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

MANAGEMENT SCIENCE
Vol. 66, No. 3, March 2020, pp. 1025–1044
ISSN 0025-1909 (print), ISSN 1526-5501 (online)

From Predictive to Prescriptive Analytics

Dimitris Bertsimas,[a] **Nathan Kallus**[b]

[a] Operations Research Center, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139; [b] Cornell Tech and School of Operations Research and Information Engineering, Cornell University, New York, New York 10044

Contact: dbertsim@mit.edu, http://orcid.org/0000-0002-1985-1003 (DB); kallus@cornell.edu, http://orcid.org/0000-0003-1672-0507 (NK)

Received: August 12, 2015
Revised: November 21, 2016; September 17, 2017; November 13, 2018
Accepted: November 13, 2018
Published Online in Articles in Advance: August 23, 2019

https://doi.org/10.1287/mnsc.2018.3253

Abstract. We combine ideas from machine learning (ML) and operations research and management science (OR/MS) in developing a framework, along with specific methods, for using data to prescribe optimal decisions in OR/MS problems. In a departure from other work on data-driven optimization, we consider data consisting, not only of observations of quantities with direct effect on costs/revenues, such as demand or returns, but also predominantly of observations of associated auxiliary quantities. The main problem of interest is a conditional stochastic optimization problem, given imperfect observations, where the joint probability distributions that specify the problem are unknown. We demonstrate how our proposed methods are generally applicable to a wide range of decision problems and prove that they are computationally tractable and asymptotically optimal under mild conditions, even when data are not independent and identically distributed and for censored observations. We extend these to the case in which some decision variables, such as price, may affect uncertainty and their causal effects are unknown. We develop the coefficient of prescriptiveness P to measure the prescriptive content of data and the efficacy of a policy from an operations perspective. We demonstrate our approach in an inventory management problem faced by the distribution arm of a large media company, shipping 1 billion units yearly. We leverage both internal data and public data harvested from IMDb, Rotten Tomatoes, and Google to prescribe operational decisions that outperform baseline measures. Specifically, the data we collect, leveraged by our methods, account for an 88% improvement as measured by our coefficient of prescriptiveness.

History: Accepted by Noah Gans, optimization.
Funding: This paper is based upon work supported by the National Science Foundation Graduate Research Fellowship [Grant 1122374].
Supplemental Material: The e-companion is available at https://doi.org/10.1287/mnsc.2018.3253.

Keywords: data-driven decision making • machine learning • stochastic optimization

1. Introduction

In today's data-rich world, many problems of operations research and management science (OR/MS) can be characterized by three primitives:

a. Data $\{y^1, \dots, y^N\}$ on uncertain quantities of interest $Y \in \mathcal{Y} \subset \mathbb{R}^{d_y}$, such as simultaneous demands.

b. Auxiliary data $\{x^1, \dots, x^N\}$ on associated covariates $X \in \mathcal{X} \subset \mathbb{R}^{d_x}$, such as recent sale figures, volume of Google searches for a products or company, news coverage, or user reviews, where x^i is concurrently observed with y^i.

c. A decision z constrained in $\mathcal{Z} \subset \mathbb{R}^{d_z}$ made after some observation $X = x$ with the objective of minimizing the *uncertain* costs $c(z; Y)$.

Traditionally, decision making under uncertainty in OR/MS has largely focused on the problem

$$v^{\text{stoch}} = \min_{z \in \mathcal{Z}} \mathbb{E}[c(z; Y)],$$
$$z^{\text{stoch}} \in \arg\min_{z \in \mathcal{Z}} \mathbb{E}[c(z; Y)] \qquad (1)$$

and its multiperiod generalizations and addressed its solution under a priori assumptions about the distribution μ_Y of Y (cf. Birge and Louveaux 2011), or, at times, in the presence of data $\{y^1, \dots, y^n\}$ in the assumed form of independent and identically distributed (iid) observations drawn from μ_Y (cf. Kleywegt et al. 2002, Shapiro 2003, Shapiro and Nemirovski 2005). (We will discuss examples of (1) in Section 1.1.) By and large, auxiliary data $\{x^1, \dots, x^N\}$ has not been extensively incorporated into OR/MS modeling, despite its growing influence in practice.

From its foundation, machine learning (ML), on the other hand, has largely focused on supervised learning, or the prediction of a quantity Y (usually univariate) as a function of X, based on data $\{(x^1, y^1), \dots, (x^N, y^N)\}$. By and large, ML does not address optimal decision-making under uncertainty that is appropriate for OR/MS problems.

At the same time, an explosion in the availability and accessibility of data and advances in ML have enabled applications that predict, for example, consumer demand for video games (Y) based on online web-search queries (X) (Choi and Varian 2012) or box-office ticket sales (Y) based on Twitter chatter (X) (Asur and Huberman 2010). There are many other applications of ML that proceed in a similar manner: use large-scale auxiliary data to generate predictions of a quantity that is of interest to OR/MS applications (Gruhl et al. 2005, Goel et al. 2010, Da et al. 2011, Kallus 2014). However, it is not clear how to go from a good prediction to a good decision. A good decision must take into account uncertainty wherever present. For example, in the absence of auxiliary data, solving (1) based on data $\{y^1, \ldots, y^n\}$ but using only the sample mean $\bar{y} = \sum_{i=1}^N y^i/N \approx \mathbb{E}[Y]$ and ignoring all other aspects of the data would generally lead to inadequate solutions to (1) and an unacceptable waste of good data.

In this paper, we combine ideas from ML and OR/MS in developing a framework, along with specific methods, for using data to prescribe optimal decisions in OR/MS problems that leverage auxiliary observations. Specifically, the problem of interest is

$$v^*(x) = \min_{z \in \mathcal{Z}} \mathbb{E}\left[c(z;Y)|X = x\right],$$
$$z^*(x) \in \mathcal{Z}^*(x) = \arg\min_{z \in \mathcal{Z}} \mathbb{E}\left[c(z;Y)|X = x\right], \qquad (2)$$

where the underlying distributions are unknown and only data $S_N = \{(x^1, y^1), \ldots, (x^N, y^N)\}$ is available. The solution $z^*(x)$ to (2) represents the full-information optimal decision, which, via full knowledge of the unknown joint distribution $\mu_{X,Y}$ of (X, Y), leverages the observation $X = x$ to the fullest possible extent in minimizing costs. We use the term *predictive prescription* for any function $z(x)$ that prescribes a decision in anticipation of the future given the observation $X = x$. Our task is to use S_N to construct a data-driven predictive prescription $\hat{z}_N(x)$. Our aim is that its performance in practice, $\mathbb{E}[c(\hat{z}_N(x);Y)|X = x]$, is close to the full-information optimum, $v^*(x)$.

Our key contributions include the following.

a. We propose various ways for constructing predictive prescriptions $\hat{z}_N(x)$ The focus of the paper is predictive prescriptions that have the form

$$\hat{z}_N(x) \in \arg\min_{z \in \mathcal{Z}} \sum_{i=1}^N w_{N,i}(x) c(z; y^i), \qquad (3)$$

where $w_{N,i}(x)$ are weight functions derived from the data. We motivate specific constructions inspired by a variety of predictive ML methods. We briefly summarize a selection of these constructions that we find the most effective below.

b. We also consider a construction motivated by the traditional empirical risk minimization (ERM) approach to ML. This construction has the form

$$\hat{z}_N(\cdot) \in \arg\min_{z(\cdot) \in \mathcal{F}} \frac{1}{N} \sum_{i=1}^N c(z(x^i); y^i), \qquad (4)$$

where $\mathcal{F}$ is some class of functions. We extend the standard ML theory of out-of-sample guarantees for ERM to the case of multivariate-valued decisions encountered in OR/MS problems. We find, however, that in the specific context of OR/MS problems, the construction (4) suffers from some limitations that do not plague the predictive prescriptions derived from (3).

c. We show that that our proposals are computationally tractable under mild conditions.

d. We study the asymptotics of our proposals under sampling assumptions more general than iid by leveraging universal law-of-large-number results of Walk (2010). Under appropriate conditions and for certain predictive prescriptions $\hat{z}_N(x)$ we show that costs with respect to the true distributions converge to the full information optimum, that is, $\lim_{N\to\infty} \mathbb{E} \cdot \left[c(\hat{z}_N(x);Y)|X = x\right] = v^*(x)$, and that prescriptions converge to true full information optimizers, that is, $\lim_{N\to\infty} \inf_{z \in Z^*(x)} \|z - \hat{z}_N(x)\| = 0$, both for almost everywhere x and almost surely. We extend our results to the case of censored data (such as observing demand via sales).

e. We extend the above results to the case in which some of the decision variables may affect the uncertain variable in unknown ways not encapsulated in the known cost function. In this case, the uncertain variable $Y(z)$ will be different depending on the decision and the problem of interest becomes $\min_{z \in \mathcal{Z}} \mathbb{E} \cdot \left[c(z; Y(z))|X = x\right]$. Complicating the construction of a data-driven predictive prescription, however, is that the data only includes the realizations $Y_i = Y_i(Z_i)$ corresponding to historic decisions. For example, in problems that also involve pricing decisions, price has an unknown causal effect on demand that must be determined in order to optimize the full decision z. We show that under certain conditions our methods can be extended to this case while preserving favorable asymptotic properties.

f. We introduce a new metric P, termed *the coefficient of prescriptiveness*, in order to measure the efficacy of a predictive prescription and to assess the prescriptive content of covariates X, that is, the extent to which observing X is helpful in reducing costs. An analogue to the coefficient of determination R^2 of predictive analytics, P is a unitless quantity that is (eventually) bounded between 0 (not prescriptive) and 1 (highly prescriptive).

g. We demonstrate in a real-world setting the power of our approach. We study an inventory management problem faced by the distribution arm of an international media conglomerate. This entity manages over 1/2 million unique items at some 50,000 retail locations around the world, with which it has vendor-managed inventory (VMI) and scan-based trading (SBT) agreements. On average it ships about 1 billion units a year. We leverage both internal company data and, in the spirit of the aforementioned ML applications, large-scale public data harvested from online sources, including IMDb, Rotten Tomatoes, and Google Trends. These data combined, leveraged by our approach, lead to large improvements in comparison with baseline measures, in particular accounting for an 88% improvement toward the deterministic perfect-foresight counterpart.

Of our proposed constructions of predictive prescriptions $\hat{z}_N(x)$, the ones that we find to be generally the most broadly and practically effective are the following:

a. Motivated by k-nearest-neighbors regression (kNN; Trevor et al. 2001, chap. 13),

$$\hat{z}_N^{k\mathrm{NN}}(x) \in \arg\min_{z \in \mathcal{Z}} \sum_{i \in \mathcal{N}_k(x)} c(z; y^i), \tag{5}$$

where $\mathcal{N}_k(x) = \{i = 1, \ldots, N : \sum_{j=1}^N \mathbb{I}[\|x - x_i\| \geq \|x - x_j\|] \leq k\}$ is the neighborhood of the k data points that are closest to x.

b. Motivated by local linear regression (LOESS; Cleveland and Devlin 1988),

$$\hat{z}_N^{\mathrm{LOESS}^*}(x) \in \arg\min_{z \in \mathcal{Z}} \sum_{i=1}^n k_i(x) \max\Bigg\{1 - \sum_{j=1}^n k_j(x) \cdot (x^j - x)^T \Xi(x)^{-1} (x^i - x), 0\Bigg\} c(z; y^i), \tag{6}$$

where $\Xi(x) = \sum_{i=1}^n k_i(x)(x^i - x)(x^i - x)^T$, $k_i(x) = (1 - (\|x^i - x\|/h_N(x))^3)^3 \mathbb{I}[\|x^i - x\| \leq h_N(x)]$, and $h_N(x) > 0$ is the distance to the k-nearest point from x.

c. Motivated by classification and regression trees (CART; Breiman et al. 1984),

$$\hat{z}_N^{\mathrm{CART}}(x) \in \arg\min_{z \in \mathcal{Z}} \sum_{i : R(x^i) = R(x)} c(z; y^i), \tag{7}$$

where $R(x) \in \{1, \ldots, r\}$ is the leaf corresponding to input x in a regression tree trained on S_N.

d. Motivated by random forests (RF; Breiman 2001),

$$\hat{z}_N^{\mathrm{RF}}(x) \in \arg\min_{z \in \mathcal{Z}} \sum_{t=1}^T \frac{1}{|\{j : R^t(x^j) = R^t(x)\}|} \cdot \sum_{i : R^t(x^i) = R^t(x)} c(z; y^i), \tag{8}$$

where $R^t(x)$ is the leaf map for the t^{th} tree in the random forest ensemble trained on S_N.

Further detail and other constructions are given in Section 2 and supplemental Section EC.1.

1.1. An Illustrative Example

In this section, we discuss different approaches to problem (2) and compare them in a two-stage linear decision-making problem, illustrating the value of auxiliary data and the methodological gap to be addressed. We illustrate this with synthetic data but, in Section 6, we study a real-world problem and use real-world data.

