Integrated Conditional Estimation-Optimization

Meng Qi
SC Johnson College of Business, Cornell University, Ithaca, NY, 14850,
mq56@cornell.edu

Paul Grigas
Department of Industrial Engineering and Operations Research, UC Berkeley, Berkeley, CA, 94720,
pgrigas@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 2022). 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 applies 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. We study 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 consistency in terms of the ICEO risk, induced decisions, and the learned hypothesis. We also provide generalization bounds to quantify the out-of-sample performance when data is limited to a finite sample.

In general, there are fundamental differences between the cases of linear and nonlinear objective functions in CSO problems and the two problem classes require completely different solution methods. As pointed out in the prior literature (for example, Bertsimas and Kallus (2020), Elmachtoub and Grigas (2022), Sadana et al. (2023) and Qi and Shen (2022)), the linear objective assumption implies that point predictions are sufficient to address the CSO problem. However, in the general nonlinear case, one needs to utilize a distributional estimation of the conditional distribution. As such, ICEO seeks a distributional estimation that directly models the relationship between the conditional distribution of the random parameter and the contextual features and then estimates the probabilistic model with a training objective that aligns with the downstream optimization problem. Our methodological and corresponding theoretical contributions are entirely novel and significant.

It is worthwhile to point out that, in the ICEO method, we adopt a strongly convex decision regularization function inside the optimization oracle in order to stabilize the decision and induce uniqueness. Due to the interplay between the downstream optimization oracle, decision regularization, and the model, the overall training objective is more complicated than typical prediction error losses and as such standard uniform convergence and consistency do not directly apply even with i.i.d. samples. Instead, such results can only be obtained after establishing the convergence and Lipschitzness of the regularized oracle. Similarly, finite sample bounds for ICEO are generalization bounds based on multi-variate Rademacher complexity. However, as ICEO adopts a more complicated training objective that incorporates the regularized oracle to take the downstream problem into account, ICEO deviates from standard learning procedures that use simple loss functions. Tackling this difficulty also relies on establishing a Lipschitz property of the regularized oracle, which further translates to the training objective.

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. To the best of our knowledge, among all integrated approaches for general classes of nonlinear CSO problems with convex objective functions, ICEO is the first that provides both asymptotic and finite sample performance guarantees.
2. We prove the asymptotic consistency of the ICEO method when the model is 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. To show asymptotic consistency, standard statistical learning approaches do not apply because of the presence of the regularized oracle. Tackling these difficulties requires establishing certain properties of the regularized oracle such as uniform convergence and uniform Lipschitzness.
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). Again, standard statistical learning results do not directly apply since taking the downstream optimization into account leads to a more complicated loss function than the typical (e.g., MSE) losses. In particular, our generalization bound requires first establishing the Lipschitz property of the regularized optimal decision oracle (Proposition 2).Furthermore, our generalization bound intuitively indicates that ICEO can be more flexible/expressive than policy optimization methods while still achieving a comparable sample (Rademacher) complexity.
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 3, Proposition 4). 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 5).

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 performances of the ICEO method is demonstrated in Section 5. Moreover, Appendix A provides supplementary materials to support the asymptotic consistency result in Section 3.1. B presents supplementary lemmas and proofs for Sections 3.2 and 4. Appendix A.1.1 presents conditions that ensure an "automatic crossover" behavior of the regularized oracle, which further supports the assumptions made in Section 3.1. Appendix C provides two detailed examples of using polynomial functions to approximate the optimal solution mapping. When both the oracle approximation and objective functions are polynomial functions, the ICEO problem can be reformulated as a semi-algebraic problem thus can be solved efficiently. Appendix D demonstrates supplementary materials for Section 5. In Appendix E, we provide more intuition of the advantages of ICEO by comparing it to policy optimization approaches.

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. In one of the first works along these lines, Hannah et al. (2010) use non-parametric methods to estimate the density function conditioned on state variables to solve convex stochastic optimization problems. They consider weights on the empirical distribution based on kernels and the Dirichlet process. Bertsimas and Kallus (2020) also 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. 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. 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. Different from Kallus and Mao (2020) 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.

Ho and Hanasusanto (2019) considers the regularized Nadaraya-Watson approach and establish performance guarantees using moderate deviations theory. We refer to Qi and Shen (2022) for a tutorial of these methods.

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 (2022) 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 strong numerical performance and improved model complexity over standard decision tree methods (e.g., CART) that minimize prediction error.

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.

Other related works include Kao et al. (2009) which investigates the case of model misspecification when features are not perfect. This work propose a method of directed regression which combines the merits of regression and empirical optimization. Later Kao and Van Roy (2012) extended this setting to a directed time series regression. 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. A recent work Poursoltani et al. (2023) investigates the coefficient of prescriptiveness which can be used to quantify the performance of different CSO approaches and how to directly optimize it. Bennouna and Van Parys (2021) took a different perspective on data-driven solutions and investigated the optimal formulation that guaranteeing a certain level of out-of-sample performance.

There are also a number of decision-focused learning contributions that address practically impactful applications. Chung et al. (2022) investigate the problem of allocating limited supply in health supply chains and proposed a decision-aware learning method that uses the decision cost to inform training. Qi et al. (2020b) focus on a multi-period inventory management problem with random demand and lead-times, 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. Cristian et al. (2022) proposes a neural network architecture that approximately solves linear programs in an end-to-end way. The authors also analyze applications of the proposed approach to a multi-warehouse inventory management problem with cross-fulfillment. Chehrazi and Weber (2010) consider the problem of optimal debt settlement and consider an approach that combines estimating the objective function and the optimization problem.

Another stream of work investigates the performance of the classic separated two-step PTO and ETO approach. Hu et al. (2022) demonstrate that, when the optimization problem has a linear objective and linear constraints, the two-step predict-then-optimize approach leads to faster convergence in terms of expected risk compared to the integrated policy optimization approach. Elmachtoub et al. (2023) take a stochastic dominance perspective and demonstrate that when the model class is well-specified, the predict-then-optimize approach outperforms the integrated approach in a strong sense. More broadly, there is a rapidly growing literature regarding data-driven methods for CSO problem. We have mainly included a discussion of work closely related to our paper and refer readers to Sadana et al. (2023) for a comprehensive survey of various approaches to address CSO as well the connection among previously stated streams of literature.

There are also other studies that explore seemingly similar but different problem settings as CSO and ICEO. For example, 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 between CSO and JEO 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 CSO 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 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.

Lastly, there are other works that explore data-driven solutions for stochastic optimization. For example, data-driven methods for robust optimization (Bertsimas et al. (2018a), Wang et al. (2023), Bertsimas et al. (2018b), Hong et al. (2021)) and distributionally robust optimization (DRO) (Delage and Ye (2010), Blanchet et al. (2019), Gao and Kleywegt (2023), Wiesemann et al. (2014), Van Parys et al. (2021)). These lines of research emphasize the robustness of data-driven solutions. Among the data-driven DRO literature, the stream of residual-based DRO is most relevant to our work (Kannan et al. (2020), Qi et al. (2022)). Residual-based DRO considers the conditional distribution based on residuals. A similar approach involving residual-SAA has been studied by Deng and Sen (2022), Liu et al. (2022), Kannan et al. (2022). Besides methodologies that solves the standard CSO, there are other works considering other settings different from the classical CSO problem, for example, the small-data large-scale regime Gupta and Rusmevichientong (2021), Gupta et al. (2022).

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 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.

In this work, in order to directly model the conditional distribution, 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

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

Note that there might be multiple optimal solutions and we use the notation $W(p)$ to refer to the set of such optimal solutions, i.e., $W(p) := \arg\min_{w \in S} \sum_{k=1}^K p_k c_k(w)$.

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 C.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 C.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, compared to the policy learning approaches which learn 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. More detailed intuitions can be found in Appendix E. 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 $\mathcal{H}$.

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. 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 1. These regularity properties induced by the use of the regularization term $\phi(\cdot)$, which is crucial for developing our ICEO methodology and for providing associated theoretical guarantees. We further assume that $w_\rho(p)$ can be efficiently computed in practice for any $p \in \Delta_K$ and $\rho > 0$. Possible examples include using a commercial solver or utilizing 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 1-strongly convex.

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 in-sample cost induced by the regularized oracle:

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

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 $\hat{f}(\cdot)$. We remark that one may consider a variant of ICEO-ρ that, in the objective function, additionally includes decision regularization terms $\rho\phi(w_i)$ for each sample. Although these decision regularization terms appear more aligned with (3), and were included in an earlier version of this paper, they are in fact a purely optional component of our model and we choose to not include them here for simplicity. Indeed, our entire analysis also will carry through with or without the additional decision regularization terms. The only difference is that the rate of convergence in the finite-sample bound is slightly faster when these terms are excluded.

Let us contrast the ICEO approach with the PTO and ETO approaches. 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 $\arg\min_{w \in S} c(w, \hat{g}_{\text{PTO}}(x))$. Note that, as pointed out in Section 1, because of the nonlinearity of the objective function, a point estimate for a prediction of ξ given x does not provide enough information about the conditional distribution to produce a reasonable solution of (1). Different from PTO, 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 $\ell_{\text{ce}}(f(x), \xi = \tilde{z}_k) := -\log(f_k(x))$.

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 $\{w \circ f : f \in \mathcal{H}\}$. For $f, g \in \mathcal{H}$, recall the sup-norm is defined as as $\|f - g\|_\infty := \sup_{x \in \mathcal{X}} |f(x) - g(x)|$. 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 $K-1$-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 $\mathrm{Var}(\sum_{l=1}^{d} w_l \xi_l)$ 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

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

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 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

$$c(w, \xi) = \sum_{l=1}^{d} g_l(w_l, \xi_l)$$

where g_l is a convex function of w_l and g_l can be different for different coordinates. 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:

1. $\hat{R}_n(\pi)$: The empirical risk with respect to a given sample $\{(x_i,\xi_i)\}_{i=1}^n$, i.e.,

$$\hat{R}_n(\pi) := \frac{1}{n}\sum_{i=1}^n c(\pi(x_i), \xi_i).$$

2. $R(\pi)$: The expected regularized risk with respect to the underlying joint distribution $\mathcal{D}$ of x and ξ, i.e.,

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

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

Note that $\hat{R}_n(\cdot)$ is the objective function of (ICEO-ρ). We also use the notation $\mathcal{D}_x$ to refer to the marginal distribution of the features x. 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.,

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

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

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

and we let f_ρ^* denote an optimal hypothesis.

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.,

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

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

3.1. Asymptotic Consistency

We first demonstrate the asymptotic consistency of the ICEO approach. The consistency of ICEO is three-fold: the consistency of the ICEO risk, the consistency of the ICEO decisions, and the consistency of the ICEO hypothesis. The asymptotic consistency results build upon the convergence of the regularized oracle to the original optimization oracle, as stated in Proposition 1.

Proposition 1 **(Convergence of Regularized Oracle)**. *Suppose $w_\rho(\cdot)$ is the regularized optimal solution mapping for any $\rho > 0$, as defined in* (3)*. For any positive sequence $\{\rho_n\}$ that satisfies $\lim_{n\to\infty}\rho_n = 0$, and for any $p \in \Delta^K$, we have $\mathrm{dist}(w_{\rho_n}(p), W(p)) \to 0$ as $n \to \infty$.*

To establish asymptotic consistency, we require the following assumptions concerning the hypothesis class $\mathcal{H}$ and the regularized oracle $w_\rho(\cdot)$. A standard sufficient (but not necessary) condition to ensure uniform convergence and thus asymptotic consistency in learning theory is for the hypothesis class to have a finite bracketing number (some representative references are Vaart and Wellner (2023)). We build upon this natural sufficient condition by introducing the *multivariate bracketing number* of the hypothesis class $\mathcal{H}$. In other words, we present a modest extension of the standard definition of bracketing number for real-valued functions to the multivariate case. First, we define a multivariate bracket of $\mathcal{H}$ relative to any norm $\|\cdot\|$ on $\mathcal{H}$.

Definition 1 (Multivariate ϵ-bracket). Given two functions $l : \mathcal{X} \to \mathbb{R}^K$ and $u : \mathcal{X} \to \mathbb{R}^K$, the bracket $[l,u]$ is the set of all functions $f \in \mathcal{H}$ with $l_k(x) \le f_k(x) \le u_k(x)$ for each coordinate $k = 1,\ldots,K$ and for all $x \in \mathcal{X}$. An ϵ-bracket is a bracket $[l,u]$ with $\|l-u\| < \epsilon$.

