Integrated Conditional Estimation-Optimization

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

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

Max Shen
Department of Industrial Engineering and Operations Research, UC Berkeley, Berkeley, CA, 94720,
Faculty of Engineering & Faculty of Business and Economics, University of Hong Kong, China,
maxshen@berkeley.edu

Many real-world optimization problems involve uncertain parameters with probability distributions that can be estimated using contextual feature information. In contrast to the standard approach of first estimating the distribution of uncertain parameters and then optimizing the objective based on the estimation, we propose an *integrated conditional estimation-optimization* (ICEO) framework that estimates the underlying conditional distribution of the random parameter while considering the structure of the optimization problem. We directly model the relationship between the conditional distribution of the random parameter and the contextual features, and then estimate the probabilistic model with an objective that aligns with the downstream optimization problem. We show that our ICEO approach is asymptotically consistent under moderate regularity conditions and further provide finite performance guarantees in the form of generalization bounds. Computationally, performing estimation with the ICEO approach is a non-convex and often non-differentiable optimization problem. We propose a general methodology for approximating the potentially non-differentiable mapping from estimated conditional distribution to optimal decision by a differentiable function, which greatly improves the performance of gradient-based algorithms applied to the non-convex problem. We also provide a polynomial optimization solution approach in the semi-algebraic case. Numerical experiments are also conducted to show the empirical success of our approach in different situations including with limited data samples and model mismatches.

Key words: contextual stochastic optimization; prescriptive analytics; statistical learning theory

1. Introduction

Two fundamental aspects of decision-making under uncertainty are estimation and optimization. Classically these two aspects are treated separately, with statistical and/or machine learning methodologies used to estimate the distributions of uncertain parameters based on data, resulting in a stochastic optimization problem to be solved for making a decision. In recent years, researchers and practitioners have increasingly recognized the significance of considering estimation and optimization in tandem (Bertsimas and Kallus 2020, Kao et al. 2009, Donti et al. 2017, Elmachtoub

and Grigas 2021). Another salient feature of modern decision-making under uncertainty is the presence of *contextual* information, usually in the form of features/covariates, that can be leveraged to improve the estimation of the uncertain parameters. For example, contextual information such as temporal information, the presence of promotions, and economic indicators can be leveraged to refine the estimation of uncertain demand for products. The refined demand distribution estimates would then be used for making inventory and supply chain decisions through optimization models. *Contextual stochastic optimization (CSO)* has recently emerged as a general paradigm describing this situation, with applications in supply chain management, finance, transportation, energy systems, and many other areas.

In this work, we consider the CSO problem in a data-driven setting where one has available historical data consisting of realizations of the uncertain parameters paired with contextual feature information. As mentioned, the classical method of solving CSO given data is a two-step procedure, where in the first step either a point prediction of the parameter or an estimation of its distribution is built based on data. (Although the phrases "prediction" and "estimation" are often synonymous or not clearly distinguished in the literature, herein we specifically let "prediction" denote point predictions of the random parameter and let "estimation" refer to any methodology, either parametric or non-parametric, for estimating the conditional distribution of the random parameter given the context.) Modern machine learning techniques are often utilized in the first step to provide more granular results, and these models are usually fit based on statistical objectives such as measures of prediction error or likelihood. Then in the second step, given the prediction or estimation, an optimization problem is solved. A major drawback of these standard predict-then-optimize (PTO) and estimate-then-optimize (ETO) approaches is that they do not consider the decision error – the cost with respect to the downstream optimization problem due to an imperfect prediction – when fitting a statistical model.

We propose an integrated conditional estimation-optimization (ICEO) approach that estimates the underlying conditional distribution of the random parameter based on minimizing the ultimate decision error. We propose a highly flexible framework that models the conditional distribution using a hypothesis class and apply ideas from statistical learning to do estimation. As compared to existing approaches, our approach uses a generic learning framework based on specifying a hypothesis class and applies to a broad class of convex contextual stochastic optimization problems with uncertainty in the objective. Many previous approaches either rely heavily on the structure of the downstream problem, for example a linear (Elmachtoub and Grigas 2021) or newsvendor (Ban and Rudin 2019) problem, or propose a variation on a specific learning algorithm like random forests (Kallus and Mao 2020). In addition, related approaches based on "end-to-end learning" (see, e.g., Donti et al. (2017), Wilder et al. (2019b)) usually do not directly model the conditional distribution

of uncertain parameters as we do in the ICEO framework and lack strong theoretical guarantees. We further discuss the relationship between the ICEO framework and existing approaches in Section 1.1.

In addition to proposing the flexible ICEO framework that models the conditional distribution of uncertain parameters in a way that accounts for the downstream optimization cost, we consider the statistical and computational properties of our approach. In particular, we prove asymptotic consistency in terms of risks, decisions and hypotheses. Asymptotic consistency is highly desired for data-driven methods because it guarantees that, as the amount of data increases, our solutions and estimated models converge to their optimums given full information of the true distribution of contextual features and uncertain parameters. We prove asymptotic consistence of the ICEO risk and induced decisions only under the assumption that the hypothesis class is compact, and consistency of the hypothesis requires an additional assumption related to the uniqueness of the true hypothesis. We also provide generalization bounds to quantify the out-of-sample performance when data is limited to a finite sample. To induce desirable generalization properties, we introduce a strongly convex decision regularization function to stabilize the ICEO decision and to eliminate potential multiple optimal decisions. The strongly convex regularization guarantees the Lipschitz property of the regularized optimal solution mapping, and the resulting generalization bounds are constructed based on multi-variate Rademacher complexity. In terms of computation, the core training problem of the ICEO framework is non-convex and even non-differentiable in many cases. In fact, due to the presence of constraints in the downstream problem, it is often the case that the optimal decision oracle has a piece-wise constant shape, which leads to poor local minima that are very hard to escape when applying gradient-based methods. For these reasons, we propose two computational approaches: *(i)* a highly practical approach that involves approximating the regularized optimal solution oracle with a smooth function and then applying gradient algorithms, and *(ii)* a polynomial optimization approach when the downstream problem has a semi-algebraic objective and we approximate the optimal solution oracle with a polynomial function.

Our key contributions are summarized as follows:

1. We propose the ICEO framework, wherein we directly estimate the underlying conditional distribution of uncertain parameters given contextual information using a hypothesis class. In contrast to two-step ETO methods, we learn the conditional distribution in a way that integrates with the downstream optimization goal. ICEO offers more flexibility compared to most existing related approaches.
2. We prove asymptotic consistency of the ICEO method when the model i specified correctly (Theorem 1). More specifically, we show the consistency of ICEO risk, ICEO decisions, and

ICEO hypothesis when the hypothesis class contains the correct conditional distribution function. The consistency in risk and decisions hold for arbitrary compact hypothesis classes. To guarantee the consistency in hypothesis, we require an additional assumption related to the uniqueness of the true hypothesis.

3. To quantify the out-of-sample performance with finite samples, we provide generalization bounds for the ICEO method based on the multi-variate Rademacher complexity of the hypothesis class used to learn the conditional distribution (Theorem 3). The generalization bound is based on the Lipschitz property of the regularized optimal decision oracle (Proposition 1).
4. The ICEO training problem is non-convex and non-differentiable. Non-differentiability poses a serious concern when applying gradient-based algorithms, like (stochastic) gradient descent, to solve the ICEO training problem as the presence of constraints can lead to local minima that are hard to escape (visually illustrated in Figure 1). To address this issue, we approximate the oracle using differentiable function classes with a guaranteed approximation error (Proposition 2, Proposition 3). We then provide corresponding generalization bounds when training ICEO method using the approximated oracle (Theorem 4). In addition, for the case where the nominal optimization problem is semi-algebraic, we propose an exact solution algorithm (Proposition 4).

The remainder of this paper is organized as follows. In Section 1.1, we review related methods in literature. The details of our proposed ICEO framework are introduced in Section 2. In Section 3, we provide performance guarantees in terms of asymptotic consistency and generalization bounds. In Section 4, we discuss the main difficulties in solving the ICEO formulation and provide solution methods. Empirical performance of the ICEO method is demonstrated in Section 5.

1.1. Relevant Literature

The fusion of prediction models based on data and the optimization problems has become more and more widespread in recent years. In the remainder of this section, we will discuss existing works related to this topic and contrast them with our proposed ICEO approach.

The first stream of research focuses on providing a prescriptive solution by approximating the conditional distribution of the random parameter given a feature vector, with the help of various machine learning tools. Bertsimas and Kallus (2020) first proposed prescriptive models that approximate the conditional distribution with a weighted empirical distribution of the uncertainty. The weights can be achieved based on multiple machine learning models, including k-nearest neighbors (KNN), kernel methods, tree-based methods, etc. A later work Bertsimas and McCord (2019) investigates these prescriptive methods in the multi-period problem setting. Bertsimas et al. (2019)

follows the same idea and propose a tree-based algorithm that balances the optimality of the prescription and accuracy of the prediction. Kallus and Mao (2020) consider a random forest model for the prescriptive solution. In contrast to the standard way of splitting the feature space, the authors consider the down stream optimization quality while constructing the partitions. Ho and Hanasusanto (2019) considers the regularized Nadaraya-Watson approach and establish performance guarantees using moderate deviations theory.

Another stream of related work investigates adjusting the loss function to meet the ultimate optimization goal while training the machine learning models to predict the random parameters. Ban and Rudin (2019) investigates the Newsvendor problem, which is inherently equivalent to a quantile prediction problem. The authors learn the feature-to-decision mapping from data by adopting a loss function that characterizes the newsvendor inventory cost, and is equivalent to the quantile loss function. Following a similar setting, Qi et al. (2020a) discuss the performance guarantees of such an approach when there are inter-temporal dependencies and non-stationarities. Elmachtoub and Grigas (2021) consider the case when the downstream optimization problem has a linear objective. The authors propose a "smart predict-then-optimize" (SPO) framework with a tractable convex surrogate loss function (SPO+) to integrate the ultimate optimization problem structure. They prove Fisher consistency of SPO+ and demonstrate its strong numerical performance on different problem classes. Balghiti et al. (2019) later provide finite-sample performance guarantee of the SPO loss in the form of generalization bounds. Recently, Liu and Grigas (2021) have strengthened the consistency of SPO+ by providing risk guarantees and a calibration analysis in the polyhedral and strongly convex cases. Elmachtoub et al. (2020) propose a method to train decision trees using the SPO loss and demonstrate its excellent numerical performance and lower model complexity.

Other existing studies aim to learn the task-based end-to-end learning models with differentiable optimization layers. Donti et al. (2017) consider a general setting where the optimization stage involves a convex optimization problem and adopt the objective in the optimization stage as the loss function to achieve an end-to-end training for the machine learning models. The main issue in such end-to-end learning models is to address the non-differentiability of the optimal solution mapping (the mapping from a contextual feature vector to the optimal decision). Amos and Kolter (2017) introduce the differentiable optimization layers for the end-to-end training approaches and propose a method of approximating the gradient of the optimal solution mapping by the solution of a group of equations representing the KKT conditions. Agrawal et al. (2019) further provide a method to convert convex programs to the canonical forms that can be implemented at the optimization layer and implemented their grammar in CVXPY for ease of use. Wilder et al. (2019a) and Wilder et al. (2019b) further consider more difficult combinatorial problems. They propose end-to-end models

that map from the graph structure to a feasible solution and train them with the quality of the solution. Wilder et al. (2019a) consider continuous relaxations of the discrete problem to propagate gradients through the optimization procedure. Mandi and Guns (2020) consider mixed integer linear programs and consider a homogeneous self-dual formulation of the LP and show that the gradients are related to an interior point step. Berthet et al. (2020) instead consider stochastically perturbed optimizers to evaluate the gradients required for back-propagation. Mandi et al. (2020), Ferber et al. (2020), Pogančić et al. (2019) also discuss how to approximate the gradients when training end-to-end models for combinatorial problems. As our work focuses on convex optimization problems, we skip the details and refer to Kotary et al. (2021) for a detailed survey. Although demonstrated to be competitive in numerical experiments, these end-to-end learning models based on optimization layers and their extensions to combinatorial cases lack strong performance guarantees in theory. Moreover, learning the feature-to-decision mapping lacks flexibility in the way that it handles constraints. Indeed, constraints restricts the hypothesis class that can be used to learn the data-to-decision mapping. In contrast, our ICEO framework learns the conditional distribution and use the optimal solution mapping to obtain the decision, which is more flexible in handling constraints.

We also comment on the difference between the problem setting of ICEO and the joint estimation-optimization (JEO) model (Jiang and Shanbhag (2013), Ahmadi and Shanbhag (2014), Jiang and Shanbhag (2016), Ho-Nguyen and Kılınç-Karzan (2019)). The major difference is that, in the JEO model, there is no contextual information considered as predictors of the uncertainty. Besides, several JEO models focus on solving an online convex optimization problem in the optimization stage, while we consider a stochastic optimization problem. We would also like to point out the differences in the problem setting of ICEO and the operational statistics method, in which the downstream optimization goal is considered in finding the optimal operational statistic (Liyanage and Shanthikumar (2005), Chu et al. (2008), Ramamurthy et al. (2012)). We include the contextual information in our problem setting which is not considered in the classic operational statistics literature. Moreover, we aim to learn the underlying conditional distribution rather than finding the best statistic. We also consider constraints in the downstream optimization problem.

Other related works include Ho-Nguyen and Kılınç-Karzan (2020), which investigates the relationship between the prediction part to the performance of the optimization part, mainly in the case of the least squares loss function. Butler and Kwon (2021) focuses on the mean-variance portfolio optimization problem and integrates regression based predictive models with the optimization setting. The authors provide closed-form analytical solutions for the unconstrained cases. Qi et al. (2020b) instead focuses on a multi-period inventory management problem with random demand and leadtime, and provide a practical end-to-end learning framework empowered by deep learning models. The authors demonstrate the empirical success of this approach in practice by conducting a field experiment in industry.

2. Contextual Stochastic Optimization and the ICEO Approach

In this section, we review the basic ingredients of contextual stochastic optimization problems, which is a fundamental model for applying machine learning in many operational contexts, and we formally describe our ICEO approach. We consider a convex CSO, which models a downstream decision-making task. The feasible region for the decision variable $w \in \mathbb{R}^d$, denoted by $S \subset \mathbb{R}^d$, is assumed to be known with certainty. We additionally assume that S is a convex and compact set. Although the feasible region of our optimization task is known with certainty, the objective function $c(\cdot, \xi) : S \to \mathbb{R}$ is stochastic and depends on a random parameter ξ. We assume that, for all values of ξ, $c(\cdot, \xi)$ is a convex function of w. While the precise value of ξ is not known at the time when a decision must be made, we assume that the decision maker observes an associated contextual feature vector $x \in \mathcal{X} \subseteq \mathbb{R}^p$ (sometimes the components of x are referred to as covariates) that can be used to learn information about the objective function. Let $\mathcal{D}$ denote the joint distribution of x and ξ. Then, given an observed $x \in \mathbb{R}^p$, the decision maker's goal is to solve the contextual stochastic optimization problem:

$$\min_{w \in S} \ \mathbb{E}_\xi[c(w, \xi)|x], \tag{1}$$

where the expectation above is with respect to the *conditional distribution* of ξ given x.

It is important to emphasize that the distribution $\mathcal{D}$, and hence the conditional distribution of ξ given any x, is typically unavailable in practice. Instead, a data-driven approach to solving (1) is much more viable. Indeed, one often has available a training dataset $\{(x_i, \xi_i)\}_{i=1}^n$ consisting of historically observed pairs of feature vectors $x_i \in \mathcal{X}$ and associated parameter values ξ_i. If the dataset $\{(x_i, \xi_i)\}_{i=1}^n$ is an independent and identically distributed sample from the distribution $\mathcal{D}$, for example, then it may be possible to learn enough information about the conditional distribution to solve problem (1). Note also that, as pointed out by Bertsimas and Kallus (2020) for example, due to the nonlinearity of the objective function, a point estimate for a prediction of ξ given x usually does not provide enough information about the conditional distribution to produce a reasonable solution of (1). In general, without any additional structural assumptions, e.g., on either the random parameter ξ, the distribution $\mathcal{D}$, or the cost functions $c(\cdot, \cdot)$, adequately learning the conditional distribution for all relevant x may be an intractable problem.

