Management Science

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

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

Nam Ho-Nguyen, Fatma Kılınç-Karzan

To cite this article:
Nam Ho-Nguyen, Fatma Kılınç-Karzan (2022) Risk Guarantees for End-to-End Prediction and Optimization Processes. Management Science 68(12):8680-8698. https://doi.org/10.1287/mnsc.2022.4321

https://pubsonline.informs.org/journal/mnsc

MANAGEMENT SCIENCE
Vol. 68, No. 12, December 2022, pp. 8680–8698
ISSN 0025-1909 (print), ISSN 1526-5501 (online)

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

Nam Ho-Nguyen,[a,*] Fatma Kılınç-Karzan[b]

[a] Discipline of Business Analytics, The University of Sydney, Sydney, New South Wales 2006, Australia; [b] Tepper School of Business, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213
*Corresponding author

Contact: nam.ho-nguyen@sydney.edu.au, https://orcid.org/0000-0003-4464-7730 (NH-N); fkilinc@andrew.cmu.edu, https://orcid.org/0000-0001-5939-4575 (FK-K)

Received: June 4, 2019
Revised: December 29, 2020; September 9, 2021
Accepted: October 30, 2021
Published Online in Articles in Advance: March 11, 2022

https://doi.org/10.1287/mnsc.2022.4321

Abstract. Prediction methods are often employed to estimate parameters of optimization models. Although the goal in an end-to-end framework is to achieve good performance on the subsequent optimization model, a formal understanding of the ways in which prediction methods can affect optimization performance is notably lacking. This paper identifies conditions on prediction methods that can guarantee good optimization performance. We provide two types of results: asymptotic guarantees under a well-known Fisher consistency criterion and nonasymptotic performance bounds under a more stringent criterion. We use these results to analyze optimization performance for several existing prediction methods and show that in certain settings, methods tailored to the optimization problem can fail to guarantee good performance. Conversely, optimization-agnostic methods can sometimes, surprisingly, have good guarantees. In a computational study on portfolio optimization, fractional knapsack, and multiclass classification problems, we compare the optimization performance of several prediction methods. We demonstrate that lack of Fisher consistency of the prediction method can indeed have a detrimental effect on performance.

History: Accepted by Chung Piaw Teo, optimization.
Funding: This work was supported by the National Science Foundation, Division of Civil, Mechanical and Manufacturing Innovation [Grant 1454548].
Supplemental Material: Data and the e-companion are available at https://doi.org/10.1287/mnsc.2022.4321.

Keywords: stochastic optimization • prediction • end-to-end

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 historical data, often using statistical inference or machine learning. One can choose from a range of prediction methods, and this choice should be informed by the subsequent optimization task. The natural measure of optimization performance is the optimality gap between the solution obtained based on the predicted parameters and the full-information solution. However, the exact effect that the prediction method has on optimization performance is not well understood. In this paper, we consider a joint *end-to-end* view of the prediction and optimization processes and identify the critical properties of prediction methods that enable a guarantee on the optimality gap in the subsequent optimization problem.

More formally, we consider an optimization problem of the form

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

where $X \subset \mathbb{R}^m$ is a convex compact domain, and $f : X \to \mathbb{R}$ is a convex function. In our setting, the linear vector c is not known exactly, but instead is governed via covariates (or side information) 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 aim to solve (1) for c, yet we are only given w. We model the dependency of c on w with a function $g : W \to \mathbb{R}^m$, which we call a *prediction model*; given w, our prediction of c is $g(w)$.

Although we do not know the distribution $\mathbb{P}$, we have access to historical data $H_n := \{(w_i, c_i) : i \in [n]\}$, where the pairs (w_i, c_i) are realizations of independent and identically distributed random variables from the unknown distribution $\mathbb{P}$. The end-to-end prediction and optimization process that we wish to examine is as follows: first, use the data H_n to select a prediction model $g : W \to \mathbb{R}^m$. Second, when presented with a covariate w, solve the optimization model (1) with c replaced by the prediction $g(w)$. This process is commonly used for a variety of decision-making problems.

Below, we give a few 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. Covariates w of c to use for prediction can include seasonality, demand, supply, and other economic factors.

Example 2. Routing algorithms are often used to solve logistics problems found in a variety of businesses. For example, food delivery services must allocate orders to delivery drivers in a way that minimizes delivery times. This can be formulated in several ways (using, e.g., bipartite matching or vehicle routing models), but with a common underlying linear model type $\min_{x\in X} c^\top x$ (so $f(x) = 0$), where X represents the space of possible allocations. Here, the vector c summarizes information that determines the efficiency of the orders. This information may include elements such as what time an order is ready for pickup, what time a driver can start their route, and how long it takes a driver to deliver all items. This information is not readily available, but instead is predicted from covariates w, which can include the current location of the driver, restaurant addresses, delivery addresses, and order characteristics (e.g., number of items in the order).

Example 3. 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); $f(x) = \gamma x^\top \Sigma x$, where $x^\top \Sigma x$ is the variance of the portfolio return with Σ being the covariance matrix of the returns between the assets; and $c = -\mu$, where μ is the vector of mean returns for each asset. Here, γ is a weighting parameter that balances the trade-off between minimizing variance and maximizing returns. In many settings, Σ is assumed to be stable, and μ is predicted through market factors (e.g., liquidity, value, momentum, volume), which are the covariates w.

Example 4. Structured prediction is a form of multiclass classification designed to predict structured objects, such as sequences or graphs, from feature data; see, for example, Goh and Jaillet (2016), Osokin et al. (2017), and references therein. In structured prediction, given covariates w, one wishes to predict a structured object $\tilde{x}^*(w)$ from some output space $\tilde{X}$. Typically, $\tilde{X}$ is a combinatorial space. To build the prediction, a function $\tilde{g}(\tilde{x}; w)$ is constructed, then the prediction $\tilde{x}(w)$ is obtained by solving $\min_{\tilde{x}\in\tilde{X}} \tilde{g}(\tilde{x}; w)$. The prediction $\tilde{x}(w)$ can then be compared with $\tilde{x}^*(w)$. This fits into our end-to-end setting by taking X to be a simplex whose vertices correspond to objects in $\tilde{X}$, c to be the negative of the vertex of X corresponding to $x^*(w)$, and a specific prediction model given by $g(w) := \{\tilde{g}(\tilde{x}; w)\}_{\tilde{x}\in\tilde{X}}$.

Examples 1–4 demonstrate the ubiquity of end-to-end prediction and optimization problems. A key component of this is to select the prediction model g. We will refer to any technique/algorithm that uses the data H_n to select a prediction model g as a *prediction method*. There are a variety of prediction methods we may choose from, such as least squares minimization, lasso, elastic net, trimmed regression, and support vector machines. Which prediction method should we use in order to obtain a prediction model g that will best suit the subsequent optimization problem? The way in which we assess the *quality* of a prediction method becomes critical to answering this question. Given the subsequent optimization task, the most natural performance measure of a prediction $g(w)$ for c is the *optimality gap* for (1): $L(g(w), c) := f(x^*(g(w))) + c^\top x^*(g(w)) - (f(x^*(c)) + c^\top x^*(c))$, where $x^*(g(w))$ is the optimal solution of (1) using predicted parameters $g(w)$, and $x^*(c)$ is the optimal solution using true parameters c. We refer to the expected optimality gap $\mathbb{E}[L(g(w), c)]$ over $(w, c) \sim \mathbb{P}$ as the *optimization performance*; the smaller it is, the better the optimization performance the prediction method will achieve.

One strategy to obtain good optimization performance is to directly minimize $\mathbb{E}[L(g(w), c)]$ (or the empirical expectation on the data H_n). However, this is often intractable to implement, because in many settings, L is highly nonsmooth and nonconvex (see Example 5). Therefore, its minimization is not amenable to common numerical techniques. Recent work (Kao et al. 2009, Elmachtoub and Grigas 2022) has examined numerically tractable proxies for the optimality gap L, giving rise to prediction methods that are moderately tailored to the subsequent optimization problems.