Then the multivariate bracketing number can be defined as follows.

Definition 2 (Multivariate bracketing number). The multivariate bracketing number $N_{[]}(\epsilon,\mathcal{H},\|\cdot\|)$ is the minimum number N of ϵ-brackets $[l_1,u_1],\ldots,[l_N,u_N]$ that cover $\mathcal{H}$, i.e., with the property that for all $f \in \mathcal{H}$ there exists $i \in \{1,\ldots,N\}$ such that $f \in [l_i,u_i]$.

Then, the asymptotic consistency of our proposed ICEO method requires that $\mathcal{H}$ be compact and that its multivariate bracketing number is finite. Note that the metric defined on f is the sup-norm, $\|f-g\|_\infty := \sup_{x\in\mathcal{X}} |f(x) - g(x)|$.

Assumption 1. *For the compact hypothesis class $\mathcal{H}$, we assume the following properties:*

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

B. *(Finite bracketing number of $\mathcal{H}$.) The multivariate bracketing number, $N_{[]}(\epsilon, \mathcal{H}, \|\cdot\|_\infty)$, as defined in Definition 2, is finite for any $\epsilon \in (0,1)$.*

C. *(Unique optimal hypotheses.) 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 $x \in \mathcal{X}$. For $\rho > 0$ and for all $f' \neq f^*_\rho$ in $\mathcal{H}$, there exists $\epsilon > 0$ with $\mathbb{P}_x(\|w_\rho(f'(x)) - w_\rho(f^*_\rho(x))\| \geq \epsilon) > 0$.*

Assumption 2 **(Uniform Properties of the Regularized Oracle)**. *For the regularized oracle, $w_\rho(\cdot)$ with $\rho \in (0, \rho_0]$ for some $\rho_0 > 0$, we assume the following properties:*

A. *(Lipschitz in ρ.) For all $p \in \Delta_K$, $\|w_{\rho_1}(p) - w_{\rho_2}(p)\| \leq L_\rho |\rho_1 - \rho_2|$ for all $\rho_1, \rho_2 \in (0, \rho_0]$.*

B. *(Lipschitz in p.) For all $\rho \in (0, \rho_0]$, $\|w_\rho(p) - w_\rho(p)\| \leq L_w \|p_1 - p_2\|$.*

C. *(Unique optimal decisions.) For any $\epsilon > 0$, there exists $\delta > 0$, so that for all $\rho \in (0, \rho_0]$ and $f \in \mathcal{H}$ it holds that*

$$\mathbb{P}_x(\|w_\rho(f(x)) - w_\rho(f^*_\rho(x))\| \geq \epsilon) > 0 \;\Rightarrow\; R(w_\rho \circ f) - R(w_\rho \circ f^*_\rho) \geq \delta.$$

Recall that if $c(\cdot)$ is not strongly convex, there may exist multiple optimal solutions denoted by $W(p)$ for any input probability vector $p \in \Delta_K$. In this case, we assume that the oracle $w(p)$ outputs one element from the set of optimal solutions $W(p)$. In Lemma 2 in the Appendix, we demonstrate that the value of the unregularized optimal risk J^* does not depend on the particular choice of $w(\cdot)$ and that f^* is always the minimizer that achieves J^*. To establish the consistency result of the ICEO method, we further assume that the regularized oracle is uniformly Lipschitz, as presented in Assumptions 2.A-2.B. In Appendix A, we study a reasonable sufficient condition to guarantee these assumptions of uniform Lipschitzness. Namely, we demonstrate that when the underlying contextual stochastic optimization problem (2) satisfies a linear growth condition away from the optimal solution set, then the regularized oracle satisfies an “automatic crossover” property and the uniform Lipschitz assumptions hold. It is worth noting that, being uniformly Lipschitz implies the uniform equicontinuity which allows one to generalizes the pointwise convergence of the oracle (Proposition 1) to uniform convergence.

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 $\rho_n \in (0, \rho_0]$ satisfies $\lim_{n\to\infty} \rho_n = 0$. Then, under Assumptions 1.A - 1.B and Assumptions 2.A - 2.B, 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) Additionally, with Assumptions 1.C and 2.C, $\mathcal{D}_x$*-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.,* $\operatorname{dist}(w_{\rho_n}(\hat{f}^n_{\rho_n}(x)), W(f^*(x))) \to 0$ *with probability 1.*

(iii) Additionally, with Assumptions 1.C and 2.C, the sequence of ICEO hypotheses converges to the true hypothesis, i.e, $\hat{f}^n_{\rho_n} \to f^*$ *with probability 1.*

The proof of Theorem 1 can be found in Appendix A. The idea of this proof is to establish the convergence of J^*_ρ to J^* and the convergence of $\hat{J}^n_\rho$ to J^*_ρ separately. Note that the convergence of $\hat{J}^n_\rho$ to J^*_ρ has to hold uniformly for all ρ.

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 define asymptotic optimality in terms of the decisions reaching the best objective function 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 property of Bertsimas and Kallus (2020).

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 3. *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 3.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 2 **(Lipschitz Properties of $w_\rho(\cdot)$ and $c(\cdot)$).** *Suppose Assumption 3 holds and note that $\mathbb{R}^K_+ := \{p \in \mathbb{R}^K : p_k \geq 0, \forall k = 1, \ldots, K\}$. Then, for any $\rho > 0$, the optimal solution mapping $w_\rho(\cdot)$ is $(\frac{L_c}{\rho})$-Lipschitz:*

$$\|w_\rho(p) - w_\rho(p')\| \leq \frac{L_c}{\rho}\|p - p'\|_2, \qquad \forall p, p' \in \mathbb{R}^K_+ \tag{6}$$

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

$$\|c(w_\rho(p)) - c(w_\rho(p'))\|_\infty \leq \|c(w_\rho(p)) - c(w_\rho(p'))\|_2 \leq \frac{L_c^2}{\rho}\|p - p'\|_2, \qquad \forall p, p' \in \mathbb{R}^K_+ \tag{7}$$

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). A detailed proof can be found in Appendix B.1.

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 $\pi \in \Pi$:*

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

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 3 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 $f \in \mathcal{H}$:*

$$R(w_{\rho_n} \circ f) \ \leq \ \hat{R}_n(w_{\rho_n} \circ f) + \frac{\sqrt{2} L_c^2}{\rho_n} \mathfrak{R}_n(\mathcal{H}) + \bar{c} \sqrt{\frac{\log(\frac{1}{\delta})}{2n}}. \tag{8}$$

Note that the right-hand side of the inequality in Theorem 3 involves the non-regularized empirical risk, which may be evaluated for any $f \in \mathcal{H}$. This is also 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 (8) converge to zero. We now state the proof of Theorem 3. Proof of of Theorem 3 utilizes Proposition 2 and can be found in Appendix B.1

Leveraging Theorem 3, we obtain the following corollary that provides an upper bound for the suboptimality gap, in terms of out-of-sample risk evaluation, for the ICEO minimizer $\hat{f}^n_{\rho_n}$ relative to the out-of-sample risk of the optimal in-class hypothesis and the true hypothesis.

Corollary 1. *Suppose Assumption 3 holds and that the hypothesis class $\mathcal{H}$ has bounded multivariate Rademacher complexity $\mathfrak{R}_n(\mathcal{H})$. Let $f^*_{\mathcal{H}} \in \arg\min_{f\in\mathcal{H}} R(w \circ f)$ denote an optimal in-class hypothesis. 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 inequality holds:*

$$R(w_{\rho_n} \circ \hat{f}^n_{\rho_n}) - R(w_{\rho_n} \circ f^*_{\mathcal{H}}) \;\le\; \frac{\sqrt{2}L_c^2}{\rho_n}\mathfrak{R}_n(\mathcal{H}) + \frac{3\bar{c}}{2}\sqrt{\frac{2\log(\frac{2}{\delta})}{n}}.$$

If, in addition, we have $f^* \in \mathcal{H}$, then we have

$$R(w_{\rho_n} \circ \hat{f}^n_{\rho_n}) - R(w_{\rho_n} \circ f^*) \;\le\; \frac{\sqrt{2}L_c^2}{\rho_n}\mathfrak{R}_n(\mathcal{H}) + \frac{3\bar{c}}{2}\sqrt{\frac{2\log(\frac{2}{\delta})}{n}}.$$

Remark 1. The generalization bound provided in Theorem 3 involves the Rademacher complexity $\mathfrak{R}_n(\mathcal{H})$ of the hypothesis class $\mathcal{H}$, which returns outputs in Δ_K. It is noteworthy that, if $\mathcal{H}$ involves a softmax operator, i.e., $\mathcal{H} = \text{soft} \circ \tilde{\mathcal{H}}$ for some hypothesis class $\tilde{\mathcal{H}}$ outputting in $\mathbb{R}^K$, then it is possible to obtain an identical generalization bound involving the Rademacher complexity $\mathfrak{R}_n(\tilde{\mathcal{H}})$ of the hypothesis class $\tilde{\mathcal{H}}$. Indeed, since the softmax function soft is Lipschitz (Gao and Pavel (2017)) with constant 1, the Lipschitz bound (7) can be extended to a bound involving the outputs of $\tilde{\mathcal{H}}$ with the same (L_c^2/ρ) constant. Alternatively and equivalently, the vector contraction inequality of Maurer (2016) can be extended to a Lipschitz mapping (such as the softmax function) that has both multivariate input and output. This extended vector contraction inequality can be obtained by a straightforward extension of Lemma 7 and Theorem 3 of Maurer (2016) to the multivariate case.

4. Computational Methods

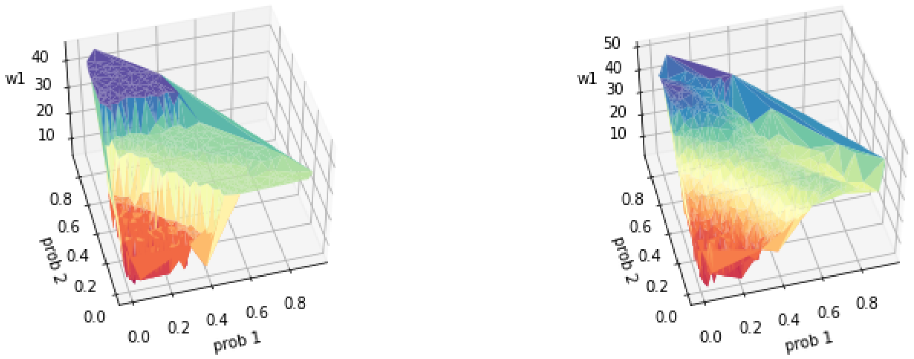
In this section, we discuss the computational difficulties of solving the ICEO formulation (ICEO-ρ) and present multiple approaches to address them.

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 in Appendix B.2. This example shows that, 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 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 2, 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-differentiable 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.

To address the issue of non-differentiability and its consequences leading to poor local optima and slow convergence, 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 enhance



(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.

convergence to a good local optimum. 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. Approximating Optimal Solution Mappings

In this section, we propose a general methodology for approximating the potentially non-differentiable optimal solution mapping. In Appendix C, In we provide two examples of using polynomial functions to approximate the optimal solution mapping.

