

Statistical Foundations for Diagnostics and Efficiency in Data-Driven Optimization

1 Problem Setup

We consider data-driven stochastic optimization problem of the form $\min_{z \in \mathcal{Z}} \mathbb{E}_{\mathbb{P}}[\ell(z; y)]$, where $\ell(\cdot; \cdot)$ is a known cost function, $\mathcal{Z} \subseteq \mathbb{R}^{D_z}$ is the set of feasible decisions, and y is the uncertainty variable. The goal is to use the data $\mathcal{D}$ sampled from $\mathbb{P}$ to identify a decision with the lowest expected cost with respect to the *true* distribution $\mathbb{P}$. This problem is ubiquitous – from empirical risk minimization in machine learning [15], to decision making in supply chain management [31], revenue management [32] and portfolio optimization [3].

Given the importance of the problem, a range of different approaches have been proposed to identify good data-driven decisions. A natural approach is empirical optimization, or sample average approximation, where the expectation with respect to the unknown distribution is replaced with its empirical estimate [30]. This approach can lead to significant overfitting in when the cost function ℓ has high complexity or in the contextual optimization setting where the uncertainty variable depends on contextual covariates. Here we accommodate contextual optimization by extending our notation $z^*(\theta)$ to $z^*(\theta, x)$ where x denotes the contextual covariate. In such situations, one often resorts to a parameterized decision rule $\{z^*(\theta, x) \in \mathcal{Z} | \theta \in \Theta\}$, where $\Theta \subset \mathbb{R}^q$, and the goal reduces to computing the estimate $\hat{\theta} \in \Theta$ using the data $\mathcal{D}$. This parametrized decision rule can be indirectly induced by assuming a parametric model for the uncertainty y [17, 11] or an end-to-end architecture [10] that maps the data into a decision where Θ is the set of parameters that defines the function approximation architecture. Furthermore, one can exploit regularized distributionally robust optimization (DRO) [22] to control the error propagation due to parameter estimation and model bias. In spite of a diversity of solutions, their real-world deployment remains limited without a principled framework that integrates learning algorithms with downstream decisions tailored to specific problem instances. This gap raises several fundamental questions:

How should we evaluate and select among competing learning algorithms under decision-focused objectives? How can we design optimization methods that remain reliable when data are limited, noisy, or subject to change?

My thesis aims to bridge this gap by developing a statistical foundation for evaluating and designing data-driven decision-making methods. In the following, we provide three perspectives on providing diagnostics

across existing data-driven optimization methods and developing new methods with efficiency. (i) **Evaluation Tools:** In Section 2, we compare existing evaluation procedures and develop new tools for estimating and comparing the decision performance; (ii) **Empirical Benchmarks:** In Section 3, we provide the empirical foundation in building benchmarks and software packages to understand decision performances in practice; (iii) **Optimization Design:** Motivated by both efficient evaluation tools and empirical findings, we study algorithms tailored to specific problem structures and data environments with significant statistical gains over general-purpose approaches in Section 4.1 and investigate a real-world Mount Sinai medication inventory control problem in Mount Sinai in Section 4.2.

Setup. Unless specified, we discuss all results under the contextual optimization framework [11, 29]. Here the distribution of $y \in \mathbb{R}^{D_y}$ is a function of $x \in \mathbb{R}^{D_x}$. The true distribution $\mathbb{P}_{y|x}$ is unknown; instead, the decision-maker only has access to $\mathcal{D} = \{(x_i, y_i)\}_{i \in [n]}$ consisting of iid samples from $\mathbb{P} = \mathbb{P}_X \times \mathbb{P}_{Y|X}$ and observes the covariate $X = x$ before making the decision $z \in \mathcal{Z} = \{z \in \mathbb{R}^{D_z} : F_j(z) \leq 0, \forall j \in J_1, F_j(z) = 0, \forall j \in J_2\}$. We summarize data-driven optimization procedures considered in this paper:

Definition 1 (General Data-Driven Optimization Procedures). *Each procedure below outputs a decision $\hat{z}(x)$. More specifically, in (a)(b), $\hat{z}(x) = z^*(\hat{\theta}, x)$ with some parameteric θ .*

(a) *The conditional distribution $Y|x$ is parametrized within the conditional probability distributions $\mathbb{P}_{\Theta|x} := \{\mathbb{P}_{\theta|x} | \theta \in \Theta\}$, and $z^*(\theta, x) \in \operatorname{argmin}_{z \in \mathcal{Z}} \mathbb{E}_{\mathbb{P}_{\theta|x}}[\ell(z; Y)]$.*

(1) *Estimate-then-Optimize (ETO): $\hat{\theta} \in \operatorname{argmin}_{\theta \in \Theta} \mathbb{E}_{\hat{\mathbb{P}}}[\phi(\theta; (x, y))]$, where $\phi(\theta; (x, y)) \in \mathbb{R}$ is a function that does not depend on the objective ℓ . A canonical choice is maximum likelihood estimation, i.e., $\phi(\theta; (x, y)) = -\ln p_{\theta}(y|x)$.*

(2) *Integrated Estimation-and-Optimization (IEO): $\hat{\theta} \in \operatorname{argmin}_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^n \ell(z^*(\theta, x_i); y_i)$, i.e., it optimizes the objective performance among the decision class instead of just model fitting.*

(b) *$z^*(\theta, x)$ is any parameterized family of decision functions specified by the decision-maker. Let $\hat{\mathbb{P}} := \frac{1}{n} \sum_{i=1}^n \delta_{(x_i, y_i)}$ denote the empirical joint distribution.*

(1) *(Parametric) Empirical Optimization (EO): $\hat{\theta} \in \operatorname{argmin}_{\theta} \mathbb{E}_{(X,Y) \sim \hat{\mathbb{P}}}[\ell(z^*(\theta, X); Y)]$;*

(2) *Regularized Empirical Optimization (RO): $\hat{\theta} \in \operatorname{argmin}_{\theta} \{\mathbb{E}_{(X,Y) \sim \hat{\mathbb{P}}}[\ell(z^*(\theta, X); Y)] + \lambda R(z^*(\theta, X))\}$ for some regularizer $R(\cdot)$;*

(3) *Distributionally Robust Optimization (DRO): $\hat{\theta} \in \operatorname{argmin}_{\theta} \max_{d(\mathbb{P}, \hat{\mathbb{P}}) \leq \lambda} \mathbb{E}_{(X,Y) \sim \mathbb{P}}[\ell(z^*(\theta, X); Y)]$ for some distance metric d .*

(c) $\hat{z}(x)$ is some nonparametrized decision rule determined by the weight estimates. Let $\tilde{\mathbb{P}}_x := \sum_{i=1}^n w_{n,i}(x) \delta_{(x_i, y_i)}$ denote the estimated conditional distribution with weights $\{w_{n,i}(x)\}_{i \in [n]}$ depending on $\mathcal{D}$ and x . For each $x \in \mathcal{X}$, $\hat{z}(x)$ is obtained by solving the same optimization problem as (b)(1) - (3) above, while replacing $\hat{\mathbb{P}}$ and $z^*(\theta, X)$ with $\tilde{\mathbb{P}}_x$ and z respectively. For example in Nonparametric-EO, $\hat{z}(x) \in \operatorname{argmin}_{z \in \mathcal{Z}} \sum_{i \in [n]} w_{n,i}(x) \ell(z; y_i)$

Above, the parts (a) and (b) in Definition 1 differ in how $z^*(\theta, x)$ is constructed. In (a), the primitive is a parametric family of distributions and two sub-categories correspond to whether (IEO) or not (ETO) the downstream objective $\ell(\cdot; \cdot)$ is taken into consideration in estimating $\hat{\theta}$. In (b), the primitive focus is the decision function $z^*(\theta, x)$ and $\hat{\theta}$ can be estimated via empirical or robust formulations. In (c), there is no parameter to estimate and the decision rule is a function mapping by solving an weighted EO for each covariate. Next, we show how to apply the various decision rules in Definition 1(a) – (c) in a newsvendor problem (a) – (c) as follows.

Example 1 (Multi-Product Newsvendor). $\ell(z; y) = \sum_{i=1}^{D_y} (c_i z_i - p_i \min\{z_i, y^{(i)}\})$ and $\mathcal{Z} = \{z \in \mathbb{R}_+^{D_y} | \mathbf{1}^\top z \leq C\}$, where c_i, p_i are the known cost and price per unit for the i -th product, $y \in \mathbb{R}^{D_y}$ denotes the product demand and $z \in \mathbb{R}^{D_y}$ denotes the ordering quantity with a capacity constraint.

(a) Assume the demand variables y are independent exponential random variables with parameters $\theta_i \in \mathbb{R}^{D_x}, \forall i \in [D_y]$, i.e., $p_\theta(y|x) = \prod_{i=1}^{D_y} ((\theta_i^\top x) e^{-(\theta_i^\top x) y^{(i)}})$, and $z^*(\theta, x) = \operatorname{argmin}_{\mathbb{P}_\theta(y|x)} [\ell(z; y)]$ Then $\hat{\theta}$ can be estimated using ETO approach (i.e., MLE) or IEO approach above.

(b) EO: Consider $z^*(\theta, x) = \theta^\top x$ with further constraints on θ to ensure feasibility. $\hat{\theta} \in \operatorname{argmin}_\theta \sum_{i=1}^n \ell(z^*(\theta, x_i); y_i)$. We can also apply RO/DRO in an analogous manner to calibrate $\hat{\theta}$.

(c) kNN-EO: $w_{n,i}(x) = \mathbf{1}_{\{x_i \text{ is the } k_n \text{ closest to } x \text{ among } \{x_i\}_{i \in [n]}\}}$ with a hyperparameter k_n .

2 Diagnostic Tools

Given so many potential data-driven optimization models in Definition 1, one fundamental question is: how should we estimate the performance of each decision accurately and efficiently such that we can pick the best one? This helps support decision evaluation and selection for practitioners.

For each decision $\hat{z}(\cdot)$ in Definition 1, we define two terms to evaluate the out-of-sample performance $\mathbb{E}_{\mathbb{P}}[\ell(\hat{z}(X); Y)]$. They are:

- a point estimator that has a *small bias* compared with the expected cost $A := \mathbb{E}_{\mathcal{D}} \mathbb{E}_{\mathbb{P}}[\ell(\hat{z}(X); Y)]$.

- an interval estimator $I(\alpha)$ that *accurately covers* $\mathbb{E}_{\mathbb{P}}[\ell(\hat{z}(X); Y)]$ with $1 - \alpha$ probability, i.e., we aim to satisfy $\mathbb{P}_{\mathcal{D}}(\mathbb{E}_{\mathbb{P}}[\ell(\hat{z}(X); Y)] \in I(\alpha)) \approx 1 - \alpha$, where the outer probability $\mathbb{P}_{\mathcal{D}}$ is with respect to the data $\mathcal{D}$ used to construct $\hat{z}(\cdot)$.