In this work, we consider the case where the random parameter ξ has finite discrete support, i.e., $\xi \in \Xi := \{\tilde{z}_1, \tilde{z}_2, \ldots, \tilde{z}_K\}$. Then, for any $x \in \mathcal{X}$, the conditional distribution of ξ given x is characterized by a probability vector $p^*(x) \in \Delta_K$, where $\Delta_K := \{p \in \mathbb{R}^K : \sum_{k=1}^K p_k = 1, p \geq 0\}$ denotes the $(K-1)$-dimensional unit simplex. That is, $p_k^*(x)$, the k-th component of $p^*(x)$, is defined by $p_k^*(x) = \mathbb{P}_\xi(\xi = \tilde{z}_k|x)$, for all $k = 1, \ldots, K$. Using this notation as well as the shorthand notation $c_k(\cdot) := c(\cdot, \tilde{z}_k)$ for all $k = 1, \ldots, K$, problem (1) can be equivalently written as

$$\min_{w \in S} \ \mathbb{E}_\xi[c(w, \xi)|x] \ = \ \min_{w \in S} \ \sum_{k=1}^K p_k^*(x) c_k(w). \tag{2}$$

2.1. ICEO Approach

Let us now describe the major ingredients of our ICEO approach, as well as the formulation of our ICEO training problem.

Hypothesis Class of Conditional Probability Estimators It is evident from the right side of (2) that learning the conditional distribution $p^*(x)$ is the most critical part of our contextual stochastic optimization setting. We adopt standard ideas from learning theory to learn $p^*(x)$, whereby we employ a compact hypothesis class $\mathcal{H}$ of conditional probability estimators. That is, $\mathcal{H}$ is a compact set (e.g., with respect to the uniform norm) of functions $f : \mathcal{X} \to \Delta_K$. The hypothesis class $\mathcal{H}$ is the first major ingredient of our ICEO approach. Note that the constraint on the output of $f \in \mathcal{H}$, namely $f(x) \in \Delta_K$ for all $x \in \mathcal{X}$, is not standard in most learning problems but is necessitated by our setting. Fortunately, this constraint can be accommodated in a number of ways. A straightforward approach is to consider the softmax operator $\text{soft} : \mathbb{R}^K \to \mathbb{R}^K$ defined by $\text{soft}_k(v) = \frac{\exp(v_k)}{\sum_{j=1}^K \exp(v_j)}$ for $v \in \mathbb{R}^K$. Then, given *any* hypothesis class $\tilde{\mathcal{H}}$ of unconstrained functions $\tilde{f} : \mathcal{X} \to \mathbb{R}^K$, we can define $\mathcal{H}$ as the composition class $\text{soft} \circ \tilde{\mathcal{H}}$. Note that, due to the differentiability properties of the softmax operator, $\text{soft} \circ \tilde{\mathcal{H}}$ naturally inherits differentiability properties from $\tilde{\mathcal{H}}$, which can be very useful from a computational perspective. For another example, consider $\mathcal{H}$ defined by a decision tree partitioning algorithm. Then, for any given x, $f(x)$ can be constructed from the empirical distribution of ξ restricted to the subset of the partition of the training data for which x lies in. Finally, a third approach, which we expand upon in Section 4.3, is to let $\mathcal{H}$ be a constrained linear hypothesis class whereby $\mathcal{H} = \{f : f(x) = Bx \in \Delta_K \text{ for all } x \in \mathcal{X}\}$. Depending on the structure of $\mathcal{X}$, it may be possible to efficiently model the constraint $Bx \in \Delta_K$ for all $x \in \mathcal{X}$, and we discuss specific examples in Section 4.3. We would like to emphasize two points about our approach for estimating the conditional distribution using a hypothesis class $\mathcal{H}$. First, by directly estimating the conditional probability our proposed method has more flexibility in handling constraints as compared to methods that learn a mapping π directly from features x to decisions w. In particular, the approach of learning a mapping from features to decisions requires that the output of the mapping π be feasible in the region S, which may severely constrain the feasible set of π. On the other hand, our approach of composing a user-specified hypothesis class $\mathcal{H}$ with the regularized optimal solution mapping $w_\rho(\cdot)$ allows for a very general selection of $\mathcal{H}$. In particular, the only requirement to achieve asymptotic consistency is compactness of $\mathcal{H}$, and, to provide a generalization bound, we further need $\mathcal{H}$ to have bounded multivariate Rademacher complexity.

Regularized Optimization Oracle. As mentioned previously, we assume that the functions $c_k(\cdot) = c(\cdot, \tilde{z}_k)$, for all $k = 1, \ldots, K$, are all convex functions of w on the convex and compact feasible region S. We additionally assume that these functions are computationally tractable in practice, in the sense that any weighted combination of these functions can be efficiently optimized. Furthermore,

we presume that we can additionally work with a *decision regularization function* $\phi(\cdot) : S \to \mathbb{R}$, which is non-negative and strongly convex with respect to some norm $\|\cdot\|$ on $\mathbb{R}^d$. Given any $p \in \Delta_K$ and $\rho > 0$, define the regularized optimal solution mapping:

$$w_\rho(p) := \arg\min_{w \in S} \sum_{k=1}^{K} p_k c_k(w) + \rho\phi(w). \tag{3}$$

Note that, due to the strong convexity of $\phi(\cdot)$, $w_\rho(p)$ is uniquely defined. Furthermore, we can show that $w_\rho(\cdot)$ is a continuous mapping as demonstrated in Lemma 2. These regularity properties induced by the use of the regularization term $\phi(\cdot)$ are crucial for developing our ICEO methodology as well as for proving associated theoretical guarantees. For computational purposes, we assume that $w_\rho(p)$ can be efficiently computed in practice for any $p \in \Delta_K$ and $\rho > 0$. For example, we may compute $w_\rho(p)$ using a commercial solver or a specialized algorithm that depends on the structure of the $c_k(\cdot)$ and $\phi(\cdot)$ functions. Ideally, the function $\phi(\cdot)$ should be chosen so that the complexity of computing $w_\rho(p)$ is not greatly increased as compared to when $\rho = 0$. Note that our performance guarantees developed in Section 3 hold for any choice of $\phi(\cdot)$ that is strongly convex. When $\phi(\cdot)$ is not present there may be multiple optimal solutions of (3), and we use the notation $W(p)$ to refer to the set of such optimal solutions, i.e., $W(p) := \arg\min_{w \in S} \sum_{k=1}^{K} p_k c_k(w)$.

ICEO Methodology. We are now ready to describe our ICEO methodology and corresponding training problem, whereby we consider an integrated approach that estimates a hypothesis $f \in \mathcal{H}$ in consideration of the downstream optimization goal. We presume that we have collected a training dataset $\{(x_i, \xi_i)\}_{i=1}^n$ consisting of historically observed pairs of feature vectors $x_i \in \mathcal{X}$ and associated parameter values ξ_i. We also presume that the decision maker uses the regularized optimal solution oracle $w_\rho(\cdot)$ defined in (3). We adopt the empirical risk minimization (ERM) principle with respect to the regularized in-sample cost induced by the regularized oracle:

$$\begin{aligned} \min_{f \in \mathcal{H}, w_1, \ldots, w_n \in S} \quad & \frac{1}{n} \sum_{i=1}^{n} c(w_i, \xi_i) + \rho\phi(w_i) \\ \text{s.t.} \quad & w_i = w_\rho(f(x_i)), \end{aligned} \tag{ICEO-ρ}$$

where $\rho > 0$ is a given value of the decision regularization parameter, which can be chosen with cross validation for example. Let $\hat{f} \in \mathcal{H}$ denote a computed optimal solution of (ICEO-ρ). Then, for any newly observed feature vector $x \in \mathcal{X}$, the decision maker implements the decision $w_\rho(\hat{f}(x)) \in S$ formed by composing $w_\rho(\cdot)$ with $\hat{f}(\cdot)$.

Let us contrast the ICEO approach with two more standard approaches: predict-then-optimize (PTO) and estimate-then-optimize (ETO). Note that the phrases "predict" and "estimate" are closely tied in the literature and there is no agreed upon consistent way to distinguish between the two. In our context, we specifically use "predict" to refer to point predictions of the random

parameter ξ and "estimate" to refer to any methodology for estimating the conditional distribution of ξ given any $x \in \mathcal{X}$. Specifically, in the PTO approach, a machine learning model $\hat{g}_{\text{PTO}} : \mathcal{X} \to \Xi$ is built, using the training data, to predict the parameter ξ based on the feature vector x. Then, given any new $x \in \mathcal{X}$, the decision maker implements a decision from the optimal solution set $\arg\min_{w \in S} c(w, \hat{g}_{\text{PTO}}(x))$. As mentioned previously, due to the nonlinearity of the objective, a point estimate for a prediction of ξ is generally too simplistic to provide a reasonable solution of (1). Indeed, unless the conditional distribution is guaranteed to be a point mass, then the PTO approach is not suitable for nonlinear problems. In the case when the objective is linear, a point estimate is actually sufficient and PTO approach is viable. In this linear case, Elmachtoub and Grigas (2021) propose a "smart predict-then-optimize (SPO)" approach that aims to minimize the downstream optimization cost. Furthermore, in this linear case the ICEO approach proposed herein (without regularization) reduces to the SPO problem proposed by Elmachtoub and Grigas (2021).

Returning to the nonlinear case studied herein, the ETO approach learns a model $\hat{f}_{\text{ETO}} : \mathcal{X} \to \Delta_K$ for estimating the conditional distribution of ξ given x. Then, given any new $x \in \mathcal{X}$, the decision maker implements a decision from the optimal solution set $W(\hat{f}_{\text{ETO}}(x))$. Thus, the ETO approach is more aligned with the ICEO approach. The main distinction is that the traditional ETO approach learns the model $\hat{f}_{\text{ETO}}$ in a way that is completely oblivious to the downstream optimization task. For instance, given a hypothesis class $\mathcal{H}$, the ETO approach might select the hypothesis by minimizing the empirical cross-entropy loss, defined for any $f \in \mathcal{H}$ and any observed $(x, \xi = \tilde{z}_k)$ by $\ell_{\text{ce}}(f(x), \xi = \tilde{z}_k) := -\log(f_k(x))$. Alternatively, one may consider a purely non-parametric method for estimating the conditional distribution such as the k-nearest neighbors or CART algorithms for example. In these cases, and several others, Bertsimas and Kallus (2020) demonstrate asymptotic consistency properties of the ETO approach. Kallus and Mao (2020) consider using a (non-parametric) random forests estimator of the conditional distribution in a way that is trained with respect to the cost of the downstream optimization task, akin to the ICEO approach. On the other hand, the ICEO approach directly models the underlying conditional distribution using a hypothesis class $\mathcal{H}$. Thus, while Kallus and Mao (2020) only provide asymptotic consistency results, we are able to prove both asymptotic consistency and generalization bounds for a wide variety of hypothesis classes.

Additional Notation. Due to the compactness of S, the cost function $c(\cdot, \cdot)$ is bounded and we define $\bar{c} := \sup_{w \in S, \xi \in \Xi} c(w, \xi)$. Because of the compactness of S, we can define diameters of S. We let $\text{diam}_j(S) := \sup_{u,v \in S} |u_j - v_j|$ to denote the coordinate-wise diameter of the feasible region S. We further let $\text{diam}(S) := \sum_{j=1}^{d} \text{diam}_j(S)$ denote the summation of the coordinate-wise diameter of all coordinates. Given a norm $\|\cdot\|$ defined on $\mathbb{R}^d$, the distance from a point $w \in \mathbb{R}^d$ to a set $W \subseteq \mathbb{R}^d$ is denoted by $\text{dist}(w, W) := \inf_{u \in W} \|w - u\|$. For a convex function $h(\cdot) : S \to \mathbb{R}$, we let

$\partial h(w)$ denote the set of subgradients of $h(\cdot)$ at w. Let $\circ$ denote the composition of functions. For example, with $f : \mathcal{X} \to \Delta_K$ and $w : \Delta_K \to S$, then $w \circ f$ is the function from $\mathcal{X}$ to S with $(w \circ f)(x) := w(f(x))$ for all $x \in \mathcal{X}$. This function composition notation also extends naturally to function classes. For example, for a class $\mathcal{H}$ of functions $f : \mathcal{X} \to \Delta_K$, we let $w \circ \mathcal{H}$ denote the class of functions $\{w \circ f : f \in \mathcal{H}\}$. We denote the set of non-negative integers a $\mathbb{N}_0$ and let $\mathbb{N}_0^k$ denote the set of all k-dimensional vectors with each component is a non-negative integer. $\mathbb{1}$ denotes the K-dimensional vector with all coordinates taking the value of one. We let $\text{TV}(\mathcal{P}, \mathcal{Q})$ denote the total variation between two probability measures $\mathcal{P}$ and $\mathcal{Q}$ supported on the K-dimensional simplex Δ_K. $\text{TV}(\mathcal{P}, \mathcal{Q}) := \sum_{A \in \mathcal{B}} |\mathcal{P}(A) - \mathcal{Q}(A)|$ where $\mathcal{B}$ denote the class of Borel sets in Δ_K. In Section 4.1, we will use an equivalent expression of $\text{TV}(\mathcal{P}, \mathcal{Q}) = \frac{1}{2} \sup_{f:\Delta \to [-1,1]} (\int_{\Delta_K} f(p) d\mathcal{P}(p) - \int_{\Delta_K} f(p) d\mathcal{Q}(p))$.

2.2. Motivating Examples

In this section, we present a few motivating examples for the ICEO framework, some of which will be revisited in our numerical experiments in Section 5.

Example 1 (Multi-item Newsvendor). The multi-item Newsvendor problem aims to find the optimal replenishment quantities for d different products. We let $\xi := (\xi_1, \ldots, \xi_d)$ denote the random demand of d products and let $w \in \mathbb{R}^d$ denote the associated order quantities. The demand values ξ might be related to contextual information such as promotions, holiday seasons, brand information, etc. The objective of this problem is the total inventory cost including the holding costs h_l and stockout costs b_l, which characterize the over-stock and under-stock, respectively. The objective cost can be formulated as

$$c(w, \xi) := \sum_{l=1}^{d} h_l (w_l - \xi_l)^+ + b_l (\xi_l - w_l)^+, \tag{4}$$

where the function $(\cdot)^+$ is defined as $\max\{\cdot, 0\}$. Moreover, we consider a budget capacity constraint $C > 0$ on the total order quantities and formulate the feasible set as

$$S := \{w : \sum_{l=1}^{d} w_l \leq C, w \geq 0\}.$$

Example 2 (Risk-Averse Portfolio Optimization). We consider the problem of finding an optimal risk-averse portfolio of d assets. We denote the random vector of asset returns by $\xi \in \mathbb{R}^d$, which may be associated with the contextual information such as economic indicators, news headlines, etc. The decision maker aims to find the best allocation of assets $w \in \mathbb{R}^d$ that optimizes a weighted combination of the expected return and variance of the portfolio. By introducing an auxiliary variable $w_0 \in \mathbb{R}$, we formulate the objective as

$$c(w, w_0, \xi) := \alpha \left(\sum_{l=1}^{d} w_l \xi_l - w_0 \right)^2 - \sum_{l=1}^{d} w_l \xi_l, \tag{5}$$

where $\alpha > 0$ is a trade off parameter. Note that the expectation of the first term in (5) is $\alpha\mathbb{E}_\xi\left[(\sum_{l=1}^d w_l\xi_l - w_0)^2\right]$, which represents the variance of the investment return $\text{Var}(\sum_{l=1}^d w_l\xi_l)$ when w_0 is optimally selected as $w_0 = \mathbb{E}_\xi\left[\sum_{l=1}^d w_l\xi_l\right]$, while the second term is the return of the portfolio. Therefore, $\mathbb{E}_\xi[c(w, w_0, \xi)]$ trades off between minimizing the variance and maximizing the expected return of the portfolio. As is standard in the classical portfolio optimization problems, we constrain the portfolio decision in the simplex $\Delta_d = \{w \in \mathbb{R}^d : \sum_{l=1}^d w_l = 1, w \geq 0\}$ and we have

$$S := \{(w, w_0) : w \in \Delta_d, w_0 \geq 0, 0 \leq w_0 \leq \bar{\Xi}\}, \tag{6}$$

where $\bar{\Xi} \geq 0$ is a known upper bound the maximum of the returns $\|\xi\|_\infty$.

