \(2017\)](#), [Ho and Hanasusanto \(2019\)](#) who all gave various performance guarantees. However, density estimation-based methods are known to require much more data than parametric prediction-based methods. As a result, when a reasonable parametric prediction model is available, it is advantageous to exploit it. Hence, density estimation methods are not the focus of this paper.

3. Risk Minimization and Consistency for Prediction and Optimization

Given an vector d , recall from Section 1 that we assess the quality of using d in place of a true cost vector c in (1) via the optimality gap of the solution obtained with d on the true objective vector c , which we define to be the *true loss function*

$$L(d, c) := f(x^*(d)) + c^\top x^*(d) - \min_{x \in X} \{f(x) + c^\top x\}, \quad (2)$$

where $x^*(d) \in \arg \min_{x \in X} \{f(x) + d^\top x\}$ is as described in Notation subsection. Note that given any $c \in \mathbb{R}^m$, $L(d, c) \geq 0$ for all $d \in \mathbb{R}^m$, and $L(c, c) = 0$.

REMARK 1. By definition the true loss function L depends on the function x^* , i.e., the algorithm that we use to solve $\min_{x \in X} \{f(x) + d^\top x\}$ for different $d \in \mathbb{R}^m$. We will take x^* to be fixed throughout the paper. Note, however, that the specific choice of x^* only affects our results up to measurability concerns; we show in Lemma EC.4 that any x^* is Lebesgue measurable, so we can safely fix x^* without changing the results as long as our distribution \mathbb{P} is Lebesgue measurable. In practice, any distribution we encounter will be Lebesgue measurable; we explicitly impose this in Assumption EC.1. Henceforth, when measurability of functions is discussed, we will understand this to be in the sense of Lebesgue. ■

Our setting of interest is prediction in the context of solving the optimization problem (1). Specifically, instead of a solitary (random) cost vector c , we are interested in random pairs $(w, c) \in W \times \mathbb{R}^m$ drawn from a distribution \mathbb{P} . We are then interested in learning a prediction function $g: W \rightarrow \mathbb{R}^m$ which predicts c with $g(w)$. Since $(w, c) \sim \mathbb{P}$ is random, we assess the performance of the prediction function g in terms of the expected true loss, which we call the *true risk*

$$R(g, \mathbb{P}) := \mathbb{E}[L(g(w), c)]. \quad (3)$$

The best possible true risk we can achieve is

$$R(\mathbb{P}) := \inf_g \{R(g, \mathbb{P}) : g \text{ measurable}\}. \quad (4)$$

A naïve attempt to solve (4) is to directly minimize $R(g, \mathbb{P})$. However, as we show in Lemma 1 below, $L(d, c)$ is not convex in d . Thus, in general it is not expected to obtain a polynomial-time approach to minimize the true risk $R(g, \mathbb{P})$.

LEMMA 1. *Suppose that $f(x) = 0$, X is such that it has at least two extreme points and $c \neq 0$ is such that $\min_{x \in X} c^\top x \neq \max_{x \in X} c^\top x$. Then, the loss function $L(d, c)$ is not convex in d .*

Given Lemma 1, in order to examine practical polynomial-time solution methodologies, we now describe an alternative approach based on *surrogate loss functions*. We first state a basic fact about (4), which is analogous to (Elmachtoub and Grigas 2017, Proposition 5) adapted to our setup.

LEMMA 2. *The function $g^*(w) := \mathbb{E}[c | w]$ minimizes (4). Furthermore,*

$$R(\mathbb{P}) = \mathbb{E} \left[\min_{d' \in \mathbb{R}^m} \mathbb{E}[L(d', c) | w] \right].$$

Lemma 2 shows that the conditional expectation is a minimizer of (4). There are several regression techniques which aim to recover the conditional expectation. These have a similar structure, which we now describe. First, we specify some loss function $\ell(d, c)$ measuring the discrepancy between vectors d and c . As before, we are in the setting where we have random pairs $(w, c) \sim \mathbb{P}$ and we wish to do prediction via a function $g(w) \approx c$. We then define the *surrogate risk* as:

$$R_\ell(g, \mathbb{P}) := \mathbb{E}[\ell(g(w), c)] \tag{5}$$

as well as the best possible surrogate risk

$$R_\ell(\mathbb{P}) := \inf_g \{R_\ell(g, \mathbb{P}) : g \text{ measurable}\}. \tag{6}$$

Now, instead of seeking to minimize the true risk $R(g, \mathbb{P})$, we seek to minimize the surrogate risk $R_\ell(g, \mathbb{P})$. Indeed, regression methods often use tractable convex losses ℓ , hence minimizing the surrogate risk is much more tractable than the true risk. Here, we use the term ‘surrogate’ since, in a sense, the loss function ℓ can be thought of as a *surrogate loss* for L , i.e., in order to maintain computational tractability, we replace the difficult loss L with a more computationally friendly surrogate ℓ . The use of surrogate loss functions to ensure algorithmic tractability is very common in machine learning. For example, convex surrogates such as hinge loss are used instead of the non-convex true 0-1 loss in classification problems.

It is not yet well-understood how minimizing the surrogate risk (5) can impact the true (3). A good surrogate loss function ℓ should mimic the natural properties of the true loss function L , i.e., $\ell(c, c) = 0$, $\ell(d, c) \geq 0$ for any d, c . However, the most important feature of a surrogate loss function

is how its risk bound relates to the true risk (3). More precisely, if one were to obtain a prediction function \hat{g} with low excess surrogate risk $R_\ell(\hat{g}, \mathbb{P}) - R_\ell(\mathbb{P})$, will it be the case that \hat{g} also has low excess true risk $R(\hat{g}, \mathbb{P}) - R(\mathbb{P})$?

Consequently, in this paper we will explore this relationship and identify important properties of surrogate loss functions that enable us to derive guarantees on the true risk. We would like to identify essential properties of surrogate loss functions $\ell(g(w), c)$ such that they can accurately, in some sense, assess the quality of using $g(w)$ in place of c for the true risk (3) related to (1), while remaining computationally tractable to optimize (e.g., being convex in $g(w)$).

While the concepts we explore are related to the more traditional notion of *statistical consistency*, they are of a slightly different nature, which we elaborate on now. In practice, the distribution \mathbb{P} is not given explicitly, but instead we only have access to historical data $H_n = \{(w_i, c_i) : i \in [n]\}$. To obtain a predictor $g : W \rightarrow \mathbb{R}^m$, we optimize the *empirical* surrogate risk

$$\hat{R}_\ell(g, H_n) := \frac{1}{n} \sum_{i=1}^n \ell(g(w_i), c_i).$$

Statistical learning theory has rich literature on relating \hat{R}_ℓ to R_ℓ ; see, e.g., [Bousquet et al. \(2004\)](#). In particular, it has several results on the following notion of consistency.

DEFINITION 1. Given a (deterministically expanding) sequence of classes of predictors $\{\mathcal{G}_n\}_{n \in \mathbb{N}}$, let $\hat{g}_n := \arg \min_{g \in \mathcal{G}_n} \hat{R}_\ell(g, H_n)$. We say that the (random) sequence of predictors $\{\hat{g}_n\}_{n \in \mathbb{N}}$ is *statistically consistent with respect to loss ℓ* if

$$R_\ell(\hat{g}_n, \mathbb{P}) \rightarrow R_\ell(\mathbb{P}) \text{ in probability.}$$

(Convergence in probability is used due to the randomness in H_n , which translates to randomness of \hat{g}_n .) This states that, for large n , we can get high-probability bounds on the excess surrogate risk $R_\ell(\hat{g}_n, \mathbb{P}) - R_\ell(\mathbb{P})$ of a predictor \hat{g}_n . Whenever $\{\mathcal{G}_n\}_{n \in \mathbb{N}}$ and ℓ are mildly regular, the consistency of the predictors \hat{g}_n holds in a wide variety of settings.

However, since we will use \hat{g}_n for optimization, we are actually interested in the excess true risk $R(\hat{g}_n, \mathbb{P}) - R(\mathbb{P})$ which depends on (1) explicitly. Thus, in this paper, we give relationships between the excess surrogate risk and the true risk. More precisely, we will explore conditions on the surrogate loss function ℓ that ensure the following property holds:

DEFINITION 2. Given a class of distributions \mathcal{P} , we say that ℓ is (\mathcal{P}, L) -consistent if, for all $\mathbb{P} \in \mathcal{P}$, whenever we have a sequence of predictors $\{g_n\}_{n \in \mathbb{N}}$ such that $R_\ell(g_n, \mathbb{P}) \rightarrow R_\ell(\mathbb{P})$, we will also imply $R(g_n, \mathbb{P}) \rightarrow R(\mathbb{P})$.

If ℓ satisfies Definition 2 then this means that *any* sequence that is statistically consistent (in the sense of Definition 1) with respect to the surrogate loss ℓ is also statistically consistent with respect to the true loss L .

REMARK 2. Note that Definition 2 does not depend on the classes of predictors $\{\mathcal{G}_n\}_{n \in \mathbb{N}}$ or the sequence of predictors $\{\hat{g}_n\}_{n \in \mathbb{N}}$ obtained by minimizing $\hat{R}_\ell(g, H_n)$, even though these are important to relate the empirical surrogate risk \hat{R}_ℓ to the population surrogate risk R_ℓ , as well as for computational considerations of optimizing the surrogate risk. Because of this, our results are also naturally independent of the choice of $\{\mathcal{G}_n\}_{n \in \mathbb{N}}$. This is important for the application of our theory: by keeping the $\{\mathcal{G}_n\}_{n \in \mathbb{N}}$ unspecified, our results are applicable to all settings. ■

4. Risk Minimization via Fisher Consistent Surrogate Loss Functions

As discussed in Section 3, we are interested in properties of the surrogate loss ℓ which ensures consistency in the sense of Definition 2 holds. In order to understand the kind of results that we are after, let us explore the negation of this. In this case, we have $R_\ell(g_n, \mathbb{P}) - R_\ell(\mathbb{P}) \rightarrow 0$ but, for some $\epsilon > 0$, $R(g_n, \mathbb{P}) - R(\mathbb{P}) > \epsilon$ for infinitely many n . In other words, there exists $\epsilon > 0$ such that for all $\delta > 0$, there exists g_n such that $R_\ell(g_n, \mathbb{P}) - R_\ell(\mathbb{P}) \leq \delta$ but $R(g_n, \mathbb{P}) - R(\mathbb{P}) > \epsilon$. To prevent this bad outcome, we want to guarantee the following relationship between the risks:

for all $\epsilon > 0$, there exists $\delta > 0$ such that: (7)

if $g : W \rightarrow \mathbb{R}^m$ satisfies $R_\ell(g, \mathbb{P}) - R_\ell(\mathbb{P}) \leq \delta$, then $R(g, \mathbb{P}) - R(\mathbb{P}) \leq \epsilon$.

We will show that (7) can be guaranteed by checking a simpler condition on the losses ℓ and L called *calibration*. This was introduced by Bartlett et al. (2006) for binary classification and extended by Steinwart (2007) for other machine learning applications. We extend this concept to the context of prediction and optimization.

DEFINITION 3. A surrogate loss function ℓ for L is *calibrated* with respect to a distribution \mathbb{P} , or *\mathbb{P} -calibrated*, if, for all $w \in W$ and $\epsilon > 0$, there exists $\delta > 0$ (which may depend on w) such that if $d \in \mathbb{R}^m$ satisfies $\mathbb{E}[\ell(d, c) | w] - \min_{d' \in \mathbb{R}^m} \mathbb{E}[\ell(d', c) | w] < \delta$, then $\mathbb{E}[L(d, c) | w] - \min_{d' \in \mathbb{R}^m} \mathbb{E}[L(d', c) | w] < \epsilon$.

Observe that Definition 3 is very similar to (7), except that predictors g (i.e., functions mapping onto vectors) are replaced with vectors $d \in \mathbb{R}^m$, and that conditional expectations given w are taken. This makes Definition 3 verifiable, i.e., given a class of probability distributions \mathbb{P} and a surrogate loss ℓ , we can check whether Definition 3 holds or not. Of course, we then need to check that Definition 3 is sufficient to obtain risk bounds. Steinwart (2007, Theorem 2.8) provides a result to obtain such risk bounds, and we apply it to obtain Theorem 1 below. More precisely, we verify that necessary measurability and boundedness conditions on certain conditional risk quantities are met in order to apply the proof technique of Steinwart (2007) in our prediction and optimization context. We use the following technical assumption:

ASSUMPTION 1. Let the probability distribution \mathbb{P} and the surrogate loss function ℓ be given. For any fixed $c \in C$, the surrogate loss function $\ell(d, c)$ is convex in $d \in \mathbb{R}^m$. For any $w \in W$ and $d \in \mathbb{R}^m$, the set $\arg \min_{d' \in \mathbb{R}^m} \mathbb{E}[\ell(d', c) | w]$ is non-empty and bounded, and $\mathbb{E}[\ell(d, c) | w] < \infty$. Furthermore, c is an integrable random vector (that is, each component is integrable) so that $\mathbb{E}[\|c\|_1] < \infty$.

THEOREM 1. Suppose that ℓ is \mathbb{P} -calibrated, and that Assumption 1 holds. Then for all $\epsilon > 0$, there exists a $\delta > 0$ such that

$$R_\ell(g, \mathbb{P}) \leq R_\ell(\mathbb{P}) + \delta \implies R(g, \mathbb{P}) \leq R(\mathbb{P}) + \epsilon.$$

We give the proof in Section EC.3.

In general, checking that a given surrogate loss ℓ is \mathbb{P} -calibrated may not be straightforward. A much simpler condition to check is Fisher consistency, stated in Definition 4 below. Note that Fisher consistency relates to the *minimizers* of the loss functions, instead of approximate minimizers as in Definition 3. In this section, we show that Fisher consistency is equivalent to calibration, thus allowing us to check the simpler condition to verify Theorem 1. We also discuss some different loss functions and their Fisher consistency properties.

DEFINITION 4. A surrogate loss function ℓ is *Fisher consistent* with respect to a distribution \mathbb{P} , or *\mathbb{P} -Fisher consistent*, if for all w ,

$$\arg \min_d \mathbb{E}[\ell(d, c) | w] \subseteq \arg \min_d \mathbb{E}[L(d, c) | w].$$

REMARK 3. Since the objective for our optimization problem is of the form $f(x) + c^\top x$, we proved in Lemma 2 that $\mathbb{E}[c | w] \in \arg \min_d \mathbb{E}[L(d, c) | w]$. Thus, one way to check that ℓ is Fisher consistent is to verify that $\arg \min_d \mathbb{E}[\ell(d, c) | w] = \{\mathbb{E}[c | w]\}$ (and this is the approach taken in some of the examples below). However, we opt not to simply take $\arg \min_d \mathbb{E}[\ell(d, c) | w] = \{\mathbb{E}[c | w]\}$ as the definition of Fisher consistency because we recognize, particularly for non-smooth optimization objectives, that there can be other vectors besides $\mathbb{E}[c | w]$ that minimize $\mathbb{E}[L(d, c) | w]$. Furthermore, the current form of Definition 4 will also allow us to encompass settings when the objective is of a more general form than $f(x) + c^\top x$ (although this is not the focus of the current paper). ■

In the following theorem we show that Fisher consistency is equivalent to calibration. Of course, the fact that calibration implies Fisher consistency is straightforward; the main challenge is to show the other direction.

THEOREM 2. Given a distribution \mathbb{P} , let $\ell(d, c)$ be a loss function that satisfies Assumption 1. Then ℓ is \mathbb{P} -calibrated if and only if ℓ is \mathbb{P} -Fisher consistent.

The key tool that we exploit in proving Theorem 2 is upper semi-continuity of the multivalued argmin mapping $X^*(\cdot)$ (see EC.1). Informally, this states that if we are given $X^*(d)$ for some vector d , and we are interested in vectors d' for which $X^*(d')$ does not move ‘too far away’ from $X^*(d)$, then we can guarantee that when d' is sufficiently close to d , this will indeed be the case. In particular, in the context of proving Theorem 2, we use this to show that when $\mathbb{E}[L(d, c) | w]$ is large, then vectors close by to d will also have large true expected loss. The full proof of Theorem 2 is in EC.4.

Armed with Theorem 2, we have the following corollaries, which are straightforward consequences of our results discussed so far.

COROLLARY 1. *Suppose that ℓ is \mathbb{P} -Fisher consistent, and that Assumption 1 holds. Then for all $\epsilon > 0$, there exists a $\delta > 0$ such that*

$$R_\ell(g, \mathbb{P}) \leq R_\ell(\mathbb{P}) + \delta \implies R(g, \mathbb{P}) \leq R(\mathbb{P}) + \epsilon.$$

COROLLARY 2. *Suppose that ℓ is \mathbb{P} -Fisher consistent, and that Assumption 1 holds. If we have a sequence of functions g_n such that $R_\ell(g_n, \mathbb{P}) \rightarrow R_\ell(\mathbb{P})$. Then $R(g_n, \mathbb{P}) \rightarrow R(\mathbb{P})$.*

We now examine several different loss functions and their Fisher consistency properties. Before doing so, let us summarize the properties on ℓ and \mathbb{P} in order to get risk guarantees of the form (7) through Theorem 2. These are:

1. the surrogate loss $\ell(\cdot, c)$ is convex for any fixed $c \in C$.
2. for any $w \in W, d \in \mathbb{R}^m$, the expected loss $\mathbb{E}[\ell(d, c) | w]$ is finite.
3. for any $w \in W$, the set of minimizers $\arg \min_{d' \in \mathbb{R}^m} \mathbb{E}[\ell(d', c) | w]$ is non-empty and bounded.
4. the surrogate loss ℓ is \mathbb{P} -Fisher consistent according to Definition 4.

We first examine the squared loss function, namely $\ell_{\text{LS}}(d, c) = \|d - c\|_2^2$, that is Fisher consistent for any class of distributions. (We use the ‘LS’ subscript as shorthand for ‘least squares’.)

EXAMPLE 4. Consider the squared loss $\ell_{\text{LS}}(d, c) = \|d - c\|_2^2$. Then ℓ_{LS} is \mathbb{P} -Fisher consistent for any distribution \mathbb{P} over $W \times C$. Note that

$$\mathbb{E}[\ell_{\text{LS}}(d, c) | w] = \mathbb{E}[\|d - c\|_2^2 | w] = \|d - \mathbb{E}[c | w]\|_2^2 + \mathbb{E}[\|c\|_2^2 | w] - \|\mathbb{E}[c | w]\|_2^2.$$

Thus, the unique minimizer of $\mathbb{E}[\ell_{\text{LS}}(d, c) | w]$ is $d^* = \mathbb{E}[c | w]$. Since we know this is also a minimizer of $\mathbb{E}[L(d, c) | w]$, this gives us \mathbb{P} -Fisher consistency of the squared loss; verifying Property 4.

Also note that Properties 1 and 3 are clearly satisfied. Property 2 will be satisfied if the conditional distribution $\mathbb{P}[\cdot | w]$ is square integrable for every $w \in W$. ■

A common loss function used in regression to safeguard against outliers is the absolute deviation loss, namely $\ell_{\text{AD}}(d, c) := \|d - c\|_1$. We next examine this loss function.