That said, there are compelling reasons to use "optimization-agnostic" prediction methods that are not tailored to any optimization model. One setting is when the optimization objective is not known exactly ahead of time. This can occur in Example 3 when one is undecided about how to balance the portfolio mean

and the variance. Optimization-agnostic prediction methods, such as least squares minimization, are far more prevalent, and they are quite well developed in comparison. From a practical point of view, it is much easier to implement a decision-making system with an optimization-agnostic prediction method, as the technical prerequisites are lower.

In either case, it is important to understand when and how a *generic* prediction method can guarantee good optimization performance, regardless of whether it is tailored to the optimization problem or not. In this paper, we seek to develop this understanding by identifying general conditions that allow us to establish explicit guarantees on the optimization performance $\mathbb{E}[L(g(w),c)]$ of a prediction method. We then use these results to examine performance guarantees of several well-known prediction methods.

1.1. Contributions and Outline

Our contributions in this paper can be summarized as follows:

- Given probability distribution $\mathbb{P}$, we show that asymptotic optimization performance guarantees can be obtained for end-to-end prediction and optimization if our prediction method satisfies a condition called *Fisher consistency* (see Definitions 3 and 4 and Theorems 1 and 2). This provides a tool for easily checking which methods lead to optimization performance guarantees.
- We establish that *nonasymptotic* and *distribution-independent* optimization performance bounds can be obtained under more stringent conditions than Fisher consistency (see Definition 5 and Theorem 3). We develop a framework to check whether these conditions are satisfied for a given prediction method.
- We demonstrate the use of our results above by theoretically analyzing performance guarantees for several common prediction methods, including least squares minimization, and a prediction method proposed by Elmachtoub and Grigas (2022) that is tailored to the subsequent optimization problem, namely, the SPO+ method (Examples 6–11, Sections 5.3 and 5.4). Furthermore, we provide numerical evidence that violating these conditions can have a detrimental effect on optimization performance in practice. Specifically, we find that prediction methods tailored to optimization problems are not guaranteed to outperform optimization-agnostic ones in certain settings. Consequently, our study highlights the managerial insight that when designing end-to-end prediction and optimization frameworks, it is important to take the conditions that guarantee optimization performance into account in the design process.

Section 2 outlines the related literature. Section 3 formally defines the framework we study, and outlines the challenges. Section 4 develops the results on asymptotic performance guarantees, and Section 5 develops the results on explicit nonasymptotic performance bounds. Section 6 describes our computational study using linear prediction models on three problem classes with real and simulated data: portfolio optimization, fractional knapsack, and multiclass classification. We also present some managerial insights from our numerical study at the beginning of Section 6 and close with brief conclusions in Section 7. We relegate all of the proofs to the corresponding sections of the e-companion.

1.2. Notation

We use the following notation. Given a positive integer N, we use the shorthand $[N] := \{1,\dots,N\}$. For a subset C of some Euclidean space, we let $\mathrm{Conv}(C)$ denote its convex hull. 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 $X^*(d) := \arg\min_{x\in X}\{f(x)+d^\top x\}$ be the argmin mapping, and let $x^*(d)$ denote some selection from $X^*(d)$, selected in a deterministic manner. More precisely, $x^* : \mathbb{R}^m \to X$ 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 classification problem from machine learning is relevant to our end-to-end setting. In this problem, w is again a covariate vector, and $c \in \{1,\dots,k\}$ denotes a class label from a finite number of classes. These classes can be simple labels or more exotic constructs. For example, structured prediction (see Example 4) is a type of classification problem in which there are potentially an exponential number of structured objects, such as a sequence of (finitely many) base objects, and each structured object forms a class. In classification, a prediction model $g : W \to \{1,\dots,k\}$ will predict the class label from the covariate w. We ideally wish to minimize the misclassification probability $\mathbb{P}[g(w) \neq c]$, but this is known to be nonconvex and NP-hard to solve. Thus, most prediction methods, such as support vector machines or logistic regression, select a prediction model by minimizing a convex surrogate of the misclassification probability. Consequently, characterizing the relationship between the chosen method and the misclassification probability $\mathbb{P}[g(w) \neq c]$ becomes of interest. This topic is well studied and understood; for example, Steinwart (2002a, b, 2005), Lin (2004), Zhang (2004a, b), Bartlett et al. (2006), Tewari and Bartlett (2007), Osokin et al. (2017), and Duchi et al. (2018) provide general theoretical results on this relationship.

Classification can be seen as a special case of the end-to-end prediction and optimization setting for a

specific domain X, function f, and support set C (see Examples 9 and 11 for a detailed description of this). In particular, we may create an analogy between the general end-to-end setup with an optimization problem and the classification problem by thinking about each feasible point $x \in X$ as a class, which leads to a classification problem with a possibly uncountably infinite number of classes. This analogy highlights that the general end-to-end setting is much more complicated than classification. However, the optimization structure makes this general end-to-end setting amenable to analysis by providing us with a concrete mechanism for choosing between the classes (i.e., by solving (1)) as well as a toolkit (i.e., perturbation and sensitivity analysis) for developing our fundamental results.

Steinwart (2007) extends the relationship between classification prediction methods and the misclassification probability $\mathbb{P}[g(w) \neq c]$ to other learning-based settings, such as regression, density estimation, and density level detection. In particular, the work examines prediction tasks where various methods can be used in the selection of a prediction model, and develops a general theory to compare the different methods. Our results on optimization performance guarantees, namely, our Theorems 1 and 3, utilize some of this theory. In particular, Theorems 1 and 3 leverage the optimization structure in our end-to-end framework to establish conditions under which we can apply the theory of Steinwart (2007) (namely, Assumptions 1 and 2 together with Assumption EC.1 in the e-companion). In the context of our prediction and optimization setting, we further expand this line of work by outlining specific strategies to analyze the performance of prediction methods. For example, we show (in Theorem 2) that because of the optimization structure, the concept of calibration in Theorem 1 and Fisher consistency are equivalent. This then enables verification of optimization performance guarantees by simply checking Fisher consistency of a prediction method. See also Zhang (2004a, theorem 3) and Tewari and Bartlett (2007, theorem 2) for a special case of this result for the classification problem. In contrast to these results from classification, because of the much more complicated structure of the prediction and optimization setting, in the proof of Theorem 2 we leverage tools from perturbation analysis of optimization problems. Similarly, our Lemmas 3 and 4 utilize the optimization structure to provide a strategy to derive stronger nonasymptotic performance bounds via Theorem 3. We employ these results to analyze the least squares minimization prediction method, as well as the SPO+ method.

The end-to-end use of prediction methods within decision-making models has been examined by only 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 the newsvendor inventory control problem, Liyanage and Shanthikumar (2005) showed that, using the 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), Donti et al. (2017) and Elmachtoub and Grigas (2022) 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) studied the specialized setting where $X = \mathbb{R}^m$, f is a strongly convex quadratic function, and the prediction model g is restricted to be linear, and presented theoretical guarantees under a particular data distribution. Donti et al. (2017) proposed a scheme to directly differentiate the optimality gap, which gives rise to a stochastic gradient descent scheme for minimizing the expected optimality gap. This scheme was shown to demonstrate good numerical performance, but no theoretical guarantees are provided, and the differentiation scheme works only when the objective is smooth.

In a setting closest to ours, Elmachtoub and Grigas (2022) proposed a convex surrogate for the optimality gap, referred to as the SPO+ method. This gave rise to the only polynomial-time (thus far) prediction method that is tailored to the optimization problem in the end-to-end setting. They proved some useful theoretical properties of the SPO+ method, such as Fisher consistency under certain distributional assumptions, and demonstrated its good numerical performance in several settings, but did not give explicit relationships on its prediction performance and the subsequent optimization performance. In contrast to their work, the main goals of our paper are to identify properties of a *generic* prediction method that ensure good optimization performance and to quantify the optimization performance explicitly for *general* classes of prediction methods. Consequently, our results can be used to provide optimization performance guarantees for both optimization-agnostic methods and the SPO+ method. For example, we show that Fisher consistency provides asymptotic optimization performance guarantees for prediction methods; therefore, the Fisher consistency result of Elmachtoub and Grigas (2022) for the SPO+ method (which holds under certain distributional assumptions) enjoys these same guarantees.