Example 3 (Minimum Convex Cost Flow Problem). Many applications such as urban traffic system and area transfers in communication networks can be formulated as a minimum convex cost flow problem (we refer to Chapter 14 of Ahuja et al. (1988) for more details). In the minimum convex cost flow problem, the decision-maker aims to find the maximum flow that minimizes the associated cost on the edges. The cost is a convex function of flow and depends on a random parameter. Suppose we consider a directed graph with d edges and the random parameter $\xi \in \mathbb{R}^d$. In this problem, we consider the objective function

$$c(w, \xi) = \sum_{i=1}^d g(w_i, \xi_i)$$

where g is a convex function of w_i. Similar to the standard network flow problem, we let the matrix A denotes the node-arc incidence matrix of the graph and restrict the flow on each edge in the region $[l, u]$. Therefore, we have the feasible region

$$S = \{w \in \mathbb{R}^d : Aw = 0, w \in [l, u]^d\}.$$

3. Performance Guarantees

In this section, we demonstrate asymptotic consistency and finite-sample performance guarantees of the ICEO approach. Let us first introduce some additional notation. We state our results in terms of arbitrary policy mappings $\pi : \mathcal{X} \to S$, which represent any mapping from the feature space $\mathcal{X}$ to the set of feasible decisions S. Our main interest herein is the class of policies that combine the optimal solution mapping and hypothesis f, i.e., $\Pi = w_\rho \circ \mathcal{H}$. This class of policies includes the policy learned by the ICEO approach as well as policies learned by ETO approaches. In the remaining part of this work, we let $f^* : \mathcal{X} \to \Delta_K$ denote the function that maps from x to the true conditional distribution $p^*(x)$. We refer to f^* as the true hypothesis. Moreover, we define $w(\cdot) : \Delta_K \to S$ as a function that arbitrarily outputs a value from the optimal solution set $W(\cdot)$, i.e., $w(p) \in W(p) = \arg\min_{w \in S} \sum_{k=1}^K p_k c_k(w)$ for all $p \in \Delta_K$.

To quantify our performance guarantees, we define the following risk functions for any policy π and given regularization parameter $\rho \geq 0$:

1. $\hat{R}_n(\pi;\rho)$: The empirical regularized risk, for any given regularization parameter $\rho \geq 0$, with respect to a given sample $\{(x_i,\xi_i)\}_{i=1}^n$, i.e.,

$$\hat{R}_n(\pi;\rho) := \frac{1}{n}\sum_{i=1}^{n} c(\pi(x_i),\xi_i) + \rho\phi(\pi(x_i)). \tag{7}$$

2. $R(\pi;\rho)$: The expected regularized risk, for any given regularization parameter $\rho \geq 0$, with respect to the underlying joint distribution $\mathcal{D}$ of x and ξ, i.e.,

$$R(\pi;\rho) := \mathbb{E}_{x,\xi}\left[c(\pi(x),\xi) + \rho\phi(\pi(x))\right] = \mathbb{E}_x\left[\sum_{k=1}^{K} p_k^*(x)c_k(\pi(x)) + \rho\phi(\pi(x))\right], \tag{8}$$

where $p_k^*(x) = \mathbb{P}_\xi(\xi = \tilde{z}_k|x)$ for all $k = 1,\ldots,K$.

We also use the short hand notation $\hat{R}_n(\pi) := \hat{R}_n(\pi;0)$ and $R(\pi) := R(\pi;0)$ to denote the unregularized empirical and expected risks, respectively. Note that $\hat{R}_n(\cdot;\rho)$ is the objective function of (ICEO-ρ). We further define the optimal risk values for the class of policies $\Pi = w_\rho \circ \mathcal{H}$ that we consider herein.

1. J^*: the optimal expected unregularized risk, i.e.,

$$J^* := \min_{f\in\mathcal{H}} \mathbb{E}_x\left[\sum_{k=1}^{K} p_k^*(x)c_k(w(f(x)))\right] = \min_{f\in\mathcal{H}} R(w\circ f;0). \tag{9}$$

2. J_ρ^*: the optimal expected regularized risk for any given regularization parameter $\rho > 0$, i.e.,

$$J_\rho^* := \min_{f\in\mathcal{H}} \mathbb{E}_x\left[\sum_{k=1}^{K} p_k^*(x)c_k(w_\rho(f(x))) + \rho\phi(w_\rho(f(x)))\right] = \min_{f\in\mathcal{H}} R(w_\rho\circ f;\rho).$$

3. $\hat{J}_\rho^n$: the optimal empirical regularized risk with any given sample S_n and a given regularization parameter $\rho > 0$, i.e.,

$$\hat{J}_\rho^n := \min_{f\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} c(w_\rho(f(x_i)),\xi_i) + \rho\phi(w_\rho(f(x_i))) = \min_{f\in\mathcal{H}} \hat{R}_n(w_\rho\circ f;\rho),$$

and we let $\hat{f}_\rho^n$ denote its optimal solution.

3.1. Asymptotic Consistency

We first demonstrate the asymptotic consistency of the ICEO approach. The consistency of our approach is three-fold: the consistency of the ICEO risk, the consistency of the ICEO decisions, and the consistency of the ICEO hypothesis. Our asymptotic consistency results require the following conditions:

ASSUMPTION 1. *For the compact hypothesis class $\mathcal{H}$, we have the following:*

A. (Model Specification) The hypothesis class $\mathcal{H}$ includes the true hypothesis f^* i.e., $f^* \in \mathcal{H}$.

B. (Unique true hypothesis.) Suppose that the training data of features, $\{x_i\}_{i=1}^n$, is an i.i.d. sequence generated from the distribution $\mathcal{D}_x$. There does not exists a hypothesis $f \neq f^*$ in $\mathcal{H}$ such that $W(f(x)) \cap W(f^*(x)) \neq \emptyset$, $\mathcal{D}_x$-almost surely for all $x \in \mathcal{X}$.

To guarantee the consistency of the ICEO method, we consider a sequence of regularization parameters ρ_n, depending on the sample size n, such that ρ_n converges to zero as n grows to infinity. Theorem 1 below demonstrates the three-levels of consistency.

THEOREM 1 **(Asymptotic Consistency of ICEO)**. *Suppose that the training data (x_i, ξ_i) is an i.i.d. sequence from the distribution $\mathcal{D}$ and that the sequence of regularization parameters ρ_n satisfies $\lim_{n\to\infty} \rho_n = 0$. Then, under Assumption 1.A, we have the following:*

(i) The optimal empirical regularized risk converges to the optimal expected risk, i.e., $\hat{J}^n_{\rho_n} \to J^*$ with probability 1.

(ii) $\mathcal{D}_x$-almost surely for all $x \in \mathcal{X}$, the sequence of ICEO decisions $w_{\rho_n}(\hat{f}^n_{\rho_n}(x))$ converges to the true set of optimal decisions $W(f^*(x))$, i.e., $\text{dist}(w_{\rho_n}(\hat{f}^n_{\rho_n}(x)), W(f^*(x))) \to 0$ with probability 1.

(iii) Additionally, with Assumption 1.B, the sequence of ICEO hypotheses converges to the true hypothesis, i.e, $\hat{f}^n_{\rho_n} \to f^*$ with probability 1.

We would like to clarify the relationship between the asymptotic consistency stated in Theorem 1 and the asymptotic optimality defined in Bertsimas and Kallus (2020). In Bertsimas and Kallus (2020), the authors provide the asymptotic optimality as the ICEO decisions reaching the best performance possible. Because of the continuity of the cost function c, the convergence of ICEO decisions, as stated in (ii) of Theorem 1, implies the asymptotic optimality stated in Bertsimas and Kallus (2020).

Proof of Theorem 1 In this proof, we slightly abuse the notations and let $R(f;\rho)$ denote $R(w_\rho \circ f;\rho)$ for any $\rho \geq 0$ and $\hat{R}_n(f;\rho)$ denote $\hat{R}_n(w_\rho \circ f;\rho)$. We first show that $\lim_{\rho\to 0} J^*_\rho = J^*$. Recall that $J^* = R(f^*;0)$ and $J^*_\rho = R(f^*_\rho; rho)$, and we have:

$$R(f^*;0) \leq R(f^*_\rho;0) \tag{10}$$

$$\leq R(f^*_\rho;\rho) \tag{11}$$

$$\leq R(f^*;\rho), \tag{12}$$

where (10) holds because the true hypothesis f^* achieves the optimal value J^*. (11) follows from the fact that ϕ is a non-negative function and (12) follows from the fact that f^*_ρ is the optimizer of $R(\cdot;\rho)$. In the meanwhile, $R(f^*;\rho) \to R(f^*;0)$ as $\rho \to 0$. Thus,

$$J^*_\rho := R(f^*_\rho;\rho) \to R(f^*;0) = J^*.$$

Then we want to show that $\hat{J}^n_\rho \to J^*_\rho$ with probability 1 as $n \to \infty$. Let $l_i(f,\rho) := c(w_\rho(f(x_i)),\xi_i) + \rho\phi(w_\rho(f(x_i)))$, then $l_i(\cdot,\rho)$ are i.i.d. random functions on the compact hypothesis class $\mathcal{H}$. Due to the compactness of S, both $R(\cdot;\rho)$ and $l_i(\cdot,\rho)$ are bounded. Then we can apply the main theorem in Rubin (1956) and have that $\frac{1}{n}\sum_{i=1}^n l_i(\cdot,\rho) \to R(\cdot;\rho)$ with probability 1 for all f uniformly. Moreover, together with the continuity of $R(\cdot;\rho)$, the uniform convergence of $\hat{R}_n$ leads to the Γ-convergence of $\hat{l}(\cdot,\rho)$ to $R(\cdot;\rho)$ in probability (Braides (2006)). Then we can apply the Fundamental Theorem of Γ-convergence and leads to the convergence of minimum values $\hat{J}^n_\rho$ converges to J^*_ρ (Braides et al. (2002)). Then for any sequence of $\rho_n > 0$ and$\rho_n \to 0$, as $n \to \infty$, we can find a sequence of $\epsilon_n > 0$ that satisfies $\lim_{n\to\infty}\epsilon_n = 0$ and the following conditions: $J^*_{\rho_n} - J^* \le \epsilon_n$ and $|\hat{J}^n_{\rho_n} - J^*_{\rho_n}| < \epsilon_n$ with probability 1. Thus, $|\hat{J}^n_{\rho_n} - J^*| \le 2\epsilon_n$ for all n, which leads to $\hat{J}^n_{\rho_n} \to J^*$ with probability 1 and (i) is proved.

Due to the compactness of S and $\mathcal{H}$, with any sequence of $\rho_i \to 0$, the sequence $w_{\rho_i}(f^*_{\rho_i}(x))$ has accumulation points. Let $w_{\rho_t}(f^*_{\rho_t}(x))$ be any subsequence converging to an accumulation point $w_{\rho_\infty}(f^*_{\rho_\infty}(x)) \in S$. Note that we have

$$J^* = \lim_{t\to\infty} \mathbb{E}_x[\sum_{k=1}^K f^*_k(x)c_k(w_{\rho_t}(f^*_{\rho_t}(x))) + \rho_t\phi(w_{\rho_t}(f^*_{\rho_t}(x)))] \tag{13}$$

$$\ge \lim_{t\to\infty} \mathbb{E}_x[\sum_{k=1}^K f^*_k(x)c_k(w_{\rho_t}(f^*_{\rho_t}(x)))] \tag{14}$$

$$\ge \mathbb{E}_x[\sum_{k=1}^K f^*_k(x)c_k(w_{\rho_\infty}(f^*_{\rho_\infty}(x)))]. \tag{15}$$

(13) follow from $J^*_\rho \to J^*$ when $\rho \to 0$ and (14) holds due to the non-negativity of the regularization term ϕ. Then by Fatou's lemma, we have the last inequality (15). Since $W(f^*(x))$ is defined the set of optimal solutions of $\sum_{k=1}^K f^*_k(x)c_k(\cdot)$ for all x, then $\mathcal{D}_x$ almost surely for all x, $w_{\rho_\infty}(f^*_{\rho_\infty}(x))$ must lie in the set $w(f^*(x))$. Then $\text{dist}(w_{\rho_t}(f^*_{\rho_t}(x)), W(f^*(x))) \le \|w_{\rho_t}(f^*_{\rho_t}(x)) - w_{\rho_\infty}(f^*_{\rho_\infty}(x))\|$. By the continuity of the norm $\|\cdot\|$, $\lim_{t\to\infty}\|w_{\rho_t}(f^*_{\rho_t}(x)) - w_{\rho_\infty}(f^*_{\rho_\infty}(x))\| = \|\lim_{t\to\infty} w_{\rho_t}(f^*_{\rho_t}(x)) - w_{\rho_\infty}(f^*_{\rho_\infty}(x))\| = 0 \ge \text{dist}(w_{\rho_t}(f^*_{\rho_t}(x)), W(f^*(x)))$. Therefore, $\mathcal{D}_x$-almost surely for all x, $\text{dist}(w_\rho(\hat{f}^n_\rho(x)), W(f^*(x))) \to 0$ as $\rho \to 0$.

Moreover, for any fixed $\rho > 0$, due to the compactness of $\mathcal{H}$ and the fact that $\frac{1}{n}\sum_{i=1}^n l_i(\cdot,\rho)$ converges to $R(\cdot;\rho)$ uniformly with probability 1, we can apply the fundamental theorem of Γ-convergence again and conclude that any accumulation point of $\hat{f}^n_\rho$, denoted by f^∞_ρ, minimizes $R(\cdot;\rho)$ with probability 1. Because of the strongly convexity of $c_k(\cdot) + \rho\phi(\cdot)$, if f^∞_ρ minimizes $R(\cdot;\rho)$, then $w(f^\infty_\rho(x)) = w(f^*_\rho(x))$ $\mathcal{D}_x$-almost surely for all $x \in \mathcal{X}$. Therefore, given any sequence $\rho_n \to 0$, we can find a sequence of $\delta_n \ge 0$ such that $\lim_{n\to\infty}\delta_n = 0$ and satisfies both $\|w(f^*_{\rho_n}(x)) - w_{\rho_n}(\hat{f}^n_{\rho_n}(x))\| \le \delta_n$ and $\text{dist}(w_{\rho_n}(f^*_{\rho_n}(x)), W(f^*(x))) \le \delta_n$, with probability 1. Therefore, with probability 1, we have

$$\text{dist}(w_{\rho_n}(\hat{f}^n_{\rho_n}(x)), W(f^*(x)) = \inf_{u\in W(f^*(x))} \|w_{\rho_n}(\hat{f}^n_{\rho_n}(x)) - w_{\rho_n}(f^*_{\rho_n}(x)) + w_{\rho_n}(f^*_{\rho_n}(x)) - u\|$$

$$\leq \|w_{\rho_n}(\hat{f}^n_{\rho_n}(x)) - w_{\rho_n}(f^*_{\rho_n}(x))\| + \inf_{u \in W(f^*(x))} \|w_{\rho_n}(f^*_{\rho_n}(x)) - u\| \quad (16)$$

$$\leq 2\delta_n, \qquad \mathcal{D}_x - \text{almost surely}, \quad (17)$$

where (16) follows from the triangle inequality. (17) further leads to the conclusion in (ii).

