Management Science

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

Fast Rates for Contextual Linear Optimization

Yichun Hu, Nathan Kallus, Xiaojie Mao

To cite this article:
Yichun Hu, Nathan Kallus, Xiaojie Mao (2022) Fast Rates for Contextual Linear Optimization. Management Science

Published online in Articles in Advance 29 Mar 2022

. https://doi.org/10.1287/mnsc.2022.4383

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

MANAGEMENT SCIENCE
Articles in Advance, pp. 1–10
ISSN 0025-1909 (print), ISSN 1526-5501 (online)

Fast Rates for Contextual Linear Optimization

Yichun Hu,[a] Nathan Kallus,[a,*] Xiaojie Mao[b]

[a] School of Operations Research and Information Engineering and Cornell Tech, Cornell University, New York, New York 10044; [b] School of Economics and Management, Tsinghua University, Beijing 100084, China
*Corresponding author

Contact: yh767@cornell.edu, https://orcid.org/0000-0002-5826-9665 (YH); kallus@cornell.edu, https://orcid.org/0000-0003-1672-0507 (NK); maoxj@sem.tsinghua.edu.cn, https://orcid.org/0000-0003-2985-1741 (XM)

Received: March 17, 2021
Revised: August 31, 2021
Accepted: November 16, 2021
Published Online in Articles in Advance: March 29, 2022

https://doi.org/10.1287/mnsc.2022.4383

Abstract. Incorporating side observations in decision making can reduce uncertainty and boost performance, but it also requires that we tackle a potentially complex predictive relationship. Although one may use off-the-shelf machine learning methods to separately learn a predictive model and plug it in, a variety of recent methods instead integrate estimation and optimization by fitting the model to directly optimize downstream decision performance. Surprisingly, in the case of contextual linear optimization, we show that the naïve plug-in approach actually achieves regret convergence rates that are significantly faster than methods that directly optimize downstream decision performance. We show this by leveraging the fact that specific problem instances do not have arbitrarily bad near-dual-degeneracy. Although there are other pros and cons to consider as we discuss and illustrate numerically, our results highlight a nuanced landscape for the enterprise to integrate estimation and optimization. Our results are overall positive for practice: predictive models are easy and fast to train using existing tools; simple to interpret; and, as we show, lead to decisions that perform very well.

History: Accepted by Hamid Nazerzadeh, data science.
Funding: This material is based on work supported by the National Science Foundation [Grant 1846210].
Supplemental Material: Data and the e-companion are available at https://doi.org/10.1287/mnsc.2022.4383.

Keywords: contextual stochastic optimization • personalized decision making • end-to-end optimization • estimate and then optimize

1. Introduction

A central tenet of machine learning is the use of rich feature data to reduce uncertainty in an unknown variable of interest, whether it is the content of an image, medical outcomes, or future stock price. Recent work in data-driven optimization has highlighted the potential for rich features to similarly reduce uncertainty in decision-making problems with uncertain objectives and thus improve resulting decisions' performance (Donti et al. 2017, El Balghiti et al. 2019, Estes and Richard 2019, Ho and Hanasusanto 2019, Vahn and Rudin 2019, Bertsimas and Kallus 2020, Diao and Sen 2020, Ho-Nguyen and Kilinç-Karzan 2020, Kallus and Mao 2020, Loke et al. 2020, Chen et al. 2021, Elmachtoub and Grigas 2021, Notz and Pibernik 2021). For decision-making problems modeled by linear optimization with uncertain coefficients, this is captured by the contextual linear optimization (CLO) problem, defined as follows:

$$\pi^*(x) \in \mathcal{Z}^*(x) = \arg\min_{z\in\mathcal{Z}} f^*(x)^\top z, \quad f^*(x) = \mathbb{E}[Y \mid X = x],$$
$$\mathcal{Z} = \{z \in \mathbb{R}^d : Az \le b\}. \tag{1}$$

Here, $X \in \mathbb{R}^p$ represents the contextual features, $z \in \mathcal{Z} \subseteq \mathbb{R}^d$ linearly constrained decisions, and $Y \in \mathbb{R}^d$ the random coefficients. Examples of CLO are vehicle routing with uncertain travel times, portfolio optimization with uncertain security returns, and supply chain management with uncertain shipment costs. In each case, X represents anything that we can observe before making a decision z that can help reduce uncertainty in the random coefficients Y, such as recent traffic or market trends. The decision policy $\pi^*(x)$ optimizes the conditional expected costs, given the observation $X = x$. (We reserve X, Y for random variables and x, y for their values.) We assume throughout that $\mathcal{Z}$ is a polytope ($\sup_{z\in\mathcal{Z}}\|z\| \le B$) and Y is bounded (without loss of generality, $Y \in \mathcal{Y} = \{y : \|y\| \le 1\}$), and we let $\mathcal{Z}^\angle$ denote the set of extreme points of $\mathcal{Z}$.

Nominally, we only do better by taking features X into consideration when making decisions:

$$\min_{z\in\mathcal{Z}} \mathbb{E}[Y^\top z] \ge \mathbb{E}[\min_{z\in\mathcal{Z}} \mathbb{E}[Y^\top z \mid X]] = \mathbb{E}[\, f^*(X)^\top \pi^*(X)],$$

and the more Y-uncertainty explained by X the larger the gap. That is, at least if we knew the true conditional expectation function f^*. In practice, we do not; we only have data $\mathcal{D} = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$, which we assume consist of n independent draws of (X, Y). The task is then to use these data to come up with a well-performing data-driven policy $\hat{\pi}(x)$ for the

decision we will make when observing $X = x$, namely, one having low average regret:

$$\text{Regret}(\hat{\pi}) = \mathbb{E}_{\mathcal{D}}\mathbb{E}_X[\, f^*(X)^\top(\hat{\pi}(X) - \pi^*(X))], \tag{2}$$

where we marginalize *both* over new features X and over the sampling of the data $\mathcal{D}$ (i.e., over $\hat{\pi}$).

One approach is the naïve plug-in method, also known as estimate and then optimize (ETO). Since f^* is the regression of Y on X, we can estimate it using a variety of off-the-shelf methods, whether parametric regression such as ordinary least squares or generalized linear models, nonparametric regression such as k-nearest neighbors or local polynomial regression, or machine learning methods such as random forests or neural networks. Given an estimate $\hat{f}$ of f^*, we can construct the induced policy $\pi_{\hat{f}}$, where for any generic $f : \mathbb{R}^p \to \mathbb{R}^d$ we define the plug-f-in policy

$$\pi_f(x) \in \arg\min_{z\in\mathcal{Z}} f(x)^\top z. \tag{3}$$

Notice that given f, π_f need not be unique; we restrict to choices $\pi_f(x) \in \mathcal{Z}^\angle$ that break ties arbitrarily but consistently (i.e., by some ordering over $\mathcal{Z}^\angle$). Notice also that $\pi_{f^*}(x) \in \mathcal{Z}^*(x)$. Given a hypothesis class $\mathcal{F} \subseteq [\mathbb{R}^p \to \mathcal{Y}]$ for f^*, we can for example choose $\hat{f}$ by least-squares regression:

$$\hat{f}_{\mathcal{F}} \in \arg\min_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \|Y_i - f(X_i)\|^2. \tag{4}$$

We let $\hat{\pi}^{\text{ETO}}_{\mathcal{F}} = \pi_{\hat{f}_{\mathcal{F}}}$ be the ETO policy corresponding to least-squares regression over $\mathcal{F}$. ETO has appealing practical benefits. It is easily implemented using tried-and-true, off-the-shelf, potentially flexible prediction methods. More crucially, it easily adapts to decision support, which is often the reality for quantitative decision-making tools: rather than a black box prescription, it provides decision makers with a prediction that they may judge and eventually use as they see fit.