In the contextual optimization literature, the model guarantees are established using bespoke analyses relying on finite-sample generalization bounds [29]. Consequently, the performance bounds tend to be overly conservative for any particular problem instance do not allow one to use them to determine the best solution approach for a given problem instance. Recent work [17, 23, 12] uses convergence rates or stochastic dominance results to compare different data-driven decision schemes; however, these approaches impose assumptions that can be difficult to verify and arguably idealized. In a nutshell, these results do not translate directly into practical methods for selecting data-driven decisions for a particular problem instance and data $\mathcal{D}$.

In statistics and machine learning, to estimate the model performance, cross-validation (CV) is considered the default choice. Leave-one-out CV (LOOCV), which repeatedly evaluates models trained using all but one observation on the left-out observation, is a prime approach; however, it is computationally demanding as it requires model re-training for the same number of times as the sample size. It is unclear whether we can find estimators to reduce computational costs significantly. Because of this, K -fold CV, which reduces the number of model re-training down to K times (where K is typically 5–10), becomes a popular substitute. Despite their wide usage, the statistical benefits of CV have remained understood mostly for specific cost functions (e.g., linear) [2, 5] or fast model convergence rates [35]. In nonparametric regimes, especially those involving slow model convergence rates, i.e., Definition 1(c), their general performances as well as comparisons with the naive empirical estimate (EM) approach, i.e., simply reuses all the same training data for model evaluation, stay essentially open.

Our goal in this section is to fill in the challenging regimes beyond these established results, and as such answer the question: *Are LOOCV and K -fold CV a “must-use” in estimating out-of-sample model performance in general and, if not, then under what situations are they worthwhile? Furthermore, can we design better evaluation tools than all existing counterparts from problem knowledge in the general data-driven optimization?*

In Section 2.1, we conduct a systematic analysis to compare the accuracies in estimating the bias and coverage accuracy of the associated interval estimate for the out-of-sample evaluation using LOOCV, K -fold CV and EM. Our main messages are: First, in terms of these asymptotic criteria, *K -fold CV never outperforms empirical estimate, regardless of the rate at which a parametric or nonparametric model converges*. Second, while LOOCV can have a smaller bias than EM, this bias improvement can be negligible compared to the evaluation variability and therefore, *in a range of important cases, LOOCV again does not outperform EM*. In

particular, we show that decision rules in Definition 1(a)(b), as well as some Definition 1(c) including random forests and kNN with sufficient smoothness, fall into this range. Since LOOCV requires significantly more computation resources, this raises the caution that its use is not always necessary in *decision assessment*.

In Section 2.2, we further investigate the issue of *decision selection* beyond assessment. We propose *Optimizer’s Information Criterion (OIC)* as a new methodology to evaluate the true decision performance for a very general class of decision rules by correcting its bias associated with the EM approach. OIC is a generalization of the celebrated Akaike Information Criterion (AIC) [1], used for model selection, to selecting decision rules in data-driven optimization. OIC directly approximates and eliminates the first-order bias arising from the interplay between model fitting and downstream optimization (Definition 5), and is both computationally efficient, in the sense that, unlike other methods such as cross-validation, one does not need to solve additional optimization problems, and is statistically efficient, in the sense that it removes bias without increasing variability (Theorem 4). Furthermore, OIC serves as an effective decision selection criterion due to its similarity to LOOCV.

Before going to our theoretical results, we discuss some assumptions of the problem here:

Assumption 1 (Problem Smoothness). *The cost function $\ell(z; Y)$ and all function constraints $\{F_j(z)\}_{j \in J_1 \cup J_2}$ are twice differentiable with respect to z for any $z \in \mathcal{Z}$, $\mathbb{P}$ -a.e. For each procedure in Definition 1(a)(b), $z^*(\theta, x)$ is twice differentiable with respect to θ , $\forall x \in \mathcal{X}$.*

Assumption 2 (Computing $\hat{z}(\cdot)$). *For each procedure in Definition 1, the output decision $\hat{z}(x)$ is $o_p(n^{-1})$ -accurate in minimizing its corresponding objective.*

Assumption 3 (Regularity Conditions). *For each procedure in Definition 1, at each $x \in \mathcal{X}$, consider $z_o^*(x)$ as the limiting decision of $\hat{z}(x)$ with infinite data points. Then $z_o^*(x)$ is a KKT point for its corresponding optimization problem. Furthermore, $\nabla_{zz} \mathbb{E}_{\mathbb{P}_{Y|x}}[\ell(z_o^*(x); x)]$ is invertible (or positive definite in Section 2.1).*

Assumptions 1 and 3 require the smoothness of the composite cost function and constraints, which is satisfied for various objectives in OR/ML applications coupled with continuous distributions of $\mathbb{P}$. Assumption 2 requires that the solution to be solved to near-optimality and is satisfied by variants of stochastic approximation algorithms with polynomial number of iterations. These are commonly used in stochastic optimization.

2.1 Existing Evaluation Procedures: Is Cross-Validation always Good?

We present our main results on the evaluation bias and interval coverage for EM, K -fold CV and LOOCV. We denote $c(z)$ as the out-of-sample cost $\mathbb{E}_{\mathbb{P}}[\ell(z(X); Y)]$. For each method $m \in \mathcal{M} = \{EM, LOOCV, KCV\}$ refer-

ring to EM, LOOCV and K -fold CV, we present our considered point estimate $\hat{A}_m$ and $(1-\alpha)$ -interval estimate $I_m(\alpha)$, where the latter are all written in the form $I_m(\alpha) = [\hat{A}_m - z_{1-\alpha/2}\hat{\sigma}_m/\sqrt{n}, \hat{A}_m + z_{1-\alpha/2}\hat{\sigma}_m/\sqrt{n}]$, $\forall m \in \mathcal{M}$ with:

$$\hat{A}_{EM} = c_n(\hat{z}), \quad \hat{\sigma}_p = \sqrt{\frac{1}{n} \sum_{i \in [n]} (\ell(\hat{z}(x_i); y_i) - \hat{A}_{EM})^2}, \quad (1)$$

$$\hat{A}_{KCV} = \frac{1}{n} \sum_{k \in [K]} \sum_{i \in N_k} \ell(\hat{z}^{(-N_k)}(x_i); y_i), \quad \hat{\sigma}_{KCV} = \sqrt{\frac{1}{n} \sum_{k \in [K]} \sum_{i \in N_k} (\ell(\hat{z}^{(-N_k)}(x_i); y_i) - \hat{A}_{KCV})^2}, \quad (2)$$

where $z_{1-\alpha/2}$ is the $(1 - \alpha/2)$ -quantile of the standard normal distribution, $\{N_k, k \in [K]\}$ is the collection of K equal-length partitions of $[n]$ and $\hat{z}^{(-N_k)}(\cdot) := \mathcal{A}(\mathcal{D}_{(-N_k)}; \cdot)$, with $\mathcal{D}_{(-N_k)}$ denoting the data set that leaves out $\{(x_i, y_i)\}_{i \in N_k}$. For the K -fold CV estimates, $\hat{A}_{KCV}$ and I_{KCV} , we always assume K is fixed with respect to n (e.g., $K = 2, 5, 10$). On the other hand, $\hat{A}_{LOOCV}$ and $I_{LOOCV}(\alpha)$ are defined by setting $K = n$ in (2). There are alternative approaches to construct the interval estimates (e.g., nested cross-validation in [6]), but the above are the most natural.

Definition 2 (Convergence Rate). *For $\hat{z}(\cdot)$, we say it has a convergence rate of order $\gamma \in (0, 1/2]$ if $\mathbb{E}_{\mathcal{D}}[\|\hat{z}(x) - z_o^*(x)\|_2] = \Theta(n^{-\gamma})$ for almost every x . Furthermore, we say $\hat{z}(\cdot)$ has a bias and variability convergence rate γ_b, γ_v respectively if $\mathbb{E}_{\mathcal{D}}[\|\mathbb{E}_{\mathcal{D}}[\hat{z}(x)] - z_o^*(x)\|_2] = \Theta(n^{-\gamma_b})$ and $\mathbb{E}_{\mathcal{D}}[\|\hat{z}(x) - \mathbb{E}_{\mathcal{D}}[\hat{z}(x)]\|_2] = \Theta(n^{-\gamma_v})$ for almost every x . Consequently, $\gamma = \min\{\gamma_b, \gamma_v\}$.*

The overall convergence order γ of $\hat{z}(\cdot)$ is determined by both its bias γ_b and variability γ_v , whichever dominates. For parametric decisions in Definition 1(a)(b), we naturally have $\gamma = \gamma_v = 1/2$. For nonparametric decisions in Definition 1(c), both γ_b and γ_v depend on the hyperparameter configuration and are often smaller than $1/2$.

Theorem 1 (Evaluation Bias). *Suppose Assumptions 1, 2 and 3 hold. Recall γ, γ_v in Definition 2. $\mathbb{E}[c(\hat{z}) - \hat{A}_{EM}] = \Theta(n^{-2\gamma_v}) > 0$, $\mathbb{E}[c(\hat{z}) - \hat{A}_{KCV}] = \Theta(n^{-2\gamma}) < 0$, $\mathbb{E}[c(\hat{z}) - \hat{A}_{LOOCV}] = o(n^{-1}) < 0$.*

Theorem 2 (Coverage Validity of Interval Estimate). *Suppose Assumptions 1, 2 and 3 hold.*

- If $\gamma > 1/4$, then: $\lim_{n \rightarrow \infty} \mathbb{P}(c(\hat{z}) \in I_m) = 1 - \alpha$, for $m \in \{EM, KCV, LOOCV\}$.
- If $\gamma \leq 1/4$, then $\lim_{n \rightarrow \infty} \mathbb{P}(c(\hat{z}) \in I_{KCV}) < 1 - \alpha$, $\lim_{n \rightarrow \infty} \mathbb{P}(c(\hat{z}) \in I_{LOOCV}) = 1 - \alpha$. For EM, $\lim_{n \rightarrow \infty} \mathbb{P}(c(\hat{z}) \in I_{EM}) \leq 1 - \alpha$, where equality holds if and only if $\gamma_v > 1/4$.

In the following, we use Theorems 1 and 2 to compare EM, K -fold CV and LOOCV, and highlight our novelty relative to what is known in the literature. In a nutshell, the regime $\gamma \leq 1/4$ has been wide open and comprises our major technical contribution.

Comparing EM and K -fold CV. Theorems 1 and 2 together stipulate that *EM is always no worse than K -fold CV* in terms of estimating out-of-sample performance. In terms of *bias*, both procedures are $\Theta(n^{-1})$ in parametric models. However, this bias size is unknown for general nonparametric models in the literature. For these models, our results show that the EM bias is $\Theta(n^{-2\gamma_v})$ which is no bigger than that of K -fold CV since $\gamma \leq \gamma_v$. This behavior arises because, even though EM incurs an underestimation of $\Theta(n^{-2\gamma_v})$ due to the reuse of training and evaluation set, K -fold CV loses efficiency due to a loss of training sample from the data splitting, thus leading to an even larger bias of $\Theta(n^{-2\gamma})$. In terms of *interval coverage*, While EM and K -fold CV both exhibit asymptotically exact coverage in the case $\gamma > 1/4$, their coverages differ for nonparametric models, with EM still always no worse than K -fold CV. Specifically, when $\gamma \leq 1/4$, K -fold CV incurs invalid coverage, whereas EM still yields valid coverage as long as $\gamma_v > 1/4$. This is because, in this regime, the bias of K -fold CV is bigger than its variability to affect coverage significantly while the bias of EM remains small enough to retain coverage validity. In the literature, [34, 2] further show CV intervals can provide coverage guarantees when $\gamma > 1/4$, but they do not touch on the case $\gamma \leq 1/4$. Moreover, all the literature above do not demonstrate an explicit difference between K -fold and LOOCV in their results. From these, our results on the regime $\gamma \leq 1/4$ where we characterize and conclude the difference between K -fold and LOOCV appear the first in the literature.

