Bootstrap Robust Prescriptive Analytics

Dimitris Bertsimas*[1] and Bart Van Parys†[1]

[1] *Operations Research Center, Massachusetts Institute of Technology*

February 23, 2019

Abstract

We address the problem of prescribing an optimal decision in a framework where its cost depends on uncertain problem parameters y that need to be learned from data. Earlier work, e.g., **hannah2010nonparametric**; **bertsimas2014predictive**, proposes prescriptive formulations based on classical machine learning methods using supervised training data $[(\bar{x}_1, \bar{y}_1), \ldots, (\bar{x}_n, \bar{y}_n)]$. These prescriptive methods factor in additional observed contextual information $x = x_0$ on a potentially large number of covariates to take context specific actions which are superior to any static decision. Such naive use of limited training data may, however, lead to gullible decisions which are over-calibrated to one particular data set. The corresponding phenomenon in prediction problems is well known as overfitting. In this paper, we combine ideas from robust optimization and the statistical bootstrap by **efron1982jackknife** to propose a novel prescriptive method based on balloon estimation learning which natively safeguards against overfitting. Our robust prescriptive method reduces to a tractable convex optimization problem. We illustrate our data-driven decision-making framework and our novel robustness notion on several numerical examples.

Keywords: Data Analytics, Distributionally Robust Optimization, Statistical Bootstrap, Nadaraya-Watson Learning, Nearest Neighbors Learning

1 Introduction

In practice decisions need to be taken despite the fact that the cost $L(z, y)$ of any potential choice z also depends on an unknown stochastic parameter y. Stocking goods under uncertain customer demand or devising profitable investment strategies when facing volatile returns would be two practical examples. In a setting where decisions need to be made repeatedly, it is argued by **shapiro2014lectures** to take action based on the stochastic optimization formulation

$$z^\star \in \arg\min_z \mathbb{E}_{Y^\star}\left[L(z, y)\right]. \tag{1}$$

The action ultimately chosen achieves a minimal cost as measured by the expected value of its loss $L(z^\star, y)$ with respect to the parameter y distributed as $Y^\star$. We make the following standing assumption as to ensure that our stochastic optimization problem is a well-posed convex problem.

Assumption 1 (Loss function)**.** *The loss function $L(\bar{z}, \bar{y})$ in $\mathrm{R}_+ \cup \{+\infty\}$ is convex in $\bar{z}$ for any $\bar{y}$ and a measurable function of $\bar{y}$ for any $\bar{z}$.*

We assume here implicitly that the minimization over z is only over $\mathrm{dom}(\mathbb{E}_{Y^\star}[L(z, y)])$. For the sake of simplicity, we also assume that the decision $z \in \mathrm{R}^{\dim(z)}$, and the random variables $y \in \mathrm{R}^{\dim(y)}$ and $x \in \mathrm{R}^{\dim(x)}$ take values in finite dimensional vector spaces. This stochastic optimization formulation makes

*dbertsim@mit.edu

†vanparys@mit.edu

sense only when nothing beyond the distribution of the stochastic variable y is known before a decision is to be made. Often however, additional information concerning a potential large number of covariates x such as weather forecasts, Twitter feeds, Google Trends data, ..., can be obtained before we need to commit to our decision. After observing a particular context $x = x_0$, the decision should take this additional information into account. A portfolio manager may for instance alter strategy when given prior notice about a relevant Twitter storm surrounding one of his assets. If decisions are to be made in a particular observed context, encoded here as $x = x_0$, the problem in need of attention should not be the classical stochastic optimization problem (1) but rather

$$z^\star(x_0) \in \arg\min_z \mathbb{E}_{D^\star}\left[L(z, y)|x = x_0\right] \tag{2}$$

The distributional model $D^\star$ must represent here the joint distribution of both the parameters y as well as the auxiliary covariates x. The action $z^\star(x_0)$ has hence minimal cost as measured by the expected value of its loss $L(z^\star, y)$ conditioned on all observed covariate contextual information. Technically, the conditional expectation $\mathbb{E}_{D^\star}\left[L(z, y)|x = x_0\right]$ is a random variable and ambiguous for events of measure zero. That is, the inclusion in (2) can hold merely almost surely; see for instance **billingsley2008probability**. All statements in this paper involving the observation x_0 should be interpreted to hold hence as $X^\star$-almost surely where $X^\star$ is the distribution of the covariates x. In a special case when $X^\star$ is finitely supported then (2) simply holds for all observations in the support. To avoid trivialities we do assume that the support counts more than one covariate observation.

The prescription problem (2) has been studied extensively since at least **wald1950statistical** in the context of statistical decision theory. As statistical learning is most often concerned with prediction problems, loss functions such as $L(\bar{z}, \bar{y}) = (\bar{z} - \bar{y})^2$ are of particular concern. The textbook solution as found for instance in **friedman2001elements** is in this particular case given as the conditional expectation $z^\star(x_0) = \mathbb{E}_{D^\star}\left[y|x = x_0\right]$. The prescription problem (2) can also be reduced to the stochastic optimization problem (1) by taking as $Y^\star$ instead the conditional distribution, denoted $D^\star(x = x_0)$, of the uncertain problem parameters y in the particular observed covariate context $x = x_0$. As the prescription problem requires stochastic optimization it inherits its shortcomings. **nemirovski2006convex** have shown that stochastic optimization problems tend to be computationally unattractive even despite their convex nature for all but the simplest of problems. Even worse, no distributional model $D^\star$ can reasonable be expected to be observed directly in practice. Indeed, distributions are a product of modeling uncertainty rather than being a directly observable primitive. This make the classical optimization formulation (2) not particularly well suited for modern decision-making.

Only historical data concerning both uncertain parameters and contexts is typically ever directly available in practice. Data instead of distributions should hence be the primitive for decision-making under uncertainty. The primitive object on which to base decisions is in practice often

$$\mathrm{tr} := [\bar{d}_1 = (\bar{x}_1, \bar{y}_1), \bar{d}_2 = (\bar{x}_2, \bar{y}_2), \dots]. \tag{3}$$

consisting of historically observed uncertain outcomes and contexts. **chen2015statistical** discusses specific situations in revenue management in which previous decisions are treated as historical data as well. Unlike the stochastic parameters y and covariates x, the data points in the training data are deterministic historical observations. We make this important distinction explicit with the symbol "$\bar{\cdot}$" Typically only a limited number n of historical observations can be used as a training data set. We will denote with $\mathrm{tr}[n]$ the training data set consisting of the first n observations. As opposed to a time series, the order of the data points in a training set is of no importance. The statistical information of the first n training data samples is thus captured without loss by its empirical distribution $D_{\mathrm{tr}[n]} := \frac{1}{n}\sum_{(\bar{x},\bar{y})\in\mathrm{tr}[n]} \delta_{(\bar{x},\bar{y})}$ and the total number of data points n. This particular notation has the benefit that it alludes to the fact that whereas decisions are classically based on a distributional model $D^\star$, data-driven approaches should favor its empirical counterpart $D_{\mathrm{tr}[n]}$ based on training data. Whereas the distributional model $D^\star$ has support Ω, the training set contains observations of only at most n distinct outcomes. We denote with $\mathcal{D}$ and $\mathcal{D}_n$ the sets of all distributions supported on Ω and Ω_n, respectively. We have hence $D_{\mathrm{tr}[n]}[\Omega_n] = \sum_{(\bar{x},\bar{y})\in\Omega_n} D_{\mathrm{tr}[n]}[\bar{x}, \bar{y}] = D^\star(\Omega) = 1$ where $\Omega_n \subseteq \Omega$ with cardinality $|\Omega_n| \leq n$.

Data-Driven Formulations

Making decisions based on supervised data is a timely topic and has received considerable attention in the literature. Much of the results in the early line of work based on the pioneering results of **wald1950statistical** are primarily focused on the existence of minimax optimal decision functions $z_{\text{tr}[n]}(\cdot)$ with computational tractability being only a secondary concern. We refer to **berger2013statistical** for a recent exposition on statistical decision theory. The need for more computationally oriented strategies was greatly advanced by the work on empirical risk minimization by **vapnik2013nature**.

Empirical-Risk-Minimization: The prescription problem (2) can alternatively be defined through the risk minimization formulation $z^\star(\cdot) \in \arg\min_{z(\cdot)\in\mathcal{F}} \mathbb{E}_{D^\star}\left[L(z(x),y)\right]$ over the set of all functions $\mathcal{F}$ mapping covariates to decisions. One could hence consider a data-driven prescriptor mapping an observed context to a decision as the solution to

$$z_{\text{tr}[n]}(\cdot) \in \arg\min_{z(\cdot)\in\mathcal{C}} \; \frac{1}{n}\sum_{(\bar{x},\bar{y})\in\text{tr}[n]} L(z(\bar{x}),\bar{y}) \tag{4}$$

where $\mathcal{C}$ is a set of functions with suitable properties. Here $\mathcal{C}$ must be a strict subset of the set of all functions $\mathcal{F}$ as to avoid overfitting and encourage generalization. **rudin2014big** consider in particular the class $\mathcal{C} = \mathcal{F}_{\text{lin}}$ consisting of all linear functions. As an added benefit, the empirical risk minimization problem (4) then reduces to a convex problem in the coefficients describing the linear functions in $\mathcal{F}_{\text{lin}}$. Empirical-risk-minimization formulations enjoy a finite sample generalization performance which depends explicitly on the complexity of the class $\mathcal{C}$; see for instance **rudin2014big**; **bertsimas2014predictive**. By explicitly constraining the prescriptor to be an element of $\mathcal{F}_{\text{lin}}$, bias is introduced when $z^\star(\cdot) \notin \mathcal{F}_{\text{lin}}$. There is indeed no reason in general to expect that $z^\star(\cdot)$ should possess a linear structure. As pointed out by **bertsimas2014predictive** constraints on the decision z may take gives rise to nonlinearities which are particularly challenging to deal with. **rudin2014big** do consider nonlinear prescriptions $z_{\text{tr}[n]}(\cdot)$ implicitly via the introduction of auxiliary nonlinear transformed covariates. This however comes at an increased computational burden and ultimately limits the applicability of this formulation as a general data-driven decision tool. In this paper we will focus on formulations of an entirely different nature instead.

Estimate-Then-Optimize: An alternative way to make decisions based on data is to first construct an estimate $D^n_{\text{tr}[n]}(x = x_0)$ based on the given training data set of the conditional distribution $D^\star(x = x_0)$ of the uncertain problem parameter y in context of interest $x = x_0$. Such estimates can be obtained using a variety of conditional density estimation methods. **hannah2010nonparametric** consider a classical learning method proposed independently by **watson1964smooth**; **nadaraya1964estimating**. **bertsimas2014predictive** consider additional formulations based on nearest-neighbors learning discussed in **altman1992introduction**, and regression trees and random forest learning proposed by **breiman2001random**. Denote with $\mathrm{E}^n_{D_{\text{tr}[n]}}\left[L(\bar{z},y)|x=x_0\right] = \int L(\bar{z},y)\,\mathrm{d}D^n_{\text{tr}[n]}(x=x_0)$ the expected loss of a given decision $\bar{z}$ consistent with the obtained conditional estimate of the uncertain problem parameter y in context of interest $x = x_0$ based on the given training data set. Our notation here alludes to the fact that $\mathrm{E}^n_{D_{\text{tr}[n]}}\left[L(\bar{z},y)|x=x_0\right]$ is an asymptotically unbiased estimate of the unknown cost $\mathbb{E}_{D^\star}\left[L(\bar{z},y)|x=x_0\right]$ based on supervised training data. Estimate-then-optimize formulations prescribe an action which minimizes the estimated cost, i.e.,

$$z_{\text{tr}[n]}(x_0) \in \arg\min_{z} \; \mathrm{E}^n_{D_{\text{tr}[n]}}\left[L(z,y)|x=x_0\right]. \tag{5}$$

We will consider in this paper estimate-then-optimize formulations based on either Nadaraya-Watson and nearest-neighbors learning. In Section 2 we in fact indicate that both can be generalized using the balloon estimator discussed by **sain2002multivariate** and stated in Definition 3. **bertsimas2014predictive** prove that both the Nadaraya-Watson and nearest-neighbors formulations are asymptotically consistent under mild assumptions on the training data and loss function. This is a clear advantage over the empirical-risk-minimization formulation of **rudin2014big** which is biased whenever $z^\star(\cdot) \notin \mathcal{C}$. Unlike empirical-risk-minimization formulations in which we can control the size of $\mathcal{C}$, estimate-then-optimize formulations do not have any natural defense mechanism against overfitting. It is indeed well known that unbiased estimators based on Nadaraya-Watson and nearest neighbors learning typically suffer a large variance. Subsequent minimization of the unbaised estimate $\mathrm{E}^n_{D_{\text{tr}[n]}}\left[L(z,y)|x=x_0\right]$ only amplifies this issue and can results in

overly optimistic prescriptors as pointed out by **michaud1989markowitz**. The performance of estimate-then-optimize formulations based on one particular training data set on out-of-sample data can consequently be very poor.

Budget-Minimization: It is clear that when given only a limited amount of training data, data-driven formulations must be guarded against such overcalibration to one specific training data set. Overfitting can be discouraged by minimizing a budget function

$$z_{\mathrm{tr}[n]}(x_0) \in \arg\min_z \left\{ c_n(z, D_{\mathrm{tr}[n]}, x_0) = \mathrm{E}^n_{D_{\mathrm{tr}[n]}} \left[L(z,y) | x = x_0 \right] + J_{\mathrm{tr}[n]}(z, x_0) \right\} \tag{6}$$

consisting of both an asymptotically unbiased cost estimate and an additional non-negative regularization term. Regularization has played a predominant role in statistics since at least the work of **tikhonov1963solution** on ill-posed linear systems. While regularization has a detrimental effect on the predicted performance on the training data as compared to its nominal counterpart (5), its role is to encourage decisions who perform well on out-of-sample data too. We discuss in Section 3 how to regularize the Nadaraya-Watson formulation of **hannah2010nonparametric** and the nearest-neighbors formulations of **bertsimas2014predictive** in order to encourage out-of-sample performance while preserving computational tractability.

Out-of-Sample Performance Metrics?

We discuss here first the limitations of the most popular out-of-sample performance metrics found in the literature. Out-of-sample performance metrics are typically defined in the context of the training data $\mathrm{tr}[n]$ as one particular realization of the first n points of the random process

$$\mathrm{data} := [d_1 = (x_1, y_1) \sim_{\mathrm{iid}} D^\star,\ d_2 = (x_2, y_2) \sim_{\mathrm{iid}} D^\star,\ \ldots] \sim D^{\star\infty}. \tag{7}$$

Each data point is here taken as an independent sample from a common distribution which is completely unknown to the decision maker. The distribution $D^{\star\infty}$ of the data generation process hence consists of a sequence of independent copies of the distributional model $D^\star$. Perhaps the most straightforward measure of out-of-sample performance is captured by the difference

$$\mathrm{R}_n(z_{\mathrm{data}[n]}, D^\star, x_0) := \mathbb{E}_{D^{\star\infty}} \left[\mathbb{E}_{D^\star} \left[L(z_{\mathrm{data}[n]}(x_0), y) | x = x_0 \right] - \min_z \mathbb{E}_{D^\star} \left[L(z,y) | x = x_0 \right] \right]$$

between the cost of our decision based on training data and the best decision made with the benefit on hindsight averaged over the random training data$[n]$. The term $\mathrm{R}_n(z_{\mathrm{data}[n]}, D^\star, x_0)$ hence quantifies the expected regret which will be experienced when committing to $z_{\mathrm{data}[n]}$ in comparison to an optimizer having access to all information. Online optimization, see for instance **hazan2016introduction**, directly attempts to minimize the asymptotic speed with which $\mathrm{R}_n(z_{\mathrm{data}[n]}, D^\star, x_0)$ goes to zero as the number of samples n increases. As the distribution generating the data is unknown, we need a uniformly small regret in the generating distribution $D^\star$ to claim good out-of-sample performance in practice. However, this is not possible without prior knowledge on the distribution $D^\star$. This negative result holds even if the covariate distribution $X^\star$ is finitely supported. Fix indeed a particular covariate context $x = \bar{x}_0$. Notice that the event $\mathcal{E}$ in which the context $x = \bar{x}_0$ is not observed at all in the training data, i.e., $\mathcal{E} := \{\forall (\bar{x}, \bar{y}) \in \mathrm{data}[n] : \bar{x} \neq \bar{x}_0\}$, has always a positive probability $D^{\star\infty}[\mathcal{E}] = (1 - D^\star[\{x = \bar{x}_0\}])^n$. The regret is hence at least $\mathrm{R}_n(z_{\mathrm{data}[n]}, D^\star, \bar{x}_0) \geq \mathbb{E}_{D^{\star\infty}} \left[\mathbb{E}_{D^\star} \left[L(z_{\mathrm{data}[n]}(x_0), y) | x = \bar{x}_0 \right] - \mathbb{E}_{D^\star} \left[L(z^\star(x_0), y) | x = \bar{x}_0 \right] | \mathcal{E} \right] (1 - D^\star[\{x = \bar{x}_0\}])^n$ the regret experienced in this particularly adverse event. As the loss function L is convex in the decision z we have by Jensen's inequality $\mathrm{R}_n(z_{\mathrm{data}[n]}, D^\star, \bar{x}_0) \geq (\mathbb{E}_{D^\star} \left[L(\bar{z}, y) | x = \bar{x}_0 \right] - \mathbb{E}_{D^\star} \left[L(z^\star(\bar{x}_0), y) | x = \bar{x}_0 \right]) (1 - D^\star[\{x = \bar{x}_0\}])^n$ for the average decision $\bar{z} := \mathbb{E}_{D^{\star\infty}} \left[z_{\mathrm{data}[n]}(\bar{x}_0) | \mathcal{E} \right]$. Let us define an ambiguity set $\mathcal{A} = \{D : D[B \cap \{x \neq \bar{x}_0\}] = D^\star[B \cap \{x \neq \bar{x}_0\}]\ \forall B\}$ as the set consisting of all distributions identical to the distributional model $D^\star$ with the exception of events in the context of interest $x = \bar{x}_0$. By construction we have then that $\bar{z} = \mathbb{E}_{D^\infty} \left[z_{\mathrm{data}[n]}(\bar{x}_0) | \mathcal{E} \right]$ for all D in $\mathcal{A}$ as in the event $\mathcal{E}$ the prescriptor $z_{\mathrm{data}}(\bar{x}_0)$ is a function of data precisely outside the context of interest $x = x_0$. We trivially have that the conditional distribution $D(x = \bar{x}_0)$ of distributions in $\mathcal{A}$ is left arbitrary, i.e., $\{D(x = \bar{x}_0) : D \in \mathcal{A}\}$ is the set of all distributions. Thus,

$$\sup_{D \in \mathcal{A}} \mathrm{R}_n(z_{\mathrm{data}[n]}, D, \bar{x}_0) \geq \max_Y \left(\mathbb{E}_Y \left[L(\bar{z}, y) \right] - \min_z \mathbb{E}_Y \left[L(z,y) \right] \right) (1 - D^\star[\{x = \bar{x}_0\}])^n.$$

The worst-case regret over $\mathcal{D} \supseteq \mathcal{A}$ is hence nonzero unless the loss function is independent of the decision. Given a finite number n of training data points, obtaining a small regret uniform in $D^\star$ seems to be too much to ask for. We can only get uniform small regret at the expense of demanding exotic conditions on $D^\star$. We could for instance constrain $D^\star[\{x = \bar{x}_0\}]$ to be bounded from below.