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:

$$\begin{aligned}\min_{f\in\mathcal{H}} \quad & \frac{1}{n}\sum_{i=1}^{n} c(w_i,\xi_i) \\ \text{s.t.} \quad & w_i = \tilde{w}_\rho(f(x_i))\end{aligned} \tag{Approx-ICEO-ρ}$$

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 4, and can be achieved by an interpolation method such as the Bernstein polynomial method as described in Section C.1. Note that, for each $j = 1, \dots, K$ 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 4 **(Uniform Error Bound)**. *For each $j = 1, \dots, K$, 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 4 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 4, 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 5 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 C.2. In Assumption 5, 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 5 **(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 $j = 1, \dots, K$, 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 4 or Assumption 5. 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 3 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 4, then with probability at least $1-\delta$ over i.i.d. data $\{(x_i,\xi_i)\}_{i=1}^n$ drawn from the distribution $\mathcal{D}$, the following inequalities hold for all $f \in \mathcal{H}$:

$$R(w_{\rho_n} \circ f) \le \hat{R}_n(\tilde{w}_{\rho_n} \circ f) + \frac{\sqrt{2}L_c^2}{\rho_n}\mathfrak{R}_n(\mathcal{H}) + \bar{c}\sqrt{\frac{\log(\frac{1}{\delta})}{2n}} + L_c \sum_{j=1}^d \mathcal{E}_j^{\text{unif}} \tag{9}$$

(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 5, 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 $f \in \mathcal{H}$:

$$\begin{aligned} R(w_{\rho_n} \circ f) \le{} & \hat{R}_n(\tilde{w}_{\rho_n} \circ f) + L_c \sum_{j=1}^d \left[\frac{1}{m}\sum_{i=1}^m |w_{\rho_n,j}(p_i) - \tilde{w}_{\rho_n,j}(p_i)| + \mathcal{E}_j^{\text{prob}}(n,\delta/2d;\mathcal{G}) + \mathcal{E}_j^{\text{prob}}(m,\delta/2d;\mathcal{G}) \right] \\ & + \frac{\sqrt{2}L_c^2}{\rho_n}\mathfrak{R}_n(\mathcal{H}) + \operatorname{diam}(S) L_c \operatorname{TV}(\mathcal{D}_{f(x)}, \mathcal{D}_p) + \bar{c}\sqrt{\frac{\log(\frac{1}{\delta})}{2n}} \end{aligned} \tag{10}$$

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 and minimizes the average of cost functions using the empirical distribution of the observations of the random parameter.
2. The two-step estimate-then-optimize (ETO) method is based on cross-entropy loss (ETO-Entropy). In this benchmark, we estimate the hypothesis $f \in \mathcal{H}$ using the cross-entropy loss function (a standard loss function for multi-class classification) instead of the downstream optimization goal.

3. The prescriptive method (PRES) proposed by Bertsimas and Kallus (2020). We consider the KNN-based (PRES-KNN) and kernel-based (PRES-Kernel) variants.

Moreover, we consider two different types of ICEO methods.

1. Vanilla ICEO (ICEO). ICEO method that solves (ICEO-ρ).
2. The ICEO method regularized by cross-entropy loss (ICEO-Entropy). This is the ICEO method incorporating an additional cross-entropy loss term as a regularization component in conjunction with the objective function described in (ICEO-ρ). Adding the cross-entropy term as a regularization component aligns with practical intuition provided by existing literature (Kao and Van Roy (2012), Elmachtoub and Grigas (2022)).

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, given K scenarios $\Xi = \{\tilde{z}_1, \ldots, \tilde{z}_K\}$, the corresponding conditional probability vector is generated according to a randomly initialized neural network, denoted as f^*_{NN}, composed with the softmax function. The softmax function is implemented by adding a softmax layer as the output layer. Subsequently, ξ_i takes on the value of $\tilde{z}_k$ with a probability of $p^*_k(x) = \text{soft} \circ f^*_{\text{NN}}$ 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 the hypothesis class $\mathcal{H} := \text{soft} \circ \tilde{\mathcal{H}}$ where $\tilde{\mathcal{H}}$ represents a neural network. If $\tilde{\mathcal{H}}$ contains the true hypothesis f^*_{NN}, then $\mathcal{H}$ is well-specified. However, if $\tilde{\mathcal{H}}$ only includes neural networks with fewer layers or hidden nodes, it indicates the case of model misspecification. For both cases of well-specification and misspecification, we employ the Adam optimization algorithm (Kingma and Ba (2014)) to solve (Approx-ICEO-ρ) and learn the hypothesis.

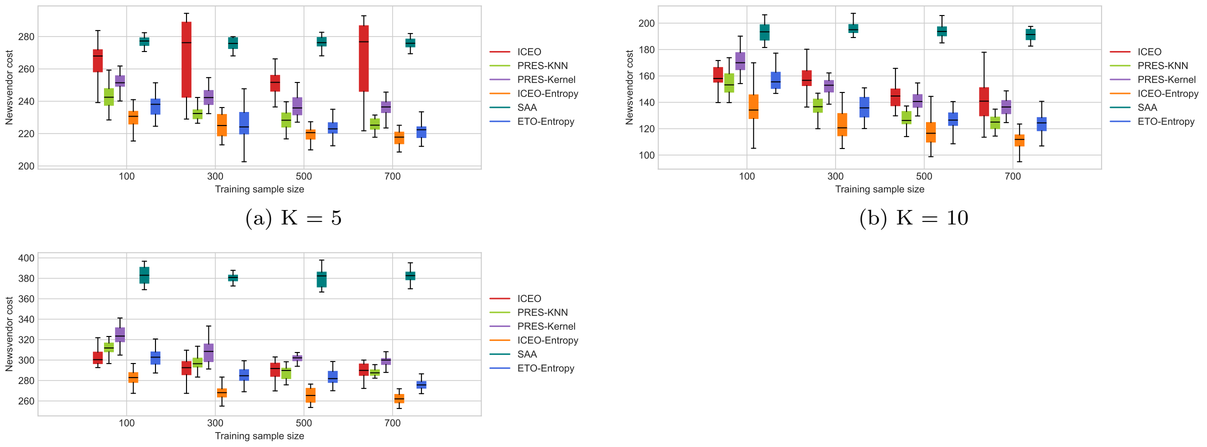
5.1. Multi-item Newsvendor Problem

We consider the multi-item newsvendor problem, as in Example 1, with synthetic data. 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.

Results and Comparisons with Benchmarks. In this experiment, we vary the number of scenarios $K \in 5, 10, 15$. The realizations for each scenario, $\tilde{z}_1, \ldots, \tilde{z}_K$, are generated randomly. Further details on the scenario generation process can be found in Appendix D. We set the regularization coefficient $\rho = 0.01$ and consider multiple training set sizes $n \in \{100, 300, 500, 700\}$. For every value of n, we run 25 simulations. We use a validation set to tune the hyper-parameters for PRES-KNN, PRES-Kernel, ETO-Entropy 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 evaluate performance using the newsvendor cost (4).

Figure 2 illustrates the performance of the ICEO methods and the non-parametric benchmarks across different numbers of scenarios. Each sub-figure demonstrates that the ICEO-Entropy method consistently outperforms all benchmarks, regardless of the training set sizes and numbers of scenarios. Although the vanilla ICEO method sometimes falls short of the benchmarks except for SAA, incorporating the cross-entropy loss in ICEO leads to the superior performance of ICEO-Entropy, which outperforms all benchmarks, even the ETO entropy, in every scenario. A comparison between vanilla ICEO and ICEO-Entropy reveals that ICEO-Entropy outperforms the former. The possible reason might be that incorporating the oracle in the ICEO objective induces a higher level of non-convexity and increases the likelihood of converging to poor local minimum, since the ICEO objective is non-convex in the predicted distribution. By adding a small component of cross-entropy loss, the gradient-based method suffers less from non-convexity while still being influenced by the ICEO objective. Despite this computational issue, the ICEO objective demonstrates the advantages of considering the ultimate optimization goal, as evident when comparing ICEO-Entropy with ETO-Entropy. Moreover, when compared with non-parametric prescriptive methods, the superior performances of the two ICEO methods and the ETO-Entropy method highlight the benefit of modeling the underlying conditional distribution.

Results on Model Misspecification We then examine the performance of ICEO and benchmark methods in the case of model misspecification. When the model is properly specified, the data generation neural network f^*_{NN} consists of two hidden layers, whereas the hypothesis class $\mathcal{H}$ consists of linear models. Since the ICEO methods and ETO-Entropy are the only approaches that utilize this hypothesis class to model the underlying conditional distribution, we compare the performance of these three methods to study the effect of model misspecification. To evaluate their effectiveness, we utilize the newsvendor cost on a test set comprising 1000 samples in each simulation. Figure 3 illustrates the performance of vanilla ICEO and ICEO-Entropy in comparison to the two-step ETO-Entropy method. As we can see, under model misspecification, both ICEO methods consistently outperform the two-step ETO-Entropy approach. This finding demonstrates the advantage of considering the ultimate optimization goal under model misspecification while estimating the conditional distribution.



(a) K = 5

(b) K = 10

(c) K = 15

Figure 2 Comparison of ICEO with benchmark methods for multi-item newsvendor problem.

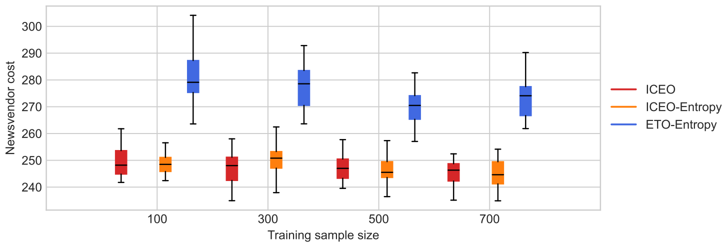


Figure 3 Comparison between ICEO and ETO-Entropy under model misspecification on the multi-item newsvendor problem.

5.2. Quadratic Cost Network Flow Problem

In this part, we consider the minimum cost network flow problem. The problem formulation is as stated in Example 3, where $g_d(w_d, \xi_d) = c_d(w_d - \xi_d)^2$. We consider a simple network demonstrated in Figure 4, where there are two source nodes 1 and 2, and two sink nodes, 3 and 4. The amount of flow that sources out of each of the source nodes 1 and 2 must be no less than a threshold equal to 10. Similarly, the amount of flow that goes into each of the sink nodes 3 and 4 must be at least 10. We let w_1, w_2, w_3, w_4 denote the amount of flow on arcs $(1,3), (1,4), (2,3), (2,4)$, respectively.

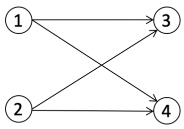
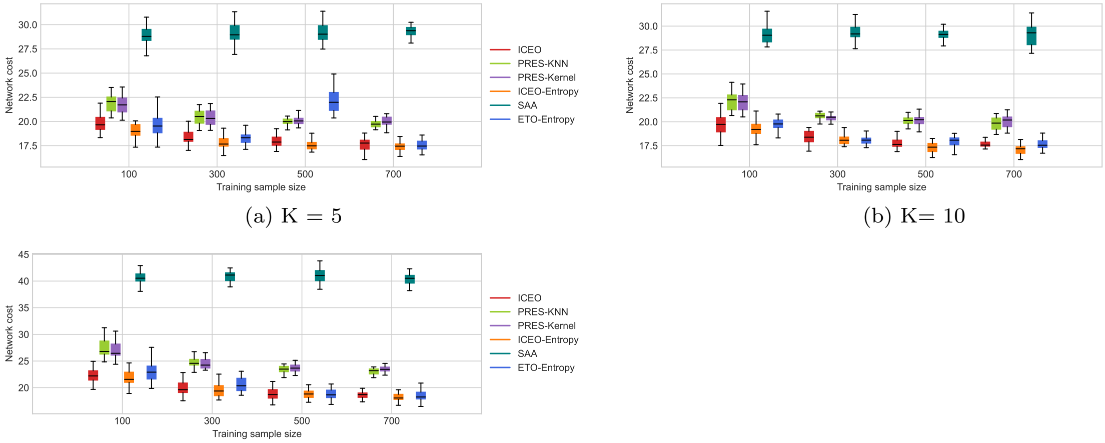


Figure 4 Network graph

Results and Comparisons with Benchmarks. In this experiment, we consider $K \in \{5, 10, 15\}$. The realizations of each scenario, $\tilde{z}_1, \ldots, \tilde{z}_K$, are generated randomly. Again, more details regarding the scenario generation process can be found in Appendix D. Furthermore, the weights of each arc c_1, c_2, c_3, c_4 take the values of $1, 3, 2, 2$ respectively. The regularization coefficient $\rho = 0.01$. As in Section 5.1, we consider multiple training set sizes $n \in \{100, 300, 500, 700\}$, we run 25 simulations for each sample size, and the test set includes 1000 samples in each simulation. To tune the hyper-parameters for ICEO, ETO-Entropy, and the two prescriptive methods, PRES-KNN and PRES-Kernel, we use a validation set including 1000 samples.



(a) K = 5

(b) K= 10

(c) K = 15

Figure 5 Comparison of ICEO with benchmark methods on the multi-item newsvendor problem.

Figure 5 compare the test-set performance of the ICEO method and benchmarks across different numbers of scenarios and sample sizes. Each sub-figure demonstrates that either the vanilla ICEO method or the ICEO-Entropy method outperformns all benchmarks in this experiment. Compared to the best benchmark (PRES-KNN), there is a vague trend that the advantage of ICEO methods are more significant when sample size is limited.