If we have an accumulation points of $\hat{f}^n_{\rho_n}$, denoted as $\bar{f}$, that does not equal to f^*, then by (ii), we have $W(\bar{f}(x)) \cap W(f^*(x)) \neq \emptyset$ for all $x \in \mathcal{X}$ almost surely, which contradict to Assumption 1.B. Thus with the uniqueness assumption, the true hypothesis f^* can be recovered by $\hat{f}^n_{\rho_n}$. □

3.2. Finite Sample Performance Guarantees

We now provide finite sample performance guarantees of the ICEO solution $\hat{f}^n_{\rho_n}$ in the form of generalization bounds based on Rademacher complexities. In particular, our overall strategy is as follows: *(i)* we demonstrate that, due to the presence of the strongly convex decision regularization function $\phi(\cdot)$, the optimal solution mapping $w_\rho(\cdot)$ is Lipschitz, *(ii)* we use the result of Maurer (2016) to bound the Rademacher complexity with respect to the cost function of the ICEO framework by the multivariate Rademacher complexity of the underlying hypothesis class $\mathcal{H}$. In addition, we slightly abuse the notation and let $c(\cdot) : S \to \mathbb{R}^K$ denote a vector-valued mapping, where each component $c_k(w)$ denotes the cost $c(w, \xi = \tilde{z}_k)$ for all scenarios $k = 1, \ldots, K$, as defined earlier in Section 2.

Before we investigate the Rademacher complexities, we first demonstrate the Lipschitz property of the regularized optimal solution mapping $w_\rho(\cdot)$ for any positive parameter ρ, based on the following assumption regarding the Lipschitz property of the cost function $c(w)$ and the strong convexity constant of the decision regularization function $\phi(\cdot)$.

Assumption 2. *The cost function $c(\cdot)$ and the decision regularization function $\phi(\cdot)$ satisfy the following conditions:*

A. *$c(\cdot)$ is L_c-Lipschitz with respect to the decision $w \in S$, i.e., it holds that $\|c(w_1) - c(w_2)\|_2 \leq L_c\|w_1 - w_2\|$ for all $w_1, w_2 \in S$.*
B. *The decision regularization function $\phi(\cdot)$ is a 1-strongly convex function on the compact set S.*

Note that we use the ℓ_2 norm as the norm on the space of outputs of the cost functions $c(\cdot)$, while the norm on the space of decisions w remains the generic norm $\|\cdot\|$. The reason for focusing on the ℓ_2 norm is that we can apply the elegant vector contraction inequality of Maurer (2016) when analyzing the Rademacher complexity. It is also worth mentioning that the Lipschitz condition in Assumption 2.A implies that the cost functions $c_k(\cdot)$ are uniformly L_c-Lipschitz, i.e., $\|c(w_1) - c(w_2)\|_\infty \leq L_c\|w_1 - w_2\|$.

PROPOSITION 1 **(Lipschitz Properties of $w_\rho(\cdot)$ and $c(\cdot)$).** *Suppose Assumption 2 holds. Then, for any $\rho > 0$, the optimal solution mapping $w_\rho(\cdot)$ is $(\frac{L_c}{\rho})$-Lipschitz:*

$$\|w_\rho(p) - w_\rho(p')\| \leq \frac{L_c}{\rho}\|p - p'\|_2, \qquad \forall p, p' \in \Delta_K. \tag{18}$$

Furthermore, $c(w_\rho(\cdot))$ is $(\frac{L_c^2}{\rho})$-Lipschitz:

$$\|c(w_\rho(p)) - c(w_\rho(p'))\|_\infty \leq \|c(w_\rho(p)) - c(w_\rho(p'))\|_2 \leq \frac{L_c^2}{\rho}\|p - p'\|_2, \qquad \forall p, p' \in \Delta_K. \tag{19}$$

The proof of this Proposition follows standard arguments of Nesterov's smoothing technique (Nesterov (2003)), and a related result with a similar proof style appears in Gupta and Kallus (2021).

Proof of Proposition 1 Let $p, p' \in \Delta_K$ be fixed. We let $h_\rho(\cdot, p) : S \to \mathbb{R}$ be defined by $h_\rho(w, p) := \sum_{k=1}^K p_k c_k(w) + \rho\phi(w)$. Since $\phi(\cdot)$ is a 1-strongly convex function, then $h_\rho(\cdot, p)$ is ρ-strongly convex and it holds for all $w \in S$ and $g \in \partial_w h_\rho(w, p)$ that

$$h_\rho(w', p) - h_\rho(w, p) \;\geq\; g^T(w' - w) + \frac{\rho}{2}\|w' - w\|^2 \qquad \forall w' \in S. \tag{20}$$

Since $w_\rho(p) = \arg\min_{w \in S} h_\rho(w, p)$, the first-order optimality condition implies there exists a subgradient $g \in \partial h(w_\rho(p), p)$ such that $g^T(w' - w_\rho(p)) \geq 0$ for all $w' \in S$. Applying this condition in (20) with $w \leftarrow w_\rho(p)$ $w' \leftarrow w_\rho(p')$ yields

$$h_\rho(w_\rho(p'), p) - h_\rho(w_\rho(p), p) \;\geq\; \frac{\rho}{2}\|w_\rho(p') - w_\rho(p)\|^2.$$

Switching the role of p and p' yields

$$h_\rho(w_\rho(p), p') - h_\rho(w_\rho(p'), p') \;\geq\; \frac{\rho}{2}\|w_\rho(p) - w_\rho(p')\|^2.$$

Adding the above two inequalities together yields

$$\begin{aligned}
\rho\|w_\rho(p) - w_\rho(p')\|^2 \;&\leq\; h_\rho(w_\rho(p), p') - h_\rho(w_\rho(p'), p') + h_\rho(w_\rho(p'), p) - h_\rho(w_\rho(p), p) \\
&=\; [h_\rho(w_\rho(p), p') - h_\rho(w_\rho(p), p)] - [h_\rho(w_\rho(p'), p') - h_\rho(w_\rho(p'), p)] \\
&=\; \sum_{k=1}^K (p'_k - p_k) c_k(w_\rho(p)) - \sum_{k=1}^K (p'_k - p_k) c_k(w_\rho(p')) \\
&=\; \sum_{k=1}^K (p'_k - p_k)(c_k(w_\rho(p)) - c_k(w_\rho(p'))) \\
&\leq\; \|p - p'\|_2 \|c(w_\rho(p)) - c(w_\rho(p'))\|_2 \\
&\leq\; L_c\|p - p'\|_2 \|w_\rho(p) - w_\rho(p')\|,
\end{aligned}$$

where the last inequality uses Assumption (2.A). Dividing by $\|w_\rho(p) - w_\rho(p')\|$ leads to (18), and combining the resulting inequality again with (2.A) yields (19). □

To establish the generalization bound for the ICEO risk, we rely on both regular single-variate and multi-variate Rademacher complexity. In the ICEO setting, given a class of policies Π, where $\pi : \mathcal{X} \to S$ for all $\pi \in \Pi$, we can apply generalization bounds that directly use the Rademacher complexity of the function class $c \circ \Pi$. Given a sample $\{(x_i, \xi_i)\}_{i=1}^n$ the *empirical Rademacher complexity* $\hat{\mathfrak{R}}_n(c \circ \Pi)$ of the function class $c \circ \Pi$ is defined by

$$\hat{\mathfrak{R}}_n(c \circ \Pi) \ := \ \mathbb{E}_\sigma \left[\frac{2}{n} \sup_{g \in c \circ \Pi} \sum_{i=1}^n \sigma_i g(x_i, \xi_i) \right] = \mathbb{E}_\sigma \left[\frac{2}{n} \sup_{\pi \in \Pi} \sum_{i=1}^n \sigma_i c(\pi(x_i), \xi_i) \right],$$

where σ_i are independent random variables drawn from the Rademacher distribution, i.e. $\Pr(\sigma_i = +1) = \Pr(\sigma_i = -1) = \frac{1}{2}$ for all $i = 1, 2, \ldots, n$. The *expected Rademacher complexity* $\mathfrak{R}_n(c \circ \Pi)$ is then defined as the expectation of $\hat{\mathfrak{R}}_n(c \circ \Pi)$ with respect to the i.i.d. sample $\{(x_i, \xi_i)\}_{i=1}^n$ drawn from the distribution $\mathcal{D}$:

$$\mathfrak{R}_n(c \circ \Pi) = \mathbb{E}_{(x_i, \xi_i) \sim \mathcal{D}}[\hat{\mathfrak{R}}_n(c \circ \Pi)].$$

Next, we introduce the multivariate Rademacher complexity as a generalization of the regular Rademacher complexity to a class of vector-valued functions. In the ICEO context, we focus on the hypothesis class $\mathcal{H}$ which takes values in Δ_K. Following Bertsimas and Kallus (2020), Maurer (2016) and Balghiti et al. (2019), the *empirical multivariate Rademacher complexity* $\hat{\mathfrak{R}}_n(\mathcal{H})$ is defined in our context as

$$\hat{\mathfrak{R}}_n(\mathcal{H}) = \mathbb{E}_\sigma \left[\frac{2}{n} \sup_{f \in \mathcal{H}} \sum_{i=1}^n \sum_{k=1}^K \sigma_{ik} f_k(x_i) \right],$$

where σ_{ik} are also independent random variables drawn from the Rademacher distribution for all $i = 1, 2, \ldots, n$ and $k = 1, \ldots, K$. Correspondingly, the *expected multivariate Rademacher complexity* $\mathfrak{R}_n(\mathcal{H})$ is then defined as

$$\mathfrak{R}_n(\mathcal{H}) = \mathbb{E}_{x_i \sim \mathcal{D}_x}[\hat{\mathfrak{R}}_n(\mathcal{H})],$$

In the remainder of this section, we provide generalization bounds with respect to the expected single-variate and multi-variate Rademacher complexities. We note that similar results can be achieved with respect to the empirical versions of the Rademacher complexities. Our focus on the expected versions is justified since, for many hypothesis classes $\mathcal{H}$, we can bound $\mathfrak{R}_n(\mathcal{H})$ by a term that converges to 0 as the sample size n grows. For example, Balghiti et al. (2019) establish upper bounds of $\mathfrak{R}_n(\mathcal{H})$ for regularized linear hypothesis classes with the rate of $\mathcal{O}(\frac{1}{\sqrt{n}})$, where the $\mathcal{O}(\cdot)$ notation hides dimension dependent constants that depend on the type of regularization used.

Given a sample $\{(x_i, \xi_i)\}_{i=1}^n$, we aim to provide a high-probability bound on the out-of-sample risk $R(w_{\rho_n} \circ f)$, given the in-sample risks $\hat{R}_n(w_{\rho_n} \circ f)$ and $\hat{R}_n(w_{\rho_n} \circ f; \rho_n)$, that holds uniformly for any hypothesis $f \in \mathcal{H}$. As such, our generalization bound is constructed based on the the classic generalization bound with Rademacher complexity due to Bartlett and Mendelson (2002), which we restate below as specialized to the ICEO setting. Recall that $\bar{c} := \sup_{w \in S, \xi \in \Xi} c(w, \xi)$.

THEOREM 2 **(Bartlett and Mendelson (2002))**. *Let Π be a family of functions mapping from $\mathcal{X}$ to S with bounded Rademacher complexity $\mathfrak{R}_n(c\circ\Pi)$. Then, for any $\delta\in(0,1]$, with probability at least $1-\delta$ over i.i.d. data $\{(x_i,\xi_i)\}_{i=1}^n$ drawn from the distribution $\mathcal{D}$, the following inequality holds for all $\pi\in\Pi$:*

$$R(\pi) \;\leq\; \hat{R}_n(\pi)+\mathfrak{R}_n(c\circ\Pi)+\bar{c}\sqrt{\frac{\log(\frac{1}{\delta})}{2n}}. \tag{21}$$

The next step of our analysis is to apply the vector contraction inequality of Maurer (2016) to derive a generalization bound that depends directly on the multi-variate Rademacher complexity of the hypothesis class $\mathcal{H}$.

THEOREM 3 **(Generalization of ICEO)**. *Suppose Assumption 2 holds and that the hypothesis class $\mathcal{H}$ has bounded multi-variate Rademacher complexity $\mathfrak{R}_n(\mathcal{H})$. Then, for any $\delta\in(0,1]$ and $\rho_n>0$, with probability at least $1-\delta$ over i.i.d. data $\{(x_i,\xi_i)\}_{i=1}^n$ drawn from the distribution $\mathcal{D}$, the following inequalities hold for all $f\in\mathcal{H}$:*

$$\begin{aligned} R(w_{\rho_n}\circ f) \;&\leq\; \hat{R}_n(w_{\rho_n}\circ f)+\frac{\sqrt{2}L_c^2}{\rho_n}\mathfrak{R}_n(\mathcal{H})+\bar{c}\sqrt{\frac{\log(\frac{1}{\delta})}{2n}} \\ &\leq\; \hat{R}_n(w_{\rho_n}\circ f;\rho_n)+\frac{\sqrt{2}L_c^2}{\rho_n}\mathfrak{R}_n(\mathcal{H})+\bar{c}\sqrt{\frac{\log(\frac{1}{\delta})}{2n}}. \end{aligned}$$

Note that the right hand side of the first inequality in Theorem 3 involves the unregularized empirical risk, which may be evaluated for any $f\in\mathcal{H}$. The right hand side of the second inequality in Theorem 3 invovles the regularized empirical risk, which is precisely the objective function of (ICEO-ρ). As mentioned previously, one can often establish upper bounds on $\mathfrak{R}_n(\mathcal{H})$ that converge to zero, for example at the rate $\mathcal{O}(\frac{1}{\sqrt{n}})$. Therefore, Theorem 3 suggests that we should set the sequence of regularization parameters ρ_n so that $\mathfrak{R}_n(\mathcal{H})/\rho_n$ converges to zero as well, in which case the remainder terms on the right-hand side of (**??**) converge to zero. We conclude this section with the proof of Theorem 3.

Proof of Theorem 3 Due to Proposition 1, in particular the Lipschitz property of $c(\cdot)$ in (19), we can apply the vector contraction inequality from Maurer (2016) which, stated in terms of empirical Rademacher complexities, yields

$$\hat{\mathfrak{R}}_n(c\circ w_{\rho_n}\circ\mathcal{H}) \;\leq\; \frac{\sqrt{2}L_c^2}{\rho_n}\hat{\mathfrak{R}}_n(\mathcal{H}).$$

Taking expectations of both sides of the above inequality, with respect to i.i.d. data $\{(x_i,\xi_i)\}_{i=1}^n$ drawn from the distribution $\mathcal{D}$, yields

$$\mathfrak{R}_n(c\circ w_{\rho_n}\circ\mathcal{H})\leq\frac{\sqrt{2}L_c^2}{\rho_n}\mathfrak{R}_n(\mathcal{H}).$$

Then, a direct application of Theorem 2 yields the first inequality. Finally, for any $\rho_n > 0$, note that $R(w_{\rho_n} \circ f) \leq R(w_{\rho_n} \circ f; \rho_n)$ due to the non-negativity of the decision regularization function $\phi(\cdot)$, which yields the second inequality. $\square$

4. Computational Methods

In this section, we discuss the computational difficulties of solving the ICEO formulation (ICEO-ρ) and present multiple approaches to address them.