An alternative out-of-sample performance metric in the context of budget minimization formulations is the disappointment of our decision $z_{\text{data}[n]}(x_0)$ defined as the random difference between its actual cost and its predicted budget

$$\mathrm{D}_n(z_{\text{data}[n]}, D^\star, x_0) = \mathbb{E}_{D^\star}\left[L(z_{\text{data}[n]}(x_0), y)|x = x_0\right] - c_n(z_{\text{data}[n]}(x_0), D_{\text{data}[n]}, x_0).$$

By keeping the probability $D^{\star\infty}\left[\{\mathrm{D}_n(z_{\text{data}[n]}, D^\star, x_0) > 0\}\right]$ uniformly small in $D^\star$, we are guaranteed that the cost of the decision we have committed to will not often break our predicted budget. Unfortunately also this requirement is too demanding even if $D^\star$ is finitely supported. Fix again a particular covariate context $x = \bar{x}_0$. **vanparys2017data** prove in this context that in order to have a disappointment rate $\lim_{n\to\infty} \frac{1}{n} \log D^{\star\infty}\big(\mathrm{D}_n(z_{\text{data}[n]}, D^\star) > 0\big) \leq -r$ uniformly in $D^\star$, the projected cost or budget used to make decisions needs to be for all z at least as large as

$$c_n(z, D_{\text{data}[n]}, \bar{x}_0) \geq \max\left\{\mathbb{E}_Y\left[L(z, y)\right] \;:\; Y \text{ s.t. } B(D_{\text{data}[n]}(x = \bar{x}_0), Y) \leq r/D_{\text{data}[n]}[x = \bar{x}_0]\right\}$$

where B is the relative entropy distance between distributions proposed first by **kullback1951information**. When in particular the context of interest is not observed in the data set, i.e., the event $\mathcal{E}$, we must resort to a budget as the worst-case cost $c_n(z, D_{\text{data}[n]}, \bar{x}_0) = \max_y L(z, y)$ which is clearly undesirable and may not even be well defined if L is not bounded. Notice that this adverse event happens with probability $D^{\star\infty}[\mathcal{E}] = (1 - D^\star[\{x = \bar{x}_0\}])^n$. As was the case for regret, obtaining a small disappointment uniform in $D^\star$ seems to be too much to ask for. We can only get uniform small disappointment at the expense of demanding exotic conditions on $D^\star$. Constraining $D^\star[\{x = \bar{x}_0\}]$ to be bounded from below is again a possibility. As a contribution we propose here to use a more practical notion of out-of-sample performance instead.

Contributions

A standard practice in machine learning to quantify the out-of-sample performance of a data-driven method is instead to compare its training performance with its performance on validation data. If its performance on validation data is comparable to its training performance, a data-driven method is then expected to generalize well to out-of-sample test data as well. Let us now consider the budget minimization formulation (6) with budget function $c_n(z, D_{\text{tr}}, x_0) = \mathrm{E}^n_{D_{\text{tr}[n]}}\left[L(z, y)|x = x_0\right] + J_{\text{tr}[n]}(z, x_0)$ as our best attempt to regularize the estimate-then-optimize formulation (5). Suppose we obtain access to further validation sets $\text{vd}[n] \in V$ each containing the same number as data points as the training data set. As the true cost $\mathbb{E}_{D^\star}\left[L(z_{\text{tr}[n]}(x_0), y)|x = x_0\right]$ of our decision can ultimately not be known, any practical measure of out-of-sample disappointment must rather measure disappointment relative to a best unbiased estimate $\mathrm{E}^n_{D_{\text{vd}[n]}}\left[L(z_{\text{tr}[n]}(x_0), y)|x = x_0\right]$ instead. A practical measure of out-of-sample disappointment would hence be the fraction

$$\frac{1}{|V|}\sum\nolimits_{\text{vd}[n]\in V} \mathbb{1}\{\mathrm{E}^n_{D_{\text{vd}[n]}}\left[L(z_{\text{tr}[n]}(x_0), y)|x = x_0\right] \geq c_n(z_{\text{tr}[n]}(x_0), D_{\text{tr}[n]}, x_0)\} \tag{8}$$

of all validation sets based on which the new estimated cost of the decision we have committed to breaks the training budget. Instead of attempting to obtain a theoretical guaranteed out-of-sample performance, which we argued is impossible, this validation metric provides a practical alternative. Notice however that our validation metric critically really on our ability to obtain large quantities of validation data. However, it is clear that obtaining large quantities of additional validation data in practice is not a viable approach. We want to be able to quantify the statistical accuracy of the cost estimate $c_n(z_{\text{tr}[n]}(x_0), D_{\text{tr}[n]}, x_0)$ of our decision without access to independent validation data. In statistics this problem is commonly addressed using resampling methods such as the statistical bootstrap of **efron1982jackknife**.

Bootstrapping is the process of resampling with replacement a validation data set from the original training data set. It can also be described formally as the stochastic process

$$\text{bs} := [(x_{1,r}, y_{1,r}) \sim_{\text{iid}} D_{\text{tr}[n]}, M_{2,r} = (x_{2,r}, y_{2,r}) \sim_{\text{iid}} D_{\text{tr}[n]}, \dots] \sim D^{\infty}_{\text{tr}[n]} \tag{9}$$

of resampling n independent data points from the empirical distribution of the training data. The bootstrap method can hence generate its own synthetic validation data sets $\text{bs}[n]$ directly from the training data set. Bootstrap data sets so obtained are synthetic as actual validation data $\text{vd}[n]$ ought to be drawn from the same data generation process as the training data defined in (7). Notice indeed that the bootstrap data and the training data share the same exact data points modulo their frequency. That is, the empirical distributions of the bootstrap data $D_{\text{bs}[n]}$ is always supported on a subset of the training data points Ω_n observed in the training data set. In the context of budget minimization formulations, we can sample bootstrap data sets $\text{bs} \in B$ and use the fraction

$$\frac{1}{|B|} \sum_{\text{bs}[n] \in B} \mathbb{1}\{\mathrm{E}^n_{D_{\text{bs}[n]}} \left[L(z_{\text{tr}[n]}(x_0), y) | x = x_0 \right] \geq c_n(z_{\text{tr}[n]}(x_0), D_{\text{tr}[n]}, x_0)\} \tag{10}$$

as a proxy to the out-of-sample disappointment (8) on actual validation data. As bootstrapping can be done at a low computational cost and does not require any validation data it is a practically viable approach to quantify the out-of-sample performance of budget minimization formulations. In what follows we will in fact consider the exact bootstrap out-of-sample disappointment which lets the number of bootstrap resamples $|B|$ tend to infinity and can formally be defined as

$$D^{\infty}_{\text{tr}[n]}[\mathrm{E}^n_{D_{\text{bs}[n]}} \left[L(z_{\text{tr}[n]}(x_0), y) | x = x_0 \right] \geq c_n(z_{\text{tr}[n]}(x_0), D_{\text{tr}[n]}, x_0)\}] \tag{11}$$

as our metric of out-of-sample performance. The main technical innovations presented in this work to enable bootstrap robust prescriptions are discussed in the remainder of this section.

We present in this paper the first budget minimization formulations who suffer small bootstrap disappointment by design. To do so we introduce our novel notion of the bootstrap robust counterpart of an estimate-then-predict formulation.

Definition 1 (Bootstrap Robust Counterpart)**.** *The budget function c_n and its associated prescriptor $z_{\text{tr}[n]}(x_0) \in \arg\min_z c_n(z, D_{\text{tr}}, x_0)$ are said to be the bootstrap robust counterpart with disappointment $b \in [0, 1)$ of an estimate-then-optimize formulation with estimator $\mathrm{E}^n_{D_{\text{tr}[n]}} \left[L(z_{\text{tr}[n]}(x_0), y) | x = x_0 \right]$ if we have*

$$D^{\infty}_{\text{tr}[n]} \left[\mathrm{E}^n_{D_{\text{bs}[n]}} \left[L(z_{\text{tr}[n]}(x_0), y) | x = x_0 \right] > c_n(z_{\text{tr}[n]}(x_0), D_{\text{tr}[n]}, x_0) \right] \leq b. \tag{12}$$

Budget minimization formulations which are bootstrap robust counterparts do not possess any theoretical out-of-sample performance as classically defined either through regret or disappointment. We have indeed argued that no data-driven method can provide such guarantees without resorting to stringent claims on how the data was generated. Instead, our robustness is defined directly through low disappointment on bootstrap data serving as our best attempt as to make synthetic validation data and hence by proxy hopefully on actual out-of-sample validation data as well. Notice however that the bootstrap out-of-sample metric and counterpart as described before are entirely descriptive and do not suggest how to ensure a budget minimization approach does in fact enjoy such performance on synthetic validation data.

We will make both the classical Nadaraya-Watson and nearest neighbors formulations bootstrap robust by formulating their distributionally robust counterpart. In fact we will do so by treating either as a special case of a more general balloon estimation formulation. We prove that the resulting budget minimization formulations are computationally as tractable as their nominal counterparts. When for instance the estimate-then-optimization formulation reduces to a tractable convex optimization problem then so will its robust counterpart. The previous crucial observation makes our robust budget minimization formulations practically viable. One particular distributionally robust counterpart based on the relative entropy distance is proven to safeguard against bootstrap overfitting as stated in Definition 1. We derive practical finite sample bootstrap performance guarantees as in (12) regarding the resulting robust supervised learning formulation. For this

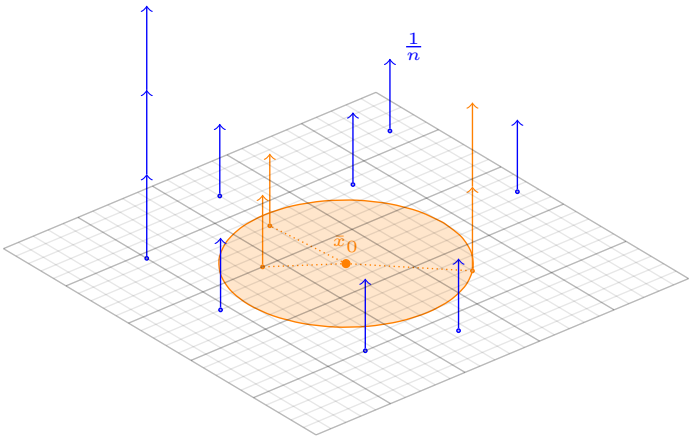


Figure 1: The three and four nearest neighbors (in orange) of the context of interest x_0. We depict the neighborhood set $N_n^3(x_0)$ as the orange circles in the support set Ω_n. This neighborhood contains both the three and four nearest neighbors around x_0 as the most distant nearest neighbor was seen twice in the training data. The orange circle visualizes the metric d implicit in the concept of nearest neighbors learning.

particular bootstrap robust counterpart we derive a more explicit tractable reformulation based on convex duality.

Finally, we present the efficacy of our three proposed data-driven formulations on a small news vendor problem as well as a small portfolio allocation problem. We published a `Julia` implementation of the ideas and examples in this work at https://gitlab.com/vanparys/BootstrapRobustAnalytics.jl.

2 Estimate-Then-Optimize Prescriptions

The estimate-then-optimize formulations which we defined in (5) are distinct only in so far as they are based on a different estimate $\mathrm{E}^n_{D_{\mathrm{tr}[n]}}[L(\bar{z},y)|x=x_0]$ of the actual cost of actions $\bar{z}$ based on supervised training data. A large variety of methods in machine learning can provides such estimates. We will focus here solely on the Nadaraya-Watson formulation of **hannah2010nonparametric** and the nearest-neighbors formulation introduced by **bertsimas2014predictive**. In fact, we will treat here both formulations simultaneously by considering a generalization based on the balloon estimator discussed in **sain2002multivariate**. As the balloon estimator is a local learning method it is based on the concept of neighborhoods depicted visually in Figure 1.

Definition 2 (Distance & Neighborhood)**.** *Let us consider a function which assigns a positive distance* $\mathrm{dist}(\bar{d},x_0)$ *for all* $\bar{d}\in\Omega$*. Assume the distance function enjoys the discrimination property* $\mathrm{dist}(\bar{d},\bar{x}) = \mathrm{dist}(\bar{d}',\bar{x}) \iff \bar{d}=\bar{d}'$ *for all* $\bar{d}$, $\bar{d}'$ *in* Ω*. We divide the training data in increasingly larger nested sets*

$$N_n^j(x_0) := \left\{\bar{d}\in \mathrm{tr}[n] \;:\; \mathrm{dist}(\bar{d},x_0)\leq R_n^j\right\}, \text{ with}$$
$$R_n^j := \min\left\{R\geq 0 \;:\; \left|\left\{\bar{d}\in \mathrm{tr}[n] \;:\; \mathrm{dist}(\bar{d},x_0)\leq R\right\}\right|\geq j\right\}$$

each containing those j *distinct points in the training data closest to* x_0*.*

The particular distance $\mathrm{dist}(\bar{d},x_0)$ should ideally reflect how relevant an observation $\bar{d}$ is to our context of interest $x=x_0$. A common choice is to define the distance as a monotonically decreasing function of the Euclidean distance $\|\bar{x}-x_0\|_2$ between the covariates. Notice that this distance function does not possess the discrimination property. Indeed, when $\|\bar{x}_i-x_0\|_2 = \|\bar{x}_j-x_0\|$ for $i\neq j$ we have a tie. The lack of discrimination property translates in ambiguously defined neighborhood sets N_n^j. The discrimination property may be recovered by deterministically breaking ties based for instance on the value of $\bar{y}$. **gyorfi2006distribution** propose a random alternative by augmenting the covariates with an independent auxiliary uniformly distributed random variable on $[0,1]$. They prove that by doing so ties occur with probability zero. Hence,

Name	Smoother S
Uniform	$\frac{1}{2}\mathbb{1}_{\|\Delta x\|\leq 1}$
Epanechnikov	$\frac{3}{4}(1-\|\Delta x\|^2)\mathbb{1}_{\|\Delta x\|\leq 1}$
Tricubic	$\frac{70}{81}(1-\|\Delta x\|^3)^3\mathbb{1}_{\|\Delta x\|\leq 1}$
Gaussian	$\exp(-\|\Delta x\|^2/2)/\sqrt{2\pi}$

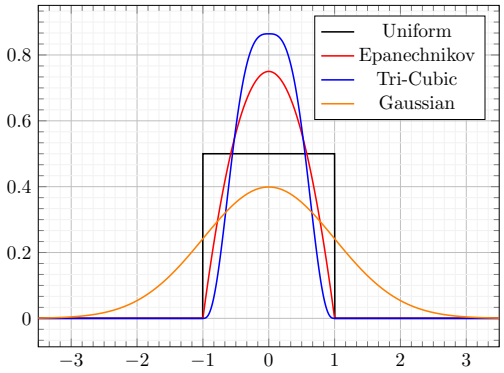


Figure 2: A comparison of popular common smoother functions S. The tricubic smoother has compact support and has two continuous derivatives at the boundary of its support, while the Epanechnikov smoother has none. The Gaussian smoother is continuously differentiable, but has infinite support.

with random tie breaking the neighborhood sets satisfy $\left|N_n^j(x_0)\right| = j$ for all $j \in [n]$ with probability one. As far as the theoretical results in this paper are concerned, we are quite flexible with regards to the particular distance function considered. In fact, we do not even need the distance function to be a metric distance. The discrimination property of the distance function allows us to order distinct data points on proximity to the covariate context of interest $x = x_0$ in a unique way.