Results on Model Misspecification. As in Section 5.1, we investigate the case of model misspecification. The model misspecification is again introduced by having the data generation neural network f^*_{NN} consists of two hidden layers, whereas the hypothesis class $\mathcal{H}$ consists of linear models. We only compare the two ICEO methods and ETO-Entropy because these are the only approaches that utilize this hypothesis class to model the underlying conditional distribution. Figure 3 illustrates the performance of vanilla ICEO, ICEO-Entropy, and ETO-Entropy methods. It is shown that again, both vanilla ICEO and ICEO-Entropy consistently outperform ETO-Entropy. Moreover, one may observe a possible trend that this advantage of ICEO methods are stronger when there are less training samples. This finding again verifies the advantage of considering the ultimate optimization goal while estimating the conditional distribution.

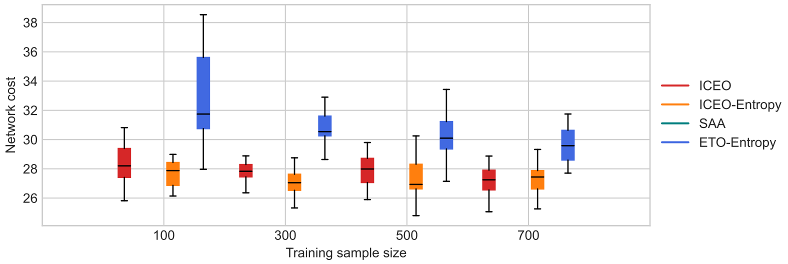


Figure 6 **Comparison between ICEO and ETO-Entropy under model misspecification on the multi-item newsvendor problem.**

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 to learn by incorporating the downstream optimization goal and applies readily to the case where the random parameter is a discrete random variable and the nominal optimization problem is convex. 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 approximation error bounds and the corresponding generalization bounds when using the approximated oracle. Among others, a natural direction for future research is to move further beyond the assumption of finite suport for the random parameter by considering modeling extensions and/or theory.

Acknowledgments

PG acknowledges the support of NSF AI Institute for Advances in Optimization Award 211253.

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.

Bennouna M, Van Parys BP (2021) Learning and decision-making with data: Optimal formulations and phase transitions. *arXiv preprint arXiv:2109.06911* .

Berge C (1877) *Topological spaces: Including a treatment of multi-valued functions, vector spaces and convexity* (Oliver & Boyd).

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, Gupta V, Kallus N (2018a) Data-driven robust optimization. *Mathematical Programming* 167:235–292.

Bertsimas D, Gupta V, Kallus N (2018b) Robust sample average approximation. *Mathematical Programming* 171:217–282.

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* .

Blanchet J, Kang Y, Murthy K (2019) Robust wasserstein profile inference and applications to machine learning. *Journal of Applied Probability* 56(3):830–857.

Chehrazi N, Weber TA (2010) Monotone approximation of decision problems. *Operations Research* 58(4-part-2):1158–1177.

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

Chung TH, Rostami V, Bastani H, Bastani O (2022) Decision-aware learning for optimizing health supply chains. *arXiv preprint arXiv:2211.08507* .

Cristian R, Harsha P, Perakis G, Quanz BL, Spantidakis I (2022) End-to-end learning via constraint-enforcing approximators for linear programs with applications to supply chains.

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.

Delage E, Ye Y (2010) Distributionally robust optimization under moment uncertainty with application to data-driven problems. *Operations research* 58(3):595–612.

Deng Y, Sen S (2022) Predictive stochastic programming. *Computational Management Science* 1–34.

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 (2022) Smart "predict, then optimize". *Management Science* 68(1):9–26.

Elmachtoub AN, Lam H, Zhang H, Zhao Y (2023) Estimate-then-optimize versus integrated-estimation-optimization: A stochastic dominance perspective. *arXiv preprint arXiv:2304.06833* .

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.

Gao B, Pavel L (2017) On the properties of the softmax function with application in game theory and reinforcement learning. *arXiv preprint arXiv:1704.00805* .

Gao R, Kleywegt A (2023) Distributionally robust stochastic optimization with wasserstein distance. *Mathematics of Operations Research* 48(2):603–655.

Gupta V, Huang M, Rusmevichientong P (2022) Debiasing in-sample policy performance for small-data, large-scale optimization. *Operations Research* .

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

Gupta V, Rusmevichientong P (2021) Small-data, large-scale linear optimization with uncertain objectives. *Management Science* 67(1):220–241.

Hannah L, Powell W, Blei D (2010) Nonparametric density estimation for stochastic optimization with an observable state variable. *Advances in Neural Information Processing Systems* 23.

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* .

Hong LJ, Huang Z, Lam H (2021) Learning-based robust optimization: Procedures and statistical guarantees. *Management Science* 67(6):3447–3467.

Hu Y, Kallus N, Mao X (2022) Fast rates for contextual linear optimization. *Management Science* 68(6):4236–4245.

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* .

Kannan R, Bayraksan G, Luedtke JR (2020) Residuals-based distributionally robust optimization with covariate information. *arXiv preprint arXiv:2012.01088* .

Kannan R, Bayraksan G, Luedtke JR (2022) Data-driven sample average approximation with covariate information. *arXiv preprint arXiv:2207.13554* .

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

Kao YH, Van Roy B (2012) Directed time series regression for control. *arXiv preprint arXiv:1206.6141* .

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* .

Liu J, Li G, Sen S (2022) Coupled learning enabled stochastic programming with endogenous uncertainty. *Mathematics of Operations Research* 47(2):1681–1705.

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.*

Poursoltani M, Delage E, Georghiou A (2023) Robust data-driven prescriptiveness optimization. *arXiv preprint arXiv:2306.05937* .

Qi M, Cao Y, Shen ZJ (2022) Distributionally robust conditional quantile prediction with fixed design. *Management Science* 68(3):1639–1658.

Qi M, Shen ZJ (2022) Integrating prediction/estimation and optimization with applications in operations management. *Tutorials in Operations Research: Emerging and Impactful Topics in Operations*, 36–58 (INFORMS).

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.

Sadana U, Chenreddy A, Delage E, Forel A, Frejinger E, Vidal T (2023) A survey of contextual optimization methods for decision making under uncertainty. *arXiv preprint arXiv:2306.10374* .

Sen B (2018) A gentle introduction to empirical process theory and applications. *Lecture Notes, Columbia University* 11:28–29.

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

Vaart Avd, Wellner JA (2023) Empirical processes. *Weak Convergence and Empirical Processes: With Applications to Statistics*, 127–384 (Springer).

Van der Vaart AW (2000) *Asymptotic statistics*, volume 3 (Cambridge university press).

Van Parys BP, Esfahani PM, Kuhn D (2021) From data to decisions: Distributionally robust optimization is optimal. *Management Science* 67(6):3387–3402.

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

Wang I, Becker C, Van Parys B, Stellato B (2023) Learning for robust optimization. *arXiv preprint arXiv:2305.19225* .

Wiesemann W, Kuhn D, Sim M (2014) Distributionally robust convex optimization. *Operations research* 62(6):1358–1376.

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: Lemmas and Proofs for Section 3.1

LEMMA 1. *For any $\rho > 0$, $w_\rho(\cdot)$ is a single-valued continuous function of $p \in \Delta_K$.*

Proof of Lemma 1 From part 2 in Theorem 9.17 of Sundaram et al. (1996), the mapping $w_\rho(p) : \Delta_k \to S$ is a continuous function of p. □

LEMMA 2. *Suppose that Assumptions 1.A and 1.C hold, and let $w(\cdot) : \Delta_K \to S$ be an optimal solution mapping such that $w(p) \in W(p)$ for all $p \in \Delta_K$. Then, we have that: (i) f^* is the unique minimizer of $\min_{f \in \mathcal{H}} R(w \circ f; 0)$; (ii) the optimal expected unregularized risk is $J^* := \min_{f \in \mathcal{H}} R(w \circ f; 0) = R(w \circ f^*, 0)$, and the value of J^* is the same for all valid optimal solution mappings $w(\cdot)$.*

Proof of Lemma 2 We first prove part *(i)*. Suppose that there exists $\bar{f} \neq f^*$ in $\mathcal{H}$ such that

$$\mathbb{E}_x\left[\sum_{k=1}^K p_k^*(x) c_k(w(\bar{f}(x)))\right] \leq \mathbb{E}_x\left[\sum_{k=1}^K p_k^*(x) c_k(w(f^*(x)))\right].$$

Since $w(f^*(x)) \in W(f^*(x))$ and $f^*(x)$ is the true hypothesis, i.e., $f_k^*(x) = p_k^*(x)$, we have

$$\sum_{k=1}^K p_k^*(x) c_k(w(\bar{f}(x))) \geq \sum_{k=1}^K p_k^*(x) c_k(w(f^*(x)))$$

for all $x \in \mathcal{X}$. Combining the above two inequalities yields

$$\sum_{k=1}^K p_k^*(x) c_k(w(\bar{f}(x))) = \sum_{k=1}^K p_k^*(x) c_k(w(f^*(x))),$$

almost surely for all $x \in \mathcal{X}$. Hence, $w(\bar{f}(x)) \in W(f^*(x))$ almost surely, which contradicts Assumption 1.C. Therefore, under Assumptions 1.A and 1.C, f^* is the unique minimizer of $\min_{f \in \mathcal{H}} R(w \circ f; 0)$.

We now show *(ii)*. Let $w(\cdot), u(\cdot) : \Delta_K \to S$ be two valid optimal solution mappings such that $w(p) \in W(p)$ and $u(p) \in W(p)$ for all $p \in \Delta_K$. Thus, $w(f^*(x)) \in W(f^*(x))$ and $u(f^*(x)) \in W(f^*(x))$ for all x, which yields $\sum_{k=1}^K p_k^*(x) c_k(w(f^*(x))) = \sum_{k=1}^K p_k^*(x) c_k(u(f^*(x)))$ by the definition of $W(\cdot)$. Therefore, taking expectation with respect to x yields $J^* = R(w \circ f^*, 0) = R(u \circ f^*, 0)$. □

Proof of Proposition 1 Let $p \in \Delta_K$ be given. We define the correspondence $g_p(\rho)$ by $g_p(\rho) := w_\rho(p)$ for $\rho > 0$ and $g_p(0) := W(p)$. Here, $g_p(\cdot)$ is a function that maps a value of $\rho \in [0, \infty)$ to a subset of S (which is just a single point $w_\rho(p)$ when $\rho > 0$). Given the sequence $\{\rho_n > 0\}$ that converges to zero, define a corresponding sequence w_n where $w_n = w_\rho(p) \in g_p(\rho_n)$. If $\text{dist}(w_{\rho_n}(p), W(p)) \not\to 0$ as $n \to \infty$, then there exists a constant $\epsilon > 0$, such that for some subsequence $w_{n(m)}$ we have $\text{dist}(w_{n(m)}, W(p)) > \epsilon$ for all m. By the Weierstrass Theorem and the compactness of S, assume without loss of generality that $w_{n(m)} \to \bar{w} \in S$ as $m \to \infty$. Due to the continuity of the objective function in (3) with respect to ρ, we apply the maximum theorem (Berge (1877)), which guarantees that $g_p(\rho)$ is upper hemicontinuous in ρ. Thus, since $\rho_{n(m)} \to 0$ as $m \to \infty$, we have that $\bar{w} \in g_p(0) = W(p)$ according to the definition of upper hemicontinuity. This directly contradicts the claim that $\text{dist}(w_{n(m)}, W(p)) > \epsilon$ for all m. □

LEMMA 3. *Consider the class of functions $g : \mathcal{X} \times \Xi \to \mathbb{R}$ defined by $\mathcal{F} := \{g : g = c \circ w_\rho \circ f$ for some $\rho \in (0, \rho_0]$ and $f \in \mathcal{H}\}$. If Assumptions 1.B and 2.A-2.B hold, then for any $\delta \in (0,1)$, the 3δ-bracketing number $N(3\delta; \mathcal{F}, \|\cdot\|_\infty)$, with respect to the sup-norm $\|\cdot\|_\infty$, is finite.*

Proof for Lemma 3 Assumption 1.B guarantees that, for any $\delta \in (0,1)$, we can find a $\frac{\delta}{2L_cL_w}$-bracket of $\mathcal{H}$. That said, for each $f \in \mathcal{H}$, there exist an $i \in \{1,\ldots,N_1\}$ such that $l_k^i(x) \le f_k^i(x) \le u_k^i(x)$ for all $k = 1,\ldots,K$ and $x \in \mathcal{X}$ and with $\|l^i - u^i\|_\infty \le \frac{\delta}{2L_cL_w}$.