Comparing EM and LOOCV. When comparing with EM, LOOCV has a smaller, and pessimistic, bias $o(1/n)$. However, this bias improvement can be negligible compared to the evaluation variability captured in interval coverage, specifically when $\gamma > 1/4$ (which includes all parametric models) and when $\gamma \leq 1/4, \gamma_v > 1/4$. In the latter case in particular, the bias of EM, even though larger than LOOCV, is small enough to ensure valid coverage.

We visually summarize our discussions on biases and interval coverages in Figure 1. In particular, Figure 1(a)(b) display our new contributions on both EM and CVs under a slow rate γ . We also see that EM intervals provide valid coverages for (b)(c), while K -fold CV is only valid for (c) and LOOCV is valid across (a)(b)(c).

We exemplify the above insights with specific models in Example 1. In the EO solution from Example 1(b), $\gamma = \gamma_v = 1/2$, and all of EM, K -fold CV and LOOCV provide valid coverages for $c(\hat{z})$; In the kNN solution from Example 1(c), under the hyperparameter k_n , $\gamma_v = -\log k_n / (2 \log n)$ and $\gamma_b = \log(k_n/n) / (4 \log n)$ (from Chapter 6 in [14]). If $k_n = \omega(\sqrt{n})$, then LOOCV and EM provide valid coverages for $c(\hat{z})$; otherwise, only LOOCV provide valid coverages. Besides:

Example 2 (Forest Learner). *Consider a forest $\mathcal{F} = \{\tau_1, \dots, \tau_T\}$, where each $\tau_i : \mathbb{R}^{D_x} \mapsto \{1, \dots, L_i\}$ is a (tree) partition of $\mathbb{R}^{D_x}$ into L_i regions. Then $w_{n,i}(x) = \sum_{j=1}^T \mathbf{1}_{\{\tau_j(X_i)=\tau_j(x)\}}/T$ where the hyperparameter is*

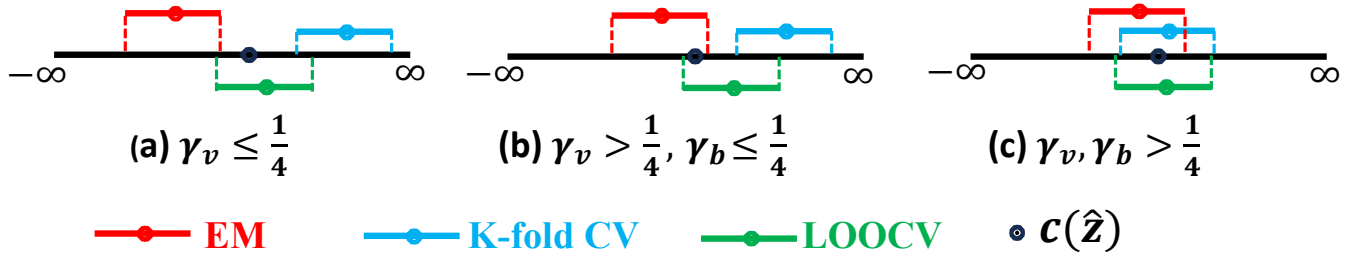


Figure 1: Concept plots of interval coverages of $c(\hat{z})$, for our considered approaches across decision rates, where the black line represents the value of the expected cost and a point is considered covered if it falls within the corresponding interval.

the subsampling ratio $\beta < 2/3$ in each tree. Specifically, $\gamma_v = (1 - \beta)/2$ from [35] and $\gamma_b < 1/6$. Then LOOCV and EM provide valid coverages for $c(\hat{z})$.

Our results suggest the following practical guidance in choosing evaluation procedures for decision assessment. In terms of the magnitude of bias, LOOCV is always smaller than EM, while EM is no larger than K -fold CV. Despite this bias ordering, the adoption of a method over another should also take into account the variability and computational demand. For parametric decisions and nonparametric decisions with a fast convergence rate, EM is preferred both computationally and statistically in both point estimates and uncertainty quantification. In the full version [20], we provide discussions of other CV procedures and extensions to distribution shifts.

We caution the distinction between decision assessment and selection, namely the selection of hyperparameters among a class of data-driven decision and between decision classes. Our decision assessment comparisons may or may not translate into the performances in decision selection. Intuitively, the evaluation bias may be correlated among different hyperparameter values and ultimately leading to a low error in the selection task. In these cases, LOOCV has the smallest evaluation bias is preferred and leads to better decision performance under the same variability.

2.2 New Tool: Optimizer's Information Criterion

In this section, we introduce a new point estimator for $A = \mathbb{E}_{\mathcal{D}} \mathbb{E}_{\mathbb{P}}[\ell(z^*(X); Y)]$, namely Optimizer's Information Criterion (OIC), that evaluates decision performance and improves the *decision assessment* and *selection* for a wide range of decision procedures in Definition 1. We first describe the results for Definition 1(a)(b) and highlight extensions for Definition 1(c).

Before we present OIC, we define influence functions that quantify the impact of a single data point on the decision procedure to express the evaluation bias of different procedures unifiedly.

Definition 3 (Influence Function). *The influence function $\text{IF}(\cdot)$ is defined as follows:*

$$\text{IF}((x, y); T, \mathbb{P}) = \lim_{\epsilon \rightarrow 0^+} \frac{T(\epsilon \delta_{(x, y)} + (1 - \epsilon)\mathbb{P}) - T(\mathbb{P})}{\epsilon},$$

where the calibration function $T : M(\mathcal{X} \times \mathcal{Y}) \mapsto \Theta$ corresponds to the optimization procedure in Definition 1(a)(b) such that the estimate $\theta = T(\mathbb{P})$ when $(X, Y) \sim \mathbb{P}$.

Influence functions were introduced originally in the robust statistics literature [9]. Recall that the calibration function T is different for each cost function and the optimization procedure. Therefore, the corresponding $\text{IF}(\cdot)$ is also different in each case. We abbreviate $\text{IF}((x, y); T, \mathbb{P})$ as $\text{IF}(x, y)$ when there is no confusion. The influence function has a closed-form expression across different procedures. Below, we provide expressions for the influence functions for IEO/EO and defer full closed-form expressions of other procedures in [19].

Definition 4 (Influence Functions of Procedures in IEO/EO). $\text{IF}(x, y) := -P(I_{\ell, \alpha}(\theta^*))^\dagger P \nabla_\theta \ell(z^*(\theta^*, x); y)$, where $I_{\ell, \alpha}(\theta^*) = \nabla_{\theta\theta}^2 \mathbb{E}_{\mathbb{P}}[\ell(z^*(\theta^*, X); Y)] + \sum_{j \in B^*} \alpha_j^* \nabla_{\theta\theta}^2 F_j(z^*(\theta^*, x))$, α is the Lagrangian multiplier corresponding to the binding constraint $B^* = \{j \in J : F_j(z^*(\theta^*, x)) = 0\}$, and $P = I - C^\top (CC^\top)^\dagger C$, with $C = [\nabla_\theta F_j(z^*(\theta^*, x))]_{j \in B^*}$ and † denotes the pseudo-inverse. Specifically, when $\mathcal{Z} = \mathbb{R}^{D_z}$, $\text{IF}(x, y) := -I_\ell(\theta^*)^{-1} \nabla_\theta \ell(z^*(\theta^*, x); y)$.

This characterizes the dominating term of the bias for each procedure:

Theorem 3 (Bias Characterization). $A = \mathbb{E}[\hat{A}_{EM}] - \mathbb{E}_{\mathbb{P}}[\nabla_\theta \ell(z^*(\theta^*, x); y)^\top \text{IF}(x, y)] + o(1/n)$.

Denote $A_c = -\mathbb{E}_{\mathbb{P}}[\nabla_\theta \ell(z^*(\theta^*, x); y)^\top \text{IF}(x, y)]$. In unconstrained EO, $A_c = n^{-1} \text{Tr}[I_\ell(\theta^*)^{-1} J_\ell(\theta^*)]$ with $J_\ell(\theta^*) = \mathbb{E}_{\mathbb{P}}[\nabla_\theta \ell(z^*(\theta^*); y) \nabla_\theta \ell(z^*(\theta^*); y)^\top]$. When $\ell(z^*(\theta); y)$ is convex in θ , $I_\ell(\theta^*)$ is positive semidefinite, and therefore $A_c > 0$, i.e., the bias in $\hat{A}_{EM}$ is optimistic. In general, however, the bias A_c can take any sign and our approach applies regardless of its sign.

Motivated by Theorem 3, we define our bias-corrected estimator OIC $\hat{A}_{OIC}$ for A by replacing the unknown θ^* , IF and $\mathbb{P}$ with empirical or estimated counterparts.

Definition 5 (Optimizer's Information Criterion (OIC)). *The OIC*

$$\hat{A}_{OIC} := \hat{A}_{EM} - \underbrace{\frac{1}{n^2} \sum_{i=1}^n \nabla_\theta \ell(z^*(\hat{\theta}, x_i); y_i)^\top \widehat{\text{IF}}(x_i, y_i)}_{\hat{A}_c} \quad (3)$$

where $\widehat{\text{IF}}(x, y) := \text{IF}((x, y); \hat{T}, \hat{\mathbb{P}})$, and $\hat{T}$ is the map that takes distribution $\mathbb{P}$ to a $o_p(n^{-1})$ -optimal parameter $\hat{\theta}$.

In Definition 5, $\widehat{\text{IF}}$ is defined generally in terms of the approximate map $\hat{T}$ and can be expressed more explicitly for decision procedures described in Definition 1.

Definition 6 (Estimated Influence Functions of Procedures in IEO/EO). $\widehat{\text{IF}}(x, y) = -\hat{P}I_{\ell, \hat{\alpha}}(\hat{\theta})^\dagger \hat{P}\nabla_\theta \ell(z^*(\hat{\theta}, x); y)$, where $\hat{P} = I - \hat{C}^\top (\hat{C}\hat{C}^\top)^\dagger \hat{C}$, $\hat{C} = [\nabla_\theta F_j(z^*(\hat{\theta}, x))]_{j \in \hat{B}}$ and $\hat{B} = \{j \in J : F_j(z^*(\hat{\theta}, x)) = 0\}$ denotes the set of active constraints at $\hat{\theta}$. Any consistent approximation $I_{\ell, \alpha}(\hat{\theta})$ is valid. When ℓ is twice differentiable, $\hat{I}_{\ell, \alpha}(\hat{\theta}) = \frac{1}{n} \sum_{i=1}^n \nabla_{\theta\theta}^2 \ell(z^*(\hat{\theta}, x_i); y_i) + \sum_{j \in \hat{B}} \hat{\alpha}_j \nabla^2 F_j(z^*(\hat{\theta}, x_i))$, where $\hat{\alpha}$ denote the Lagrange multipliers, and reduces to $\frac{1}{n} \sum_{i=1}^n \nabla_{\theta\theta}^2 \ell(z^*(\hat{\theta}, x_i); y_i)$ when $\mathcal{Z} = \mathbb{R}^{D_z}$.