Definition 3 (Balloon estimation formulation)**.** *The balloon estimation formulation minimizes the weighted average*

$$z_{\mathrm{tr}[n]}(x_0) \in \arg\min_z \mathrm{E}^n_{D_{\mathrm{tr}[n]}}\left[L(z,y)|x=x_0\right] := \frac{\mathbb{E}_{D_{\mathrm{tr}[n]}}\left[L(z,y)\cdot w_n(x,x_0)\cdot \mathbb{1}\{(x,y)\in N_n^k(x_0)\}\right]}{\mathbb{E}_{D_{\mathrm{tr}[n]}}\left[w_n(x,x_0)\cdot \mathbb{1}\{(x,y)\in N_n^k(x_0)\}\right]} \tag{13}$$

using a positive weighing function w_n over the smallest data neighborhood around the context $x = x_0$ containing no less than k out of n observations.

A balloon estimator hence consider only data within the smallest neighborhood around its context of interest $x = x_0$ containing no less than k out of n observations and is completely blind to other data outside. Within the neighborhood $N_n^k(x_0)$ each data point $(\bar{x}, \bar{y})$ is weighted as $w_n(\bar{x}, \bar{y}) \geq 0$ with the help of a weighing function. It can easily be verified that the balloon estimator is insensitive to the order of the data points in the training data set. Hence, the balloon estimator is indeed only a function the training data through its empirical distribution and the total number of samples n.

2.1 Statistical Consistency

Our balloon estimation formulation has several hyper-parameters which have to be chosen with care; the distance function, the number of neighbors, and the weighing function. The statistical properties of the balloon estimator will depend on how these hyper-paramteres are chosen. We do not try to establish here the statistical consistency of the balloon formulation in its greatest generality. Rather we briefly discuss the statistical consistency of its two most common special cases.

The Nadaraya-Watson formulation of **hannah2010nonparametric** reduces to our balloon estimation formulation for the particular choice $k(n) = n$. The Nadaraya Watson formulation minimizes indeed the weighted average using a positive weighting function w_n over all observations

$$z_{\mathrm{tr}[n]}(x_0) \in \arg\min_z \mathrm{E}^n_{D_{\mathrm{tr}[n]}}\left[L(z,y)|x=x_0\right] := \frac{\mathbb{E}_{D_{\mathrm{tr}[n]}}\left[L(z,y)\cdot w_n(x,x_0)\right]}{\mathbb{E}_{D_{\mathrm{tr}[n]}}\left[w_n(x,x_0)\right]}. \tag{14}$$

Typically, the weighing function is taken to be $w_n(\bar{x}, x_0) = S(\|\bar{x} - x_0\|_2 / h(n))$ with the help of a positive smoothing function S and bandwidth parameter $h(n)$. Some common popular choices of smoothers are given

in Figure 2. The Nadaraya-Watson formulation is particularly amenable to theoretical analysis due mostly to its simplicity. Nadaraya-Watson estimation can indeed be shown to be point-wise consistent when using an appropriately scaled bandwidth parameter $h(n)$ for any of the smoother functions listed in Figure 2.

Theorem 1 (**walk2010strong**). *Let us have a loss function satisfying* $\mathbb{E}_{D^\star}\left[|L(\bar{z},y)|\cdot\max\{\log(|L(\bar{z},y)|),0\}\right]<\infty$ *for all* $\bar{z}$. *Let the bandwidth* $h(n)=cn^{-\delta}$ *for some* $c>0$ *and* $\delta\in(0,1/\dim(x))$. *Let* S *be any of the smoother functions listed in Figure 2 and the weighing function is taken to be* $w_n(x,x_0)=S(\|x-x_0\|_2/h(n))$. *Then, the balloon estimation formulation (with* $k(n)=n$*) is asymptotically consistent for any* $D^\star$, *i.e., with probability one we have*

$$\lim_{n\to\infty}\mathrm{E}^n_{D_{\mathrm{data}[n]}}\left[L(\bar{z},y)|x=x_0\right]=\mathbb{E}_{D^\star}\left[L(\bar{z},y)|x=x_0\right]\qquad\forall\bar{z}.$$

Nearest neighbors learning is one of the most fundamental yet very simple learning methods and is discussed in virtually any textbook on machine learning. It is a common choice for learning when there is a lot of data but little or no prior knowledge about the distribution of that data. Choosing a weighing function $w_n(x,x_0)=1$ in our balloon estimation formulation yields precisely the nearest-neighbors formulation discussed in **bertsimas2014predictive**. The nearest neighbors formulation indeed minimizes average

$$z_{\mathrm{tr}[n]}(x_0)\in\arg\min_z\ \mathrm{E}^n_{D_{\mathrm{tr}[n]}}\left[L(z,y)|x=x_0\right]:=\frac{\mathbb{E}_{D_{\mathrm{tr}[n]}}\left[L(z,y)\cdot\mathbb{1}\{(x,y)\in N^k_n(x_0)\}\right]}{\mathbb{E}_{D_{\mathrm{tr}[n]}}\left[\mathbb{1}\{(x,y)\in N^k_n(x_0)\}\right]}\tag{15}$$

restricted to the smallest data neighborhood around the context $x=x_0$ containing no less than k observations. Nearest neighbors estimation is consistent under very mild technical conditions provided that the number of neighbors is scaled appropriately with the number of training data samples.

Theorem 2 (**walk2010strong**). *Assume* $\mathrm{dist}(\bar{d}=(\bar{x},\bar{y}),x_0)=\|\bar{x}-x_0\|_2$ *and follow the random tie breaking rule discussed in* **gyorfi2006distribution**. *Let* $k(n)=\lceil\min\{cn^\delta,n\}\rceil$ *for some* $c>0$ *and* $\delta\in(0,1)$. *Then, the balloon estimation formulation (with* $w_n(\bar{x},x_0)=1$*) is asymptotically consistent for any* $D^\star$, *i.e., with probability one we have*

$$\lim_{n\to\infty}\mathrm{E}^n_{D_{\mathrm{data}[n]}}\left[L(\bar{z},y)|x=x_0\right]=\mathbb{E}_{D^\star}\left[L(\bar{z},y)|x=x_0\right]\quad\forall\bar{z}.$$

Theorems 1 and 2 merely establish the point-wise consistency of the balloon estimate of the cost for a fixed decision $\bar{z}$ in two particular cases related to Nadaraya-Watson ($k(n)=n$) and nearest neighbors ($w_n(\bar{x},x_0)=1$) learning. Point-wise consistency of cost estimates crucially does not establish consistency of the associated budget minimization formulations, i.e., $\lim_{n\to\infty}\mathbb{E}_{D^\star}\left[L(z_{\mathrm{data}[n]},y)|x=x_0\right]=\min_z\ \mathbb{E}_{D^\star}\left[L(z,y)|x=x_0\right]$. For the latter uniform convergence of the cost estimates needs to be shown. **bertsimas2014predictive** do so under rather mild technical assumptions. Indeed, it suffices to assume that the family of loss functions $\{L(\cdot,\bar{y})\}_{\bar{y}}$ is equicontinuous. This equicontinuity assumption is rather mild as any family of functions with common Lipschitz constant is equicontinuous. Lemma 4 of **bertsimas2014predictive** establishes that if the cost estimate converges, it does so uniformly in the decision $\bar{z}$, i.e., $|\mathrm{E}^n_{D_{\mathrm{data}[n]}}\left[L(\bar{z},y)|x=x_0\right]-\mathbb{E}_{D^\star}\left[L(\bar{z},y)|x=x_0\right]|\leq\epsilon(n)$ with probability one and $\lim_{n\to\infty}\epsilon(n)=0$ for all $\bar{z}$ in a compact feasible decision set $\mathrm{dom}(\mathbb{E}_{Y^\star}\left[L(z,y)\right])$. It is not hard to see that the previous uniform bound in turn implies the consistency of the budget minimization formulation. Indeed, with probability one we have

$$\begin{aligned}
\min_z\ \mathbb{E}_{D^\star}\left[L(z,y)|x=x_0\right]+\epsilon(n)&\geq\mathrm{E}^n_{D_{\mathrm{data}[n]}}\left[L(z^\star(x_0),y)|x=x_0\right]\\
\mathrm{E}^n_{D_{\mathrm{data}[n]}}\left[L(z^\star(x_0),y)|x=x_0\right]&\geq\mathrm{E}^n_{D_{\mathrm{data}[n]}}\left[L(z_{\mathrm{data}[n]},y)|x=x_0\right]\\
\mathrm{E}^n_{D_{\mathrm{data}[n]}}\left[L(z_{\mathrm{data}[n]},y)|x=x_0\right]&\geq\mathbb{E}_{D^\star}\left[L(z_{\mathrm{data}[n]},y)|x=x_0\right]-\epsilon(n)\\
\implies 0\leq\mathbb{E}_{D^\star}\left[L(z_{\mathrm{data}[n]},y)|x=x_0\right]-\min_z\ \mathbb{E}_{D^\star}\left[L(z,y)|x=x_0\right]&\leq 2\epsilon(n).
\end{aligned}$$

Here the optimality gap between the cost of the full information decision $z^\star(x_0)$ and the cost of the data-driven decision $z_{\mathrm{data}[n]}$ is bounded by $\epsilon(n)$ which in turn converges to zero. It is noteworthy to remark that consistency of the balloon estimation formulation is a property which holds uniformly for all distributions $D^\star$ which may have generated our data. As we argued in the introduction consistency does not mean good out-of-sample performance when only a finite amount of training data is available.

2.2 Bootstrapping Balloon Estimation

In this section, we show that carrying out balloon estimation on bootstrap data can be posed as a convex optimization problem. The significance of the following result will become clear in Section 3. Recall that any bootstrap data set $\mathrm{bs}[n]$ counts the same number of observations and shares its observations with the training data set $\mathrm{tr}[n]$ from which it is resampled. In terms of its empirical distribution $D_{\mathrm{bs}[n]}$ this translates to

$$D_{\mathrm{bs}[n]} \in \mathcal{D}_{n,n} := \left\{ D \;:\; \textstyle\sum_{(\bar{x},\bar{y})\in\Omega_n} D[\bar{x},\bar{y}] = 1,\; n \cdot D[\bar{x},\bar{y}] \in \{0,1,2,\ldots,n\} \quad \forall (\bar{x},\bar{y}) \in \Omega_n \right\} \subset \mathcal{D}_n.$$

We would like to characterize our balloon estimator $\mathrm{E}_D^n\left[L(\bar{z},y)|x=x_0\right]$ as an explicit function of such empirical bootstrap distributions $D \in \mathcal{D}_{n,n}$. That is, an explicit function predicting the cost of decisions $\bar{z}$ in the context $x = x_0$ based on bootstrap data. We can do so by first associating a partial predictor to each of the $j \in [n]$ neighborhoods sets $N_n^j(x_0)$ with the help of a linear optimization problem

$$\begin{array}{rll} \mathrm{E}_D^{n,j}\left[L(\bar{z},y)|x=\bar{x}\right] := \sup\limits_{s>0,P} & \textstyle\sum_{(\bar{x},\bar{y})\in N_n^j(x_0)} w_n(\bar{x},x_0)\cdot L(\bar{z},\bar{y})\cdot P[\bar{x},\bar{y}] & \\ \text{s.t.} & P[\bar{x},\bar{y}] \geq 0,\; s\cdot D[\bar{x},\bar{y}] = P[\bar{x},\bar{y}] \quad \forall (\bar{x},\bar{y}) \in \Omega_n & \\ & \textstyle\sum_{(\bar{x},\bar{y})\in\Omega_n} P[\bar{x},\bar{y}] = s,\;\; \sum_{(\bar{x},\bar{y})\in N_n^j(x_0)} w_n(\bar{x},x_0)\cdot P[\bar{x},\bar{y}] = 1, & \\ & \textstyle\sum_{(\bar{x},\bar{y})\in N_n^j(x_0)} P[\bar{x},\bar{y}] \geq \frac{k}{n}\cdot s,\;\; \sum_{N_n^{j-1}(x_0)} P[\bar{x},\bar{y}] \leq \frac{k-1}{n}\cdot s. & \end{array} \tag{16}$$

Here an optimization variable $P[\bar{x},\bar{y}]$ for each distinct observed data point $(\bar{x},\bar{y})$ in the training data $\mathrm{tr}[n]$ is introduced together with an additional variable s. Note that the domain of the partial estimators as a function of the distribution D satisfies

$$\begin{array}{rl} & \mathrm{dom}\ \mathrm{E}_D^{n,j}\left[L(z,y)|x=\bar{x}\right] \\ \subseteq & \mathcal{D}_n^j := \left\{ D \in \mathcal{D}_n : \textstyle\sum_{N_n^{j-1}(x_0)} D[\bar{x},\bar{y}] \leq (k-1)/n < k/n \leq \sum_{(\bar{x},\bar{y})\in N_n^j(x_0)} D[\bar{x},\bar{y}] \right\}. \end{array} \tag{17}$$

For a distribution D to be in the domain of the partial estimator the constraints in the optimization formulation given in equation (16) must indeed be feasible. In other words, there must exists some P and $s > 0$ for which $s \cdot D[\bar{x},\bar{y}] = P[\bar{x},\bar{y}]$ for all $(\bar{x},\bar{y}) \in \Omega_n$. The last two constraints in equation (16) then imply that any such $D \in \mathcal{D}_n$ must also be in $\mathcal{D}_n^j$. We follow here the standard convention that the supremum over an empty set to be unbounded from below. The balloon estimator can be decomposed using this convention as the maximum of each of the partial estimators previously defined. We refer for the proof of the following result to Appendix A.1.

Theorem 3 (An Equivalent Optimization Characterization)**.** *We have that the balloon estimator can be decomposed as*

$$\begin{aligned} \mathrm{E}_D^n\left[L(z,y)|x=x_0\right] &= \begin{cases} \mathrm{E}_D^{n,1}[L(z,y)|x=x_0] & \text{if } \mathrm{D} \in \mathcal{D}_{\mathrm{n}}^1, \\ \vdots & \vdots \\ \mathrm{E}_D^{n,n}[L(z,y)|x=x_0] & \text{if } \mathrm{D} \in \mathcal{D}_{\mathrm{n}}^{\mathrm{n}}, \end{cases} \\ &= \max_{j\in[n]} \mathrm{E}_D^{n,j}\left[L(z,y)|x=x_0\right] \end{aligned}$$

for all empirical distributions $D \in \mathcal{D}_{n,n}$ constructed from n samples contained in the training data set.

3 Distributionally Robust Prescriptions

When working with data instead of models, one should safeguard against making decisions which display promising training performance, but lead to out-of-sample disappointment. The nominal supervised learning formulations discussed before are indeed gullible and tend to be over-calibrated to one particular data set. It is clear that when given only a limited amount of training data, any data-driven method must be guarded against such overfitting phenomena. Distributionally robust optimization has attracted

Type	Model distance function $R(D, D') =$
Total Variation	$\sum_{(\bar{x},\bar{y})\in\Omega_n} \lvert D[\bar{x},\bar{y}] - D'[\bar{x},\bar{y}]\rvert$
Pearson	$\sum_{(\bar{x},\bar{y})\in\Omega_n} (D[\bar{x},\bar{y}] - D'[\bar{x},\bar{y}])^2/D'[\bar{x},\bar{y}]$
Entropy	$\sum_{(\bar{x},\bar{y})\in\Omega_n} \log\left(D[\bar{x},\bar{y}]/D'[\bar{x},\bar{y}]\right) D[\bar{x},\bar{y}]$
Burg Entropy	$\sum_{(\bar{x},\bar{y})\in\Omega_n} \log\left(D'[\bar{x},\bar{y}]/D[\bar{x},\bar{y}]\right) D'[\bar{x},\bar{y}]$
f-Divergence	$\sum_{(\bar{x},\bar{y})\in\Omega_n} f\left(D[\bar{x},\bar{y}]/D'[\bar{x},\bar{y}]\right) D'[\bar{x},\bar{y}]$

Table 1: Model distance functions based on popular probability divergence metrices. The f-divergences give rise to a model distance function for convex functions f with $f(1) = 0$. The Pearson and Burg entropy are particular cases for $f(t) = t^2 - 1$ and $f(t) = -\log(t)$, respectively. **postek2016computationally** provide and discus many more probability divergences in great detail.

significant attention as it provides the sample average formulation with a disciplined safeguard mechanism against overfitting. By using a robust counterpart with respect to an ambiguity set of distributions around an estimated nominal one, they were shown by **vanparys2017data** to be minimally biased while still enjoying statistical out-of-sample guarantees. Many interesting choices of the ambiguity set furthermore result in a tractable overall decision-making approach. The ambiguity set can be defined, for example, through confidence intervals for the distribution's moments as done by **delage2010distributionally**; **vanparys2016generalized**; **vanparys2015distributionally**; **stellato2016multivariate**. Alternatively, **wang2009likelihood** use an ambiguity set that contains all distributions that achieve a prescribed level of likelihood, while **bertsimas2014robust** based theirs on models which pass a statistical hypothesis test. Distance-based ambiguity sets contain all models sufficiently close to a reference with respect to probability metrics such as the Prokhorov metric (**erdogan2006**), the Wasserstein distance (**pflug2007**; **esfahani2015data**), or the total variation distance (**sun2016**).