By the uniform Lipschitzness of w_ρ in ρ (Assumption 2.A), we have that for any ρ_1, ρ_2 and p, $\|w_{\rho_1}(p) - w_{\rho_2}(p)\| \le L_\rho|\rho_1 - \rho_2|$. Therefore, $|c(w_{\rho_1}(p),\xi) - c(w_{\rho_2}(p),\xi)| \le L_cL_\rho|\rho_1 - \rho_2|$. Then we consider $T = \lfloor \frac{\rho_0 L_c L_w}{2\delta} \rfloor$ and have a collection of points $\{\rho^0,\ldots,\rho^{T+1}\}$, where $\rho^i := \frac{\delta}{L_cL_\rho}i$, for $i = 0,\ldots,T$, and $\rho^{T+1} := \rho_0$.

Now we show that $\{[c \circ w_{\rho^i} \circ l^j - \delta, \quad c \circ w_{\rho^i} \circ u^j + \delta] \quad : \quad \forall i = 0,\ldots,T+1, \forall j = 1,\ldots,N_1\}$ forms a 3δ-cover of $\mathcal{F}$. We first show that for any $\rho \in (0,\rho_0]$ and $f \in \mathcal{H}$, there exists i and j such that

$$
\begin{aligned}
&|c(w_\rho(f(x)),\xi) - c(w_{\rho^i}(u^j(x)),\xi)| \\
\le{}& |c(w_\rho(f(x)),\xi) - c(w_{\rho^i}(f(x)),\xi)| + |c(w_{\rho^i}(f(x)),\xi) - c(w_{\rho^i}(u^j(x)),\xi)| \\
\le{}& L_cL_\rho|\rho - \rho_i| + L_c\|w_{\rho^i}(u^j(x)) - w_{\rho^i}(f(x))\| \\
\le{}& \frac{\delta}{2} + L_cL_w\|u^j(x) - f(x)\| \\
\le{}& \frac{\delta}{2} + \frac{\delta}{2} = \delta.
\end{aligned}
$$

The second inequality holds by the Lipschitz Assumptions 2.A and the the fact that the objective function $c(\cdot,\xi)$ is L_c-Lipschitz for any ξ. Then the third inequality follows from Assumption 2.B. Therefore, $c(w_\rho(f(x)),\xi) \le c(w_{\rho^i}(u^j(x)),\xi) + \delta$. The same reasoning leads to the fact that $c(w_\rho(f(x)),\xi) \ge c(w_{\rho^i}(l^j(x)),\xi) - \delta$.

Note that for any x and ξ

$$
\begin{aligned}
&|(c(w_{\rho^i}(u^j(x)),\xi) + \delta) - (c(w_{\rho^i}(l^j(x)),\xi) - \delta)| \\
\le{}& 2\delta + |c(w_{\rho^i}(u^j(x)) - c(w_{\rho^i}(l^j(x)),\xi))| \\
\le{}& 2\delta + L_cL_w\|u^j(x) - l^j(x)\|.
\end{aligned}
$$

Therefore, $\|(c \circ w_{\rho^i} \circ u^j + \delta) - (c \circ w_{\rho^i} \circ l^j)\| \le 2\delta + 0.5\delta \le 3\delta$ and we have fund a 3δ-bracket with a size of $N_1 \times T + 1 < \infty$. □

A.1. Justifying Assumptions for Theorem 1

In this subsection, we justify the validity of the uniform Lipschitzness of the regularized oracle (Assumptions 2.A-2.B). To achieve that, we start by showing that the desired property can be guaranteed by a common "automatic crossover" phenomenon of regularized oracle $w_\rho(\cdot)$.

Assumption 6 **(Automatic Crossover of the Regularized Oracle)**. *There exists a positive number $\rho_{c,\phi}$, so that for all $\rho \le \rho_{c,\phi}$ and all $p \in \Delta_K$, $w_\rho(p) \in W^*(p)$. The universal constant $\rho_{c,\phi}$ only depends on the property of the objective function $c(\cdot,\cdot)$ and the regularization $\phi(\cdot)$.*

This assumption indicates that, if ρ is smaller than a threshold $\rho_{c,\phi}$, the output of the regularized oracle is in the optimal solution set of the unregularized oracle.

Combining the automatic crossover assumption and Proposition 2, we have a universal Lipschitz constant for any $\rho \ge 0$.

Corollary 2. *Suppose Assumption 6 holds, then for all $\rho \geq 0$, and any p_1 and p_2 from Δ_K, $\|w_\rho(p_1) - w_\rho(p_2)\| \leq \frac{L_c}{\rho_{c,\phi}}\|p_1 - p_2\|_2$.*

Proof of Corollary 2 Directly follows by combining Proposition 2 and Assumption 6. □

The previous results further imply the uniform Lipschitzness with respect to the regularization parameter ρ for any $p \in \Delta_K$.

Corollary 3. *Suppose Assumption 6 holds, then for all p from Δ_k, and any ρ_1 and ρ_2 from $(0, \rho_0]$, $\|w_{\rho_1}(p) - w_{\rho_2}(p)\| \leq \frac{L_c\sqrt{K}}{\rho_{c,\phi}^2}|\rho_1 - \rho_2|$.*

Proof of Corollary 3 For any $p \in \Delta_K$, and any $\rho > 0$, $w_\rho(p)$ is also the unique minimizer of

$$\sum_{k=1}^{K} \frac{p_k}{\rho} c_K(w) + \phi(w).$$

For simplicity, we assume that $\rho_1 < \rho_2$. Then since $\rho_1 \geq \rho_{c,\phi}$, we apply Proposition 2, we have

$$\|w_{\rho_1}(p) - w_{\rho_2}(p)\| \leq L_c \left\| \frac{p}{\rho_1} - \frac{p}{\rho_2} \right\|_2 \leq L_c \|p\|_2 \left| \frac{1}{\rho_1} - \frac{1}{\rho_2} \right| \leq L_c \frac{\sqrt{K}}{\rho_{c,\phi}^2} |\rho_1 - \rho_2|.$$

The first inequality holds by applying Proposition 2 and the last inequality follows from the the fact that Δ_K is compact and Assumption 6. Similarly, if $\rho_1 < \rho_{c,\phi} \leq \rho_2$, the above analysis also applies because $w_{\rho_1}(p) = w_{\rho_{c,\phi}}(p)$ for any $p \in \Delta_K$ by Assumption 6 and that $|\rho_1 - \rho_2| \geq |\rho_{c,\phi} - \rho_2|$. In the last case where $\rho_1 < \rho_2 \leq \rho_{c,\phi}$, then $w_{\rho_1}(p) = w_{\rho_2}(p) = w_{\rho_{c,\phi}}(p)$ for any $p \in \Delta_K$. Thus the desired result still trivially holds. □

A.1.1. A Sufficient Condition for the Automatic Crossover Property In this part, we let $h(w,p) := \sum_{k=1}^{K} p_k c_k(w)$ for simplicity. $W^*(p) := \arg\min_{w \in S} h(w,p)$ and $h^*(p) := \min_{w \in S} h(w,p)$. Then we demonstrate that the automatic crossover property is quite common, for example, it holds for any convex objective $c(\cdot,\cdot)$ as long as it satisfies the following linear growth condition stated in Assumption 7.

Assumption 7 **(Linear Growth)**. *There exists $\mu > 0$ such that for any $w \in S$, it holds that $h(w,p) - h^*(p) \geq \mu \text{dist}(w, W^*(p))$.*

Note that, for example, any piece-wise linear function satisfies the linear growth condition.

Theorem 5. *Suppose Assumption 7 holds. Then, for all $\rho \in (0, \rho_{c,\phi}]$, the regularized oracle $w_\rho(p)$ automatically crosses over (Assumption 6) with the unregularized solution set $W^*(p)$, i.e., $w_\rho(p) \in W^*(p)$. The phase transitioning constant of the crossover is $\rho_{c,\phi} := \frac{\mu}{\overline{\nabla\phi}}$, where $\overline{\nabla\phi} := \sup_{w \in S} \|\nabla\phi(w)\|_* < \infty$.*

Proof of Theorem 5 We let $\bar{w}_\rho(p) := \arg\min_{w \in W^*(p)} \|w - w_\rho(p)\|$ denote the projection of $w_\rho(p)$ to the unregularized optimal solution set $W^*(p)$. For any $p \in \Delta_K$ and any $\rho \geq 0$, we have

$$\begin{aligned}
0 &\geq [h(w_\rho(p),p) + \rho\phi(w_\rho(p))] - [h(\bar{w}_\rho(p),p) + \rho\phi(\bar{w}_\rho(p))] \\
&= [h(w_\rho(p),p) - h(\bar{w}_\rho(p),p)] + \rho[\phi(w_\rho(p)) - \phi(\bar{w}_\rho(p))] \\
&= [h(w_\rho(p),p) - h(\bar{w}_\rho(p),p)] + \rho[\phi(w_\rho(p)) - \phi(\bar{w}_\rho(p)) - \nabla\phi(\bar{w}_\rho(p))^T(w_\rho(p) - \bar{w}_\rho(p))] \\
&\quad + \rho\nabla\phi(\bar{w}_\rho(p))^T(w_\rho(p) - \bar{w}_\rho(p)) \\
&\geq \mu\|w_\rho(p) - \bar{w}_\rho(p)\| + \frac{\rho}{2}\|w_\rho(p) - \bar{w}_\rho(p)\|^2 - \rho\|\nabla\phi(\bar{w}_\rho(p))\|_*\|w_\rho(p) - \bar{w}_\rho(p)\| \\
&\geq \|w_\rho(p) - \bar{w}_\rho(p)\|(\mu + \frac{\rho}{2}\|w_\rho(p) - \bar{w}_\rho(p)\| - \rho\overline{\nabla\phi}).
\end{aligned}$$

The first inequality follows from optimality of $w_\rho(p)$. The first term of the second inequality holds because of Assumption 7, and the second term in the second inequality holds due to the 1-strong convexity of ϕ. If $\rho \le \frac{\mu}{\overline{\nabla\phi}} := \rho_{c,\phi}$, then $\mu + \frac{\rho}{2}\|w_\rho(p) - \bar{w}_\rho(p)\| - \rho\overline{\nabla\phi} \ge 0$. Thus, $\|w_\rho(p) - \bar{w}_\rho(p)\|$ has to be zero; otherwise the right side is a positive number and we have a contradiction. □

A.2. Proof of Theorem 1

Proof of Theorem 1. We first show that $\lim_{\rho\to 0} J^*_\rho = J^*$. Recall that $J^* = R(w \circ f^*)$ and $J^*_\rho = R(w_\rho \circ f^*_\rho)$. As defined at the beginning of Section 3, the function $w(\cdot) : \Delta_K \to S$ arbitrarily selects a value from the optimal solution set given by $W(\cdot)$. Specifically, for any given $p \in \Delta_K$, $w(p)$ outputs an arbitrary value from the set $W(p) = \arg\min_{w\in S} \sum_{k=1}^K p_k c_k(w)$.