Non-convexity. First, we point out that the ICEO formulation, (ICEO-ρ), is not a convex optimization problem even in a very simple case where both the objective and constraints of the nominal optimization problem are linear and the decision regularization is quadratic, as stated in Example 4.

Example 4 (Linear nominal optimization problem). Consider an example with a linear objective function in the optimization stage, i.e., $c_j(w)$ is a linear function $c_j^T w$ for some $c_j \in \mathbb{R}^d$ for all $j = 1, \ldots, K$. Suppose we use the decision regularization function $\phi(w) := \frac{1}{2}\|w\|_2^2$. For any $p \in \Delta_K$, let $\bar{c}(p) := \sum_{j=1}^K p_j c_j$. Then, note that

$$w_\rho(p) = \arg\min_{w \in S} \left\{ \bar{c}(p)^T w + \tfrac{\rho}{2}\|w\|_2^2 \right\} = \arg\min_{w \in S} \left\{ \tfrac{\rho}{2}\|(\bar{c}(p)/\rho) - w\|_2^2 \right\} = \Pi_S(\bar{c}(p)/\rho),$$

where $\Pi_S(\cdot)$ is the Euclidean projection operator onto S. Then, (ICEO-ρ) is the problem of minimizing a sum of linear functions composed with projection operators, which is generally non-convex. At best, when S is a polyhedron, i.e., $S := \{w \in \mathbb{R}^d : Aw \leq b\}$ and when we adopt a linear hypothesis class $\mathcal{H} = \{x \mapsto Bx \in \Delta_K : B \in \mathbb{R}^{K \times p}\}$, we can formulate (ICEO-ρ) as a bilinear quadratic optimization problem. Indeed, (ICEO-ρ) can be reformulated as

$$\begin{aligned}
\min_{B, w_i, \lambda_i} \quad & \frac{1}{n}\sum_{i=1}^n \sum_{j=1}^K (\mathbb{1}\{\xi_i = \tilde{z}_j\}(Bx_i)_j c_j^T w_i + (\rho/2) w_i^T w_i) && \text{(ICEO-}\rho\text{-LP)} \\
\text{s.t.} \quad & \frac{\rho}{2} w_i^T w_i + \sum_{j=1}^K (Bx_i)_j c_j^T w_i + \frac{1}{2}\Big(\sum_{j=1}^K (Bx_i)_j c_j + A^T\lambda\Big)^T \Big(\sum_{j=1}^K (Bx_i)_j c_j + A^T\lambda\Big) \\
& + \lambda_i^T b \leq 0, \quad \forall i = 1, \ldots, n \\
& Aw_i \leq b, \quad \forall i = 1, \ldots, n \\
& \lambda_i \geq 0, \quad \forall i = 1, \ldots, n
\end{aligned}$$

Note that the dual function of the nominal quadratic optimization is

$$-\frac{1}{2}\Big(\sum_{j=1}^K (B^T x_i)_j c_j + A^T\lambda\Big)^T \Big(\sum_{j=1}^K (B^T x_i)_j c_j + A^T\lambda\Big) - \lambda^T b$$

thus the dual problem becomes

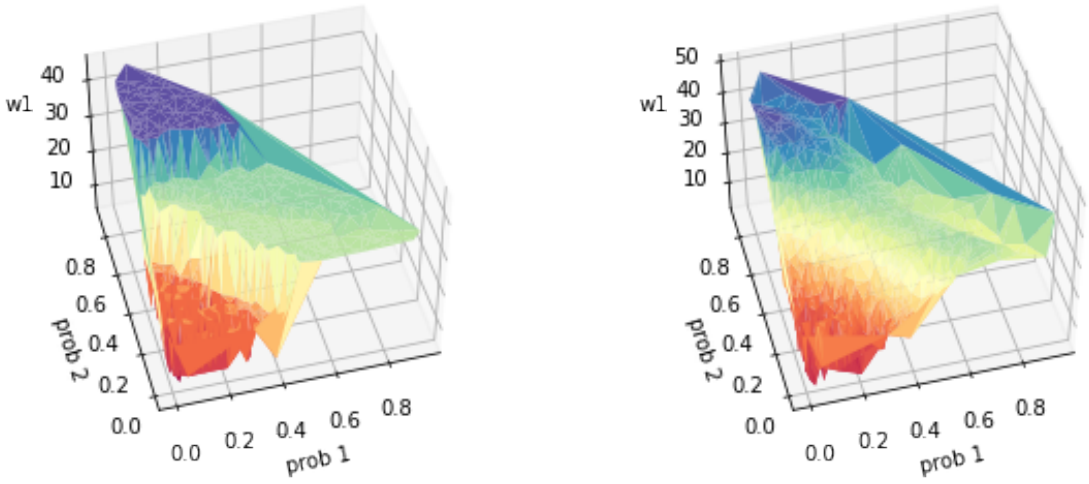
$$\min_{\lambda \geq 0} \frac{1}{2}(\sum_{j=1}^{K}(B^T x_i)_j c_j + A^T\lambda)^T(\sum_{j=1}^{K}(B^T x_i)_j c_j + A^T\lambda)) + \lambda^T b. \tag{22}$$

The first two group of constraints in (ICEO-ρ-LP) is to guarantee that w_i and λ_i are the optimal primal and dual solutions. The second and third group of constraints are for the primal and dual feasibility. $\square$

As demonstrated above, even in this simplest case, (ICEO-ρ-LP) is not a convex optimization problem. Besides non-convexity, a more serious issue from a practical standpoint is the potential of non-differentiability of the optimal solution mapping.

Non-differentiability. To solve the non-convex ICEO problem (ICEO-ρ), a default approach in machine learning is to use a gradient-based algorithm such as the basic stochastic gradient descent method. Indeed, in practice gradient-based algorithms are often able to deliver high quality solutions for machine learning problems, especially in high dimensions. Unfortunately, applying these basic gradient-based methods to solve the ICEO formulation poses an additional major difficulty due to the non-differentiability of the optimal solution mapping $w_\rho(\cdot)$. Although $w_\rho(\cdot)$ is a continuous function, as guaranteed for example by Proposition 1, it is generally not differentiable. The non-differentiability leads to major difficulties in applying gradient-based method while solving ICEO-ρ. As reviewed in Section 1.1, existing studies that focused on directly learning the optimal solution mapping $w(f^*(x))$ also encounter the same issue of non-differentiability. Wilder et al. (2019b) does not discuss much about this. Donti et al. (2017) use the output of an automatic gradient function calculated by back propagation of a neural network. Agrawal et al. (2019) approximate the gradient by solving a group of linear equations based on KKT conditions. However, all existing methods fail to demonstrate theoretical reliability or performance guarantees in approximating the gradient.

The non-differentiability of the optimal solution mapping mainly arises from the constraints and the points of discontinuity occur where there is a "jump" in the optimal solution, e.g., in the polyhedral case as in Example 4. Therefore, a non-differentiability optimal solution map may also have regions where it is constant (or close to constant), resulting in the gradient of the ICEO objective being equal to zero. We demonstrate this poor behavior in Figure 1a, where we plot the second coordinate of the optimal solution mapping $w_\rho(\cdot)$ with respect to the first two coordinates of the input probability vector, for the multi-product newsvendor problem in Example 1, demonstrating the piece-wise constant shape. Such piece-wise constant shapes will greatly impede the performance of gradient-based methods, even if the gradient is easily calculated. This is because the gradient of the optimal solution mapping is zero in flat regions creating poor local minima that are very difficult to escape.



(a) 3-D plot of the optimal oracle. It is clear that the landscape of the optimal solution mapping has cliffs and platforms.

(b) 3-D plot of the approximated oracle constructed using polynomial functions. The piece-wise constant shape is smoothed out.

Figure 1 The landscape of the optimal and the approximated oracle.

To address the issue of non-differentiability and its consequences leading to poor local optima, we develop a framework for approximating the mapping $w_\rho(\cdot)$ with a differentiable function $\tilde{w}_\rho(\cdot)$, which allows us to smooth out the optimal solution mapping and eliminate those poor local minima. Figure 1b is an example of smoothing out the piece-wise constant shape by constructing an approximate oracle using polynomial kernel regression. As noted before, gradient-based methods are often highly effective at delivering high quality solutions to non-convex machine learning problems in practice. Thus, in a practical sense, the non-differentiability of the optimal solution mapping is a much more serious concern than the non-convexity. Our general strategy of approximating the optimal solution mapping with a differentiable function, for which we expand upon and give examples in Section 4.1, greatly increases the practical viability of the ICEO approach.

4.1. Approximate Optimal Solution Mappings

As stated in the previous section, the major computational difficulty in solving the ICEO training problem (ICEO-ρ) in practice is the non-differentiability of the mapping $w_\rho(\cdot)$. To overcome this difficulty, for any given ρ, we approximate the function $w_\rho(\cdot)$ with a differentiable function $\tilde{w}_\rho(\cdot): \Delta_K \to S$. Then instead of (ICEO-$\rho$), we solve the following problem:

$$\begin{array}{ll} \min\limits_{f\in\mathcal{H}} & \dfrac{1}{n}\sum\limits_{i=1}^{n} c(w_i,\xi_i) + \rho\phi(w_i) \\ \text{s.t.} & w_i = \tilde{w}_\rho(f(x_i)) \end{array} \qquad \text{(Approx-ICEO-}\rho\text{)}$$

To construct such an approximation $\tilde{w}_\rho(\cdot)$, we rely on the ability to evaluate the optimal solution mapping $w_\rho(p)$ for any given $p \in \Delta_K$, as stated in Section 2. We can then generate a sequence

of samples $(p_i, w_\rho(p_i))$ and build an approximation function $\tilde{w}_\rho(\cdot)$ using any class of continuous functions with enough representation power, such as polynomial functions or neural networks.

We consider two generic types of approximation schemes for building the mapping $\tilde{w}_\rho(\cdot)$: *(i)* uniform approximations, and *(ii)* high-probability approximations. Uniform approximation schemes satisfy a uniform error bound, as formalized below in Assumption 3, and can be achieved by an interpolation method such as the Bernstein polynomial method as described in Section 4.2.1. Note that, for each $j = 1, \ldots, K$ and $p \in \Delta_K$, we use the notation $w_{\rho,j}(p)$ and $\tilde{w}_{\rho,j}(p)$ to refer to the j^{th} coordinates of $w_\rho(p)$ and $\tilde{w}_\rho(p)$, respectively.

Assumption 3 **(Uniform Error Bound)**. *For each $j = 1, \ldots, K$, there exists a constant $\mathcal{E}_j^{\text{unif}} \geq 0$ such that the approximate optimal solution mapping $\tilde{w}_\rho(\cdot) : \Delta_K \to S$ satisfies:*

$$|w_{\rho,j}(p) - \tilde{w}_{\rho,j}(p)| \;\leq\; \mathcal{E}_j^{\text{unif}}, \qquad \forall p \in \Delta_K.$$

The uniform error bound in Assumption 3 provides guarantees for the approximation error over all probability vectors from the simplex Δ_K. There are two main drawbacks that apply to all known approaches for achieving a uniform error bound. First, achieving a tight uniform error bound requires exact or near-exact computations of the optimal solution mapping $w_\rho(p)$ for all $p \in \Delta_K$. In practice, we may only have an approximate optimal solution mapping available. Second, the sample size required by a method that achieves Assumption 3, for example an interpolations scheme, may be prohibitively large. For these reasons we are motivated to consider a high-probability error bound, which would hold for the more realistic approach of using a regression method, possibly with noise in the output of $w_\rho(\cdot)$, to fit the approximate optimal solution mapping. We consider a generic approach that uses a hypothesis class $\mathcal{G}$ for the approximate optimal solution mappings. Assumption 4 below formalizes our high-probability error bound, which holds for a wide range of regression methods including, for example, the polynomial kernel regression method considered in Section 4.2.2. In Assumption 4, we work with a *reference distribution* $\mathcal{D}_p$ on Δ_K that we use to generate samples $\{p_i\}_{i=1}^m$ to feed into a regression method. In addition, for any $f \in \mathcal{H}$, we later use the notation $\mathcal{D}_{f(x)}$ to refer to the distribution on Δ_K induced by the marginal distribution $\mathcal{D}_x$ of $x \in \mathcal{X}$.

Assumption 4 **(High-probability Error Bound)**. *Let $\mathcal{G}$ be a family of candidate approximate optimal solution mappings whereby $\tilde{w}_\rho(\cdot) : \Delta_K \to S$ for all $\tilde{w}_\rho(\cdot) \in \mathcal{G}$. For each $j = 1, \ldots, K$, there exists a function $\mathcal{E}_j^{\text{prob}}(\cdot, \cdot; \mathcal{G}) : \mathbb{N} \times [0, 1) \to [0, \infty)$ such that, for any distribution $\mathcal{D}_p$ on Δ_K and for any $\delta \in (0, 1]$, with probability at least $1 - \delta$ over m independent samples drawn from $\mathcal{D}_p$ with empirical distribution $\hat{\mathcal{D}}_p^m$, it holds for all $\tilde{w}_\rho(\cdot) \in \mathcal{G}$ that:*

$$\left| \mathbb{E}_{\mathcal{D}_p}[|w_{\rho,j}(p) - \tilde{w}_{\rho,j}(p)|] - \mathbb{E}_{\hat{\mathcal{D}}_p^m}[|w_{\rho,j}(p) - \tilde{w}_{\rho,j}(p)|] \right| \;\leq\; \mathcal{E}_j^{\text{prob}}(m, \delta; \mathcal{G}).$$

When using an approximate optimal solution mapping with either a uniform or a high-probability error bound guarantee, a natural question is: do the performance guarantees of the ICEO approach developed in Section 3 extend to problem (Approx-ICEO-ρ)? We now answer this question affirmatively by extending the generalization bounds of Theorem 3 to situations with approximate mappings satisfying either Assumption 3 or Assumption 4. We make an implicit assumption that, after solving problem (Approx-ICEO-ρ), the decision-maker uses the correct optimal solution mapping $w_\rho(\cdot)$ to make decisions. Therefore, the left hand side of our bounds involve the true risk $R(w_\rho \circ f)$ with the correct mapping while the right hand sides involve the empirical risk $\hat{R}_n(\tilde{w}_\rho \circ f)$ with the approximation (and the reguarlized version thereof).