Then, the bias formula of each procedure in Definition 1 can be directly calculated by plugging each estimated influence function into (3). We now present the main properties of OIC as follows.

Theorem 4 (Statistical Properties of OIC). *Suppose Assumptions 1, 2 and 3 hold. Then the OIC estimator $\hat{A}$ satisfies $\mathbb{E}_{\mathcal{D}}[\hat{A}_{OIC}] = A + o(1/n)$ and the estimated bias term $\text{Var}[\hat{A}_c] = \Theta(1/n^2)$.*

Theorem 4 states that $\hat{A}$ is an unbiased estimate for A up to $o(1/n)$ as $\hat{A}_c$ is an unbiased estimator of the bias up to the same order of error with a small variability.

Interpretation. We explain some insights on the bias correction term in OIC, in particular highlighting the interaction between the parameter space Θ and the decision space $\mathcal{Z}$. First, we can rewrite: $\hat{A}_c = \mathbb{E}_{\hat{\mathbb{P}}} \left[\|\nabla_\theta \ell(z^*(\hat{\theta}, x); y)\|_2 \|\widehat{\text{IF}}(x, y)\|_2 \cos(-\nabla_\theta \ell(z^*(\hat{\theta}, x); y), \widehat{\text{IF}}(x, y)) \right]$, where $\cos(-\nabla_\theta \ell(z^*(\hat{\theta}, x); y), \widehat{\text{IF}}(x, y))$ is the cosine of the angle between $-\nabla_\theta \ell(z^*(\hat{\theta}, x); y)$ and $\widehat{\text{IF}}(x, y)$. This expansion provides a useful interpretation of OIC as correcting the self-influence of the training sample on the evaluated decision. The vector $\widehat{\text{IF}}(x_i, y_i)$ captures how much the i -th sample changes the fitted parameter $\hat{\theta}$, while $\nabla_\theta \ell(z^*(\hat{\theta}); y_i)$ captures how sensitive the evaluated objective at y_i is to this change. Their inner product therefore quantifies the extent to which the same observation both influences the decision and benefits from being used in the evaluated objective.

This decomposition also clarifies when the bias is large. First, the bias is possibly high when the decision rule is “complex” in that $\{\|\nabla_\theta \ell(z^*(\theta); \cdot)\|_2 : \theta \in \Theta\}$ has a high complexity measure, i.e., a large Vapnik–Chervonenkis dimension, metric entropy, or localized Rademacher complexity such that $\mathbb{E}_{\hat{\mathbb{P}}}[\|\nabla_\theta \ell(z^*(\hat{\theta}); y)\|_2]$ is large. Second, a high-dimensional parametric space Θ also causes a large bias, since it typically increases $\mathbb{E}_{\hat{\mathbb{P}}}[\|\widehat{\text{IF}}(y)\|_2]$, as a single sample can significantly affect the parameter estimate. Finally, the magnitude of the bias depends on the cosine similarity between $\widehat{\text{IF}}(y)$ and $\nabla_\theta \ell(z^*(\hat{\theta}); y)$. The bias can be small if the vector flows $\nabla_\theta \ell(z^*(\hat{\theta}); y)$ and $\widehat{\text{IF}}(y)$ are approximately orthogonal, which corresponds to the case where the estimation of $z^*(\theta)$ and $\hat{\theta}$ are nearly uncorrelated.

Comparing OIC with existing methods. In this part, we describe the rationale and benefits of OIC in more detail by elaborating its comparison against existing approaches.

Proposition 1 (Relationship with other Evaluation Procedure). *We have the following:*

1. *Akaike Information Criterion: Consider the parametric family of densities $\mathcal{P} = \{p_\theta(\cdot) \mid \theta \in \Theta\}$. Suppose $\mathbb{P}$ is well-specified, i.e., the corresponding p.d.f. $p^* \in \mathcal{P}$. Then AIC is OIC applied to EO (see Definition 1(b)(1)) using the function $\ell(z^*(\theta, x); y) = -\ln p_\theta(y)$, and $A_c = D_\theta/n$, where D_θ denotes the dimension of θ .*
2. *LOOCV: Across Definition 1(a)(b), $\hat{A}_{LOOCV} - \hat{A}_{OIC} = o_p(1/n)$.*

As a summary, we provide a complete computational-statistical tradeoff picture when using OIC, cross-validation, empirical estimate for comparing decision procedures in Definition 1(a)(b). OIC is the most general bias correction formula known to-date that applies across different data-driven optimization problems with general decision rules and cost functions. It recovers previous similar formulas like AIC and its variants as special cases, and therefore, inherits the benefits of these previous formulas, but with the additional benefit of being substantially more general in that it can be used for decision selection instead of just model selection. Since OIC is similar to LOOCV, it can also be used to select optimal hyperparameters well.

OIC requires access to the gradient to compute the bias, and sometimes the Hessian of the cost function and the influence function of the decision rule with respect to its parametrization. Nonetheless, in the full version [19], we show that for a wide range of data-driven optimization formulations, these quantities are explicitly computable so that OIC is an efficiently implementable approach. Also, note that OIC does not improve the mean squared error over the naive empirical estimator $\hat{A}_{EM}$ – this is because any estimator $\tilde{A}$ based on $\mathcal{D}$ incurs $\mathbb{E}_{\mathcal{D}}[(\tilde{A} - A)^2] = \Omega(1/n)$ [36] due to the statistical variability in $\mathcal{D}$. Nonetheless, from Theorem 4 we have that OIC alleviates the evaluation bias $A - \mathbb{E}[\hat{A}_{EM}]$ without hurting the mean squared error. The bias correction leads to a marked improvement in decision selection. Without the bias correction provided by OIC, the empirical estimate $\hat{A}_{EM}$ always favors EO which is not the optimal decision rule when the sample size is only moderately large [21, 33].

Extension. OIC can also be applied to decision procedures in Definition 1(c) or some cost functions that violate Assumption 3. In fact, the proof of OIC relies on the second-order Taylor expansion at θ^* of both $(1/n) \sum_{i=1}^n (\ell(z^*(\hat{\theta}, x_i), y_i) - \ell(z^*(\theta^*, x_i), y_i))$ and $\mathbb{E}_{\mathbb{P}}[\ell(z^*(\theta^*, x), y) - \ell(z^*(\hat{\theta}, x), y)]$, which further builds on two following key equations.

$$\exists \theta^*, \quad \hat{\theta} - \theta^* = \frac{1}{n} \sum_{i=1}^n \text{IF}(x_i, y_i) + o_p(n^{-1/2}). \quad (4)$$

$$\mathbb{E}_{\mathcal{D}}[\|\widehat{\text{IF}}(x, y)\|_2^2] - \|\text{IF}(x, y)\|_2^2 = o(1), \forall (x, y) \in \mathcal{X} \times \mathcal{Y}. \quad (5)$$

Table 1: Comparisons among different evaluation methods. “ $\#\hat{\theta}$ to compute” and “ $\#z^*(\cdot)$ to compute” refer to the numbers of times we need to compute $\hat{\theta}$ using (some subset of) data and to evaluate $z^*(\theta)$ at different θ respectively, “Model-free” means that the evaluation method does not require knowledge of the cost function and the decision rule.

	LOOCV	K-Fold CV	AIC-type	OIC
$\#\hat{\theta}$ to compute in Definition 1(a)(b)	n	K	1	1
$\#z^*(\cdot)$ to compute	n	K	1	1
Model-free	✓	✓	✗	✗
Remove $\Theta(n^{-\gamma})$ bias	✓	✗	✓	✓
Encompass constrained problems	✓	✓	✗	✓
Explicit bias formula	✗	✗	✓	✓

(4) is regarding the asymptotic linear expansion of $\hat{\theta}$ and (5) is regarding the approximation of the influence function with the estimated one. Across general procedures beyond Definition 1(b)(c) under smooth conditions, as long as (4) and (5) are satisfied, the bias correction expressions are valid.

First, we show how OIC can be applied to the kNN-EO in Example 1(c):

Example 3 (OIC for kNN-EO). Recall the kNN-EO in Example 1(c). When $\mathcal{Z} = \mathbb{R}^{D_z}$, define

$$\hat{A}_{OIC} := \frac{1}{n} \sum_{i=1}^n \ell(\hat{z}(x_i); y_i) - \frac{1}{n^2} \text{Tr} \left[\nabla_{zz}^2 \mathbb{E}_{\mathbb{P}}[w_{n,i}(x) \ell(\hat{z}(x); Y)] \left(\sum_{i=1}^n w_{n,i}(x_i) \nabla_z \ell(\hat{z}(x_i); y_i) \nabla_z \ell(\hat{z}(x_i); y_i)^\top \right) \right]. \quad (6)$$

Then $\mathbb{E}_{\mathcal{D}}[\hat{A}_{OIC}] = A + o(n^{-2\gamma_v})$. When $\ell(z; y)$ is twice differentiable with respect to z , $\nabla_{zz}^2 \mathbb{E}_{\mathbb{P}}[w_{n,i}(x) \ell(\hat{z}(x); Y)]$ in (6) can be replaced by: $\frac{1}{n} \sum_{i=1}^n w_{n,i}(x_i) \nabla_{zz}^2 \ell(\hat{z}(x_i); y_i)$.

When verifying the result, (4) is replaced with: $\hat{z}(x) - z^*(x) = (1/n) \sum_{i=1}^n \text{IF}(x_i, y_i) + o_p(n^{-\gamma_v})$. Certainly, the order of the bias, $O(n^{-2\gamma_v})$ is specific to each procedure, and can be much larger than that for parametric optimization procedures since γ_v can be much smaller than $\frac{1}{2}$. In general, to apply OIC to procedures in Definition 1(c), we need to get access to how $\{w_{n,i}(x)\}_{i \in [n]}$ are computed, which is difficult to obtain in Example 2.