We first show that, for any sequence $\{\rho_n\}$ that converges to zero and for any given x,

$$\sum_{k=1}^K f^*_k(x) c_k(w_{\rho_n}(f^*(x))) \to \sum_{k=1}^K f^*_k(x) c_k(w(f^*(x))) \text{ point-wise as } n \to \infty.$$

Let $\epsilon > 0$ be fixed. Note that $c_k(\cdot)$ is Lipschitz and, therefore, uniformly continuous. Thus, there exists a $\delta > 0$ such that, for any $\bar{w}_n$, if $\|w_{\rho_n}(f^*(x)) - \bar{w}_n\| < \delta$, then $|c_k(w_{\rho_n}(f^*(x))) - c_k(\bar{w}_n)| < \epsilon$. By utilizing the convergence result provided by Proposition 1, there is a value of N such that for all $n > N$, it holds that $\text{dist}(w_{\rho_n}(f^*(x)), W(f^*(x))) < \delta$. Then we let $\bar{w}_n := \arg\min_{w\in W(f^*(x))} \|w_{\rho_n}(f^*(x)) - w\|$. It is important to note that for any $\bar{w}_1, \bar{w}_2 \in W(f^*(x))$, we have $\sum_{k=1}^K f^*_k(x)c_k(\bar{w}_1) = \sum_{k=1}^K f^*_k(x)c_k(\bar{w}_2)$. Therefore, for all $n > N$, the following holds

$$|\sum_{k=1}^K f^*_k(x) c_k(w_{\rho_n}(f^*(x))) - \sum_{k=1}^K f^*_k(x) c_k(w(f^*(x)))| < \epsilon,$$

where again note that $w(\cdot) : \Delta_K \to S$ arbitrarily selects a value from the optimal solution set. This shows that $\sum_{k=1}^K f^*_k(x)c_k(w_{\rho_n}(f^*(x)))$ converges to $\sum_{k=1}^K f^*_k(x)c_k(w(f^*(x)))$ point-wise as $n \to \infty$.

Then, by the dominated convergence theorem we have that

$$\begin{aligned} R(w_{\rho_n} \circ f^*) &:= \mathbb{E}_x\left[\sum_{k=1}^K f^*_k(x) c_k(w_{\rho_n}(f^*(x)))\right] \\ &\to R(w \circ f^*) := \mathbb{E}_x\left[\sum_{k=1}^K f^*_k(x) c_k(w(f^*(x)))\right] \quad \text{as } n \to \infty. \end{aligned} \tag{11}$$

Note that

$$\begin{aligned} R(w_{\rho_n} \circ f^*) &:= \mathbb{E}_x\left[\sum_{k=1}^K f^*_k(x) c_k(w_{\rho_n}(f^*(x)))\right] \\ &\le \mathbb{E}_x\left[\sum_{k=1}^K f^*_k(x) c_k(w_{\rho_n}(f^*(x))) + \rho_n \phi(w_{\rho_n}(f^*(x)))\right] \\ &\le \mathbb{E}_x\left[\sum_{k=1}^K f^*_k(x) c_k(w_{\rho_n}(f^*_{\rho_n}(x))) + \rho_n \phi(w_{\rho_n}(f^*_{\rho_n}(x)))\right] \\ &\le \mathbb{E}_x\left[\sum_{k=1}^K f^*_k(x) c_k(w_{\rho_n}(f^*_{\rho_n}(x))) + \rho_n \bar{\phi}\right] \\ &= R(w_{\rho_n} \circ f^*_{\rho_n}) + \rho_n \bar{\phi} \\ &\le \mathbb{E}_x\left[\sum_{k=1}^K f^*_k(x) c_k(w_{\rho_n}(f^*(x)))\right] + \rho_n \bar{\phi} \\ &= R(w_{\rho_n} \circ f^*) + \rho_n \bar{\phi}. \end{aligned}$$

The first inequality holds due to the presence of an additional non-negative regularization term. The second inequality arises from the definition of $w_{\rho_n}(\cdot)$. Specifically, $w_{\rho_n}(f^*(x))$ is the minimizer of $\min_{w\in S}\sum_{k=1}^K f_k^*(x)c_k(w)+\rho_n\phi(w)$ for any given x while $w_{\rho_n}(f^*_{\rho_n}(x))$ is not. The third inequality holds because the decision regularization is upper-bounded by $\bar{\phi}$. Then the last inequality holds because $f^*_{\rho_n}$ is the minimizer of $\min_{f\in\mathcal{H}}\mathbb{E}_x\sum_{k=1}^K f_k^*(x)c_k(w_{\rho_n}(f(x)))$. Notice that the above chain of inequalities demonstrates that $R(w_{\rho_n}\circ f^*_{\rho_n})\in[R(w_{\rho_n}\circ f^*)-\rho_n\bar{\phi}, R(w_{\rho_n}\circ f^*)+\rho_n\bar{\phi}]$. Now let $\epsilon>0$ again be fixed. Since $\rho_n\to 0$, there exists N_1 such that, for all $n>N_1$, $\rho_n<\frac{\epsilon}{2\bar{\phi}}$ and therefore $|R(w_{\rho_n}\circ f^*_{\rho_n})-R(w_{\rho_n}\circ f^*)|\le\epsilon/2$. At the same time, by (11), there exists N_2 such that, for all $n>N_2$, we have $|R(w_{\rho_n}\circ f^*)-R(w\circ f^*)|<\frac{\epsilon}{2}$. Therefore, considering $N=\max\{N_1,N_2\}$, we have that for all $n>N$,

$$\begin{aligned}|R(w_{\rho_n}\circ f^*_{\rho_n})-R(w\circ f^*)| &\le |R(w_{\rho_n}\circ f^*_{\rho_n})-R(w_{\rho_n}\circ f^*)|+|R(w_{\rho_n}\circ f^*)-R(w\circ f^*)|\\ &\le\frac{\epsilon}{2}+\frac{\epsilon}{2}=\epsilon.\end{aligned}$$

Therefore, as ρ_n converges to zero, we have

$$J^*_{\rho_n}:=R(w_{\rho_n}\circ f^*_{\rho_n})\to R(w\circ f^*)=J^*.$$

If Assumptions 1.B and 2.A-2.B hold, Lemma 3 shows that, the class of function $\mathcal{F}:=\{c\circ w_\rho\circ f:\mathcal{X}\times\Xi\to\mathbb{R}|\rho\in(0,\rho_0], f\in\mathcal{H}\}$ has a finite bracketing number. Therefore, we apply Theorem 3.2 in Sen (2018)(see also Wainwright (2019) and Van der Vaart (2000)) and have

$$\hat{R}_n(w_\rho\circ f)\to R(w_\rho\circ f)\ \text{ with probability 1,} \tag{12}$$

uniformly for $\rho\in(0,\rho_0]$ and $f\in\mathcal{H}$. Then by (12), we have

$$\begin{aligned}|\hat{J}^n_\rho-J^*_\rho| &=\max\left\{\hat{R}_n(w_\rho\circ\hat{f}^n_\rho)-R(w_\rho\circ f^*_\rho), R(w_\rho\circ f^*_\rho)-\hat{R}_n(w_\rho\circ\hat{f}^n_\rho)\right\}\\ &\le\max\left\{\hat{R}_n(w_\rho\circ f^*_\rho)-R(w_\rho\circ f^*_\rho), R(w_\rho\circ\hat{f}^n_\rho)-\hat{R}_n(w_\rho\circ\hat{f}^n_\rho)\right\}\\ &\le\max\left\{|\hat{R}_n(w_\rho\circ f^*_\rho)-R(w_\rho\circ f^*_\rho)|, |R(w_\rho\circ\hat{f}^n_\rho)-\hat{R}_n(w_\rho\circ\hat{f}^n_\rho)|\right\},\end{aligned}$$

where the second inequality holds due to the fact that $\hat{R}_n(w_\rho\circ f^*_\rho)\ge\hat{R}_n(w_\rho\circ\hat{f}^n_\rho)$ and that $R(w_\rho\circ\hat{f}^n_\rho)\ge R(w_\rho\circ f^*_\rho)$, because $\hat{f}^n_\rho$ minimizes $\hat{R}_n(w_\rho\circ f)$ and f^*_ρ minimizes $R(w_\rho\circ f)$. By the uniform convergence of the empirical risk in $\rho\in(0,\rho_0]$ and $f\in\mathcal{H}$, guaranteed by (12), for any positive $\epsilon>0$, there exists N such that for all $n>N$, $\sup_{\rho\in(0,\rho_0]}|\hat{R}_n(w_\rho\circ f^*_\rho)-R(w_\rho\circ f^*_\rho)|<\epsilon$ and $\sup_{\rho\in(0,\rho_0]}|R(w_\rho\circ\hat{f}^n_\rho)-\hat{R}_n(w_\rho\circ\hat{f}^n_\rho)|<\epsilon$ with probability 1. Thus, we conclude that $\sup_{\rho\in(0,\rho_0]}|\hat{J}^n_\rho-J^*_\rho|\to 0$ with probability 1 as $n\to\infty$. Therefore, since $\{\rho_n\}_{n=1}^\infty$ satisfies $\rho_n\in(0,\rho_0]$ and $\rho_n\to 0$ as $n\to\infty$, we have that

$$|\hat{J}^n_{\rho_n}-J^*|\ \le\ |J^*_{\rho_n}-J^*|+\sup_{\rho\in(0,\rho_0]}|\hat{J}^n_\rho-J^*_\rho|\ \to\ 0 \text{ with probability 1, as } n\to\infty.$$

Now we prove *(ii)* and *(iii)* simultaneously. Noting that

$$\begin{aligned}\text{dist}(w_{\rho_n}(\hat{f}^n_{\rho_n}(x)), W(f^*(x)) &=\inf_{u\in W(f^*(x))}\|w_{\rho_n}(\hat{f}^n_{\rho_n}(x))-w_{\rho_n}(f^*_{\rho_n}(x))+w_{\rho_n}(f^*_{\rho_n}(x))-u\|\\ &\le\|w_{\rho_n}(\hat{f}^n_{\rho_n}(x))-w_{\rho_n}(f^*_{\rho_n}(x))\|+\inf_{u\in W(f^*(x))}\|w_{\rho_n}(f^*_{\rho_n}(x))-u\|\\ &\le\|w_{\rho_n}(\hat{f}^n_{\rho_n}(x))-w_{\rho_n}(f^*_{\rho_n}(x))\|+\text{dist}(w_{\rho_n}(f^*_{\rho_n}(x)), W(f^*(x))).\end{aligned} \tag{13}$$

The first inequality follows from the triangle inequality. The second inequality holds by the definition of $\mathrm{dist}(\cdot,\cdot)$. In the remaining part of this proof, we separately establish the convergence in probability of the first and second terms above, both holding $\mathcal{D}_x$-almost surely over x.

Due to the compactness of S, with any sequence of ρ_n converging to zero, the sequence $w_{\rho_n}(f^*_{\rho_n}(x))$ has accumulation points. Let $w_{\rho_{n(m)}}(f^*_{\rho_{n(m)}}(x))$ be any subsequence converging to an accumulation point $\bar{w}(x) \in S$. Note that we have

$$J^* = \lim_{m\to\infty} \mathbb{E}_x\left[\sum_{k=1}^{K} f^*_k(x) c_k(w_{\rho_{n(m)}}(f^*_{\rho_{n(m)}}(x)))\right] \geq \mathbb{E}_x\left[\sum_{k=1}^{K} f^*_k(x) c_k(\bar{w}(x))\right],$$

where the equality follows from the convergence of $J^*_\rho \to J^*$ when $\rho \to 0$, and the first inequality holds by Fatou's lemma. Therefore, $\mathcal{D}_x$ almost surely for all x, $\bar{w}(x)$ must lie in the set $W(f^*(x))$. Then $\mathrm{dist}(w_{\rho_{n(m)}}(f^*_{\rho_{n(m)}}(x)), W(f^*(x))) \leq \|w_{\rho_{n(m)}}(f^*_{\rho_{n(m)}}(x)) - \bar{w}(x)\| = 0$. The aforementioned reasoning holds for every converging subsequence $n(m)$, then the original sequence $\mathrm{dist}(w_{\rho_n}(f^*_{\rho_n}(x)), W(f^*(x)))$ also converges to zero $\mathcal{D}_x$-almost surely for all x.

Now we aim to demonstrate the convergence of $\hat{f}^n_\rho$ to f^*_ρ for any fixed $\rho \in (0, \rho_0]$. Note that the first condition required by Theorem 5.7 in Van der Vaart (2000) is the uniform convergence in f, which is already established in (12). The second condition, which requires f^*_ρ to be a well-separated point given each ρ, is implied by Assumption 1.C combined with Assumption 2.C. Then we apply Theorem 5.7 in Van der Vaart (2000) and have that, for each $\rho \in (0, \rho_0]$, $\hat{f}^n_\rho$ convergences to f^*_ρ in probability as $n \to \infty$. Due to the continuity of $w_\rho(\cdot)$, we have $\|w_\rho(\hat{f}^n_\rho(x)) - w_\rho(f^*_\rho(x))\|$ converges to zero in probability for all x.