Nonetheless, a criticism of this approach is that Equation (4) uses the *wrong* loss function as it does not consider the impact of $\hat{f}$ on the downstream performance of the policy $\pi_{\hat{f}}$ and in a sense ignores the decision-making problem. The alternative empirical risk minimization *(ERM) method* directly minimizes an empirical estimate of the average costs of a policy: given a policy class $\Pi \subseteq [\mathbb{R}^p \to \mathcal{Z}]$,

$$\hat{\pi}^{\text{ERM}}_{\Pi} \in \arg\min_{\pi\in\Pi} \frac{1}{n}\sum_{i=1}^{n} Y_i^\top \pi(X_i). \tag{5}$$

In particular, a hypothesis class $\mathcal{F}$ induces the plug-in policy class $\Pi_{\mathcal{F}} = \{\pi_f : f \in \mathcal{F}\}$, and ERM over $\Pi_{\mathcal{F}}$ corresponds to optimizing the empirical risk of π_f over choices $f \in \mathcal{F}$, yielding a *different* criterion from Equation (4) for choosing $f \in \mathcal{F}$. We call this the induced ERM *(IERM) method*, which thus *integrates* the estimation and optimization aspects of the problem into one, sometimes referred to as *end-to-end estimation*. We let $\hat{\pi}^{\text{IERM}}_{\mathcal{F}} = \hat{\pi}^{\text{ERM}}_{\Pi_{\mathcal{F}}}$ denote the IERM policy induced by $\mathcal{F}$.

Although the latter IERM approach appears to much more correctly and directly deal with the decision-making problem of interest, in this paper we demonstrate a surprising fact:

> *Estimate-and-then-optimize approaches can have much faster regret-convergence rates.*

To theoretically characterize this phenomenon, we develop regret bounds for ETO and IERM when $f^* \in \mathcal{F}$. Without further assumptions beyond such well-specification (which is necessary for any hope of vanishing regret), we show that the regret convergence rate $1/\sqrt{n}$ reigns. However, appropriately limiting how degenerate an instance can be uncovers faster rates and a divergence between ETO and IERM that favors ETO. This can be attributed to ETO leveraging structure in $\mathcal{F}$ compared with IERM using only what is implied about $\Pi_{\mathcal{F}}$. Numerical examples corroborate our theory's predictions and demonstrate the conclusions extend to flexible/nonparametric specifications while highlighting the benefits of IERM for simple/interpretable models that are bound to be misspecified. We provide a detailed discussion on how this fits into the larger practical considerations of choosing between ETO and end-to-end methods such as IERM for developing decision-making and decision-support systems.

1.1. Background and Relevant Literature

1.1.1. Contextual Linear and Stochastic Optimization.

The IERM problem is generally nonconvex in $f \in \mathcal{F}$. For this reason, Elmachtoub and Grigas (2021) develop a convex surrogate loss they call SPO+, which they show is Fisher consistent under certain regularity conditions in that if $f^* \in \mathcal{F}$, then the solution to the convex surrogate problem solves the nonconvex IERM problem. El Balghiti et al. (2019) prove an $O(\log(|\mathcal{Z}^\angle|n)/\sqrt{n})$ regret bound for IERM when $\mathcal{F}$ is linear functions. Both El Balghiti et al. (2019) and Elmachtoub and Grigas (2021) advocate for the integrated IERM approach to CLO, referring to it as *smart* in comparison with the naïve ETO method.

CLO is a special case of the more general contextual stochastic optimization (CSO) problem, $\pi^*(x) \in \arg\min_{z\in\mathcal{Z}} \mathbb{E}[c(z; Y) \mid X = x]$. Bertsimas and Kallus (2020) study ETO approaches to CSO where the distribution of $Y \mid X = x$ is estimated by a reweighted empirical distribution of Y_i, for which they establish asymptotic optimality. Diao and Sen (2020) study stochastic gradient descent (SGD) approaches to solving the resulting problems. Ho and Hanasusanto (2019) propose to add variance regularization to this ETO rule to account

for errors in this estimate. Bertsimas and Kallus (2020) additionally study ERM approaches to CSO and provide generic regret bounds (see their appendix EC.1). Notz and Pibernik (2021) apply these bounds to reproducing kernel Hilbert spaces (RKHS) in a capacity planning application. Vahn and Rudin (2019) study ERM with a sparse linear model for the newsvendor problem. Kallus and Mao (2020) construct forest policies for CSO by using optimization perturbation analysis to approximate the generally intractable problem of ERM for CSO over trees; they also prove asymptotic optimality. Many other works that study CSO generally advocate for *end-to-end* solutions that *integrate* or *harmonize* estimation and optimization (Donti et al. 2017, Estes and Richard 2019, Ho-Nguyen and Kilinç-Karzan 2020, Loke et al. 2020).

1.1.2. Classification. Classification is a specific case of CLO with $Y \in \{-1,1\}$ and $\mathcal{Z} = [-1,1]$. Then $\frac{1}{2}\text{Regret}(\hat{\pi}) = \mathbb{P}(Y \neq \hat{\pi}(X)) - \mathbb{P}(Y \neq \pi^*(X))$ is the excess error rate. Vapnik and Chervonenkis (1974), Tsybakov (2004), Bartlett et al. (2005), Koltchinskii et al. (2006), and Massart and Nédélec (2006), among others, study regret and generalization bounds for ERM approaches, convexifications, and related approaches. Our work is partly inspired by Audibert and Tsybakov (2007), who compare such ERM classification approaches to methods that estimate $\mathbb{P}(Y = 1 \mid X)$ and then classify by thresholding at $1/2$ and show that these can enjoy fast regret convergence rates under a noise condition (also known as margin) that quantifies the concentration of $\mathbb{P}(Y = 1|X)$ near $1/2$. In contrast to Audibert and Tsybakov (2007), we study fast rates for the more general CLO problem as our aim is to shed light on data-driven optimization; we use complexity notions that allow direct comparison of ETO and IERM (rather than ERM) using the same hypothesis class (whereas entropy conditions for ERM and plug-in used by Audibert and Tsybakov 2007 are incomparable); and we provide lower bounds that rigorously show the gap between IERM and ETO for any given polytope (the lower bounds of Audibert and Tsybakov 2007 only apply to Hölder-smooth functions and classification and they show the optimality of plug-in methods rather than the *sub*optimality of ERM). Similar noise or margin conditions have also been used in contextual bandits (Rigollet and Zeevi 2010, Goldenshluger and Zeevi 2013, Perchet and Rigollet 2013, Bastani and Bayati 2020, Hu et al. 2020). Our condition is similar to these but adapted to CLO.