EXAMPLE 5. Consider the absolute deviation loss $\ell_{\text{AD}}(d, c) = \|d - c\|_1$. We claim that ℓ_{AD} is \mathbb{P} -Fisher consistent as long as, for every w , $\mathbb{P}[\cdot | w]$ is centrally symmetric about some vector d_w . A distribution \mathbb{P} is centrally symmetric about d if, for a random variable $c \sim \mathbb{P}$, $c - d$ has the same distribution as $d - c$. Note that $\arg \min_{d' \in \mathbb{R}^m} \mathbb{E}[\|d' - c\|_1 | w]$ recovers the vector of coordinate-wise medians, which for a centrally symmetric distribution will be the point of symmetry d_w , i.e., d_w minimizes $\mathbb{E}[\|d - c\|_1 | w]$. Furthermore, we have $\mathbb{E}[c | w] = d_w$ also. Therefore, d_w minimizes $\mathbb{E}[L(d, c) | w]$. ■

We now discuss the SPO+ loss function proposed in [Elmachtoub and Grigas \(2017\)](#), which aims to incorporate knowledge of the domain X into the loss, in the hopes of achieving low true risk R , which is based on the optimization problem.

EXAMPLE 6. In the setting when $f(x) = 0$ for all $x \in X$, [Elmachtoub and Grigas \(2017, Definition 3\)](#) defined the following loss function:

$$\ell_{\text{SPO}+}(d, c) := (2d - c)^\top x^*(c) - \min_{x \in X} (2d - c)^\top x = L(c, 2d - c). \quad (8)$$

[Elmachtoub and Grigas \(2017, Proposition 6\)](#) shows that $\ell_{\text{SPO}+}$ is Fisher consistent as long as $\mathbb{P}[c | w]$ is centrally symmetric and continuous. We remark also that [Elmachtoub and Grigas \(2017\)](#) achieve good numerical results, particularly when the hypothesis class is misspecified versus the true distribution. ■

We now highlight some positive and negative aspects of the loss function of [Elmachtoub and Grigas \(2017\)](#). We start with an example below to review an important observation made in [Elmachtoub and Grigas \(2017, Proposition 1\)](#) that in the case of binary classification, by carefully choosing the set C and domain X , the true loss L from (2) becomes the 0-1 loss. In addition, their surrogate loss $\ell_{\text{SPO}+}$ (8) also has a familiar interpretation and admits Fisher consistency in this setting.

EXAMPLE 7. Let $m = 1$, $C = \{-1, 1\}$, $X = [-1/2, 1/2]$ and $f(x) = 0$ for all $x \in X$. Then, $x^*(d) = -\text{sign}(d)/2$, and $\min_{x \in X} c^\top x = -1/2$ for any $c \in C$, so

$$L(d, c) = \frac{c \text{sign}(d) - 1}{2} = \begin{cases} 0, & c = \text{sign}(d) \\ 1, & c \neq \text{sign}(d). \end{cases}$$

That is, the 0-1 loss for classification is exactly equivalent to the true loss function L . [Elmachtoub and Grigas \(2017, Proposition 4\)](#) shows that the loss from (8) reduces to the hinge loss in this case: since $x^*(c) = -c/2$ for $c \in C$ and $\min_{x \in X} d^\top x = -|d|/2$,

$$\ell_{\text{SPO}+}(d, c) = \frac{|2d - c| - (2d - c)c}{2} = \frac{|1 - 2dc| + 1 - 2dc}{2} = \max\{0, 1 - 2dc\}.$$

Moreover, [Lin \(2004, Theorem 3.1\)](#) states that the hinge loss, and thus $\ell_{\text{SPO}+}$, is Fisher consistent for any distribution over $C = \{-1, 1\}$ except the uniform one. ■

In contrast to this, we next demonstrate with the following two general examples that the loss function ℓ_{SPO_+} of [Elmachtoub and Grigas \(2017\)](#) is not Fisher consistent in some very natural settings.

EXAMPLE 8. Consider the setting where $m = 1$, $X = [-1/2, 1/2]$ and $f(x) = 0$ for all $x \in X$, but C is an arbitrary subset of \mathbb{R} . Then $x^*(c) = -\text{sign}(c)/2$, $\min_{x \in X} d^\top x = -|d|/2$, hence the loss function from (8) becomes

$$\ell_{\text{SPO}_+}(d, c) = \frac{|2d - c| - (2d - c) \text{sign}(c)}{2} = \frac{|2d - c| - 2d \text{sign}(c) + |c|}{2}.$$

Let \mathbb{P} be a distribution over $W \times C$. For any $w \in W$, note that the minimizers of $\mathbb{E}[L(d, c) | w]$ are $D_w^* = \{d \in \mathbb{R} : \text{sign}(d) = \text{sign}(\mathbb{E}[c | w])\}$. Thus, checking \mathbb{P} -Fisher consistency requires showing that $\arg \min_{d' \in \mathbb{R}} \mathbb{E}[\ell_{\text{SPO}_+}(d', c) | w] \subseteq D_w^*$ for every $w \in W$, i.e., we need to show that the minimizers have the same sign as the mean $\mathbb{E}[c | w]$. However, we show (in [EC.4](#)) that the minimizer of the loss function $\ell_{\text{SPO}_+}(d, c)$ has the same sign as the median. Therefore, for distributions where the mean and median have different signs, this loss function is not Fisher consistent. ■

EXAMPLE 9. In [Example 7](#), we examined binary classification and showed that for appropriately chosen X , f and C , L specializes to the 0-1 loss and ℓ_{SPO_+} specializes to the hinge loss. Thus, ℓ_{SPO_+} defined in (8) can be seen as a generalization of the hinge loss for optimization problems. We next show that the multiclass classification loss admits a similar representation, i.e., by choosing X and C appropriately we can make L represent the 0-1 loss for multiclass classification. However, we also establish that the generalization of hinge loss given by (8) to this setting is not Fisher consistent.

Suppose we have pairs (w, c) , where w are features, and $c \in C'$ is a label from one of $m \in \mathbb{N}$ different classes, i.e., $C' = [m]$. We want a predictor $g' : W \rightarrow C'$ which classifies w according to $g'(w)$. If we classify w incorrectly (i.e., $g'(w)$ is in a different class to c) we suffer a loss of 1; otherwise, our loss is 0. We can capture this in our optimization framework as follows.

Consider $C = \{c_j := \mathbf{1}_m - e_j : j \in [m]\} \subset \mathbb{R}^m$, $X = \text{Conv}\{e_j : j \in [m]\} \subset \mathbb{R}^m$ and $f(x) = 0$ for all $x \in X$. Then $\min_{x \in X} d^\top x = \min_{j' \in [m]} d_{j'}$, $\min_{x \in X} c_j^\top x = 0$ and $x^*(d) = e_j$ for $j \in \arg \min_{j' \in [m]} d_{j'}$, so for any $j \in [m]$ and vector d with unique minimum entry

$$L(d, c_j) = \begin{cases} 0, & \arg \min_{j' \in [m]} d_{j'} = j \\ 1, & \arg \min_{j' \in [m]} d_{j'} \neq j. \end{cases}$$

In other words, if we have a function $g : W \rightarrow \mathbb{R}^m$, we can use it to build a classifier $g' : W \rightarrow C'$ by classifying w according to the minimum entry of $g(w) \in \mathbb{R}^m$. Then L is exactly the 0-1 loss for this classifier. Suppose that we have a distribution $\mathbb{P}[c = c_j] = p_j > 0$, $\sum_{j \in [m]} p_j = 1$. Then, letting $j^*(d) = \arg \min_{j' \in [m]} d_{j'}$,

$$\mathbb{E}[L(d, c)] = 1 - p_{j^*(d)},$$

so the vectors d which minimize $\mathbb{E}[L(d, c)]$ must satisfy $j^*(d) \in \arg \max_{j' \in [m]} p_{j'}$.

The loss (8) becomes

$$\ell_{\text{SPO}^+}(d, c_j) = (2d - c_j)^\top e_j - \min_{j' \in [m]} \{2d_{j'} - c_{j'}\} = 2d_j - \min_{j' \in [m]} \{2d_{j'} - \mathbf{1}(j' \neq j)\}.$$

In EC.4, we show that for distributions \mathbb{P} with $\max_{j' \in [m]} p_{j'} < 1/2$, ℓ is not \mathbb{P} -Fisher consistent, since the set of minimizers of $\mathbb{E}[\ell_{\text{SPO}^+}(d, c)]$ are the vectors $d_\alpha = \alpha \mathbf{1}_m$, $\alpha \in \mathbb{R}$, which cannot in general pick out the maximum probability class $j \in [m]$, i.e., the highest p_j . ■

5. Non-Asymptotic Risk Guarantees via Uniform Calibration

Corollary 2 is an asymptotic result, that is, it asserts only that minimizing the surrogate risk will minimize the true risk in the limit. This does not present much insight about the rate of convergence of these quantities, which is governed by the relationship between ϵ and δ in Corollary 1. Moreover, the δ in Corollary 1 depends on the distribution \mathbb{P} . In general, this is undesirable, since often in statistical learning, we assume minimal knowledge of \mathbb{P} . Furthermore, when given n data points $\{(w_i, c_i) : i \in [n]\}$ we can build a predictor g_n with quantified guarantees on the excess surrogate risk $R_\ell(g_n, \mathbb{P}) - R_\ell(\mathbb{P})$ via standard learning theoretic results. We would ideally like to translate these into quantified guarantees on the excess true risk $R(g_n, \mathbb{P}) - R(\mathbb{P})$.

Steinwart (2007) builds a theory for non-asymptotic relationships between true and surrogate risk for various types of learning problems, such as classification, regression, and density estimation, giving necessary and sufficient conditions for the existence of distribution-independent guarantees. In this section, building on the results from Steinwart (2007), we provide conditions for the existence of similar guarantees in the prediction and optimization context. Using these conditions, we identify a non-asymptotic distribution-independent guarantee between the risk of the surrogate squared loss function ℓ_{LS} and the true optimality gap risk. We then provide risk guarantees for a class of symmetric loss functions by appealing to existing results on their risk relationships to the squared loss ℓ_{LS} . Finally, we study the a special case of the ℓ_{SPO^+} loss function (8) of Elmachroub and Grigas (2017), and provide positive and negative results on its risk guarantees.

5.1. Outline of Key Idea

In order to provide guarantees on the true risk implied by the surrogate risk, in this section, our aim is to identify an increasing function $\eta : [0, \infty) \rightarrow [0, \infty)$ with $\eta(0) = 0$ such that for any distribution \mathbb{P} , we have

$$\eta(R(g, \mathbb{P}) - R(\mathbb{P})) \leq R_\ell(g, \mathbb{P}) - R_\ell(\mathbb{P}).$$

Thus, any bound on the excess surrogate risk $R_\ell(g, \mathbb{P}) - R_\ell(\mathbb{P})$ translates to a bound on the excess true risk $R(g, \mathbb{P}) - R(\mathbb{P})$. Let us explore how we would derive such bounds. First, suppose that η and ℓ are chosen so that η is convex and that for any $w \in W$ and $d \in \mathbb{R}^m$, we have

$$\eta \left(\mathbb{E}[L(d, c) | w] - \min_{d' \in \mathbb{R}^m} \mathbb{E}[L(d', c) | w] \right) \leq \mathbb{E}[\ell(d, c) | w] - \min_{d' \in \mathbb{R}^m} \mathbb{E}[\ell(d', c) | w]. \quad (9)$$

Then, we have

$$\begin{aligned}
\eta(R(g, \mathbb{P}) - R(\mathbb{P})) &= \eta \left(\mathbb{E} \left[\mathbb{E}[L(g(w), c) | w] - \min_{d' \in \mathbb{R}^m} \mathbb{E}[L(d', c) | w] \right] \right) \\
&\leq \mathbb{E} \left[\eta \left(\mathbb{E}[L(g(w), c) | w] - \min_{d' \in \mathbb{R}^m} \mathbb{E}[L(d', c) | w] \right) \right] \\
&\leq \mathbb{E} \left[\mathbb{E}[\ell(g(w), c) | w] - \min_{d' \in \mathbb{R}^m} \mathbb{E}[\ell(d', c) | w] \right] \\
&= R_\ell(g, \mathbb{P}) - R_\ell(\mathbb{P}),
\end{aligned}$$

where the first inequality follows from Jensen's inequality, and the second follows from (9).

As a first attempt to choose such η and ℓ , we define

$$\delta_\ell(\epsilon, w; \mathbb{P}) := \inf_{d \in \mathbb{R}^m} \left\{ \mathbb{E}[\ell(d, c) | w] - \min_{d' \in \mathbb{R}^m} \mathbb{E}[\ell(d', c) | w] : \mathbb{E}[L(d, c) | w] - \min_{d' \in \mathbb{R}^m} \mathbb{E}[L(d', c) | w] > \epsilon \right\}. \quad (10)$$

REMARK 4. Note that $\delta_\ell(\epsilon, w; \mathbb{P})$ is simply giving an explicit representation of the δ that appears in Definition 3 as a function of ϵ and w . In particular, $\delta_\ell(\epsilon, w; \mathbb{P}) > 0$ for $\epsilon > 0$ whenever ℓ is \mathbb{P} -calibrated.

To see this, suppose that ℓ is \mathbb{P} -calibrated. Fix some $w \in W$ and $\epsilon > 0$. Then (by the contrapositive statement of the implication in Definition 3) there exists $\delta > 0$ such that whenever $\mathbb{E}[L(d, c) | w] - \min_{d' \in \mathbb{R}^m} \mathbb{E}[L(d', c) | w] > \epsilon$ we have $\mathbb{E}[\ell(d, c) | w] - \min_{d' \in \mathbb{R}^m} \mathbb{E}[\ell(d', c) | w] > \delta$. Taking the infimum of $\mathbb{E}[\ell(d, c) | w] - \min_{d' \in \mathbb{R}^m} \mathbb{E}[\ell(d', c) | w]$ over d such that $\mathbb{E}[L(d, c) | w] - \min_{d' \in \mathbb{R}^m} \mathbb{E}[L(d', c) | w] > \epsilon$ gives exactly $\delta_\ell(\epsilon, w; \mathbb{P})$ as defined in (10), and we know that $\mathbb{E}[\ell(d, c) | w] - \min_{d' \in \mathbb{R}^m} \mathbb{E}[\ell(d', c) | w] > \delta$ for such d , hence $\delta_\ell(\epsilon, w; \mathbb{P}) \geq \delta > 0$. ■

Fixing $w \in W$, consider $d \in \mathbb{R}^m$ such that $\mathbb{E}[L(d, c) | w] - \min_{d' \in \mathbb{R}^m} \mathbb{E}[L(d', c) | w] = \epsilon$. Then

$$\begin{aligned}
\mathbb{E}[\ell(d, c) | w] - \min_{d' \in \mathbb{R}^m} \mathbb{E}[\ell(d', c) | w] &\geq \delta_\ell(\epsilon, w; \mathbb{P}) \\
&= \delta_\ell \left(\mathbb{E}[L(d, c) | w] - \min_{d' \in \mathbb{R}^m} \mathbb{E}[L(d', c) | w], w; \mathbb{P} \right).
\end{aligned}$$

This relation then inspires us to select $\eta = \delta_\ell$. But, such a choice of $\eta = \delta_\ell$ may not be feasible as we cannot ensure that δ_ℓ is convex in general. Instead, we can use $\eta = \delta_\ell^{**}$, where, given a function $h: \mathbb{R} \rightarrow \mathbb{R} \cup \{\infty\}$,

$$h^{**}(\epsilon) = \sup_{h'} \{h'(\epsilon) : h' \text{ convex function on } \mathbb{R}, h' \leq h \text{ pointwise}\}.$$

Clearly, h^{**} is convex since it is a supremum of convex functions, and it can be obtained via convex conjugacy (however, we will not need to appeal to this representation for our results).

Note that δ_ℓ is only defined for $\epsilon > 0$, so we define $\delta_\ell(\epsilon, w; \mathbb{P}) = 0$ when $\epsilon = 0$ and $\delta_\ell(\epsilon, w; \mathbb{P}) = +\infty$ when $\epsilon < 0$. Using $\eta = \delta_\ell^{**}$ guarantees both convexity of η and also that $\eta(\epsilon, w; \mathbb{P}) \leq \delta_\ell(\epsilon, w; \mathbb{P})$, hence the desired inequality (9) holds. Now, by the definition (10), we have δ_ℓ is non-decreasing in ϵ and positive for \mathbb{P} -calibrated ℓ . However, ℓ could be such that $\delta_\ell(\epsilon, w; \mathbb{P})$ does not increase once ϵ is sufficiently large, or only increases at a sublinear rate; in this case $\eta = \delta_\ell^{**}$ is going to be 0 for $\epsilon \geq 0$, so the inequality (9) will be useless. To prevent this, we make the assumption that $\mathbb{E}[L(d, c) | w] - \min_{d' \in \mathbb{R}^m} \mathbb{E}[L(d', c) | w] \leq B$ for all $w \in W$, $d \in \mathbb{R}^m$. We can then re-define $\delta_\ell(\epsilon, w; \mathbb{P}) = \infty$ for $\epsilon > B$, and take $\eta = \delta_\ell^{**}$. This ensures that $\eta(\epsilon) > 0$ for $\epsilon \in (0, B]$. To ensure that such a B exists, we define the following quantities:

$$B_X := \max_{x, x' \in X} \|x - x'\|_2, \quad B_f := \max_{x, x' \in X} \{f(x) - f(x')\}, \quad B_C := \max_{c \in C} \|c\|_2. \quad (11)$$

Note that since X is compact and f is continuous on X , $B_X, B_f < \infty$.

ASSUMPTION 2. *The quantity $B_C < \infty$. (This means that $\mathbb{E}[c | w] \in \text{Conv}(C)$ is uniformly bounded over $w \in W$.)*

REMARK 5. Under Assumption 2 and using the fact that X is compact, we have

$$\begin{aligned} & \mathbb{E}[L(d, c) | w] - \min_{d' \in \mathbb{R}^m} \mathbb{E}[L(d', c) | w] \\ &= f(x^*(d)) - f(x^*(\mathbb{E}[c | w])) + \mathbb{E}[c | w]^\top (x^*(d) - x^*(\mathbb{E}[c | w])) \\ &\leq f(x^*(d)) - f(x^*(\mathbb{E}[c | w])) + \|\mathbb{E}[c | w]\|_2 \|x^*(d) - x^*(\mathbb{E}[c | w])\|_2 \\ &\leq B_f + B_C B_X < \infty, \end{aligned}$$

where the first inequality follows from Cauchy-Schwarz. ■

Another subtlety that we need to consider is that there needs to be a *single* fixed η for which (9) holds for all $w \in W$. Because of this, the definition $\eta = \delta_\ell^{**}$ is not well-defined as δ_ℓ in (10) depends on $w \in W$. To get around this, we need to strengthen the definition of calibration to be uniform across $w \in W$. In summary, the additions we need to make to the assumptions from Section 4 are Assumption 2, which ensures a uniform bound on the expected true loss, and a stronger definition of calibration, which we give next. Notice, however, that since our proof technique is different to that of Theorem 1, we need only measurability of ℓ , and not necessarily its convexity in d . In practice, however, convexity of ℓ in d gives us implementable algorithms with performance guarantees.