As a major contribution in this paper we generalize distributionally robust optimization to estimate-then-optimize formulations. We construct generic robust balloon estimation formulations with the help of a model distance function. The resulting robust balloon estimation formulations should suffer only a limited out-of-sample disappointment (12) on the bootstrap data bs$[n]$ generated as defined in (9). Generic robust estimate-then-optimize formulations are not necessarily robust in the sense put forward in Definition 1. In the next section we will show that such bootstrap robustness guarantees can be obtained by considering a very particular bootstrap distance function. However, we will concern ourselves in this section only with showing the practical viability of generic robust estimate-then-optimize formulations with respect to any model distance function in terms of computational tractability.

Definition 4 (Model Distance Function)**.** *A model distance function R is a function quantifying the distance between two empirical distributions supported on Ω_n enjoying the following property:*

(i) Discrimination: $R(D, D') \geq 0$ *for all* D *and* D'*, while* $R(D', D) = 0$ *if and only if* $D' = D$*.*

(ii) Convexity: $R(D, D')$ *is a convex function of* D *in* D *for all fixed* D'*.*

We define first a generic robust counterpart to our nominal balloon estimation formulation with respect to the ambiguity set $\left\{D \;:\; R(D, D_{\mathrm{tr}[n]}) \leq r\right\}$ consisting of all empirical models at distance not exceeding r.

Definition 5 (Distributionally Robust Budget Formulations)**.** *The distributionally robust counterpart to an estimate-then-optimize formulation with respect to the model distance function D is defined as*

$$\begin{aligned} z^{\mathrm{r}}_{\mathrm{tr}[n]}(x_0) \in \arg\min_z \; c_n(z, D_{\mathrm{tr}[n]}, x_0) := \sup_D \quad & \mathrm{E}^n_D\left[L(z, Y) | x = x_0\right] \\ \text{s.t.} \quad & D \in \mathcal{D}_n, \\ & R(D, D_{\mathrm{tr}[n]}) \leq r. \end{aligned} \tag{18}$$

Due to the discrimination property of the model distance function, the nominal supervised learning formulation is recovered when the robustness radius tends towards zero. In that case we are indeed merely robust with respect to the singleton $\{D \ : \ R(D, D_{\mathrm{tr}[n]}) \leq 0\} = \{D_{\mathrm{tr}[n]}\}$. Using a robust counterpart instead of nominal supervised learning formulations will help us protect against making prescriptions which do well on the training data set but tend to disappoint on unseen data. The robust training prescription $z^{\mathrm{r}}_{\mathrm{tr}[n]}(x_0)$ indeed does well not on one particular empirical training distribution $D_{\mathrm{tr}[n]}$ but rather on all distributions $\{D \ : \ R(D, D_{\mathrm{tr}[n]}) \leq r\}$ at distance less than r simultaneously. The particular distance function D dictates which distributions are close to the nominal training model and consequently should be chosen with care. Several popular choices are listed in Table 1. In the next section, we will single out one particularly relevant model distance function in the context of the bootstrap disappointment defined in (12).

In Section 2, we have divided the training data into the nested neighborhoods $N^j_n(x_0)$. Each of these neighborhoods sets contains those data points closest to the context of interest $x = x_0$. We associated with each of these neighborhoods partial estimators. Theorem 3 shows how the balloon estimator can be decomposed as the maximum of these n partial estimators Here, we likewise first introduce partial robust budget function through the optimization problem

$$\begin{aligned} c^j_n(z, D, x_0) := \sup_{s>0, P} \quad & \textstyle\sum_{(\bar{x},\bar{y})\in N^j_n(x_0)} w_n(\bar{x}, x_0) \cdot L(z, \bar{y}) \cdot P[\bar{x}, \bar{y}] \\ \text{s.t.} \quad & s \cdot R(P/s, D) \leq s \cdot r, \\ & \textstyle\sum_{(\bar{x},\bar{y})\in\Omega_n} P[\bar{x}, \bar{y}] = s, \ \ \sum_{(\bar{x},\bar{y})\in N^j_n(x_0)} w_n(\bar{x}, x_0) \cdot P[\bar{x}, \bar{y}] = 1, \\ & \textstyle\sum_{(\bar{x},\bar{y})\in N^j_n(x_0)} P[\bar{x}, \bar{y}] \geq s \cdot k/n, \ \ \sum_{(\bar{x},\bar{y})\in N^{j-1}_n(x_0)} P[\bar{x}, \bar{y}] \leq s \cdot (k-1)/n. \end{aligned} \tag{19}$$

The previous maximization problem characterizing the robust balloon estimators is concave. Its first optimization variable s is merely one dimensional, while an additional optimization variable $P(\bar{x}, \bar{y})$ is added for each distinct training data point in the support Ω_n. Its ultimate constraint is the only nonlinear one and is convex as the perspective function $s \cdot R(P/s, D)$ is convex jointly in both variables whenever the model distance function R is convex. Remark again that the smoother weights $w_n(\bar{x}, x_0)$ are assumed to be non-negative. As a result, the robust estimator $c^j_n(z, D, x_0)$ is a convex function of the decision z.

Theorem 4 (Robust Balloon Estimation Formulation)**.** *The robust balloon estimation formulation can be reformulated as the convex optimization problem*

$$z^{\mathrm{r}}_{\mathrm{tr}[n]}(x_0) \in \arg\min_z \ c_n(z, D_{\mathrm{tr}[n]}, x_0) := \max_{j\in[n]} \ c^j_n(z, D_{\mathrm{tr}[n]}, x_0). \tag{20}$$

Proof. The chain of equalities

$$\begin{aligned} \{(s, P) \ : \ \exists D \text{ s.t. } s \cdot D = P, \ R(D, D_{\mathrm{tr}[n]}) \leq r\} &= \{(s, P) \ : \ R(P/s, D_{\mathrm{tr}[n]}) \leq r\} \\ &= \{(s, P) \ : \ s \cdot R(P/s, D_{\mathrm{tr}[n]}) \leq s \cdot r\} \end{aligned}$$

imply that the robust partial budget function $c^j_n(z, D_{\mathrm{tr}[n]}, x_0)$ correspond exactly to the robust version $\sup\{\mathrm{E}^{n,j}_D[L(z, y)|x = x_0] : R(D, D_{\mathrm{tr}[n]}) \leq r\}$ of the partial balloon estimators. We have from Theorem 3 the composition $c_n(z, D_{\mathrm{tr}[n]}, x_0) := \max_{j\in[n]} c^j_n(z, D_{\mathrm{tr}[n]}, x_0)$. The final optimization formulation is convex as it consists of minimizing the maximum of the individually convex partial budget functions $c^j_n(z, D_{\mathrm{tr}[n]}, x_0)$. □

The previous theorem establishes the computational tractability of robust balloon formulations and in particular entails both robust Nadaraya-Watson and nearest neighbors formulations. We indicate that in case of the former a less involved formulation can be obtained. We refer for its proof to Appendix A.2.

Corollary 1. *The robust balloon estimation formulation ($k(n) = n$) can be reformulated as the convex*

optimization problem $z^{\mathrm{r}}_{\mathrm{tr}[n]}(x_0) \in \arg\min_z \; c_n(z, D_{\mathrm{tr}[n]}, x_0)$ *with*

$$\begin{aligned} c_n(z, D_{\mathrm{tr}[n]}, x_0) := \quad & \sup_{D} \quad \textstyle\sum_{(\bar{x},\bar{y})\in\Omega_n} w_n(\bar{x}, \bar{x}_0) \cdot L(z, \bar{y}) \cdot D[\bar{x}, \bar{y}] / \sum_{(\bar{x},\bar{y})\in\Omega_n} w_n(\bar{x}, \bar{x}_0) \cdot D[\bar{x}, \bar{y}] \\ & \text{s.t.} \quad R(D, D_{\mathrm{tr}[n]}) \leq r \\ = \quad & \sup_{s>0, P} \quad \textstyle\sum_{(\bar{x},\bar{y})\in\Omega_n} w_n(\bar{x}, x_0) \cdot L(z, \bar{y}) \cdot P[\bar{x}, \bar{y}] \\ & \text{s.t.} \quad s \cdot R(P/s, D_{\mathrm{tr}[n]}) \leq s \cdot r, \\ & \quad \textstyle\sum_{(\bar{x},\bar{y})\in\Omega_n} P[\bar{x}, \bar{y}] = s, \;\; \sum_{(\bar{x},\bar{y})\in\Omega_n} w_n(\bar{x}, x_0) \cdot P[\bar{x}, \bar{y}] = 1. \end{aligned}$$

Robust Nadaraya-Watson formulation have been proposed before, most notably by **hanasusanto2013robust** in the context of robust dynamic programming. We briefly take here the opportunity to point out that our robust Nadaraya-Watson formulation is fundamentally different. The robust Nadaraya-Watson formulations found in **hanasusanto2013robust** correspond directly to the more restricted alternative

$$\begin{aligned} c'_n(z, D_{\mathrm{tr}[n]}, x_0) := \sup_{D} \quad & \textstyle\sum_{(\bar{x},\bar{y})\in\Omega_n} w_n(\bar{x}, x_0) \cdot L(z, \bar{y}) \cdot D[\bar{x}, \bar{y}] / \sum_{(\bar{x},\bar{y})\in\Omega_n} w_n(\bar{x}, x_0) \cdot D[\bar{x}, \bar{y}] \\ \text{s.t.} \quad & R(D, D_{\mathrm{tr}[n]}) \leq r, \\ & \textstyle\sum_{(\bar{x},\bar{y})\in\Omega_n} w_n(\bar{x}, x_0) \cdot D[\bar{x}, \bar{y}] = \sum_{(\bar{x},\bar{y})\in\Omega_n} w_n(\bar{x}, x_0) \cdot D_{\mathrm{tr}[n]}[\bar{x}, \bar{y}]. \end{aligned}$$

In **hanasusanto2013robust** one particular convex model distance function R based on a scaled version of the Pearson distance mentioned in Table 1 is singled out. In terms of the convex reformulation given in Corollary 1, this alternative can be easily seen to correspond to simply restricting the optimization variable $s > 0$ to take fixed value $s = 1/\sum_{(\bar{x},\bar{y})\in\Omega_n} w_n(\bar{x}, x_0) \cdot D_{\mathrm{tr}[n]}[\bar{x}, \bar{y}]$. Even in the particular context of the Nadaraya-Watson formulation, our notion of distributional robustness hence seems to be novel.

In this section, we were merely interested in the practical viability of the generic robust supervised learning formulations stated in Definition 5. We argued that for arbitrary convex model distance functions this is indeed the case. Most convex model distance functions do not necessarily guarantee that the corresponding generic robust supervised learning formulations perform well on the out-of-sample bootstrap data. In the next section, we single out one particular distance function for which this is not the case. Correspondingly, we will come to denote this special model distance function as the bootstrap distance function.

4 Bootstrap Robust Prescriptions

In the previous section, we indicated that a robust balloon estimation formulation with respect to any arbitrary model distance function is not necessarily bootstrap robust in the sense of Definition 1. In the remainder of this section we will indicate that for the following particular model distance function this is however not the case. We also provide an even more practical representation of the particular budget minimization formulations based on convex duality specific to this particular bootstrap distance function.

Definition 6 (The Bootstrap Distance Function)**.** *For two empirical models D and D' in $\mathcal{D}_n$ we define their bootstrap distance as*

$$B(D, D') := \textstyle\sum_{(\bar{x},\bar{y})\in\Omega_n} D[\bar{x}, \bar{y}] \cdot \log\left(\frac{D\,[\bar{x},\bar{y}]}{D'[\bar{x},\bar{y}]}\right). \tag{21}$$

The bootstrap distance between two empirical models is recognized as the relative entropy distance for discrete distributions as stated in Table 1. The relative entropy is also known as information for discrimination, cross-entropy, information gain or Kullback-Leibler divergence (**kullback1951information**). We first prove that in fact the resulting particular robust balloon estimation formulation is in fact statistically consistent under mild conditions.

4.1 Statistical consistency

The bootstrap robust Nadaraya-Watson estimation is consistent for bounded loss functions when the robustness radius $r(n)$ is appropriately scaled with respect to the bandwidth parameter $h(n)$. To establish consistency of the associated bootstrap robust balloon estimation, it is noteworthy to know that the total variation distance is related to the bootstrap distance as

$$\begin{aligned}\|D - D'\|_1 &:= \max_{\mathcal{E}\subseteq\Omega_n} |D[\mathcal{E}] - D'[\mathcal{E}]| \\ &= \sum_{(\bar{x},\bar{y})\in\Omega_n} |D[\bar{x},\bar{y}] - D'[\bar{x},\bar{y}]| \\ &\leq \sqrt{1/2 B(D,D')}\end{aligned}$$

which better known as to Pinkser's inequality.

Theorem 5 (Bootstrap Robust Nadaraya-Watson Esitmation)**.** *Assume a bounded loss function $L(\bar{z},\bar{y}) < \bar{L} < \infty$ for all feasible decisions $\bar{z}$ and parameters $\bar{y}$. Let $h(n) = cn^{-\delta}$ for some $c > 0$ and $\delta \in (0, 1/\dim(x))$. Let S be any of the smoother functions listed in Figure 2 and the weighing function is taken to be $w_n(\bar{x}, x_0) = S(\|\bar{x} - x_0\|_2 / h(n))$. Let the robustness radius satisfy $\lim_{n\to\infty} \sqrt{r(n)}/h(n)^{\dim(x)} = 0$ for model distance function $R = B$. Then, bootstrap robust balloon estimation (with $k(n) = n$) is asymptotically consistent for any $D^\star$, i.e., with probability one*

$$\lim_{n\to\infty} \mathbb{E}_{D^\star}\left[L(z^r_{\text{data}[n]}, y)|x = x_0\right] = \min_z \mathbb{E}_{D^\star}\left[L(z,y)|x = x_0\right].$$

We refer to the proof of the previous result based on Pinsker's inequality to Appendix A.3. A similar result holds for bootstrap robust nearest neighbors estimation. Here the robustness radius $r(n)$ needs to be scaled appropriately as a function of the number of nearest neighbors $k(n)$. Again the proof is technically based on Pinsker's inequality and referred to Appendix A.4.

Theorem 6 (Bootstrap Robust nearest neighbors Estimation)**.** *Assume a bounded loss function $L(z,\bar{y}) < \bar{L} < \infty$ for all feasible decisions $\bar{z}$ and parameters $\bar{y}$. Let $\text{dist}(\bar{d} = (\bar{x},\bar{y}), x_0) = \|\bar{x} - x_0\|_2$ and follow the tie breaking rule discussed in* **gyorfi2006distribution**. *Let $k(n) = \lceil \min\{cn^\delta, n\} \rceil$ for some $c > 0$ and $\delta \in (0,1)$. Let the robustness radius satisfy $\lim_{n\to\infty} n\sqrt{r(n)}/k(n) = 0$ for model distance function $R = B$. Then, bootstrap robust balloon estimation ($w_n(\bar{x}, x_0) = 1$) is asymptotically consistent for any $D^\star$, i.e., with probability one*

$$\lim_{n\to\infty} \mathbb{E}_{D^\star}\left[L(z^r_{\text{data}[n]}, y)|x = x_0\right] = \min_z \mathbb{E}_{D^\star}\left[L(z,y)|x = x_0\right].$$

4.2 Bootstrap Performance

We will now see that a particular robustness radius is advisable if a constant bootstrap performance is required. To establish that besides consistency, the bootstrap robust balloon estimation formulation suffers only a limited bootstrap disappointment, we will need one elementary result from large deviation theory. The following theorem characterizes the essential large deviation behavior of the empirical distribution $D_{\text{bs}[n]}$ of the bootstrap data resampled from the training data as outlined in (9). This result forms the backbone of most of the theoretical results in this paper concerning the out-of-sample properties of our robust balloon estimation formulation.

Theorem 7 (The Bootstrap Inequality (**csiszar1984sanov**))**.** *The probability that the random bootstrap distribution $D_{\text{bs}[n]}$ realizes in a convex set of models $\mathcal{C}$ satisfies the finite sample inequality*

$$D^\infty_{\text{tr}[n]}\left[D_{\text{bs}[n]} \in \mathcal{C}\right] \leq \exp\left(-n \cdot \inf_{D\in\mathcal{C}} B(D, D_{\text{tr}[n]})\right), \quad \forall n \geq 1. \tag{22}$$

The geometry of the bootstrap inequality is visualized in Figure 3. The bootstrap inequality is of high-quality and is asymptotically exact in the exponential case. Large deviation theory provides the corresponding lower bound

$$-\inf_{D\in\text{int}\,\mathcal{C}} B(D, D_{\text{tr}[n]}) \leq \liminf_{n\to\infty} \frac{1}{n} \log D^\infty_{\text{tr}[n]}\left[D_{\text{bs}[n]} \in \mathcal{C}\right], \tag{23}$$

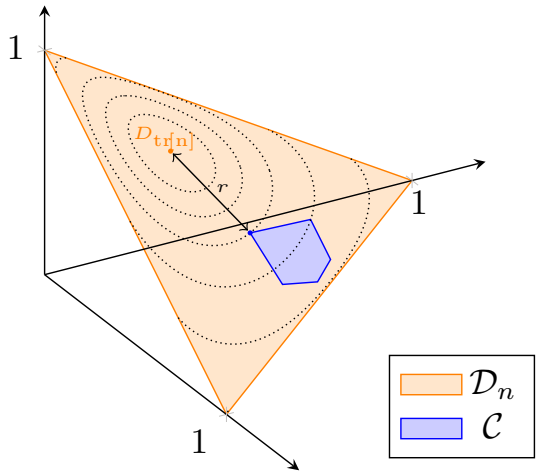


Figure 3: Visualization of the bootstrap inequality (22) in Theorem 7. The probability $D^{\infty}_{\mathrm{tr}[n]}(D_{\mathrm{bs}[n]} \in \mathcal{C})$ decays at the exponential rate $r := \inf_{D \in \mathcal{C}} B\big(D, D_{\mathrm{tr}[n]}\big)$, which can be viewed as the bootstrap *distance* of the empirical training model $D_{\mathrm{tr}[n]}$ to the set of interest $\mathcal{C}$.

which meets the upper bound (22) asymptotically in the exponential case for regular event sets $\mathcal{C} = \mathrm{cl}\,\mathrm{int}\,\mathcal{C}$ as the bootstrap distance function is continuous in its first argument when its second argument $D_{\mathrm{tr}[n]}[\bar{x}, \bar{y}] \geq 1/n$ happens to be positive for any $(\bar{x}, \bar{y}) \in \Omega_n$. For further discussion on large deviation theory we refer the reader to the work of **csiszar1984sanov**.