1.2. A Simple Example

We start with a simple illustrative example. Consider univariate decisions, $\mathcal{Z} = [-1,1]$, univariate features, $X \sim \text{Unif}[-1,1]$, and a simple linear relationship, $f^*(X) = X$, with noise $Y - f^*(X) = (U-1)\sigma$, where $U \sim \text{Exp}(1)$ and $\sigma \geq 0$. Let us default to $z = -1$ under ties. Then, given a hypothesis set $\mathcal{F} = \{f_\theta(x) = x - \theta : \theta \in [-1,1]\}$, we have $\pi_{f_\theta}(x) = 2\mathbb{I}[x \leq \theta] - 1$. Let us default to smaller θ under ties. We can compute $\mathbb{E}_X[f^*(X)^\top(\pi_{f_\theta}(X) - \pi^*(X))] = \frac{1}{2}\theta^2$. We can also see that $\hat{\pi}^{\text{ETO}}_{\mathcal{F}}(x) = 2\mathbb{I}[x \leq \hat{\theta}_{\text{OLS}}] - 1$, where $\hat{\theta}_{\text{OLS}} = \frac{1}{n}\sum_{i=1}^n (X_i - Y_i)$, which has second moment $\frac{\sigma^2}{n}$. Hence, $\text{Regret}(\hat{\pi}^{\text{ETO}}_{\mathcal{F}}) = \frac{\sigma^2}{2n}$.

Unfortunately, $\hat{\pi}^{\text{IERM}}_{\mathcal{F}}$ and its regret is harder to compute. We can instead study it empirically. Figure 1(a) displays results for 500 replications for each of $n = 32, 38, 45, \ldots, 2{,}048$ with $\sigma = 1$. The plot is shown on a log-log scale with linear trend fits. The slope for ETO is –1.05 and for IERM is –0.665. (We also plot "IERM-mid" where we choose the midpoint of the argmin set for θ rather than left endpoint to show that this changes little; and we plot SPO+, whose regret does not converge to zero as it is only a surrogate for IERM. In the special case of $\sigma = 0$, we can actually analytically derive $\text{Regret}(\hat{\pi}^{\text{IERM}}_{\mathcal{F}}) = \Theta(1/n^2)$, infinitely slower than $\text{Regret}(\hat{\pi}^{\text{ETO}}_{\mathcal{F}}) = 0$; see Online Appendix E.1.)

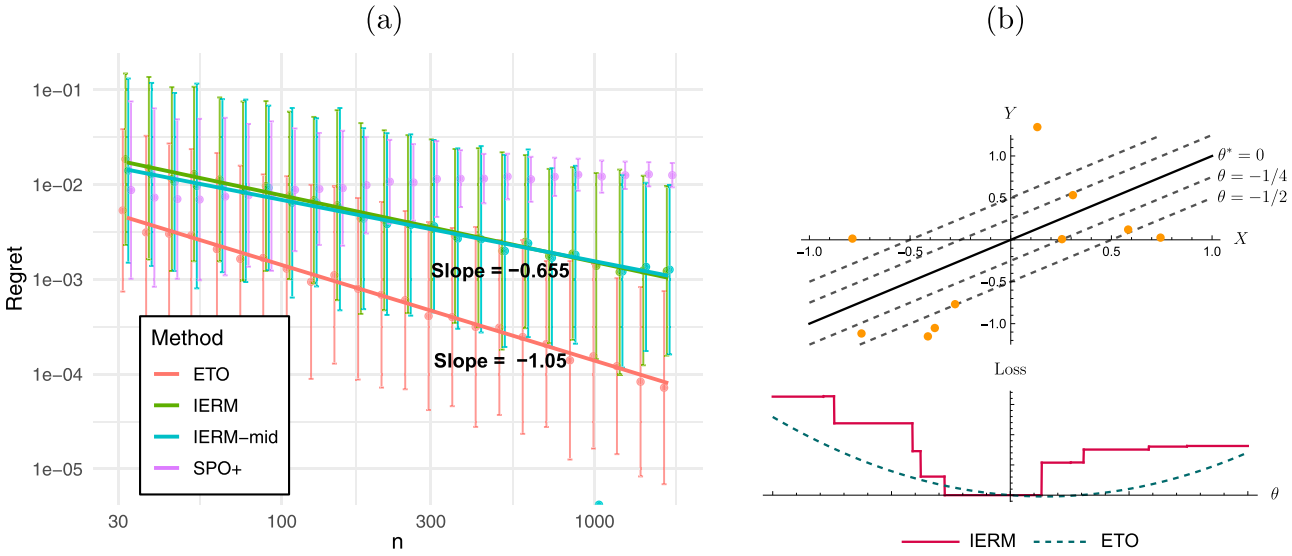
The first thing to note is that *both* slopes are *steeper* than the usual $1/\sqrt{n}$ convergence rate (i.e., –0.5 slope), such as El Balghiti et al. (2019) gives for IERM. This suggests the usual theory does not correctly predict the behavior in practice. The second thing to note is that the slope for ETO is steeper than for IERM, with an apparent rate of convergence of n^{-1} as compared with $n^{-2/3}$. Although ETO is leveraging all the information about $\mathcal{F}$, IERM is only leveraging what is implied about $\Pi_{\mathcal{F}}$, so it cannot, for example, distinguish between θ values lying between two consecutive observations of X (see Figure 1(b)). Our fast (noise-dependent) rates will *exactly* predict this divergent regret behavior. Note this very simple example is only aimed to illustrate this convergence phenomenon and need not be representative of real problems, which we explore further in Sections 4 and 5.

2. Slow (Noise-Independent) Rates

Our aim is to obtain regret bounds in terms of *primitive* quantities that are *common* to *both* the ETO and IERM approaches. To compare them, we will consider implications of our general results for the case of a correctly specified hypothesis class $\mathcal{F}$ with bounded *complexity*. One standard notion of the complexity for scalar-valued functions $\mathcal{F} \subseteq [\mathbb{R}^p \to \mathbb{R}]$ is the VC (Vapnik-Chervonenkis)-subgraph dimension (Dudley 1987). No commonly accepted notions appear to exist for vector-valued classes of functions. Here we define and use an apparently new, natural extension of VC-subgraph dimension.

Definition 1. The VC-linear-subgraph dimension of a class of functions $\mathcal{F} \subseteq [\mathbb{R}^p \to \mathbb{R}^d]$ is the VC dimension of the sets $\mathcal{F}^\circ = \{\{(x,\beta,t) : \beta^\top f(x) \leq t\} : f \in \mathcal{F}\}$ in $\mathbb{R}^{p+d+1}$, that is, the largest integer ν for which there exist

Figure 1. (Color online) A Simple Example Illustrating the Fast Rates for IERM and Even Faster Rates for ETO



Notes. (a) Regret convergence rates. Shown is average regret by n plus/minus one standard deviation. Solid lines are log-log linear fits. (b) A draw of $n = 10$ data points ($\sigma = 1$) and the corresponding loss surfaces for IERM and for ETO (least squares).

$x_1, \ldots, x_\nu \in \mathbb{R}^p$, $\beta_1, \ldots, \beta_\nu \in \mathbb{R}^d$, $t_1 \in \mathbb{R}, \ldots, t_\nu \in \mathbb{R}$ such that

$$\{(\mathbb{I}[\beta_1^\top f(x_1) \le t_1], \ldots, \mathbb{I}[\beta_\nu^\top f(x_\nu) \le t_\nu]) : f \in \mathcal{F}\} = \{0,1\}^\nu.$$

Our standing assumption will be that $f^* \in \mathcal{F}$ where $\mathcal{F}$ has bounded VC-linear-subgraph dimension. (In Online Appendices C.4 and D we study other functions classes, including RKHS and Hölder functions.)