The specific problem we consider is a two-stage shipment planning problem. We have a network of d_z warehouses that we use in order to satisfy the demand for a product at d_y locations. We consider two stages of the problem. In the first stage, some time in advance, we choose amounts $z_i \geq 0$ of units of product to produce and store at each warehouse i, at a cost of $p_1 > 0$ per unit produced. In the second stage, demand $Y \in \mathbb{R}^{d_y}$ realizes at the locations and we must ship units to satisfy it. We can ship from warehouse i to location j at a cost of c_{ij} per unit shipped (recourse variable $s_{ij} \geq 0$) and we have the option of using last-minute production at a cost of $p_2 > p_1$ per unit (recourse variable t_i). The overall problem has the cost function and feasible set

$$c(z; y) = p_1 \sum_{i=1}^{d_z} z_i + \min_{(t,s) \in \mathcal{Q}(z,y)} \left(p_2 \sum_{i=1}^{d_z} t_i + \sum_{i=1}^{d_z} \sum_{j=1}^{d_y} c_{ij} s_{ij} \right), \quad \mathcal{Z} = \{z \in \mathbb{R}^{d_z} \ : \ z \geq 0\},$$

where $\mathcal{Q}(z, y) = \{(s, t) \in \mathbb{R}^{(d_z \times d_y) \times d_z} : t \geq 0,\ s \geq 0, \sum_{i=1}^{d_z} s_{ij} \geq y_j \, \forall j,\ \sum_{j=1}^{d_y} s_{ij} \leq z_i + t_i \, \forall i\}$.

The key concern is that we do not know Y or its distribution. We consider the situation where we only have data $S_N = ((x^1, y^1), \ldots, (x^N, y^N))$ consisting of observations of Y along with concurrent observations of some auxiliary quantities X that may be associated with the future value of Y. For example, X may include past product sales at each of the different locations, weather forecasts at the locations, or volume of Google searches for a product to measure consumer attention.

We consider two possible existing data-driven approaches to leveraging such data for making a decision. One approach is the sample average approximation of stochastic optimization (SAA). SAA is only concerned with the marginal distribution of Y, thus ignoring data on X, and solves the following data-driven optimization problem

$$\hat{z}_N^{\mathrm{SAA}} \in \arg\min_{z \in \mathcal{Z}} \frac{1}{N} \sum_{i=1}^N c(z; y^i), \tag{9}$$

whose objective approximates $\mathbb{E}\left[c(z; Y)\right]$.

Machine learning, on the other hand, leverages the data on X as it tries to predict Y given observations $X = x$. Consider for example a random forest trained on the data S_N. It provides a point prediction $\hat{m}_N(x)$ for the value of Y when $X = x$. Given this prediction, one possibility is to consider the approximation of the random variable Y by our best-guess value $\hat{m}_N(x)$ and solve the corresponding optimization problem,

$$\hat{z}_N^{\text{point-pred}} \in \arg\min_{z\in\mathcal{Z}} c(z;\hat{m}_N(x)). \tag{10}$$

The objective approximates $c(z;\mathbb{E}[Y|X=x])$. We call (10) a point-prediction-driven decision.

If we knew the full joint distribution of Y and X, then the optimal decision having observed $X = x$ is given by (2). Let us compare SAA and the point-prediction-driven decision (using a random forest) to this optimal decision in the two decision problems presented. Let us also consider our proposals (5)–(8) and others that will be introduced in Section 2.

We consider a particular instance of the two-stage shipment planning problem with $d_z = 5$ warehouses and $d_y = 12$ locations, where we observe some features predictive of demand. In both cases we consider $d_x = 3$ and data S_N that, instead of iid, is sampled from a multidimensional evolving process in order to simulate real-world data collection. We give the particular parameters of the problems in Section EC.6 of the e-companion. In Figure 1(a), we report the average performance of the various solutions with respect to the *true* distributions.

The full-information optimum clearly does the best with respect to the true distributions, as expected. The SAA and point-prediction-driven decisions have performances that quickly converge to suboptimal values. The former because it does not use observations on X and the latter because it does not take into account the remaining uncertainty after observing $X = x$.[1] In comparison, we find that our proposals converge upon the full-information optimum given sufficient data. In Section 4.3, we study the general asymptotics of our proposals and prove that the convergence observed here empirically is generally guaranteed under only mild conditions.

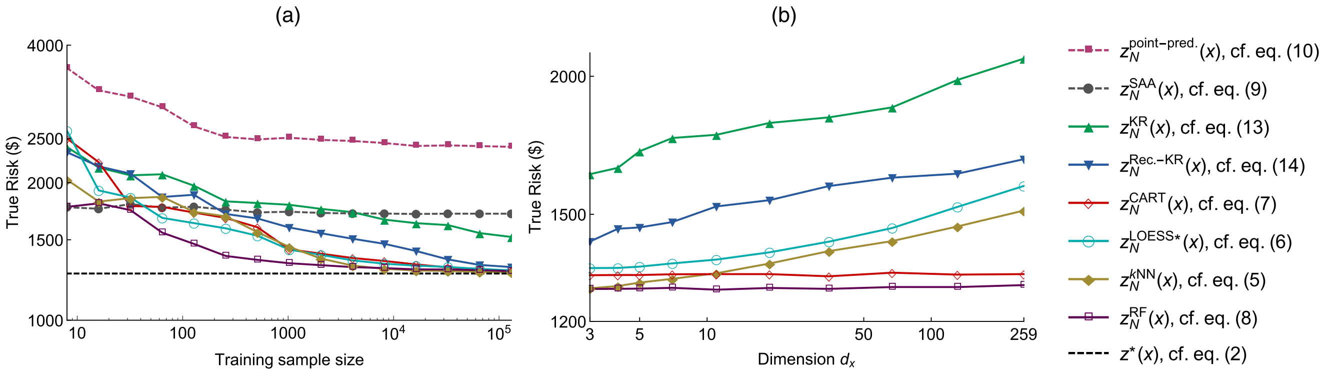
Inspecting the figure further, it seems that ignoring X and using only the data on Y, as SAA does, is appropriate when there is very little data; in both examples, SAA outperforms other data-driven approaches for N smaller than ~64. Past that point, our constructions of predictive prescriptions, in particular (5)–(8), leverage the auxiliary data effectively and achieve better, and eventually optimal, performance. The predictive prescription motivated by RF is notable in particular for performing no worse than SAA in the small N regime, and better in the large N regime.

In this example, the dimension d_x of the observations x was relatively small at $d_x = 3$. In many practical problems, this dimension may well be bigger, potentially inhibiting performance. For example, in our real-world application in Section 6, we have $d_x = 91$. To study the effect of the dimension of x on the performance of our proposals, we consider polluting x with additional dimensions of uninformative components distributed as independent normals. The results, shown in Figure 1(b), show that while some of the predictive prescriptions show deteriorating performance with growing dimension d_x, the predictive prescriptions based on CART and RF are largely unaffected, seemingly able to detect the 3-dimensional subset of features that truly matter. In the supplemental Section EC.6.2 we also consider an alternative setting of this experiment where additional dimensions carry marginal predictive power.

1.2. Relevant Literature

Stochastic optimization like in (1) has long been the focus of decision making under uncertainty in OR/MS problems (cf. Birge and Louveaux 2011) as has its multiperiod generalization known commonly as dynamic programming (cf. Bertsekas 1995). The solution of stochastic optimization problems like in (1)

Figure 1. Performance of Various Prescriptions with Respect to True Distributions, Averaged over Samples and New Observations x (Lower Is Better)



Varying sample size N $(d_x = 3)$.

Varying dimension d_x $(N = 2^{14})$.

Note. Note the horizontal and vertical log scales.

in the presence of data $\{y^1, \ldots, y^N\}$ on the quantity of interest is a topic of active research. The traditional approach is the sample average approximation (SAA) where the true distribution is replaced by the empirical one (cf. Kleywegt et al. 2002, Shapiro 2003, Shapiro and Nemirovski 2005). Other approaches include stochastic approximation (cf. Robbins and Monro 1951, Nemirovski et al. 2009), robust SAA (cf. Bertsimas et al. 2018b), and data-driven mean-variance distributionally robust optimization (cf. Delage and Ye 2010). A notable alternative approach to decision making under uncertainty in OR/MS problems is robust optimization (cf. Ben-Tal et al. 2009) and its data-driven variants (cf. Bertsimas et al. 2018a). A vast literature considers the trade-off between the collection of data and optimization as informed by data collected so far (cf. Robbins 1952, Lai and Robbins 1985, Besbes and Zeevi 2009). In all these methods for data-driven decision-making under uncertainty, the focus is on data in the assumed form of iid observations of the parameter of interest Y. On the other hand, ML has attached great importance to the problem of supervised learning of the conditional expectation (regression) or mode (classification) of target quantities Y given auxiliary observations $X = x$ (cf. Trevor et al. 2001, Mohri et al. 2012).

Statistical decision theory is generally concerned with the optimal selection of statistical estimators (cf. Berger 1985, Lehmann and Casella 1998). Following the early work of Wald (1949), a loss function, such as the sum of squared errors or of absolute deviations, is specified and the corresponding admissibility, minimax-optimality, or Bayes-optimality are of main interest. Statistical decision theory and ML intersect most profoundly in the realm of regression via empirical risk minimization (ERM), where a regression model is selected on the criterion of minimizing empirical average of loss. A range of ML methods arise from ERM applied to certain function classes and extensive theory on function-class complexity has been developed to analyze these (cf. Vapnik 1992, Bartlett and Mendelson 2003). Such ML methods include ordinary linear regression, ridge regression, the LASSO of Tibshirani (1996), quantile regression, and ℓ_1-regularized quantile regression of Belloni and Chernozhukov (2011). ERM is also closely connected with M-estimation (Geer 2000), which estimates a distributional parameter that maximizes an average of a function of the parameter by the estimate that maximizes the corresponding empirical average. Unlike M-estimation theory, which is concerned with estimation and inference, ERM theory is only concerned with out-of-sample performance and can be applied more flexibly with less assumptions.

In certain OR/MS decision problems, one can employ ERM to select a decision policy, conceiving of the loss as costs. Indeed, the loss function used in quantile regression is exactly equal to the cost function of the newsvendor problem. Ban and Rudin (2018) consider this loss function and the selection of a univariate-valued linear function with coefficients restricted in ℓ_1-norm in order to solve a newsvendor problem with auxiliary data, resulting in a method similar to Belloni and Chernozhukov (2011). Kao et al. (2009) study finding a convex combination of two ERM solutions, the least-cost decision and the least-squares predictor, which they find to be useful when costs are quadratic. In the supplemental Section EC.1, we generalize the ERM approach to general decision problems, where decisions may be multivariate, and prove a performance guarantee.

Our main predictive-prescription proposals are motivated more by a strain of nonparametric ML methods based on local learning, where predictions are made based on the mean or mode of past observations that are in some way similar to the one at hand. The most basic such method is kNN (cf. Trevor et al. 2001, chap. 13), which defines the prediction as a locally constant function depending on which k data points lie closest. A related method is Nadaraya-Watson kernel regression (KR; Nadaraya 1964, Watson 1964). KR weighting for solving conditional stochastic optimization problems like in (2) has been considered in Hanasusanto and Kuhn (2013) and Hannah et al. (2010), but these have not considered the more general connection to a great variety of ML methods used in practice nor have they considered asymptotic optimality. A generalization of KR is local polynomial regression (Cameron and Trivedi 2005, p. 311), of which the LOESS method of Cleveland and Devlin (1988) is a specific case. Local learning also includes recursive partitioning methods, which are most often in the form of trees like the CART method (Breiman et al. 1984). Ensembles of trees, most notably RF of Breiman (2001), are also a form of local learning and are known to be flexible learner with competitive performance in a great range of prediction problems.

2. From Data to Predictive Prescriptions

Recall that we are interested in the conditional-stochastic optimization problem (2) of minimizing uncertain costs $c(z; Y)$ after observing $X = x$. The key difficulty is that the true joint distribution $\mu_{X,Y}$, which specifies problem (2), is unknown and only data S_N is available. One approach may be to approximate $\mu_{X,Y}$ by the empirical distribution $\hat{\mu}_N$ over the data S_N where each datapoint (x^i, y^i) is assigned mass $1/N$. This, however, will in general fail unless X has small and finite support; otherwise, either $X = x$ has not been observed and the conditional expectation is undefined with respect to $\hat{\mu}_N$ or it has been observed, $X = x = x^i$ for some i, and the conditional distribution is a degenerate distribution with a single atom at y^i

without any uncertainty. Therefore, we require some way to generalize the data to reasonably estimate the conditional expected costs for any x. In some ways this is similar to, but more intricate than, the prediction problem where $\mathbb{E}[Y|X = x]$ is estimated from data for any possible $x \in \mathcal{X}$. Therefore, we are motivated to consider predictive methods and their adaptation to our cause.

In the next subsections we propose a selection of constructions of predictive prescriptions $\hat{z}_N(x)$, each motivated by a local-learning predictive methodology. All the constructions in this section will take the common form of defining some data-driven weights $w_{N,i}(x)$ and optimizing the decision $\hat{z}_N$ against a reweighting of the data, like in (3):

$$\hat{z}_N^{\text{local}}(x) \in \arg\min_{z \in \mathcal{Z}} \sum_{i=1}^{N} w_{N,i}(x) c(z; y^i). \tag{11}$$

Whenever the weights are nonnegative, they can be understood to correspond to an estimated conditional distribution of Y given $X = x$.