Notice that for the optimization problem defining the partial estimators c^j_n to be nontrivial for the training model $D_{\mathrm{tr}[n]}$, the robustness radius r needs to be bigger than the minimum robustness radius

$$\begin{aligned} r^j_n := \inf \quad & R(D, D_{\mathrm{tr}[n]}) \\ \text{s.t.} \quad & D \in \mathcal{D}^j_n \end{aligned} \tag{24}$$

where $\mathcal{D}^j_n$ was defined in (17). If this is the case, then the feasible set of the optimization problem (19) defining the partial cost c^j_n is indeed non-empty. Also these extremal bootstrap radii are characterized as the solution of a tractable convex optimization problem. These bootstrap radii will come to play an important role in the characterization of the bootstrap disappointment suffered by our robust balloon estimator formulation.

Theorem 8 (Performance of the Bootstrap Robust Balloon Estimation Formulation)**.** *The robust balloon estimation formulation with bootstrap distance function ($R = B$) and robustness radius r suffers a bootstrap disappointment defined in (12) at most*

$$b = \textstyle\sum_{j \in [n]} \exp\left(-n \cdot \max\{r, r^j_n\}\right).$$

Proof. Let us fix a training data set with empirical distribution $D_{\mathrm{tr}[n]}$ and consider a given decision $\bar{z}$. Let $\bar{c}^j_n := \mathrm{E}^{n,j}_{D_{\mathrm{tr}[n]}}[L(\bar{z}, y)|x = x_0]$ with $\bar{c}_n = \max_{j \in [n]} \bar{c}^j_n$ be the budgeted cost based on the training data. In order to prove the theorem, it suffices to characterize the probability of the event that the empirical distribution $D_{\mathrm{bs}[n]}$ of random bootstrap data resampled from the training data realizes in the set $\mathcal{C} = \{D \in \mathcal{D}_n \ : \ c_n(\bar{z}, D, x_0) > \bar{c}_n\} = \cup_{j \in [n]} \mathcal{C}_j$ with

$$\begin{aligned} \mathcal{C}_j := \ & \left\{ D \in \mathcal{D}^j_n \ : \ \mathrm{E}^{n,j}_D\left[L(\bar{z}, y)|x = x_0\right] > \bar{c}_n \right\} \\ = \ & \{D \in \mathcal{D}^j_n : \textstyle\sum_{(\bar{x}, \bar{y}) \in N^j_n(x_0)} w_n(\bar{x}, x_0) \cdot L(\bar{z}, y) \cdot D[\bar{x}, \bar{y}] > \bar{c}_n \cdot \sum_{(\bar{x}, \bar{y}) \in N^j_n(x_0)} w_n(\bar{x}, x_0) \cdot D[\bar{x}, \bar{y}]\}. \end{aligned}$$

Each of the partial sets $\mathcal{C}_j$ is a convex polyhedron. We can use the union bound to establish

$$D^{\infty}_{\mathrm{tr}[n]}(D_{\mathrm{bs}[n]} \in \mathcal{C}) \leq \textstyle\sum_{j \in [n]} D^{n}_{\mathrm{tr}[n]}(D_{\mathrm{bs}[n]} \in \mathcal{C}_j).$$

The robust budget cost $\bar{c}_n$ is constructed precisely as to ensure that $\inf_{D\in\mathcal{C}_j} R(D, D_{\mathrm{tr}[n]}) > r$. Indeed, we have the implication $\bar{D} \in \mathcal{C}_j \implies \mathrm{E}_{\bar{D}}^{n,j}[L(\bar{z}, y)|x = x_0] > \bar{c}_n \geq \sup\{\mathrm{E}_{D}^{n,j}[L(\bar{z}, y)|x = x_0] \mid R(D, D_{\mathrm{tr}[n]}) \leq r,\, D \in \mathcal{D}_n\}$ which in turn itself implies $R(\bar{D}, D_{\mathrm{tr}[n]}) > r$. By virtue of the inclusion $\mathcal{C}_j \subseteq \mathcal{D}_n^j$, evidently, we must also have that $\inf_{D\in\mathcal{C}_j} R(D, D_{\mathrm{tr}[n]}) > r_n^j := \inf\{R(D, D_{\mathrm{tr}[n]}) : D \in \mathcal{D}_n^j\}$. Hence, the result follows from the bootstrap inequality (22) applied to each of the probabilities $D_{\mathrm{tr}[n]}^{\infty}(D_{\mathrm{bs}[n]} \in \mathcal{C}_j)$ as in this particular case the employed model distance function ($R = B$) coincides with the bootstrap distance function. □

The previous theorem gives an explicit characterization of the bootstrap performance of the general balloon estimation formulation. Choosing the robustness radius $r(n)$ yielding a desired bootstrap disappointment b can not be done analytically, but thanks to the convex characterization (24) of the minimum bootstrap radii r_n^j it can nevertheless be carried out numerically in a tractable fashion. It is also trivial to see that adding robustness to the extend $r(n) \geq (\log(n) + \log(1/b))/n$ suffices to have bootstrap disappointment at most b. When we scale the robustness radius in such a way that

$$\lim_{n\to\infty} \frac{r(n) \cdot n}{\log(n)} = \infty$$

then the disappointment on bootstrap data asymptotically converges to zero when the number of training data points n tends to infinity. To be asymptotically consistent the robust nearest neighbors formulation needs to satisfy $\lim_{n\to\infty} n\sqrt{r(n)}/k(n) = 0$ as pointed out in Theorem 6. Asymptotically vanishing disappointment on bootstrap data can be combined with consistency by taking a number of nearest neighbors $k(n) = \lceil \min\{cn^{\delta}, n\} \rceil$ for some $c > 0$ and $\delta \in (0, 1)$ while at the same time scaling the robustness radius as $r(n) = tn^{\gamma}$ with $-1 < \gamma < 2(\delta - 1)$ for any $t > 0$. For the Nadaraya-Watson formulation a slightly improved result can in fact be stated. We omit the proof and refer to Appendix A.5.

Corollary 2 (Bootstrap Performance of the Nadaraya-Watson Formulation)**.** *The robust balloon estimation formulation (with $k(n) = n$) with bootstrap distance function ($D = B$) suffers bootstrap disappointment as defined in (12) at most*

$$b = \exp(-n \cdot r).$$

Adding robustness to the extend $r(n) \geq \log(1/b)/n$ already suffices here to have bootstrap disappointment at most b. When we scale the robustness radius in such a way now that $\lim_{n\to\infty} r(n) \cdot n = \infty$ then the disappointment on bootstrap data asymptotically converges to zero when the number of training data points n tends to infinity. To be asymptotically consistent the robust Nadaraya-Watson formulation needs to satisfy $\lim_{n\to\infty} \sqrt{r(n)}/h(n)^{\dim(x)} = 0$ as pointed out in Theorem 5. Asympotically vanishing disappointment on bootstrap data can be combined with consistency by scaling the bandwidth parameter $h(n) = cn^{\delta}$ for some $c > 0$ and $\delta \in (0, 1/\dim(x))$ while at the same time scaling the robustness radius as $r(n) = tn^{\gamma}$ with $-1 < \gamma < 2\delta \cdot \dim(x)$ for any $t > 0$.

4.3 Dual Perspective

Despite all previous encouraging results regarding the bootstrap performance and consistency of the robust balloon estimation formulation, it is still stated as the solution to a saddle point problem in (20) which may be awkward to handle practically. Both the size and the number of the maximization problems constituting the bootstrap robust nearest neighbors formulation grows linearly with the amount of training data samples. The following lemma tries to alleviate one of these concerns by considering a dual formulation of the maximization problem characterizing the partial robust cost functions. We refer its proof to Appendix A.6.

Lemma 1 (Dual Representation balloon estimation)**.** *The partial bootstrap robust budget $c_n^j(\bar{z}, D, x_0)$ can*

be represented using a dual convex optimization problem as

$$\begin{array}{ll} \inf & \alpha \\ \text{s.t.} & \alpha \in \mathrm{R},\ \eta \in \mathrm{R}_+^2,\ \nu \in \mathrm{R}_+, \\ & \nu \log\Big(\sum_{(\bar{x},\bar{y})\in N_n^{j-1}(x_0)} \exp([(L(\bar{z},y)-\alpha)\cdot w_n(\bar{x},x_0)+\eta_1-\eta_2]/\nu)\cdot D[\bar{x},\bar{y}] \\ & \quad + \sum_{(\bar{x},\bar{y})\in N_n^{j}(x_0)\setminus N_n^{j-1}(x_0)} \exp([(L(\bar{z},y)-\alpha)\cdot w_n(\bar{x},x_0)+\eta_1]/\nu)\cdot D[\bar{x},\bar{y}] \\ & \quad + \sum_{\Omega_n\setminus N_n^{j}(x_0)} D[\bar{x},\bar{y}]\Big) + r\cdot\nu - \frac{k}{n}(\eta_1-\eta_2) - \frac{\eta_2}{n} \leq 0. \end{array} \tag{25}$$

when the robustness radius satisfies $r > r_n^j$ for all $D \in \mathcal{D}_n$.

The main advantage of using the previous convex dual formulation of the robust nearest neighbors formulation is that finding the optimal prescription $z^r_{\mathrm{tr}[n]}(x_0)$ now merely requires the solution of a convex optimization problem over the decision z and three additional dual variables α, β, and η, instead of a saddle point problem with variables of a dimension which may scale linearly in the amount of training data. This dependence on the amount of training data is again not completely eliminated as the constraint in the dual characterization (25) of the partial robust budget c_n^j still counts j terms. As the robust nearest neighbors cost function c_n consists of the maximum of all of these partial robust cost functions we still have to account for a total number of $O(n^2)$ such terms. Only in the special case of Nadaraya-Watson estimation can the quadratic size in terms of the number of optimization variables in n be avoided. We refer to the proof of next result to Appendix A.7.

Lemma 2 (Dual Representation Nadaraya-Watson estimation)**.** *The partial bootstrap robust cost $c_n^j(\bar{z}, D, x_0)$ can be represented using a dual convex optimization problem as*

$$\begin{array}{ll} \inf & \alpha \\ \text{s.t.} & \alpha \in \mathrm{R},\ \nu \in \mathrm{R}_+, \\ & \nu\cdot\log\Big(\sum_{(\bar{x},\bar{y})\in\Omega_n} \exp\left((L(\bar{z},y)-\alpha)\cdot w_n(\bar{x},x_0)/\nu\right)\cdot D[\bar{x},\bar{y}]\Big) + r\cdot\nu \leq 0. \end{array} \tag{26}$$

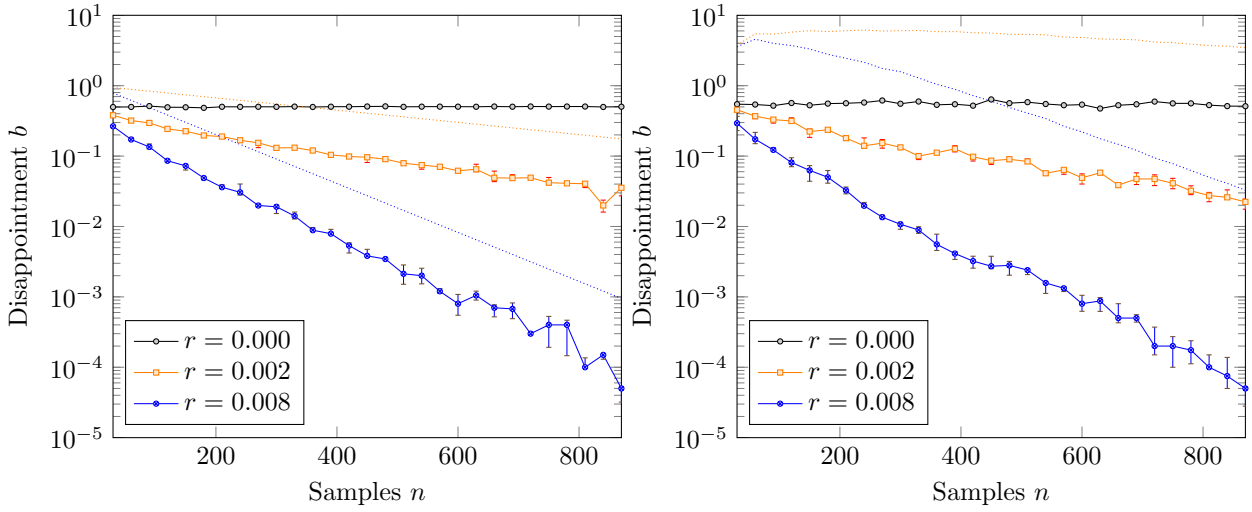
for all $D \in \mathcal{D}_n$.

5 Numerical Examples

We discuss a data-driven news vendor problem in Section 5.1 and a data-driven portfolio allocation problem in Section 5.2. Both of these problems are prescriptive analytics problems stated generally in (2) for a particular loss function L. For both problems, we consider the nominal and bootstrap robust supervised learning formulations discussed in this paper. We briefly discuss first how our supervised learning formulations were solved and trained in practice. All algorithms were implemented in `Julia` discussed developed by **bezanson2017julia**.

The nominal Nadaraya-Watson and nearest neighbors formulations of **bertsimas2014predictive** were implemented with the help of the `Convex` package developed by **udell2014convex**. Taking advantage of the dual representations given in Lemmas 2 and 1, the same procedure was followed for their robust counterparts with respect to the bootstrap distance function as well. The corresponding exponential cone optimization problems were solved numerically with the `ECOS` interior point solver by **domahidi2013ECOS**.

Both the Nadaraya-Watson and nearest neighbors formulations require several hyper parameters such as the smoother function S or the number of neighbors to be learned from data. We considered a Nadaraya-Watson formulation using the Gaussian smoother function given in Figure 2. Likewise, we considered the classical nearest neighbors formulation with the Mahalanobis distance metric $d(m=(x,y),\bar{x}) = (x-\bar{x})^\top \Sigma^{-1}_{\mathrm{tr}[n]}(x-\bar{x})$ based on the empirical variance $\Sigma_{\mathrm{tr}[n]} := \sum_{(\bar{x},\bar{y})\in\mathrm{tr}[n]}(\bar{x}-\mu_{\mathrm{tr}[n]})\cdot(\bar{x}-\mu_{\mathrm{tr}[n]})^\top/n$ and the empirical mean of the auxiliary data $\mu_{\mathrm{tr}[n]} := \sum_{(\bar{x},\bar{y})\in\mathrm{tr}[n]} x/n$. Potential ties among equidistant points were broken based on the method discussed by **gyorfi2006distribution**. The bandwidth parameter $h(n)$ and the number of nearest



(a) Nadaraya-Watson Formulations

(b) Nearest Neighbors Formulations

Figure 4: The empirical bootstrap disappointment b of the Nadaraya-Watson and nearest neighbors formulations in function of the number of samples n. The nominal Nadaraya-Watson and nearest neighbors formulation corresponds to the case $r = 0$. Such nominal formulations do not safeguard against over-calibration as they disappoint on random bootstrap data about half ($b \approx \frac{1}{2}$) the time. The dotted lines visualize the upper bounds concerning the bootstrap disappointment of the bootstrap robust Nadaraya-Watson and nearest neighbors formulation given in Theorem 2 and Theorem 8, respectively. Large deviation theory (**csiszar1984sanov**) indicates that these bootstrap upper bounds and the actual bootstrap disappointments of either formulation drop to zero at the same exponential rate r.

neighbors $k(n)$ were determined based on the squared prediction loss performance of the corresponding Nadaraya-Watson or nearest neighbors predictive learner on ten data sets cross validated from the training data.

5.1 A News Vendor Problem

A company sells a perishable good and needs to make an order $z \in \mathrm{R}$. Ideally, the company would of course like to order exactly $z = y$ where y is the demand of the perishable good. Unfortunately, a decision on the order quantity needs to be made before the demand is observed. Fortunately, however, the company can observe before making the order several covariates $x = x_0$ which may correlate with the uncertain demand. The company may consider the day of the week $w \in \{\text{Monday}, \ldots, \text{Sunday}\}$ to capture weekly cyclical demand, and the outside temperature $t \in \mathrm{R}$ which can influence demand as well. Here, only two covariates are considered where in practice many more covariates may be taken into consideration. For repeated sales, a sensible goal is to order a quantity that minimizes the total expected cost according to

$$z^\star(x_0) \in \arg\inf \quad \mathbb{E}_{D^\star}\left[L(z,y) := b \cdot (y-z)^+ + h \cdot (z-y)^+ \middle| x = x_0\right].$$

The constants $b = 10 \in \mathrm{R}_+$ and $h = 1 \in \mathrm{R}_+$ represent here the marginal cost in dollars of back ordering and holding goods, respectively. If the model distribution $D^\star$ is known, then a classical result states that the optimal decision is then given by the quantile $z^\star(x_0) := \inf\left\{z \;:\; \mathbb{E}_{D^\star}\left[\mathbb{1}\{y \le z\}\middle| x = x_0\right] \ge b/(b+h)\right\}$ of the demand distribution in the covariate context of interest. The classical news vendor formulation assumes the joint distribution $D^\star$ between returns and covariates to be known. In practice however this is almost never the case.