5.2. Risk Bounds via Uniform Calibration

We consider the following strengthening of Definition 3.

DEFINITION 5. We say that a loss function ℓ is *uniformly calibrated* with respect to a class of distributions \mathcal{P} on $W \times C$, or *\mathcal{P} -uniformly calibrated*, if, for all $\epsilon > 0$, there exists $\delta > 0$ such that for all $\mathbb{P} \in \mathcal{P}$, $w \in W$ and $d \in \mathbb{R}^m$, we have

$$\mathbb{E}[\ell(d, c) | w] - \inf_{d'} \mathbb{E}[\ell(d', c) | w] < \delta \implies \mathbb{E}[L(d, c)] - \inf_{d'} \mathbb{E}[L(d', c)] < \epsilon. \quad (12)$$

Note that Definition 5 considers a class of distributions \mathcal{P} so that we can get distribution-independent guarantees. This is due to practical considerations where knowledge of \mathbb{P} may not be available explicitly, but rather we may know that \mathbb{P} belongs to some class \mathcal{P} , so we may aim to get guarantees on the class \mathcal{P} .

If ℓ is \mathcal{P} -uniformly calibrated, then we define

$$\delta_\ell(\epsilon; \mathcal{P}) := \inf_{\substack{d \in \mathbb{R}^m \\ w \in W \\ \mathbb{P} \in \mathcal{P}}} \left\{ \mathbb{E}[\ell(d, c) | w] - \min_{d' \in \mathbb{R}^m} \mathbb{E}[\ell(d', c) | w] : \mathbb{E}[L(d, c)] - \min_{d' \in \mathbb{R}^m} \mathbb{E}[L(d', c)] \geq \epsilon \right\}. \quad (13)$$

REMARK 6. If ℓ is \mathcal{P} -calibrated, then $\delta_\ell(\epsilon; \mathcal{P}) > 0$ for all $\epsilon > 0$ by taking the contrapositive of (12), and is non-decreasing in ϵ . In addition, if Assumption 2 holds, then $\delta_\ell(\epsilon; \mathcal{P}) = \infty$ for $\epsilon > B_f + B_C B_X$ since the infimum is infeasible. Also, $\delta_\ell(\epsilon; \mathcal{P}) = 0$ for $\epsilon < 0$. Furthermore, measurability of $\delta_\ell(\cdot; \mathcal{P})$ follows by a similar proof to Lemma EC.5. ■

Remark 6 shows that positivity of δ_ℓ is necessary for \mathcal{P} -uniform calibration. We next establish that it is also sufficient.

LEMMA 3. *A surrogate loss function ℓ is \mathcal{P} -uniformly calibrated if and only if $\delta_\ell(\epsilon; \mathcal{P}) > 0$ for all $\epsilon > 0$.*

We now have the tools to prove the risk guarantee for uniform calibration. This is presented as Theorem 3 below, and we utilize a result of Steinwart (2007, Theorem 2.13) to prove it. Remark 5 allows us to apply this result in the prediction and optimization context. In this proof, it is crucial to ensure that the risk guarantee is non-trivial, i.e., verifying that δ_ℓ^{**} is positive on its domain. We utilize Lemma 3 for this purpose.

THEOREM 3. *Suppose that ℓ is \mathcal{P} -uniformly calibrated, and that Assumption 2 holds. Define*

$$\delta_\ell^{**}(\epsilon; \mathcal{P}) := \sup_{h'} \{h'(\epsilon) : h' \text{ convex function on } \mathbb{R}, h' \leq \delta_\ell(\cdot; \mathcal{P}) \text{ pointwise on } (0, B_f + B_C B_X)\},$$

where B_f, B_C, B_X are defined in (11). Then $\delta_\ell^{**}(\epsilon; \mathcal{P})$ is positive for $\epsilon \in (0, B_f + B_C B_X]$, and for any $\mathbb{P} \in \mathcal{P}$, $g : W \rightarrow \mathbb{R}^m$,

$$\delta_\ell^{**}(R(g, \mathbb{P}) - R(\mathbb{P}); \mathcal{P}) \leq R_\ell(g, \mathbb{P}) - R_\ell(\mathbb{P}).$$

In general, ensuring uniform calibration of a loss function is much harder than showing Fisher consistency. To end this section, we outline a general strategy to show uniform calibration for generic loss functions, which involves lower-bounding δ_ℓ defined in (13). In Section 5.3, we demonstrate this strategy for the squared loss ℓ_{LS} for the general class of square-integrable distributions, and then, invoking results from Steinwart (2007), we show uniform calibration for the class of separable loss functions with respect to the class of symmetric distributions. In Section 5.4, we show that for $m = 1$, uniform calibration can fail for the SPO+ loss function of Elmachoub and Grigas (2017) even when Fisher consistency is satisfied, and we give a sufficient condition on the class of continuous symmetric distributions that guarantees uniform calibration.

We first present an alternative form for δ_ℓ .

LEMMA 4. Consider δ_ℓ defined in (13). We have

$$\delta_\ell(\epsilon; \mathcal{P}) = \inf_{x, x' \in X} \inf_{\substack{d: x^*(d)=x \\ \bar{c}: x^*(\bar{c})=x'}} \inf_{\substack{\mathbb{P} \in \mathcal{P} \\ w \in W \\ \mathbb{E}[c|w]=\bar{c}}} \left\{ \mathbb{E}[\ell(d, c) | w] - \min_{d' \in \mathbb{R}^m} \mathbb{E}[\ell(d', c) | w] : f(x) - f(x') + \bar{c}^\top (x - x') \geq \epsilon \right\}. \quad (14)$$

We now give a bound on the distance between d and \bar{c} in the second infimum in (14).

LEMMA 5. Fix distinct $x, x' \in X$. Let d and \bar{c} be such that $x^*(d) = x$, $x^*(\bar{c}) = x'$. Then

$$\|d - \bar{c}\|_2 \geq \frac{\max\{0, f(x) - f(x') + \bar{c}^\top (x - x')\}}{\|x - x'\|_2}.$$

The strategy to prove \mathcal{P} -calibration of ℓ is as follows. First, fixing x, x' , notice that if \bar{c} and d are chosen according to the conditions of Lemma 5, together with the condition that $f(x) - f(x') + \bar{c}^\top (x - x') > \epsilon$, then $\|d - \bar{c}\|_2 > \epsilon / \|x - x'\|_2 \geq \epsilon / B_X > 0$ holds, where B_X is the Euclidean diameter of X defined in (11). Then, we want to give a positive lower bound for $\mathbb{E}[\ell(d, c) | w] - \min_{d' \in \mathbb{R}^m} \mathbb{E}[\ell(d', c) | w]$ over all distributions $\mathbb{P} \in \mathcal{P}$ and $w \in W$ such that $\mathbb{E}[c | w] = \bar{c}$. To this end, we will exploit the fact that $\arg \min_{d' \in \mathbb{R}^m} \mathbb{E}[\ell(d', c) | w]$ is close to $\mathbb{E}[c | w] = \bar{c}$, and the fact that $\|d - \bar{c}\|_2 > \epsilon / B_X$.

5.3. Uniform Calibration of the Squared Loss and Related Loss Functions

We now specifically consider the squared loss function:

$$\begin{aligned} \ell_{\text{LS}}(d, c) &:= \|d - c\|_2^2 \\ \mathcal{P} &:= \{\mathbb{P} : \forall w \in W, \mathbb{P}[\cdot | w] \text{ is square integrable, and } \mathbb{E}[c | w] \in \text{Conv}(C)\}. \end{aligned}$$

Due to the bias-variance decomposition of the squared loss, we can write $\delta_{\ell_{\text{LS}}}$ entirely as a geometric quantity, without any probabilistic terms.

LEMMA 6. Consider the case of the squared loss ℓ_{LS} and \mathcal{P} as defined above. Then, we have

$$\delta_{\ell_{\text{LS}}}(\epsilon; \mathcal{P}) = \inf_{x, x' \in X} \inf_{\substack{d \in \mathbb{R}^m: x^*(d) = x \\ \bar{c} \in \text{Conv}(C): x^*(\bar{c}) = x'}} \left\{ \|d - \bar{c}\|_2^2 : f(x) - f(x') + \bar{c}^\top (x - x') \geq \epsilon \right\}.$$

Using Lemmas 5 and 6, we derive \mathcal{P} -uniform calibration of the squared loss ℓ_{LS} .

THEOREM 4. The squared loss ℓ_{LS} is \mathcal{P} -uniformly calibrated, with

$$\delta_{\ell_{\text{LS}}}(\epsilon; \mathcal{P}) \geq \frac{\epsilon^2}{B_X^2} > 0 \quad \text{for all } \epsilon > 0.$$

COROLLARY 3. For the squared loss ℓ_{LS} , we have

$$\frac{1}{B_X^2} (R(g, \mathbb{P}) - R(\mathbb{P}))^2 \leq R_{\ell_{\text{LS}}}(g, \mathbb{P}) - R_{\ell_{\text{LS}}}(\mathbb{P}).$$

REMARK 7. Theorem 4 and Corollary 3 show that bounding the risk of the squared loss of a predictor $g : W \rightarrow \mathbb{R}^m$ is enough to bound the true risk. Intriguingly, this holds despite the fact that the squared loss contains no information about the optimization problem at hand (i.e., f or X). This means that minimization of the true risk can be achieved by training a predictor g without any information on the optimization problem, which is quite counter-intuitive. Furthermore, let $g = (g_1, \dots, g_m)$ where each $g_j : W \rightarrow \mathbb{R}$, and observe that

$$R_{\ell_{\text{LS}}}(g, \mathbb{P}) - R_{\ell}(\mathbb{P}) = \sum_{j \in [m]} \left(\mathbb{E}[(g_j(w) - c_j)^2] - \inf_{g'_j} \mathbb{E}[(g'_j(w) - c_j)^2] \right).$$

Thus, the excess squared loss risk is separable in the coefficients $j \in [m]$, hence we can train individual predictors $g_j : W \rightarrow \mathbb{R}$ to predict each coefficient c_j . Our results state that individual squared error risk bounds are enough to obtain bounds on the true risk $R(g, \mathbb{P})$. In particular, invoking Corollary 3 gives

$$R(g, \mathbb{P}) - R(\mathbb{P}) \leq B_X \sqrt{\sum_{j \in [m]} \left(\mathbb{E}[(g_j(w) - c_j)^2] - \inf_{g'_j} \mathbb{E}[(g'_j(w) - c_j)^2] \right)}.$$

Again, this is quite counter-intuitive, since we know that a small change in only one coefficient of d can change the optimal solution $x^*(d)$. ■

Remark 7 states that squared error risk bounds on individual coefficients $j \in [m]$ are enough to bound the true optimality gap risk, which essentially states that one-dimensional least squares regression on each coefficient $j \in [m]$ is sufficient for end-to-end prediction and optimization. Several other loss functions have been utilized in regression, due to their superior finite-sample performance. For example, the absolute deviation loss from Example 5 or the Huber loss have been used for heavy-tailed data due to their reduced sensitivity to outliers. Steinwart (2007, Section 4.3) studies the use of alternate loss functions in regression, and their risk relationships to the squared loss risk. By invoking these results, we can correspondingly obtain bounds on the true risk. More precisely, we have the following result.

LEMMA 7. For each $j \in [m]$, let $\ell_j : \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}$ be a loss function such that there exists a non-decreasing function $\eta_j : (0, \infty) \rightarrow (0, \infty)$ that satisfies

$$\mathbb{E}[(g_j(w) - c_j)^2] - \inf_{g'_j} \mathbb{E}[(g'_j(w) - c_j)^2] \leq \eta_j \left(\mathbb{E}[\ell_j(g_j(w), c_j)] - \inf_{g'_j} \mathbb{E}[\ell_j(g_j(w), c_j)] \right) \quad (15)$$

for any $g_j : W \rightarrow \mathbb{R}$ and $\mathbb{P} \in \mathcal{P}$. Then, denoting $g = (g_1, \dots, g_m)$,

$$R(g, \mathbb{P}) - R(\mathbb{P}) \leq B_X \sqrt{\sum_{j \in [m]} \eta_j \left(\mathbb{E}[\ell_j(g_j(w), c_j)] - \inf_{g'_j} \mathbb{E}[\ell_j(g_j(w), c_j)] \right)}.$$

Thus, when we use a separable loss function $\ell(d, c) = \sum_{j \in [m]} \ell_j(d_j, c_j)$ to train a predictor g , we can obtain true risk bounds by deriving bounds on each $\mathbb{E}[\ell_j(g_j(w), c_j)] - \inf_{g'_j} \mathbb{E}[\ell_j(g_j(w), c_j)]$. Conditions for the existence of the functions η_j are given by results from [Steinwart \(2007\)](#). To obtain them, we need to restrict the class of distributions. Precisely, we define \mathcal{P}_{sym} to be the class of square integrable distributions such that for all $w \in W$ and $j \in [m]$, $\mathbb{P}[c_j | w]$ is a symmetric distribution, i.e., $c_j - \mathbb{E}[c_j | w]$ has the same conditional distribution as $\mathbb{E}[c_j | w] - c_j$.

THEOREM 5 ([Steinwart \(2007, Theorems 4.19, 4.20\(ii\)\)](#)). Fix any $j \in [m]$. Let $\ell_j(d_j, c_j) = \psi_j(d_j - c_j)$ where $\psi_j : \mathbb{R} \rightarrow [0, \infty)$ is symmetric, i.e., $\psi_j(r) = \psi_j(-r)$, and uniformly convex, i.e., there exists some non-decreasing $\delta_j : [0, \infty) \rightarrow [0, \infty)$ with $\eta_j(0) = 0$ such that for all $\alpha \in [0, 1]$ and $r, r' \in \mathbb{R}$,

$$\alpha \psi_j(r) + (1 - \alpha) \psi_j(r') - \psi_j(\alpha r + (1 - \alpha)r') \geq \alpha(1 - \alpha) \delta_j(|r - r'|^2).$$

Then, for any $g_j : W \rightarrow \mathbb{R}$ and $\mathbb{P} \in \mathcal{P}_{\text{sym}}$, we have

$$\frac{1}{4} \delta_j^{**} \left(\mathbb{E}[(g_j(w) - c_j)^2] - \inf_{g'_j} \mathbb{E}[(g'_j(w) - c_j)^2] \right) \leq \mathbb{E}[\psi_j(g_j(w) - c_j)] - \inf_{g'_j} \mathbb{E}[\psi_j(g_j(w) - c_j)].$$

While the proof of Theorem 5 can be found in the relevant sections of [Steinwart \(2007\)](#), we give a more concise version in [EC.5](#).

5.4. Uniform Calibration of the SPO+ Loss in Example 8

Recall our Example 6 that studied the SPO+ loss (8) introduced in [Elmachtoub and Grigas \(2017, Definition 3\)](#). It was shown in [Elmachtoub and Grigas \(2017, Theorem 1\)](#) that this loss is Fisher consistent, hence by Theorem 2 it is \mathbb{P} -calibrated whenever $\mathbb{P}[c | w]$ is centrally symmetric and continuous for all $w \in W$. On the other hand, the uniform calibration of the SPO+ loss (8) has not yet been studied. In this section, we examine its uniform calibration for the special one-dimensional case $m = 1$, i.e., Example 8; to our knowledge, the general m case remains open.

Recall Example 8 has $m = 1$, $f = 0$, $X = [-1/2, 1/2]$, and we will take $C = \mathbb{R}$. In this case, recall that the loss function (8) becomes

$$\ell_{\text{SPO}+}(d, c) := \frac{1}{2} (|2d - c| - 2d \text{sign}(c) + |c|).$$

For this loss function, [Elmachtoub and Grigas \(2017\)](#) studied a particular class of probability distributions that are symmetric and continuous over \mathbb{R} . Recall that a continuous distribution is one such that the probability density function (w.r.t. Lebesgue measure) is positive over all of \mathbb{R} . For simplicity, we consider the same class of symmetric, continuous distributions over \mathbb{R} , i.e.,

$$\mathcal{P}_{\text{cont, sym}} := \{\mathbb{P} : \forall w \in W, \mathbb{P}[c | w] \text{ is continuous and symmetric}\}.$$

For this class of distributions, in [Elmachtoub and Grigas \(2017\)](#) it was shown that the (conditional) mean is the unique minimizer of $\min_{d' \in \mathbb{R}} \mathbb{E}[\ell(d', c) | w]$.

LEMMA 8 ([Elmachtoub and Grigas \(2017, Theorem 1\)](#)). *Let $\mathbb{P} \in \mathcal{P}_{\text{cont, sym}}$. Then for any $w \in W$, the unique minimizer of $\min_{d' \in \mathbb{R}} \mathbb{E}[\ell_{\text{SPO}^+}(d', c) | w]$ is $d^* = \mathbb{E}[c | w]$.*

Using [Lemmas 4 and 8](#), and noting

$$x^*(d) = \begin{cases} -1/2, & d > 0 \\ 0, & d = 0 \\ 1/2, & d < 0, \end{cases}$$

we have

$$\delta_{\ell_{\text{SPO}^+}}(\epsilon; \mathcal{P}_{\text{cont, sym}}) = \inf_{w \in W} \inf_{d, \bar{c} \in \mathbb{R}: d\bar{c} < 0} \inf_{\substack{\mathbb{P} \in \mathcal{P}_{\text{cont, sym}} \\ \mathbb{E}[c | w] = \bar{c}}} \{\mathbb{E}[\ell_{\text{SPO}^+}(d, c) | w] - \mathbb{E}[\ell_{\text{SPO}^+}(\bar{c}, c) | w] : |\bar{c}| > \epsilon\}.$$

Fixing $w \in W$, assume that $\mathbb{E}[c | w] = \bar{c} > 0$, hence $d < 0 < \epsilon < \bar{c}$. Since the function $\mathbb{E}[\ell_{\text{SPO}^+}(d, c) | w] = \frac{1}{2}(\mathbb{E}[|2d - c| | w] - 2d(\mathbb{P}[c > 0 | w] - \mathbb{P}[c < 0 | w]) + \mathbb{E}[|c| | w])$ is convex in d and hence continuous, when restricting $d < 0$, the closest $\mathbb{E}[\ell_{\text{SPO}^+}(d, c) | w]$ can get to the minimizer $\mathbb{E}[\ell_{\text{SPO}^+}(\bar{c}, c) | w]$ is at $d = 0$, i.e., $\mathbb{E}[\ell_{\text{SPO}^+}(0, c) | w] = \mathbb{E}[|c| | w]$. A similar argument holds for $\bar{c} < 0$. Therefore,

$$\begin{aligned} \delta_{\ell_{\text{SPO}^+}}(\epsilon; \mathcal{P}_{\text{cont, sym}}) &= \inf_{w \in W} \inf_{\substack{|\bar{c}| > \epsilon \\ \mathbb{P} \in \mathcal{P}_{\text{cont, sym}} \\ \mathbb{E}[c | w] = \bar{c}}} \{\mathbb{E}[\ell_{\text{SPO}^+}(0, c) | w] - \mathbb{E}[\ell_{\text{SPO}^+}(\bar{c}, c) | w]\} \\ &= \inf_{w \in W} \inf_{\substack{|\bar{c}| > \epsilon \\ \mathbb{P} \in \mathcal{P}_{\text{cont, sym}} \\ \mathbb{E}[c | w] = \bar{c}}} \left\{ \mathbb{E}[|c| | w] - \frac{1}{2} \left(\mathbb{E}[|2\bar{c} - c| | w] + 2\bar{c}(\mathbb{P}[c > 0 | w] - \mathbb{P}[c < 0 | w]) + \mathbb{E}[|c| | w] \right) \right\} \\ &= \inf_{w \in W} \inf_{\substack{\mathbb{P} \in \mathcal{P}_{\text{cont, sym}} \\ |\mathbb{E}[c | w]| > \epsilon}} \{\mathbb{E}[c | w] (\mathbb{P}[c > 0 | w] - \mathbb{P}[c < 0 | w])\}, \end{aligned}$$

where the third equality follows because $\mathbb{P}[c | w]$ is symmetric, so $2\bar{c} - c$ has the same conditional distribution as c , thus $\mathbb{E}[|2\bar{c} - c| | w] = \mathbb{E}[|c| | w]$. Unfortunately, we can show that $\delta_{\ell}(\epsilon; \mathcal{P}_{\text{cont, sym}}) = 0$ for all $\epsilon > 0$. This is due to the following result.