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, we estimate the conditional distribution $\mathbb{P}[c \mid w]$ using a

kernel: $\mathbb{P}[\cdot \mid w] \approx \sum_{i\in[n]} k_{w_i}(w)\delta_{c_i}(\cdot)$, where $k_{w_i}(w)$ are convex combination weights that 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 (2020), Ban and Rudin (2019), Bertsimas and Van Parys (2021), and Srivastava et al. 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 class 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

Recall from Section 1 that we are given data $H_n := \{(w_i, c_i) : i \in [n]\}$, where each (w_i, c_i) is drawn from some unknown $\mathbb{P}$. Our end-to-end prediction and optimization framework is as follows: first, using the data H_n, we select a prediction model $g : W \to \mathbb{R}^m$; then, given $w \in W$, we make a prediction $g(w)$ for c and solve (1) with c replaced by $g(w)$. The central issue of this framework is that there is a discrepancy in how the prediction method selects a prediction model g, and how we assess its quality. We explain these two factors below.

Given (w, c) and the prediction vector $d = g(w)$, the quality of d is measured by the *true loss function*

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

where $x^*(d) \in \arg\min_{x\in X}\{f(x) + d^\top x\}$ is as defined in Section 1.2. This is simply the optimality gap for (1) of the solution $x^*(d)$ obtained with d instead of c. The overall quality of a prediction model in the end-to-end setting is the expectation of the true loss over an unseen data point $(w, c) \sim \mathbb{P}$, which we define as the *true risk*:

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

Ideally, we wish to obtain a prediction model for which $R(g, \mathbb{P})$ is small.

On the other hand, usually the selection of a prediction model g involves an alternative measure of the discrepancy between vectors d and c, called a *surrogate loss function* $\ell(d, c)$. Examples of surrogate loss functions include $\ell(d,c) = \|d - c\|$, where $\|\cdot\|$ is a norm on $\mathbb{R}^m$. Concretely, we consider prediction methods that select $g : W \to \mathbb{R}^m$ by minimizing the *empirical surrogate risk*

$$\hat{R}_\ell(g;H_n) := \frac{1}{n}\sum_{i\in[n]} \ell(g(w_i),c_i). \tag{4}$$

Choosing $\ell(d,c) = \|d - c\|_2^2$ results in the well-known least squares minimization method, whereas $\ell(d,c) = \|d - c\|_1$ is least absolute deviations. Under certain regularity conditions, minimizing $\hat{R}_\ell(g; H_n)$ gives guarantees on the *surrogate risk*

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

This is done via results from statistical learning theory (Bousquet et al. 2004), which uses the following notion of consistency.

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

$$R_\ell(\hat{g}_n,\mathbb{P}) \to \inf_g\{R_\ell(g,\mathbb{P}) : g \text{ measurable}\} \text{ in probability.}$$

(Convergence in probability is used because of the randomness in H_n, which translates to randomness of $\hat{g}_n$.) Whenever $\{\mathcal{G}_n\}_{n\in\mathbb{N}}$ and ℓ are mildly regular, the statistical consistency of the predictors $\hat{g}_n$ is guaranteed. Therefore, prediction methods that select g by minimizing (4) can provide guarantees on the surrogate risk $R_\ell(g, \mathbb{P})$ in (5), but it is not clear what impact this has on the true risk $R(g, \mathbb{P})$ in (3). A visual summary of our setup is given in Figure 1.

Instead of the traditional notion of statistical consistency in Definition 1, we will explore conditions on the surrogate loss function ℓ that ensure the following notion of consistency holds.

Definition 2. Let

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

$$R_\ell(\mathbb{P}) := \inf_g\{R_\ell(g,\mathbb{P}) : g \text{ measurable}\}. \tag{7}$$

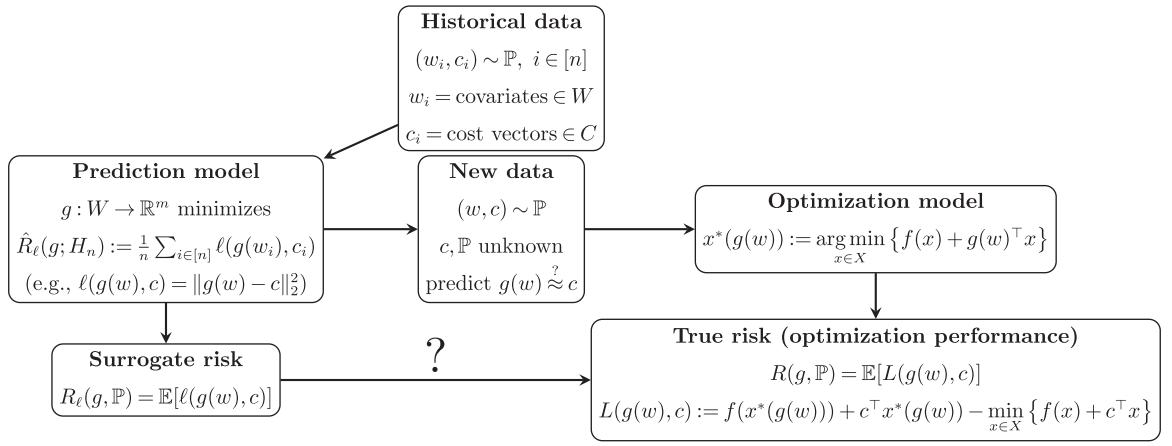
Given a class of distributions $\mathcal{P}$, we say that ℓ is $(\mathcal{P}, L)$-*consistent* if, for all $\mathbb{P} \in \mathcal{P}$, whenever we have a sequence of predictors $\{g_n\}_{n\in\mathbb{N}}$ such that $R_\ell(g_n, \mathbb{P}) \to R_\ell(\mathbb{P})$, we will also have $R(g_n, \mathbb{P})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

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.

If we are able to obtain g by minimizing (4) with $\ell = L$ directly, then the results are trivial, because L is of course $(\mathcal{P}, L)$-consistent. In practice, however, doing so may be intractable because $L(d, c)$ is not convex in d in general. We formalize this in Example 5. This is why we wish to consider numerically tractable surrogate loss functions ℓ instead, for example, convex loss functions.

Example 5. Suppose that $f(x) = 0$, X is a polytope with at least two extreme points, and $c \neq 0$ is such that $\min_{x\in X} c^\top x < \max_{x\in X} c^\top x$. Then, the loss function $L(d, c)$ is not convex in d. To see this, consider two extreme points of X, x_0 and x_1, with $c^\top x_0 = \max_{x\in X} c^\top x >$

Figure 1. The End-to-End Prediction and Optimization Framework



Note. The relationship between surrogate risk and true risk will be explored in this paper.

$c^\top x_1 = \min_{x\in X} c^\top x$. 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.

Remark 1. Note that Definition 2 does not depend on the data H_n, 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 $\hat{R}_\ell$ to the 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 $\{\mathcal{G}_n\}_{n\in\mathbb{N}}$ and H_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.

Remark 2. The true loss function L depends on the function x^*, that is, 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 fix x^* throughout the paper, which is equivalent to fixing the algorithm. Note, however, that the specific choice of x^* only affects our results up to measurability concerns; we show in Lemma EC.4 in the e-companion 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 in the e-companion. Henceforth, when measurability of functions is discussed, we will understand this to be in the sense of Lebesgue.