Then we want to show the following claim: $\sup_{\rho\in(0,\rho_0]} \|w_\rho(f^*_\rho(x)) - w_\rho(\hat{f}^n_\rho(x))\|$ converges to zero in probability, $\mathcal{D}_x$-almost surely for all x. By the uniform convergence of $|\hat{R}(w_\rho \circ f) - R(w_\rho \circ f)|$ presented in (12), for any $\delta > 0$, there exists N, such that for any $n \geq N$, $\sup_{\rho\in(0,\rho_0]} |\hat{R}(w_\rho \circ \hat{f}^n_\rho) - R(w_\rho \circ \hat{f}^n_\rho)| < \delta/2$. If the desired claim is not true, there exists $\epsilon > 0$ and a subsequence $\{m(n)\}$ with $\mathbb{P}_x(\sup_{\rho\in(0,\rho_0]} \|w_\rho(f^*_\rho(x)) - w_\rho(\hat{f}^{m(n)}_\rho(x))\| > \epsilon) > 0$ for all $n = 1, \ldots, \infty$. Then for each n, one may choose a ρ that achieves $\mathbb{P}_x(\|w_\rho(f^*_\rho(x)) - w_\rho(\hat{f}^{m(n)}_\rho(x))\| \geq \epsilon) > 0$. Then by Assumption 2.C, there must exist a $\delta > 0$ such that $\sup_{\rho\in(0,\rho_0]} |R(w_\rho \circ f^*_\rho) - R(w_\rho \circ \hat{f}^{m(n)}_\rho)| \geq \delta$. Then we notice that, for this subsequence $\{m(n)\}$, so that there exists a positive δ such that

$$\begin{aligned}
&\sup_{\rho\in(0,\rho_0]} |\hat{J}^{m(n)}_\rho - J^*_\rho| \\
&= \sup_{\rho\in(0,\rho_0]} |\hat{R}(w_\rho \circ \hat{f}^{m(n)}_\rho) - R(w_\rho \circ f^*_\rho)| \\
&\geq \sup_{\rho\in(0,\rho_0]} |R(w_\rho \circ \hat{f}^{m(n)}_\rho) - R(w_\rho \circ f^*_\rho)| - \sup_{\rho\in(0,\rho_0]} |\hat{R}(w_\rho(\hat{f}^{m(n)}_\rho)) - R(w_\rho(\hat{f}^{m(n)}_\rho))| \\
&\geq \delta - \delta/2 = \delta/2 > 0.
\end{aligned}$$

This contradicts the uniform convergence of $\hat{J}^n_\rho$ to J^*_ρ in probability. Therefore, for any decreasing sequence $\{\rho_n\}_{n=1}^\infty$ with $\rho_n > 0$ and $\rho_n \to 0$, we have $\|w_{\rho_n}(\hat{f}^n_{\rho_n}(x)) - w_{\rho_n}(f^*_{\rho_n}(x))\| \leq \sup_{\rho\in(0,\rho_0]} \|w_\rho(\hat{f}^n_\rho(x)) - w_\rho(f^*_\rho(x))\| \to 0$ with probability 1 for $\mathcal{D}_x$-almost surely. Then, returning to (13), we have

$$\begin{aligned}
\mathrm{dist}(w_{\rho_n}(\hat{f}^n_{\rho_n}(x)), W(f^*(x)) &\leq \|w_{\rho_n}(\hat{f}^n_{\rho_n}(x)) - w_{\rho_n}(f^*_{\rho_n}(x))\| + \mathrm{dist}(w_{\rho_n}(f^*_{\rho_n}(x)), W(f^*(x))) \\
&\to 0, \qquad \mathcal{D}_x - \text{almost surely, with probability } 1.
\end{aligned}$$

Thus *(ii)* is proved.

Finally, if we have an accumulation point of $\hat{f}^n_{\rho_n}$, denoted as $\bar{f}$, we have

$$\begin{aligned}\mathrm{dist}(w_{\rho_n}(\hat{f}^n_{\rho_n}(x)), W(\bar{f}(x))) &\leq \|w_{\rho_n}(\hat{f}^n_{\rho_n}(x)) - w_{\rho_n}(\bar{f}(x))\| + \mathrm{dist}(w_{\rho_n}(\bar{f}(x)), W(\bar{f}(x))) \\ &\leq L_w \|\hat{f}^n_{\rho_n}(x) - \bar{f}(x)\| + \mathrm{dist}(w_{\rho_n}(\bar{f}(x)), W(\bar{f}(x))) \\ &\to 0, \qquad \text{as } n \to \infty \text{ for all } x \in \mathcal{X}.\end{aligned}$$

Knowing that $\mathrm{dist}(w_{\rho_n}(\bar{f}(x)), W(\bar{f}(x))) \to 0$ holds by Proposition 1. Then by *(ii)*, $\mathrm{dist}(w_{\rho_n}(\hat{f}^n_{\rho_n}(x)), W(f^*(x)) \to 0$, $\mathcal{D}_x$-almost surely with probability 1, as $n \to \infty$. Thus, for any accumulation point of $w_{\rho_n}(\hat{f}^n_{\rho_n}(x))$, denoted by $\bar{w}$, it holds that $\bar{w} \in W(\bar{f}(x))$ for all $x \in \mathcal{X}$ and that $\bar{w} \in W(f^*(x))$ $\mathcal{D}_x$-almost surely with probability 1. That is, $W(\bar{f}(x)) \cap W(f^*(x)) \neq \emptyset$, $\mathcal{D}_x$-almost surely with probability 1. Then if $\bar{f} \neq f^*$, it contradicts with Assumption 1.C. Thus with the uniqueness assumption, the true hypothesis f^* can be recovered by $\hat{f}^n_{\rho_n}$. □

Appendix B: Supplementary Lemmas and Proofs for Sections 3.2 and 4

B.1. Lemmas and Proofs for Section 3.2

Proof of Proposition 2 Let $p, p' \in \mathbb{R}^K_+$ 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

$$h_\rho(w', p) - h_\rho(w,p) \ \geq \ g^T(w' - w) + \frac{\rho}{2}\|w' - w\|^2 \qquad \forall w' \in S. \tag{14}$$

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 (14) 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 (3.A). Dividing by $\|w_\rho(p) - w_\rho(p')\|$ leads to (6), and combining the resulting inequality again with (3.A) yields (7). □

Proof of Theorem 3 Due to Proposition 2, in particular, the Lipschitz property of $c(\cdot)$ in (7), 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 desired result. □

Proof of Corollary 1 According to Theorem 3, we have

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

with probability at least $1-\frac{\delta}{2}$. Then we apply Hoeffding's inequality and have

$$\hat{R}_n(w_{\rho_n} \circ f^*) \leq R(w_{\rho_n} \circ f^*) + \bar{c}\sqrt{\frac{2\log(\frac{2}{\delta})}{n}}$$

with probability at least $1-\frac{\delta}{2}$. Note that

$$\hat{R}_n(w_{\rho_n} \circ \hat{f}^n_{\rho_n}) \leq \hat{R}_n(w_{\rho_n} \circ f^*),$$

then we have

$$R(w_{\rho_n} \circ \hat{f}^n_{\rho_n}) \leq R(w_{\rho_n} \circ f^*) + \frac{\sqrt{2}L_c^2}{\rho_n}\mathfrak{R}_n(\mathcal{H}) + \frac{3\bar{c}}{2}\sqrt{\frac{2\log(\frac{2}{\delta})}{n}}$$

with probability at least $1-\frac{\delta}{2}$. □

B.2. Proofs and Supplementary Results for Section 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

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

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

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

Note that the dual function of the nominal quadratic optimization is

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

thus the dual problem becomes

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

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. □

Proof of Proposition 4 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 (4). 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 (4). □

Proof of Corollary 4 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. □

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

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

Noted that

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} |c(w_\rho(f(x_i)), \xi_i) - c(\tilde{w}_\rho(f(x_i)), \xi_i)| &\leq L_c \frac{1}{n}\sum_{i=1}^{n} \|w_\rho(f(x_i))) - \tilde{w}_\rho(f(x_i))\|_1 \\ &= L_c \sum_{j=1}^{d} \frac{1}{n}\sum_{i=1}^{n} |w_{\rho,j}(f(x_i)) - \tilde{w}_{\rho,j}(f(x_i))| \qquad (15)\end{aligned}$$

When the approximated oracle has a uniform error, we combine (15) and Assumption 4 to have

$$\frac{1}{n}\sum_{i=1}^{n} |c(w_\rho(f(x_i)), \xi_i) - c(\tilde{w}_\rho(f(x_i)), \xi_i)| \leq L_c \sum_{j=1}^{d} \mathcal{E}_j^{\mathrm{unif}}. \qquad (16)$$

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 (5) with distribution $\mathcal{D}_{f(x)}$ and $\mathcal{D}_p$ respectively and have

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

and

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

each with probability at least $1 - \frac{\delta}{2d}$. Considering the total variation between $\mathcal{D}_{f(x)}$ and $\mathcal{D}_p$, we have the following

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

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,

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

with probability at least $1 - \delta$. Therefore, we have

$$(15) \leq L_c \sum_{j=1}^{d} \left[\frac{1}{m}\sum_{i=1}^{m} |w_{i,j} - \tilde{w}_{\rho,j}(p_i)| + \mathcal{E}_j^{\text{prob}}(n, \delta/2d; \mathcal{G}) + \mathcal{E}_j^{\text{prob}}(m, \delta/2d; \mathcal{G})\right] + \text{diam}(S) L_c \text{TV}(\mathcal{D}_{f(x)}, \mathcal{D}_p), \tag{17}$$

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 (9) and (10) follow from combining (16) and (17) with Theorem 3. □

Appendix C: Approximating 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.

C.1. Bernstein Polynomials

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

Definition 3 (**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:

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

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 4.

Proposition 3. *For a given $\rho > 0$, suppose that we use the Bernstein approximation method (Definition 3) 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 4 with $\mathcal{E}_j^{\text{unif}} = \frac{\Omega L_c}{\rho\sqrt{s}}$,where $\Omega > 0$ is an absolute constant.*

Proof of Proposition 3 This result directly follows from Theorem 3.2 in De Klerk et al. (2008) together with the Lipschitz property from Proposition 2. □

Given the result in Proposition 3, 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.

C.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 5. 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 $\tau_j(\mathcal{G}, \{p_i\}_{i=1}^m, r) := \inf_{u \in \mathcal{G}: \|u\|_{\mathcal{G}} \leq r} (\frac{1}{m}\sum_{i=1}^m (u(p_i) - w_{\rho,j}(p_i))^2)^{1/2}$ 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:

$$\min_{u \in \mathcal{G}: \|u\|_{\mathcal{G}} \leq r} \frac{1}{m} \sum_{i=1}^m (u(p_i) - w_{i,j})^2 \tag{18}$$

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

Proposition 4. *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* (18) *with a user-defined radius $r > 0$. Then, there exist absolute constants $\bar{c}, \bar{c}'$ such that for all $\delta_m \geq \bar{c}_0 \frac{\sigma}{r} \frac{(s-1+K)!}{(s-1)!K!} \frac{1}{m}$, we have*

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

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 $\theta_m \geq \bar{c}_3 \sqrt{\frac{1}{m} \frac{(s-1+K)!}{(s-1)!K!}}$, if it also holds that $m\theta_m^2 \geq \bar{c}_0 \log(4 \log(\frac{1}{\theta_m}))$, then

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

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 body of the proof is a generalization of the result in Example 13.19 of Wainwright (2019). We have the corresponding generalization bound in the following corollary.

Corollary 4. *Suppose Assumption 3 holds and that the hypothesis class $\mathcal{H}$ has bounded multi-variate Rademacher complexity $\mathfrak{R}_n(\mathcal{H})$. Suppose further that we employ kernel ridge regression* (18) *using a function class $\mathcal{G}$ induced by a polynomial kernel of degree s under the same conditions as in Proposition 4. Then, for any $\delta \in (0,1]$ and $\rho_n > 0$, the following inequalities hold for all $f \in \mathcal{H}$:*

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

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 $\delta' = \delta + \bar{c}_2\exp(-\bar{c}_2'\frac{mr^2}{\sigma^2}\delta_m^2) + \bar{c}_4(\exp(-\bar{c}_4'\frac{m\theta_m^2}{r^2}) + \exp(-\bar{c}_4'\frac{n\theta_n^2}{r^2}))$ and δ_m, θ_m, and θ_n are chosen to satisfy the conditions in Proposition 4.