PROPOSITION 1. *For any $\epsilon > 0$, we can construct a sequence of symmetric, continuous distributions $\{\mathbb{P}^{(k)}\}_{k \in \mathbb{N}}$ on \mathbb{R} with $|\mathbb{E}^{(k)}[c]| \geq \epsilon$ such that $\mathbb{E}^{(k)}[c] (\mathbb{P}^{(k)}[c > 0] - \mathbb{P}^{(k)}[c < 0]) \rightarrow 0$. Therefore, by [Lemma 3](#), ℓ_{SPO^+} is not $\mathcal{P}_{\text{cont, sym}}$ -calibrated even in the restricted $m = 1$ setting.*

In contrast to this, we close this section by establishing a uniform calibration result for $\ell_{\text{SPO}+}$ for the case of the more restrictive class of continuous and symmetric distributions with uniformly bounded margin $|\mathbb{P}[c > 0 | w] - \mathbb{P}[c < 0 | w]|$.

PROPOSITION 2. For $\alpha > 0$, let

$$\mathcal{P}_{\text{cont,sym},\alpha} := \left\{ \mathbb{P} : \forall w \in W, \begin{array}{l} \mathbb{P}[c | w] \text{ is continuous and symmetric} \\ |\mathbb{P}[c > 0 | w] - \mathbb{P}[c < 0 | w]| \geq \alpha \end{array} \right\}.$$

Then, $\ell_{\text{SPO}+}$ is $\mathcal{P}_{\text{cont,sym},\alpha}$ -calibrated, and we have

$$R(g, \mathbb{P}) - R(\mathbb{P}) \leq \frac{1}{\alpha} (R_{\ell_{\text{SPO}+}}(g, \mathbb{P}) - R_{\ell_{\text{SPO}+}}(\mathbb{P})).$$

6. Computational Study

In this section, we conduct a computational study in order to investigate the effect of consistency in end-to-end prediction and optimization frameworks, and the effect of using a loss function that takes into account the optimization problem information. For this purpose, we examine the squared loss ℓ_{LS} and the SPO+ loss $\ell_{\text{SPO}+}$ in our experiments. Recall that the squared loss ℓ_{LS} does not take into account any information about the optimization problem, e.g., f or X , yet in Theorem 4 and Corollary 3 we provided true risk bounds in terms of the surrogate squared loss risk bounds. In contrast, [Elmachtoub and Grigas \(2017\)](#) proposed the SPO+ loss function $\ell_{\text{SPO}+}$ (8), which incorporates information about the optimization problem, and is known to be Fisher consistent with respect to certain distributions (see Example 6) but has weaker calibration properties than ℓ_{LS} (see Section 5.4).

Recall also that the squared loss ℓ_{LS} and the SPO+ loss $\ell_{\text{SPO}+}$ are defined as follows:

$$\begin{aligned} \ell_{\text{LS}}(d, c) &:= \|d - c\|_2^2 \\ \ell_{\text{SPO}+}(d, c) &:= f(x^*(c)) + (2d - c)^\top x^*(c) - \min_{x \in X} \{f(x) + (2d - c)^\top x\} = L(c, 2d - c). \end{aligned}$$

Note that [Elmachtoub and Grigas \(2017, Section 3.2\)](#) originally defined the SPO+ loss for linear objectives $c^\top x$ only, with $f = 0$. However, the above definition is a straightforward extension of their derivation for the objective $f(x) + c^\top x$.

We investigate three problem classes. First, we examine portfolio optimization using real-world data, where consistency is not known a priori. Second, we examine the fractional knapsack problem on simulated data, where some chosen parameters control the degree of non-linearity of the underlying data model, and thereby the consistency of certain loss functions. Third, we examine multiclass classification on simulated data, where the SPO+ loss is provably inconsistent (see Example 9), but the squared loss is consistent.

In all problem classes, we compare linear predictors $w \mapsto g(w) := Vw$ where V is obtained by solving the empirical risk minimization problem

$$\min_{V \in \mathbb{R}^{m \times k}} \frac{1}{n} \sum_{i \in [n]} \ell(Vw_i, c_i) \quad (16)$$

for different loss functions ℓ on the same historical data $\{(w_i, c_i)\}_{i \in [n]}$.

Our results suggest the following key managerial insights:

- On portfolio instances constructed from real data, there is no significant difference between using ℓ_{LS} loss and the more high-powered SPO+ loss which takes into account optimization problem information.

- Recall that we have shown theoretically that on the multiclass classification problem, the SPO+ loss is provably inconsistent (Example 9). Moreover, our numerical results on these problem instances show that the performance of SPO+ loss (expectedly) deteriorates. This highlights the importance of ensuring a consistent loss function is used in practice whenever possible.

- On the fractional knapsack instances, we observe that constructing a close approximation of the true loss by regularization, while conceptually reasonable, does not provide good empirical results due to the considerable increase in computational effort required to find the corresponding estimator. Therefore, computational efficiency plays an important role in the prediction and optimization context.

- Our experiments on the fractional knapsack instances also highlight that there are further properties besides consistency and calibration that can be investigated, such as robustness to model misspecification, where SPO+ has an advantage.

6.1. Mean-variance portfolio optimization

The mean-variance portfolio optimization problem can be expressed as the following constrained quadratic optimization problem

$$\min_{x \in X} \{f(x) - c^\top x\}, \quad \text{where } f(x) = \frac{1}{2}x^\top Qx, \quad X := \{x \in \mathbb{R}^m : p^\top x = b, x \geq 0\},$$

and $Q \succ 0$ is positive definite matrix. This problem arises from portfolio optimization: x denotes a vector of weights for each asset which specifies what proportion of our wealth to invest in each one, the random vector c represents returns of each stock, the quadratic term $f(x) = \frac{1}{2}x^\top Qx$ represents the risk of the portfolio (usually its variance), and a wealth constraint is imposed with $p = \mathbf{1}$ and $b = 1$.

In our study, we assume that c is uncertain but Q is fixed and known. We follow the common hypothesis in portfolio optimization that the expected cost vector can be described via a linear model $\mathbb{E}[c | \tilde{w}] = \tilde{b} + \tilde{V}\tilde{w}$, where \tilde{w} are market factors (see (Fama and French 1992)). In this setting,

\tilde{b} is the mean vector and \tilde{V} is called the ‘factor loading matrix.’ The goal in this problem is to estimate both \tilde{b} and \tilde{V} . To simplify notation, we append a 1 to each feature vector and denote $w = (\tilde{w}, 1)$. Similarly, we add \tilde{b} as a column to \tilde{V} , and denote $V = (\tilde{V}, \tilde{b})$. Thus, our model is $\mathbb{E}[c | w] = Vw$, and we aim to estimate V . We do this by again minimizing (16) where we take ℓ to be ℓ_{LS} or $\ell_{\text{SPO+}}$. Note that for objectives of type $f(x) - c^\top x$, the SPO+ loss is

$$\ell_{\text{SPO+}}(d, c) = L(c, 2d - c) = f(x^*(c)) - (2d - c)^\top x^*(c) - \min_{x \in X} \{f(x) - (2d - c)^\top x\}.$$

Usually, in portfolio optimization, we are permitted to have entries of x negative, which means we short-sell some assets. We show that if we redefine the domain to be $X := \{x \in \mathbb{R}^m : p^\top x = b\}$ without the non-negativity constraints, the SPO+ loss and the true loss are equivalent.

PROPOSITION 3. *Let $X := \{x \in \mathbb{R}^m : p^\top x = b\}$ and $A := Q^{-1} - \frac{1}{p^\top Q^{-1} p} Q^{-1} p (Q^{-1} p)^\top$. Then, for any d , the optimal solution to $\min_{x \in X} \{\frac{1}{2} x^\top Q x - d^\top x\}$ is*

$$x^*(d) = Ad + \frac{b}{p^\top Q^{-1} p} Q^{-1} p.$$

Furthermore,

$$L(d, c) = \frac{1}{2} x^*(d)^\top Q x^*(d) - c^\top x^*(d) - \min_{x \in X} \left\{ \frac{1}{2} x^\top Q x - c^\top x \right\} = \frac{1}{2} (d - c)^\top A (d - c).$$

Consequently,

$$\ell_{\text{SPO+}}(d, c) = L(c, 2d - c) = 2(c - d)^\top A (c - d) = 4L(d, c).$$

When we consider linear predictors $w \mapsto Vw$, we can show that the least squares loss $\ell_{\text{LS}}(d, c) = \frac{1}{2} \|d - c\|_2^2$ also optimizes the true loss. More precisely, given data $\{(w_i, c_i) : i \in [n]\}$, a solution to $\frac{1}{n} \sum_{i \in [n]} L(Vw_i, c_i)$ can be obtained by minimizing $\frac{1}{n} \sum_{i \in [n]} \ell_{\text{LS}}(Vw_i, c_i)$.

PROPOSITION 4. *Given a matrix $A \succeq 0$ and random variables $(w, c) \sim \mathbb{P}$ such that $\mathbb{E}[ww^\top]$ is invertible, we have*

$$\arg \min_V \mathbb{E} \left[\frac{1}{2} (Vw - c)^\top A (Vw - c) \right] = \arg \min_V \mathbb{E} \left[\frac{1}{2} \|Vw - c\|_2^2 \right] + \left\{ \tilde{V} : A\tilde{V} = \mathbf{0} \right\}.$$

Consequently, when $X = \{x \in \mathbb{R}^m : p^\top x = b\}$ and $A = Q^{-1} - \frac{1}{p^\top Q^{-1} p} Q^{-1} p (Q^{-1} p)^\top$, the minimizers of $\mathbb{E}[\ell_{\text{LS}}(Vw, c)]$ are also minimizers of $\mathbb{E}[L(Vw, c)]$.

Proofs of Propositions 3 and 4 are in Section EC.6.

For this reason, in our numerical study we henceforth impose non-negativity constraints on our decision variables $X := \{x \in \mathbb{R}^m : x \geq 0, p^\top x = b\}$. We generate instances from data on stocks that remained in the S&P 500 index for all 1258 trading days between January 1, 2003 and December

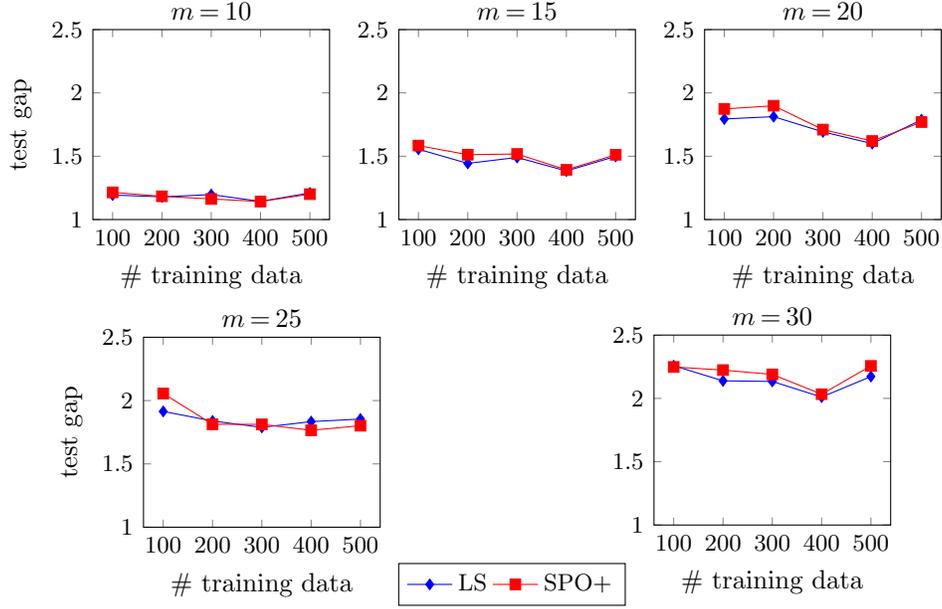


Figure 2 Median test optimality gap for different m for portfolio optimization.

31, 2007. We also collected data on the three Fama-French factors for these trading days, these are our feature vectors, with a 1 appended, so $k = 4$.

We consider $m \in \{10, 15, \dots, 30\}$, and for each m , we generate 100 random instances by choosing m random stocks. For each instance, we collect $n \in \{100, \dots, 500\}$ consecutive days of stock returns for the set of chosen stocks; stock returns for a particular day are recorded as the percentage increase/decrease of that day's price from the previous day's price. The matrix Q is the $m \times m$ sample covariance matrix of the stock returns computed from the n training days. We then estimate V from the n days of stock returns data via optimizing the least squares loss and the SPO+ loss. We evaluate the performance of our estimated V on the next $N = 10$ days after the n -day window in the training data, by first taking the factor data w for each test day, computing Vw , using that to compute a portfolio $x^*(Vw)$, then computing the objective of that portfolio on the actual $f(x^*(Vw)) - c^\top x^*(Vw)$ for that day. We report the median optimality gap $L(Vw, c) = f(x^*(Vw)) - c^\top x^*(Vw) - (f(x^*(c)) - c^\top x^*(c))$ (so lower is better) in Figure 2, which shows little difference between using the SPO+ loss and least squares on this class of problems with real data.

6.2. Fractional knapsack problem

In the case of fractional knapsack linear programs, we have

$$\max_{x \in X} d^\top x, \quad \text{where } X := \{x \in [0, 1]^m : p^\top x \leq B\}, \quad \text{and } f(x) = 0. \quad (17)$$

Here, $p \in \mathbb{R}^m$ is some fixed positive vector, and $B > 0$ is the capacity of the knapsack. As before, we test ℓ_{LS} and $\ell_{\text{SPO}+}$. Note that due to the max-type optimization problem, the SPO+ loss becomes

$$\ell_{\text{SPO}+}(d, c) = \max_{x \in X} (2d - c)^\top x - (2d - c)^\top x^*(c) = L(c, 2d - c).$$

For this problem class, we also test an additional loss function

$$\ell_{\text{reg},\lambda}(d, c) := c^\top x^*(c) - c^\top x_\lambda^*(d), \quad x_\lambda^*(d) := \arg \max_{x \in X} \left\{ d^\top x - \frac{\lambda}{2} \|x\|_2^2 \right\}.$$

Note that the loss function $\ell_{\text{reg},\lambda}$ is nothing but the *exact* optimality gap evaluated at the *unique* solution to the regularized knapsack problem with the objective function $d^\top x - \frac{\lambda}{2} \|x\|_2^2$ that includes a regularization term. We consider the regularized objective due to the fact that the set of optimal solutions $X^*(d)$ for the unregularized problem does not admit a simple model. By adding a regularizer, however, we can show that $\ell_{\text{reg},\lambda}(d, c)$ is mixed-integer linear representable.

PROPOSITION 5. *For fixed c and λ , The set $\{(d, t) : \ell_{\text{reg},\lambda}(d, c) \leq t\}$ admits a mixed-integer linear representation. Consequently, the empirical risk minimization problem (16) with $\ell = \ell_{\text{reg},\lambda}$ can be formulated as a mixed-integer linear program.*

The proof of Proposition 5 is rather standard; thus we give the details in Section EC.6.

We generate and test knapsack instances with $m = 10$ with synthetic data as follows. Each item weight p_j is a random integer between 1 and 1000. Then, B is a random integer between l and u , where $l = \max_{j \in [m]} p_j$, $u = (rl/\mathbf{1}^\top p + 1 - l/\mathbf{1}^\top p)\mathbf{1}^\top p$, where r is uniformly distributed on $[0, 1]$. For $m = 10$, $k = 5$, we generate 30 knapsack instances in this way, each paired with a randomly chosen ground truth coefficient matrix $V_0 \in \mathbb{R}^{m \times k}$. To generate data from V_0 , we use a similar scheme to that of [Elmachtoub and Grigas \(2017\)](#). The feature support set is $W := [-1, 1]^k$, and each w_i is drawn uniformly at random from W , except that the last entry $w_{ik} = 1$ always (in this way we can model a constant term in our predictor). Then, given hyper-parameters $\delta \geq 1, \epsilon \in (0, 1)$, each c_i is generated as

$$c_{ij} := \tilde{\epsilon}_{ij} (v_{0,j}^\top w)^\delta + \eta_{ij}, \quad j \in [m]$$

where $\tilde{\epsilon}_{ij}$ is uniformly distributed on $[1 - \epsilon, 1 + \epsilon]$ and $2\eta_{ij} + 1$ is an exponential random variable with scale parameter $\lambda = 1$ (thus η_{ij} has zero mean). Note that the exponentiation by δ is entry-wise, and that when $\delta = 1$ this means we have a linear model with random noise. We test $\delta = 1, 3, 5$ and $\epsilon = 0.1$ for each instance. We consider datasets of size $n = 100, 200, 300, 400, 500$ generated in this way. We trained $\ell_{\text{reg},\lambda}$ with $\lambda = 0.01$.