4. Risk Minimization via Fisher-Consistent Surrogate Loss Functions

As discussed in Section 3, we are interested in properties of the surrogate loss ℓ that ensure 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 $R_\ell(g_n, \mathbb{P}) - R_\ell(\mathbb{P}) \to 0$ 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 $R_\ell(g_n, \mathbb{P}) - R_\ell(\mathbb{P}) \le \delta$ 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:

$$\begin{aligned}&\text{for all } \epsilon > 0, \text{ there exists } \delta > 0 \text{ such that}\\ &\quad \text{if } R_\ell(g, \mathbb{P}) - R_\ell(\mathbb{P}) \le \delta, \text{ then } R(g, \mathbb{P}) - R(\mathbb{P}) \le \epsilon.\end{aligned} \tag{8}$$

We will show that (8) 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 including density level detection and density estimation. 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

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

Observe that Definition 3 is very similar to (8), except that predictors g (i.e., functions mapping onto

vectors) are replaced with vectors $d \in \mathbb{R}^m$, and that expectations are conditioned on w. This makes Definition 3 verifiable; that is, 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 $\arg\min_{d' \in \mathbb{R}^m} \mathbb{E}[\ell(d',c) \mid w]$ is nonempty and bounded, and $\mathbb{E}[\ell(d,c) \mid w] < \infty$. Furthermore, c is an integrable random vector (i.e., 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 if $R_\ell(g,\mathbb{P}) - R_\ell(\mathbb{P}) \le \delta$, then $R(g,\mathbb{P}) - R(\mathbb{P}) \le \epsilon$.*

We give the proof in Section EC.3 of the e-companion.

In general, checking that a given surrogate loss ℓ is $\mathbb{P}$-calibrated may not be straightforward. A potentially simpler condition to check is Fisher consistency, stated in Definition 4 below.

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,

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

Note that Definition 4 relates to the *minimizers* of the loss functions, instead of approximate minimizers as in Definition 3. Often, analyzing the minimizers of $\mathbb{E}[\ell(d,c) \mid w]$ and $\mathbb{E}[L(d,c) \mid w]$ can be easier than analyzing their approximate minimizers. For example, we have the following basic fact about (6), which is analogous to Elmachtoub and Grigas (2022, proposition 5) adapted to our setup, showing that the conditional expectation is a minimizer of (6).

Lemma 1. *The function $g^*(w) := \mathbb{E}[c \mid w]$ minimizes (6). Furthermore,*

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

Remark 3. Because the objective for our optimization problem is of the form $f(x) + c^\top x$, we proved in Lemma 1 that $\mathbb{E}[c \mid w] \in \arg\min_d \mathbb{E}[L(d,c) \mid w]$. Thus, one way to check that ℓ is Fisher consistent is to verify that $\arg\min_d \mathbb{E}[\ell(d,c) \mid w] = \{\mathbb{E}[c \mid w]\}$ (and this is the approach taken in some of the examples below). We opt not to simply take $\arg\min_d \mathbb{E}[\ell(d,c) \mid w] = \{\mathbb{E}[c \mid w]\}$ as the definition of Fisher consistency because we recognize, particularly for nonsmooth optimization objectives, that there can be other vectors besides $\mathbb{E}[c \mid w]$ that minimize $\mathbb{E}[L(d,c) \mid w]$. 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).

Although analyzing Fisher consistency may be conceptually simpler than calibration, it turns out that, theoretically, the two are *equivalent*. Importantly, we may then use Fisher consistency to verify Theorem 1.

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 fact that calibration implies Fisher consistency is straightforward; the main challenge is to show the other direction. The optimization structure is what allows us to do this. More precisely, we exploit upper semicontinuity of the multivalued argmin mapping $X^*(\cdot)$ (see Section EC.1 of the e-companion), which holds under mild regularity conditions. 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 $\mathbb{E}[L(d,c) \mid w]$ is large, then vectors close by to d will also have large true expected loss. The full proof of Theorem 2 is in Section EC.4 of the e-companion.

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 if $R_\ell(g,\mathbb{P}) - R_\ell(\mathbb{P}) \le \delta$, then $R(g,\mathbb{P}) - : R(\mathbb{P}) \le \epsilon$.*

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 $R_\ell(g_n,\mathbb{P}) \to R_\ell(\mathbb{P})$, then $R(g_n,\mathbb{P}) \to R(\mathbb{P})$.*

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 (8) through Theorem 2. These are as follows:

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 $\mathbb{E}[\ell(d,c) \mid w]$ is finite.
3. For any $w \in W$, the set of minimizers $\arg\min_{d' \in \mathbb{R}^m} \mathbb{E}[\ell(d',c) \mid w]$ is nonempty 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 6. 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

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

Thus, the unique minimizer of $\mathbb{E}[\ell_{\text{LS}}(d,c) \mid w]$ is $d^* = \mathbb{E}[c \mid w]$. Because we know this is also a minimizer of $\mathbb{E}[L(d,c) \mid w]$, 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 $\mathbb{P}[\cdot \mid w]$ 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 7. 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, $\mathbb{P}[\cdot \mid w]$ 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 $\arg\min_{d' \in \mathbb{R}^m} \mathbb{E}[\|d'-c\|_1 \mid w]$ recovers the vector of coordinate-wise medians, which, for a centrally symmetric distribution, will be the point of symmetry d_w; that is, d_w minimizes $\mathbb{E}[\|d-c\|_1 \mid w]$. Furthermore, we have $\mathbb{E}[c \mid w] = d_w$ also. Therefore, d_w minimizes $\mathbb{E}[L(d,c) \mid w]$.

We now discuss the SPO+ loss function proposed in Elmachtoub and Grigas (2022), 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 8. In the setting when $f(x) = 0$ for all $x \in X$, Elmachtoub and Grigas (2022, definition 3) defined the following loss function:

$$\ell_{\text{SPO+}}(d,c) := (2d-c)^\top x^*(c) - \min_{x \in X} (2d-c)^\top x = L(c, 2d-c). \tag{9}$$

Elmachtoub and Grigas (2022, proposition 6) shows that $\ell_{\text{SPO+}}$ is Fisher consistent as long as $\mathbb{P}[c \mid w]$ is centrally symmetric and continuous. Note also that Elmachtoub and Grigas (2022) presents 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 (2022). We start with an example below to review an important observation made in Elmachtoub and Grigas (2022, 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 $\ell_{\text{SPO+}}$ (9) also has a familiar interpretation and admits Fisher consistency in this setting.

Example 9. 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 (2022, proposition 4) shows that the loss from (9) reduces to the hinge loss in this case: because $x^*(c) = -c/2$ 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 $\ell_{\text{SPO+}}$ of Elmachtoub and Grigas (2022) is not Fisher consistent in some very natural settings.

Example 10. 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 $x^*(c) = -\text{sign}(c)/2$, $\min_{x \in X} d^\top x = -|d|/2$, and hence the loss function from (9) 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 $\mathbb{E}[L(d,c) \mid w]$ are $D_w^* = \{d \in \mathbb{R} : \text{sign}(d) = \text{sign}(\mathbb{E}[c \mid w])\}$. Thus, checking $\mathbb{P}$-Fisher consistency requires showing that $\arg\min_{d' \in \mathbb{R}} \mathbb{E}[\ell_{\text{SPO+}}(d', c) \mid w] \subseteq D_w^*$ for every $w \in W$; that is, we need to show that the minimizers have the same sign as the mean $\mathbb{E}[c \mid w]$. However, we show (in Section EC.4 of the e-companion) 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 11. In Example 9, we examined binary classification and showed that for appropriately chosen X, f, and C, L specializes to the 0-1 loss and $\ell_{\text{SPO+}}$ specializes to the hinge loss. Thus, $\ell_{\text{SPO+}}$ defined in (9) 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; that is, 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 (9) 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, that is, $C' = [m]$. We want a predictor $g' : W \to C'$ that 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 one; otherwise, our loss is zero. 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 $g : W \to \mathbb{R}^m$, we can use it to build a classifier $g' : W \to C'$ 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 that minimize $\mathbb{E}[L(d, c)]$ must satisfy $j^*(d) \in \text{argmax}_{j' \in [m]} p_{j'}$.

The loss (9) becomes

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

In Section EC.4 of the e-companion, we show that for distributions $\mathbb{P}$ with $\max_{j' \in [m]} p_{j'} < 1/2$, ℓ is not $\mathbb{P}$-Fisher consistent, because the 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 $j \in [m]$, that is, the highest p_j.

5. Nonasymptotic 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, because 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 nonasymptotic 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 nonasymptotic 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 special case of the $\ell_{\text{SPO+}}$ loss function (9) of Elmachtoub and Grigas (2022) and provide positive and negative results on its risk guarantees.