THEOREM 4. *Suppose Assumption 2 holds and that the hypothesis class $\mathcal{H}$ has bounded multivariate Rademacher complexity $\mathfrak{R}_n(\mathcal{H})$. Then, for any $\delta \in (0,1]$ and $\rho_n > 0$, we have the following:*

(i) If the approximate optimal solution mapping $\tilde{w}_\rho(\cdot)$ satisfies the uniform error bound as stated in Assumption 3, then with probability at least $1-\delta$ over i.i.d. data $\{(x_i,\xi_i)\}_{i=1}^n$ drawn from the distribution $\mathcal{D}$, the following inequalities hold for all $f \in \mathcal{H}$:

$$R(w_{\rho_n} \circ f) \leq \hat{R}_n(\tilde{w}_{\rho_n} \circ f) + \frac{\sqrt{2}L_c^2}{\rho_n}\mathfrak{R}_n(\mathcal{H}) + \bar{c}\sqrt{\frac{\log(\frac{1}{\delta})}{2n}} + L_c\sum_{j=1}^d \mathcal{E}_j^{\text{unif}} \tag{23}$$

$$\leq \hat{R}_n(\tilde{w}_{\rho_n} \circ f; \rho_n) + \frac{\sqrt{2}L_c^2}{\rho_n}\mathfrak{R}_n(\mathcal{H}) + \bar{c}\sqrt{\frac{\log(\frac{1}{\delta})}{2n}} + L_c\sum_{j=1}^d \mathcal{E}_j^{\text{unif}} \tag{24}$$

(ii) If the approximate optimal solution mapping $\tilde{w}_\rho(\cdot)$ comes from a family $\mathcal{G}$ satisfying the high probability error bound as stated in Assumption 4, then with probability at least $1-\delta$ over i.i.d. data $\{(x_i,\xi_i)\}_{i=1}^n$ drawn from the distribution $\mathcal{D}$ and over m independent samples $\{p_i\}_{i=1}^m$ drawn from a reference distribution $\mathcal{D}_p$ on Δ_K, the following inequalities hold for all $f \in \mathcal{H}$:

$$R(w_{\rho_n} \circ f) \leq \hat{R}_n(\tilde{w}_{\rho_n} \circ f) + L_c\sum_{j=1}^d \left[\frac{1}{m}\sum_{i=1}^m |w_{\rho,j}(p_i) - \tilde{w}_{\rho,j}(p_i)| + \mathcal{E}_j^{\text{prob}}(n,\delta/2d;\mathcal{G}) + \mathcal{E}_j^{\text{prob}}(m,\delta/2d;\mathcal{G})\right]$$
$$+ \frac{\sqrt{2}L_c^2}{\rho}\mathfrak{R}_n(\mathcal{H}) + \text{diam}(S)L_c\text{TV}(\mathcal{D}_{f(x)},\mathcal{D}_p) + \bar{c}\sqrt{\frac{\log(\frac{1}{\delta})}{2n}} \tag{25}$$

$$\leq \hat{R}_n(\tilde{w}_{\rho_n} \circ f; \rho_n) + L_c\sum_{j=1}^d \left[\frac{1}{m}\sum_{i=1}^m |w_{\rho,j}(p_i) - \tilde{w}_{\rho,j}(p_i)| + \mathcal{E}_j^{\text{prob}}(n,\delta/2d;\mathcal{G}) + \mathcal{E}_j^{\text{prob}}(m,\delta/2d;\mathcal{G})\right]$$
$$+ \frac{\sqrt{2}L_c^2}{\rho}\mathfrak{R}_n(\mathcal{H}) + \text{diam}(S)L_c\text{TV}(\mathcal{D}_{f(x)},\mathcal{D}_p) + \bar{c}\sqrt{\frac{\log(\frac{1}{\delta})}{2n}} \tag{26}$$

Proof of Theorem 4 By Theorem 3 (i), for any ρ, we have

$$R(w_\rho \circ f; \rho) \leq \hat{R}_n(w_\rho \circ f; \rho) + \frac{\sqrt{2}L_c^2}{\rho}\mathfrak{R}_n(\mathcal{H}) + \bar{c}\sqrt{\frac{\log(\frac{1}{\delta})}{2n}}.$$

Noted that

$$\frac{1}{n}\sum_{i=1}^{n}|c(w_\rho(f(x_i)),\xi_i)-c(\tilde{w}_\rho(f(x_i)),\xi_i)|\le L_c\frac{1}{n}\sum_{i=1}^{n}\|w_\rho(f(x_i)))-\tilde{w}_\rho(f(x_i))\|_1 \tag{27}$$

$$=L_c\sum_{j=1}^{d}\frac{1}{n}\sum_{i=1}^{n}|w_{\rho,j}(f(x_i))-\tilde{w}_{\rho,j}(f(x_i))| \tag{28}$$

When the approximated oracle has a uniform error, we combine (27) and Assumption 3 to have

$$\frac{1}{n}\sum_{i=1}^{n}|c(w_\rho(f(x_i)),\xi_i)-c(\tilde{w}_\rho(f(x_i)),\xi_i)|\le L_c\sum_{j=1}^{d}\mathcal{E}_j^{\text{unif}}. \tag{29}$$

When the oracle is noised, we consider two different distributions $\mathcal{D}_{f(x)}$ and $\mathcal{D}_p$. We let $\mathcal{D}_{f(x)}$ denote the distribution of $f(x)$ given a hypothesis f and the distribution of x, $\mathcal{D}_x$. Moreover, we let $\mathcal{D}_p$ denote the distribution used to generate training samples $\{(p_i,w_i)\}_{i=1}^m$ for oracle approximation. Then, we apply the error bound (4) with distribution $\mathcal{D}_{f(x)}$ and $\mathcal{D}_p$ respectively and have

$$\frac{1}{n}\sum_{i=1}^{n}|w_{\rho,j}(f(x_i))-\tilde{w}_{\rho,j}(f(x_i))|\le\mathbb{E}_{\mathcal{D}_{f(x)}}[|w_{\rho,j}(p))-\tilde{w}_{\rho,j}(p)|]+\mathcal{E}_j^{\text{prob}}(n,\delta/2d;\mathcal{G}),$$
$$\text{w.p.}\quad 1-\delta/2d,$$

and

$$\mathbb{E}_{\mathcal{D}_p}[|w_{\rho,j}(p))-\tilde{w}_{\rho,j}(p)|]\le\frac{1}{m}\sum_{i=1}^{m}|w_{\rho,j}(p_i))-\tilde{w}_{\rho,j}(p_i)|+\mathcal{E}_j^{\text{prob}}(m,\delta/2d;\mathcal{G}),$$
$$\text{w.p.}\quad 1-\delta/2d.$$

Considering the total variation between $\mathcal{D}_{f(x)}$ and $\mathcal{D}_p$, we have the following

$$\mathbb{E}_{\mathcal{D}_{f(x)}}[|w_{\rho,j}(p))-\tilde{w}_{\rho,j}(p)|]\le\mathbb{E}_{\mathcal{D}_p}[|w_{\rho,j}(p))-\tilde{w}_{\rho,j}(p)|]+\text{diam}_j(S)\text{TV}(\mathcal{D}_{f(x)},\mathcal{D}_p),$$

where $\text{diam}_j(S)$ is the diameter of S in the j-th coordinate and TV denotes the total variation. This result holds because $|w_{\rho,j}(\cdot)-\tilde{w}_{\rho,j}(\cdot)|$ is continuous and bounded by $\text{diam}_j(S)$. Thus,

$$\frac{1}{n}\sum_{i=1}^{n}\|w_\rho(f(x_i)))-\tilde{w}_\rho(f(x_i))\|_1\le\sum_{j=1}^{d}\left[\frac{1}{m}\sum_{i=1}^{m}|w_{\rho,j}(p_i)-\tilde{w}_{\rho,j}(p_i)|+\mathcal{E}_j^{\text{prob}}(n,\delta/2d;\mathcal{G})+\mathcal{E}_j^{\text{prob}}(m/2d,\delta;\mathcal{G})\right]$$
$$+\text{diam}_j(S)\text{TV}(\mathcal{D}_{f(x)},\mathcal{D}_p)\quad\text{w.p.}\quad 1-\delta,$$

so we have

$$(27)\le L_c\sum_{j=1}^{d}\left[\frac{1}{m}\sum_{i=1}^{m}|w_{i,j}-\tilde{w}_{\rho,j}(p_i)|+\mathcal{E}_j^{\text{prob}}(n,\delta/2d;\mathcal{G})+\mathcal{E}_j^{\text{prob}}(m,\delta/2d;\mathcal{G})\right]+\text{diam}(S)L_c\text{TV}(\mathcal{D}_{f(x)},\mathcal{D}_p)$$

with probability at least $1-\delta$. Here we slightly abuse the notation and let $\text{diam}(S)$ denote the summation of coordinate-wise diameter along all coordinates.

Then (23) and (25) follow from combining (29) and (**??**) with Theorem 3. (24) and (26) follow from the non-negativity of the regularization term. □

4.2. Approximate the optimal solution oracle by polynomials

In this section, we provide two examples of using polynomial functions to approximate the optimal solution mapping: *(i)* interpolation using Bernstein polynominals, which satisfies a uniform error bound, and *(ii)* polynomial kernel regression, which satisfies a high-probability error bound.

4.2.1. Bernstein polynomials. One example for approximating the optimal solution mapping is interpolation using Bernstein polynomials, for which we review the definition below.

DEFINITION 1 (**Bernstein approximation (De Klerk et al. (2008))**). For a given function $\omega : \Delta_K \to \mathbb{R}$, the Bernstein approximation with order s, $B_s(\omega) : \Delta_K \to \mathbb{R}$, is defined by:

$$B_s(\omega)(p) \ := \sum_{\alpha \in I(K,s)} \bar{w}\left(\frac{\alpha}{s}\right) \frac{s!}{\alpha!} p^\alpha, \quad \forall p \in \Delta_K, \tag{30}$$

where $I(K,s) := \{\alpha \in \mathbb{N}_0^K \,|\, \sum_{i=1}^K \alpha_i = s\}$, $\alpha! := \Pi_i \alpha_i!$, and $p^\alpha := p_1^{\alpha_1} \cdots p_K^{\alpha_K}$. □

Using Bernstein polynomials, based on a result of De Klerk et al. (2008), we can achieve a uniform bound of the approximation error as described in Assumption 3.

PROPOSITION 2. *For a given $\rho > 0$, suppose that we use the Bernstein approximation method (Definition 1) applied separately to each coordinate function $w_{\rho,j}(\cdot)$ to construct an approximate optimal solution mapping $\tilde{w}_\rho(\cdot)$. Then, $\tilde{w}_\rho(\cdot)$ satisfies the uniform error bound in Assumption 3 with*

$$\mathcal{E}_j^{\text{unif}} \ = \ \frac{\Omega L_c}{\rho\sqrt{s}},$$

where $\Omega > 0$ is an absolute constant.

Proof of Proposition 2 This result directly follows from Theorem 3.2 in De Klerk et al. (2008) together with the Lipschitz property from Proposition 1. □

Given the result in Proposition 2, we can immediately obtain a generalization bound for the Bernstein approximation method by applying item *(i)* of Theorem 4. While the Bernstein polynomial method provides a strong uniform error bound guarantee, there is a significant drawback in the number of samples required to obtain this bound. Indeed, to accomplish this approximation, it involves knowing function values of $w_{\rho,j}(\cdot)$ on the grid $\Delta_{K,s} := \{w \in \Delta_K : sw \in \mathbb{N}_0^K\}$ which has $\binom{K+s}{K}$ many points in total. As such, the number of calculations of $w_\rho(\cdot)$ may be prohibitively large, which motivates the use of regression methods.

4.2.2. Polynomial kernel regression. In this section, we consider using the less computationally prohibitive regression methods that lead to high-probability bounds as in Assumption 4. As an exemplary case, we consider the polynomial kernel regression method. In this setting, we allow for the possibility of a "noised oracle" whereby the optimal solution mapping is not

computed exactly. Specifically, the noised oracle outputs $w_\rho(p) + \sigma\varepsilon$ instead of $w_\rho(p)$, where ε is a d-dimensional standard Gaussian random vector and σ is a scalar that represents the standard deviation of the noise. The approximate oracle $\tilde{w}_\rho(\cdot)$ is constructed on independent samples $\{(p_i, w_i)\}_{i=1}^m$ where p_i is drawn from the reference distribution $\mathcal{D}_p$ and w_i is computed from the noised oracle. That is, we assume that $w_i = w_\rho(p_i) + \sigma\varepsilon_i$ for $\{\varepsilon_i\}_{i=1}^m$ that are i.i.d. realizations of Gaussian random variables. These samples can be achieved by first generating $\{p_i\}_{i=1}^m$ randomly from following any user-chosen distribution $\mathcal{D}_p$ over the simplex Δ_K and then calculating $\{w_i\}_{i=1}^m$ from a (possibly randomized) algorithm for approximating $w_\rho(\cdot)$. Note that we do assume that the noise is Gaussian, which may be a reasonable assumption for some algorithmic schemes for approximating $w_\rho(\cdot)$.

The approximate optimal solution mapping is learned using polynomial kernels $k(p, p') = (c + p^T p')^s$, where $s \in \mathbb{N}$ is the degree parameter. In the remaining part of this section, we let $\mathcal{G}$ denote a function class induced by a polynomial kernel of degree s and let $\|\cdot\|_\mathcal{G}$ denote any norm defined on $\mathcal{G}$. Note that $\mathcal{G}$ is a convex, star-shaped function class Wainwright (2019). For the function class $\mathcal{G}$ and a given sample $\{p_i\}_{i=1}^m$, let $\tau_j(\mathcal{G}, \{p_i\}_{i=1}^m, r) := \inf_{u\in\mathcal{G}:\|u\|_\mathcal{G}\le r}(\frac{1}{m}\sum_{i=1}^m (u(p_i) - w_{\rho,j}(p_i))^2)^{1/2}$ denote the fitting ability for $w_{\rho,j}$ using the kernel function class $\mathcal{G}$ within a user-defined radius r. Given the function class $\mathcal{G}$ and a given sample $\{(p_i, w_i)\}_{i=1}^m$, the method of kernel ridge regression estimates the approximate optimal solution mapping $\tilde{w}_\rho(\cdot)$ by solving:

$$\min_{u\in\mathcal{G}:\|u\|_\mathcal{G}\le r} \frac{1}{m}\sum_{i=1}^m (u(p_i) - w_{i,j})^2 \tag{31}$$

The corresponding high-probability approximation error bound for learning the noised oracle using polynomial kernel ridge regression.

PROPOSITION 3. *Let $\mathcal{G}$ denote a function class induced by a polynomial kernel of degree s, suppose that the noise of the output has standard deviation σ, and that we construct the approximate solution mapping $\tilde{w}_\rho(\cdot)$ using kernel ridge regression* (31) *with a user-defined radius $r > 0$. Then, there exist absolute constants $\bar{c}, \bar{c}'$ such that for all $\delta_m \ge \bar{c}_0 \frac{\sigma}{r}\frac{(s-1+K)!}{(s-1)!K!}\frac{1}{m}$, we have*

$$\frac{1}{m}\sum_{i=1}^m (\tilde{w}_{\rho,j}(p_i) - w_{\rho,j}(p_i))^2 \;\le\; \bar{c}_1(\bar{c}_1'\tau_j(\mathcal{G}, \{p_i\}_{i=1}^m, r) + r^2\delta_m^2),$$

with probability at least $1 - \bar{c}_2\exp(-\bar{c}_2'\frac{mr^2}{\sigma^2}\delta_m^2)$ for each coordinate $j = 1, \ldots, K$. Moreover, for any θ_m that satisfies $\theta_m \ge \bar{c}_3\sqrt{\frac{1}{m}\frac{(s-1+K)!}{(s-1)!K!}}$, if it also holds that $m\theta_m^2 \ge \bar{c}_0\log(4\log(\frac{1}{\theta_m}))$, then

$$\left|\mathbb{E}_p[(\tilde{w}_{\rho,j}(p) - w_{\rho,j}(p))^2]^{1/2} - \left(\frac{1}{m}\sum_{i=1}^m (\tilde{w}_{\rho,j}(p_i) - w_{\rho,j}(p_i))^2\right)^{1/2}\right| \;\le\; \bar{c}_3 r^2\theta_m$$

with probability at least $1 - \bar{c}_4\exp(-\bar{c}_4'\frac{m\theta_m^2}{r^2})$ for each coordinate $j = 1, \ldots, K$.

The main main body of this proof is a generalization of the result in Example 13.19 of Wainwright (2019).

Proof of Proposition 3 Considering the kernel function $\mathcal{K}(p,p') = (c + p^T p')^s$ with $p \in \mathbb{R}^K$, we first generalize the result of Example 13.19 from Wainwright (2019). When the input p and p' are K-dimensional vectors, the empirical kernel matrix can have rank at most $\frac{(s-1+K)!}{(s-1)!K!}$. Therefore, the left-hand side of Inequality (13.56) from Wainwright (2019) can be upper-bounded by $\delta_m \sqrt{\frac{1}{m}\frac{(s-1+K)!}{(s-1)!K!}}$. Then we can apply Theorem 13.17 from Wainwright (2019) and set $\lambda_m = 2\delta_m^2$ to achieve inequality (3). Moreover, the empirical Rademacher complexity can be upper-bounded by $\bar{c}\sqrt{\frac{1}{m}\frac{(s-1+K)!}{(s-1)!K!}}$ with some constant $\bar{c}$. Then if we have $\theta_m \geq \bar{c}_3 b \sqrt{\frac{1}{m}\frac{(s-1+K)!}{(s-1)!K!}}$, we can apply Theorem 14.1 from Wainwright (2019) and therefore have the desired result inequality (3). □

Finally, we have the corresponding generalization bound in the following corollary.