To test our predictors, we generate 10,000 points from the same distribution for each hyper-parameter setting and V_0 , and evaluate the average optimality gap using L on the test set for our

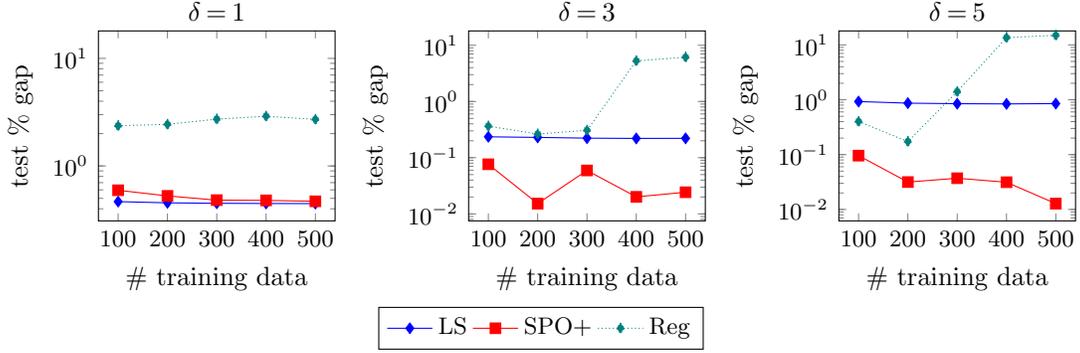


Figure 3 Average test relative optimality gap for different δ and $\epsilon = 0.2$ for the continuous knapsack problem.

predictors. Our results are shown in Figure 3 where we measure the average percentage optimality gap $\tilde{L}(d, c) = c^\top(x^*(c) - x^*(d)) / (c^\top x^*(c))$ (so lower is better).

First, it is clear that $\ell_{\text{reg}, \lambda}$ has poorer performance than ℓ_{LS} and $\ell_{\text{SPO+}}$. We attribute this to the fact that very few problems were solved to optimality within the five minute time limit. Therefore, this brings up the insight that despite $\ell_{\text{reg}, \lambda}$ being a close approximation to the true loss L on paper, computational considerations must be taken into account during training. Second, notice that for higher values of δ (i.e., as the true model becomes more non-linear), $\ell_{\text{SPO+}}$ outperforms ℓ_{LS} , which points to a ‘robustness to prediction model misspecification’ property that $\ell_{\text{SPO+}}$ might satisfy, and suggests that taking into account optimization information may increase performance under model misspecification. This phenomenon of $\ell_{\text{SPO+}}$ is currently unexplained by the theoretical results, and is an interesting direction for future research.

6.3. Multiclass classification

In our last class of examples, we consider the setting of multiclass classification from Example 9 with $C = \{c_j := \mathbf{1}_m - e_j : j \in [m]\} \subset \mathbb{R}^m$, $X = \text{Conv}\{e_j : j \in [m]\} \subset \mathbb{R}^m$ and $f(x) = 0$ for all $x \in X$. Recall that $e_j \in \mathbb{R}^m$ denotes the j th standard basis vector for $j \in [m]$. The SPO+ loss for this problem class is given by

$$\ell_{\text{SPO+}}(d, c_j) = (2d - c_j)^\top e_j - \min_{x \in X} (2d - c_j)^\top x = 2d_j - \min_{x \in X} \left(2d_j x_j + \sum_{j' \in [m], j' \neq j} (2d_{j'} - 1)x_{j'} \right).$$

A lifted representation of the SPO+ loss is given by the following proposition.

PROPOSITION 6. For fixed $c = c_j$, the set $\{(d, t) : \ell_{\text{SPO+}}(d, c_j) \leq t\}$ has a lifted representation

$$\left\{ \begin{array}{l} 2d_j - \gamma \leq t \\ (d, t, \gamma) : \gamma \leq 2d_j \\ \gamma \leq 2d_{j'} - 1, j' \in [m] \setminus \{j\} \end{array} \right\}.$$

Recall that in Example 9 we have shown SPO+ to be inconsistent for this problem theoretically. We next numerically compare the performance of ℓ_{LS} with $\ell_{\text{SPO+}}$ to investigate the effects of using an inconsistent loss function versus a consistent one. We use simulated data generated under the following multinomial logit model with parameters $v_1, \dots, v_m \in \mathbb{R}^k$:

$$\mathbb{P}[c = c_j | w] = \frac{\exp(-v_j^\top w)}{\sum_{j' \in [m]} \exp(-v_{j'}^\top w)} \quad j \in [m].$$

Under this model, given w , choosing the most likely class is equivalent to choosing the index j which gives the smallest $v_j^\top w$. We generate w uniformly at random from the unit cube $[0, 1]^k$. We fix $k = 4$ and $m = 4$. We do 100 repetitions of the following:

- Generate a true coefficient matrix $V^{\text{true}} \in \mathbb{R}^{m \times k}$ where each entry is distributed as a standard normal random variable.
- Generate test features $\{w_i^{\text{test}}\}_{i \in [N]}$ where $N = 100,000$. and compute the true probabilities $\{p_j(w_i^{\text{test}}) := \mathbb{P}[c = c_j | w_i^{\text{test}}]\}_{j \in [m], i \in [N]}$ using the true parameters V^{true} .
- For each $n \in \{100, 200, \dots, 1000\}$:
 - Generate training data $\{w_i^{\text{train}}, c_i^{\text{train}}\}_{i \in [n]}$ according to the true model.
 - Estimate the parameters using the two proposed methods to obtain $V_{\text{LS}}, V_{\text{SPO+}}$.
 - Use the test data to evaluate estimated parameters \hat{V} by computing

$$\frac{1}{N} \sum_{i \in [N]} \left(1 - \frac{1}{|\arg \min_{j' \in [m]} \hat{v}_{j'}^\top w_i^{\text{test}}|} \sum_{j \in \arg \min_{j' \in [m]} \hat{v}_{j'}^\top w_i^{\text{test}}} p_j(w_i^{\text{test}}) \right).$$

Note that the term in the outer summand is simply $\mathbb{E} \left[L(\hat{V} w_i^{\text{test}}, c) | w_i^{\text{test}} \right]$, the expected true loss of plugging in the vector $\hat{V} w_i^{\text{test}}$ into the optimization problem, and if it has a non-unique minimizer then one is chosen at random from the set of minimizers. We can estimate the best possible loss if we had true knowledge of the distribution, i.e., the Bayes loss, as

$$L_{\text{Bayes}} = \frac{1}{N} \sum_{i \in [N]} \left(1 - \max_{j \in [m]} p_j(w_i^{\text{test}}) \right).$$

In Figure 4 we plot the mean and a two standard deviation band for the gap of the true losses for each predictor relative to the Bayes loss across 100 runs, that is we plot statistics for the following quantity:

$$\frac{\mathbb{E}[L(\hat{V} w, c)] - L_{\text{Bayes}}}{L_{\text{Bayes}}}.$$

It is clear from Figure 4 that the SPO+ loss performs noticeably worse than the least squares loss. This observation is perhaps expected from our theoretical findings since we established that the

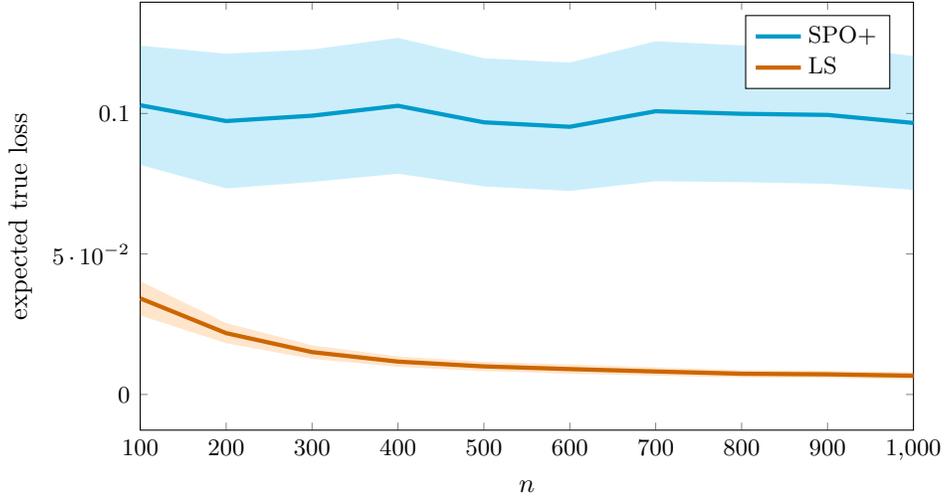


Figure 4 Mean (line) and a two standard deviation range (shaded region) of expected true loss across 100 runs for multiclass classification with $m = k = 4$.

SPO+ loss is inconsistent for this problem class. However, note that this performance difference between LS and SPO+ losses is still interesting because the true (conditional) expected cost vector

$$\mathbb{E}[c | w] = \left\{ 1 - \frac{\exp(-v_j^\top w)}{\sum_{j' \in [m]} \exp(-v_{j'}^\top w)} \right\}_{j \in [m]}$$

is a highly non-linear function of w and restricting it to a linear model, such as the case of $V_{LS}w$, may prevent us from learning the true functional form of $\mathbb{E}[c | w]$. A potential reason for the superior performance of the least squares loss is that it is consistent. In particular, for a given w , even though $V_{LS}w$ may not exactly be $\mathbb{E}[c | w]$, the minimal entry may still coincide. On the other hand, we showed in Example 9 in Section EC.4 that the true minimizer of $\mathbb{E}[\ell_{\text{SPO+}}(d, c) | w]$ is a constant vector, which we know will not give us the correct minimal entry of $\mathbb{E}[c | w]$. Our experiments thus highlight an important insight: consistency of a loss function matters more than whether the loss function takes into account optimization problem information. In particular, despite the fact that the SPO+ loss takes into account information from the optimization problem, its inconsistency for this problem class resulted in poor performance.

Notice also that there is no downward trend in the expected true loss of SPO+ as n increases. This is because of its inconsistency. In fact, a closer look at the estimated $V_{\text{SPO+}}$ reveals that it often estimates a zero matrix, which predicts the zero vector $V_{\text{SPO+}}w = \mathbf{0}$. This is consistent with the theoretical analysis of Example 9 in Section EC.4, where it shows that constant vectors d minimize $\mathbb{E}[\ell_{\text{SPO+}}(d, c)]$ when $\max_{j \in [m]} p_j < 1/2$.

7. Conclusion

In this paper, we explored risk guarantees for end-to-end prediction and optimization processes, which are prevalent in practice. We showed that the true non-convex optimality gap risk can be

minimized via minimizing the surrogate risk as long as the surrogate loss function is appropriately calibrated, and provided precise relationships between the two risks under these assumptions. We provided an equivalence result (Theorem 2) that allows us to easily check the weaker \mathbb{P} -calibration condition via Fisher consistency, and used it to explore calibration conditions for certain loss functions in Section 4. We also examined a stronger notion of uniform calibration for the least squares ℓ_{LS} and SPO+ loss ℓ_{SPO+} in Section 5. We found that the least squares loss satisfies Fisher consistency and uniform calibration under fairly general conditions, but in contrast the SPO+ loss fails to satisfy these conditions in some fairly natural settings. Our numerical results in Section 6.3 demonstrate that lack of consistency of the loss function can indeed have a detrimental effect on performance.

That said, our results in Section 6.2 re-affirm Elmachtoub and Grigas (2017)'s finding that the SPO+ loss performs well under model misspecification, e.g., when we restrict ourselves to learning a linear predictor but the true underlying data generation model is nonlinear. This suggests a future research direction to build our understanding of robustness to model misspecification of loss functions in the prediction and optimization context. Our findings from Sections 6.2 and 6.3 call for the design of new loss function that are consistent on broad problem classes and take into account optimization problem information as well. Some other interesting future directions include further exploration of uniform calibration for loss functions besides ℓ_{LS} and ℓ_{SPO+} and investigating theoretical and numerical performance of calibration on objective functions $f(x, c)$ depending non-linearly on c .

Acknowledgments

This research was supported by NSF grant CMMI 1454548. We would like to thank the review team for their suggestions that lead to significant improvements in terms of the presentation of the material.

References

- Ban GY, Rudin C (2019) The big data newsvendor: Practical insights from machine learning. *Operations Research* 67(1):90–108.
- Bartlett PL, Jordan MI, McAuliffe JD (2006) Convexity, classification, and risk bounds. *Journal of the American Statistical Association* 101(473):138–156, ISSN 01621459.
- Bengio Y (1997) Using a financial training criterion rather than a prediction criterion. *International Journal of Neural Systems* 8(04):433–443.
- Bertsimas D, Kallus N (2014) From Predictive to Prescriptive Analytics. *arXiv e-prints* arXiv:1402.5481.
- Bertsimas D, Van Parys B (2017) Bootstrap Robust Prescriptive Analytics. *arXiv e-prints* arXiv:1711.09974.
- Bogachev V (2007) *Measure Theory* (Springer-Verlag Berlin Heidelberg), ISBN 978-3-540-34514-5.

- Bonnans JF, Shapiro A (2000) *Perturbation analysis of optimization problems*. Springer Series in Operations Research (Springer, New York, NY), ISBN 978-1-4612-1394-9.
- Bousquet O, Boucheron S, Lugosi G (2004) *Introduction to Statistical Learning Theory*, 169–207 (Berlin, Heidelberg: Springer Berlin Heidelberg), ISBN 978-3-540-28650-9, URL http://dx.doi.org/10.1007/978-3-540-28650-9_8.
- Donti P, Amos B, Kolter JZ (2017) Task-based end-to-end model learning in stochastic optimization. *Advances in Neural Information Processing Systems 30*, 5484–5494 (Curran Associates, Inc.).
- Drusvyatskiy D, Lewis AS (2011) Generic nondegeneracy in convex optimization. *Proceedings of the American Mathematical Society* 139(7):2519–2527, ISSN 00029939, 10886826.
- Elmachtoub AN, Grigas P (2017) Smart “Predict, then Optimize”. Technical report, URL <https://arxiv.org/abs/1710.08005>.
- Fama EF, French KR (1992) The cross-section of expected stock returns. *The Journal of Finance* 47(2):427–465, URL <http://dx.doi.org/10.1111/j.1540-6261.1992.tb04398.x>.
- Goh CY, Jaillet P (2016) Structured Prediction by Conditional Risk Minimization. Technical report, URL <https://arxiv.org/abs/1611.07096>.
- Hanasusanto GA, Kuhn D (2013) Robust data-driven dynamic programming. *Advances in Neural Information Processing Systems 26*, 827–835 (Curran Associates, Inc.).
- Hannah L, Powell W, Blei DM (2010) Nonparametric density estimation for stochastic optimization with an observable state variable. *Advances in Neural Information Processing Systems 23*, 820–828 (Curran Associates, Inc.).
- Hiriart-Urruty JB, Lemaréchal C (2001) *Fundamentals of Convex Analysis* (Springer-Verlag Berlin Heidelberg), ISBN 978-3-642-56468-0.
- Ho CP, Hanasusanto GA (2019) On data-driven prescriptive analytics with side information: A regularized nadaraya-watson approach. Technical report.
- Kao Y, Roy BV, Yan X (2009) Directed regression. *Advances in Neural Information Processing Systems 22*, 889–897 (Curran Associates, Inc.).
- Lin Y (2004) A note on margin-based loss functions in classification. *Statistics & Probability Letters* 68(1):73–82, ISSN 0167-7152.
- Liyanage LH, Shanthikumar JG (2005) A practical inventory control policy using operational statistics. *Operations Research Letters* 33(4):341–348.
- Osokin A, Bach F, Lacoste-Julien S (2017) On structured prediction theory with calibrated convex surrogate losses. *Advances in Neural Information Processing Systems*, 302–313.
- Rockafellar RT (1970) *Convex analysis*. Princeton Mathematical Series (Princeton, N. J.: Princeton University Press).

- Stein EM, Shakarchi R (2009) *Real analysis: measure theory, integration, and Hilbert spaces* (Princeton University Press).
- Steinwart I (2002a) On the influence of the kernel on the consistency of support vector machines. *J. Mach. Learn. Res.* 2:67–93, ISSN 1532-4435.
- Steinwart I (2002b) Support vector machines are universally consistent. *Journal of Complexity* 18(3):768 – 791, ISSN 0885-064X.
- Steinwart I (2005) Consistency of support vector machines and other regularized kernel classifiers. *IEEE Transactions on Information Theory* 51(1):128–142, ISSN 0018-9448.
- Steinwart I (2007) How to compare different loss functions and their risks. *Constructive Approximation* 26(2):225–287, ISSN 1432-0940.
- Zhang T (2004) Statistical behavior and consistency of classification methods based on convex risk minimization. *Ann. Statist.* 32(1):56–85.

Electronic Companion to *Risk Guarantees for End-to-End Prediction and Optimization Processes*

EC.1. A Note on the Regularity of X^* and x^*

We denote the power set, the collection of all subsets of X , as 2^X . An important property that we exploit is that the argmin mapping $X^*(d)$ is, in a sense, well-behaved as we change d . More precisely, the sense of regularity that we use is upper semicontinuity, which stems from a result in perturbation analysis ([Bonnans and Shapiro 2000](#)).

DEFINITION EC.1. A multivalued function $F : \mathbb{R}^m \rightarrow 2^X$ is *upper semi-continuous* at a point $d \in \mathbb{R}^m$ if, for any open set U containing $F(d)$, there exists an open set U_d containing d such that for all $d' \in U_d$, $F(d') \subseteq U$. Equivalently, F is upper semi-continuous if, for any closed set V , the following set is closed:

$$\{d \in \mathbb{R}^m : F(d) \cap V \neq \emptyset\}.$$

LEMMA EC.1. *Suppose X is compact. Then the multivalued mapping $X^* : \mathbb{R}^m \rightarrow 2^X$ is upper semi-continuous.*

Proof. This follows immediately from verifying the conditions of [Bonnans and Shapiro \(2000, Proposition 4.4\)](#), which are straightforward to check due to the fact that the domain X does not change with the vector d . \square

We can use Lemma [EC.1](#) to show the existence of a measurable selection $x^*(d) \in X^*(d)$ via an application of the Kuratowski–Ryll–Nardzewski theorem on the existence of measurable selectors for multivalued mappings. We use the version stated in [Bogachev \(2007, Theorem 6.9.3\)](#).

LEMMA EC.2. *Suppose X is compact. Then there exists a measurable mapping $x^* : \mathbb{R}^m \rightarrow X$ such that $x^*(d) \in X^*(d)$ for all $d \in \mathbb{R}^m$.*

Proof. Consider the multivalued function $X^* : \mathbb{R}^m \rightarrow 2^X$ defined by $X^*(d) = \arg \min_{x \in X} d^\top x$. Note that since $d^\top x$ is continuous, $X^*(d) = \{x \in X : d^\top x = \min_{x' \in X} d^\top x'\}$ is closed (it is the inverse of a singleton). Now consider an open set U , and the sets

$$\hat{X}^*(U) := \{d \in \mathbb{R}^m : X^*(d) \cap U \neq \emptyset\}.$$

It is known that U can be represented as the countable union of closed sets: $U = \bigcup_{k \in \mathbb{N}} V_k$ where V_k are closed. Thus, we can write

$$\hat{X}^*(U) = \{d \in \mathbb{R}^m : \exists k \in \mathbb{N} \text{ s.t. } X^*(d) \cap V_k \neq \emptyset\} = \bigcup_{k \in \mathbb{N}} \{d \in \mathbb{R}^m : X^*(d) \cap V_k \neq \emptyset\}.$$

Now, since $X^*(d)$ is upper semicontinuous, $\{d \in \mathbb{R}^m : X^*(d) \cap U_k \neq \emptyset\}$ is closed, hence $\hat{X}^*(U)$ is a countable union of closed sets, hence measurable. This shows that $X^*(\cdot)$ satisfies the conditions of [Bogachev \(2007, Theorem 6.9.3\)](#), therefore there exists a measurable selection $x^*(d) \in X^*(d)$ for all $d \in \mathbb{R}^m$. \square

Furthermore, we can show that *any* selection x^* is at least *Lebesgue* measurable, using the following result of [Drusvyatskiy and Lewis \(2011\)](#).