5.1. Outline of the 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 $\eta : [0, \infty) \to [0, \infty)$ with $\eta(0) = 0$ such that for any distribution $\mathbb{P}$, we have

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

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

$$\begin{aligned} &\eta\left(\mathbb{E}[L(d, c) \mid w] - \min_{d' \in \mathbb{R}^m} \mathbb{E}[L(d', c) \mid w]\right) \\ &\quad \leq \mathbb{E}[\ell(d, c) \mid w] - \min_{d' \in \mathbb{R}^m} \mathbb{E}[\ell(d', c) \mid w]. \end{aligned} \tag{10}$$

Then, we have

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

where the first inequality follows from Jensen's inequality, and the second follows from (10).

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

$$\delta_\ell(\epsilon, w; \mathbb{P}) := \inf_{d \in \mathbb{R}^m} \Big\{ \mathbb{E}[\ell(d,c) \mid w] - \min_{d' \in \mathbb{R}^m} \mathbb{E}[\ell(d',c) \mid w] : \mathbb{E}[L(d,c) \mid w] - \min_{d' \in \mathbb{R}^m} \mathbb{E}[L(d',c) \mid w] > \epsilon \Big\}. \tag{11}$$

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 $\mathbb{E}[L(d,c) \mid w] - \min_{d' \in \mathbb{R}^m} \mathbb{E}[L(d',c) \mid w] > \epsilon$, we have $\mathbb{E}[\ell(d,c) \mid w] - \min_{d' \in \mathbb{R}^m} \mathbb{E}[\ell(d',c) \mid w] > \delta$. Taking the infimum of $\mathbb{E}[\ell(d,c) \mid w] - \min_{d' \in \mathbb{R}^m} \mathbb{E}[\ell(d',c) \mid w]$ over d such that $\mathbb{E}[L(d,c) \mid w] - \min_{d' \in \mathbb{R}^m} \mathbb{E}[L(d',c) \mid w] > \epsilon$ gives exactly $\delta_\ell(\epsilon, w; \mathbb{P})$ as defined in (11), and we know that $\mathbb{E}[\ell(d,c) \mid w] - \min_{d' \in \mathbb{R}^m} \mathbb{E}[\ell(d',c) \mid w] > \delta$ 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 $\mathbb{E}[L(d,c) \mid w] - \min_{d' \in \mathbb{R}^m} \mathbb{E}[L(d',c) \mid w] = \epsilon$. Then

$$\mathbb{E}[\ell(d,c) \mid w] - \min_{d' \in \mathbb{R}^m} \mathbb{E}[\ell(d',c) \mid w] \geq \delta_\ell(\epsilon, w; \mathbb{P}) = \delta_\ell\Big(\mathbb{E}[L(d,c) \mid w] - \min_{d' \in \mathbb{R}^m} \mathbb{E}[L(d',c) \mid w], w; \mathbb{P}\Big).$$

This relation then inspires us to select $\eta = \delta_\ell(\cdot, w; \mathbb{P})$ for a given $\mathbb{P}$. However, there needs to be a *single* fixed η for which (10) holds for all $w \in W$. Therefore, we will instead take $\eta(\cdot; \mathbb{P}) = \bar{\delta}_\ell(\cdot; \mathbb{P}) := \inf_{w \in W} \delta_\ell(\cdot, w; \mathbb{P})$; that is, we need to strengthen the notion of calibration.

The choice of $\eta = \bar{\delta}_\ell$ may not be convex in general. Instead, we can use $\eta = \bar{\delta}_\ell^{**}$, where, given a function $h : \mathbb{R} \to \mathbb{R} \cup \{\infty\}$,

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

Clearly, h^{**} is convex because 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 $\bar{\delta}_\ell$ is only defined for $\epsilon > 0$, so we define $\bar{\delta}_\ell(\epsilon; \mathbb{P}) = 0$ when $\epsilon = 0$ and $\bar{\delta}_\ell(\epsilon; \mathbb{P}) = +\infty$ when $\epsilon < 0$. Using $\eta = \bar{\delta}_\ell^{**}$ guarantees both convexity of η and also that $\eta(\epsilon; \mathbb{P}) \leq \bar{\delta}_\ell(\epsilon; \mathbb{P})$; hence, the desired inequality (10) holds. Now, by the definition (11), we have that $\bar{\delta}_\ell$ is nondecreasing in ϵ and positive for $\mathbb{P}$-calibrated ℓ. However, ℓ could be such that $\bar{\delta}_\ell(\epsilon; \mathbb{P})$ does not increase once ϵ is sufficiently large, or only increases at a sublinear rate; in this case $\eta = \bar{\delta}_\ell^{**}$ is going to be zero for $\epsilon \geq 0$, so the inequality (10) will be useless. To prevent this, we make the assumption that $\mathbb{E}[L(d,c) \mid w] - \min_{d' \in \mathbb{R}^m} \mathbb{E}[L(d',c) \mid w] \leq B$ for all $w \in W$, $d \in \mathbb{R}^m$. We can then redefine $\bar{\delta}_\ell(\epsilon; \mathbb{P}) = \infty$ for $\epsilon > B$, and take $\eta = \bar{\delta}_\ell^{**}$. This ensures that $\eta(\epsilon) > 0$ for $\epsilon \in (0, B]$. To ensure that such a B exists, we define the following quantities:

$$B_X := \max_{x, x' \in X} \|x - x'\|_2, \quad B_f := \max_{x, x' \in X} \{f(x) - f(x')\}, \quad B_C := \max_{c \in C} \|c\|_2. \tag{12}$$

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

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

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

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

where the first inequality follows from the Cauchy-Schwarz inequality.

In summary, the additions we need to make to the assumptions from Section 4 are a stronger definition of calibration (which gives $\bar{\delta}_\ell(\epsilon; \mathbb{P}) > 0$; see Remark 6 in the next section) and Assumption 2. Notice, however, that because our proof technique is different from 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

Recall from the discussion of the previous section that in order to obtain nonasymptotic risk bounds, we need to include some stronger assumptions to the ones from Section 4. In this section, we formalize in Definition 5 below the strengthened notion of calibration that is required to obtain such bounds. We then formally present the risk bound obtainable under this definition and outline a strategy to check this definition in practice.

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

$$\mathbb{E}[\ell(d,c) \mid w] - \inf_{d'} \mathbb{E}[\ell(d',c) \mid w] < \delta \Rightarrow \mathbb{E}[L(d,c)] - \inf_{d'} \mathbb{E}[L(d',c)] < \epsilon. \tag{13}$$

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 $\mathcal{P}$.

For a given $\mathcal{P}$, we define

$$\delta_\ell(\epsilon;\mathcal{P}) := \inf_{\substack{d\in\mathbb{R}^m \\ w\in W \\ \mathbb{P}\in\mathcal{P}}} \Big\{ \mathbb{E}[\ell(d,c)\,|\,w] - \min_{d'\in\mathbb{R}^m} \mathbb{E}[\ell(d',c)\,|\,w] : \mathbb{E}[L(d,c)\,|\,w] - \min_{d'\in\mathbb{R}^m} \mathbb{E}[L(d',c)\,|\,w] \geq \epsilon \Big\}. \quad (14)$$

Remark 6. If ℓ is $\mathcal{P}$-calibrated, then $\delta(\epsilon;\mathcal{P}) > 0$ for all $\epsilon > 0$ by taking the contrapositive of (13), and is nondecreasing in ϵ. In addition, if Assumption 2 holds, then $\delta_\ell(\epsilon;\mathcal{P}) = \infty$ for $\epsilon > B_f + B_C B_X$ because 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 proof similar to that of Lemma EC.5 in the e-companion.