Second, we show how OIC can be applied to linear objectives by verifying (4) and (5) hold for linear objectives suitably regularized by a proximal function. We consider the non-contextual setup for simplicity, with the EO solution in the linear objective: $\hat{z} \in \arg\min_z \{\mathbb{E}_{\hat{\mathbb{P}}}[Y]^\top z, \text{ s.t., } Fz = g, z \geq 0\}$. It is clear that the objective of the linear programming violates Assumption 3. Instead, consider the following entropy-regularized convex optimization problem

$$\hat{z}(\eta_n) \in \arg\min_{\theta} \left\{ \mathbb{E}_{\hat{\mathbb{P}}_n}[Y]^\top z + \eta_n^{-1} \left(\sum_i z_i \log z_i \right), \text{ s.t. } Fz = g \right\}, \quad (7)$$

where $\eta_n > 0$ is a penalization parameter. The entropy-regularized objective $\tilde{\ell}(z; y) = y^\top z + \eta_n^{-1} \sum_i z_i \log z_i$ satisfies Assumption 3 and Eqns (4) and (5) hold for some $\hat{z}$ and z^* . Moreover, for some appropriate η_n , the entropy-regularized solution is close to the original LP solution. That is:

Example 4 (OIC for LP). Let $\hat{z}(\eta_n)$ denote the solution of (7). Define $\hat{A}_{OIC}(\eta_n) = \mathbb{E}_{\hat{\mathbb{P}}_n}[Y]^\top \hat{z}(\eta_n) - \frac{1}{n^2} \sum_{i=1}^n y_i^\top \widehat{\mathbf{IF}}(y_i) - \eta_n^{-1} \sum_{j \in [D_\xi]} \hat{z}_j(\eta_n) \log \hat{z}_j(\eta_n)$, where η_n of order $\log(n)$ and $\widehat{\mathbf{IF}}(y) = -\eta_n \cdot \text{diag}(\hat{z}(\eta_n)) \cdot (I - F^\top (F \text{diag}(\hat{z}(\eta_n)) F^\top)^{-1} F \text{diag}(\hat{z}(\eta_n))) (y - \mathbb{E}_{\hat{\mathbb{P}}_n}[Y])$. Then $\mathbb{E}[\hat{A}_{OIC}(\eta_n)] = \mathbb{E}_{\mathcal{D}} \mathbb{E}_{\mathbb{P}}[y^\top \hat{z}] + \tilde{o}(1/n)$, where $\hat{z}$ is the EO solution of LP.

Note that the Lagrangian multiplier of this constrained LP problem has a closed-form representation of the problem parameters, which allows us to simplify the influence function formula.

In the full version [19], we compare OIC with other evaluation procedures including Approximate-Leave-One Out (ALO, [7]) and extend to the generalized contextual optimization and risk functions. Overall, for practitioners to conduct decision assessment and evaluation, compared with all other evaluation procedures, OIC is particularly useful when (i) the solution of the repeated optimization problem is substantially more expensive than computing gradients, Hessians, or influence-function approximations and (ii) one is interested in evaluating the contextual optimization and risk functions, and in cases not covered by existing bias formulas.

3 Empirical Benchmarks

Beyond the statistical evaluation needs to find the best one, a holistic empirical benchmark is essential for diagnosing algorithmic performance in real-world deployments. To date, many existing data-driven optimization design lack rigorous empirical validation, making it difficult to understand their performance limits or to identify the true drivers of performance improvements in practice.

Consider a canonical data-driven optimization problem, i.e., binary classification in the machine learning. Here, X denotes features, Y denotes labels, $\hat{z}(\cdot) : \mathcal{X} \mapsto \mathcal{Y}$ denotes the predictor trained on the training data $\mathcal{D}$ from the source distribution $\mathbb{P}$ and $\ell(\cdot, \cdot) \in \{0, 1\}$ denotes the 0-1 loss. In the training, we replace it with some training loss $\ell_{tr} \in [0, 1]$. Within this problem, practitioners may often want to find the best classifier that maximize the accuracy on some target distribution $\mathbb{Q} \neq \mathbb{P}$, expressed as $\mathbb{E}_{\mathbb{Q}}[\ell(\hat{z}(x); Y)]$, a phenomenon called *distribution shifts*. For example, in healthcare services, predictors of patient mortality trained using data from one hospital in the past may not reflect their real deployment environments in other hospitals in the future.

Algorithmic solutions in Definition 1(b)(2) or (b)(3), like DRO, are proposed to handle distribution shifts. These methods typically require modelers to (i) articulate the set of distribution shifts of interest through some *ambiguity sets* $\{\mathbb{P} : d(\mathbb{P}, \hat{\mathbb{P}}) \leq \lambda\}$ and (ii) explicitly optimize worst-case performance within these sets. Empirically, they show better performance under specific distribution shifts posited by researchers. These shifts typically constructed by specifying a priori assumptions about how distributions may change rather than

being informed by empirical data. When modeling assumptions are not validated, the resulting predictions can perform poorly [16].

Motivated by empirical benchmarking practices that have driven progress in machine learning [28], we advocate for a fresh perspective in this section. Rather than starting with mathematical assumptions, we study distribution shifts as they manifest in the data, and then evaluate every available method across diverse model classes to empirically determine what works. Section 3.1 describes the benchmark design including the data source, method design and software implementation; Section 3.2 answers some empirical questions based on our benchmark: (i) *how distribution shifts look like in OR datasets*; (ii) *how algorithms behave under these distribution shifts* and (iii) *what components drives algorithmic performance improvements*. This inspires us for better algorithmic development through a data-driven understanding of distribution shifts.

3.1 Benchmark Setup

We focus on the data structured in a table format (“tabular data”), where each row corresponds to a data instance (e.g., a person) and each column corresponds to a specific feature describing the instance (e.g., age, salary, product category). Such data modalities dominate electronic healthcare records, financial data, online advertising and many OR/OM domains.

Dataset Overview. We design our distribution shift benchmark to cover diverse prediction tasks in *socioeconomic, traffic, education and healthcare* systems from 8 real-world tabular datasets. For economic data, we use **ACS Income**, **ACS Mobility**, and **ACS Public Coverage** datasets from the US-wide ACS PUMS data that predicts individual’s income, residential address and health insurance. We primarily focus on spatial shifts across different states/cities. For traffic data, we use **US Accident** and **Taxi** that predicts the severity of accidents and taxi duration. To complement spatial shifts, we derive an **ACS Time** task based on the **ACS Public Coverage** dataset, where there are temporal shifts between different years (2010 to 2021). We also extract data from **Diabetes**, **ASSISTments** and **College** datasets to represent demographic and socioeconomic/cultural shifts in education and healthcare systems. In alignment with the desiderata, we carefully design 10 settings, summarized in Table 2, each with one source domain and multiple target domains, resulting in *172 source-target pairs*. This helps capture diverse spatiotemporal distribution shifts on geographic, demographic, temporal and socioeconomic diversity and ensures that our evaluations leads to robust insights into model behavior under diverse conditions. we also selected a single target domain in each setting that exhibits relatively large distribution shifts, referred to as the *10 selected source-target pairs*.

Distribution Shift Types. Here, we decompose distribution shifts between $\mathbb{P}$ and $\mathbb{Q}$ further into X -shifts, where the covariate marginal distribution changes, and $Y|X$ -shifts, where the conditional distribution changes.

Table 2: Overview of datasets and 10 settings

Desiderata	ID	Dataset	Source Domain	Selected Target Domain	Total Number of Target Domains	Shift Pattern	Outcome
Geographic	1	ACS Income	California (CA)	Puerto Rico (PR)	50	$Y X \gg X$	Income $\geq 50k$
	2	ACS Mobility	Mississippi (MS)	Hawaii (HI)	50	$Y X \gg X$	Residential Address
	3	Taxi	New York City (NYC)	Botogá (BOG)	3	$Y X \gg X$	Duration ≥ 30 min
	4	ACS Pub.Cov	Nebraska (NE)	Louisiana (LA)	50	$Y X > X$	Public Ins. Coverage
	5	US Accident	California (CA)	Oregon (OR)	12	$Y X > X$	Severity of Accident
Temporal	6	ACS Pub.Cov	2010 (NY)	2017 (NY)	3	$Y X < X$	Public Ins. Coverage
Demographic	7	ACS Income	Younger	Older	1	$Y X \ll X$	Income $\geq 50k$
	8	Diabetes	White	Others	1	$Y X \gg X$	Been told have diabetes
Socioeconomic	9	ASSISTments	700 Schools	1 New School	1	$Y X > X$	Next Answer Correct
	10	College	Mainstream Institutions	Non-traditional Institutions	1	$Y X > X$	Completion rate $\geq 50\%$
Total		8 datasets	10 settings	10 <i>selected</i> source-target pairs	172 <i>all</i> source-target pairs	8 prediction goals	

Considering all the 172 source-target pairs in Table 2, we find that performance degradation under shifts is overwhelmingly attributed to $Y|X$ -shifts. For each source-target pair, we use DISDE [8] to attribute the total performance drop to X -shifts and $Y|X$ -shifts. Among source-target pairs whose performance degradation is larger than 8 percentage points (67 out of 172 pairs), 87.2% of them have over 50% of the performance degradation attributed to $Y|X$ -shifts (and 70.2% of them have over 60% of the gap attributed to $Y|X$ -shifts). Unlike previous OR/OM experiments that construct *controlled* shifts where datasets are perturbed to match specific ambiguity set assumptions, the distribution shifts we examine are larger in scale compared to those typically considered in the literature (172 distribution pairs) and quantitatively more complex.

Method Overview. We selected 45 methods that cover a wide range of learning approaches and model classes for tabular data. These methods, listed in Table 3, are organized into eight categories given their model classes $z^*(\cdot)$ and distance choices d in their distributionally robust specifications. Equipped with the 10 settings in Table 2 that represent diverse distribution shift patterns across X - and $Y|X$ -shifts, we conduct an extensive grid search across its hyperparameters for each method and evaluate the performance of over 90,000 different configurations in total.

Table 3: Method specification of benchmark. Each bracket reports the number of methods in that category. Balanced/Fairness learning methods are included due to their widespread use in ML.

Category Name	$\hat{z}(\cdot)$	d
Basic EO (6)	LR, SVM, Kernel-SVM, NN (different numbers of layers)	NA
Tree-based EO (4)	GBM, Randomforest, LightGBM, XGBoost	NA
Linear-DRO (13)	Linear	Wasserstein, χ^2 , KL, CVaR,...
Kernel-DRO (4)	RKHS Kernel	Wasserstein, χ^2 , KL, CVaR
Tree-DRO (4)	LightGBM, XGBoost	CVaR, KL
NN-DRO (8)	NN with different numbers of layers	CVaR, χ^2
Balanced-Learning (4)	LightGBM, XGBoost	Subsampling and Reweighting
Fairness-Learning (4)	LightGBM, XGBoost	Feature Intervention

Software Implementation To support reproducible benchmarking, we developed an open-source Python

package, `dro` [25], that implements modular DRO formulations for modern ML workflows. The package integrates with `scikit-learn` and `PyTorch`, supports flexible losses and model classes, and enables scalable experimentation across the benchmark settings. In `python-dro.org`, we provide comprehensive documentation, a modular architecture, and seamless integration with modern ML workflows in the library to both researchers and practitioners.

3.2 Primary Findings

The benchmark enable three key contributions that advance our understanding of both robust algorithms and distribution shifts:

Understanding when blind benchmarking fails: the role of shift type. To illustrate the need for careful modeling of distribution shifts, we first show that even widely cited empirical findings in the ML literature fail to generalize without a grounded understanding of distribution shifts. In particular, Miller et al. [26] observed that across vision and language tasks, out-of-distribution performance strongly correlates with in-distribution performance (“accuracy-on-the-line”), suggesting practitioners can reliably use validation performance to predict deployment performance.