LEMMA EC.3 ([Drusvyatskiy and Lewis \(2011, Corollary 3.5\)](#)). *The set*

$$D := \{d \in \mathbb{R}^m : X^*(d) \text{ is not a singleton}\}$$

has Lebesgue measure zero.

LEMMA EC.4. *Any selection $x^* : \mathbb{R}^m \rightarrow X$ such that $x^*(d) \in X^*(d)$ for all $d \in \mathbb{R}^m$ is Lebesgue measurable.*

Proof. Lemma [EC.2](#) tells us that there exists one such measurable selection \bar{x}^* . Consider another selection x^* . Then by Lemma [EC.3](#), \bar{x}^* and x^* differ on at most a set D with Lebesgue measure 0, which is Lebesgue measurable. Furthermore, all subsets of D are also Lebesgue measurable, so x^* must be Lebesgue measurable. \square

In order for our expectations to be well-defined, we make the following assumption.

ASSUMPTION EC.1. *Any probability distribution \mathbb{P} is defined on the σ -algebra of Lebesgue measurable sets.*

This is not practically restrictive, since any probability distribution we encounter in practice can be written as a mixture of a distribution which is absolutely continuous with respect to Lebesgue measure (i.e., it has a density function), and a discrete distribution supported on a countable set. Such a probability distribution is Lebesgue measurable.

EC.2. Proof of Results from Section 3

Proof of Lemma 1. Consider two extreme points of X , x_0, x_1 with $c^\top x_0 > c^\top x_1$. Choose d_0, d_1 such that minimizing $d_k^\top x$ over $x \in X$ results in the unique minimum x_k for $k = 0, 1$. Now note that $L(d_0, c) - L(d_1, c) = c^\top x_0 - c^\top x_1 > 0$. Let us now consider $d_\gamma = (1 - \gamma)d_0 + \gamma d_1$ for very small $\gamma \in (0, 1)$. When γ is sufficiently small, then d_γ will also have x_0 as a unique minimizer, so $L(d_\gamma, c) = L(d_0, c)$. Then because $L(d_0, c) > L(d_1, c)$, we have $L(d_\gamma, c) = L(d_0, c) > (1 - \gamma)L(d_0, c) + \gamma L(d_1, c)$. Hence, $L(d, c)$ is not convex in d for any such c . \square

Proof of Lemma 2. The measurability of $w \mapsto \mathbb{E}[c | w]$ is obvious by definition of the conditional expectation. Fix some measurable $g : W \rightarrow \mathbb{R}^m$. Observe that for $w \in W$,

$$\begin{aligned}
\mathbb{E}[L(g(w), c) | w] &= \mathbb{E} \left[f(x^*(g(w))) + c^\top x^*(g(w)) - \min_{x \in X} \{f(x) + c^\top x\} | w \right] \\
&= \mathbb{E} [f(x^*(g(w))) | w] + \mathbb{E}[c | w]^\top x^*(g(w)) - \mathbb{E} \left[\min_{x \in X} \{f(x) + c^\top x\} | w \right] \\
&= \mathbb{E} [f(x^*(g(w))) | w] + g^*(w)^\top x^*(g(w)) - \mathbb{E} \left[\min_{x \in X} \{f(x) + c^\top x\} | w \right] \\
&\geq f(x^*(g^*(w))) + g^*(w)^\top x^*(g^*(w)) - \mathbb{E} \left[\min_{x \in X} \{f(x) + c^\top x\} | w \right] \\
&= f(x^*(g^*(w))) + \mathbb{E}[c | w]^\top x^*(g^*(w)) - \mathbb{E} \left[\min_{x \in X} \{f(x) + c^\top x\} | w \right] \\
&= \mathbb{E} \left[f(x^*(g^*(w))) + c^\top x^*(g^*(w)) - \min_{x \in X} \{f(x) + c^\top x\} | w \right] \\
&= \mathbb{E}[L(g^*(w), c) | w],
\end{aligned}$$

where the inequality follows from the definition of $x^*(\cdot)$. Integrating both sides of this relation over $w \in W$ gives $R(g, \mathbb{P}) \geq R(g^*, \mathbb{P})$. Thus, g^* is the minimizer of $R(g, \mathbb{P})$.

The second result follows because

$$\begin{aligned}
\min_{d' \in \mathbb{R}^m} \mathbb{E}[L(d', c) | w] &= \min_{d' \in \mathbb{R}^m} \{f(x^*(d')) + \mathbb{E}[c | w]^\top x^*(d')\} - \mathbb{E} \left[\min_{x \in X} \{f(x) + c^\top x\} | w \right] \\
&= \min_{d' \in \mathbb{R}^m} \{f(x^*(d')) + g^*(w)^\top x^*(d')\} - \mathbb{E} \left[\min_{x \in X} \{f(x) + c^\top x\} | w \right] \\
&= f(x^*(g^*(w))) + g^*(w)^\top x^*(g^*(w)) - \mathbb{E} \left[\min_{x \in X} \{f(x) + c^\top x\} | w \right] \\
&= f(x^*(g^*(w))) + \mathbb{E}[c | w]^\top x^*(g^*(w)) - \mathbb{E} \left[\min_{x \in X} \{f(x) + c^\top x\} | w \right] \\
&= \mathbb{E}[L(g^*(w), c) | w],
\end{aligned}$$

and then integrating both sides over $w \in W$ gives $\mathbb{E}[\min_{d' \in \mathbb{R}^m} \mathbb{E}[L(d', c) | w]] = \mathbb{E}[L(g^*(w), c)] = R(\mathbb{P})$. \square

EC.3. Proof of Theorem 1

Define

$$\delta_\ell(\epsilon, w; \mathbb{P}) := \inf_{d \in \mathbb{R}^m} \left\{ \mathbb{E}[\ell(d, c) | w] - \min_{d' \in \mathbb{R}^m} \mathbb{E}[\ell(d', c) | w] : \mathbb{E}[L(d, c) | w] - \min_{d' \in \mathbb{R}^m} \mathbb{E}[L(d', c) | w] \geq \epsilon \right\}. \tag{EC.1}$$

Note that if ℓ is \mathbb{P} -calibrated, then $\delta_\ell(\epsilon, w; \mathbb{P}) > 0$ for all $\epsilon > 0, w \in W$ by taking the contrapositive of the implication in Definition 3. In order to prove Theorem 1, we first verify measurability for δ_ℓ .

LEMMA EC.5. *Suppose ℓ is measurable and satisfies Assumption 1, and that X is compact. For any $\epsilon > 0$, the function $\delta_\ell(\epsilon, \cdot; \mathbb{P}) : W \rightarrow \mathbb{R}$ is measurable.*

Proof. Consider the set

$$W_r := \{w \in W : \delta_\ell(\epsilon, w; \mathbb{P}) \leq r\}.$$

Showing measurability of $\delta_\ell(\epsilon, \cdot; \mathbb{P})$ boils down to showing that W_r is measurable. Rewrite

$$\begin{aligned} W_r &= \left\{ w \in W : \forall k \in \mathbb{N}, \exists d \in \mathbb{R}^m \text{ s.t. } \begin{array}{l} \mathbb{E}[\ell(d, c) | w] - \min_{d' \in \mathbb{R}^m} \mathbb{E}[\ell(d', c) | w] \leq r + 1/k \\ \mathbb{E}[L(d, c) | w] - \min_{d' \in \mathbb{R}^m} \mathbb{E}[L(d', c) | w] \geq \epsilon \end{array} \right\} \\ &= \bigcap_{k \in \mathbb{N}} \left\{ w \in W : \exists d \in \mathbb{R}^m \text{ s.t. } \begin{array}{l} \mathbb{E}[\ell(d, c) | w] - \min_{d' \in \mathbb{R}^m} \mathbb{E}[\ell(d', c) | w] \leq r + 1/k \\ \mathbb{E}[L(d, c) | w] - \min_{d' \in \mathbb{R}^m} \mathbb{E}[L(d', c) | w] \geq \epsilon \end{array} \right\} \end{aligned}$$

To this end, first consider the subset

$$\begin{aligned} W_L(\epsilon) &= \left\{ (w, d) \in W \times \mathbb{R}^m : \mathbb{E}[L(d, c) | w] - \min_{d' \in \mathbb{R}^m} \mathbb{E}[L(d', c) | w] \geq \epsilon \right\} \\ &= \left\{ (w, d) \in W \times \mathbb{R}^m : f(x^*(d)) + \mathbb{E}[c | w]^\top x^*(d) - \min_{x \in X} \{f(x) + \mathbb{E}[c | w]^\top x\} \geq \epsilon \right\}. \end{aligned}$$

This is measurable since $\mathbb{E}[c | w]$ is measurable in w by definition of conditional expectation, f is continuous hence measurable, and we have assumed $x^*(d)$ is measurable in d , which is possible by Lemma [EC.2](#).

Now consider the subset

$$W_\ell(\alpha) = \left\{ (w, d) \in W \times \mathbb{R}^m : \mathbb{E}[\ell(d, c) | w] - \min_{d' \in \mathbb{R}^m} \mathbb{E}[\ell(d', c) | w] \leq \alpha \right\}.$$

First observe that the function h defined by $h(w, d) = \mathbb{E}[\ell(d, c) | w]$ is continuous in d and measurable in w . Continuity in d follows because $\ell(d, c)$ is convex in d , and $h(w, d)$ is finite for any w by Assumption [1](#), and all convex functions are continuous in the relative interiors of their domains (see e.g., [Rockafellar \(1970, Theorem 10.1\)](#)). Measurability follows from measurability of ℓ and the definition of conditional expectation.

We now show that h is jointly measurable in (w, d) by showing that it is a pointwise limit of measurable functions. For $k \in \mathbb{N}$, consider the box $B_k := [-k, k]^m \subset \mathbb{R}^m$ and a finite set of grid points $G_k \subset B_k$ such that any point $d \in B_k$ is at most distance $1/k$ away from a grid point in Euclidean norm. If $d \in B_k$, define $h_k(w, d) = h(w, g)$ where $g \in B_k$ is the closest grid point to d (with ties broken arbitrarily), and if $d \notin B_k$ define $h_k(w, d) = 0$. Note that fixing g , $w \mapsto h_g(w) := h(w, g)$ is measurable in w . Now, h_k is the sum of finitely many functions of the form $\mathbf{1}_D(d)h_g(w)$ for some measurable set D and grid point g . It is easy to check that this is measurable, therefore h_k is measurable. Furthermore, by continuity of h in d , $h_k(w, d) \rightarrow h(w, d)$ pointwise. Therefore, h is measurable. Finally, the function $(w, d) \mapsto \min_{d' \in \mathbb{R}^m} h(w, d')$ is measurable because by continuity of h in d , we can write

$$\left\{ (w, d) : \min_{d' \in \mathbb{R}^m} h(w, d') \leq \alpha \right\} = \bigcup_{d' \in D_*, k \in \mathbb{N}} \{(w, d) : h(w, d') \leq \alpha + 1/k\}$$

where D_* is a countable dense subset of \mathbb{R}^m (e.g., \mathbb{Q}^m). This shows that $W_\ell(\alpha)$ is measurable because the function $h(w, d) - \min_{d' \in \mathbb{R}^m} h(w, d') = \mathbb{E}[\ell(d, c) | w] - \min_{d' \in \mathbb{R}^m} \mathbb{E}[\ell(d', c) | w]$ is measurable.

Now notice that the set

$$\left\{ (w, d) \in W \times \mathbb{R}^m : \begin{array}{l} \mathbb{E}[\ell(d, c) | w] - \min_{d' \in \mathbb{R}^m} \mathbb{E}[\ell(d', c) | w] \leq r + 1/k \\ \mathbb{E}[L(d, c) | w] - \min_{d' \in \mathbb{R}^m} \mathbb{E}[L(d', c) | w] \geq \epsilon \end{array} \right\} = W_\ell(r + 1/k) \cap W_L(\epsilon)$$

is measurable. Therefore, its projection onto W is measurable, which is

$$\left\{ w \in W : \exists d \in \mathbb{R}^m \text{ s.t. } \begin{array}{l} \mathbb{E}[\ell(d, c) | w] - \min_{d' \in \mathbb{R}^m} \mathbb{E}[\ell(d', c) | w] \leq r + 1/k \\ \mathbb{E}[L(d, c) | w] - \min_{d' \in \mathbb{R}^m} \mathbb{E}[L(d', c) | w] \geq \epsilon \end{array} \right\}.$$

This shows that W_r is measurable, concluding our proof. \square

Proof of Theorem 1. We wish to apply a result of [Steinwart \(2007, Theorem 2.8\)](#), for which we need to show that there exists measurable functions $b : W \rightarrow \mathbb{R}$ and $\delta : (0, \infty) \times W \rightarrow (0, \infty)$ such that $\mathbb{E}[|b(w)|] < \infty$, for any $w \in W$

$$\mathbb{E}[L(d, c) | w] - \min_{d' \in \mathbb{R}^m} \mathbb{E}[L(d', c) | w] \leq b(w),$$

and for any $\epsilon > 0$ and predictor $g : W \rightarrow \mathbb{R}^m$,

$$\begin{aligned} & \left\{ w \in W : \mathbb{E}[\ell(g(w), c) | w] - \min_{d' \in \mathbb{R}^m} \mathbb{E}[\ell(d', c) | w] < \delta(\epsilon, w) \right\} \\ & \subseteq \left\{ w \in W : \mathbb{E}[L(g(w), c) | w] - \min_{d' \in \mathbb{R}^m} \mathbb{E}[L(d', c) | w] < \epsilon \right\}. \end{aligned}$$

We first find b . Let Ω be the ℓ_∞ -diameter of the set X , which is finite since X is compact. Observe that for any $d \in \mathbb{R}^m$,

$$\begin{aligned} \mathbb{E}[L(d, c) | w] - \min_{d' \in \mathbb{R}^m} \mathbb{E}[L(d', c) | w] &= f(x^*(d)) + \mathbb{E}[c | w]^\top x^*(d) - \min_{x \in X} \{f(x) + \mathbb{E}[c | w]^\top x\} \\ &\leq \max_{x, x' \in X} \{f(x) - f(x') + \mathbb{E}[c | w]^\top (x' - x)\} \\ &\leq \max_{x, x' \in X} \{f(x) - f(x') + \|\mathbb{E}[c | w]\|_1 \|x' - x\|_\infty\} \\ &\leq \Omega \|\mathbb{E}[c | w]\|_1 + \max_{x, x' \in X} \{f(x) - f(x')\}. \end{aligned}$$

Therefore, we can define $b(w) := \Omega \|\mathbb{E}[c | w]\|_1 + \max_{x, x' \in X} \{f(x) - f(x')\}$ for each $w \in W$, which is integrable as $\mathbb{E}[\|\mathbb{E}[c | w]\|_1] \leq \mathbb{E}[\mathbb{E}[\|c\|_1 | w]] = \mathbb{E}[\|c\|_1] < \infty$ by [Assumption 1](#).

We will take $\delta := \delta_\ell(\cdot; \mathbb{P})$ defined in [\(EC.1\)](#), which is measurable by [Lemma EC.5](#). For any $g : W \rightarrow \mathbb{R}^m$, and $w \in W$ such that $\mathbb{E}[L(g(w), c) | w] - \min_{d' \in \mathbb{R}^m} \mathbb{E}[L(d', c) | w] \geq \epsilon$, by \mathbb{P} -calibration and definition of δ_ℓ we have $\mathbb{E}[\ell(g(w), c) | w] - \min_{d' \in \mathbb{R}^m} \mathbb{E}[\ell(d', c) | w] \geq \delta_\ell(\epsilon, w; \mathbb{P})$, therefore the required property for δ is satisfied.

Applying the result of [Steinwart \(2007, Theorem 2.8\)](#) then gives the risk bound. \square

EC.4. Proofs of Results from Section 4

Proof of Theorem 2. Denote

$$D_\ell(\alpha; w) := \left\{ d \in \mathbb{R}^m : \mathbb{E}[\ell(d, c) | w] - \min_{d' \in \mathbb{R}^m} \mathbb{E}[\ell(d', c) | w] < \alpha \right\}$$

$$D(\alpha; w) := \left\{ d \in \mathbb{R}^m : \mathbb{E}[L(d, c) | w] - \min_{d' \in \mathbb{R}^m} \mathbb{E}[L(d', c) | w] < \alpha \right\}.$$

Note that

$$\arg \min_{d' \in \mathbb{R}^m} \mathbb{E}[\ell(d', c) | w] = \bigcap_{\alpha > 0} D_\ell(\alpha; w), \quad \arg \min_{d' \in \mathbb{R}^m} \mathbb{E}[L(d', c) | w] = \bigcap_{\alpha > 0} D(\alpha; w).$$

Suppose first that ℓ is \mathbb{P} -calibrated. Then for any $\epsilon > 0$, there exists $\delta > 0$ (which can depend on w) such that $D_\ell(\delta; w) \subseteq D(\epsilon; w)$. In particular, since $D_\ell(\alpha; w) \subseteq D_\ell(\alpha'; w)$ for $\alpha \leq \alpha'$, we have

$$\arg \min_{d' \in \mathbb{R}^m} \mathbb{E}[\ell(d', c) | w] = \bigcap_{0 < \alpha \leq \delta} D_\ell(\alpha; w) \subseteq D(\epsilon; w).$$

Taking the intersection of the right hand side over $\epsilon > 0$, we have

$$\arg \min_{d' \in \mathbb{R}^m} \mathbb{E}[\ell(d', c) | w] \subseteq \arg \min_{d' \in \mathbb{R}^m} \mathbb{E}[L(d', c) | w],$$

hence ℓ is \mathbb{P} -Fisher consistent.

Suppose now that ℓ is not \mathbb{P} -calibrated. We show that it is also not \mathbb{P} -Fisher consistent. Fix an arbitrary $w \in W$. Note that the function $h : \mathbb{R}^m \rightarrow \mathbb{R}$ defined by $h(d) = \mathbb{E}[\ell(d, c) | w]$ is convex by convexity of $\ell(d, c)$, and hence under Assumption 1, it is continuous (see e.g., [Rockafellar \(1970, Theorem 10.1\)](#)).

Since ℓ is not \mathbb{P} -calibrated, there exists $w \in W$ and $\epsilon > 0$ such that for all $\delta > 0$, there exists $d(\delta) \in \mathbb{R}^m$ such that $h(d(\delta)) - \min_{d' \in \mathbb{R}^m} h(d') < \delta$ but $\mathbb{E}[L(d(\delta), c) | w] - \min_{d' \in \mathbb{R}^m} \mathbb{E}[L(d', c) | w] \geq \epsilon$.