Assumption 1 (Hypothesis Class). *The VC-linear-subgraph dimension of $\mathcal{F}$ is at most ν, where $f^* \in \mathcal{F}$.*

Example 1 (Vector-Valued Linear Functions). Suppose $\mathcal{F} \subseteq \{Wx : W \in \mathbb{R}^{d \times p}\}$. (Note we can always pretransform x.) Because $\beta^\top f(x) = \mathrm{vec}(W)^\top \mathrm{vec}(\beta x^\top)$, the VC-linear-subgraph dimension of $\mathcal{F}$ is at most the usual VC-subgraph dimension of $\{v \mapsto w^\top v : w \in \mathbb{R}^{dp}\}$, which is dp.

Example 2 (Trees). Suppose $\mathcal{F}$ consists of all binary trees of depth at most D, where each internal node queries "$w^\top x \le \theta$?" for a choice of $w \in \mathbb{R}^p$, $\beta_0 \in \mathbb{R}$ for each internal node, splitting left if true and right otherwise, and each leaf node assigns the output v to x that reach it, for any choice of $v \in \mathbb{R}^d$ for each leaf node. (In particular, this is a superset of restricting w to be a vector of all zeros except for a single one so that the splits are axis aligned.) Then, $\mathcal{F}^\circ$ is contained in the disjunction over leaf nodes of the classes of sets representable by a leaf, which is the conjunction over internal nodes' half-spaces on the path to the leaf and over the final query of $\beta^\top v \le t$. Because there are at most 2^D leaf nodes and at most D internal nodes on the path to each, applying (Van Der Vaart and Wellner 2009, theorem 1.1) twice, the VC dimension of $\mathcal{F}^\circ$ is at most $22(D^2 p + Dd)\, 2^D \log(8D)$.

2.1. Slow Rates for ERM and IERM

We first establish a generalization result for generic ERM for CLO and then apply it to IERM.

Definition 2. The Natarajan dimension of a class of functions $\mathcal{G} \subseteq [\mathbb{R}^p \to \mathcal{S}]$ with codomain $\mathcal{S}$ is the largest integer η for which there exist $x_1, \ldots, x_\eta \in \mathbb{R}^p$, $s_1 \ne s_1', \ldots, s_\eta \ne s_\eta' \in \mathcal{S}$ such that

$$\{(\mathbb{I}[g(x_1) = s_1], \ldots, \mathbb{I}[g(x_\eta) = s_\eta]) : g \in \mathcal{G}, g(x_1) \in \{s_1, s_1'\}, \ldots, g(x_\eta) \in \{s_\eta, s_\eta'\}\} = \{0,1\}^\eta.$$

Theorem 1. *Suppose $\Pi \subseteq [\mathbb{R}^p \to \mathcal{Z}^\angle]$ has Natarajan dimension at most η. Then, for a universal constant C, with probability at least $1 - \delta$,*

$$\sup_{\pi \in \Pi} \left| \frac{1}{n} \sum_{i=1}^n Y_i^\top \pi(X_i) - \mathbb{E}_X[f^*(X)^\top \pi(X)] \right| \le CB \sqrt{\frac{\eta \log(|\mathcal{Z}^\angle| + 1) \log(5/\delta)}{n}}. \tag{6}$$

Equation (6) immediately implies that the excess loss to the best-in-class policy (which need not be π^* in the absence well specification), that is, $\inf_{\pi \in \Pi} \mathbb{E}_X[f^*(X)^\top (\hat{\pi}_\Pi^{\mathrm{ERM}}(X) - \pi(X))]$, is bounded by twice the right-hand side of Equation (6) with probability at least $1 - \delta$. Note El Balghiti et al. (2019) prove a similar result to Theorem 1 but with an additional suboptimal dependence on $\sqrt{\log(n)}$.

To study IERM, we next relate VC-linear-subgraph dimension to Natarajan dimension.

Theorem 2. *The VC-linear-subgraph dimension of $\mathcal{F}$ bounds the Natarajan dimension of $\Pi_{\mathcal{F}}$.*

Corollary 1. *Suppose Assumption 1 holds. Then, for a universal constant C,*

$$\text{Regret}(\hat{\pi}_{\mathcal{F}}^{\text{IERM}}) \le CB\sqrt{\frac{\nu\log(|\mathcal{Z}^\angle|+1)}{n}}.$$

We can in fact show that the rate in Theorem 1 is optimal in n and η by showing any algorithm must suffer at least this rate on some example. When $\mathcal{Z}=[-1,1]$, our result reduces to that of Devroye and Lugosi (1995) for binary classification; but we tackle CLO with *any* polytope $\mathcal{Z}$.

Theorem 3. *Fix any polytope $\mathcal{Z}$. Fix any $\Pi \subseteq [\mathbb{R}^p \to \mathcal{Z}^\angle]$ with Natarajan dimension at least η. Fix any algorithm mapping $\mathcal{D} \mapsto \hat{\pi} \in \Pi$. Then there exists a distribution $\mathbb{P}$ on $(X,Y) \in \mathbb{R}^p \times \mathcal{Y}$ satisfying $\pi^* \in \Pi$ such that for any $n \ge 4\eta$, when $\mathcal{D} \sim \mathbb{P}^n$, we have*

$$\text{Regret}(\hat{\pi}) \ge \frac{\rho(\mathcal{Z})}{2e^4}\sqrt{\frac{\eta}{n}},$$

where $\rho(\mathcal{Z}) = \inf_{z\in\mathcal{Z}^\angle, z'\in\text{conv}(\mathcal{Z}^\angle\setminus\{z\})}\|z-z'\|$ (which is positive by definition).

In general, Theorem 3 also shows that the rate in Corollary 1 is optimal in n when we only assume $\pi^* \in \Pi_{\mathcal{F}}$, but not necessarily in ν, because Theorem 2 is only an upper bound. In many cases, however, it can be very tight. In Example 1, we upper bounded the VC-linear-subgraph dimension of $\mathcal{F}$ by dp, whereas corollary 29.8 of Shalev-Shwartz and Ben-David (2014) shows the Natarajan dimension of $\Pi_{\mathcal{F}}$ is at least $(d-1)(p-1)$ when $\mathcal{Z}$ is the simplex, so the gap is very small.

2.2. Slow Rates for ETO

We next establish comparable rates for ETO. The following is immediate from Cauchy-Schwartz.

Theorem 4. *Let $\hat{f}$ be given. Then,*

$$\text{Regret}(\pi_{\hat{f}}) \le 2B\mathbb{E}_{\mathcal{D}}\mathbb{E}_X\|f^*(X)-\hat{f}(X)\|.$$

To study ETO under Assumption 1, we next establish a convergence rate for $\hat{f}_{\mathcal{F}}$ to plug in above.