2.1. *k*NN

Motivated by kNN regression, we propose

$$w_{N,i}^{k\text{NN}}(x) = \tfrac{1}{k}\mathbb{I}[x^i \text{ is a } k\text{NN of } x], \tag{12}$$

giving rise to the predictive prescription (5). Ties among equidistant data points are broken either randomly or by a lower-index-first rule. Finding the kNNs of x can clearly be done in $O(Nd)$ time. This can be sped up by precomputation (Bentley 1975) or by approximation (Arya et al. 1998).

2.2. Kernel Methods

Motivated by KR, which uses a kernel K to measure distances in x, we propose

$$w_{N,i}^{\text{KR}}(x) = \frac{K\big((x^i - x)/h_N\big)}{\sum_{j=1}^{N} K\big((x^j - x)/h_N\big)}, \tag{13}$$

where $K : \mathbb{R}^d \to \mathbb{R}$ satisfies $\int K < \infty$ and $h_N > 0$, known as the bandwidth. Our weights (13) also can be thought of as the ratio of the Parzen-window density estimates (Parzen 1962) of $\mu_{X,Y}$ and μ_X. We restrict our attention to the following common kernels: $K(x) = \mathbb{I}\,[\|x\| \le 1]$ (Naïve), $K(x) = (1 - \|x\|^2)\mathbb{I}\,[\|x\| \le 1]$ (Epanechnikov), and $K(x) = (1 - \|x\|^3)^3\mathbb{I}\,[\|x\| \le 1]$ (Tricubic). For example, the naïve kernel uniformly weights all neighbors of x that are within a radius h_N.

We also propose a variant with varying bandwidths, motivated by Devroye and Wagner (1980):

$$w_{N,i}^{\text{recursive-KR}}(x) = \frac{K\big((x^i - x)/h_i\big)}{\sum_{j=1}^{N} K\big((x^j - x)/h_j\big)}. \tag{14}$$

2.3. Local Linear Methods

Motivated by LOESS (Cleveland and Devlin 1988), we propose

$$\begin{aligned} w_{N,i}^{\text{LOESS}}(x) &= \frac{\tilde{w}_{N,i}(x)}{\sum_{j=1}^{N} \tilde{w}_{N,j}(x)}, \\ \tilde{w}_{N,i}(x) &= k_i(x)\Big(1 - \sum_{j=1}^{n} k_j(x)(x^j - x)^T \Xi(x)^{-1} \\ &\qquad \cdot (x^i - x)\Big), \end{aligned} \tag{15}$$

where $\Xi(x) = \sum_{i=1}^{n} k_i(x)(x^i - x)(x^i - x)^T$ and $k_i(x) = K((x^i - x)/h_N(x))$. This corresponds to the idea of approximating $\mathbb{E}\big[c(z; Y)\big|X = x\big]$ locally by a linear function in x (Cameron and Trivedi 2005, p. 311). Because the weights in (15) may be negative, we also propose a modification that only uses the nonnegative weights, which we show helps with tractability (Section 4.1):

$$\begin{aligned} w_{N,i}^{\text{LOESS*}}(x) &= \frac{\tilde{w}_{N,i}(x)}{\sum_{j=1}^{N} \tilde{w}_{N,j}(x)}, \\ \tilde{w}_{N,i}(x) &= k_i(x) \max\Big\{1 - \sum_{j=1}^{n} k_j(x)(x^j - x)^T \\ &\qquad \cdot \Xi(x)^{-1}(x^i - x), 0\Big\}. \end{aligned} \tag{16}$$

2.4. Trees

Motivated by tree-based methods, given any map $\mathcal{R} : \mathcal{X} \to \{1, \ldots, r\}$, we propose

$$w_{N,i}^{\text{CART}}(x) = \frac{\mathbb{I}[\mathcal{R}(x) = \mathcal{R}(x^i)]}{\big|\{j : R(x^j) = R(x)\}\big|}. \tag{17}$$

The map $\mathcal{R}$ corresponds to the disjoint partition $\mathcal{X} = \mathcal{R}^{-1}(1) \sqcup \cdots \sqcup \mathcal{R}^{-1}(r)$. In particular, we propose using the partition generated by training CART (Breiman et al. 1984) on the data S_n, so that $\mathcal{R}(x)$ is the identity of the leaf that an input x is assigned to.[2] Notice that the weights (17) are piecewise constant over the partitions and therefore the recommended optimal decision $\hat{z}_N(x)$ is also piecewise constant. Therefore, solving r optimization problems after the recursive partitioning process, the resultant predictive prescription can be fully compiled into a decision tree, with the decisions that are truly decisions. This also retains CART's interpretability.

2.5. Ensembles

Motivated by tree-ensemble methods, given T maps $\mathcal{R}^t : \mathcal{X} \to \{1, \ldots, r^t\}$, $t = 1, \ldots, T$, we propose

$$w_{N,i}^{\text{RF}}(x) = \frac{1}{T}\sum_{t=1}^{T} \frac{\mathbb{I}[\mathcal{R}^t(x) = \mathcal{R}^t(x^i)]}{\big|\{j : R^t(x^j) = R^t(x)\}\big|}. \tag{18}$$

In particular, we propose to use the binning rules from the individual trees of a RF (Breiman 2001) trained on the data S_n. Based on the performance of this approach seen in Section 1.1, we choose this predictive prescription in our real-world application in Section 6.

3. From Data to Predictive Prescriptions When Decisions Affect Uncertainty

Up to now, we have assumed that the effect of the decision z on costs is wholly encapsulated in the cost function and that the choice of z does not directly affect the realization of uncertainty Y. However, in some settings, such as in the presence of pricing decisions, this assumption clearly does not hold. As one increases a price control, demand diminishes, and the *causal* effect of pricing on demand is not known a priori (e.g., can be abstracted in the cost function) and must be derived from data. In such cases, we must take into account the effect of our decision z on the uncertainty Y by considering historical data $\{(x^1, y^1, z^1), \ldots, (x^N, y^N, z^N)\}$, where we have also recorded historical observations of the variable Z, which represents the historical decision taken in each instance. Using potential outcomes, we let $Y(z)$ denote the value of the uncertain variable that would be observed if decision z were chosen. (For detail on potential outcomes and history, see Imbens and Rubin 2015, chap. 1–2.) For each data point i, only the realization corresponding to the chosen decision z^i is revealed, $y^i = y^i(z^i)$. The main issue we face in dealing with such is known as the *fundamental problem of causal inference*: we do not observe the counterfactuals $y^i(z)$ for any $z \neq z^i$. Thus, when trying to score how a new decision z would have done in a particular historical instance in the data, we are faced with the problem that we *do not know* what our cost $c(z; y^i(z))$ would have been. This cost function is *distinct* from the observed cost function $c(z; y^i) = c(z; y^i(z^i))$, in stark contrast to the setting considered in the previous sections.

Because only some parts of our decision may have unknown effects on uncertainty, we decompose our decision variable into the part with unknown effect (e.g., pricing decisions) and known effect (e.g., production decisions) in the following way:

Assumption 1 (Decomposition of Decision). *For some decomposition $z = (z_1, z_2)$ only $z_1 \in \mathbb{R}^{d_{z_1}}$ affects uncertainty, that is,*

$$Y(z_1, z_2) = Y(z_1, z_2') \quad \forall (z_1, z_2), (z_1, z_2') \in \mathcal{Z}.$$

For brevity, we write $Y(z) = Y(z_1)$. And we let $\mathcal{Z}_1(z_2) = \{z_1 : (z_1, z_2) \in \mathcal{Z}\}$, $\mathcal{Z}_1 = \{z_1 : \exists z_2\ (z_1, z_2) \in \mathcal{Z}\}$, $\mathcal{Z}_2(z_1) = \{z_2 : (z_1, z_2) \in \mathcal{Z}\}$, $\mathcal{Z}_2 = \{z_2 : \exists z_1\ (z_1, z_2) \in \mathcal{Z}\}$.

For example, in pricing, if $z_1 \in [0, \infty)$ represents a price control for a product and Y represents realized demand, then $\{(z_1, Y(z_1)) : z_1 \in [0, \infty)\}$ represents the *random* demand curve. If in the i^{th} data point the price was z_1^i, then we only observe the single point $(z_1^i, y^i(z_1^i))$ on this random curve. Decision components z_2 could represent, for example, a production and shipment plan, which does not affect demand but does affect final costs as encapsulated in the cost function $c(z; y)$.

The immediate generalization of problem (2) to this setting is

$$v^*(x) = \min_{z \in \mathcal{Z}} \mathbb{E}\left[c(z; Y(z)) \middle| X = x\right],$$
$$z^*(x) \in \mathcal{Z}^*(x) = \arg\min_{z \in \mathcal{Z}} \mathbb{E}\left[c(z; Y(z)) \middle| X = x\right]. \quad (19)$$

This problem depends on understanding the joint distribution of $(X, Y(z))$ for each $z \in \mathcal{Z}$ and, in this full information setting, chooses z for least expected cost given the observation $X = x$ *and* the effect z would have on the uncertainty $Y(z)$. Assumption 1 allows problem (19) to encompass the standard conditional stochastic optimization problem (2) by letting $z = z_2$ and $d_{z_1} = 0$. On the other hand, Assumption 1 is nonrestrictive in the sense that it can be as general as necessary by letting $z = z_1$, that is, no decomposing into parts of unknown effect and known no effect. For these reasons, we maintain the notation $v^*(x)$, $z^*(x)$, $\mathcal{Z}^*(x)$.

Given only the data (x^i, y^i, z^i) on (X, Y, Z) and without any assumptions, problem (19) is in fact not well specified because of the missing data on the counterfactuals. In particular, the most we can hope to learn from observations of (X, Y, Z) is the joint distribution of (X, Y, Z). However, there may be many joint distributions of $(X, \{Y(z) : z \in \mathcal{Z}\})$, each giving rise to a different solution $z^*(x)$ in problem (19), that all agree with a given single joint distribution of $(X, Y = Y(Z), Z)$ under some choice of distribution for Z (cf. Bertsimas and Kallus 2016). For example, in the most general setting, it may impossible to discern whether high demand was due to low prices or other factors, such as consumer interest, season, or advertising.

To eliminate this issue, we must make additional assumptions about the data. Here, we assume that controlling for X is sufficient for isolating the effect of z on Y.

Assumption 2 (Ignorability). *For every $z \in \mathcal{Z}$, $Y(z)$ is independent of Z conditioned on X.*

In words, Assumption 2 says that, historically, X accounts for all the features influencing managerial decision making that may also correlate with the potential outcomes $Y(z)$. In causal inference, this assumption is standard for ensuring identifiability of causal effects (Rosenbaum and Rubin 1983).

In stark contrast to many situations in causal inference dealing with latent self-selection, Assumption 2 is particularly defensible in our setting. In the setting we consider, Z represents historical managerial decisions and, like future decisions to be made by the learned predictive prescription, these decisions must have been made based on observable quantities available to the manager. As long as these quantities were also recorded as part of X, then Assumption 2 is guaranteed to hold. Alternatively, were decisions Z taken at random for exploration, then Assumption 2 holds trivially.

3.1. Adapting Local-Learning Methods

We now show how to generalize the predictive prescriptions from Section 2 to solve problem (19) when decisions affect uncertainty based on data on (X, Y, Z). We begin with a rephrasing of problem (19) based on Assumptions 1 and 2. The proof is given in the e-companion.

Theorem 1. *Under Assumptions 1 and 2, problem (19) is equivalent to*

$$\min_{(z_1,z_2)\in\mathcal{Z}} \mathbb{E}\left[c(z;Y)\middle|X=x,\ Z_1=z_1\right]. \tag{20}$$

Note that problem (20) depends on the distribution of the data $(X,\ Y,\ Z)$, does not involve unknown counterfactuals, and has the form of a conditional stochastic optimization problem. Correspondingly, all predictive-prescriptive local-learning methods from Section 2 can be adapted to this problem by simply augmenting the data x^i with z_1^i. In particular, we can consider data-driven predictive prescriptions of the form

$$\hat{z}_N(x) \in \arg\min_{z\in\mathcal{Z}} \sum_{i=1}^{N} w_{N,i}(x,z_1)c(z;y^i), \tag{21}$$

where $w_{N,i}(x,z_1)$ are weight functions derived from the data by taking the same approach used in Section 2 but treating z_1 as part of the X data. Thus, the objective of problem (21) can be solely computed from the observed data, given a weighting scheme $w_{N,i}(\tilde{x})$ and a cost function $c(z;y)$. In particular, for each method in Section 2, to compute the objective for a given $z=(z_1,z_2)$, we let $\tilde{x}^i=(x^i,z_1^i)$, construct weights $w_{N,i}(\tilde{x})$ based on data $\tilde{S}_N=\{(\tilde{x}^1,y^1),\dots,(\tilde{x}^N,y^N)\}$, and plug $w_{N,i}(\tilde{x})$ into (21). For example, our kNN approach applied to (19) has the form (21) with weights

$$w_{N,i}^{k\mathrm{NN}}(x,z_1)=\tfrac{1}{k}\mathbb{I}\left[(x^i,z_1^i)\text{ is a }k\text{NN of }(x,z_1)\right].$$

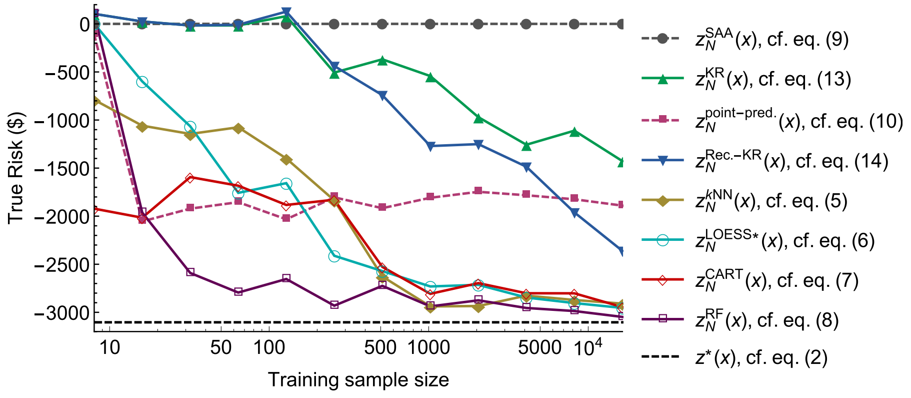
Then $\hat{z}_N(x)$ would be given by optimizing this objective over $z\in\mathcal{Z}$. As we discuss in Section 4.2, there is an increased computational burden in this last step of optimizing problem (21) over z when decisions affect uncertainty, compared with our standard predictive prescriptions from Section 2. To address this, we propose both a discretization approach and a specialized algorithm for the case of tree-based weights. Furthermore, we show in Section 4.4 that this approach produces prescriptions that are asymptotically optimal even when our decisions have an unknown effect on uncertainty.