A supervised data version of this news vendor problem is discussed by **rudin2014big**. We consider synthetic training data drawn as independent samples from the synthetic model $D^\star$ with a Gaussian conditional

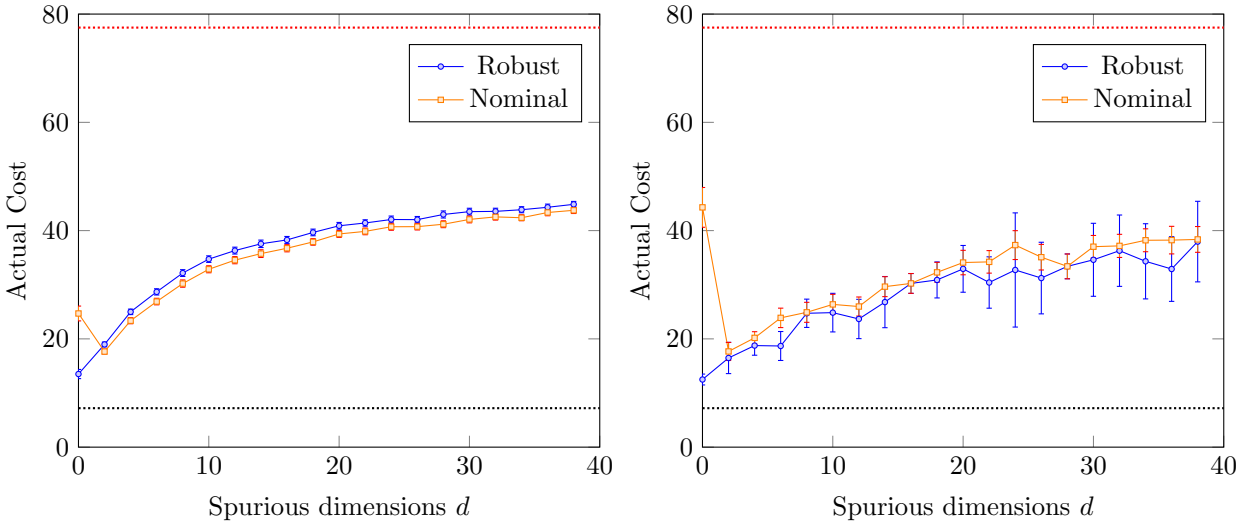
distribution

$$D^\star(x_0 = (\bar{t}, \bar{w})) = N(100 + (\bar{t} - 20) + 20 \cdot \mathbb{I}(\bar{w} \in \{\text{Weekend}\}), 16)$$

and where the day of the week and outside temperature are independent random variables distributed uniformly and normally as $N(20, 4)$, respectively. In our synthetic example, the oracle solution $z^\star(x_0)$ is not linear in x_0 and hence, as discussed in the introduction, the empirical-risk-minimization formulation developed in **rudin2014big** will be biased and hence not directly applicable. We ignored here for the sake of simplicity that the demand typically is not observed directly but needs to be estimated from censored data as discussed for instance in **ferreira2015analytics**. Also for simplicity, we assume that the decision is static in contrast to recent work by **ban2018dynamic**. We shall use this synthetic big data news vendor problem to illustrate the bootstrap disappointment of the robust Nadaraya-Watson and nearest neighbors formulations in a particular context of interest, i.e., $\bar{x} = (\bar{t}, \bar{w}) = (10°\text{C}, \text{Friday})$.

We would like to investigate to what extent our bootstrap robust formulations prevent against overfitting the training data set. Given an action $z^r_{\text{tr}[n]}$ calibrated to this training data set and its budgeted cost $c_n(z^r_{\text{tr}[n]}, D_{\text{tr}[n]}, x_0)$, we approximate its bootstrap disappointment as stated in Definition 1 using a large number $|B| = 20,000$ of bootstrap resamples as suggested in (10). In Figure 4, we present this empirical bootstrap disappointment as a function of the number n of training samples for the nominal and robust Nadaraya-Watson and nearest neighbors formulations. The nominal Nadaraya-Watson of **hannah2010nonparametric** and the nearest neighbors formulation by **bertsimas2014predictive** corresponds to the cases depicted with $r = 0$. Such nominal formulations do not safeguard against over-calibration as they disappoint on random bootstrap data about half the time. The dotted lines visualize the upper bounds concerning the bootstrap disappointment of the bootstrap robust Nadaraya-Watson and nearest neighbors formulation given in Corollary 2 and Theorem 8, respectively. The guarantee in case of the nearest neighbors formulation is not as tight as its Nadaraya-Watson counterpart mostly due to the use of the union bound in the proof of Theorem 8. Nevertheless, large deviation theory via (23) ensures that the empirical bootstrap disappointments and their corresponding theoretical upper bound in either formulation drop to zero at the same exponential rate r.

Working with synthetic data gives us the opportunity to compare the robust and nominal formulations in terms of their true performance. Indeed, as the distribution $D^\star$ is known we can compare the actual cost $\mathbb{E}_{D^\star}\left[L(z_{\text{data}[n]}, y)|x = x_0\right]$ of any proposed decisions $z_{\text{data}[n]}$. We are also interested in seeing how each of the methods fares if we augment the covariates with a number d of irrelevant spurious observations generated from independently sampling a standard normal distribution. It should not come as a surprise that when no spurious covariates are introduced the robust approaches in blue outperform their nominal counterparts in orange significantly as can be seen in Figure 5. The reported cost is the average among $20,000$ random training sets each of length $n = 200$. Here, the amount of robustification r used in either formulation was determined based on ten fold cross validation. Given only $n = 200$ training samples, both the robust Nadaraya-Watson and nearest neighbor formulation do indeed come close to the theoretically minimal cost $\min_z \mathbb{E}_{D^\star}[L(z, y)|x = x_0]$ visualized as the black line for reference. As more spurious covariates are injected in the training data set, the performance of any data-driven method must evidently degrade. Ultimately as d tends to infinity, the noise completely drowns out the signal. The performance of data-driven methods is expected to tend to the optimal cost $\min_z \mathbb{E}_{D^\star}[L(z, y)]$ without using any covariate information and is visualized as the red line. It is curious to observe that the performance of the nominal Nadaraya-Watson and nearest neighbors decisions behave differently than their robust counterparts. Counter-intuitively, the performance of both nominal formulations initially improves with the introduction of spurious covariates. We observed empirically that when no spurious covariates are introduced the nominal cost estimates $\mathrm{E}^n_{D_{\text{tr}[n]}}[L(\bar{z}, y)|x = x_0]$ for all $\bar{z}$ on which the nominal formulations are based tend to overcalibrate the training data. Consequently, the nominal actions suffer poor out-of-sample performance. Our robust formulations do a good job in alleviating this adverse phenomenon without much performance loss. The introduction of artificial noise in the data through spurious covariates seems to provide implicit protection against over-calibration of $\mathrm{E}^n_{D_{\text{tr}[n]}}[L(\bar{z}, y)|x = x_0]$ to the training data by preventing the selection, at least when using cross-validation, of a small bandwidth parameter $h(n)$ or a small number of neighbors $k(n)$ in the Nadaraya-Watson and nearest neighbor formulation, respectively. The addition of spurious covariates has a similar effect as robustification but crucially does comes with a loss of performance.



(a) Nadaraya-Watson Formulations

(b) Nearest Neighbors Formulations

Figure 5: The actual cost $\mathbb{E}_{D^\star}\left[L(z,y)|x=x_0\right]$ of the decisions proposed by the nominal and robust Nadaraya-Watson formulation as well as the nominal and robust nearest neighbors formulation. The reported cost is the average among $20,000$ random training sets each of length $n=200$. The error bars indicate 95% intersample variation. The dimension d reflects the number of spurious covariates introduced. The black line correspond to the optimal full information cost $\min_z \mathbb{E}_{D^\star}\left[L(z,y)|x=x_0\right]$, while the red line represents the optimal no information cost $\min_z \mathbb{E}_{D^\star}\left[L(z,y)\right]$. The robust formulations clearly dominate their nominal counterparts when no spurious covariates ($d=0$) are introduced. When spurious covariates are introduced ($d>0$) the cost predictions $\mathrm{E}^n_{D_{\mathrm{tr}[n]}}\left[L(z,y)|x=x_0\right]$ on which all formulations are based tend to be under-calibrated leading to both a loss of performance and a limited room for improvement through robustification.

5.2 A Portfolio Selection Problem

We consider a small portfolio allocation problem in which a decision $z \in \mathrm{R}^4_+$ needs to be taken in how to split a limited weekly investment budget, i.e., $\mathbb{1}^\top z = 1$, among four stocks $\mathtt{S} := \{\mathtt{Apple}, \mathtt{Airbus}, \mathtt{Boeing}, \mathtt{Facebook}\}$. We decide each Sunday which stocks to buy on Monday and sell on Friday during the subsequent trading week. We obtained weekly trading data for each of the mentioned stocks from Alpha Vantage[1] for all 261 weeks between the 2th of February 2014 and the 27th of January 2019. The weekly returns of each of those stocks may evidently be affected by a large number of covariates. We decided to use the popularity of the search terms {`Cambridge Analytica`, `IBM Redhat`} as obtained from Google Trends[2] as well as the date as potential covariates x. The mentioned covariates may have an indirect impact on the weekly returns $y \in \mathrm{R}^4$ of each of the stocks in our portfolio. A responsible portfolio manager would hence do good to take into account this additional information in the selection of stocks to invest in.

Portfolio selection via optimization has a long history dating back to the pioneering work of **markowitz1991foundations**. A plethora of methods have been developed in the literature which address portfolio selection problems, see **demiguel2007optimal** and references therein. More recently, **ban2016machine** use machine learning to incorporate covariate information when optimization the portfolio selection. The purpose of this section is not to show that the our data-driven formulations compares favorable to all other portfolio selection strategies. Given the amount of prior art in portfolio selection, this seems indeed unlikely to be the case. Rather, we would like to illustrate the value of robustness in the context of real data.

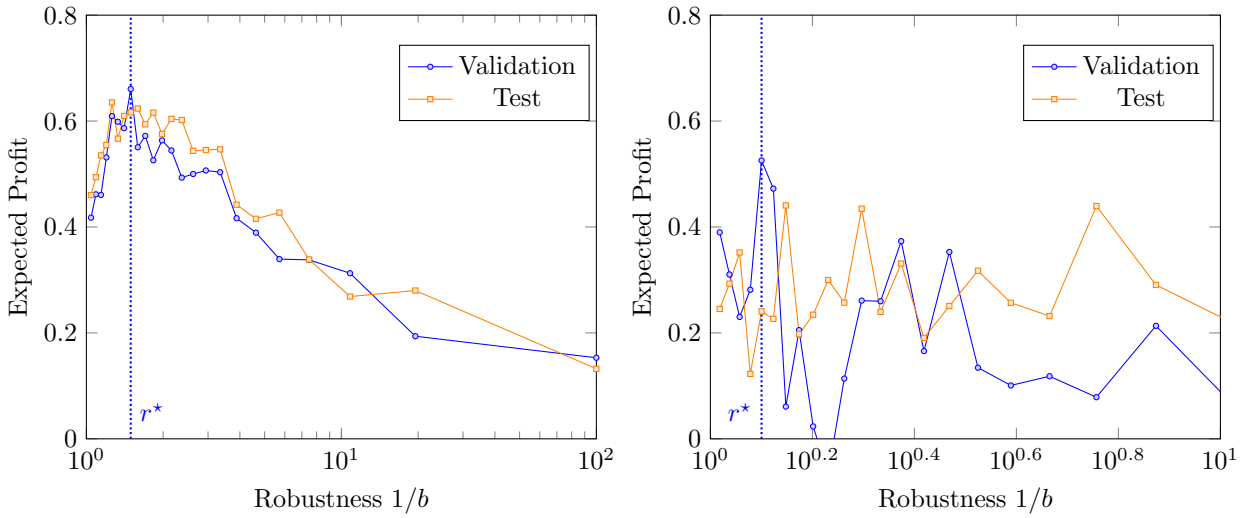
Typically an portfolio selection is made to balance both expected return and some measure of risk. Popular risk measures are the value-at-risk (VAR) and the conditional value-at-risk (CVaR); see for instance **lim2011conditional**. To simplify the exposition here, we imagine a portfolio manager which is a assigned a budget to manage risky investments where the primary objective is high expected returns in the long run. Here, the primary focus is the raw profit $L(z, y) = -z^\top y$ without regard for risk assessment. Ideally, the portfolio manager would like to select

$$\begin{array}{rl} z^\star(x_0) = \arg\min_z & \mathbb{E}_{D^\star}\left[-z^\top y | x = x_0\right] \\ \text{s.t.} & z \in \mathrm{R}^4_+,\ \mathbb{1}^\top z = 1. \end{array}$$

where $D^\star$ is the unknown distribution of the returns and covariates. Let us denote the expected return of each stock as $r(x_0) = \mathbb{E}_{D^\star}[y|x = x_0] \in \mathrm{R}^4$. Evidently, when primarily caring about long term expected profit the manager will assign the entire budget to the stocks $\arg\max_{i \in \mathtt{S}} r_i(x_0)$ with highest expected return.

However, in practice the distribution $D^\star$ is not known and our training data set of 261 historical losses and covariates will have to do instead. We use instead the data-driven formulations we consider in this paper again. To evaluate the performance of a data-driven formulation in view of the fact that the actual expected profits are unknown, we partition the data into $n = 200$ training data points, 31 validation data points and 30 test data point. As market return data is highly noisy, we consider 200 such partitions randomly and report as the validation performance and test performance of any data-driven formulation, its profit averaged over each of these 200 validation and test partitions; see Figure 6. Besides the nominal Nadaraya-Watson and nearest neighbors formulations ($r = 0$), we consider their robust counterparts with $r(n) = 1/n \log(1/b)$ where b ranges between 1 and 0.01. The final $r^\star$ is chosen as to maximize the validation profit. It should first be remarked that Nadaraya-Watson learning seems more appropriate here as it seems both to provide better average profits and is less affected by the highly variable return and search data. From Figure 6 it is clear that robustness in Nadaraya-Watson formulations select portfolios which have better performance on both the validation and crucially the test data sets. We close by remarking that robustness only helps if the nominal formulations are appropriate for the data set. Indeed, the nominal nearest neighbors formulation does not seem to be competitive, and simply adding robustness is not a miracle cure.

[1] https://www.alphavantage.co
[2] https://trends.google.com



(a) Nadaraya-Watson Formulations

(b) Nearest Neighbors Formulations

Figure 6: The average profit of the portfolios on 200 random train-learn-test partitions for both the Nadaraya-Watson and nearest neighbors formulations. From first glance it is clear that the Nadaraya-Watson formulations seem more appropriate here as they seems both to provide better average profits and are less affected by the highly variable return and search data. Adding robustness is most effective when the nominal formulation performs well. Robustness is not a miracle cure when the data-driven formulation is not appropriate as the results for the nearest neighbors formulation illustrate.

6 Conclusion

We discussed in this paper prescriptive analytics problems where cost optimal decisions are to be adapted to a specific covariate context using only supervised data. Our balloon estimation formulation allows for superior context specific decision-making when compared to the naive sample average formulation. As all data-driven methods are prone to adverse overfitting phenomena we must safeguard against over-calibration to one particular training data set. To that end we introduced a novel notion of robustness which guards against overfitting and crucially is itself completely data-driven. Our notion of bootstrap robustness is inspired by the statistical bootstrap, and does not pose any statistical assumption on training data. We derived a novel bootstrap robust balloon estimation formulation which is as tractable as its nominal counterpart based on ideas from distributionally robust optimization. Finally, we have illustrated the benefits of bootstrap robust decisions empirically in terms of their superior out-of-sample performance on two small numerical examples.

Acknowledgments

The second author is generously supported by the Early Post.Mobility fellowship No. 165226 of the Swiss National Science Foundation.

A Proofs

A.1 Proof of Theorem 3

Proof. First note that the domain of the partial estimators as a function of the distribution D satisfies

$$\operatorname{dom} \mathrm{E}_{\mathrm{D}}^{\mathrm{n,j}}\left[\mathrm{L}(\mathrm{z},\mathrm{Y})|\mathrm{x}=\bar{\mathrm{x}}\right] \subseteq \mathcal{D}_{\mathrm{n}}^{\mathrm{j}}.$$

For a distribution D to be in the domain of the partial estimator the constraints in equation (16) must indeed be feasible. In other words, there must exists some P and $s > 0$ for which $s \cdot D[\bar{x}, \bar{y}] = P[\bar{x}, \bar{y}]$ for all $(\bar{x}, \bar{y}) \in \Omega_n$. The last two constraints in equation (16) then imply that any such $D \in \mathcal{D}_n$ must also be in $\mathcal{D}_n^j$.