Our analysis reveals that the “accuracy-on-the-line” phenomenon fundamentally breaks down on the benchmark we consider. To quantify the relationship between source and target accuracy, we fit a linear regression line by regressing target accuracy on source accuracy using the data points and report the coefficient of determination, R^2 across all 172 pairs in 10 settings. In Figure 2(a)-(d), we plot the source accuracy (on the x -axis) against the target accuracy (on the y -axis) within one selected pair in 4 of 10 settings.

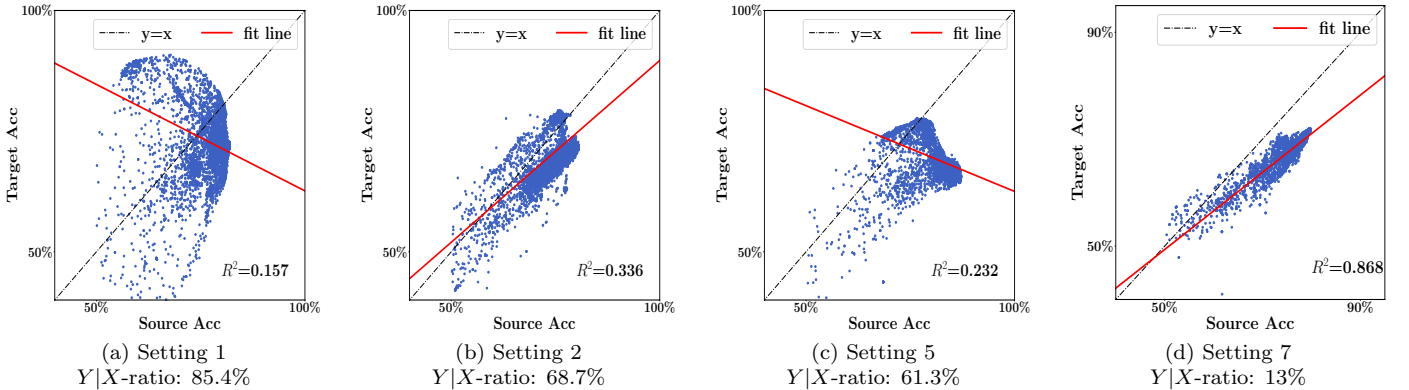


Figure 2: Target (y -axis) vs. source (x -axis) accuracies for different kinds of methods and datasets. Each sub-figure corresponds to one selected source-target pair in Table 2. In each sub-figure, the corresponding R^2 value for a linear fit (red line) is reported on the bottom right. Each blue point represents one hyperparameter configuration of one method.

As shown in Figure 2(a)-(c), we tend to see poor linear fits (low R^2) when the bulk of the performance

degradation is attributed to $Y|X$ -shifts. Notably, the R^2 of the linear trend on image datasets [28, 27] is above 0.8, highlighting the contrast. The weak correlation between source and target performance across various methods highlights the challenges of improving performance solely based on source data under $Y|X$ -shifts. The key distinction lies in the type of the distribution shifts. Under X -shifts (Figure 2(d)), model rankings remain stable and accuracy-on-the-line holds, justifying the standard ML practice of benchmarking out-of-distribution performance with the expectation of a single best model (e.g., images classified by humans). However, $Y|X$ -shifts are common in our tabular data because of changes in unobserved covariates across environments.

This introduces considerable performance variation, leading to different relationships between in-distribution and out-of-distribution performance across datasets, and ultimately causes the breakdown of accuracy-on-the-line. To further quantify this instability, we aggregate the source–target pairs with more than 5% degradation by their $Y|X$ -shift ratio. The variability of the linear fit increases markedly as the $Y|X$ -shift becomes larger: pairs with small $Y|X$ -shift ratios have relatively stable and high R^2 values, while pairs with larger $Y|X$ -shift ratios exhibit substantially higher dispersion and lower average fit. This supports the diagnostic conclusion that, under strong $Y|X$ -shifts, source accuracy is no longer a reliable predictor of target accuracy.

Evaluating robust algorithms I: fluctuations and limited performance improvements. We investigate whether DRO algorithms perform well on the semi-synthetic shifts we construct. In Figure 3, for each method within each setting (a single source domain paired with *multiple* target domains), we present the average target accuracy along with the standard deviation calculated across the multiple target domains. The results yield the following observations: (i) The overall performance of DRO methods in each setting does not demonstrate significant improvement compared to other categories, particularly when compared with Basic EO and tree-ensemble methods; In Settings 4 and 6, Kernel-DRO methods achieve significant improvements over the EO counterparts (i.e., Kernel-SVM), averaged across target domains. This occurs because the regularization in DRO methods helps control the high variance that can arise in the projection of kernel methods. Nevertheless, switching to tree-based ensemble models or linear models yields even greater performance gains, suggesting that model class selection has a larger impact than the choice of learning methods. (ii) For settings with multiple target domains, the relatively large standard deviation indicates considerable variability in generalization performance across different target domains, with DRO methods showing limited enhancements in this regard.

Evaluating robust algorithms II: model class dominates algorithmic choice. To understand which aspects of algorithmic design impacts performance, we conduct extensive empirical analysis where we carefully

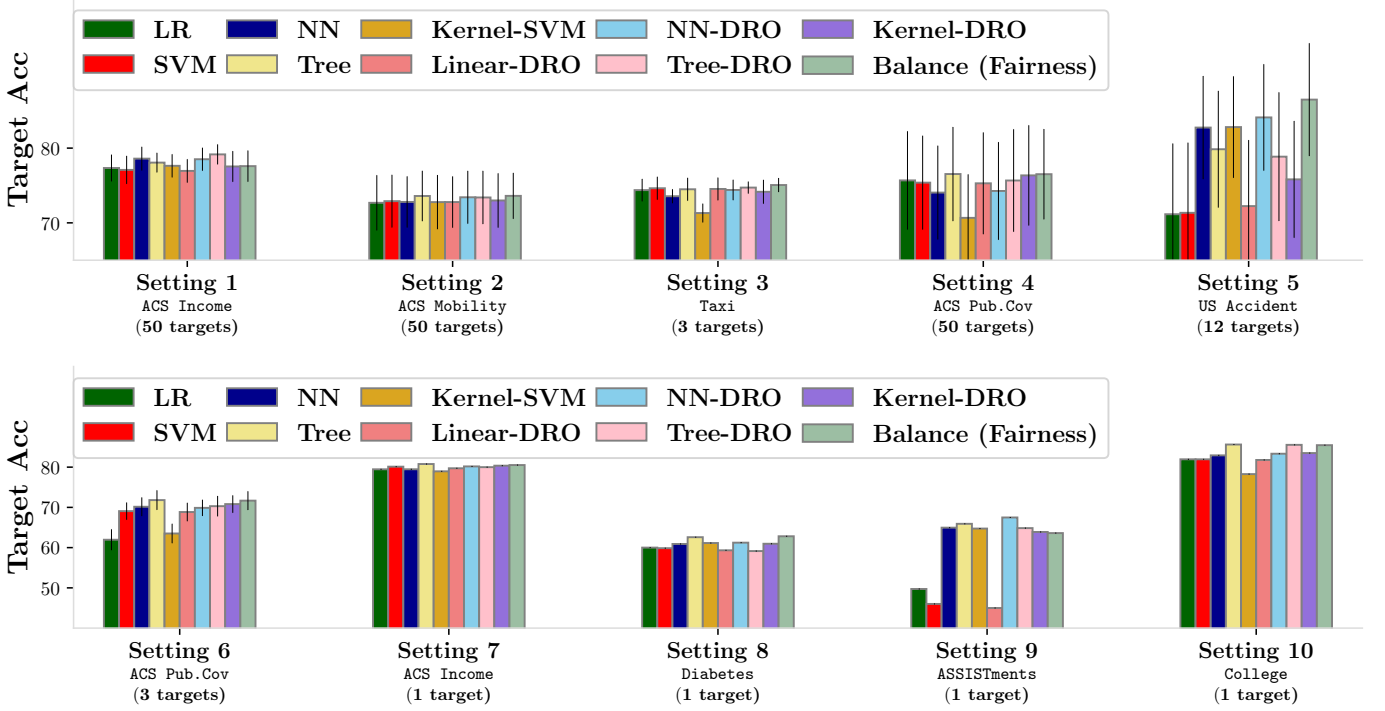


Figure 3: Overall performances of all methods on the target data in our 10 settings in Table 2. Here we calculate the average accuracy (as well as the standard deviation) among multiple target domains in each setting. Note that for Settings 3, 5 and 9, we do not include fairness-enhancing methods since there are no sensitive attributes. Within each method class, we report only the method with median average accuracy.

control for confounding factors such as the type of distribution shift. Our main finding is that implementation details often overlooked by methodology researchers can have outsize impact. For example, the choice of the model class (e.g., neural network vs. LightGBM) has a significantly greater impact on the performance of DRO methods compared to the choice of the ambiguity set, its radius, and even the use of DRO itself.

We use a simple linear regression analysis to examine the relative impact of various algorithm design elements for all source-target pairs in the benchmark. We regress the model accuracy in target domains on variables such as model class types, DRO distance types, and other explanatory variables. We include 24 methods in all settings, including basic EO methods (SVM, Kernel-SVM, NN-2, NN-3, NN-4); Linear-DRO methods (Wasserstein, CVaR, χ^2 , TV, KL, OT-Distance); Tree-based ensemble methods (LGBM); Kernel-DRO methods (Wasserstein, CVaR, χ^2 , KL); Tree-DRO methods (KL, CVaR); NN-DRO methods (CVaR, χ^2) varying the layer number as 2, 3, 4. We present aggregate results across all settings, as well as target domains within the setting. We estimate the following equation via ordinary least squares (OLS):

$$\text{Accuracy}_{i,j,s,t} = \alpha + \beta_1^\top X_{i,j,s} + \beta_2^\top D_{i,j,s,t} + \beta_3 Z_{i,j} + \beta_4^\top V_{i,j} + \mu_i + \tau_j + \eta_{i,j,s,t}, \quad (8)$$

where i, j, s, t represent the setting, target domain, model class, and specific configurations respectively. The

explanatory variables in the linear regression model (8) are separated into four groups. The variables in the first group, $X_{i,j,s}$, include dummy variables of model types indicating whether the model class of the method is NN or LGBM or Kernel (with the default defined as the linear class). The variables in the second group, $D_{i,j,s,t}$, include dummy variables for the DRO hyperparameter, indicating whether the DRO methods use a particular distance metric and one variable representing the scaled robustness radius. The default is the method without any robustness intervention, i.e., the EO method. The variable in the third group, $Z_{i,j}$, denotes the $Y|X$ -shift percentage for the target domain j of the setting i . This allows us to control for the type of distribution shift. The variables in the fourth group $V_{i,j}$ are two dummy variables for validation types, indicating whether we validate the best configuration using the best average-case accuracy or the worst-case accuracy (with the default is the method configuration selected by the best in-distribution accuracy). We use μ_i and τ_j as control variables representing the setting and domain fixed effects that are incorporated either in all settings or one particular setting.

We regress on the accuracy of the best configuration found per method and target domain and report regression results in Table 4. In each setting, the source domain is the corresponding selected source domain in the “Selected Source-Target” column and the target domains are all the possible target domains there in Table 2. We also include regression results on Settings 5 and 1, to see whether the findings are consistent across different settings.