Now, let $d_k = d(1/k)$ for $k \in \mathbb{N}$. Note that $\{d_k\}_{k \in \mathbb{N}} \subset D_\ell(1; w)$ which is compact since by Assumption 1 $\arg \min_{d' \in \mathbb{R}^m} h(d')$ is compact, so all level sets are bounded (see, e.g., [Rockafellar \(1970, Corollary 8.7.1\)](#)). Therefore, there exists a convergent subsequence $d'_k \rightarrow d \in \text{cl } D_\ell(1; w)$. Since h is continuous, we must have $d \in \arg \min_{d' \in \mathbb{R}^m} h(d')$.

We now want to show that $d \notin \arg \min_{d' \in \mathbb{R}^m} \mathbb{E}[L(d', c) | w]$. We know from Lemma [EC.1](#) that the argmin mapping $X^*(\cdot)$ is upper semi-continuous at d . Suppose for contradiction that $d \in \arg \min_{d' \in \mathbb{R}^m} \mathbb{E}[L(d', c) | w]$. Then we must have $X^*(d) \subseteq X^*(\mathbb{E}[c | w])$. Thus, for $\epsilon > 0$ the set

$$X^\circ(\epsilon') = \left\{ x' : f(x') + \mathbb{E}[c | w]^\top x' < \min_{x \in X} \{ f(x) + \mathbb{E}[c | w]^\top x \} + \epsilon' \right\}$$

is an ‘open’ set (as $x \mapsto f(x) + \mathbb{E}[c | w]^\top x$ is continuous) containing $X^*(d)$. Note that this is not open in \mathbb{R}^m by the usual topology, since $f(x)$ may be infinite for $x \notin X$. However, it is open when

we work with $X \subset \mathbb{R}^m$ as the *entire* topological space with the induced topology from \mathbb{R}^m . Then, by Definition EC.1 of upper semi-continuity, there exists a neighbourhood $D^\circ(\epsilon')$ of d such that for any $d^\circ \in D^\circ(\epsilon')$, $X^*(d^\circ) \subset X^\circ(\epsilon')$, which means that $\mathbb{E}[L(d^\circ, c) | w] < \min_{d' \in \mathbb{R}^m} \mathbb{E}[L(d', c) | w] + \epsilon'$ since $x^*(d^\circ) \in X^*(d^\circ) \subseteq X^\circ(\epsilon')$.

But now consider $\epsilon' < \epsilon$. Since $d'_k \rightarrow d$, $D^\circ(\epsilon')$ is open, and $d \in D^\circ(\epsilon')$, we eventually have $d'_k \in D^\circ(\epsilon')$ for sufficiently large k . But this contradicts the fact that by construction of the sequence $\{d_k\}_{k \in \mathbb{N}}$ we have $\epsilon' < \epsilon < \mathbb{E}[L(d'_k, c) | w] - \min_{d' \in \mathbb{R}^m} \mathbb{E}[L(d', c) | w] = \mathbb{E}[c | w]^\top x^*(d'_k) - \min_{x \in X} \mathbb{E}[c | w]^\top x$. \square

Proof of Corollary 2. Fix some $\epsilon > 0$. Take $\delta > 0$ corresponding to ϵ in Corollary 1. Since $R_\ell(g_n, \mathbb{P}) \rightarrow R_\ell(\mathbb{P})$, we have $R_\ell(g_n, \mathbb{P}) \leq R_\ell(\mathbb{P}) + \delta$ eventually. By Theorem 1, we will also have $R(g_n, \mathbb{P}) \rightarrow R(\mathbb{P}) + \epsilon$ eventually. \square

Proof of Example 8. Let us explore what $\operatorname{argmin}_{d' \in \mathbb{R}} \mathbb{E}[\ell(d, c) | w]$ is for our setting. For convenience, we fix $w \in W$, and omit the w in the notation, so that $D^* = D_w^*$, $\mathbb{E}[\cdot] = \mathbb{E}[\cdot | w]$ and $\mathbb{P}[\cdot] = \mathbb{P}[\cdot | w]$. Then

$$2\mathbb{E}[\ell(d, c)] = \mathbb{E}[|2d - c|] - 2d\mathbb{E}[\operatorname{sign}(c)] + \mathbb{E}[|c|] = \mathbb{E}[|2d - c|] + 2d(\mathbb{P}[c < 0] - \mathbb{P}[c > 0]) + \mathbb{E}[|c|].$$

This is a convex function in d , so we look at the subdifferential to determine its minimizers. Note that

$$\partial_d \mathbb{E}[|2d - c|] = \{2(\mathbb{P}[c < 2d] - \mathbb{P}[c > 2d]) + s\mathbb{P}[c = 2d] : s \in [-1, 1]\},$$

so

$$\partial_d \mathbb{E}[\ell(d, c)] = \{\mathbb{P}[c < 2d] - \mathbb{P}[c > 2d] + \mathbb{P}[c < 0] - \mathbb{P}[c > 0] + s\mathbb{P}[c = 2d] : s \in [-1, 1]\}.$$

For simplicity, let us assume that $\mathbb{P}[c = 2d] = 0$ for any d (many such distributions exist). Then $\mathbb{E}[\ell(d, c)]$ is differentiable with

$$\nabla_d \mathbb{E}[\ell(d, c)] = \mathbb{P}[c < 2d] - \mathbb{P}[c > 2d] + \mathbb{P}[c < 0] - \mathbb{P}[c > 0].$$

Denote d^* to be a minimizer of $\mathbb{E}[\ell(d, c)]$. If $\mathbb{P}[c < 0] = \mathbb{P}[c > 0]$, then setting $d = 0$ gives $\nabla_d \mathbb{E}[\ell(d, c)] = 0$, so $d^* = c = 0$. If $\mathbb{P}[c < 0] - \mathbb{P}[c > 0] < 0$, then $\nabla_d \mathbb{E}[\ell(d, c)]|_{d=0} < 0$, so increasing d from 0 will decrease $\mathbb{E}[\ell(d, c)]$. Thus, $d^* > 0$. However, note that $\mathbb{P}[c < 0] - \mathbb{P}[c > 0] < 0$ implies that the median of c is also > 0 . If $\mathbb{P}[c < 0] - \mathbb{P}[c > 0] > 0$, then $\nabla_d \mathbb{E}[\ell(d, c)]|_{d=0} > 0$, so decreasing d from 0 will decrease $\mathbb{E}[\ell(d, c)]$. Thus, $d^* < 0$. However, note that $\mathbb{P}[c < 0] - \mathbb{P}[c > 0] > 0$ implies that the median of c is also < 0 . In all cases, the minimizer d^* is of the same sign as the median of c . Now, if \mathbb{P} is a symmetric distribution, then the mean $\mathbb{E}[c]$ is equal to the median, and thus d^* has the same sign as $\mathbb{E}[c]$, so also minimizes $\mathbb{E}[L(d, c)]$. However, if the median has a different sign to the mean, then ℓ is not \mathbb{P} -Fisher consistent. Such distributions can be constructed by shifting a log-normal distribution, for example. \square

Proof of Example 9. With the distribution \mathbb{P} specified, $\min_{d' \in \mathbb{R}^m} \mathbb{E}[\ell_{\text{SPO}^+}(d', c)]$ can be expressed as the following linear program (making the change of variables $2d' \rightarrow d$):

$$\begin{aligned} \min_{d, \gamma} \quad & \sum_{j \in [m]} p_j (d_j - \gamma_j) \\ \text{s.t.} \quad & \gamma_j \leq d_j, \quad j \in [m] \\ & \gamma_j \leq d_k - 1, \quad j, k \in [m], k \neq j \\ & d, \gamma \in \mathbb{R}^m. \end{aligned}$$

We analyse this linear program. Fix a vector $d \in \mathbb{R}^m$. Let $j^* \in \arg \min_{j' \in [m]} d_{j'}$. Then since $p_k > 0$ for all $k \neq j^*$, the optimal choice of γ_k makes it as large as possible, so we set $\gamma_k = d_{j^*} - 1$ for $k \neq j^*$. In other words, for all but one index $j^* \in \arg \min_{j' \in [m]} d_{j'}$, we set $\gamma_j = \min_{j' \in [m]} d_{j'} - 1$. For j^* , we set $\gamma_{j^*} = \min \{d_{j^*}, \min_{j' \neq j^*} d_{j'} - 1\}$.

If there exists $j \neq j^*$ such that $d_{j^*} \leq d_j - 1$, then decreasing $d_j \downarrow d_{j^*} + 1$ does not violate any constraints since $\gamma_j = d_{j^*} - 1 < d_j$ and $\gamma_{j^*} = d_{j^*} \leq d_j - 1$, and decreases the objective. Therefore, without loss of generality, we assume that $d_j - 1 \leq d_{j^*}$ for all $j \neq j^*$. This implies that $\gamma_{j^*} = \min_{j' \neq j^*} d_{j'} - 1$.

Furthermore, if we have $j, k \in [m] \setminus \{j^*\}$, $j \neq k$ such that $d_j < d_k$, note that we can decrease $d_k \downarrow d_j$ without violating any constraints, since $\gamma_{j'} = d_{j^*} - 1 \leq d_j - 1 < d_k - 1 < d_k$ for all $j' \neq j^*$ and $\gamma_{j^*} \leq d_j - 1 < d_k - 1$. This implies that, without loss of generality, we can assume that for $j \neq j^*$, we have $d_j = \delta$ for some $\delta \in [d_{j^*}, d_{j^*} + 1]$. In particular, this implies that $\gamma_{j^*} = \delta - 1$, thus the objective becomes

$$\sum_{j \in [m]} p_j (d_j - \gamma_j) = (\delta - d_{j^*} + 1) \sum_{j \neq j^*} p_j + p_{j^*} (d_{j^*} - \delta + 1) = (1 - 2p_{j^*})(\delta - d_{j^*}) + 1.$$

This shows that if $p_{j^*} > 1/2$, then we should make δ as large as possible, i.e., $\delta = d_{j^*} + 1$. On the other hand, when $p_{j^*} < 1/2$, we set $\delta = d_{j^*}$, i.e., the optimal vector d^* is constant.

This implies that, if there exists $j^* \in [m]$ such that $p_{j^*} > 1/2$, and necessarily $j^* = \arg \max_{j' \in [m]} p_{j'}$, then the minimizers of $\mathbb{E}[\ell(d, c)]$ take the form $d_\alpha = (\alpha \mathbf{1}_m - e_{j^*})/2$ for $\alpha \in \mathbb{R}$. Clearly, $\arg \min_{j' \in [m]} d_{\alpha, j'} = j^*$, so for such distributions \mathbb{P} , ℓ_{SPO^+} is \mathbb{P} -Fisher consistent.

On the other hand, for distributions \mathbb{P} with $\max_{j' \in [m]} p_{j'} < 1/2$, ℓ_{SPO^+} is not \mathbb{P} -Fisher consistent, since the set of minimizers of $\mathbb{E}[\ell(d, c)]$ are the vectors $d_\alpha = \alpha \mathbf{1}_m$, $\alpha \in \mathbb{R}$, which cannot in general pick out the maximum probability class $j \in [m]$, i.e., the highest p_j . \square

EC.5. Proofs of Results from Section 5

Proof of Lemma 3. The ‘only if’ direction was established in Remark 6, so we only need to prove the ‘if’ direction.

When $\delta_\ell(\epsilon; \mathcal{P}) > 0$, take $0 < \delta \leq \delta_\ell(\epsilon; \mathcal{P})$, and noting that $\delta_\ell(\cdot; \mathcal{P})$ is non-decreasing, we get for any $d \in \mathbb{R}^m$, $w \in W$ and $\mathbb{P} \in \mathcal{P}$,

$$\mathbb{E}[\ell(d, c) | w] - \min_{d' \in \mathbb{R}^m} \mathbb{E}[\ell(d', c) | w] \leq \delta < \delta_\ell(\epsilon; \mathcal{P}).$$

If $d \in \mathbb{R}^m$, $w \in W$ and $\mathbb{P} \in \mathcal{P}$ were such that $\mathbb{E}[L(d, c) | w] - \min_{d' \in \mathbb{R}^m} \mathbb{E}[L(d', c) | w] > \epsilon$, we reach a contradiction since we would then by definition of $\delta_\ell(\cdot; \mathcal{P})$ in (13) have $\mathbb{E}[\ell(d, c) | w] - \min_{d' \in \mathbb{R}^m} \mathbb{E}[\ell(d', c) | w] \geq \delta_\ell(\epsilon; \mathcal{P})$. Thus, for any $w \in W$ and $\mathbb{P} \in \mathcal{P}$,

$$\mathbb{E}[L(d, c) | w] - \min_{d' \in \mathbb{R}^m} \mathbb{E}[L(d', c) | w] \leq \delta \implies \mathbb{E}[L(d, c) | w] - \min_{d' \in \mathbb{R}^m} \mathbb{E}[L(d', c) | w] \leq \epsilon.$$

□

Proof of Theorem 3. When ℓ is \mathcal{P} -uniformly calibrated, we know that $\delta_\ell(\epsilon; \mathcal{P}) > 0$ for any $\epsilon > 0$. Steinwart (2007, Lemma A.6) shows that this implies $\delta^{**}(\epsilon; \mathcal{P}) > 0$ for $\epsilon \in (0, B_f + B_C B_X]$.

We can now utilize Steinwart (2007, Theorem 2.13) to derive the risk bound. In order to do so, note that Steinwart (2007, Theorem 2.13) requires us to verify that for any $g: W \rightarrow \mathbb{R}$,

$$\operatorname{ess\,sup}_{w \in W} \mathbb{E}[L(g(w), c) | w] - \min_{d' \in \mathbb{R}^m} \mathbb{E}[L(d', c) | w] \leq B_f + B_C B_X,$$

where $\operatorname{ess\,sup}$ stands for essential supremum. The relation above follows from Remark 5 and from the definition of $\delta_\ell(\cdot; \mathcal{P})$ that ensures that for any $\epsilon > 0$ and $w \in W$, we have

$$\left\{ g: \mathbb{E}[\ell(g(w), c) | w] - \min_{d' \in \mathbb{R}^m} \mathbb{E}[\ell(d', c) | w] < \delta_\ell(\epsilon; \mathcal{P}) \right\} \subseteq \left\{ g: \mathbb{E}[L(g(w), c) | w] - \min_{d' \in \mathbb{R}^m} \mathbb{E}[L(d', c) | w] < \epsilon \right\}.$$

□

Proof of Lemma 4. Fix an arbitrary $w \in W$. Note that $\mathbb{E}[c | w] \in \{c' : x^*(c') \in X^*(\mathbb{E}[c | w])\} = \arg \min_{d' \in \mathbb{R}^m} \mathbb{E}[L(d', c) | w]$, hence we have

$$\begin{aligned} \mathbb{E}[L(d, c) | w] - \min_{d' \in \mathbb{R}^m} \mathbb{E}[L(d', c) | w] &= f(x^*(d)) + \mathbb{E}[c | w]^\top x^*(d) - \min_{d' \in \mathbb{R}^m} \{f(x^*(d')) + \mathbb{E}[c | w]^\top x^*(d')\} \\ &= f(x^*(d)) - f(x^*(\mathbb{E}[c | w])) + \mathbb{E}[c | w]^\top (x^*(d) - x^*(\mathbb{E}[c | w])). \end{aligned}$$

Hence

$$\begin{aligned} \delta_\ell(\epsilon; \mathcal{P}) &= \inf_{\substack{d \in \mathbb{R}^m \\ w \in W \\ \mathbb{P} \in \mathcal{P}}} \left\{ \mathbb{E}[\ell(d, c) | w] - \min_{d' \in \mathbb{R}^m} \mathbb{E}[\ell(d', c) | w] : \mathbb{E}[L(d, c) | w] - \min_{d' \in \mathbb{R}^m} \mathbb{E}[L(d', c) | w] \geq \epsilon \right\} \\ &= \inf_{d, \bar{c} \in \mathbb{R}^m} \inf_{\substack{w \in W \\ \mathbb{P} \in \mathcal{P} \\ \mathbb{E}[c | w] = \bar{c}}} \left\{ \mathbb{E}[\ell(d, c) | w] - \min_{d' \in \mathbb{R}^m} \mathbb{E}[\ell(d', c) | w] : f(x^*(d)) - f(x^*(\bar{c})) + \bar{c}^\top (x^*(d) - x^*(\bar{c})) \geq \epsilon \right\} \\ &= \inf_{x, x' \in X} \inf_{\substack{d: x^*(d) = x \\ \bar{c}: x^*(\bar{c}) = x'}} \inf_{\mathbb{P}: \mathbb{E}[c] = \bar{c}} \left\{ \mathbb{E}[\ell(d, c)] - \min_{d' \in \mathbb{R}^m} \mathbb{E}[\ell(d', c)] : f(x) - f(x') + \bar{c}^\top (x - x') \geq \epsilon \right\}. \end{aligned}$$

□

Proof of Lemma 5. Fix arbitrary distinct $x, x' \in X$. Consider the halfspace

$$H_0(x, x') = \{d' : f(x) - f(x') + (d')^\top(x - x') \leq 0\} \supseteq \{d' : x^*(d') = x\}.$$

Since $x^*(d) = x$, we have $d \in H_0(x, x')$. Now, if $\bar{c} \in H_0(x, x')$, then $f(x) - f(x') + \bar{c}^\top(x - x') \leq 0$, hence $\|d - \bar{c}\|_2 \geq 0$.

On the other hand, if $x^*(\bar{c}) = x'$ and $f(x) - f(x') + \bar{c}^\top(x - x') > 0$, then we have $\bar{c} \notin H_0(x, x')$, hence the distance between \bar{c} and d is bounded below by the distance between \bar{c} and the halfspace $H_0(x, x')$, which has the expression

$$\|d - \bar{c}\|_2 \geq \inf_{d' \in H_0(x, x')} \|d' - \bar{c}\|_2 = \frac{f(x) - f(x') + \bar{c}^\top(x - x')}{\|x - x'\|_2}.$$

□

Proof of Lemma 6. The usual bias-variance decomposition for squared error gives us

$$\begin{aligned} \mathbb{E}[\ell_{\text{LS}}(d, c) | w] &= \mathbb{E}[\|d - c\|_2^2 | w] \\ &= \|d - \mathbb{E}[c | w]\|_2^2 + 2\mathbb{E}[(d - \mathbb{E}[c | w])^\top(\mathbb{E}[c | w] - c)] + \mathbb{E}[\|\mathbb{E}[c | w] - c\|_2^2] \\ &= \|d - \mathbb{E}[c | w]\|_2^2 + \mathbb{E}[\|\mathbb{E}[c | w] - c\|_2^2]. \end{aligned}$$

Hence, we can minimize this by choosing $d = \mathbb{E}[c | w]$, and

$$\min_{d' \in \mathbb{R}^m} \mathbb{E}[\ell_{\text{LS}}(d', c) | w] = \mathbb{E}[\|\mathbb{E}[c | w] - c\|_2^2].$$

Therefore,

$$\mathbb{E}[\ell_{\text{LS}}(d, c) | w] - \min_{d' \in \mathbb{R}^m} \mathbb{E}[\ell_{\text{LS}}(d', c) | w] = \|d - \mathbb{E}[c | w]\|_2^2.$$

Substituting this into (14) and using the fact that the definition of \mathcal{P} tells us that $\bar{c} = \mathbb{E}[c | w]$ can take on any point in $\text{Conv}(C)$ gives the result. □

Proof of Theorem 4. Fixing distinct $x, x' \in X$, notice that if \bar{c}, d are chosen according to the conditions of Lemma 5, together with the condition that $f(x) - f(x') + \bar{c}^\top(x - x') > \epsilon$, then $\|d - \bar{c}\|_2 > \epsilon/\|x - x'\|_2 \geq \epsilon/B_X > 0$. Together with Lemma 6, we have, for all $\epsilon > 0$,

$$\delta_{\ell_{\text{LS}}}(\epsilon; \mathcal{P}) \geq \frac{\epsilon^2}{B_X^2} > 0.$$

Then \mathcal{P} -uniform calibration follows from Lemma 3. □

Proof of Corollary 3. The result follows by observing that $\epsilon^2/B_X^2 \leq \delta^{**}(\epsilon)$ since $\epsilon \mapsto \epsilon^2/B_X^2$ is already convex, and then applying Theorem 3. □

Proof of Theorem 5. Analogous to (13), define

$$\delta_j(\epsilon; \mathcal{P}_{\text{sym}}) := \inf_{\substack{d_j \in \mathbb{R} \\ w \in W \\ \mathbb{P} \in \mathcal{P}_{\text{sym}}}} \left\{ \mathbb{E}[\ell_j(d_j, c_j) | w] - \min_{d'_j \in \mathbb{R}} \mathbb{E}[\ell_j(d'_j, c_j) | w] : (d_j - \mathbb{E}[c_j | w])^2 > \epsilon \right\}.$$

We first show that $\delta_j(\epsilon; \mathcal{P}_{\text{sym}}) > 0$ for all $\epsilon > 0$.