3.2. Example: Two-Stage Shipment Planning with Pricing

Consider a pricing variation on our two-stage shipment planning problem from Section 1.1. We introduce an additional decision variable $z_1\in[0,\infty)$ for the price at which we sell the product. The uncertain demand at the d_y locations, $Y(z_1)$, depends on the price we set. In the first stage, we determine price z_1 and amounts z_2 at d_{z_2} warehouses. In the second stage, instead of shipping from warehouses to satisfy all demand, we can ship as much as we would like. Our profit is the price times number of units sold minus production and transportation costs. Assuming we behave optimally in the second stage, we can write the problem using the cost function and feasible set

$$c(z;y)=p_1\sum_{i=1}^{d_{z_2}} z_{2,i}+\min_{(t,s)\in Q(z,y)}\Big(p_2\sum_{i=1}^{d_{z_2}} t_i$$
$$+\sum_{i=1}^{d_{z_2}}\sum_{j=1}^{d_y}(c_{ij}-z_1)s_{ij}\Big),$$
$$\mathcal{Z}=\left\{(z_1,z_2)\in\mathbb{R}^{1+d_{z_2}}\ :\ z_1,z_2\geq 0\right\},$$

Figure 2. Performance of Various Prescriptions in the Two-Stage Shipment Planning with a Pricing Problem



where $\mathcal{Q}(z, y) = \{(s,t) \in \mathbb{R}^{(d_z \times d_y) \times d_z} : t \geq 0,\ s \geq 0, \sum_{i=1}^{d_z} s_{ij} \leq y_j\ \forall j, \sum_{j=1}^{d_y} s_{ij} \leq z_{2,i} + t_i\ \forall i\}$.

We now consider observing not only X and Y but also Z_1. We consider the same parameters of the problem used in Section 1.1 with an added unknown effect of price on demand so that higher prices induce lower demands. The particular parameters are given in Section EC.6 of the e-companion. In Figure 2, we report the average negative profits (production and shipment costs less revenues) of various solutions with respect to the true distributions. We include the full information optimum (19) and all our local-learning methods applied as described in Section 3.1. Again, we compare with SAA and to the point-prediction-driven decision (using a random forest to fit $\hat{m}_N(x, z_1)$, a predictive model based on both x and z_1).

We see that our local-learning methods converge upon the full-information optimum as more data becomes available. On the other hand, SAA, which considers only data y^i, will always have out-of-sample profits 0 as it will drive z_1 to infinity, where demand goes to zero faster than linear. The point-prediction-driven decision performs comparatively well for small N, learning quickly the average effect of pricing, but does not converge to the full-information optimum as we gather more data. Overall, our predictive-prescription using RF that addresses the unknown effect of pricing decisions on uncertain demand performs the best.

4. Properties of Local Predictive Prescriptions

In this section, we study two important properties of local predictive prescriptions: computational tractability and asymptotic optimality. All proofs are given in the e-companion.

4.1. Tractability

In Section 2, we considered a variety of predictive prescriptions $\hat{z}_N(x)$ that are computed by solving the optimization problem (3). An important question is whether this optimization problem is computationally tractable to solve. As an optimization problem, problem (3) differs from the problem solved by the standard SAA approach (9) only in the weights given to different observations. Therefore, it is similar in its computational complexity, and we can defer to computational studies of SAA, such as Shapiro and Nemirovski (2005), to study the complexity of solving problem (3). For completeness, we develop sufficient conditions for problem (3) to be solvable in polynomial time.

Theorem 2. *Fix x and weights $w_{N,i}(x) \geq 0$. Suppose $\mathcal{Z}$ is a closed convex set and let a separation oracle for it be given. Suppose also that $c(z; y)$ is convex in z for every fixed y, and let oracles be given for evaluation and subgradient in z. Then for any x we can find an ϵ-optimal solution to (3) in time and oracle calls polynomial in N_0, d, $\log(1/\epsilon)$ where $N_0 = \sum_{i=1}^{N} \mathbb{I}[w_{N,i}(x) > 0] \leq N$ is the effective sample size.*

Note that all weights presented in Section 2 are nonnegative, with the exception of local regression (15), which is what led us to their nonnegative modification (16).

4.2. Tractability When Decisions Affect Uncertainty

Solving problem (21) with general weights $w_N^i(x, z_1)$ is generally hard as the objective of problem (21) may be nonconvex in z. In some specific instances we can maintain tractability, while in others we can devise specialized approaches that allow us to solve problem (21) in practice.

In the simplest case, if $\mathcal{Z}_1 = \{z_{11}, \ldots, z_{1b}\}$ is discrete, then the problem can be simply solved by optimizing once for each fixed value of z_1, letting z_2 remain variable.

Theorem 3. *Fix x and weights $w_{N,i}(x, z_1) \geq 0$. Suppose $\mathcal{Z}_1 = \{z_{11}, \ldots, z_{1b}\}$ is discrete and that $\mathcal{Z}_2(z_{1j})$ is a closed convex set for each $j = 1, \ldots, b$ and let a separation oracle for it be given. Suppose also that $c((z_1, z_2); y)$ is convex in z_2 for every fixed y, z_1 and let oracles be given for evaluation and subgradient in z_2. Then for any x we can find an ϵ-optimal solution to (21) in time and oracle calls polynomial in N_0, b, d, $\log(1/\epsilon)$ where $N_0 = \sum_{i=1}^{N} \mathbb{I}[w_{N,i}(x) > 0] \leq N$ is the effective sample size.*

Note that the convexity in z_2 condition is weaker than convexity in z, which would be sufficient.

Alternatively, if $\mathcal{Z}_1$ is not discrete, we can approach the problem using discretization, which leads to exponential dependence in z_1's dimension d_{z_1} and the precision $\log(1/\epsilon)$.

Theorem 4. *Fix x and weights $w_{N,i}(x, z_1) \geq 0$. Suppose $c((z_1, z_2); y)$ is L-Lipschitz in z_1 for each $z_2 \in \mathcal{Z}_2$, that $\mathcal{Z}_1$ is bounded, and that $\mathcal{Z}_2(z_1)$ is a closed convex set for each $z_1 \in \mathcal{Z}_1$ and let a separation oracle for it be given. Suppose also that $c((z_1, z_2); y)$ is convex in z_2 for every fixed y, z_1 and let oracles be given for evaluation and subgradient in z_2. Then, for any x, we can find an ϵ-optimal solution to (21) in time and oracle calls polynomial in N_0, b, d, $\log(1/\epsilon)$, $(L/\epsilon)^{d_{z_1}}$, where $N_0 = \sum_{i=1}^{N} \mathbb{I}\left[w_{N,i}(x) > 0\right] \leq N$ is the effective sample size.*

The exponential dependence in d_{z_1} and super-logarithmic dependence in $1/\epsilon$ limit this approach to small d_{z_1} (although d_z may still be big). For example, we use this approach in our pricing example in Section 3.2, where $d_{z_1} = 1$, to successfully solve many instances of (21).

For the specific case of tree weights, we can discretize the problem *exactly*, leading to a particularly efficient algorithm in practice. Suppose we are given the CART partition rule $\mathcal{R} : \mathcal{X} \times \mathcal{Z}_1 \to \{1, \dots, r\}$, then we can solve problem (21) exactly as follows:

1. Let x be given and fix

$$w_{N,i}^{\mathrm{CART}}(x, z_1) = \frac{\mathbb{I}[\mathcal{R}(x, z_1) = \mathcal{R}(x^i, z_1^i)]}{\left|\{j : R(x^j, z_1^j) = R(x, z_1)\}\right|}.$$

2. Find the partitions that contain x, $\mathcal{J} = \{j : \exists z_1, (x, z_1) \in R^{-1}(j)\}$, and compute the constraints on z_1 in each part, $\tilde{\mathcal{Z}}_{1j} = \{z_1 : \exists x, \ (x, z_1) \in R^{(-1)}(j)\}$ for $j \in \mathcal{J}$. This is easily done by going down the tree and at each node, if the node queries the value of x we only take the branch that corresponds to the value of our given x and if the node queries the value of a component of z_1, then we take both branches and record the linear constraint on z_1 on each side.

3. For each $j \in \mathcal{J}$, solve

$$v_j = \min_{z \in \mathcal{Z} : z \in \tilde{\mathcal{Z}}_{1j}} \sum_{i : R(x^i, z_1^i) = j} c(z; y^i),$$
$$z_j = \arg\min_{z \in \mathcal{Z} : z \in \tilde{\mathcal{Z}}_{1j}} \sum_{i : R(x^i, z_1^i) = j} c(z; y^i).$$

(These can be solved for in advance for each $j = 1, \dots, r$ to reduce computation at query time.)

4. Let $j(x) = \arg\min_{j \in \mathcal{J}} v_j$ and $\hat{z}_n(x) = z_{j(x)}$.

This procedure solves (21) *exactly* for weights $w_{N,i}^{\mathrm{CART}}(x, z_1)$.

4.3. Asymptotic Optimality

In Section 1.1, we saw that our predictive prescriptions $\hat{z}_N(x)$ converged to the full-information optimum as the sample size N grew. Next, we show that this anecdotal evidence is supported by mathematics and that such convergence is guaranteed under only mild conditions. We define *asymptotic optimality* as the desirable asymptotic behavior for $\hat{z}_N(x)$.

Definition 1. We say that $\hat{z}_N(x)$ is *asymptotically optimal* if, with probability 1, we have that for μ_X-almost-everywhere $x \in \mathcal{X}$,

$$\lim_{N \to \infty} \mathbb{E}\left[c(\hat{z}_N(x); Y) \middle| X = x\right] = v^*(x).$$

We say $\hat{z}_N(x)$ is *consistent* if, with probability 1, we have that for μ_X-almost-everywhere $x \in \mathcal{X}$,

$$\lim_{N \to \infty} \|\hat{z}_N(x) - Z^*(x)\| = 0, \text{ where}$$
$$\|\hat{z}_N(x) - Z^*(x)\| = \inf_{z \in Z^*(x)} \|\hat{z}_N(x) - z\|.$$

To a decision maker, asymptotic optimality is the most critical limiting property as it says that decisions implemented will have performance reaching the best possible. Consistency refers to the consistency of $\hat{z}_N(x)$ as a statistical estimator for the full-information optimizer(s) $\mathcal{Z}^*(x)$ and is perhaps less critical for a decision maker but will be shown to hold nonetheless.

Asymptotic optimality depends on our choice of $\hat{z}_N(x)$, the structure of the decision problem (cost function and feasible set), and on how we accumulate our data S_N. The traditional assumption on data collection is that it constitutes an iid process. This is a strong assumption and is often only a modeling approximation. The velocity and variety of modern data collection often means that historical observations do not generally constitute an iid sample in any real-world application. Therefore, we are motivated to consider an alternative model for data collection, that of mixing processes. Mixing encompasses processes such as ARMA, GARCH, and Markov chains, which can correspond to sampling from evolving systems like prices in a market, daily product demands, or the volume of Google searches on a topic. Although many of our results extend to such settings via generalized strong laws of large numbers (Walk 2010), we present only the iid case in the main text to avoid cumbersome exposition and defer these extensions to the supplemental Section EC.2.2. For the rest of the section, let us assume that S_N is generated by iid sampling.

As mentioned, asymptotic optimality also depends on the structure of the decision problem. Therefore, we will also require the following conditions.