Any $D \in \mathcal{D}_{n,n}$ is the empirical distribution of some bootstrap data set, say bs$[n]$, consisting of n observations from the training data set. Each set $\mathcal{D}_{n,n}^j := \mathcal{D}_n^j \cap \mathcal{D}_{n,n}$ has a very natural interpretation in terms of $N_n^j(x_0)$ being the smallest neighborhood containing at least k observations of this associated bootstrap data set bs$[n]$. Indeed, $D \in \mathcal{D}_{n,n}^j$ is in terms of the associated data set equivalent to

$$\begin{aligned} k &\leq \textstyle\sum_{(\bar{x},\bar{y}) \in \text{bs}[n]} \mathbb{1}\{(\bar{x}, \bar{y}) \in N_n^j(x_0)\} &&= n \cdot \textstyle\sum_{(\bar{x},\bar{y}) \in N_n^j(x_0)} D[\bar{x}, \bar{y}], \\ k &> \textstyle\sum_{(\bar{x},\bar{y}) \in \text{bs}[n]} \mathbb{1}\{(\bar{x}, \bar{y}) \in N_n^{j-1}(x_0)\} &&= n \cdot \textstyle\sum_{(\bar{x},\bar{y}) \in N_n^{j-1}(x_0)} D[\bar{x}, \bar{y}]. \end{aligned}$$

The first inequality implies that the neighborhood $N_n^j(x_0)$ contains at least k observations of the associated data set. Note that the sets $N_n^j(x_0)$ are increasing with increasing j in terms of set inclusion. The latter inequality hence implies that the biggest smaller neighborhood $N_n^{j-1}(x_0)$ does not contain k observations. Both conditions taken together thus imply that $N_n^j(x_0)$ is the smallest neighborhood which contains at least k samples of the bootstrap data set. Any D in $\mathcal{D}_{n,n}$ is an element in one and only one set $\mathcal{D}_{n,n}^j$ as the smallest neighborhood containing at least k samples is uniquely defined for any data set. Formally, $\mathcal{D}_{n,n}^j \cap \mathcal{D}_{n,n}^{j'} = \emptyset$ for all $j \neq j'$ and moreover $\cup_{j \in [n]} \mathcal{D}_{n,n}^j = \mathcal{D}_{n,n}$. Notice also that the only feasible s in the constraints defining the partial predictors in equation (16) is the particular choice $s = 1/\sum_{(\bar{x},\bar{y}) \in N_n^j(x_0)} w_n(\bar{x}, x_0) \cdot P[\bar{x}, \bar{y}] > 0$. Hence, we must have that for any $D \in \mathcal{D}_{n,n}^j$ the partial estimator equates to

$$\mathrm{E}_D^{n,j}\left[L(\bar{z}, y) | x = x_0\right] = \frac{\sum_{(\bar{x},\bar{y}) \in N_n^j(x_0)} L(\bar{z}, \bar{y}) \cdot w_n(\bar{x}, x_0) \cdot D(\bar{x}, \bar{y})}{\sum_{(\bar{x},\bar{y}) \in N_n^j(x_0)} w_n(\bar{x}, x_0) \cdot D(\bar{x}, \bar{y})} \qquad \forall \bar{z}$$

which is precisely the weighted average over the neighborhood $N_n^j(x_0)$. We have hence for all $\bar{z}$ that

$$\max_{j \in [1,\ldots,n]} \mathrm{E}_D^{n,j}\left[L(\bar{z}, y) | x = x_0\right] = \begin{cases} \dfrac{\sum_{(\bar{x},\bar{y}) \in N_n^1(x_0)} L(\bar{z}, \bar{y}) \cdot w_n(\bar{x}, x_0) \cdot D[\bar{x}, \bar{y}]}{\sum_{(\bar{x},\bar{y}) \in N_n^1(x_0)} w_n(\bar{x}, x_0) \cdot D[\bar{x}, \bar{y}]} & \text{for D} \in \mathcal{D}_{\text{n}}^1, \\ \vdots & \vdots \\ \dfrac{\sum_{(\bar{x},\bar{y}) \in N_n^n(x_0)} L(\bar{z}, \bar{y}) \cdot w_n(\bar{x}, x_0) \cdot D(\bar{x}, \bar{y})}{\sum_{(\bar{x},\bar{y}) \in N_n^n(x_0)} w_n(\bar{x}, x_0) \cdot D(\bar{x}, \bar{y})} & \text{for D} \in \mathcal{D}_{\text{n}}^{\text{n}}. \end{cases}$$

$$= \mathrm{E}_D^n\left[L(\bar{z}, y) | x = x_0\right]$$

as we have argued that $D \in \mathcal{D}_{n,n}^j$ if and only if $N_n^j(x_0)$ is the smallest neighborhood containing at least k data points. As the empirical distribution D and associated data set bs$[n]$ were chosen arbitrary the result follows. □

A.2 Proof of Corollary 1

Proof. Remark that from the definition of the Nadaraya-Watson cost estimate $\mathrm{E}_{D_{\text{tr}[n]}}^n\left[L(z, y) | x = x_0\right] := \mathbb{E}_{D_{\text{tr}[n]}}\left[L(z, y) \cdot w_n(x, x_0)\right] / \mathbb{E}_{D_{\text{tr}[n]}}\left[w_n(x, x_0)\right]$ given in (14) it follows that we have

$$\begin{aligned} \mathrm{E}_{D_{\text{tr}[n]}}^n\left[L(z, y) | x = x_0\right] := \max_{s>0, P} \quad & \textstyle\sum_{(\bar{x},\bar{y}) \in \Omega_n} w_n(\bar{x}, x_0) \cdot L(z, \bar{y}) \cdot P[\bar{x}, \bar{y}] \\ \text{s.t.} \quad & P[\bar{x}, \bar{y}] = D[\bar{x}, \bar{y}] \cdot s \quad \forall (\bar{x}, \bar{y}) \in \Omega_n, \\ & \textstyle\sum_{(\bar{x},\bar{y}) \in \Omega_n} P[\bar{x}, \bar{y}] = s, \ \ \textstyle\sum_{(\bar{x},\bar{y}) \in \Omega_n} w_n(\bar{x}, x_0) \cdot P[\bar{x}, \bar{y}] = 1. \end{aligned}$$

Indeed, the only feasible s is such that $s = 1/\sum_{(\bar{x},\bar{y}) \in \Omega_n} w_n(\bar{x}, x_0) \cdot P[\bar{x}, \bar{y}]$. Hence, the only feasible P is $P = D/\sum_{(\bar{x},\bar{y}) \in \Omega_n} w_n(\bar{x}, x_0) \cdot P[\bar{x}, \bar{y}]$ and the equivalence follows. The chain of equalities

$$\begin{aligned} \left\{(s, P) \ : \ \exists D \text{ s.t. } s \cdot D = P,\, R(D, D_{\text{tr}[n]}) \leq r\right\} &= \left\{(s, P) \ : \ R(P/s, D_{\text{tr}[n]}) \leq r\right\} \\ &= \left\{(s, P) \ : \ s \cdot R(P/s, D_{\text{tr}[n]}) \leq s \cdot r\right\} \end{aligned}$$

imply that the robust budget function $c_n(z, D_{\text{tr}[n]}, x_0)$ corresponds exactly to the optimization formulation claimed in the corollary. □

A.3 Proof of Theorem 5

Proof. We first show the uniform convergence of the robust budget function to its nominal counterpart when the loss function $L(\bar{z}, \bar{y}) < \bar{L} < \infty$ is bounded. Let us first consider a given training data set $\text{tr}[n]$. Note that because of its definition as robust counterpart of the balloon estimator, we have that the robust balloon budget can be bounded as

$$\mathrm{E}^n_{D_{\text{tr}[n]}}\left[L(\bar{z}, y)|x = x_0\right] \leq c_n(\bar{z}, D_{\text{tr}[n]}, x_0) \leq \mathrm{E}^n_{D_{\text{wc}[n]}}\left[L(\bar{z}, y)|x = x_0\right] + \alpha$$

for some worst-case distributions $D_{\text{wc}[n]}$ at distance at most $B(D_{\text{wc}[n]}, D_{\text{tr}[n]}) \leq r(n)$ from the training distribution for any arbitrary $\alpha > 0$. In terms of the total variation distance, we have that $\left\|D_{\text{wc}[n]} - D_{\text{tr}[n]}\right\|_1 \leq \sqrt{B(D_{\text{wc}[n]}, D_{\text{tr}[n]})/2} \leq \sqrt{r(n)/2}$ following Pinkser's inequality.

The Nadaraya-Watson estimate based on the worst-case distribution is the fraction

$$\mathrm{E}^n_{D_{\text{wc}[n]}}\left[L(\bar{z}, y)|x = x_0\right] := \frac{\sum_{(\bar{x},\bar{y})\in\Omega_n} L(\bar{z}, \bar{y}) \cdot w_n(\bar{x}, x_0) \cdot D_{\text{wc}[n]}[\bar{x}, \bar{y}]/h(n)^d}{\sum_{(\bar{x},\bar{y})\in\Omega_n} w_n(\bar{x}, x_0) \cdot D_{\text{wc}[n]}[\bar{x}, \bar{y}]/h(n)^d}$$

where we denote here $d = \dim(x)$ for conciseness. We have that the denumerator of the Nadaraya-Watson estimator is lower bounded by

$$\begin{aligned}
&\textstyle\sum_{(\bar{x},\bar{y})\in\Omega_n} w_n(\bar{x}, x_0)/h(n)^d D_{\text{wc}[n]}[\bar{x}, \bar{y}] \\
=&\textstyle\sum_{(\bar{x},\bar{y})\in\Omega_n} w_n(\bar{x}, x_0)/h(n)^d D_{\text{tr}[n]}[\bar{x}, \bar{y}] + \sum_{(\bar{x},\bar{y})\in\Omega_n} w_n(\bar{x}, x_0)/h(n)^d \left(D_{\text{wc}[n]}[\bar{x}, \bar{y}] - D_{\text{tr}[n]}[\bar{x}, \bar{y}]\right) \\
\geq&\textstyle\sum_{(\bar{x},\bar{y})\in\Omega_n} w_n(\bar{x}, x_0)/h(n)^d D_{\text{tr}[n]}[\bar{x}, \bar{y}] - \left(\max_{(\bar{x},\bar{y})\in\Omega_n} w_n(\bar{x}, x_0)\right) \sqrt{r(n)/2}/h(n)^d \\
\geq&\textstyle\sum_{(\bar{x},\bar{y})\in\Omega_n} w_n(\bar{x}, x_0)/h(n)^d D_{\text{tr}[n]}[\bar{x}, \bar{y}] - \sqrt{r(n)/2}/h(n)^d
\end{aligned}$$

Here the first inequality follows from the Cauchy-Schwartz inequality $\left|a^\top b\right| \leq \|a\|_\infty \cdot \|b\|_1$. Notice that here the weights $0 \leq w_n(\bar{x}, x_0) \leq w_n(x_0, x_0) \leq 1$ are all non-negative and bounded from above by one for all smoother functions in Table 1. Lemma 6 in **walk2010strong** establishes that the limit

$$\liminf_{n\to\infty} \textstyle\sum_{(\bar{x},\bar{y})\in\Omega_n} w_n(\bar{x}, x_0)/h(n)^d D_{\text{data}[n]}[\bar{x}, \bar{y}] = 2d(x_0) > 0$$

is positive with probability one. Taken together with the premisse $\lim_{n\to\infty} \sqrt{r(n)}/h(n)^d = 0$ this establishes the existence of a large enough sample size n_0 such that for all $n \geq n_0$ we have that the denominator of the Nadaraya-Watson estimator satisfies $\sum_{(\bar{x},\bar{y})\in\Omega_n} w_n(\bar{x}, x_0)/h(n)^d D_{\text{wc}[n]}[\bar{x}, \bar{y}] \geq d(x_0)$ and $\sum_{(\bar{x},\bar{y})\in\Omega_n} w_n(\bar{x}, x_0)/h(n)^d D_{\text{tr}[n]}[\bar{x}, \bar{y}] \geq 2\sqrt{r(n)/2}/h(n)^d$. Similarly, the nominator of the Nadaraya-Watson estimator satisfies

$$\begin{aligned}
&\textstyle\sum_{(\bar{x},\bar{y})\in\Omega_n} L(\bar{z}, \bar{y}) \cdot w_n(\bar{x}, x_0) \cdot D_{\text{wc}[n]}[\bar{x}, \bar{y}]/h(n)^d \\
\leq\ &\textstyle\sum_{(\bar{x},\bar{y})\in\Omega_n} L(\bar{z}, \bar{y}) \cdot w_n(\bar{x}, x_0) \cdot D_{\text{tr}[n]}[\bar{x}, \bar{y}]/h(n)^d + \sqrt{r(n)/2}/h(n)^d
\end{aligned}$$

In what follows we will use the inequality $a/(b - x) \leq a/b + 2a/b \cdot x$ for all $x \leq b/2$ when $a, b > 0$. This inequality follows trivially from the definition of convexity of the function $a/(b - x)$ for all $x \leq b$ when

$a, b > 0$. Using the previous inequalities we can establish when $n \geq n_0$ the following claims

$$
\begin{aligned}
&\mathrm{E}^n_{D_{\mathrm{wc}[n]}}\left[L(\bar{z}, y)|x = x_0\right] \\
&\leq \frac{\sum_{(\bar{x},\bar{y})\in\Omega_n} L(\bar{z},\bar{y}) \cdot w_n(\bar{x}, x_0) \cdot D_{\mathrm{tr}[n]}[\bar{x},\bar{y}]/h(n)^d}{\sum_{(\bar{x},\bar{y})\in\Omega_n} w_n(\bar{x}, x_0) \cdot D_{\mathrm{wc}[n]}[\bar{x},\bar{y}]/h(n)^d} + \frac{\sqrt{r(n)/2}}{h(n)^d d(x_0)} \\
&\leq \frac{\sum_{(\bar{x},\bar{y})\in\Omega_n} L(\bar{z},\bar{y}) \cdot w_n(\bar{x}, x_0) \cdot D_{\mathrm{tr}[n]}[\bar{x},\bar{y}]/h(n)^d}{\sum_{(\bar{x},\bar{y})\in\Omega_n} w_n(\bar{x}, x_0) \cdot D_{\mathrm{tr}[n]}[\bar{x},\bar{y}]/h(n)^d - \sqrt{r(n)/2}/h(n)^d} + \frac{\sqrt{r(n)/2}}{h(n)^d d(x_0)} \\
&\leq \mathrm{E}^n_{D_{\mathrm{tr}[n]}}\left[L(\bar{z}, y)|x = x_0\right] + 2\mathrm{E}^n_{D_{\mathrm{tr}[n]}}\left[L(\bar{z}, y)|x = x_0\right]\sqrt{r(n)/2}/h(n)^d + \frac{\sqrt{r(n)/2}}{h(n)^d d(x_0)} \\
&\leq \mathrm{E}^n_{D_{\mathrm{tr}[n]}}\left[L(\bar{z}, y)|x = x_0\right] + (2\bar{L} + 1/d(x_0))\sqrt{r(n)/2}/h(n)^d.
\end{aligned}
$$

The nominal Nadaraya-Watson estimator is uniformly consistent, i.e., $|\mathrm{E}^n_{D_{\mathrm{data}[n]}}\left[L(\bar{z}, y)|x = x_0\right] - \mathbb{E}_{D^\star}\left[L(\bar{z}, y)|x = x_0\right]| \leq \epsilon(n)$ for $\lim_{n\to\infty}\epsilon(n) = 0$ as discussed before. It hence trivially follows that the robust Nadaraya-Watson estimator is uniformly consistent as well. Indeed, from the previous inequality it follows that

$$
|c_n(\bar{z}, D_{\mathrm{data}[n]}, x_0) - \mathbb{E}_{D^\star}\left[L(\bar{z}, y)|x = x_0\right]| \leq \epsilon(n) + (2\bar{L} + 1/d(x_0))\sqrt{r(n)/2}/h(n)^d + \alpha
$$

with probability one for all $\bar{z}$. This inequality holds for any arbitrary $\alpha > 0$. Uniform consistency then directly implies here an asymptotically diminishing optimality gap

$$
\mathbb{E}_{D^\star}\left[L(z^r_{\mathrm{data}[n]}, y)|x = x_0\right] - \min_z \mathbb{E}_{D^\star}\left[L(\bar{z}, y)|x = x_0\right] \leq 2\epsilon(n) + (4\bar{L} + 2/d(x_0))\sqrt{r(n)/2}/h(n)^d.
$$

Using that $\lim_{n\to\infty}\sqrt{r(n)}/h(n)^d$ yields the wanted result immediately. □