First, fix $\mathbb{P} \in \mathcal{P}_{\text{sym}}$ and $w \in W$, and observe that for any d_j

$$\begin{aligned} \mathbb{E}[\ell_j(\mathbb{E}[c_j | w] + d_j, c_j) | w] &= \mathbb{E}[\psi_j(d_j - (c_j - \mathbb{E}[c_j | w])) | w] = \mathbb{E}[\psi_j(d_j - (\mathbb{E}[c_j | w] - c_j)) | w] \\ &= \mathbb{E}[\psi_j(-d_j + \mathbb{E}[c_j | w] - c_j) | w] \\ &= \mathbb{E}[\ell_j(\mathbb{E}[c_j | w] - d_j, c_j) | w]. \end{aligned}$$

Since ψ is strictly convex, $\mathbb{E}[\ell_j(d_j, c_j) | w]$ is strictly convex in d_j , thus for any $d_j \neq 0$,

$$\begin{aligned} &\mathbb{E}[\ell_j(\mathbb{E}[c_j | w] + d_j, c_j) | w] - \mathbb{E}[\ell_j(\mathbb{E}[c_j | w], c_j) | w] \\ &= \frac{1}{2} \mathbb{E}[\ell_j(\mathbb{E}[c_j | w] + d_j, c_j) | w] + \frac{1}{2} \mathbb{E}[\ell_j(\mathbb{E}[c_j | w] - d_j, c_j) | w] - \mathbb{E}[\ell_j(\mathbb{E}[c_j | w], c_j) | w] \\ &\geq \frac{1}{4} \mathbb{E}[\delta_j(4d_j^2) | w] = \frac{1}{4} \delta_j(4d_j^2). \end{aligned}$$

This shows that $\delta_j(\epsilon; \mathcal{P}_{\text{sym}}) \geq \delta_j(2\epsilon)/4$. Now, following the outline in Section 5.1 and proceeding similarly to the proof of Theorem 3, we deduce the risk bound. \square

Proof of Proposition 1. The proof is by construction. We will fix the mean of our class to be ϵ . Let ϕ be the density function of the standard normal distribution, and Φ be the distribution function (note that $\phi(c - \epsilon)$ is the density function of a $N(\epsilon, 1)$ random variable. Let $z_\epsilon = \Phi(-\epsilon) + 1 - \Phi(\epsilon)$ denote the probability that a standard normal variable is $< -\epsilon$ or $> \epsilon$. Furthermore, let $\{h(\cdot; \alpha)\}_{\alpha \in (0,1)}$ be a class of continuous functions such that for each $\alpha \in (0, 2/3)$, $h(r; \alpha) > 0$ for $r \in [0, 1]$, $h(1; \alpha) = 1$, $\int_{r=0}^1 h(r; \alpha) dr = \alpha$. Such a class can be defined as follows:

$$h(r; \alpha) = \begin{cases} \alpha/2, & 0 \leq r \leq (2 - 3\alpha)/(2 - \alpha) \\ (2 - \alpha)^2(r - 1)/(4\alpha) + 1, & (2 - 3\alpha)/(2 - \alpha) < r \leq 1. \end{cases}$$

For each $k \in \mathbb{N}$, define the following density function $\psi^{(k)}$:

$$\psi^{(k)}(c) = \begin{cases} \frac{z_\epsilon + (1 - 1/k)(1 - z_\epsilon)}{z_\epsilon} \phi(c - \epsilon), & c \leq 0 \\ \frac{z_\epsilon + (1 - 1/k)(1 - z_\epsilon)}{z_\epsilon} \phi(-\epsilon) h\left(1 - c/\epsilon; \frac{(1 - z_\epsilon)z_\epsilon}{2k\epsilon(z_\epsilon + (1 - 1/k)(1 - z_\epsilon))\phi(-\epsilon)}\right), & 0 < c < \epsilon \\ \frac{z_\epsilon + (1 - 1/k)(1 - z_\epsilon)}{z_\epsilon} \phi(-\epsilon) h\left(c/\epsilon - 1; \frac{(1 - z_\epsilon)z_\epsilon}{2k\epsilon(z_\epsilon + (1 - 1/k)(1 - z_\epsilon))\phi(-\epsilon)}\right), & \epsilon \leq c < 2\epsilon \\ \frac{z_\epsilon + (1 - 1/k)(1 - z_\epsilon)}{z_\epsilon} \phi(\epsilon - c), & c \geq 2\epsilon. \end{cases}$$

By construction, $\psi^{(k)}(c)$ is continuous and positive for all $c \in \mathbb{R}$, and integrates to 1. Let $\mathbb{P}^{(k)}$ denote the corresponding probability distribution, and by construction we have $\mathbb{P}^{(k)}[0 \leq c \leq 2\epsilon] = (1 - z_\epsilon)/k$. Therefore $\mathbb{P}^{(k)}[c > 0] - \mathbb{P}^{(k)}[c < 0] = (1 - z_\epsilon)/k \rightarrow 0$ as $k \rightarrow \infty$, but $\mathbb{E}^{(k)}[c] = \epsilon$ since $\psi^{(k)}(c - \epsilon) = \psi(\epsilon - c)$ is symmetric about ϵ . \square

Proof of Proposition 2. We know that

$$\delta_{\ell_{\text{SPO}^+}}(\epsilon; \mathcal{P}_{\text{cont,sym},\alpha}) = \inf_{w \in W} \inf_{\substack{\mathbb{P} \in \mathcal{P}_{\text{cont,sym},\alpha} \\ |\mathbb{E}[c|w]| > \epsilon}} \{ \mathbb{E}[c|w] (\mathbb{P}[c > 0|w] - \mathbb{P}[c < 0|w]) \}.$$

Using the property of $\mathcal{P}_{\text{cont,sym},\alpha}$, we deduce that

$$\delta_{\ell_{\text{SPO}^+}}(\epsilon; \mathcal{P}_{\text{cont,sym},\alpha}) \geq \inf_{w \in W} \inf_{\substack{\mathbb{P} \in \mathcal{P}_{\text{cont,sym},\alpha} \\ |\mathbb{E}[c|w]| > \epsilon}} \alpha |\mathbb{E}[c|w]| \geq \alpha \epsilon.$$

Thus since $\delta_{\ell_{\text{SPO}^+}}(\epsilon; \mathcal{P}_{\text{cont,sym},\alpha}) > 0$ for any $\epsilon > 0$, we have uniform calibration by Lemma 3. Furthermore, Theorem 3 gives us the risk bound. \square

EC.6. Proof of Results from Section 6

Proof of Proposition 3. Using Lagrange duality we know that the dual problem is

$$\min_{\gamma} \left\{ b\gamma + \frac{1}{2} (d - \gamma p)^\top Q^{-1} (d - \gamma p) \right\} = - \min_x \left\{ \frac{1}{2} x^\top Q x - d^\top x : p^\top x = b \right\},$$

and the optimal solution is $x^*(d) = Q^{-1}(d - \gamma^* p)$ where γ^* is the optimal dual solution. The closed form solution is $\gamma^* = \frac{1}{p^\top Q^{-1} p} (p^\top Q^{-1} d - b)$, hence

$$x^*(d) = Q^{-1} d - \frac{1}{p^\top Q^{-1} p} (p^\top Q^{-1} d - b) Q^{-1} p = Ad + \frac{b}{p^\top Q^{-1} p} Q^{-1} p.$$

Observe that $Ap = 0$, so

$$\begin{aligned} x^*(d)^\top Q x^*(d) &= \left(Ad + \frac{b}{p^\top Q^{-1} p} Q^{-1} p \right)^\top \left(QAd + \frac{b}{p^\top Q^{-1} p} p \right) \\ &= d^\top A^\top QAd + \frac{b^2}{p^\top Q^{-1} p} \\ &= d^\top A^\top \left(I - \frac{p(Q^{-1} p)^\top}{p^\top Q^{-1} p} \right) d + \frac{b^2}{p^\top Q^{-1} p} \\ &= d^\top Ad + \frac{b^2}{p^\top Q^{-1} p} \\ c^\top x^*(d) &= c^\top Ad + \frac{b \cdot p^\top Q^{-1} c}{p^\top Q^{-1} p} \\ \frac{1}{2} x^*(d)^\top Q x^*(d) - c^\top x^*(d) &= \frac{1}{2} d^\top Ad - c^\top Ad + \frac{b^2/2 - b \cdot p^\top Q^{-1} c}{p^\top Q^{-1} p}. \end{aligned}$$

Clearly we have $\frac{1}{2} x^*(c)^\top Q x^*(c) - c^\top x^*(c) = -\frac{1}{2} c^\top Ac + \frac{b^2/2 - b \cdot p^\top Q^{-1} c}{p^\top Q^{-1} p}$ so therefore

$$L(d, c) = \frac{1}{2} d^\top Ad - c^\top Ad + \frac{1}{2} c^\top Ac = \frac{1}{2} (d - c)^\top A (d - c).$$

The result now follows. \square

Proof of Proposition 4. First, notice that $\mathbb{E}[\frac{1}{2}\|Vw - c\|_2^2] = \frac{1}{2}\text{Tr}(V^\top V\mathbb{E}[ww^\top]) - \text{Tr}(V^\top\mathbb{E}[cw^\top]) + \frac{1}{2}\mathbb{E}[\|c\|_2^2]$. Via standard vector calculus, we have $\nabla_V\mathbb{E}[\frac{1}{2}\|Vw - c\|_2^2] = V\mathbb{E}[ww^\top] - \mathbb{E}[cw^\top]$, therefore the optimality condition of the least squares predictor is

$$V\mathbb{E}[ww^\top] = \mathbb{E}[cw^\top].$$

Now observe that we can write $\mathbb{E}[\frac{1}{2}(Vw - c)^\top A(Vw - c)] = \frac{1}{2}\text{Tr}(V^\top A V\mathbb{E}[ww^\top]) - \text{Tr}(V^\top\mathbb{E}[Acw^\top]) + \frac{1}{2}\mathbb{E}[c^\top A c]$. The gradient is $\nabla_V\mathbb{E}[\frac{1}{2}(Vw - c)^\top A(Vw - c)] = A V\mathbb{E}[ww^\top] - A\mathbb{E}[cw^\top]$. The optimality condition is $A V\mathbb{E}[ww^\top] = A\mathbb{E}[cw^\top]$. We can alternatively represent this as

$$V\mathbb{E}[ww^\top] \in \mathbb{E}[cw^\top] + \left\{ \tilde{V} : A\tilde{V} = \mathbf{0} \right\}.$$

Since $\mathbb{E}[ww^\top]$ is invertible, the result follows. \square

Proof of Proposition 5 This result follows immediately from Lemma EC.6 below. \square

LEMMA EC.6. *Assume that $\|d\|_1 \leq M$ and that $M_\tau = M / (\min_{j \in [n]} p_j)$. The set of optimal solutions to $\max_{x \in X} \left\{ d^\top x - \frac{\lambda}{2} \|x\|_2^2 \right\}$ can be characterized as*

$$\left\{ x : \begin{array}{l} p^\top x \leq B, \mathbf{0} \leq x \leq \mathbf{1} \\ \tau \geq 0, q, z \in \{0, 1\}^n, v \in \{0, 1\} \\ \tau \leq M_\tau v, B - p^\top x \leq B(1 - v) \\ d_j - p_j \tau \leq M q_j, p_j \tau - d_j \leq (M_\tau p_j + M)(1 - q_j), j \in [m] \\ d_j - p_j \tau - \lambda \leq M z_j, \lambda + p_j \tau - d_j \leq (M_\tau p_j + M + \lambda)(1 - z_j), j \in [m] \\ x_j \leq q_j, x_j \geq z_j, j \in [m] \\ \lambda x_j \leq d_j - p_j \tau + (M + M_\tau p_j)(1 - q_j), \lambda x_j \geq d_j - p_j \tau - M z_j, j \in [m]. \end{array} \right\}.$$

Proof of Lemma EC.6. Fix d . We consider the primal-dual pair of problems for the regularized fractional knapsack:

$$\begin{aligned} & \max_x \left\{ d^\top x - \frac{\lambda}{2} \|x\|_2^2 : p^\top x \leq B, \mathbf{0} \leq x \leq \mathbf{1} \right\} \\ & = \min_{s, y, \tau} \left\{ B\tau + \mathbf{1}^\top y + \frac{1}{2\lambda} \|s\|_2^2 : s \geq d - p\tau - y, s, y, \tau \geq 0 \right\}. \end{aligned}$$

Using the complementary slackness conditions, the set of primal-dual optimal pairs (x, s, y, τ) can be written as

$$X^*(d) = \left\{ (x, s, y, \tau) : \begin{array}{l} p^\top x \leq B, \mathbf{0} \leq x \leq \mathbf{1}, \\ s \geq d - p\tau - y, s, y, \tau \geq 0, \\ \tau(B - p^\top x) = 0, \\ y_i(1 - x_i) = 0, i \in [n] \\ x_i(s_i - (d_i - p_i\tau - y_i)) = 0, i \in [n] \\ s = \lambda x \end{array} \right\}$$

$$\text{Proj}_{x, \tau}(X^*(d)) = \left\{ (x, \tau) : \begin{array}{l} p^\top x \leq B, \mathbf{0} \leq x \leq \mathbf{1}, \\ \tau \geq 0, \\ \tau(B - p^\top x) = 0, \\ \lambda x_i = \max\{0, \min\{\lambda, d_i - p_i\tau\}\}, i \in [n] \end{array} \right\}.$$

To see why the second equality holds, consider some solution $(x, s, y, \tau) \in X^*(d)$. If $d_i - p_i\tau - y_i > 0$, then we need $\lambda x_i = s_i = d_i - p_i\tau - y_i$, which follows from $s_i \geq d_i - p_i\tau - y_i$, $x_i(s_i - (d_i - p_i\tau - y_i)) = 0$ and $s_i = \lambda x_i$. If $d_i - p_i\tau - y_i \leq 0$, then since $s_i = \lambda x_i$, we would have $x_i(s_i - (d_i - p_i\tau - y_i)) > 0$ if $s_i = \lambda x_i > 0$, so we must have $s_i = \lambda x_i = 0$. Therefore $\lambda x_i = \max\{0, d_i - p_i\tau - y_i\}$. We now show that $y_i = \max\{0, d_i - p_i\tau - \lambda\}$. To see this, suppose that $d_i - p_i\tau - \lambda > 0$. We know that $y_i \geq d_i - p_i\tau - \lambda x_i > 0$, and since $y_i(1 - x_i) = 0$, we have $x_i = 1$. Since $\lambda x_i = \lambda = \max\{0, d_i - p_i\tau - y_i\} = d_i - p_i\tau - y_i$ implies that $y_i = d_i - p_i\tau - \lambda$. Now suppose that $d_i - p_i\tau - \lambda \leq 0$. If $y_i > 0$, then since $y_i(1 - x_i) = 0$, we necessarily have $x_i = 1$. But then $\lambda x_i = \lambda \leq d_i - p_i\tau - y_i$ is a contradiction. Therefore we necessarily have $y_i = 0$. Substituting $y_i = \max\{0, d_i - p_i\tau - \lambda\}$ into $\lambda x_i = \max\{0, d_i - p_i\tau - y_i\}$ gives us $\lambda x_i = \max\{0, \min\{\lambda, d_i - p_i\tau\}\}$.

We assume that $\|d\|_1 \leq M$, and that we are given an a priori bound $\tau \leq M_\tau$. We can model the constraint $\tau(B - p^\top x) = 0$ as

$$\tau \leq M_\tau v, \quad B - p^\top x \leq B(1 - v), \quad v \in \{0, 1\}.$$

We now describe how to model the constraint $\lambda x_i = \max\{0, \min\{\lambda, d_i - p_i\tau\}\}$. First, since $\lambda > 0$, we have that $\min\{\lambda, d_i - p_i\tau\} \geq 0$ if and only if $d_i - p_i\tau \geq 0$. Let q_i be an indicator variable for this event, which we model as

$$d_i - p_i\tau \leq Mq_i, \quad p_i\tau - d_i \leq (M_\tau p_i + M)(1 - q_i).$$

Let z_i be an indicator variable for the event $d_i - p_i\tau \geq \lambda$, so we need the constraints

$$d_i - p_i\tau - \lambda \leq Mz_i, \quad \lambda + p_i\tau - d_i \leq (M_\tau p_i + M + \lambda)(1 - z_i).$$

Note that implicitly, we have $q_i \geq z_i$. When $q_i = 0$, we have $\lambda x_i = 0$. When $q_i = 1$ and $z_i = 1$, we have $x_i = 1$, and when $q_i = 1$ and $z_i = 0$, we have $\lambda x_i = d_i - p_i \tau$. Therefore we need the constraints

$$x_i \leq q_i, \quad x_i \geq z_i, \quad \lambda x_i \leq d_i - p_i \tau + (M + M_\tau p_i)(1 - q_i), \quad \lambda x_i \geq d_i - p_i \tau - M z_i.$$

This shows that the proposed MIP representation is correct. \square

Proof of Proposition 6. The dual of $\min_{x \in X} \left(2d_j x_j + \sum_{j' \in [m], j' \neq j} (2d_{j'} - 1)x_j \right)$ is $\max_{\gamma} \{ \gamma : \gamma \leq 2d_j, \gamma \leq 2d_{j'} - 1, j' \in [m] \setminus \{j\} \}$. The lifted representation immediately follows from this. \square