Assumption 3 (Existence). *The full-information problem (2) is well defined: $\mathbb{E}[|c(z; Y)|] < \infty$ for every $z \in \mathcal{Z}$ and $\mathcal{Z}^*(x) \neq \varnothing$ for almost every x.*

Assumption 4 (Continuity). *$c(z; y)$ is equicontinuous in z: for any $z \in \mathcal{Z}$ and $\epsilon > 0$ there exists $\delta > 0$ such that $|c(z; y) - c(z'; y)| \leq \epsilon$ for all z' with $\|z - z'\| \leq \delta$ and $y \in \mathcal{Y}$.*

Assumption 5 (Regularity). *$\mathcal{Z}$ is closed and nonempty and in addition either*

1. *$\mathcal{Z}$ is bounded or*
2. *$\liminf_{\|z\| \to \infty} \inf_{y \in \mathcal{Y}} c(z; y) > -\infty$ and for every $x \in \mathcal{X}$, there exists $D_x \subset \mathcal{Y}$ such that $\lim_{\|z\| \to \infty} c(z; y) \to \infty$ uniformly over $y \in D_x$ and $\mathbb{P}(y \in D_x | X = x) > 0$.*

Under these conditions, we have the following sufficient conditions for asymptotic optimality, which are proven as consequences of universal pointwise convergence results of related supervised learning problem of Walk (2010) and Hansen (2008).

Theorem 5 (*k*NN). *Suppose Assumptions 3–5 hold. Let $w_{N,i}(x)$ be as in (12) with $k = \min\left\{\lceil CN^{\delta} \rceil, N - 1\right\}$ for some $C > 0$, $0 < \delta < 1$. Let $\hat{z}_N(x)$ be as in (3). Then $\hat{z}_N(x)$ is asymptotically optimal and consistent.*

Theorem 6 (Kernel Methods). *Suppose Assumptions 3–5 hold and that $\mathbb{E}\left[|c(z; Y)| \max\left\{\log |c(z; Y)|, 0\right\}\right] < \infty$ for each z. Let $w_{N,i}(x)$ be as in (13) with K being any of the kernels in*

Section 2.2 and $h_N = CN^{-\delta}$ for $C>0$, $0<\delta<1/d_x$. Let $\hat{z}_N(x)$ be as in (3). Then $\hat{z}_N(x)$ is asymptotically optimal and consistent.

Theorem 7 (Recursive Kernel Methods). *Suppose Assumptions 3–5 hold. Let $w_{N,i}(x)$ be as in (14) with K being the naïve kernel and $h_i = Ci^{-\delta}$ for some $C>0$, $0<\delta<1/(2d_x)$. Let $\hat{z}_N(x)$ be as in (3). Then $\hat{z}_N(x)$ is asymptotically optimal and consistent.*

Theorem 8 (Local Linear Methods). *Suppose Assumptions 3–5 hold, that μ_X is absolutely continuous and has density bounded away from 0 and ∞ on the support of X and twice continuously differentiable, and that costs are bounded over y for each z (i.e., $|c(z;y)| \le g(z)$) and twice continuously differentiable. Let $w_{N,i}(x)$ be as in (15) with K being any of the kernels in Section 2.2 and $h_N = CN^{-\delta}$ for some $C>0$, $0<\delta<1/d_x$. Let $\hat{z}_N(x)$ be as in (3). Then $\hat{z}_N(x)$ is asymptotically optimal and consistent.*

Theorem 9 (Nonnegative Local Linear Methods). *Suppose Assumptions 3–5 hold, that μ_X is absolutely continuous and has density bounded away from 0 and ∞ on the support of X and twice continuously differentiable,* and *that costs are bounded over y for each z (i.e., $|c(z;y)| \le g(z)$) and twice continuously differentiable. Let $w_{N,i}(x)$ be as in (16) with K being any of the kernels in Section 2.2 and $h_N = CN^{-\delta}$ for some $C>0$, $0<\delta<1/d_x$. Let $\hat{z}_N(x)$ be as in (3). Then $\hat{z}_N(x)$ is asymptotically optimal and consistent.*

Although we do not have firm theoretical results on the asymptotic optimality of the predictive prescriptions based on CART [Equation (7)] and RF [Equation (8)], we have observed them to converge empirically in Section 1.1.

4.4. Asymptotic Optimality When Decisions Affect Uncertainty

When decisions affect uncertainty, the condition for asymptotic optimality is subtly different. Under the identity $Y = Y(Z)$, Definition 1 does not accurately reflect asymptotic optimality and indeed methods that do not account for the unknown effect of the decision (e.g., if we apply our methods without regard to this effect, ignoring data on Z_1) will not reach the full-information optimum given by (19). Instead, we would like to ensure that our decisions have optimal cost when taking into account their effect on uncertainty. The desired asymptotic behavior for $\hat{z}_N(x)$ when decisions affect uncertainty is the more general condition given below.

Definition 2. We say that $\hat{z}_N(x)$ is *asymptotically optimal* if, with probability 1, we have that for μ_X-almost-everywhere $x \in \mathcal{X}$, as $N \to \infty$

$$\lim_{N\to\infty} \mathbb{E}\left[c(\hat{z}_N(x); Y(\hat{z}_N(x)))\middle| X = x\right] = \min_{z\in\mathcal{Z}} \mathbb{E}\left[c(z; Y(z))\middle| X = x\right].$$

The following theorem establishes asymptotic optimality for our predictive prescription based on either kernel methods, local linear methods, or nonnegative local linear methods as adapted to the case when decisions affect uncertainty. Like in Section 3.1, we use $\tilde{x}^i$ to denote (x^i, z_1^i) and $\tilde{S}_N = (\tilde{x}^1, y^1), \ldots, (\tilde{x}^N, y^N)$. To avoid issues of existence, we focus on weak minimizers $\hat{z}_N(x)$ of (21) and on asymptotic optimality.

Theorem 10. *Suppose Assumptions 1–5 (case 1) hold, that $\mu_{(X,Z_1)}$ is absolutely continuous and has density bounded away from 0 and ∞ on the support of X, Z_1 and twice continuously differentiable, and that costs are bounded over y for each z (i.e., $|c(z;y)| \le g(z)$) and twice continuously differentiable. Let $w_{N,i}(\tilde{x})$ be as in (13), (15), or (16) applied to $\tilde{S}_N$ with K being any of the kernels in Section 2.2 and with $h_N = CN^{-\delta}$ for some $C>0$, $0<\delta<1/(d_x + d_{z_1})$. Then for any $\epsilon_N \longrightarrow 0$, any $\hat{z}_N(x)$ that ϵ_N-minimizes (21) (has objective value within ϵ_N of the infimum) is asymptotically optimal.*

5. Metrics of Prescriptiveness

In this section, we develop a relative, unitless measure of the efficacy of a predictive prescription. An absolute measure of efficacy is marginal expected costs,

$$R(\hat{z}_N) = \mathbb{E}\left[\mathbb{E}\left[c\left(\hat{z}_N(X); Y\right)\middle| X\right]\right] = \mathbb{E}\left[c(\hat{z}_N(X); Y)\right].$$

Given a validation data set $\overline{S}_{N_v} = \left((\overline{x}^1, \overline{y}^1), \cdots, (\overline{x}^{N_v}, \overline{y}^{N_v})\right)$, we estimate $R(\hat{z}_N)$ as

$$\hat{R}_{N_v}(\hat{z}_N) = \frac{1}{N_v}\sum_{i=1}^{N_v} c\left(\hat{z}_N(\overline{x}^i); \overline{y}^i\right).$$

If $\overline{S}_{N_v}$ is disjoint and independent of the training set S_N, then this is an out-of-sample estimate that provides an unbiased estimate of $R(\hat{z}_N)$. While an absolute measure allows one to compare two predictive prescriptions for the same problem and data, a relative measure can quantify the overall prescriptive content of the data and the efficacy of a prescription on a universal scale. For example, in predictive analytics, the coefficient of determination R^2—rather than the absolute root-mean-squared error—is a unitless quantity used to quantify the overall quality of a prediction and the predictive content of data X. R^2 measures the fraction of variance of Y reduced, or "explained," by the prediction based on X. Another way of interpreting R^2 is as the fraction of the way that X and a particular predictive model take us from a data-poor prediction (the sample average) to a perfect-foresight prediction that knows Y in advance.

We define an analogous quantity for the predictive prescription problem, which we term *the coefficient of prescriptiveness*. It involves three quantities. First,

$$\hat{R}_{N_v}(\hat{z}_N) = \frac{1}{N_v}\sum_{i=1}^{N_v} c\left(\hat{z}_N(\overline{x}^i); \overline{y}^i\right)$$

is the estimated expected costs because of our predictive prescription. Second,

$$\hat{R}^*_{N_v} = \frac{1}{N_v}\sum_{i=1}^{N_v} \min_{z\in\mathcal{Z}} c(z;\bar{y}^i)$$

is the estimated expected costs in the deterministic perfect-foresight counterpart problem, in which one has foreknowledge of Y without any uncertainty (note the difference to the full-information optimum, which does have uncertainty). Third,

$$\hat{R}_{N_v}(z_N^{\mathrm{SAA}}) = \frac{1}{N_v}\sum_{i=1}^{N_v} c(\hat{z}_N^{\mathrm{SAA}};\bar{y}^i), \quad \text{where}$$

$$\hat{z}_N^{\mathrm{SAA}} \in \arg\min_{z\in\mathcal{Z}} \tfrac{1}{N}\sum_{i=1}^{N} c(z;y^i)$$

is the estimated expected costs of a data-driven prescription that is data poor, only based on Y data. This is the SAA solution to the prescription problem, which serves as the analogue to the sample average as a data-poor solution to the prediction problem. Using these three quantities, we define the coefficient of prescriptiveness P as follows:

$$P = 1 - (\hat{R}_{N_v}(\hat{z}_N) - \hat{R}^*_{N_v}) \;/\; (\hat{R}_{N_v}(z_N^{\mathrm{SAA}}) - \hat{R}^*_{N_v}). \tag{22}$$

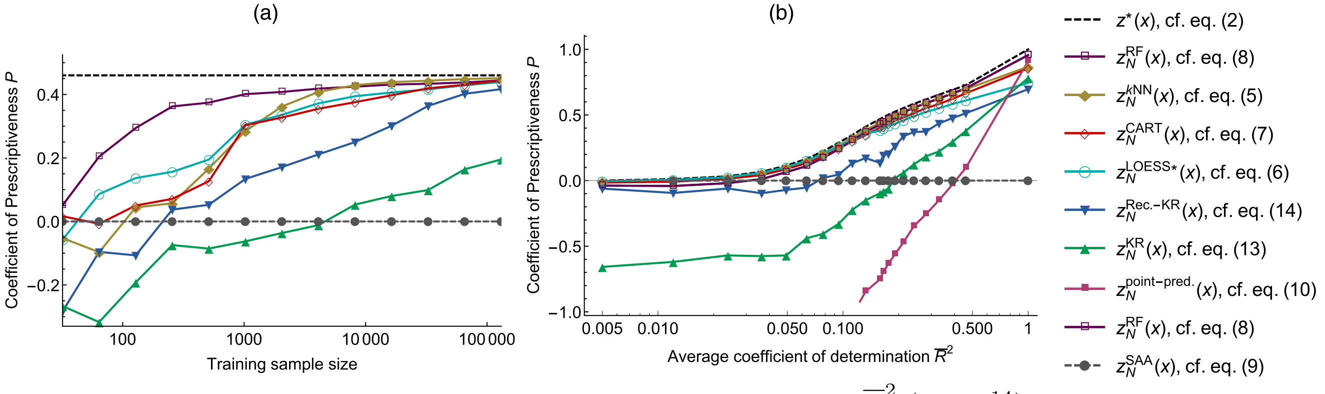
The coefficient of prescriptiveness P is a unitless quantity bounded above by 1. A low P denotes that X provides little useful information for the purpose of prescribing an optimal decision in the particular problem at hand or that $\hat{z}_N(x)$ is ineffective in leveraging the information in X. A high P denotes that taking X into consideration significantly affects reducing costs and that $\hat{z}_N(x)$ is effective in leveraging X for this purpose.

In particular, if X is independent of Y then, under appropriate conditions, $\lim_{N,N_v\to\infty} \hat{R}_{N_v}(z_N^{\mathrm{SAA}}) = \min_{z\in\mathcal{Z}} \mathbb{E}\,[c(z;Y)] = \mathbb{E}\left[\min_{z\in\mathcal{Z}} \mathbb{E}\left[c(z;Y)\middle|X\right]\right] = \lim_{N,N_v\to\infty} \cdot \hat{R}_{N_v}(\hat{z}_N)$, so as N grows, we would see P reach 0. On the other hand, if Y is measurable with respect to X, that is, Y is a function of X, then, under appropriate conditions, $\lim_{N,N_v\to\infty} \hat{R}_{N_v}(\hat{z}_N) = \mathbb{E}[\min_{z\in\mathcal{Z}} \mathbb{E}[c(z;Y)|X]] = \mathbb{E}[\min_{z\in\mathcal{Z}} c(z;Y)] = \lim_{N_v\to\infty} \hat{R}^*_{N_v}$, so as N grows, we would see P reach 1. It is also notable that in the extreme case that Y is a function of X, then $Y = m(X)$, where $m(x) = \mathbb{E}\left[Y\middle|X=x\right]$ so that $\mathbb{E}[\min_{z\in\mathcal{Z}} c(z;Y)] = \mathbb{E}[\min_{z\in\mathcal{Z}} c(z;m(X))]$, and so in this extreme case we would also see P reach 1 for $\hat{z}_N^{\text{point-pred}}$ under appropriate conditions. On the other hand, in the independent case, we would always see P reach a *nonpositive* number under $\hat{z}_N^{\text{point-pred}}$.