COROLLARY 1. *Suppose Assumption 2 holds and that the hypothesis class $\mathcal{H}$ has bounded multivariate Rademacher complexity $\mathfrak{R}_n(\mathcal{H})$. Suppose further that we employ kernel ridge regression* (31) *using a function class $\mathcal{G}$ induced by a polynomial kernel of degree s under the same conditions as in Proposition 3. Then, for any $\delta \in (0,1]$ and $\rho_n > 0$, the following inequalities hold for all $f \in \mathcal{H}$:*

$$\begin{aligned} R(w_{\rho_n} \circ f) \;\leq\; & \hat{R}_n(\tilde{w}_{\rho_n} \circ f) + L_c \sum_{j=1}^{d} [\bar{c}_3 r^2(\theta_m + \theta_n) + \tau_j(\mathcal{G}, \{p_i\}_{i=1}^m, r) + \bar{c}_1' r \delta_m] \\ & + \frac{\sqrt{2} L_c^2}{\rho_n} \mathfrak{R}_n(\mathcal{H}) + L_c \bar{w}_j \sqrt{2\mathrm{TV}(\mathcal{D}_{f(x)}, \mathcal{D}_p)} + \bar{c}\sqrt{\frac{\log(\frac{1}{\delta})}{2n}} \qquad (32) \\ \leq\; & \hat{R}_n(\tilde{w}_{\rho_n} \circ f; \rho_n) + L_c \sum_{j=1}^{d} [\bar{c}_3 r^2(\theta_m + \theta_n) + \tau_j(\mathcal{G}, \{p_i\}_{i=1}^m, r) + \bar{c}_1' r \delta_m] \\ & + \frac{\sqrt{2} L_c^2}{\rho_n} \mathfrak{R}_n(\mathcal{H}) + L_c \bar{w}_j \sqrt{2\mathrm{TV}(\mathcal{D}_{f(x)}, \mathcal{D}_p)} + \bar{c}\sqrt{\frac{\log(\frac{1}{\delta})}{2n}} \qquad (33) \end{aligned}$$

with probability at least $1 - \delta'$ over i.i.d. data $\{(x_i, \xi_i)\}_{i=1}^n$ drawn from the distribution $\mathcal{D}$ and over m independent samples $\{(p_i, w_i)\}_{i=1}^m$, where $\delta' = \delta + \bar{c}_2 \exp(-\bar{c}_2' \frac{m r^2}{\sigma^2} \delta_m^2) + \bar{c}_4(\exp(-\bar{c}_4' \frac{m\theta_m^2}{r^2}) + \exp(-\bar{c}_4' \frac{n\theta_n^2}{r^2}))$ and δ_m, θ_m, and θ_n are chosen to satisfy the conditions in Proposition 3.

Proof of Corollary 1 The proof follow from a slight modification of the proof of Theorem 4. We first consider

$$\frac{1}{n}\sum_{i=1}^{n} \|w_\rho(f(x_i))) - \tilde{w}_\rho(f(x_i))\|_1 \leq L_c \sum_{j=1}^{d} (\frac{1}{n}\sum_{i=1}^{n} |w_{\rho,j}(f(x_i)) - \tilde{w}_{\rho,j}(f(x_i))|^2)^{\frac{1}{2}}.$$

Then noted that

$$\mathbb{E}_{\mathcal{D}_{f(x)}}[|w_{\rho,j}(p)) - \tilde{w}_{\rho,j}(p)|^2] \leq \mathbb{E}_{\mathcal{D}_p}[|w_{\rho,j}(p)) - \tilde{w}_{\rho,j}(p)|^2] + 2\bar{w}_j^2 \mathrm{TV}(\mathcal{D}_{f(x)}, \mathcal{D}_p),$$

because $|w_{\rho,j}(p)) - \tilde{w}_{\rho,j}(p)|^2$ is bounded by $\bar{w}_j^2$ for all $p \in \Delta_K$. Thus,

$$\mathbb{E}_{\mathcal{D}_{f(x)}}[|w_{\rho,j}(p)) - \tilde{w}_{\rho,j}(p)|^2]^{1/2} \leq \mathbb{E}_{\mathcal{D}_p}[|w_{\rho,j}(p)) - \tilde{w}_{\rho,j}(p)|^2]^{1/2} + \bar{w}_j\sqrt{2\mathrm{TV}(\mathcal{D}_{f(x)}, \mathcal{D}_p)}.$$

Following the same reasoning of the proof in Theorem 4, the desired result follows. □

4.3. Computational Methods for the Semi-algebraic Case

In this section, we present an approach based on polynomial optimization in the case where the objective of the downstream optimization problem is semi-algebraic and we use a linear hypothesis class. In this case, when we additionally use a polynomial approximation $\tilde{w}_\rho(\cdot)$, we can reformulate the approximate ICEO formulation (Approx-ICEO-ρ) as a polynomial optimization problem, which can be solved with a hierarchy of semi-definite optimizaiton problems. Specifically we assume that both c and ϕ are semi-algebraic functions and we consider the linear hypothesis class $\mathcal{H} = \{f(x) : f(x) = Bx + b, (B, b) \in \mathcal{B}\}$ where $\mathcal{B}(\mathcal{X}) = \{(B, b) \in \mathbb{R}^{K\times p} \times \mathbb{R}^K : f(x) \in \Delta_K \ \forall x \in \mathcal{X}\}$ ensures that the output of the hypothesis returns a feasible probability vector. In this section, we demonstrate an exact solution method for the semi-algebraic case by transforming the (Approx-ICEO-ρ) to a polynomial optimization program. Before we reach the reformulated problem, we first review the definitions of semi-algebraic sets and semi-algebraic functions.

Definition 2 (**Semi-algebraic set Lasserre (2015)**). $K \subset \mathbb{R}^n$ is a basic semi-algebraic set if

$$K = \{x \in \mathbb{R}^n : g_j(x) \geq 0, j = 1, \ldots, m\} \tag{34}$$

for some polynomial functions $(g_j)_{j=1}^m$, i.e., $(g_j)_{j=1}^m \subset \mathbb{R}[x]$. □

Definition 3 (**Basic semi-algebraic function Lasserre (2015)**). Suppose a function $f : K \to \mathbb{R}^p$, where $K \subseteq \mathbb{R}^n$ is basic semi-algebraic, is in the algebra of functions generated by finitely many of dyadic operations $\{+, \times, \div, \vee, \wedge\}$ and monadic operations $|\cdot|$ and $(\cdot)^{1/q}$, $q \in \mathbb{N}$, on polynomials. We say f is *basic semi-algebraic* (b.s.a.), if there exists $s \in \mathbb{N}$, polynomials $(h_k)_{k=1}^s \subset \mathbb{R}[x, y_1, \ldots, y_p]$ and a basic semi-algebraic set $K_f = \{(x, y) \in \mathbb{R}^{n+p} : x \in K, h_k(x, y) \geq 0\}$ such that the graph of f satisfies $\{(x, f(x)) : x \in K\} = K_f$. □

Note that with the linear hypothesis class, we need an additional constraint $Bx + b \in \Delta_K$, to guarantee that the output $f(x)$ is a valid probability vector for any $x \in \mathcal{X}$. We also assume that $\mathcal{X}$ is a polyhedron, i.e. $\mathcal{X} := \{x \in \mathbb{R}^p : Ax \geq a\}$ for some $A \in \mathbb{R}^{m\times p}$ and $a \in \mathbb{R}^m$. Then the problem Approx-ICEO-ρ becomes:

$$\begin{aligned} \min_{B,b} \quad & \frac{1}{n}\sum_{i=1}^n c(w_i, \xi_i) + \frac{\rho}{2}\phi(w_i) && \text{(Poly-Approx-ICEO-}\rho_n\text{)} \\ \text{s.t.} \quad & w_i = \tilde{w}_\rho(Bx_i + b) \\ & Bx + b \in \Delta_K, \forall x \in \mathcal{X} = \{x \in \mathbb{R}^p : Ax \geq a\} \end{aligned}$$

Note that the approximated oracle $\tilde{w}_\rho(\cdot)$ is constructed by polynomial kernels, so the first group of constraints are polynomial functions. Then we show that the second group of constraints can be reformulated to a group of linear constraints using the following proposition.

PROPOSITION 4. *Suppose* $\mathcal{X} := \{x \in \mathbb{R}^p : Ax \geq a\}$ *for some* $A \in \mathbb{R}^{m \times p}$ *and* $a \in \mathbb{R}^m$*, then the constraint*

$$Bx + b \in \Delta_K, \forall x \in \mathcal{X}$$

can be rewritten as the following group of constraints by introducing new decision variables $y_k \in \mathbb{R}^m, k = 1, \ldots, K$*,* $z, u \in \mathbb{R}^m$

$$\begin{cases} a^T y_k \geq -b_k \quad \forall k = 1, \ldots, K \\ A^T y_k = B_k \quad \forall k = 1, \ldots, K \\ a^T z \geq 1 - \mathbb{1}^T b \\ A^T z = B^T \mathbb{1} \\ a^T u \geq -1 + \mathbb{1}^T b \\ A^T u = -B^T \mathbb{1} \\ y_k, z, u \geq 0 \quad \forall k = 1, \ldots, K \end{cases} \tag{35}$$

We have now shown that problem (Poly-Approx-ICEO-ρ_n) is a problem optimizing a basic semi-algebraic function on a basic semi-algebraic set which, by Proposition 11.10 of Lasserre (2015), can be reformulated as a polynomial optimization problem, which can be solved by solving a hierarchy of semi-definite problems.

5. Numerical Experiments

In this section, we demonstrate the numerical performance of our proposed ICEO framework using synthetic data. We first summarize the benchmark methods that we adopted for comparison:

1. Sample average approximation (SAA). In this benchmark, the decision-maker simply ignores the contextual features then minimizes the average of cost functions using empirical distribution of the observations of the random parameter.
2. The two-step predict-then-optimize (PTO) method. In this benchmark, we estimate the hypothesis $f \in \mathcal{H}$ using a cross-entropy loss function instead of the downstream optimization goal.
3. The prescriptive method (PRES) proposed in Bertsimas and Kallus (2020). We consider KNN-based (PRES-KNN) and kernel-based (PRES-Kernel).
4. The stochastic optimization forest (SO-Forest) proposed in Kallus and Mao (2020).

As for our proposed ICEO method, we find the best hypothesis class by solving (Approx-ICEO-ρ) using gradient-base algorithms. More details of the approximated oracle and the optimization algorithm can be fond in Section 5.1.

5.1. Multi-item Newsvendor Problem

We consider the multi-item newsvendor problem, as in Example 1, with synthetic data. In this setting, we consider $d = 2$, which is the case where the newsvendor jointly decides the order quantities of two products with an overall budget of 50. The decision variable $w \in \mathbb{R}^2$ and random demand $\xi \in \mathbb{R}^2$ are both two-dimensional, corresponding to the order quantity and demand of the two products. The newsvendor aims to minimize the total inventory cost as formulated in (4), with unit overstock costs h_1 and h_2 set to 1 and 1.3 and unit stockout cost b_1 and b_2 set to 9 and 8 for the two products, respectively.

Data Generation Process. The synthetic data is generated in the following manner. The features $x_i \in \mathbb{R}^p$ are generated independently following the multi-variate Gaussian distribution $x_i \sim N(0, MI_p)$ for some constant $M > 0$ and where I_p is an identity matrix. Then we consider $K = 4$ scenarios for the demand $\xi_i \in \mathbb{R}^2$, i.e., $\Xi = \{\tilde{z}_1, \tilde{z}_2, \tilde{z}_3, \tilde{z}_4\}$. Then, the corresponding conditional probability vector of ξ_i is generated according to $\text{soft}((B^* x_i + b^*)^{\text{deg}})$, with $B^* \in \mathbb{R}^{K \times p}$, $b^* \in \mathbb{R}^K$ and deg a positive integer being parameters set before the data generation process. Then, ξ_i takes the value of $\tilde{z}_k$ with probability $p_k^*(x) = \text{soft}((B^* x_i + b^*)^{\text{deg}})_k$ for all $k = 1, \ldots, K$.

Optimal Solution Mapping Approximation. The optimal oracle is approximated using neural networks in the experiment. We first generate a data set $\{(p_i, w_i)\}_{i=1}^m$ by uniformly sampling p_i from the simplex Δ_K and then generating $w_i := w_\rho(p_i)$. Then we train a neural network with one hidden layer to approximate the oracle. The neural network is trained with respect to the mean absolute percentage error (MAPE) loss.

ICEO Hypothesis Learning. In this experiment, we consider two candidate hypothesis classes for $\mathcal{H}$. First, we consider a softmax function composed with a linear function class, whereby $\mathcal{H}_1 := \text{soft} \circ \tilde{\mathcal{H}}_1$ and $\tilde{\mathcal{H}}_1 := \{x \mapsto Bx + b_0 : B \in \mathbb{R}^{K \times p}, b \in \mathbb{R}^K\}$. The other case is $\mathcal{H}_2 := \text{soft} \circ \tilde{\mathcal{H}}_2$ where $\tilde{\mathcal{H}}_2$ denotes a neural network with one-hidden layer. When the degree parameter deg of the data generation process is higher than one, then there is a model misspecification when learning the ICEO hypothesis with $\mathcal{H}_1$ since the true hypothesis $f^*(x) := \text{soft}((Bx_i + b)^{\text{deg}})$ is not in the hypothesis class $\mathcal{H}_1$ when $\text{deg} > 1$. For both hypothesis class, we apply Adam optimization algorithm (Kingma and Ba (2014)) while learning the hypothesis.

Comparison with Benchmarks: Results. In this experiment, we consider $\{\tilde{z}_1 := (33, 15), \tilde{z}_2 := (71, 4), \tilde{z}_3 := (17, 47), \tilde{z}_4 := (4, 43)\}$, $M = 5$, and each element of B is an integer between 0 and 150. We consider regularization coefficient $\rho = 0.01$ and $\text{deg} = 1$. We consider multiple training set sizes $n \in \{100, 300, 500, 700\}$ and for every value of n, we run 25 simulations. We use a validation set to tune the hyper-parameters for KNN, Kernel and SO-forest and the ICEO method. To evaluate out-of-sample performances of all these methods, we generate a test set including 1000 samples in

each simulation. We would like to emphasize that we evaluate the newsvendor cost (4) rather than the ICEO objective with regularization.

Figure 2 demonstrates the performance of the ICEO method and the non-parametric benchmarks. As demonstrated in this plot, the performance of ICEO method outperforms other benchmarks when the sample size is greater than 300. Even when the sample size is small, the performance of the ICEO method is comparable to the best of the benchmark (KNN). When compared with non-parametric prescriptive methods, our method shows the benefit of modeling the underlying conditional distribution.