Theorem 5. *Suppose Assumption 1 holds and that $\mathcal{F}$ is star shaped at f^* (meaning $(1-\lambda)f+\lambda f^* \in \mathcal{F}$ for any $f\in\mathcal{F}, \lambda\in[0,1]$). Then, there exist positive universal constants $C_0, C_1, C_2 > 0$ such that, for any $\delta \le (nd+1)^{-C_0}$, with probability at least $1-C_1\delta^\nu$,*

$$\mathbb{E}_X\|\hat{f}_{\mathcal{F}}(X)-f^*(X)\| \le C_2\sqrt{\frac{\nu\log(1/\delta)}{n}}.$$

In Online Appendix C, we prove a novel finite-sample guarantee for least squares with vector-valued response over a general function class $\mathcal{F}$, which is of independent interest (relying on existing results for scalar-valued response leads to suboptimal dependence on d). Theorem 5 is its application to the VC-linear-subgraph case. The star shape assumption is purely technical but, although it holds for Example 1, it does not for Example 2. We can avoid it by replacing $\mathcal{F}$ with $\bar{\mathcal{F}} = \{(1-\lambda)f+\lambda f' : f, f' \in \mathcal{F}, \lambda\in[0,1]\}$ in Equation (4) (for Example 2, we even have $\bar{\mathcal{F}} = \mathcal{F}+\mathcal{F}$), which does not affect the result, only the universal constants. We omit this because least squares over $\bar{\mathcal{F}}$ is not so standard.

Corollary 2. *Suppose the assumptions of Theorem 5 hold. Then, for a universal constant C,*

$$\text{Regret}\left(\hat{\pi}_{\mathcal{F}}^{\text{ETO}}\right) \le CB\sqrt{\frac{\nu\log(nd+1)}{n}}.$$

We can remove the term $\log(nd+1)$ in the specific case of Example 1 (see Corollary EC.1 in Online Appendix C.4). Because $\log(|\mathcal{Z}^\angle|+1)$ is generally of order d (Barvinok 2013, Henk et al. 2018), the d-dependence above may be better than in Theorem 1 even for general VC-linear-subgraph classes.

Note Corollaries 1 and 2 uniquely enable us to compare ETO and IERM using the same primitive complexity measure. In contrast, complexity measures like bounded metric entropy or Rademacher complexity on $\mathcal{F}$ may not provide similar control on the complexity of $\Pi_{\mathcal{F}}$. The slow rates for IERM and ETO are nonetheless the same (up to polylogs), suggesting no differentiation between the two. Studying finer instance characteristics beyond specification reveals the differentiation.

3. Fast (Noise-Dependent) Rates

We next show that much faster rates actually occur in any one instance. To establish this, we characterize the *noise* in an instance as the level of near-dual-degeneracy (multiplicity of solutions).

Assumption 2 (Noise Condition). *Let $\Delta(x) = \inf_{z\in\mathcal{Z}^\angle\setminus\mathcal{Z}^*(x)} f^*(x)^\top z - \inf_{z\in\mathcal{Z}} f^*(x)^\top z$ if $\mathcal{Z}^*(x) \ne \mathcal{Z}$ and otherwise $\Delta(x)=0$. Assume for some $\alpha,\gamma \ge 0$,*

$$\mathbb{P}_X(0<\Delta(X)\le\delta) \le (\gamma\delta/B)^\alpha \quad \forall\, \delta>0. \tag{7}$$

Assumption 2 controls the mass of $\Delta(X)$ near (but not at) zero. It always holds for $\alpha=0$ (with $\gamma=1$). If $\Delta(X) \ge B/\gamma$ is bounded away from zero (akin to strict separation assumptions in Massart and Nédélec 2006 and Foster et al. 2020), then Assumption 2 holds for $\alpha\to\infty$. Generically, for any one instance, Assumption 2 holds for *some* $\alpha\in(0,\infty)$. For example, if X has a bounded density and $f^*(x)$ has a Jacobian that is uniformly nonsingular (or if $f^*(x)$ is linear), then Assumption 2 holds with $\alpha=1$. In

particular, the example in Section 1.2 has $\Delta(X) = |X| \sim \text{Unif}[0,1]$ and hence $\alpha = 1$.

3.1. Fast Rates for ERM and IERM

Under Assumption 2, we can obtain a faster rate both for generic ERM and specifically for IERM.

Theorem 6. *Suppose Assumption 2 holds, $\mathbb{P}(|\mathcal{Z}^*(X)| > 1) = 0$, $\Pi \subseteq [\mathbb{R}^p \to \mathcal{Z}^\angle]$ has Natarajan dimension at most η, and $\pi^* \in \Pi$. Then, for a constant $C(\alpha,\gamma)$ depending only on α,γ,*

$$\text{Regret}\left(\hat{\pi}_\Pi^{\text{ERM}}\right) \le C(\alpha,\gamma)B\left(\frac{\eta\log(|\mathcal{Z}^\angle|+1)\log(n+1)}{n}\right)^{\frac{1+\alpha}{2+\alpha}}.$$

Whenever $\alpha > 0$, this is faster than the noise-independent rate (Theorem 1). Assuming $\mathbb{P}(|\mathcal{Z}^*(X)| > 1) = 0$ requires that, in addition to nice near-dual-degeneracy, we almost never have exact dual degeneracy.

Corollary 3. *Suppose Assumptions 1 and 2 hold and $\mathbb{P}(|\mathcal{Z}^*(X)| > 1) = 0$. Then,*

$$\text{Regret}\left(\hat{\pi}_\mathcal{F}^{\text{IERM}}\right) \le C(\alpha,\gamma)B\left(\frac{\nu\log(|\mathcal{Z}^\angle|+1)\log(n+1)}{n}\right)^{\frac{1+\alpha}{2+\alpha}}.$$

Notice that with $\alpha = 1$, this *exactly* recovers the rate behavior observed empirically in Section 1.2. We next show that the rate in n in Theorem 6 and Corollary 3 (and in η in the former) is in fact optimal (up to polylogs) under Assumption 2 when we only rely on well specification of the policy.

Theorem 7. *Fix any $\alpha \ge 0$. Fix any polytope $\mathcal{Z}$. Fix any $\Pi \subseteq [\mathbb{R}^p \to \mathcal{Z}^\angle]$ with Natarajan dimension at least η. Fix any algorithm mapping $\mathcal{D} \mapsto \hat{\pi} \in \Pi$. Then there exists a distribution $\mathbb{P}$ on $(X,Y) \in \mathbb{R}^p \times \mathcal{Y}$ satisfying $\pi^* \in \Pi$ and Assumption 2 with the given α and $\gamma = B/\rho(\mathcal{Z})$ such that for any $n \ge 2^{2+\alpha}(\eta-1)$, when $\mathcal{D} \sim \mathbb{P}^n$, we have*

$$\text{Regret}(\hat{\pi}) \ge \frac{\rho(\mathcal{Z})}{2e^4}\left(\frac{\eta-1}{n}\right)^{\frac{1+\alpha}{2+\alpha}}.$$

3.2. Fast Rates for ETO

We next show the noise-level-specific rate for ETO is *even* faster, sometimes *much* faster. Although Theorems 6 and 7 are tight if we only leverage information about the policy class, leveraging the information on $\mathcal{F}$ itself, as ETO does, can break that barrier and lead to better performance.