Let us consider the coefficient of prescriptiveness in the example from Section 1.1. For each of our predictive prescriptions and for each N, we measure the out of sample P on a validation set of size $N_v = 200$ and plot the results in Figure 3(a). Notice that even when we converge to the full-information optimum, P does not approach 1 as N grows. Instead we see that for the same methods that converged to the full-information optimum, we have a P that approaches 0.46. This number represents the extent of the potential that X has to reduce costs in this particular problem. It is the fraction of the way that knowledge of X, leveraged correctly, takes us from making a decision under full uncertainty about the value of Y to making a decision in a completely deterministic setting. As is the case with R^2, what magnitude of P denotes a successful application depends on the context. In our real-world application in Section 6, we find an out-of-sample P of 0.88.

To consider the relationship between how predictive X is of Y and the coefficient of prescriptiveness, we consider modifying the example by varying the magnitude of residual noise, fixing $N = 2^{14}$. The details are given

Figure 3. The Coefficient of Prescriptiveness P in the Example from Section 1.1, Measured out of Sample



Varying sample size N.

Varying determination $\overline{R}^2$ $(N = 2^{14})$.

Note. The dashed horizontal line denotes the theoretical limit.

in the supplemental Section EC.6. As we vary the noise, we can vary the average coefficient of determination

$$\overline{R}^2 = 1 - \frac{1}{d_y}\sum_{i=1}^{d_y} \frac{\mathbb{E}[\mathrm{Var}(Y_i | X)]}{\mathrm{Var}(Y_i)}$$

from 0 to 1. In the original example, $\overline{R}^2 = 0.16$. We plot the results in Figure 3(b), noting that the behavior matches our description of the extremes above. In particular, when X and Y are independent ($\overline{R}^2 = 0$), we see most methods having a zero coefficient of prescriptiveness, less successful methods (KR) have a somewhat negative coefficient, and the point-prediction-driven decision has a very negative coefficient. When Y is measurable with respect to X ($\overline{R}^2 = 1$), the coefficient of the optimal decision reaches 1, most methods have a coefficient near 1, and the point-prediction-driven decision also has a coefficient near 1 and beats most other methods. While neither extreme exists in practice, throughout the range in between the extremes, our predictive prescriptions perform the best and in particular the one based on RF.

6. A Real-World Application

In this section, we apply our approach to a real-world problem faced by the distribution arm of an international media conglomerate (the vendor) and demonstrate that our approach, combined with extensive data collection, leads to significant advantages. The vendor has asked us to keep its identity confidential as well as data on sale figures and specific retail locations. Therefore, some figures are shown on relative scales.

6.1. Problem Statement

The vendor sells over 0.5 million entertainment media titles on CD, DVD, and BluRay at over 50,000 retailers across the United States and Europe. On average they ship 1 billion units in a year. The retailers range from electronic home goods stores to supermarkets, gas stations, and convenience stores. These have vendor-managed inventory (VMI) and scan-based trading (SBT) agreements with the vendor. VMI means that the inventory is managed by the vendor, including replenishment (which they perform weekly) and planogramming. SBT means that the vendor owns all inventory until sold to the consumer, at which point the retailer buys the unit from the vendor and sells it to the consumer. This means that retailers have no cost of capital in holding the vendor's inventory.

The cost of a unit of entertainment media consists mainly of the cost of production of the content. Media-manufacturing and delivery costs are secondary in effect. Therefore, the primary objective of the vendor is simply to sell as many units as possible and the main limiting factor is inventory capacity at the retail locations. For example, at many of these locations, shelf space for the vendor's entertainment media is limited to an aisle endcap display and no back-of-the-store storage is available. Thus, the main loss incurred in over-stocking a particular product lies in the loss of potential sales of another product that sold out but could have sold more. In studying this problem, we will restrict our attention to the replenishment and sale of video media only and to retailers in Europe.

Apart from the limited shelf space the other main reason for the difficulty of the problem is the particularly high uncertainty inherent in the initial demand for new releases. Whereas items that have been sold for at least one period have a somewhat predictable decay in demand, determining where demand for a new release will start is a much less trivial task. At the same time, new releases present the greatest opportunity for high demand and many sales.

We now formulate the full-information problem. Let $r = 1, \ldots, R$ index the locations, $t = 1, \ldots, T$ index the replenishment periods, and $j = 1, \ldots, d$ index the products. Denote by z_j the order quantity decision for product j, by Y_j the uncertain demand for product j, and by K_r the overall inventory capacity at location r. Considering only the *main* effects on revenues and costs as discussed in the previous paragraph, the problem decomposes on a per-replenishment-period, per-location basis. We therefore wish to solve, for each t and r, the following problem:

$$\begin{aligned} v^*(x_{tr}) = \max \quad & \mathbb{E}\left[\sum_{j=1}^{d} \min\{Y_j, z_j\} \middle| X = x_{tr}\right] \\ &= \sum_{j=1}^{d} \mathbb{E}\left[\min\{Y_j, z_j\} \middle| X_j = x_{tr}\right] \\ \text{s.t.} \quad & z \geq 0, \ \sum_{j=1}^{d} z_j \leq K_r, \end{aligned} \tag{23}$$

where x_{tr} denotes auxiliary data available at the beginning of period t in the $(t, r)^{\text{th}}$ problem.

Note that had there been no capacity constraint in problem (23) and a per-unit ordering cost were added, the problem would decompose into d separate newsvendor problems, the solution to each being exactly a quantile regression on the regressors x_{tr}. As it is, the problem is coupled, but, fixing x_{tr}, the capacity constraint can be replaced with an equivalent per-unit ordering cost λ via Lagrangian duality and the optimal solution is attained by setting each z_j to the λ^{th} conditional quantile of Y_j. However, the reduction to quantile regression does not hold because the dual optimal value of λ *simultaneously* depends on all the conditional distributions of Y_j for $j = 1, \ldots, d$.

6.2. Applying Predictive Prescriptions to Censored Data

In applying our approach to problem (23), we face the issue that we have data on sales, not demand. That is, our data on the quantity of interest Y is right-censored. In this section, we develop a modification of our approach to correct for this. The results in this section apply generally.

Suppose that instead of data $\{y^1, \dots, y^N\}$ on Y, we have data $\{u^1, \dots, u^N\}$ on $U = \min\{Y, V\}$ where V is an observable random threshold, data on which we summarize via $\delta = \mathbb{I}[U < V]$. For example, in our application, V is the on-hand inventory level at the beginning of the period. Overall, our data consist of $S_N' = \{(x^1, u^1, \delta^1), \dots, (x^N, u^N, \delta^N)\}$.

One way to deal with this is by considering decisions (sock levels) as affecting uncertainty (sales). As long as demand and threshold are conditionally independent given X, Assumption 2 will be satisfied and we can use the approach (21) developed in Section 3.1. However, the particular setting of censored data has a lot structure where we actually know the mechanism of *how* decision affect uncertainty. This allows us to develop a special-purpose solution that side-steps the need to learn the structure of this dependence and computationally less tractable approaches (Section 4.2).

To correct for the fact that our observations are, in fact, censored, we develop a conditional variant of the Kaplan-Meier method (cf. Kaplan and Meier 1958, Huh et al. 2011) to transform our weights appropriately. Let (i) denote the ordering $u^{(1)} \le \cdots \le u^{(N)}$. Given the weights $w_{N,i}(x)$ generated based on the naïve assumption that $y^i = u^i$, we transform these into the weights

$$w_{N,(i)}^{\text{Kaplan-Meier}}(x) = \mathbb{I}\big[\delta^{(i)} = 1\big]\left(\frac{w_{N,(i)}(x)}{\sum_{\ell=i}^{N} w_{N,(\ell)}(x)}\right) \cdot \prod_{k \le i-1\,:\,\delta k=1} \left(\frac{\sum_{\ell=k+1}^{N} w_{N,(\ell)}(x)}{\sum_{\ell=k}^{N} w_{N,(\ell)}(x)}\right). \tag{24}$$

Next, we show that the transformation (24) preserves asymptotic optimality under certain conditions. The proof is in the e-companion.

Theorem 11. *Suppose that Y and V are conditionally independent given X, that Y and V share no atoms, that for every $x \in \mathcal{X}$ the upper support of V given $X = x$ is greater than the upper support of Y given $X = x$, and that costs are bounded over y for each z (i.e., $|c(z; y)| \le g(z)$). Let $w_{N,i}(x)$ be as in (12), (13), (14), (15), or (16) and suppose the corresponding assumptions of Theorem 5, 6, 7, 8, or 9 apply. Let $\hat{z}_N(x)$ be as in (3) but using the transformed weights (24). Then $\hat{z}_N(x)$ is asymptotically optimal and consistent.*

The assumption that Y and V share no atoms (which holds in particular if either is continuous) provides that $\delta \overset{a.s.}{=} \mathbb{I}[Y \le V]$ so that the event of censorship is observable. In applying this to problem (23), the assumption that Y and V are conditionally independent given X, which mirrors Assumption 2, will hold if X captures at least all of the information that past stocking decisions, which are made before Y is realized, may have been based on. The assumption on bounded costs applies to problem (23) because the cost (negative of the objective) is bounded in $[-K_r, 0]$.

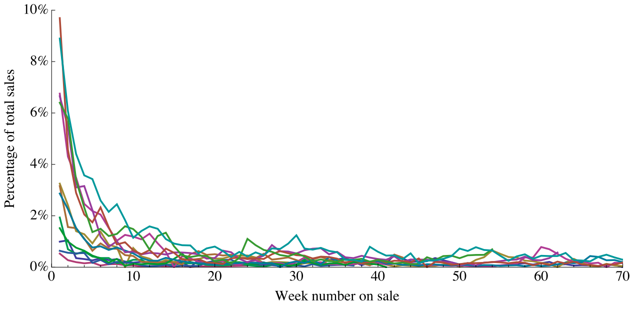
6.3. Data

In this section, we describe the data collected. To get at the best data-driven predictive prescription, we combine both internal company data and public data harvested from online sources. The predictive power of such public data has been extensively documented in the literature (cf. Gruhl et al. 2005, Asur and Huberman 2010, Goel et al. 2010, Da et al. 2011, Choi and Varian 2012, Kallus 2014). Here, we study its *prescriptive* power.

6.3.1. Internal Data. The internal company data consists of 4 years of sale and inventory records across the network of retailers, information about each of the locations, and information about each of the items.

We aggregate the sales data by week (the replenishment period of interest) for each feasible combination of location and item. As discussed above, these sales-per-week data constitute a right-censored observation of weekly demand, where censorship occurs when an item is sold out. We developed the transformed weights (24) to tackle this issue exactly. Figure 4 shows the sales life cycle of a selection of titles in terms of their marketshare when they are released to home entertainment (HE) sales and onwards. Because new releases can attract up to almost 10% of *all* sales in the week of their release, they pose a great sales opportunity, but, at the same time, significant demand uncertainty. Information about retail locations includes to which chain a location belongs and the address of the location. To parse the address and obtain a precise position of the location, including country and subdivision, we used the Google Geocoding API (Application Programming Interface).[3] Information about items include the medium (e.g.,

Figure 4. Percentage of All Sales in the German State of Berlin Taken up by Each of 13 Selected Titles, Starting from the Point of Release of Each Title to HE Sales



DVD or BluRay) and an item "title." The title is a short descriptor composed by a local marketing team in charge of distribution and sales in a particular region and may often include information beyond the title of the underlying content. For example, a hypothetical film titled *The Film* sold in France may be given the item title "THE FILM DVD + LIVRET - EDITION FR," implying that the product is a French edition of the film, sold on a DVD, and accompanied by a booklet (*livret*), whereas the same film sold in Germany on BluRay may be given the item title "FILM, THE (2012) - BR SINGLE," indicating it is sold on a single BluRay disc.

6.3.2. Public Data: Item Metadata, Box Office, and Reviews. We sought to collect additional data to characterize the items and how desirable they may be to consumers. For this we turned to the Internet Movie Database (IMDb; www.imdb.com) and Rotten Tomatoes (RT; www.rottentomatoes.com). IMDb is an online database of information on films and TV series. RT is a website that aggregates professional reviews from newspapers and online media, along with user ratings, of films and TV series.

To harvest information from these sources on the items being sold by the vendor, we first had to disambiguate the item entities and extract original content titles from the item titles. Having done so, we extract the following information from IMDb: type (film, TV, other/unknown); U.S. original release date of content (e.g., in theaters); average IMDb user rating (0–10); number of IMDb users voting on rating; number of awards (e.g., Oscars for films, Emmys for TV) won and number nominated for; the main actors (i.e., first-billed); plot summary (30–50 words); genre(s) (of 26; can be multiple); and MPAA rating (e.g., PG-13, NC-17) if applicable. And the following information from RT: professional reviewers' aggregate score; RT user aggregate rating; number of RT users voting on rating; and if a film, then American box office gross when shown in theaters.

In Figure 5, we provide scatter plots of some of these attributes against sale figures in the first week of an HE release. Notice that the number of users voting on the rating of a title is much more indicative of HE sales than the quality of a title as reported in the aggregate score of these votes.