Table 4: Regression results on algorithmic design components on the method performance

Dependent variable: Target accuracy; best configuration											
Setting	Model Class					Ambiguity Set					
	LGBM	NN2	NN3	NN4	Kernel	Wasserstein	χ^2	KL	TV	OT	Radius
All	.0081*** (.0016)	.0077*** (.0017)	.0028* (.0017)	-.0011 (.0017)	.0000 (.0013)	.0020 (.0018)	.0010 (.0012)	.0007 (.0015)	-.0049** (.0024)	-.0022 (.0024)	.0025** (.0010)
Setting 5	.0500*** (.0051)	.0730*** (.0054)	.0774*** (.0053)	.0790*** (.0052)	.0397*** (.0042)	-.0077 (.0057)	.0018 (.0037)	-.0089* (.0048)	-.0019 (.0079)	-.0062 (.0076)	-.0091 (.0057)
Setting 1	.0156*** (.0006)	.0115*** (.0007)	.0108*** (.0007)	.0124*** (.0007)	.0012** (.0006)	.0025*** (.0008)	.0010** (.0005)	.0017*** (.0006)	-.0006 (.0009)	-.0021** (.0009)	.0033*** (.0004)
Other Variables											
Setting	$Y X$ -ratio	Worst-case	Average-case	Setting FE	Domain FE	N	Adj. R^2				
All	-.0044*** (.0002)	-.0056*** (.0010)	.0054*** (.0010)	Yes	No	12527	.191				
Setting 5	-.0051*** (.0002)	-.0067** (.0033)	.0202*** (.0033)	No	Yes	935	.777				
Setting 1	.0679*** (.0013)	.0050*** (.0004)	.0078*** (.0004)	No	Yes	3671	.789				

Notes. ***, ** and * show statistical significance at the 1%, 5%, and 10% levels using two-tailed tests, respectively. Standard errors are reported in parentheses below each coefficient.

We observe the choice of the model class has by far the greatest impact on robustness. The coefficients of the model class (including LGBM and NN) are statistically significant in all settings. Furthermore, the model class effects have the largest positive coefficients when compared to the coefficients of the ambiguity set and

validation type. This implies that changing the method from the linear class to other model classes, such as tree-based ensemble methods, results in the largest improvements in the method performance, compared to building distributional robustness counterparts or changing the validation types. This finding generalizes across different settings (i.e. Settings 5 and 1 here). In each setting, the coefficients of the model class are again largest with statistical significance; coefficient for LGBM is particularly large. The hyperparameter selection mechanism also has relatively large effects on robustness against real-world distribution shifts. The positive coefficient of the average-case accuracy is usually larger than the coefficients of ambiguity set across different settings. This implies that selecting the best configuration using the average-case accuracy improves robustness more than using DRO.

On the other hand, the effects of ambiguity sets are inconsistent and relatively small across different settings. This can be seen from the small and relatively unstable coefficients on the radius of the ambiguity set (Table 4). For example, while the impact of the radius of the ambiguity set is significantly positive in Setting 1, it is not as impactful compared to the model class. And none of the distance types used in the ambiguity set is significantly positive across all the settings. This further validates that DRO does not have as much effect as the model class, making it less useful for ensuring the method generalizes well to other domains.

Through this benchmark, we advocate for bridging the gap between the OR community’s focus on modeling distribution shifts and the ML community’s emphasis on empirical validation. OR research emphasizes rich theoretical insights based on mathematically tractable model classes. In contrast, ML practice prioritizes performance on benchmarks, readily adopting complex model architectures like tree-based ensembles that have demonstrated superior performance without analytical treatment. If DRO methods are built and evaluated only with linear models, their comparatively poor performance versus more expressive tree-based ensemble methods can be attributed to the choice of model class. In other words, restricting attention to “theoretically convenient” linear models masks the real potential of DRO with richer model families.

In the full version [37], we also present case studies illustrating how a nuanced data-driven understanding of shifts from small amounts of samples from the target domain can enhance both algorithmic and data-centric interventions and improve existing EO solutions significantly.

4 Optimization Design

While diagnostics guide us in selection, it is arguably of even more primitive importance to design data-driven methods that perform well across diverse conditions, such as complex cost structures, limited or noisy data,

and changing environments. From Section 3, the performance of optimization methods varies significantly across problem instances due to heterogeneity in data and cost functions and existing robust methods do not outperform the naive one unless specific problem patterns are incorporated. In the following, I address this need by developing efficient methods that exploit problem-specific structure and data patterns to achieve significant statistical and computational gains with theoretical understanding in Section 4.1 and practical collaborations in Section 4.2.

4.1 Theory: When can we observe significant performance improvements?

We consider the problem without covariates. Then in Definition 1, the parametric EO solution reduces to $\hat{z}_{EO} \in \operatorname{argmin}_z \sum_{i=1}^n \ell(z; y_i)$ and the DRO solution reduces to $\hat{z}_{DRO} \in \operatorname{argmin}_z \max_{d(\mathbb{P}, \hat{\mathbb{P}}) \leq \epsilon} \mathbb{E}_{\mathbb{P}}[\ell(z; y)]$ for a distance d . Recall $c(z) = \mathbb{E}_{\mathbb{P}}[\ell(z; y)]$. We measure statistical improvement of a data-driven solution $\hat{z}$ by the impact on the excess risk:

$$R(z) = c(z) - c(z^*), \text{ where } z^* \in \operatorname{argmin}_{z \in \mathcal{Z}} \mathbb{E}_{\mathbb{P}}[\ell(z; y)]. \quad (9)$$

Here, our comparison centers on the expected excess risk, contrasting $\mathbb{E}_{\mathcal{D}}[R(\hat{z})]$ with $\mathbb{E}_{\mathcal{D}}[R(\hat{z}_{EO})]$, where $\hat{z}$ can be existing EO/DRO solutions or other data-optimization integrations. We develop better solutions that achieves significant performance improvements in finite-sample and asymptotic regimes given particular problem information.

4.1.1 Finite-sample Regime.

We consider *high-dimensional complex cost functions*, and use *upper bounds on $R(z)$* to analyze the behavior of data-driven optimization solutions for any n , especially when n is small.

Diagnostics. In the literature, the bounds are typically of the form: $R(\hat{z}) \leq B/n^\alpha$, where B, α are positive quantities depending on the method used to obtain $\hat{z}$. For existing EO/DRO solutions, either we have: (i) $\alpha = 1/2$ and B depends on the complexity of the function class $\mathcal{L} = \{\ell(z, \cdot) : z \in \mathcal{Z}\}$, such as local Rademacher complexity [4]; or (ii) B depends only on the true loss $\ell(z^*, \cdot)$ instead of the complexity of the hypothesis class [13], but now α typically degrades as $1/D_y$. In other words, in all the existing bounds for EO and DRO, $R(\hat{z})$ depends on *either the complexity of the cost function class or the distribution dimension*.

Improvements. To improve upon these existing solutions, we incorporate parametric estimate into the optimization model design. Our intuition is observed by real-world observations that cost functions in inventory or portfolio optimization problems are complex (e.g., involving high-order polynomials), while the available data are limited and approximately parametric, such as Gaussian or exponential. We develop a new DRO formulation called Parametric-DRO (P-DRO) centered at some parametric distribution $\hat{\mathbb{Q}}$ estimated from the data:

$\hat{z}_\lambda \in \operatorname{argmin}_z \max_{\mathbb{P}: d(\mathbb{P}, \hat{\mathbb{Q}}) \leq \lambda} \mathbb{E}_{\mathbb{P}}[\ell(z; y)]$, when the ambiguity set size λ accounts for both model misspecification and finite-sample estimation error with high probability, For this P-DRO solution, we show its upper bound as $B/\sqrt{n}$, with B only depending on the model misspecification and finite-sample estimation error. Therefore, the excess risk of our method is independent of the complexity of hypothesis class and eliminates curse of dimensionality, as long as we can identify a potential good parametric candidate through distributional model selection. In the full version [18], we generalize our P-DRO solution to more complicated settings such as distribution shifts or contextual optimization, where it achieves superior performance in data-limited regimes. We demonstrate significant performance improvements over real-world portfolio optimization problems.

4.1.2 Asymptotical Regime.

We focus on smooth cost functions that satisfy Assumption 1 and use *first-order improvement sense* to analyze the behavior of data-driven optimization solutions. Here, $\hat{z}$ provides improvements over $\hat{z}_{EO}$ in the first-order sense that $\mathbb{E}_{\mathcal{D}}[R(\hat{z})]/\mathbb{E}_{\mathcal{D}}[R(\hat{z}_{EO})]$ will converge to a constant smaller than 1 as the sample size n grows. This helps us understand whether any data-optimization integration $\hat{z}$ (including all RO/DRO methods) from the EO solution has significant performance asymptotically.

Diagnostics. We first conduct systematic dissection of existing data-driven optimization formulations using the view of “directionally perturbed” EO. We show that under smooth conditions as Assumptions 1-3, each RO / DRO solution $\hat{z}_\lambda$ in Definition 1 with regularized parameters λ admits the representation:

$$\hat{z}_\lambda = \Pi_{\mathcal{Z}} \left(\hat{z}_{EO} + \frac{1}{n} \sum_{i=1}^n H(\lambda) M(z^*; y_i) + r_n \right), \quad (10)$$

where $r_n = o_p(n^{-1/2})$. For example, if $d = \chi^2$ -divergence, then: $M(z; y) = [\nabla_z \ell(z^*; y); \nabla_z \ell(z^*; y)(\ell(z^*; y) - \mathbb{E}_{\mathbb{P}}[\ell(z^*; y)])]^\top \in \mathbb{R}^{2D_z}$, and $H(\lambda) = [H_1(\lambda); H_2(\lambda)] \in \mathbb{R}^{D_z \times 2D_z}$ for some $\|H_1(\lambda)\| \propto \lambda, \|H_2(\lambda)\| \propto \sqrt{\lambda}$ with $H_1(\lambda), H_2(\lambda) \in \mathbb{R}^{D_z \times D_z}$. That is, regardless of their specific optimization formulations, each method in Definition 1 can be unified as a directional perturbation of the EO solution plus a negligible remainder.

Following that, we show that if the following two conditions hold:

- (i) the side information is correct: $\mathbb{E}_{\mathbb{P}}[M(z^*; y)] = \mathbf{0}$;
- (ii) the non-orthogonality condition holds: $\operatorname{Cov}_{\mathbb{P}}[\operatorname{IF}(y), M(z^*; y)] \neq \mathbf{0}$,

then there exists some $\lambda = \Theta(1)$, with $\|H(\lambda)\| = \Theta(1)$ such that $\hat{z}_\lambda$ achieves a first-order improvement over $\hat{z}_{EO}$. In contrast, any $\lambda = o(1)$ can achieve at most a second-order improvement over $\hat{z}_{EO}$; If either condition (i) or (ii) fails, then $\hat{z}_\lambda$ has at most a second-order improvement over $\hat{z}_{EO}$ with the maximum improvement achieved at $\|H(\lambda)\| = \Theta(1/n)$.