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

Lemma 2. *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 nontrivial, that is, verify that δ_ℓ^{**} is positive on its domain. We utilize Lemma 2 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 as defined in (12). Then $\delta_\ell^{**}(\epsilon;\mathcal{P})$ is positive for $\epsilon \in (0, B_f + B_C B_X]$, and for any $\mathbb{P}\in\mathcal{P}$, $g : W \to \mathbb{R}^m$,

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

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 (14). 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 Elmachtoub and Grigas (2022), 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 3. *Consider δ_ℓ defined in (14). We have*

$$\delta_\ell(\epsilon;\mathcal{P}) = \inf_{x,x'\in X} \inf_{\substack{d:x^*(d)=x \\ \bar{c}:x^*(\bar{c})=x'}} \inf_{\substack{\mathbb{P}\in\mathcal{P} \\ w\in W \\ \mathbb{E}[c|w]=\bar{c}}} \Big\{ \mathbb{E}[\ell(d,c)\,|\,w] - \min_{d'\in\mathbb{R}^m} \mathbb{E}[\ell(d',c)\,|\,w] : f(x) - f(x') + \bar{c}^\top(x-x') \geq \epsilon \Big\}. \quad (15)$$

We now give a bound on the distance between d and $\bar{c}$ in the second infimum in (15).

Lemma 4. *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 4, 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 (12). 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:

$$\ell_{\text{LS}}(d,c) := \|d - c\|_2^2,$$

$$\mathcal{P} := \{ \mathbb{P} : \forall w \in W, \; \mathbb{P}[\cdot\,|\,w] \text{ is square integrable, and } \mathbb{E}[c\,|\,w] \in \text{Conv}(C) \}.$$

Because of 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 5. *Consider the case of the squared loss ℓ_{LS} and $\mathcal{P}$ as defined above. Then, we have*

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

Using Lemmas 4 and 5, 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 $g : W \to \mathbb{R}^m$ 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 counterintuitive. Furthermore, let $g = (g_1,\dots,g_m)$, where each $g_j : W \to \mathbb{R}$, 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 $g_j : W \to \mathbb{R}$ 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 counterintuitive, because 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, because of their superior finite-sample performance. For example, the absolute deviation loss from Example 7 and the Huber loss have been used for heavy-tailed data because of 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 6. *For each $j \in [m]$, let $\ell_j : \mathbb{R}\times\mathbb{R} \to \mathbb{R}$ be a loss function such that there exists a nondecreasing function $\eta_j : (0,\infty) \to (0,\infty)$ that satisfies*

$$\mathbb{E}[(g_j(w)-c_j)^2] - \inf_{g_j'} \mathbb{E}[(g_j'(w)-c_j)^2] \leq \eta_j\left(\mathbb{E}[\ell_j(g_j(w),c_j)] - \inf_{g_j'} \mathbb{E}[\ell_j(g_j(w),c_j)]\right) \tag{16}$$

for any $g_j : W \to \mathbb{R}$ 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 $\mathcal{P}_{\text{sym}}$ to be the class of square integrable distributions such that for all $w \in W$ and $j \in [m]$, $\mathbb{P}[c_j \mid w]$ is a symmetric distribution; that is, $c_j - \mathbb{E}[c_j \mid w]$ has the same conditional distribution as $\mathbb{E}[c_j \mid w] - c_j$.

Theorem 5 (Steinwart (2007, theorems 4.19 and 4.20 (ii))). *Fix any $j \in [m]$. Let $\ell_j(d_j,c_j) = \psi_j(d_j - c_j)$, where $\psi_j : \mathbb{R} \to [0,\infty)$ is symmetric, that is, $\psi_j(r) = \psi_j(-r)$, and uniformly convex, that is, there exists some nondecreasing $\delta_j : [0,\infty) \to [0,\infty)$ 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 $g_j : W \to \mathbb{R}$ and $\mathbb{P} \in \mathcal{P}_{\text{sym}}$, we have

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

Although the proof of Theorem 5 can be found in the relevant sections of Steinwart (2007), we give a more concise version in Section EC.5 of the e-companion.

5.4. Uniform Calibration of the SPO+ Loss in Example 10

Recall our Example 8, which studied the SPO+ loss (9) introduced in Elmachtoub and Grigas (2022, definition 3). It was shown in Elmachtoub and Grigas (2022, theorem 1) that this loss is Fisher consistent; hence, by Theorem 2, it is $\mathbb{P}$-calibrated whenever $\mathbb{P}[c \mid w]$ is centrally symmetric and continuous for all $w \in W$. On the other hand, the uniform calibration of the SPO+ loss (9) has

not yet been studied. In this section, we examine its uniform calibration for the special one-dimensional case $m = 1$, that is, Example 10. To our knowledge, the general m case remains open.

Recall Example 10 has $m = 1$, $f = 0$, and $X = [-1/2, 1/2]$, and we will take $C = \mathbb{R}$. In this case, recall that the loss function (9) becomes

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

For this loss function, Elmachtoub and Grigas (2022) 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 (with respect to Lebesgue measure) is positive over all of $\mathbb{R}$. For simplicity, we consider the same class of symmetric, continuous distributions over $\mathbb{R}$, that is,

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

For this class of distributions, in Elmachtoub and Grigas (2022), it was shown that the (conditional) mean is the unique minimizer of $\min_{d' \in \mathbb{R}} \mathbb{E}[\ell(d',c) \mid w]$.

Lemma 7 (Elmachtoub and Grigas (2022, theorem 1)). *Let $\mathbb{P} \in \mathcal{P}_{\text{cont,sym}}$. Then for any $w \in W$, the unique minimizer of $\min_{d' \in \mathbb{R}} \mathbb{E}[\ell_{\text{SPO+}}(d',c) \mid w]$ is $d^* = \mathbb{E}[c \mid w]$.*

Using Lemmas 3 and 7, and noting

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

we have

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

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

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

where the third equality follows because $\mathbb{P}[c \mid w]$ is symmetric, so $2\bar{c} - c$ has the same conditional distribution as c; thus, $\mathbb{E}[|2\bar{c} - c| \mid w] = \mathbb{E}[|c| \mid w]$. 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 $\mathbb{E}^{(k)}[c](\mathbb{P}^{(k)}[c > 0] - \mathbb{P}^{(k)}[c < 0]) \to 0$. Therefore, by Lemma 2, $\ell_{\text{SPO+}}$ 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 $|\mathbb{P}[c > 0 \mid w] - \mathbb{P}[c < 0 \mid w]|$.

Proposition 2. *For $\alpha > 0$, let*

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

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

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

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 information from the optimization problem. 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, for example, 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 (2022) proposed the SPO+ loss function $\ell_{\text{SPO+}}$ (9), which incorporates information about the optimization problem and is known to be Fisher consistent with respect to certain distributions (see Example 8), 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:

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

Note that Elmachtoub and Grigas (2022, 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 nonlinearity 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 11), but the squared loss is consistent.

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

$$\min_{V\in\mathbb{R}^{m\times k}} \frac{1}{n}\sum_{i\in[n]} \ell(Vw_i, c_i) \tag{17}$$

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 the least squares loss and the more high-powered SPO+ loss, which takes into account optimization problem information. This highlights that on real data, a "naïve" method such as using the least squares loss may actually perform decently in comparison with a high-powered method that utilizes optimization structure. This parallels our theoretical insight from Propositions 3 and 4, which is that there is essentially *no* difference between using the least squares loss, using the SPO+ loss, and using the optimality gap L on portfolio optimization problems without the nonnegativity constraints.
- Recall that we have established theoretically that the SPO+ loss is provably inconsistent for the multiclass classification problem (Example 11). Moreover, our numerical results on these problem instances show that the performance of SPO+ loss (expectedly) deteriorates. This highlights an important insight: consistency of a prediction method matters as much as (if not more than) whether the prediction method takes into account explicit information from the optimization problem.
- On the fractional knapsack instances, we observe that constructing a close approximation of the true loss by regularization, although conceptually reasonable, does not provide good empirical results because of the considerable increase in computational effort required to find the corresponding estimator. Therefore, these results illustrate that 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:

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

and $Q\succ 0$ is a 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, $\mathbb{E}[c\mid\tilde{w}]=\tilde{b}+\tilde{V}\tilde{w}$, where $\tilde{w}$ are market factors (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 one 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 hypothesized prediction model is $\mathbb{E}[c\mid w]=Vw$, and we aim to estimate V. We do this by again minimizing (17), 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

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

Usually, in portfolio optimization, we are permitted to have entries of x that are negative, which means we short sell some assets. We show that if we redefine the domain to be $X:=\{x\in\mathbb{R}^m : \ p^\top x=b\}$, which does not have the nonnegativity 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* $\min_{x\in X}\{\frac{1}{2}x^\top Qx-d^\top x\}$ *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,

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

Moreover, we next show that when we consider linear predictors $w \mapsto Vw$, 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*

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

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 of the e-companion.