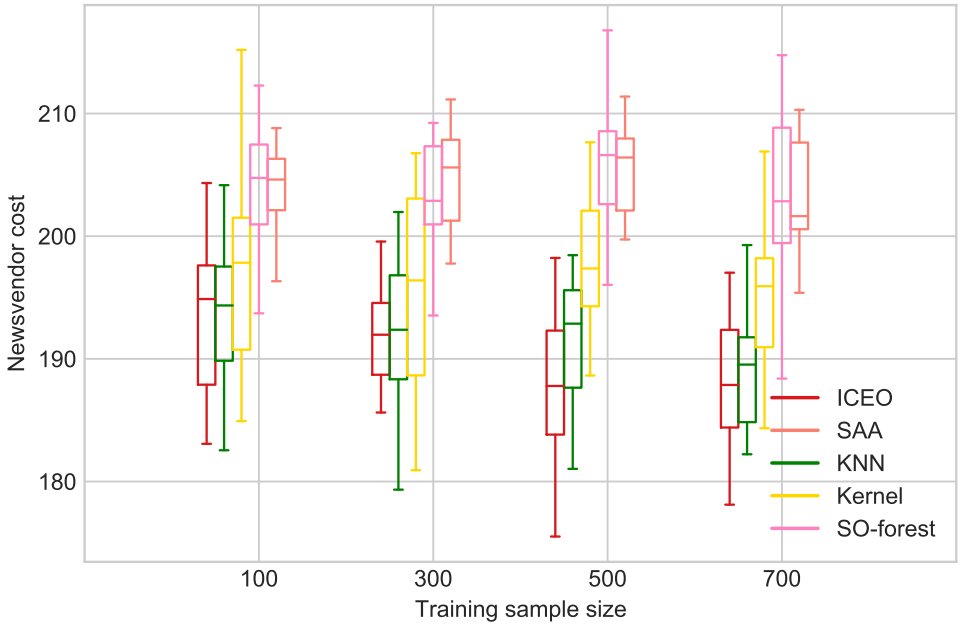
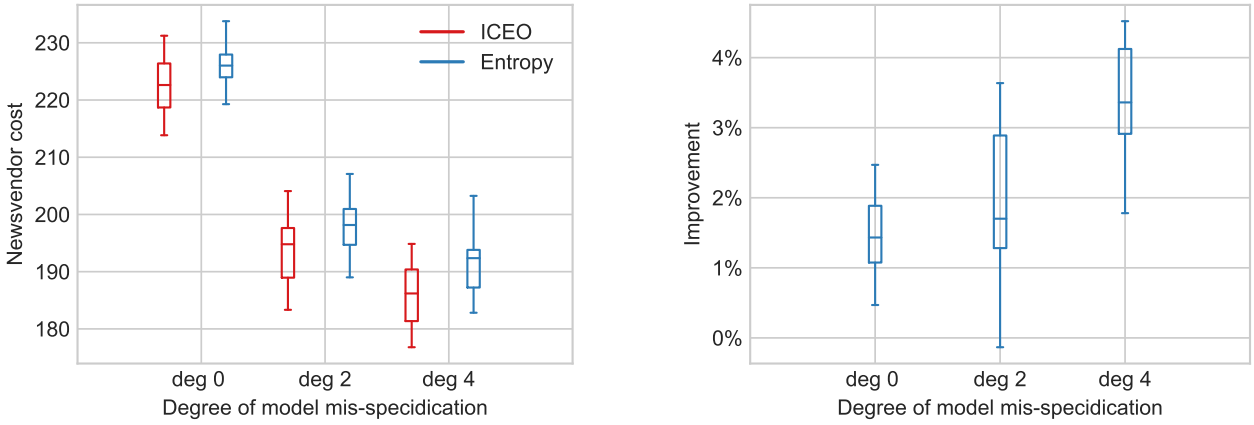


Figure 2 Comparison of ICEO with non-parametric methods

Model Misspecification: Results We also investigate the case of model misspecification when we consider the softmax linear hypothesis class $\mathcal{H}_1$. Since the ICEO method and the PTO (Entropy) methods are the only two methods that involves modeling the underlying conditional distribution using this hypothesis class, we only compare the performance of these two methods. To evaluate the performance of both methods, we consider the newsvendor cost on a test set with size 1000 in each simulation. To better quantify the improvement of ICEO compared to Entropy, we define the *improvement* $\frac{\text{cost(Entropy)}-\text{cost(ICEO)}}{\text{cost(Entropy)}}$. In Figure 3a, we show the performance of the ICEO method compared to the two-step Entropy method and Figure 3b demonstrates the improvement of the ICEO method as compared to the Entropy method. As we can see, under model misspecification, ICEO constantly outperforms two-step Entropy method and the advantage increases when the

degree of model misspecification increases. Both plots demonstrate the advantage of considering ultimate optimization goal while estimating the conditional distribution. Besides, the readers may note a decreasing trend of the out-of-sample cost for both methods in Figure 3a. It is because as the degree of model misspecification increase, the components in the probability vector tends to be binary. In other words, with higher degree if model misspecification, the demand becomes more deterministic, which makes it easier for both methods to learn the conditional demand distribution.



(a) Out-of-sample performance (b) Improvement of ICEO compared to Entropy

Figure 3 Comparison between ICEO and Entropy under model mis-specification

6. Conclusion

In this paper, we propose a new framework for estimating the underlying conditional distribution in contextual stochastic optimization. The proposed ICEO framework uses a flexible hypothesis class and learn the hypothesis by incorporating the downstream optimization goal.

The ICEO framework developed herein applies to the case where the random parameter is a discrete random variable and the nominal optimization problem is convex. To address the issue that the optimal solution oracle may have multiple outputs, we consider an additional strongly convex decision regularization function for both the oracle and the ICEO objective. We then prove that the ICEO method is asymptotically consistent and provide finite-sample analysis in the form of generalization bounds. Moreover, we investigate the non-differentiability of the regularized optimal solution oracle which often leads to computational difficulties in calculating the gradients and poor local minima that are hard to escape. We address this issue by approximating the regularized oracle using differentiable functions. We then provide possible approximation error bounds and the corresponding generalization bounds when using the approximated oracle. For the cases when the nominal optimization problem is semi-algebraic, and when the approximated oracle is constructed

using polynomial functions, the ICEO problem can be reformulated as a polynomial optimization problem and thus solved for the optimal solution up to arbitrary accuracy by solving a hierarchy of semi-definite problems.

Naturally, there are many potential directions to investigate for future work. One possible direction is to generalize the ICEO framework to the infinite dimensional case where the random parameter is a continuous random variable. Besides, one may also investigate the ICEO framework in the high-dimensional setting or when the data is non-stationary.

Acknowledgments

PG acknowledges the support of NSF Awards CCF-1755705 and CMMI-1762744.

References

Agrawal A, Amos B, Barratt S, Boyd S, Diamond S, Kolter JZ (2019) Differentiable convex optimization layers. *Advances in Neural Information Processing Systems*, 9558–9570.

Ahmadi H, Shanbhag UV (2014) Data-driven first-order methods for misspecified convex optimization problems: Global convergence and rate estimates. *53rd IEEE Conference on Decision and Control*, 4228–4233 (IEEE).

Ahuja RK, Magnanti TL, Orlin JB (1988) Network flows .

Amos B, Kolter JZ (2017) Optnet: Differentiable optimization as a layer in neural networks. *International Conference on Machine Learning*, 136–145 (PMLR).

Balghiti OE, Elmachtoub AN, Grigas P, Tewari A (2019) Generalization bounds in the predict-then-optimize framework. *arXiv preprint arXiv:1905.11488* .

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

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

Berthet Q, Blondel M, Teboul O, Cuturi M, Vert JP, Bach F (2020) Learning with differentiable perturbed optimizers. *arXiv preprint arXiv:2002.08676* .

Bertsimas D, Dunn J, Mundru N (2019) Optimal prescriptive trees. *INFORMS Journal on Optimization* 1(2):164–183.

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

Bertsimas D, McCord C (2019) From predictions to prescriptions in multistage optimization problems. *arXiv preprint arXiv:1904.11637* .

Braides A (2006) A handbook of gamma-convergence. *Handbook of Differential Equations: stationary partial differential equations*, volume 3, 101–213 (Elsevier).

Braides A, et al. (2002) *Gamma-convergence for Beginners*, volume 22 (Clarendon Press).

Butler A, Kwon R (2021) Integrating prediction in mean-variance portfolio optimization. *Available at SSRN 3788875* .

Chu LY, Shanthikumar JG, Shen ZJM (2008) Solving operational statistics via a bayesian analysis. *Operations Research Letters* 36(1):110–116.

De Klerk E, Den Hertog D, Elabwabi G (2008) On the complexity of optimization over the standard simplex. *European journal of operational research* 191(3):773–785.

Donti P, Amos B, Kolter JZ (2017) Task-based end-to-end model learning in stochastic optimization. *Advances in Neural Information Processing Systems*, 5484–5494.

Elmachtoub A, Liang JCN, McNellis R (2020) Decision trees for decision-making under the predict-then-optimize framework. *International Conference on Machine Learning*, 2858–2867 (PMLR).

Elmachtoub AN, Grigas P (2021) Smart "predict, then optimize". *Management Science* .

Ferber A, Wilder B, Dilkina B, Tambe M (2020) Mipaal: Mixed integer program as a layer. *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 1504–1511.

Gupta V, Kallus N (2021) Data pooling in stochastic optimization. *Management Science* .

Ho CP, Hanasusanto GA (2019) On data-driven prescriptive analytics with side information: A regularized nadaraya-watson approach. Technical report, Technical report, March.

Ho-Nguyen N, Kılınç-Karzan F (2019) Exploiting problem structure in optimization under uncertainty via online convex optimization. *Mathematical Programming* 177(1):113–147.

Ho-Nguyen N, Kılınç-Karzan F (2020) Risk guarantees for end-to-end prediction and optimization processes. *arXiv preprint arXiv:2012.15046* .

Jiang H, Shanbhag UV (2013) On the solution of stochastic optimization problems in imperfect information regimes. *2013 Winter Simulations Conference (WSC)*, 821–832 (IEEE).

Jiang H, Shanbhag UV (2016) On the solution of stochastic optimization and variational problems in imperfect information regimes. *SIAM Journal on Optimization* 26(4):2394–2429.

Kallus N, Mao X (2020) Stochastic optimization forests. *arXiv preprint arXiv:2008.07473* .

Kao Yh, Roy B, Yan X (2009) Directed regression. *Advances in Neural Information Processing Systems* 22:889–897.

Kingma DP, Ba J (2014) Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* .

Kotary J, Fioretto F, Van Hentenryck P, Wilder B (2021) End-to-end constrained optimization learning: A survey. *arXiv preprint arXiv:2103.16378* .

Lasserre JB (2015) *An introduction to polynomial and semi-algebraic optimization*, volume 52 (Cambridge University Press).

Liu H, Grigas P (2021) Risk bounds and calibration for a smart predict-then-optimize method. *arXiv preprint arXiv:2108.08887* .

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

Mandi J, Guns T (2020) Interior point solving for lp-based prediction+ optimisation. *arXiv preprint arXiv:2010.13943* .

Mandi J, Stuckey PJ, Guns T, et al. (2020) Smart predict-and-optimize for hard combinatorial optimization problems. *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 1603–1610.

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

Nesterov Y (2003) *Introductory lectures on convex optimization: A basic course*, volume 87 (Springer Science & Business Media).

Pogančić MV, Paulus A, Musil V, Martius G, Rolinek M (2019) Differentiation of blackbox combinatorial solvers. *International Conference on Learning Representations*.

Qi M, Shen ZJM, Zheng Z (2020a) Learning newsvendor problem with intertemporal dependence and moderate non-stationarities. *Available at SSRN 3648615* .

Qi M, Shi Y, Qi Y, Ma C, Yuan R, Wu D, Shen ZJM (2020b) A practical end-to-end inventory management model with deep learning. *Available at SSRN 3737780* .

Ramamurthy V, George Shanthikumar J, Shen ZJM (2012) Inventory policy with parametric demand: Operational statistics, linear correction, and regression. *Production and Operations Management* 21(2):291–308.

Rubin H (1956) Uniform convergence of random functions with applications to statistics. *The Annals of Mathematical Statistics* 200–203.

Sundaram RK, et al. (1996) *A first course in optimization theory* (Cambridge university press).

Wainwright MJ (2019) *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48 (Cambridge University Press).

Wilder B, Dilkina B, Tambe M (2019a) Melding the data-decisions pipeline: Decision-focused learning for combinatorial optimization. *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 1658–1665.

Wilder B, Ewing E, Dilkina B, Tambe M (2019b) End to end learning and optimization on graphs. *arXiv preprint arXiv:1905.13732* .

Appendix A: Supplementary Lemmas and Proofs

LEMMA 1. *$W(\cdot,\cdot)$, as defined in is a convex valued upper semi-continuous correspondence, i.e., $W(f,x)$ is a convex set for any fixed $x \in \mathcal{X}$ and $f \in \mathcal{H}$.*

Proof of Lemma 1 The objective function $\sum_{k=1}^{K} f_k(x)c_k(w)$ is a convex function in w for given x and f because $c(w,\tilde{z}_k)$ is convex of w for all $\tilde{z}_k$, $k=1,\ldots,K$. Since S is convex, we can apply the Maximum theorem (see part 1 of Theorem 9.17 in Sundaram et al. (1996)) to achieve the desired conclusion $\square$

LEMMA 2. *For any $\rho > 0$, $w_\rho(\cdot,\cdot)$ is a single-valued correspondence, hence a continuous function in x and f.*

Proof of Lemma 2 From part 2 in Theorem 9.17 of Sundaram et al. (1996), the mapping $w(p):\Delta_k \to S$ is a continuous function of p. As $p = f(x)$ for any $x \in \mathcal{X}$ and $f \in \mathcal{F}$, $w(\cdot,\cdot)$ is a continuous function in x and f. $\square$

Proof of Proposition 4: We first consider the constraint $B^T x + b \geq 0, \forall x, \text{s.t.} \quad Ax \geq a$ is equivalent to

$$0 \leq \min \quad B_k^T x + b_k \tag{36}$$

$$\text{s.t.} \quad Ax \geq a \tag{37}$$

for all $k = 1,\ldots,K$. Then consider the dual of (36)-(37)

$$\begin{aligned} -b_k \leq \max \quad & a^T y_k \\ \text{s.t.} \quad & A^T y_k = B_k \\ & y_k \geq 0 \end{aligned}$$

which reduces to find a feasible solution of the following group of constraints

$$\begin{cases} a^T y_k \geq -b_k \\ A^T y_k = B_k \\ y_k \geq 0 \end{cases} \tag{38}$$

for all $k = 1,\ldots,K$.

Then the normalization constraint $\mathbb{1}^T(Bx+b) \geq 1, \forall x, \text{s.t.} Ax \geq a$ is equivalent to

$$1 \leq \min \quad \mathbb{1}^T Bx + \mathbb{1}^T b \tag{39}$$

$$\text{s.t.} \quad Ax \geq a, \tag{40}$$

similarly by considering the dual problem

$$\begin{aligned} 1 - \mathbb{1}^T b \leq \max \quad & a^T z \\ \text{s.t.} \quad & A^T z = B^T \mathbb{1} \\ & z \geq 0, \end{aligned}$$

which reduces to the following group of constraints

$$\begin{cases} a^T z \geq 1 - \mathbb{1}^T b \\ A^T z = B^T \mathbb{1} \\ z \geq 0. \end{cases} \tag{41}$$

Finally, the constraint $\mathbb{1}^T(Bx+b) \leq 1, \forall x, \text{s.t.} Ax \geq a$ is equivalent to

$$1 \geq \max \quad \mathbb{1}^T Bx + \mathbb{1}^T b \tag{42}$$

$$\text{s.t.} \quad -Ax \leq -a \tag{43}$$

by strong duality, it is equivalent to

$$\begin{aligned} 1 - \mathbb{1}^T b \geq \min \quad & -a^T u \\ \text{s.t.} \quad & -A^T u = B^T \mathbb{1} \\ & u \geq 0 \end{aligned}$$

which reduces to

$$\begin{cases} a^T u \geq -1 + \mathbb{1}^T b \\ A^T u = -B^T \mathbb{1} \\ u \geq 0. \end{cases} \tag{44}$$

□