This diagnostics provides a complete characterization of when an EO+ method can achieve a first-order improvement over EO, namely that this is possible only when the induced side information is both correct and non-orthogonal to the EO fluctuation, and the perturbation magnitude measured by $H(\lambda)$ needs to be significant. This characterization implies that, in fact, existing EO+ methods are unlikely to achieve first-order improvements over EO in typical smooth functions.

Improvements. To improve upon these existing solutions and achieve a first-order gain compared with the EO solution, since existing EO+ solutions can be unified in (10), we only need to optimize over the following directionally perturbed EO solution: $\hat{z}_{H,M} = \Pi_Z \left(\hat{z}_{EO} + \frac{H}{n} \sum_{i=1}^n M(\hat{z}_{EO}; y_i) \right)$.

Suppose we are given with correct side information $M(z; y)$ that satisfies Conditions (i)(ii). We can design a principled methodology to find an data-driven adjustment matrix $\hat{H}$. The best $\hat{H}$ can be computed via either system knowledge or bootstrap resampling based on excess risk estimation. Moreover, in the full version [24], we demonstrate how this gain connects to the concept of control variate, a variance reduction technique in the Monte Carlo simulation literature, which helps explain why geometrically effective side information is necessary to elicit first-order improvements.

In the following, we present a real-world industry collaboration that demonstrates how diagnostics in data-driven optimization solutions integrated can improve the real-world operations.

4.2 Industry Collaboration: Emergency Medication Stocking in Mount Sinai

We describe an ongoing collaboration with Mount Sinai on emergency medication stocking. The operational challenge is diagnostic in nature: hospitals must distinguish whether excess cost and stockouts are driven by poor demand prediction, a rigid decision rule, or a mismatch between the evaluation criterion and the operational objective. As in the inventory setting of Example 1, the hospital must keep enough medication on hand to avoid shortages while limiting holding costs and refill frequency. The existing policy is a standard (s, S) rule calibrated to a fixed number of days of demand for each medication. This simple rule is interpretable and easy to deploy, but it does not adapt well to the high variability and changing demand patterns of emergency medications, leading to avoidable holding costs and occasional stockouts.

Our diagnostic procedure evaluates adaptive stocking rules that update (s, S) using recent demand history. We instantiate the decision procedures in Definition 1(b) and use the OIC criterion from Section 2.2 to choose among linear and neural-network decision classes $\hat{z}(\cdot)$ on each day. In this way, the policy selection is tied directly to estimated out-of-sample operational cost rather than to prediction accuracy alone. Figure 4 reports a historical backtest for one medication at a major Mount Sinai hospital. The green line shows the OIC-selected data-driven policy, and the blue line shows the historical policy. With a comparable number of

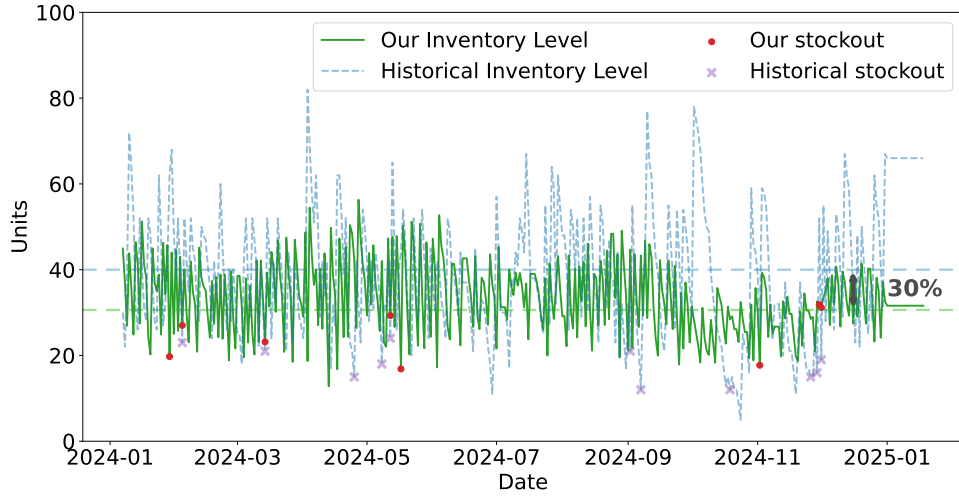


Figure 4: Holding cost comparison between historical and our policy

refill events, the diagnostic policy reduces holding cost by roughly 30% and also reduces stockouts by roughly 30%. This case study illustrates how the evaluation tools developed in this thesis can be used not only to compare algorithms, but also to guide operational decisions in a high-impact healthcare environment.

References

- [1] H. Akaike. A new look at the statistical model identification. *IEEE transactions on automatic control*, 19(6):716–723, 1974.
- [2] M. Austern and W. Zhou. Asymptotics of cross-validation. *Annales de l’Institut Henri Poincaré (B) Probabilités et statistiques*, 61(4): 2804–2865, 2025.
- [3] G.-Y. Ban, N. El Karoui, and A. E. Lim. Machine learning and portfolio optimization. *Management Science*, 64(3):1136–1154, 2018.
- [4] P. L. Bartlett, O. Bousquet, and S. Mendelson. Local Rademacher complexities. *Annals of Statistics*, 33(4):1497–1537, 2005.
- [5] S. Bates, T. Hastie, and R. Tibshirani. Cross-validation: what does it estimate and how well does it do it? *Journal of the American Statistical Association*, pages 1–12, 2023.
- [6] P. Bayle, A. Bayle, L. Janson, and L. Mackey. Cross-validation confidence intervals for test error. In *Advances in Neural Information Processing Systems*, pages 16339–16350, 2020.
- [7] A. Beirami, M. Razaviyayn, S. Shahrampour, and V. Tarokh. On Optimal Generalizability in Parametric Learning. In *Advances in Neural Information Processing Systems*, volume 30, pages 3458–3468, 2017.
- [8] T. Cai, H. Namkoong, and S. Yadowsky. Diagnosing model performance under distribution shift. *Operations Research*, 74(2):898–916, 2026.
- [9] R. D. Cook and S. Weisberg. *Residuals and influence in regression*. New York: Chapman and Hall, 1982.
- [10] P. Donti, B. Amos, and J. Z. Kolter. Task-based end-to-end model learning in stochastic optimization. In *Advances in Neural Information Processing Systems*, volume 30, pages 5490–5500, 2017.
- [11] A. N. Elmachtoub and P. Grigas. Smart “predict, then optimize”. *Management Science*, 68(1):9–26, 2022.
- [12] A. N. Elmachtoub, H. Lam, H. Zhang, and Y. Zhao. Estimate-Then-Optimize Versus Integrated-Estimation-Optimization: A Stochastic Dominance Perspective. *arXiv preprint arXiv:2304.06833*, 2023.
- [13] P. M. Esfahani and D. Kuhn. Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations. *Mathematical Programming*, 171(1):115–166, 2018.
- [14] L. Györfi, M. Kohler, A. Krzyzak, and H. Walk. *A distribution-free theory of nonparametric regression*. Springer Science & Business Media, 2006.
- [15] T. Hastie, R. Tibshirani, J. H. Friedman, and J. H. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, volume 2. Springer, 2009.

- [16] W. Hu, G. Niu, I. Sato, and M. Sugiyama. Does distributionally robust supervised learning give robust classifiers? In *International Conference on Machine Learning*, pages 2029–2037. PMLR, 2018.
- [17] Y. Hu, N. Kallus, and X. Mao. Fast rates for contextual linear optimization. *Management Science*, 68(6):4236–4245, 2022.
- [18] G. Iyengar, H. Lam, and T. Wang. Hedging complexity in generalization via a parametric distributionally robust optimization framework. *arXiv preprint arXiv:2212.01518*, 2022.
- [19] G. Iyengar, H. Lam, and T. Wang. Optimizer’s information criterion: Dissecting and correcting bias in data-driven optimization. *arXiv preprint arXiv:2306.10081*, 2023.
- [20] G. Iyengar, H. Lam, and T. Wang. Is cross-validation the gold standard to estimate out-of-sample model performance? In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, pages 94736–94775, 2024.
- [21] D. Kuhn, P. M. Esfahani, V. A. Nguyen, and S. Shafieezadeh-Abadeh. Wasserstein distributionally robust optimization: Theory and applications in machine learning. In *Operations Research & Management Science in the Age of Analytics*, pages 130–166. INFORMS, 2019.
- [22] D. Kuhn, S. Shafiee, and W. Wiesemann. Distributionally robust optimization. *Acta Numerica*, 34:579–804, 2025.
- [23] H. Lam. On the impossibility of statistically improving empirical optimization: A second-order stochastic dominance perspective. *Management Science*, 2025.
- [24] H. Lam and T. Wang. Achieving first-order statistical improvements in data-driven optimization. In *OPT 2025: Optimization for Machine Learning*, 2025.
- [25] J. Liu, T. Wang, H. Lam, H. Namkoong, and J. Blanchet. DRO: A python library for distributionally robust optimization in machine learning. *arXiv preprint arXiv:2505.23565*, 2025.
- [26] J. Miller, K. Krauth, B. Recht, and L. Schmidt. The effect of natural distribution shift on question answering models. In *International Conference on Machine Learning*, pages 6905–6916. PMLR, 2020.
- [27] J. P. Miller, R. Taori, A. Raghuathan, S. Sagawa, P. W. Koh, V. Shankar, P. Liang, Y. Carmon, and L. Schmidt. Accuracy on the Line: On the Strong Correlation Between Out-of-Distribution and In-Distribution Generalization. In *International Conference on Machine Learning*, pages 7721–7735. PMLR, 2021.
- [28] B. Recht, R. Roelofs, L. Schmidt, and V. Shankar. Do imagenet classifiers generalize to imagenet? In *International Conference on Machine Learning*, pages 5389–5400. PMLR, 2019.
- [29] U. Sadana, A. Chenreddy, E. Delage, A. Forel, E. Frejinger, and T. Vidal. A survey of contextual optimization methods for decision-making under uncertainty. *European Journal of Operational Research*, 320(2):271–289, 2025.
- [30] A. Shapiro, D. Dentcheva, and A. Ruszczyński. *Lectures on stochastic programming: modeling and theory*. SIAM, 2021.
- [31] L. V. Snyder and Z.-J. M. Shen. *Fundamentals of supply chain theory*. John Wiley & Sons, 2019.
- [32] K. T. Talluri and G. J. Van Ryzin. *The theory and practice of revenue management*, volume 68. Springer Science & Business Media, 2006.
- [33] B. P. Van Parys, P. M. Esfahani, and D. Kuhn. From data to decisions: Distributionally robust optimization is optimal. *Management Science*, 67(6):3387–3402, 2021.
- [34] S. Wager. Cross-validation, risk estimation, and model selection: Comment on a paper by rosset and tibshirani. *Journal of the American Statistical Association*, 115(529):157–160, 2020.
- [35] S. Wager and S. Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018.
- [36] M. J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*, volume 48. Cambridge University Press, 2019.
- [37] T. Wang, J. Liu, P. Cui, and H. Namkoong. Rethinking distribution shifts: Empirical analysis and inductive modeling for tabular data. *arXiv preprint arXiv:2307.05284*, 2023.