Propositions 3 and 4 imply that without nonnegativity constraints $x \geq 0$ in the definition of the domain X in the mean-variance portfolio instances, there is essentially no difference between using ℓ_{LS} and $\ell_{\text{SPO+}}$. For this reason, in our numerical study, we henceforth impose nonnegativity 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 1,258 trading days between January 1, 2003, and December 31, 2007. We also collected data on the three Fama–French factors for these trading days; these are our feature vectors, with a one appended, so $k = 4$.

We consider $m \in \{10, 15, \ldots, 30\}$, and for each m, we generate 100 random instances by choosing m random stocks. For each instance, we collect $n \in \{100, \ldots, 500\}$ 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 $x^*(Vw)$, and 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

$$\max_{x\in X} d^\top x, \text{ where } X := \{x \in [0,1]^m : p^\top x \leq B\}, \text{ and } f(x) = 0. \tag{18}$$

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 because of 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

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

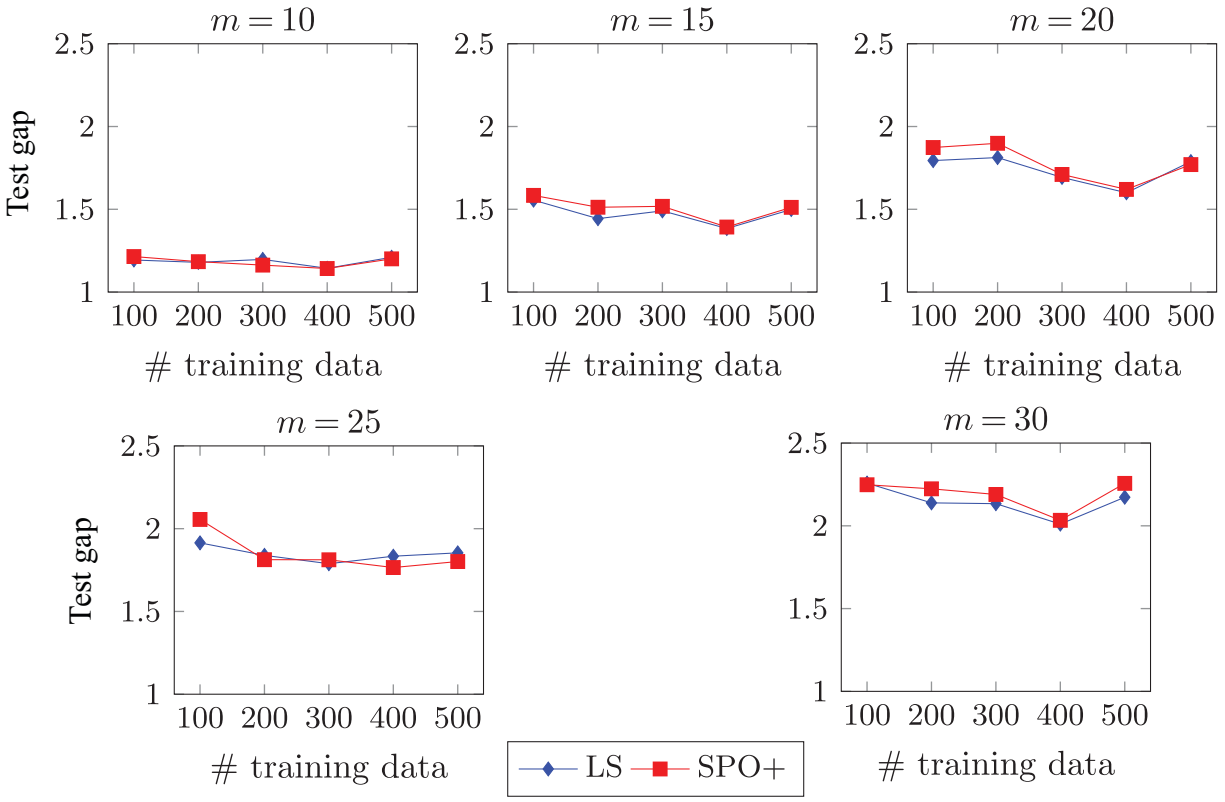
Note that the loss function $\ell_{\text{reg},\lambda}$ is nothing but the *exact* optimality gap evaluated at the *unique* solution to the regularized knapsack problem with the objective function $d^\top x - \frac{\lambda}{2}\|x\|_2^2$ that includes a regularization term. We consider the regularized objective because 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 (17) 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 of the e-companion.

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 1,000. 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

Figure 2. (Color online) Median Test Optimality Gaps for Different m for Portfolio Optimization



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 scheme similar to that of Elmachtoub and Grigas (2022). 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 hyperparameters $\delta \geq 1, \epsilon \in (0,1)$, 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 data sets of size $n = 100,200,300,400,500$ generated in this way. We trained $\ell_{\text{reg},\lambda}$ with $\lambda = 0.01$.

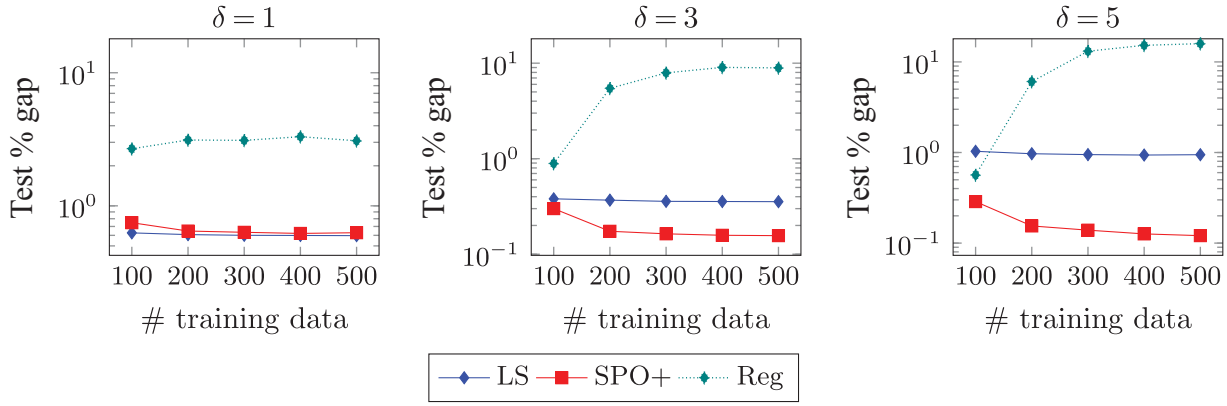
To test our predictors, we generate 10,000 points from the same distribution for each hyperparameter setting and V_0, and evaluate the average optimality gap using L on the test set for our predictors. Our results are shown in Figure 3, where we measure the average percentage optimality gap $\tilde{L}(d,c) = c^\top(x^*(c) - x^*(d))/(c^\top x^*(c))$ (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 nonlinear), $\ell_{\text{SPO+}}$ 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 11 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

Figure 3. (Color online) Average Test Relative Optimality Gaps for Different δ and $\epsilon = 0.1$ for the Continuous Knapsack Problem



the jth 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*

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

Recall that in Example 11, we showed SPO+ to be inconsistent for this problem theoretically. We next numerically compare the performance of ℓ_{LS} with that of $\ell_{\text{SPO+}}$ 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 $v_1, \ldots, v_m \in \mathbb{R}^k$:

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

Under this model, given w, choosing the most likely class is equivalent to choosing the index j that 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 $\{p_j(w_i^{\text{test}}) := \mathbb{P}[c = c_j \mid w_i^{\text{test}}]\}_{j \in [m], i \in [N]}$ using the true parameters V^{true}.
- For each $n \in \{100, 200, \ldots, 1{,}000\}$,
  — 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 $\mathbb{E}[L(\hat{V} w_i^{\text{test}}, c) \mid w_i^{\text{test}}]$, the expected true loss of plugging in the vector $\hat{V} w_i^{\text{test}}$ into the optimization problem, and if it has a nonunique 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, that is, the Bayes loss, as