Theorem 8. *Suppose Assumption 2 holds and, for universal constants C_1, C_2, and a sequence a_n, $\hat{f}$ satisfies that, for any $\delta > 0$ and almost all x, $\mathbb{P}(\|\hat{f}(x) - f^*(x)\| \ge \delta) \le C_1\exp(-C_2 a_n \delta^2)$. Then, for a constant $C(\alpha,\gamma)$ depending only on α,γ,*

$$\text{Regret}(\pi_{\hat{f}}) \le C(\alpha,\gamma)B\, a_n^{-\frac{1+\alpha}{2}}.$$

Although Theorem 4 requires $\hat{f}$ to have good average error, Theorem 8 requires $\hat{f}$ to have a point-wise tail bound on error with rate a_n. This is generally stronger but holds for a variety of estimators. For example, if $\hat{f}$ is given by, for example, a generalized linear model then we can obtain $a_n = n$ (McCullagh and Nelder 1989), which together with Theorem 8 leads to an *even* better regret rate of $n^{-\frac{1+\alpha}{2}}$.

Although such point-wise rates generally hold when $\hat{f}$ is parametric, VC-linear-subgraph dimension only characterizes average error so a comparison based on it requires we also make a recoverability assumption to study pointwise error (see also Hanneke 2011, Foster et al. 2020). In Online Appendix B, we show Assumption 3 generally holds for Examples 1 and 2 (Propositions EC.1 and EC.2).

Assumption 3 (Recovery). *There exists κ such that for all $f \in \mathcal{F}$ and almost all x,*

$$\|f(x) - f^*(x)\|^2 \le \kappa\mathbb{E}[\|f(X) - f^*(X)\|^2].$$

Corollary 4. *Suppose Assumptions 1–3 hold and $\mathcal{F}$ is star shaped at f^*. Then,*

$$\text{Regret}\left(\hat{\pi}_\mathcal{F}^{\text{ETO}}\right) \le C(\alpha,\gamma)B\kappa^{1+\alpha}\left(\frac{\nu\log(nd+1)}{n}\right)^{\frac{1+\alpha}{2}}.$$

With $\alpha = 1$, this *exactly* recovers the rate behavior observed in Section 1.2. We can also remove the $\log(nd+1)$ term in the case of Example 1 (see Corollary EC.1 in Online Appendix C.4). Compared with Theorem 3, we see the regret rate's exponent in n is faster by a factor of $1+\frac{\alpha}{2}$. This can be attributed to using all the information on $\mathcal{F}$ rather than just what is implied about $\Pi_\mathcal{F}$.

3.3. Fast Rates for Nonparametric ETO

Assumption 1 is akin to a parametric restriction, but ETO can easily be applied using any flexible nonparametric or machine learning regression. For some such methods, we can also establish theoretical results (with correct d-dependence, compared with relying on existing results for regression). If, instead of Assumption 1, we assume that f^* is β-smooth (roughly meaning it has β derivatives), then we show in Online Appendix D how to construct an estimator $\hat{f}$ satisfying the point-wise condition in Theorem 4 with $a_n = n^{\frac{2\beta}{2\beta+p}}/d$ and without a recovery assumption. This leads to a regret rate of $n^{-\frac{\beta(1+\alpha)}{2\beta+p}}$ for ETO. Although slower than the rate in Theorem 4, the restriction on f^* is nonparametric; the rate can still be arbitrarily fast as either α or β grow. In Online Appendix C.4, we also analyze estimates $\hat{f}$ based on kernel ridge regression, which we also deploy in experiments in Section 5.

4. Considerations for Choosing Separated vs. Integrated Approaches

We next provide some perspective on our results and on their implications. We frame this discussion as a comparison between IERM and ETO approaches to CLO along several aspects.

4.1. Regret Rates

Section 2 shows that the noise-level-agnostic regret rates for IERM and ETO have the same $n^{-1/2}$-rate (albeit, the ETO rate may also have better d-dependence). But this hides the fact that specific problem instances do not actually have *arbitrarily* bad near-degeneracy, that is, they satisfy Assumption 2 for *some* $\alpha > 0$. When we restrict how bad the near-degeneracy can be, we obtained fast rates in Section 3. In this regime, we showed that ETO can actually have *much* better regret rates than IERM. It is important to emphasize that, although specific instances *do* satisfy Assumption 2, this regime truly captures how these methods actually behave in practice in specific problems. Therefore, in terms of regret rates, this shows a clear preference for ETO approaches.

4.2. Specification

Our theory focused on the well-specified setting, that is, $f^* \in \mathcal{F}$. When this fails, convergence of the regret of $\hat{\pi}$ to π^* to zero is essentially hopeless for any method that focuses only on $\mathcal{F}$. ERM, nonetheless, can still provide best-in-class guarantees: regret to the best policy in Π still converges to zero. For induced policies, π_f, this means IERM gets best-in-class guarantees over $\Pi_{\mathcal{F}}$, whereas ETO may not. Given the fragility of correct specification if $\mathcal{F}$ is too simple, the ability to achieve best-in-class performance is important and may be the primary reason one might prefer (I)ERM to ETO. Nonetheless, if $\mathcal{F}$ is not well specified, it begs the question why use IERM rather than ERM directly over some policy class Π. The benefit of using $\Pi_{\mathcal{F}}$ may be that it provides an easy way to construct a reasonable policy class that respects the decision constraints, $\mathcal{Z}$.

4.3. BYOB: Bring Your Own Blackbox

Although IERM, as defined, is given by optimizing over $\mathcal{F}$ and is therefore specified by $\mathcal{F}$, ETO accommodates any regression method as a black box, not just least squares. This is perhaps most important in view of specification: many flexible regression methods, including local polynomial or gradient boosting regression, do not take the form of minimization over $\mathcal{F}$. (See Section 3.3 regarding guarantees for the former.) At the same time, there do also exist end-to-end methods that target empirical policy risk, though they are not exactly of the form of Equation (5), such as Elmachtoub et al. (2020) and Kallus and Mao (2020); the number of such tailored methods may grow as more attention is given to this area. Any benefits of these, nonetheless, may be greatest in nonlinear problems, as discussed below.

4.4. Interpretability

ETO has the benefit of an *interpretable output*: rather than just having a black box spitting out a decision with no explanation, our output has a clear interpretation as a prediction of Y. We can therefore probe this prediction and understand more broadly what other implications it has, such as what happens if we changed our constraints $\mathcal{Z}$ and other counterfactuals. This is absolutely crucial in decision-support applications, which are the most common in practice.

If we care about *model* —understanding *how* inputs lead to outputs—it may be preferable to focus on simple models like shallow trees. For these, which are likely not well specified, IERM has the benefit of at least ensuring best-in-class performance (Elmachtoub et al. 2020).

4.5. Computational Tractability

Another important consideration is tractability. For ETO, this reduces to learning $\hat{f}$, and both classic and modern prediction methods are often tractable and built to scale. On the other hand, IERM is nonconvex and may be hard to optimize. This is exactly the motivation of Elmachtoub and Grigas (2021), which develop a convex relaxation. However, it is only consistent if $\mathcal{F}$ is well specified, in which case we expect ETO has better performance.