6.3.3. Public Data: Search Engine Attention. In the above, we saw that box office gross is reasonably informative about future HE sale figures. The box office gross we are able to access, however, is for the American market and is also missing for various European titles. We therefore would like additional data to quantify the attention being given to different titles and to understand the local nature of such attention. For this, we turned to Google Trends (GT; www.google.com/trends).[4]

For each title, we measure the relative Google search volume for the search term equal to the original content title in each week from 2011 to 2014 (inclusive) over the whole world, in each European country, and in each country subdivision (states in Germany, cantons in Switzerland, autonom communities in Spain, etc.). In each such region, after normalizing against the volume of our baseline query, the measurement can be interpreted as the fraction of Google searches for the title in a given week out of all searches in the region, measured on an arbitrary but (approximately) common scale between regions.

In Figure 6, we compare this search engine attention to sales figures in Germany for two unnamed films.[5] Comparing panels (a) and (b), we first notice that the overall scale of sales correlates with the overall scale of *local* search engine attention at the time of theatrical release, whereas the global search engine attention is less meaningful (note vertical axis scales, which are common between the two figures). Looking closer at differences between regions in panel (b), we see that, while showing in cinemas, unnamed film 2 garnered more search engine attention in North Rhine-Westphalia

Figure 5. Scatter Plots of Various Data from IMDb and RT (Horizontal Axes) Against Total European Sales During the First Week of an HE Release (Vertical Axes, Rescaled to Anonymize) and Corresponding Coefficients of Correlation (ρ)

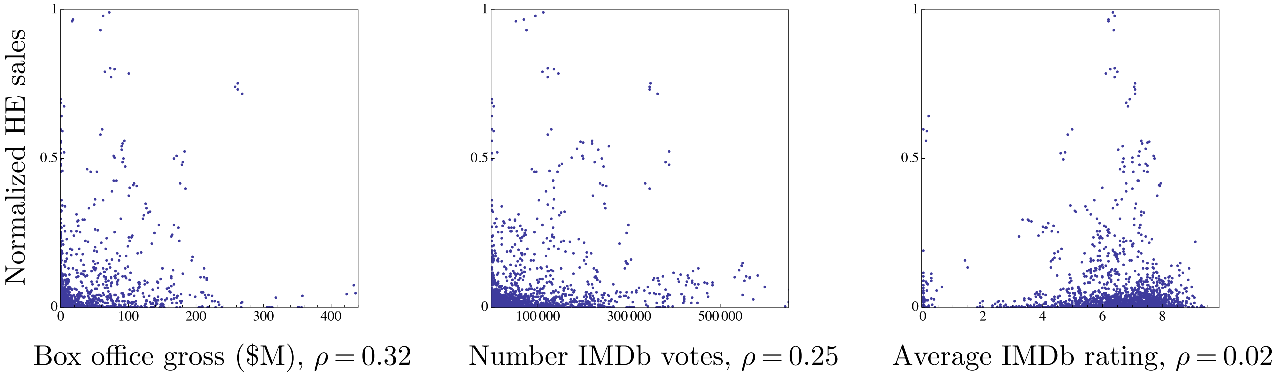
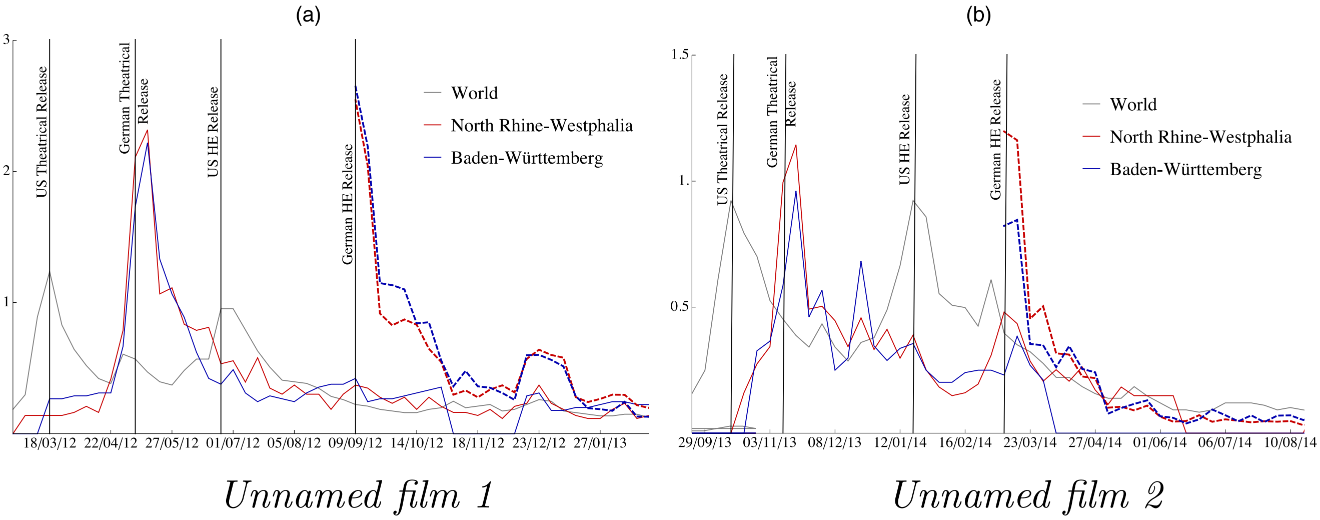


Figure 6. Weekly Search Engine Attention for Two Unnamed Films in the World and in Two Populous German States (Solid Lines) and Weekly HE Sales for the Same Films in the Same States (Dashed Lines)



Note. The scales are arbitrary but common between regions and the two plots.

(NW) than in Baden-Württemberg (BW) and, correspondingly, HE sales in NW in the first weeks after an HE release were greater than in BW. In panel (a), unnamed film 1 garnered similar search engine attention in both NW and BW and similar HE sales as well. In panel (b), we see that the search engine attention to unnamed film 2 in NW accelerated in advance of the HE release, which was particularly successful in NW. In panel (a), we see that a slight bump in search engine attention 3 months into HE sales corresponded to a slight bump in sales. These observations suggest that local search engine attention both at the time of local theatrical release and in recent weeks may be indicative of future sales volumes.

6.4. Constructing Auxiliary Data Features and a Random Forest Prediction

For each instance (t, r) of problem (23) and for each item i we construct a vector of numeric predictive features x_{tri} that consist of backward cumulative sums of the sale volume of the item i at location r over the past three weeks (as available; e.g., none for new releases), backward cumulative sums of the total sale volume at location r over the past three weeks, the overall mean sale volume at location r over the past 1 year, the number of weeks since the original release date of the content (e.g., for a new release this is the length of time between the premier in theaters to release on DVD), an indicator vector for the country of the location r, an indicator vector for the identity of chain to which the location r belongs, the total search engine attention to the title i over the first two weeks of local theatrical release globally, in the country, and in the country-subdivision of the location r, backward cumulative sums of search engine attention to the title i over the past three weeks globally, in the country, and in the country-subdivision of the location r, and features capturing item information harvested from IMDb and RT.

Much of the information harvested from IMDb and RT is unstructured in that it is not numeric features, such as plot summaries, MPAA ratings, and actor listings. To capture this information as numerical features that can be used in our framework, we use a range of clustering and community-detection techniques, which we fully describe in supplemental Section EC.7.

We end up with $d_x = 91$ numeric predictive features. Having summarized these numerically, we train a RF of 500 trees to predict sales. In training the RF, we normalize the sales in each instance by the training-set average sales in the corresponding location; we denormalize after predicting. To capture the decay in demand from time of release in stores, we train a separate RFs for sale volume on the kth week on the shelf for $k = 1, \ldots, 35$ and another RF for the "steady state" weekly sale volume after 35 weeks.

Figure 7. Top-25 x Variables in Predictive Importance

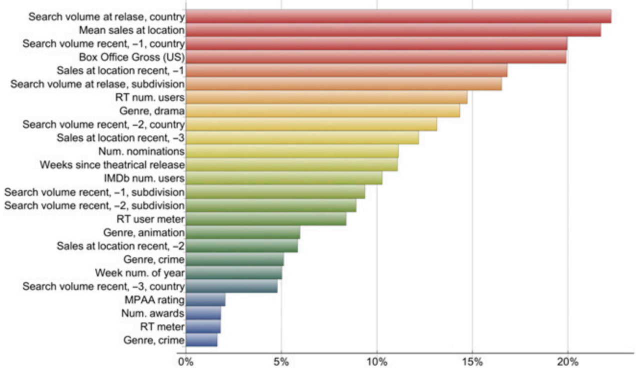
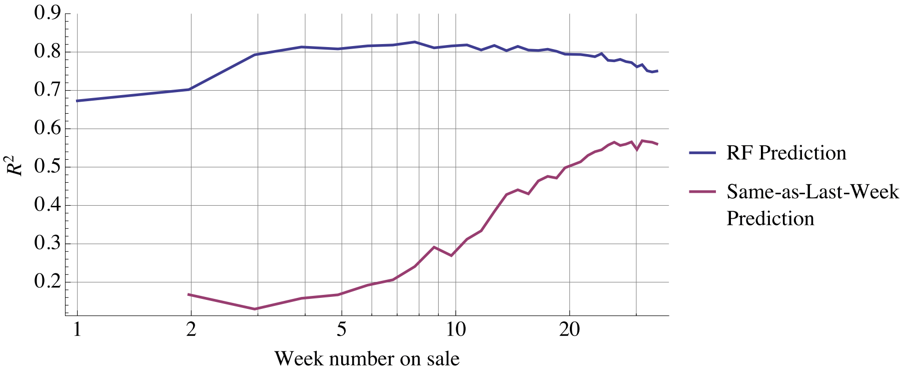


Figure 8. Out-of-Sample Coefficients of Determination R^2 for Predicting Demand Next Week



For $k = 1$, we are predicting the demand for a new release, the uncertainty of which, as discussed in Section 6.1, constitutes one of the greatest difficulties of the problem to the company. In terms of predictive quality, when measuring out-of-sample performance we obtain an $R^2 = 0.67$ for predicting sale volume for new releases. Figure 7 shows the top feature in predictive importance, measured as the average over trees of the change in mean-squared error as percentage of total variance when the value of the variables is randomly permuted among the out-of-bag training data. In Figure 8, we show the R^2 obtained also for predictions at later times in the product life cycle, compared with the performance of a baseline heuristic that always predicts this week's demand for next.

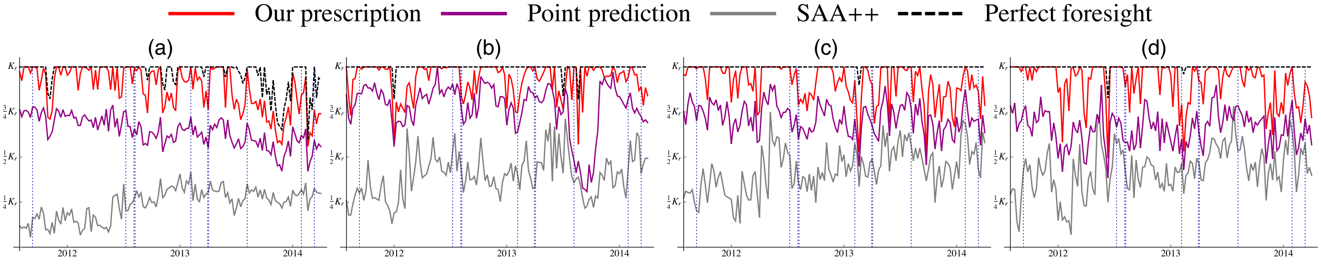
6.5. Applying Our Predictive Prescriptions to the Problem

In the last section, we discussed how we construct RFs to predict sales, but our problem of interest is to prescribe order quantities. To solve our problem (23), we use the forests we trained to construct weights $w_{N,i}(x)$ exactly like in (18), then we transform these like in (24), and, finally, we prescribe data-driven order quantities $\hat{z}_N(x)$ like in (8). Thus, we use our data to go from an observation $X = x$ of our varied auxiliary data directly to a replenishment decision on order quantities.

We would like to test how well our prescription does out-of-sample and as an actual live policy. To do this we consider what we would have done over the 150 weeks from December 19, 2011 to November 9, 2014 (inclusive). At each week, we consider only data from time prior to that week, train our RFs on this data, and apply our prescription to the current week. Then we observe what had actually materialized and score our performance.

There is one issue with this approach to scoring: our historical data only consists of sales, not demand. While we corrected for the adverse effect of demand censorship on our prescriptions using the transformation (24), we are still left with censored demand when scoring performance as described above. To have a reasonable measure of how good our method is, we therefore consider the problem (23) with capacities K_r that are a *quarter* of their nominal values. In this way, demand censorship hardly ever becomes an issue in the scoring of performance. To be clear, this correction is necessary just for a counterfactual scoring of performance; not in practice for applying the predictive prescription. The transformation (24) already corrects for prescriptions trained on

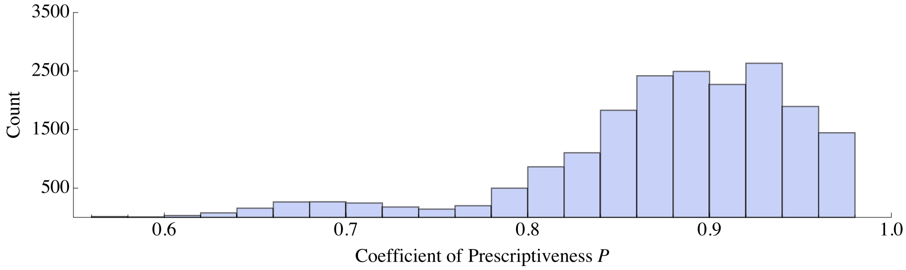
Figure 9. Performance of Our Prescription over Time



Munich, $P = 0.89$ Paris, $P = 0.90$ Waterloo, $P = 0.85$ The Hague, $P = 0.86$

Notes. Vertical dashes indicate major release dates. The vertical axis is shown in terms of the location's capacity, K_r.