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

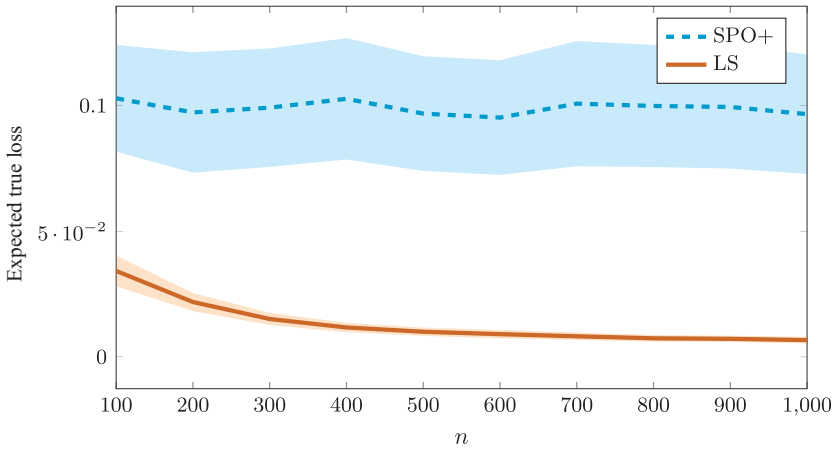
In Figure 4, we plot the mean and a two-standard-deviation band for the gap of the true losses for each predictor relative to the Bayes loss across 100 runs; that is, we plot statistics for the following quantity:

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

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 because we established that the 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

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

Figure 4. (Color online) Means (Lines) and Two-Standard-Deviation Ranges (Shaded Regions) of Expected True Loss Across 100 Runs for Multiclass Classification with $m = k = 4$



is a highly nonlinear function of w, and restricting it to a linear model, such as the case of $V_{\mathrm{LS}}w$, may prevent us from learning the true functional form of $\mathbb{E}[c \mid w]$. 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 $V_{\mathrm{LS}}w$ may not exactly be $\mathbb{E}[c \mid w]$, the minimal entry may still coincide. On the other hand, we show in the proof of Example 11 in Section EC.4 of the e-companion that the true minimizer of $\mathbb{E}[\ell_{\mathrm{SPO+}}(d,c) \mid w]$ is a constant vector, which we know in general will not give us the correct minimal entry of $\mathbb{E}[c \mid w]$. 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 $V_{\mathrm{SPO+}}$ reveals that it often estimates a zero matrix, which predicts the zero vector $V_{\mathrm{SPO+}}w = \mathbf{0}$. This is consistent with the theoretical analysis of Example 11 in Section EC.4 of the e-companion, where it is shows that constant vectors d minimize $\mathbb{E}[\ell_{\mathrm{SPO+}}(d,c)]$ 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 risk defined via the optimality gap can be minimized via the surrogate risk, as long as the surrogate loss function is appropriately calibrated. We provided precise relationships between the two risks under these assumptions. An equivalence result (Theorem 2) allows us to easily check the weaker $\mathbb{P}$-calibration condition via Fisher consistency, which we used 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 $\ell_{\mathrm{SPO+}}$ 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 its performance.

That said, our results in Section 6.2 reaffirm earlier findings from the literature that the SPO+ loss performs well under prediction model misspecification, for example, 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 of the performance guarantees in the face of prediction or optimization model misspecification. Optimization model misspecification can occur, for example, in mean-variance optimization problems when the precise risk trade-off is not decided a priori. Our findings from Sections 6.2 and 6.3 call for the design of new loss functions 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 $\ell_{\mathrm{SPO+}}$, and investigating calibration of objective functions $f(x, c)$ depending nonlinearly on c.

Acknowledgments

The authors thank the review team for their careful reading and suggestions, which significantly improved this paper.

References

Ban GY, Rudin C (2019) The big data newsvendor: Practical insights from machine learning. *Oper. Res.* 67(1):90–108.

Bartlett PL, Jordan MI, McAuliffe JD (2006) Convexity, classification, and risk bounds. *J. Amer. Statist. Assoc.* 101(473):138–156.

Bengio Y (1997) Using a financial training criterion rather than a prediction criterion. *Internat. J. Neural Systems* 8(04):433–443.

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

Bertsimas D, Van Parys B (2021) Bootstrap robust prescriptive analytics. *Math. Programming*, ePub ahead of print June 25, https://doi.org/10.1007/s10107-021-01679-2.

Bousquet O, Boucheron S, Lugosi G (2004) Introduction to statistical learning theory. Bousquet O, von Luxburg U, Rätsch G, eds. *Advanced Lectures on Machine Learning*, Lecture Notes in Computer Science, vol. 3176 (Springer, Berlin), 169–207.

Donti P, Amos B, Kolter JZ (2017) Task-based end-to-end model learning in stochastic optimization. Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, Garnett R, eds. *Advances in Neural Information Processing Systems*, vol. 30 (Curran Associates, Red Hook, NY), 5484–5494.

Duchi J, Khosravi K, Ruan F (2018) Multiclass classification, information, divergence and surrogate risk. *Annals Statist.* 46(6B): 3246–3275.

Elmachtoub AN, Grigas P (2022) Smart "predict, then optimize." *Management Sci.* 68(1):9–26.

Fama EF, French KR (1992) The cross-section of expected stock returns. *J. Finance* 47(2):427–465.

Goh CY, Jaillet P (2016) Structured prediction by conditional risk minimization. Technical report, https://arxiv.org/abs/1611.07096.

Hanasusanto GA, Kuhn D (2013) Robust data-driven dynamic programming. Burges CJC, Bottou L, Welling M, Ghahramani Z, Weinberger KQ, eds. *Advances in Neural Information Processing Systems*, vol. 26 (Curran Associates, Red Hook, NY), 827–835.

Hannah L, Powell W, Blei DM (2010) Nonparametric density estimation for stochastic optimization with an observable state variable. Lafferty JD, Williams CKI, Shawe-Taylor J, Zemel RS, Culotta A, eds. *Advances in Neural Information Processing Systems*, vol. 23 (Curran Associates, Red Hook, NY), 820–828.

Kao Y, Roy BV, Yan X (2009) Directed regression. Bengio Y, Schuurmans D, Lafferty JD, Williams CKI, Culotta A, eds. *Advances in Neural Information Processing Systems*, vol. 22 (Curran Associates, Red Hook, NY), 889–897.

Lin Y (2004) A note on margin-based loss functions in classification. *Statist. Probab. Lett.* 68(1):73–82.

Liyanage LH, Shanthikumar JG (2005) A practical inventory control policy using operational statistics. *Oper. Res. Lett.* 33(4):341–348.

Osokin A, Bach F, Lacoste-Julien S (2017) On structured prediction theory with calibrated convex surrogate losses. Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, Garnett R, eds. *Advances in Neural Information Processing Systems*, vol. 30 (Curran Associates, Red Hook, NY), 302–313.

Srivastava PR, Wang Y, Hanasusanto GA, Ho CP (2019) On data-driven prescriptive analytics with side information: A regularized {N}adaraya-{W}atson approach. Technical report, http://www.optimization-online.org/DB_HTML/2019/01/7043.html.

Steinwart I (2002a) On the influence of the kernel on the consistency of support vector machines. *J. Machine Learn. Res.* 2:67–93.

Steinwart I (2002b) Support vector machines are universally consistent. *J. Complexity* 18(3):768–791.

Steinwart I (2005) Consistency of support vector machines and other regularized kernel classifiers. *IEEE Trans. Inform. Theory* 51(1):128–142.

Steinwart I (2007) How to compare different loss functions and their risks. *Constructive Approximation* 26(2):225–287.

Tewari A, Bartlett PL (2007) On the consistency of multiclass classification methods. *J. Machine Learn. Res.* 8(36):1007–1025.

Zhang T (2004a) Statistical analysis of some multi-category large margin classification methods. *J. Machine Learn. Res.* 5(October): 1225–1251.

Zhang T (2004b) Statistical behavior and consistency of classification methods based on convex risk minimization. *Ann. Statist.* 32(1): 56–85.