4.6. Contextual Stochastic Optimization

While we focused on CLO, a question is what do our results suggest for CSO generally. CSO with a finite feasible set (or set of possibly optimal solutions), $\mathcal{Z} = \{z^{(1)}, \dots, z^{(K)}\}$, is immediately reducible to CLO by replacing $\mathcal{Z}$ with the K-simplex and Y with $(c(z^{(1)}; Y), \dots, c(z^{(K)}; Y))$. Then, our results still apply. Continuous CSO may require a different analysis to account for a nondiscrete notion of a noise condition. In either the continuous or finite setting, however, ETO would entail learning a high-dimensional object, being the conditional distribution of $Y|X = x$ (or, rather, the conditional expectations $\mathbb{E}[c(z; Y)|X = x]$ for *every* $z \in \mathcal{Z}$, whether infinite or finite and big). Although certainly methods for this exist, if $\mathcal{Z}$ has reasonable dimensions, a purely policy-based approach, such as ERM or IERM, might be more practical. For example, Kallus and Mao (2020) show that directly targeting the downstream optimization problem when training random forests significantly improves forest-based approaches to CSO. This is in contrast to the CLO case, where both the decision policy and

relevant prediction function have the same dimension, both being functions $\mathbb{R}^p \to \mathbb{R}^d$.

5. Experiments

We next demonstrate these considerations in an experiment, the replication code for which is available at https://github.com/CausalML/ContextualLPCode. We consider the stochastic shortest path problem shown in Figure 2(a). We aim to go from s to t on a 5 × 5 grid, and the cost of traveling on edge j is Y_j. There are $d = 40$ edges; $\mathcal{Z}$ is given by standard flow preservations constraints, with a source of +1 at s and a sink of –1 at t. We consider covariates with $p = 5$ dimensions and $f^*(x)$ being a degree-5 polynomial in x, as we detail in Online Appendix E.2.

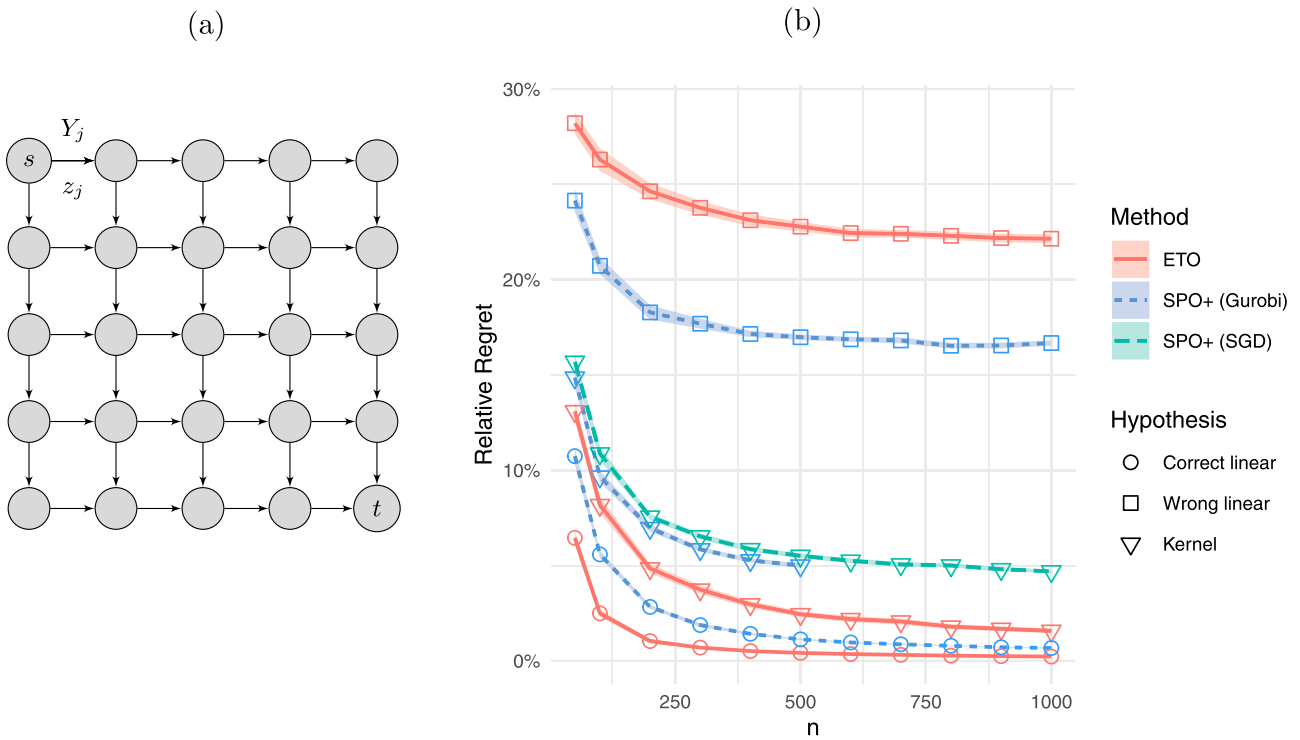
Ideally, we would like to compare ETO to IERM. However, IERM involves a difficult optimization problem that cannot feasibly be solved in practice. We therefore employ the SPO+ loss proposed by Elmachtoub and Grigas (2021), which is a convex surrogate for IERM's objective function. Like IERM, this is still an end-to-end method that integrates estimation and optimization, standing in stark contrast to ETO, which completely separates the two steps. We consider three different hypothesis classes $\mathcal{F}$ for each of ETO (using least-squares regression, $\hat{f}_{\mathcal{F}}$) and SPO+:

- Correct linear: The class $\mathcal{F}$ is as in Example 1 with $\phi(x)$ a 31-dimensional basis of monomials spanning f^*. This represents the unrealistic ideal where we have a perfectly specified parametric model.
- Wrong linear: The class $\mathcal{F}$ is as in Example 1 with $\phi(x) = x \in \mathbb{R}^5$. This represents the realistic setting where parametric models are misspecified.
- Kernel: The class $\mathcal{F}$ is the RKHS with Gaussian kernel, $\mathcal{K}(x, x') = \exp(-\rho\|x - x'\|^2)$. This represents the realistic setting of using flexible, nonparametric models.

We employ a ridge penalty in each of the above and choose ρ and this penalty by validation. We use Gurobi to solve the SPO+ optimization problem, except for the RKHS case where because of the heavy computational burden of this, we must instead use SGD for n larger than 500. See details in Online Appendix E.2. By averaging over 50 replications of $\mathcal{D}$, we estimate relative regret, $\mathbb{E}_{\mathcal{D}}\mathbb{E}_X[f^*(X)^\top(\hat{\pi}(X) - \pi^*(X))]/\mathbb{E}_{\mathcal{D}}\mathbb{E}_X[f^*(X)\pi^*(X)]$, for each method and each $n = 50, 100, \ldots, 1,000$, shown in Figure 2(b) with shaded bands for plus/minus one standard error.