Finally, in Section C.3 we provide an alternative and exact computational approach in the semi-algebraic case when we use a polynomial function to approximate the optimal oracle.

C.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 4 (**Semi-algebraic Set (Lasserre (2015))**). $K \subset \mathbb{R}^n$ is a basic semi-algebraic set if

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

for some polynomial functions $(g_j)_{j=1}^m$, i.e., $(g_j)_{j=1}^m \subset \mathbb{R}[x]$, where $\mathbb{R}[x]$ denotes the ring of real polynomials. Similarly, a semi-algebraic set is defined by not only a finite sequence of polynomial inequalities, but also equations, or finite union of these.

Definition 5 (**Semi-algebraic Function (Lasserre (2015)).**). Let K be a semi-algebraic set of $\mathbb{R}^n$. A function $f : K \to \mathbb{R}$ is a semi-algebraic function if its graph $\Psi_f := \{(x, f(x)) : x \in K\}$ is a semi-algebraic set of $\mathbb{R}^n \times \mathbb{R}$.

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 are semi-algebraic.

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:

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

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 5. *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 $y_k \in \mathbb{R}^m, k = 1,\ldots,K,$ $z, u \in \mathbb{R}^m$

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

Proof of Proposition 5: The condition of

$$Bx + b \in \Delta_K$$

can be represented by the following constraints:

$$B_k^T x + b_k \geq 0, \qquad \forall k = 1,\ldots,K \tag{19}$$

$$\mathbb{1}^T(Bx + b) \geq 1 \tag{20}$$

$$\mathbb{1}^T(Bx + b) \leq 1 \tag{21}$$

(19) represents the non-negativity constraints, while (20) and (21) consist of the normalization constraint. We first rewrite the non-negativity constraint for a component k, for all x such that $Ax \geq a$ as

$$\begin{aligned} 0 \leq \min \quad & B_k^T x + b_k \\ \text{s.t.} \quad & Ax \geq a. \end{aligned}$$

Then consider the dual problem of the above linear programming and we have:

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

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

$$\begin{cases} a^T y_k \geq -b_k \\ A^T y_k = B_k \\ y_k \geq 0. \end{cases} \tag{22}$$

Then we rewrite (20) for all $x \in \mathcal{X}$ as

$$\begin{aligned} 1 \leq \min \quad & \mathbb{1}^T Bx + \mathbb{1}^T b \\ \text{s.t.} \quad & Ax \geq a, \end{aligned}$$

similarly by considering the dual problem

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

which reduces to the following group of constraints

$$\begin{cases} a^T z \ge 1-\mathbb{1}^T b \\ A^T z = B^T \mathbb{1} \\ z \ge 0. \end{cases} \tag{23}$$

Finally, we consider the constraint (21)

$$\begin{aligned} 1 \ge \max \quad & \mathbb{1}^T Bx + \mathbb{1}^T b \\ \text{s.t.} \quad & -Ax \le -a \end{aligned}$$

by strong duality, it is equivalent to

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

which reduces to

$$\begin{cases} a^T u \ge -1+\mathbb{1}^T b \\ A^T u = -B^T \mathbb{1} \\ u \ge 0. \end{cases} \tag{24}$$

Therefore, the condition $Bx + b \in \Delta_K, \forall x \in \mathcal{X}$ can be represented by combining (22), (23), and (24). □

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.

Appendix D: Supplementary Materials for Section 5

Demand generation (multi-product newsvendor) For $K = \{5, 10, 15\}$, we generate scenarios $\{\tilde{z}_1, \ldots, \tilde{z}_K\}$ randomly in the following manner. First, we randomly generate an aggregated demand vector in the dimension of $K = 15$, where each component follows a Normal distribution with a mean of 70 and a standard deviation of 15. Then, for each k, we split the aggregated demand between two products using the weights u_k and $1 - u_k$, where u_k follows a uniform distribution $U[0, 1]$. We generated the scenarios once and used the demand scenarios listed in Table 1 for all numerical experiments in Section 5.1.

Scenarios	Product 1	Product 2
1	15.080	76.154
2	47.238	6.483
3	4.120	56.635
4	21.646	37.001
5	8.686	66.519
6	8.989	84.359
7	2.149	23.506
8	7.857	82.529
9	1.774	77.951
10	57.281	17.009
11	5.628	108.726
12	40.191	46.200
13	82.417	5.146
14	19.817	71.711
15	24.311	42.974

Table 1 Randomly generated demand scenarios for multi-product newsvendor problem

Demand generation (quadratic cost network flow) For $K = \{5, 10, 15\}$, we generate scenarios $\{\tilde{z}_1, \ldots, \tilde{z}_K\}$ randomly in the following manner. First, we randomly generate an aggregated vector that its k-th element represents $\sum_{d=1}^{4} \tilde{z}_{k,d}$. This is the sum of ξ_d over all edges. Each component of this aggregated vector follows a Normal distribution with a mean of 15 and a standard deviation of 5. Then, for each k, we split the aggregated demand across four edges using a weight vector randomly sampled from the K-th simplex Δ_K following the Dirichlet distribution. We generated the scenarios once and used the demand scenarios listed in Table 2 for all numerical experiments in Section 5.2.

Scenarios	Edge 1	Edge 2	Edge 3	Edge 4
1	1.411441	16.519779	0.548652	3.598200
2	2.919887	2.408818	2.081370	2.163493
3	0.123442	8.085430	0.276777	3.432811
4	8.932937	0.769055	1.413136	0.100291
5	4.889119	2.813464	1.819007	7.213648
6	10.608435	0.143297	11.535804	0.494919
7	0.004825	0.010799	0.184079	0.018554
8	0.493836	11.108500	3.756536	6.436532
9	0.647171	8.852180	6.165669	2.576713
10	5.981172	1.493690	7.191768	1.763514
11	8.081012	13.955062	0.332960	7.415544
12	3.645159	6.734041	6.581146	3.503203
13	9.534242	1.280481	6.823592	3.215876
14	4.553855	5.306134	2.320332	9.995780
15	0.338866	5.585264	1.949093	6.221872

Table 2 Randomly generated flow scenarios for quadratic cost network flow problem

Appendix E: Comparison with Policy Optimization

To illustrate the applicability of our performance guarantees, we briefly compare with performance guarantees of policy optimization methods. Policy optimization methods involve directly learning the policy function $\pi : \mathcal{X} \to S$, which is a mapping from the feature space to the set of feasible decisions S. This is achieved by estimating the policy function using a hypothesis $\pi \in \mathcal{P}$ based on the collected training data set $\{(x_i, \xi_i)\}_{i=1}^n$. That is, applying the ERM principle with policy optimization would lead to minimizing the empirical risk

$$\min_{\pi \in \mathcal{P}} \hat{R}_n(\pi).$$

One of the advantages of the policy optimization approach is that generalization bounds and finite sample guarantees are apparent from standard statistical learning theory. For example, in Theorem 13 of Bertsimas and Kallus (2020), a generalization bound is provided with two terms. The first term has a convergence rate of $\mathcal{O}(n^{-\frac{1}{2}})$, while the second term involves the empirical Rademacher complexity of the policy class $\mathcal{P}$.

Both policy optimization and ICEO suffers from a bias introduced by using a hypothesis class, $\mathcal{P}$ and $\mathcal{H}$, respectively. Indeed, a suitable hypothesis class should be computationally tractable and have relatively small Rademacher complexity. For example, a linear class would satisfy both of these. Unfortunately, a hypothesis class that satisfying both of these requirements may introduce a bias due to model mis-specification. For policy optimization methods, there may not exist $\pi^* \in \mathcal{P}$ that is “close enough” to some function outputting values in the set $W(f^*(x)))$. Similarly, for ICEO, there may not exist $f^*_{\mathcal{H}} \in \mathcal{H}$ that fully characteristics the true underlying hypothesis f^*. However, we want to point out that, in practice we generally expect the bias introduced by ICEO to be less than that of policy optimization. It is because ICEO methods models the conditional distribution while the latter models the conditional distribution. Noted that the optimal solution set $W(\cdot)$ may have a complicated structure and require an intricate modeling class $\mathcal{P}$. ICEO does not need to include $w(\cdot)$ as part of the model and only needs to approximate $f^*(\cdot)$ within $\mathcal{H}$, whereas policy optimization necessitates the use of a flexible enough class $\mathcal{P}$ to appropriately model the mapping $w(f^*(\cdot))$. Thus ICEO can be advantageous, especially in situations where the optimization oracle $w(\cdot)$ is complex.

In light of the above discussion, let us solidify our intuition further by comparing the generalization bound of policy optimization with the result provided in Theorem 3. First, note that, according to Proposition 1 and Theorem 1, we may take ρ_n as any sequence approaching zero, for example $\rho_n = \frac{1}{\log(n)}$ suffices. Then according to the result of Corollary 1, the left-hand side quantifies the error between the out-of-sample risk induced by the policy $w_{\rho_n} \circ \hat{f}^n_{\rho_n}$ relative to the policy $w_{\rho_n} \circ f^*_{\mathcal{H}}$ and converges to zero at rate of $\tilde{\mathcal{O}}(n^{-\frac{1}{2}})$. For comparison, policy optimization gives a similar bound of error between the out-of-sample risks induced by a learned policy $\hat{\pi}^n$ and the best policy π^* within the policy class $\mathcal{P}$ with the same convergence rate of $\mathcal{O}(n^{-\frac{1}{2}})$. In both cases, the constant in the convergence bound will be controlled by the Rademacher complexity of the corresponding class, either $\mathcal{H}$ for ICEO or $\mathcal{P}$ for policy optimization. Both methods may introduce some bias, but as we have already argued, we expect it to be easier to achieve smaller bias with smaller complexity for ICEO. In the following, we use an special case (albiet, an extreme one) to illustrate how the flexibility of ICEO can lead to less bias smaller Rademacher complexity (i.e., faster convergence of the finite-sample performance bound).

Example 5. We consider a special case where $\mathcal{X}$ is the simplex Δ_K, which means that the decision-maker has the ability to know the probability vector p directly through the contextual information. In this special case ICEO has nothing to learn, therefore the hypothesis class should be selected as the singleton identity map $\mathcal{H} = \{f : f(p) = p \ \forall p \in \Delta_K\}$. Then, the multi-variate Rademacher complexity $\mathfrak{R}_n(\mathcal{H}) = 0$, and so the regularization coefficient ρ can be set to zero, i.e., $\rho_n = 0$ for all n. On the other hand, policy optimization, which by its presumption is not allowed to use knowledge of the oracle $w(\cdot)$, requires a larger hypothesis class $\mathcal{P}$ to successfully learn the oracle $w(\cdot)$ and so we generally expect $\mathfrak{R}_n(\mathcal{P}) > \mathfrak{R}_n(\mathcal{H}) = 0$.

In addition, considering the fact that, in Corollary 1, the left-hand side quantifies the error between the out-of-sample risk induced by the policy $w_{\rho_n} \circ \hat{f}^n_{\rho_n}$ relative to the policy $w_{\rho_n} \circ f^*_{\mathcal{H}}$. One may notice that $w_{\rho_n} \circ f^*_{\mathcal{H}}$ is not the optimal policy $w \circ f^*_{\mathcal{H}}$, which introduces bias by setting the decision regularization parameter ρ as a small positive value. However, as demonstrated in Proposition 1, $w_{\rho_n} \circ f^*_{\mathcal{H}}$ converges to the optimal policy $w \circ f^*_{\mathcal{H}}$ as ρ_n converges to zero when sample size grows large.

On the other hand, an additional issue that may also limit the applicability of policy optimization is ensuring the feasibility of the output. Unlike in classical machine learning problems like regression or classification, it is difficult to ensure that the hypothesis class $\mathcal{P}$ even guarantees the feasibility of its outputs, whereby $\pi(x) \in S$ for all $\pi \in \mathcal{P}$ and $x \in \mathcal{X}$. This issue does not vanish as sample sizes grows.

Moreover, learning the conditional distribution $f^*(\cdot)$ is arguably more interpretable than policy optimization, as the ICEO approach follows with the estimate-then-optimize structure, and may be easier to accomplish with "simpler" classes $\mathcal{F}$ since the constraints on the outputs are much simpler (we only need to ensure that $f(x) \in \Delta_K$).