Figure 10. Distribution of Coefficients of Prescriptiveness P over Retail Locations



censored observations of the quantity Y that affects true costs.

We compare the performance of our method with three other quantities. One is the performance of the perfect-forecast policy, which knows future demand exactly (no distributions). Another is the performance of a data-driven policy without access to the auxiliary data (i.e., SAA). Because the decay of demand over the lifetime of a product is significant, to make it a fair comparison we let this policy depend on the distributions of product demand based on how long it has been on the market. That is, it is based on T separate data sets where each consists of the demands for a product after t weeks on the market (again, considering only past data). Because of this handicap, we term it SAA++ henceforth. The last benchmark is the performance of a point-prediction-driven policy using the RF sale prediction. Because there are a multitude of optimal solutions z_j to (23) if we were to let Y_j be deterministic and fixed as our prediction $\hat{m}_{N,j}(x)$, we have to choose a particular one for the point-prediction-driven decision. The one we choose sets order levels to match demand and scale to satisfy the capacity constraint: $\hat{z}^{\text{point-pred}}_{N,j}(x) = K_r \max\{0, \hat{m}_{N,j}(x)\} / \sum_{j'=1}^{d} \max\{0, \hat{m}_{N,j'}(x)\}$.

The ratio of the difference between our performance and that of the prescient policy and the difference between the performance of SAA++ and that of the prescient policy is the coefficient of prescriptiveness P. When measured out-of-sample over the 150-week period as these policies make live decisions, we obtain $P = 0.88$. Said another way, in terms of our objective (sales volumes), our data X and our prescription $\hat{z}_N(x)$ gets us 88% of the way from the best data-poor decision to the impossible perfect-foresight decision. This is averaged over just under 20,000 locations.

In Figure 9, we plot the performance over time at four specific locations, the city of which is noted. Blue vertical dashes in each plot indicate the release dates of the 10 biggest first-week sellers in each location, which turn out to be the same. Two pairs of these coincide on the same week. The plots show a general ordering of performance with our policy beating the point-prediction-driven policy [but not always as seen in a few days in Figure 9(b)], which, in turn, beats SAA++ [but not always as seen in a few days in Figure 9(d)]. The P of our policy specific to these locations are 0.89, 0.90, 0.85, and 0.86. The corresponding P of the point-prediction-driven policy are 0.56, 0.57, 0.50, 0.40. That the point-prediction-driven policy outperforms SAA++ (even with the handicap) and provides a significant improvement as measured by P can be attributed to the informativeness of the data collected in Section 6.3 about demand. On most major release dates, the point-prediction-driven policy does relatively worse, which can be attributed to the fact that demand for new releases has the greatest amount of (residual) uncertainty, which the point-prediction-driven policy ignores. When we leverage this data in a manner appropriate for inventory management using our approach, we nearly double the improvement. We also see that on most major release dates, our policy seizes the opportunity to match the perfect foresight performance, but on a few it falls short. In Figure 10, we plot the overall distribution of P of our policy over all retail locations in Europe.

7. Concluding Remarks

In this paper, we combined ideas from ML and OR/MS in developing a framework, along with specific methods, for using data to prescribe optimal decisions in OR/MS problems that leverage auxiliary observations. We motivate our methods based on existing predictive methodology from ML, but, in the OR/MS tradition, focus on the making of a decision and on the effect on costs, revenues, and risk. Our approach is generally applicable, tractable, asymptotically optimal, and leads to substantive and measurable improvements in a real-world context.

We believe that the above qualities, together with the growing availability of data and in particular auxiliary data in OR/MS applications, afford our proposed approach a potential for substantial impact in the practice of OR/MS.

Acknowledgments

The authors thank the anonymous reviewers and the associate editor for helpful suggestions that helped improve this paper. They also thank Ross Anderson and David Nachum from Google for assistance in obtaining access to Google Trends data.

Endnotes

[1] Note that the uncertainty of the point prediction in estimating the conditional expectation, gleaned, for example, via the bootstrap, is the wrong uncertainty to take into account, in particular because it shrinks to zero as $N \to \infty$.

[2] A more direct application of tree methods to the prescription problem would have us minimize the cost of taking the best constant decision z_j in each leaf $j = 1, \ldots, r$: $\min_R \min_{z_1,\ldots,z_r \in \mathcal{Z}} \sum_{j=1}^{r} \cdot \sum_{i:\mathcal{R}(x^i)=j} c(z_j; y^i)$. Like in CART, this can be heuristically done by recursively partitioning, at each stage minimizing the sum of costs across the candidate split. However, because we must consider splitting on each variable and at each data point to find the best split (cf. Trevor et al. 2001, p. 307), this can be overly computationally burdensome for all, except for the simplest problems that admit a closed-form solution, such as least sum of squares or the newsvendor problem. Similarly, a forest ensemble of such trees trained on bootstrap samples and random feature subsets can be used in our ensemble proposal.

[3] See https://developers.google.com/maps/documentation/geocoding for details.

[4] While GT is available publicly online, access to massive-scale querying and week-level trends data are not public. See the acknowledgments.

[5] These films must remain unnamed because a simple search can reveal their European distributor and hence the vendor who prefers their identity be kept confidential.

References

Arya S, Mount DM, Netanyahu NS, Silverman R, Wu A (1998) An optimal algorithm for approximate nearest neighbor searching in fixed dimensions. *J. ACM* 45(6):891–923.

Asur S, Huberman B (2010) Predicting the future with social media. *Proc. 2010 IEEE/WIC/ACM Internat. Conf. Web Intelligence Intelligent Agent Tech.*, vol. 1 (IEEE, Washington, DC), 492–499.

Ban G-Y, Rudin C (2018) The big data newsvendor: Practical insights from machine learning. *Oper. Res.* 67(1):90–108.

Bartlett P, Mendelson S (2003) Rademacher and Gaussian complexities: Risk bounds and structural results. *J. Machine Learn. Res.* 3(November):463–482.

Belloni A, Chernozhukov V (2011) $\ell 1$-penalized quantile regression in high-dimensional sparse models. *Ann. Stat.* 39(1):82–130.

Ben-Tal A, El Ghaoui L, Nemirovski A (2009) *Robust Optimization* (Princeton University, Princeton, NJ).

Bentley J (1975) Multidimensional binary search trees used for associative searching. *Commun. ACM* 18(9):509–517.

Berger JO (1985) *Statistical Decision Theory and Bayesian Analysis* (Springer, New York).

Bertsekas DP (1995) *Dynamic Programming and Optimal Control* (Athena Scientific, Belmont, MA).

Bertsimas D, Kallus N (2016) The power and limits of predictive approaches to observational-data-driven optimization. Preprint, submitted May 8, https://arxiv.org/abs/1605.02347.

Bertsimas D, Gupta V, Kallus N (2018a) Data-driven robust optimization. *Math. Programming* 167(2):235–292.

Bertsimas D, Gupta V, Kallus N (2018b) Robust sample average approximation. *Math. Programming* 171(1–2):217–282.

Besbes O, Zeevi A (2009) Dynamic pricing without knowing the demand function: Risk bounds and near-optimal algorithms. *Oper. Res.* 57(6):1407–1420.

Birge JR, Louveaux F (2011) *Introduction to Stochastic Programming* (Springer, New York).

Breiman L (2001) Random forests. *Mach. Learn.* 45(1):5–32.

Breiman L, Friedman J, Stone C, Olshen R (1984) *Classification and Regression Trees* (CRC Press, New York).

Cameron AC, Trivedi PK (2005) *Microeconometrics* (Cambridge University, Cambridge, UK).

Choi H, Varian H (2012) Predicting the present with Google Trends. *Econom. Rec.* 88(s1):2–9.

Cleveland WS, Devlin SJ (1988) Locally weighted regression: An approach to regression analysis by local fitting. *J. Amer. Statist. Assoc.* 83(403):596–610.

Da Z, Engelberg J, Gao P (2011) In search of attention. *J. Finance* 66(5): 1461–1499.

Delage E, Ye Y (2010) Distributionally robust optimization under moment uncertainty with application to data-driven problems. *Oper. Res.* 55(3):98–112.

Devroye LP, Wagner TJ (1980) On the L1 convergence of kernel estimators of regression functions with applications in discrimination. *Probab. Theory Related Fields* 51(1):15–25.

Geer SA (2000) *Empirical Processes in M-estimation* (Cambridge University, Cambridge, UK).

Goel S, Hofman J, Lahaie S, Pennock D, Watts D (2010) Predicting consumer behavior with web search. *Proc. Natl. Acad. Sci. USA* 107(41):17486–17490.

Gruhl D, Guha R, Kumar R, Novak J, Tomkins A (2005) The predictive power of online chatter. *Proc. 11th ACM SIGKDD Internat. Conf. Knowledge Discovery Data Mining* (ACM, New York), 78–87.

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

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

Hansen BE (2008) Uniform convergence rates for kernel estimation with dependent data. *Econometric Theory* 24(3):726–748.

Huh WT, Levi R, Rusmevichientong P, Orlin JB (2011) Adaptive data-driven inventory control with censored demand based on Kaplan-Meier estimator. *Oper. Res.* 59(4):929–941.

Imbens GW, Rubin DB (2015) *Causal Inference in Statistics, Social, and Biomedical Sciences* (Cambridge University, Cambridge, UK).

Kallus N (2014) Predicting crowd behavior with big public data. *Proc. 23rd Internat. Conf. on World Wide Web (WWW) Companion* (ACM, New York), 625–630.

Kao Y-h, Roy BV, Yan X (2009) Directed regression. Bengio Y, Schuurmans D, Lafferty JD, Williams CKI, Culotta A, eds. *Adv. Neural Inform. Proc. Systems 22* (Curran Associates, New York), 889–897.

Kaplan EL, Meier P (1958) Nonparametric estimation from incomplete observations. *J. Amer. Statist. Assoc.* 53(282):457–481.

Kleywegt A, Shapiro A, Homem-de Mello T (2002) The sample average approximation method for stochastic discrete optimization. *SIAM J. Optim.* 12(2):479–502.

Lai TL, Robbins H (1985) Asymptotically efficient adaptive allocation rules. *Adv. Appl. Math.* 6(1):4–22.

Lehmann EL, Casella G (1998) *Theory of Point Estimation* (Springer, New York).

Mohri M, Rostamizadeh A, Talwalkar A (2012) *Foundations of Machine Learning* (MIT, Cambridge, MA).

Nadaraya E (1964) On estimating regression. *Theory Probab. Appl.* 9(1):141–142.

Nemirovski A, Juditsky A, Lan G, Shapiro A (2009) Robust stochastic approximation approach to stochastic programming. *SIAM J. Optim.* 19(4):1574–1609.

Parzen E (1962) On estimation of a probability density function and mode. *Ann. Math. Statist.* 33(3):1065–1076.

Robbins H (1952) Some aspects of the sequential design of experiments. *Bull. Amer. Math. Soc.* 58(5):527–535.

Robbins H, Monro S (1951) A stochastic approximation method. *Ann. Math. Statist.* 22(3):400–407.

Rosenbaum PR, Rubin DB (1983) The central role of the propensity score in observational studies for causal effects. *Biometrika* 70(1): 41–55.

Shapiro A (2003) Monte Carlo sampling methods. Ruszczynski A, Shapiro A, eds. *Handbooks in Operations Research and Management Science*, vol. 10 (Amsterdam, Elsevier), 353–425.

Shapiro A, Nemirovski A (2005) On complexity of stochastic programming problems. Jeyakumar V, Rubinov A, eds. *Continuous Optimization: Current Trends and Modern Applications* (Springer, New York), 111–146.

Tibshirani R (1996) Regression shrinkage and selection via the lasso. *J. Royal Statist. Soc. Ser. B. Methodological* 58(1):267–288.

Trevor H, Robert T, Friedman J (2001) *The Elements of Statistical Learning* (Springer, New York).

Vapnik V (1992) Principles of risk minimization for learning theory. Moody JE, Hanson SJ, Lippmann RP, eds. *Adv. Neural Inform. Proc. Systems 4* (Curran Associates, Red Hook, NY), 831–838.

Wald A (1949) Statistical decision functions. *Ann. Math. Statist.* 20(2): 165–205.

Walk H (2010) Strong laws of large numbers and nonparametric estimation. Korn R, Karasözen B, Kohler M, Devroye L, eds. *Recent Developments in Applied Probability and Statistics* (Springer, New York), 183–214.

Watson G (1964) Smooth regression analysis. *Sankhyā: Indian J. Statist., Ser. A* 26(4):359–372.