Although the theoretical results in Sections 2 and 3 do not directly apply to SPO+, our experimental results support our overall insights. With correctly specified models, the ETO method can achieve better performance than end-to-end methods that integrate estimation with optimization (see circle markers for "Correct linear"). However, from a practical lens, perfectly specified linear models are not realistic. For misspecified linear models, our experiments illustrate how end-to-end methods can

Figure 2. (Color online) Comparing ETO and SPO+ with Well-Specified, Misspecified, and Nonparametric Hypotheses



Notes. (a) The CLO instance is a stochastic shortest path problem. We need to go from s to t. The random cost of an edge j is $Y_j \in \mathbb{R}$. Whether we choose to proceed along an edge j is $z_j \in \{0, 1\}$. (b) The regret of different methods, relative to average minimal cost. Shaded regions represent 95% confidence intervals.

account for misspecification to obtain best-in-class performance, beating the corresponding misspecified ETO method (see square markers for "Wrong linear"). At the same time, we see that such best-in-class performance may sometimes still be bad in an absolute sense. Using more flexible models can sometimes close this gap. The kernel model (triangle markers) is still misspecified in the sense that the RKHS does not contain the true regression function and can only approximate it using functions of growing RKHS norm. When using such a flexible model, we observe that ETO achieves regret converging to zero with performance just slightly worse than the correctly specified case, whereas end-to-end methods have higher regret. Therefore, even though end-to-end methods handle decision-making problems more directly, our experiments demonstrate that the more straightforward ETO approach can be better even in decision-problem performance.

6. Concluding Remarks

In this paper, we studied the regret convergence rates for two approaches to CLO: the naïve, optimization-ignorant ETO and the end-to-end, optimization-aware IERM. We arrived at a surprising fact: the convergence rate for ETO is *orders faster* than for IERM, despite its ignoring the downstream effects of estimation. We reviewed various reasons for preferring either approach. This highlights a nuanced landscape for the enterprise to integrate estimation and optimization. The practical implications, nonetheless, are positive: relying on regression as a plug-in is easy and fast to run using existing tools and simple to interpret as predictions of uncertain variables; and as our results show, it provides downstream decisions with very good performance. Beyond providing new insights with practical implications, we hope our work inspires closer investigation of the statistical behavior of data-driven and end-to-end optimization in other settings. Section 4 points out *nonlinear* CSO as one interesting setting; other settings requiring attention include partial feedback (observe $Y^\top Z$, not Y, for historical Z), sequential/dynamic optimization problems, and online learning. The unique structure of constrained optimization problems brings up new algorithmic and statistical questions; the right approach is not always immediately clear, as we showed here for CLO.

Acknowledgments

The authors are listed in alphabetical order.

References

Audibert JY, Tsybakov AB (2007) Fast learning rates for plug-in classifiers. *Ann. Statist.* 35(2):608–633.

Bartlett PL, Bousquet O, Mendelson S (2005) Local Rademacher complexities. *Ann. Statist.* 33(4):1497–1537.

Barvinok A (2013) A bound for the number of vertices of a polytope with applications. *Combinatorica* 33(1):1–10.

Bastani H, Bayati M (2020) Online decision making with high-dimensional covariates. *Oper. Res.* 68(1):276–294.

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

Chen X, Owen Z, Pixton C, Simchi-Levi D (2021) A statistical learning approach to personalization in revenue management. *Management Sci.*, ePub ahead of print January 18, https://doi.org/10.1287/mnsc.2020.3772.

Devroye L, Lugosi G (1995) Lower bounds in pattern recognition and learning. *Pattern Recognition* 28(7):1011–1018.

Diao S, Sen S (2020) Distribution-free algorithms for learning enabled predictive stochastic programming. Technical Report, http://www.optimization-online.org/.

Donti P, Amos B, Kolter JZ (2017) Task-based end-to-end model learning in stochastic optimization. *Proc. 31st Internat. Conf. Neural Inform. Processing Systems*, 5490–5500.

Dudley R (1987) Universal Donsker classes and metric entropy. *Ann. Probab.* 15(4):1306–1326.

El Balghiti O, Elmachtoub AN, Grigas P, Tewari A (2019) Generalization bounds in the predict-then-optimize framework. *Proc. 33rd Internat. Conf. Neural Inform. Processing Systems*, 14412–14421.

Elmachtoub AN, Grigas P (2021) Smart "predict, then optimize." *Management Sci.* 68(1):9–26.

Elmachtoub AN, Liang JCN, McNellis R (2020) Decision trees for decision-making under the predict-then-optimize framework. *Proc. 37th Internat. Conf. Machine Learn.*, 2858–2867.

Estes A, Richard JP (2019) Objective-aligned regression for two-stage linear programs. Preprint, submitted October 23, https://dx.doi.org/10.2139/ssrn.3469897.

Foster DJ, Rakhlin A, Simchi-Levi D, Xu Y (2020) Instance-dependent complexity of contextual bandits and reinforcement learning. Preprint, submitted October 7, https://arxiv.org/abs/2010.03104.

Goldenshluger A, Zeevi A (2013) A linear response bandit problem. *Stochastic Systems* 3(1):230–261.

Hanneke S (2011) Rates of convergence in active learning. *Ann. Statist.* 39(1):333–361.

Henk M, Richter-Gebert J, Ziegler GM (2018) Basic properties of convex polytopes. Goodman JE, O'Rourke J, Toth CD, eds. *Handbook of Discrete and Computational Geometry* (CRC Press, Boca Raton, FL), 255–382.

Ho CP, Hanasusanto GA (2019) On data-driven prescriptive analytics with side information: A regularized Nadaraya-Watson approach. Technical Report, http://www.optimization-online.org/.

Ho-Nguyen N, Kilinç-Karzan F (2020) Risk guarantees for end-to-end prediction and optimization processes. Preprint, submitted December 30, https://arxiv.org/abs/2012.15046.

Hu Y, Kallus N, Mao X (2020) Smooth contextual bandits: Bridging the parametric and non-differentiable regret regimes. *PMLR* 125:2007–2010.

Kallus N, Mao X (2020) Stochastic optimization forests. *Management Sci.*, https://pubsonline.informs.org/doi/10.1287/opre.2021.2237.

Koltchinskii V (2006) Local Rademacher complexities and oracle inequalities in risk minimization. *Ann. Statist.* 34(6):2593–2656.

Loke G, Tang Q, Xiao Y (2020) Decision-driven regularization: Harmonizing the predictive and prescriptive. Technical Report, https://www.ssrn.com/.

Massart P, Nédélec É (2006) Risk bounds for statistical learning. *Ann. Statist.* 34(5):2326–2366.

McCullagh P, Nelder JA (1989) *Generalized Linear Models* (Chapman & Hall/CRC, London).

Notz PM, Pibernik R (2021) Prescriptive analytics for flexible capacity management. *Management Sci.*, ePub ahead of print May 6, https://doi.org/10.1287/mnsc.2020.3867.

Perchet V, Rigollet P (2013) The multi-armed bandit problem with covariates. *Ann. Statist.* 41(2):693–721.

Rigollet P, Zeevi A (2010) Nonparametric bandits with covariates. Preprint, submitted March 8, https://arxiv.org/abs/1003.1630.

Shalev-Shwartz S, Ben-David S (2014) *Understanding Machine Learning: From Theory to Algorithms* (Cambridge University Press, New York).

Tsybakov AB (2004) Optimal aggregation of classifiers in statistical learning. *Ann. Statist.* 32(1):135–166.

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

Van Der Vaart A, Wellner JA (2009) A note on bounds for VC dimensions. *IMS Collections* 5:103–107.

Vapnik V, Chervonenkis A (1974) *Theory of Pattern Recognition* (Nauka, Moscow).