A.4 Proof of Theorem 6

Proof. We show the uniform convergence of the robust budget function to the unknown cost, that is, and any bounded function $L(\bar{z}, \bar{y}) < \bar{L} < \infty$ for all $\bar{z}$ and $\bar{y}$. Let us first consider a given training data set $\mathrm{tr}[n]$ without ties. That is, we have that $\left|N^j_n(x_0)\right| = j$ for all $j \in [n]$. Note that because of its definition as robust counterpart of the balloon estimator $\mathrm{E}^n_{D_{\mathrm{tr}[n]}}\left[L(\bar{z}, y)|x = x_0\right]$, we have that the robust cost can be bounded as

$$
\mathrm{E}^n_{D_{\mathrm{tr}[n]}}\left[L(\bar{z}, y)|x = x_0\right] \leq c_n(z^r_{\mathrm{tr}[n]}, D_{\mathrm{tr}[n]}, x_0) \leq \mathrm{E}^{j^\star}_{D_{\mathrm{wc}[n]}}[L(\bar{z}, y)|x = x_0] + \alpha
$$

for some worst-case distributions $D_{\mathrm{wc}[n]} \in \mathcal{D}^{j\star}_n$ at distance at most $B(D_{\mathrm{wc}[n]}, D_{\mathrm{tr}[n]}) \leq r(n)$ from the training distribution for any arbitrary $\alpha > 0$. In terms of the total variation distance, we have that $\left\|D_{\mathrm{wc}[n]} - D_{\mathrm{tr}[n]}\right\|_1 \leq \sqrt{B(D_{\mathrm{wc}[n]}, D_{\mathrm{tr}[n]})/2} \leq \sqrt{r(n)/2}$ following Pinkser's inequality. The nearest-neighbors estimate based on the worst-case distribution is defined as the fraction

$$
\mathrm{E}^n_{D_{\mathrm{wc}[n]}}\left[L(\bar{z}, y)|x = x_0\right] := \frac{\sum_{(\bar{x},\bar{y})\in N^{j^\star}_n(x_0)} L(\bar{z},\bar{y}) \cdot D_{\mathrm{wc}[n]}[\bar{x},\bar{y}]}{\sum_{(\bar{x},\bar{y})\in N^{j^\star}_n(x_0)} D_{\mathrm{wc}[n]}[\bar{x},\bar{y}]}.
$$

The neighborhood parameter $j^\star$ satisfies by definition $\sum_{(\bar{x},\bar{y})\in N^{j^\star}_n(x_0)} D_{\mathrm{wc}[n]}[\bar{x},\bar{y}] \geq k(n)/n$. The previous inequality bounds the denominator from below by $k(n)/n$. Using the Cauchy-Schwartz inequality $\left|a^\top b\right| \leq \|a\|_\infty \cdot \|b\|_1$ as well as the Pinkser inequality, the nominator of the Nadaraya-Watson estimator satisfies

$$
\begin{aligned}
&\sum_{(\bar{x},\bar{y})\in N^{j^\star}_n(x_0)} L(\bar{z},\bar{y}) \cdot D_{\mathrm{wc}[n]}[\bar{x},\bar{y}] \\
\leq\ &\sum_{(\bar{x},\bar{y})\in N^{j^\star}_n(x_0)} L(\bar{z},\bar{y}) \cdot D_{\mathrm{tr}[n]}[\bar{x},\bar{y}] + \bar{L}\sqrt{r(n)/2}.
\end{aligned}
$$

We also have that from the definition of the total variation distance that $\|D_{\mathrm{tr}[n]} - D_{\mathrm{wc}[n]}\|_1 \geq D_{\mathrm{tr}[n]}[N^{j\star}_n(x_0)] - D_{\mathrm{wc}[n]}[N^{j\star}_n(x_0)] \geq (j^\star - k(n))/n$. We have also $\|D_{\mathrm{wc}[n]} - D_{\mathrm{tr}[n]}\|_1 \geq D_{\mathrm{wc}[n]}[N^{j\star-1}_n(x_0)] - D_{\mathrm{tr}[n]}[N^{j\star-1}_n(x_0)] \geq (k(n) - j^\star)/n$. Last two inequalities imply that we can use the bound $\sqrt{r(n)/2} \geq \|D_{\mathrm{tr}[n]} - D_{\mathrm{wc}[n]}\|_1 \geq$

$|k(n) - j^\star| / n$. By applying first the Cauchy-Schwartz inequality again and then the previously obtained bounds we can obtain

$$\begin{aligned} & \textstyle\sum_{(\bar{x},\bar{y})\in N_n^{j^\star}(x_0)} L(\bar{z},\bar{y}) D_{\mathrm{tr}[n]}[\bar{x},\bar{y}] \\ \leq\ & \textstyle\sum_{(\bar{x},\bar{y})\in N_n^{k(n)}(x_0)} L(\bar{z},\bar{y}) D_{\mathrm{tr}[n]}[\bar{x},\bar{y}] + \bar{L} |D_{\mathrm{tr}[n]}\{N_n^{j^\star}(x_0)] - D_{\mathrm{tr}[n]}[N_n^{k(n)}(x_0)]| \\ \leq\ & \textstyle\sum_{(\bar{x},\bar{y})\in N_n^{k(n)}(x_0)} L(\bar{z},\bar{y}) D_{\mathrm{tr}[n]}[\bar{x},\bar{y}] + \bar{L} \left| j^\star - k(n) \right| / n \\ \leq\ & \textstyle\sum_{(\bar{x},\bar{y})\in N_n^{k(n)}(x_0)} L(\bar{z},\bar{y}) D_{\mathrm{tr}[n]}[\bar{x},\bar{y}] + \bar{L}\sqrt{r(n)/2}. \end{aligned}$$

Hence,

$$\mathrm{E}^n_{D_{\mathrm{tr}[n]}}\left[L(\bar{z},y)|x=x_0\right] \leq c_n(\bar{z}, D_{\mathrm{tr}[n]}, x_0) \leq \mathrm{E}^n_{D_{\mathrm{tr}[n]}}\left[L(\bar{z},y)|x=x_0\right] + \frac{2\bar{L}n\sqrt{r(n)/2}}{k(n)} + \alpha \tag{27}$$

for any arbitrary training data set without ties. Ties among data points when using the random tie breaking method are a probability zero event which we may ignore. We already know that the nearest neigbors estimator is uniformly consistent. That is, we have that $|\mathrm{E}^n_{D_{\mathrm{data}[n]}}\left[L(\bar{z},y)|x=x_0\right] - \mathbb{E}_{D^\star}\left[L(\bar{z},y)|x=x_0\right]| \leq \epsilon(n)$ with probability one and $\lim_{n\to\infty}\epsilon(n) = 0$. When the robustness radius does shrinks at an appropriate rate, i.e., its size compared to the bandwidth parameter is negligible ($\lim_{n\to\infty} n\sqrt{r(n)}/k(n) = 0$), then uniform consistency of budget estimator c_n follows by taking the limit for n tends to infinity for the chain of inequalities in (27) applied to $D_{\mathrm{data}[n]}$ and observing that $\alpha > 0$ is arbitrarily small. Uniform consistency of the nearest-neighbors formulation follows by the exact same argument as given in proof of Theorem 5 in case of the Nadaraya-Watson formulation. □

A.5 Proof of Corollary 2

Proof. Let us fix a training data set with empirical distribution $D_{\mathrm{tr}[n]}$ and a given decision z. Let $\bar{c}_n := \mathrm{E}^n_{D_{\mathrm{tr}[n]}}\left[L(\bar{z},y)|x=x_0\right]$ be the budgeted cost based on the training data with $k(n) = n$. In order to prove the theorem, it suffices to characterize the probability of the event that the empirical distribution $D_{\mathrm{bs}[n]}$ of random bootstrap data resampled from the training data realizes in the set

$$\mathcal{C} := \left\{ D \in \mathcal{D}_n \ : \begin{array}{l} \exists s > 0,\ s \cdot \sum_{(\bar{x},\bar{y})\in\Omega_n} w_n(\bar{x},x_0) \cdot L(\bar{z},y) \cdot D(\bar{x},\bar{y}) > \bar{c}_n, \\ s \cdot \sum_{(\bar{x},\bar{y})\in\Omega_n} w_n(\bar{x},x_0) \cdot D(\bar{x},\bar{y}) = 1 \end{array} \right\}$$

as follows from Corollary 1. After eliminating the auxiliary variable s we arrive at the description $\mathcal{C} = \{D \in \mathcal{D}_n : \sum_{(\bar{x},\bar{y})\in\Omega_n} w_n(\bar{x},x_0)\cdot L(\bar{z},y)\cdot D(\bar{x},\bar{y}) > \bar{c}_n \cdot \sum_{(\bar{x},\bar{y})\in\Omega_n} w_n(\bar{x},x_0)\cdot D(\bar{x},\bar{y})\}$. The set $\mathcal{C}$ is a convex polyhedron. The robust budget cost $\bar{c}_n$ is constructed to ensure that $\inf_{D\in\mathcal{C}} R(D, D_{\mathrm{tr}[n]}) > r$. Indeed, we have the rather direct implication $\bar{D} \in \mathcal{C} \Longrightarrow \mathrm{E}^n_{\bar{D}}\left[L(\bar{z},y)|x=x_0\right] > \bar{c}_n = \sup\{\mathrm{E}^n_D\left[L(\bar{z},y)|x=x_0\right] \mid R(D, D_{\mathrm{tr}[n]}) \leq r\}$ which in turn itself implies $R(\bar{D}, D_{\mathrm{tr}[n]}) > r$. Hence, the result follows from the bootstrap inequality (22) applied to the probability $D^\infty_{\mathrm{tr}[n]}(D_{\mathrm{bs}[n]} \in \mathcal{C})$ as in this particular case the employed model distance function ($R = B$) coincides with the bootstrap distance function. □

A.6 Proof Lemma 1

Proof. We will employ standard Lagrangian duality on the convex optimization characterization (19) of the partial nearest neighbors cost function associated $c^j_n(\bar{z}, D, x_0)$. The Lagrangian function associated with the primal optimization problem in (19) is denoted here at the function

$$\begin{aligned} & \mathcal{L}(P, s; \alpha, \beta, \eta, \nu) := \\ & \textstyle\sum_{(\bar{x},\bar{y})\in N_n^j(x_0)} w_n(\bar{x},x_0)\cdot L(z,\bar{y})\cdot P[x,y] + \left(1 - \sum_{(\bar{x},\bar{y})\in N_n^{j-1}(x_0)} w_n(\bar{x},x_0)\cdot P[\bar{x},\bar{y}]\right)\alpha \\ & + \left(\textstyle\sum_{(\bar{x},\bar{y})\in N_n^j(x_0)} P[\bar{x},\bar{y}] - \frac{k}{n}\cdot s\right)\eta_1 + \left(\frac{k-1}{n}\cdot s - \sum_{(\bar{x},\bar{y})\in N_n^{j-1}(x_0)} P[\bar{x},\bar{y}]\right)\eta_2 \\ & + \left(\textstyle\sum_{(\bar{x},\bar{y})\in\Omega_n} P[\bar{x},\bar{y}] - s\right)\beta + \left(r\cdot s - \sum_{(\bar{x},\bar{y})\in\Omega_n} P[\bar{x},\bar{y}]\log\left(\frac{P[\bar{x},\bar{y}]}{s\cdot D[\bar{x},\bar{y}]}\right)\right)\nu \end{aligned}$$

where P and s are the primal variables of the primal optimization problem (19) and α, β, η and ν the dual variables associated with each of its constraints. Collecting the relevant terms in the Lagrangian function results in $\mathcal{L}(P, s; \alpha, \beta, \nu) =$

$$\begin{aligned}
&\alpha + s(r\nu - \beta - \tfrac{k}{n}(\eta_1 - \eta_2) - \tfrac{\eta_2}{n}) \\
&\quad + \textstyle\sum_{(\bar{x},\bar{y}) \in N_n^{j-1}(x_0)} \left[P[\bar{x},\bar{y}] \left((L(z,y) - \alpha) \cdot w_n(\bar{x}, x_0) + \beta + \eta_1 - \eta_2 \right) - \nu P[\bar{x},\bar{y}] \log \left(\tfrac{P[\bar{x},\bar{y}]}{s \cdot D[\bar{x},\bar{y}]} \right) \right] \\
&\quad + \textstyle\sum_{(\bar{x},\bar{y}) \in N_n^{j}(x_0) \setminus N_n^{j-1}(x_0)} \left[P[\bar{x},\bar{y}] \left((L(z,y) - \alpha) \cdot w_n(\bar{x}, x_0) + \beta + \eta_1 \right) - \nu P[\bar{x},\bar{y}] \log \left(\tfrac{P[\bar{x},\bar{y}]}{s \cdot D[\bar{x},\bar{y}]} \right) \right] \\
&\quad + \textstyle\sum_{(\bar{x},\bar{y}) \in \Omega_n \setminus \mathtt{N}_n^{j}(x_0)} \left[P[\bar{x},\bar{y}] \beta - \nu P[\bar{x},\bar{y}] \log \left(\tfrac{P[\bar{x},\bar{y}]}{s \cdot D[\bar{x},\bar{y}]} \right) \right]
\end{aligned}$$

The dual function of the primal optimization problem (19) is identified with the concave function $g(\alpha, \beta, \eta, \nu) := \inf_{P \geq 0,\, s > 0} \mathcal{L}(P, s; \alpha, \beta, \nu)$. Using the same manipulations as presented in the proof of Lemma 2 we can express the dual function as $g(\alpha, \beta, \eta, \nu) =$

$$\begin{aligned}
= \sup_{s>0} \; & \alpha + s(r\nu - \beta - \tfrac{k}{n}(\eta_1 - \eta_2) - \tfrac{\eta_2}{n}) + s\nu \textstyle\sum_{(\bar{x},\bar{y}) \in \Omega_n \setminus N_n^{j}(x_0)} D[\bar{x},\bar{y}] \exp\left(\tfrac{\beta}{\nu} - 1 \right) \\
& + s\nu \textstyle\sum_{(\bar{x},\bar{y}) \in N_n^{j-1}(x_0)} D[\bar{x},\bar{y}] \exp\left(\tfrac{(L(z,\bar{y}) - \alpha) \cdot w_n(\bar{x}, x_0) + \beta + \eta_1 - \eta_2}{\nu} - 1 \right) \\
& + s\nu \textstyle\sum_{(\bar{x},\bar{y}) \in N_n^{j}(x_0) \setminus N_n^{j-1}(x_0)} D[\bar{x},\bar{y}] \exp\left(\tfrac{(L(z,\bar{y}) - \alpha) \cdot w_n(\bar{x}, x_0) + \beta + \eta_1}{\nu} - 1 \right).
\end{aligned}$$

Our dual function can be expressed alternatively as

$$\begin{aligned}
g(\alpha, \beta, \eta, \nu) = \Big\{ \alpha : r \cdot \nu - \tfrac{k}{n}(\eta_1 - \eta_2) - \tfrac{\eta_2}{n} & + \nu \textstyle\sum_{(\bar{x},\bar{y}) \in \Omega_n \setminus N_n^{j}(x_0)} D[\bar{x},\bar{y}] \exp\left(\tfrac{\beta}{\nu} - 1 \right) \\
& + \nu \textstyle\sum_{(\bar{x},\bar{y}) \in N_n^{j-1}(x_0)} D[\bar{x},\bar{y}] \exp\left(\tfrac{(L(z,\bar{y}) - \alpha) \cdot w_n(\bar{x}, x_0) + \beta + \eta_1 - \eta_2}{\nu} - 1 \right) \\
& + \nu \textstyle\sum_{(\bar{x},\bar{y}) \in N_n^{j}(x_0) \setminus N_n^{j-1}(x_0)} D[\bar{x},\bar{y}] \exp\left(\tfrac{(L(z,\bar{y}) - \alpha) \cdot w_n(\bar{x}, x_0) + \beta + \eta_1}{\nu} - 1 \right) \leq \beta \Big\}.
\end{aligned}$$

The dual optimization problem of the primal problem (19) is now found as $\inf_{\alpha,\beta,\nu \geq 0} \; g(\alpha, \beta, \nu)$. As the primal optimization problem in (19) is convex, strong duality holds under Slater's condition which is satisfied whenever $r > r_n^j$. Using first-order optimality conditions, the optimal $\beta^\star$ must satisfy the relationship $\beta^\star = -\nu + \nu \log(\sum_{(\bar{x},\bar{y}) \in N_n^{j-1}(x_0} D[\bar{x},\bar{y}] \exp([(L(z,\bar{y}) - \alpha) \cdot w_n(\bar{x}, x_0) + \eta_1 - \eta_2]/\nu) + \sum_{(\bar{x},\bar{y}) \in N_n^{j}(x_0) \setminus N_n^{j-1}(x_0)} D[\bar{x},\bar{y}] \exp([(L(z,\bar{y}) - \alpha) \cdot w_n(\bar{x}, x_0) + \eta_1]/\nu) + \sum_{(\bar{x},\bar{y}) \in N_n^{j}(x_0) \setminus N_n^{j-1}(x_0)} D[\bar{x},\bar{y}])$. Substituting the optimal value of $\beta^\star$ in the back in the dual optimization problem gives

$$\begin{aligned}
\inf_{\alpha,\beta,\nu \geq 0} g(\alpha, \beta, \eta, \nu) &= \inf_{\alpha,\nu \geq 0} \; g(\alpha, \beta^\star, \eta, \nu) \\
&= \inf \Big\{ \alpha \in \mathrm{R} : \exists \nu \in \mathrm{R}_+, \exists \eta \in \mathrm{R}_+^2, \; r \cdot \nu - \tfrac{k}{n}(\eta_1 - \eta_2) - \tfrac{\eta_2}{n} \cdot \nu \\
&\quad + \nu \log(\textstyle\sum_{(\bar{x},\bar{y}) \in N_n^{j-1}(x_0)} \exp([(L(z,\bar{y}) - \alpha) \cdot w_n(\bar{x}, x_0) + \eta_1 - \eta_2]/\nu) \cdot D[\bar{x},\bar{y}] \\
&\quad + \textstyle\sum_{(\bar{x},\bar{y}) \in N_n^{j}(x_0) \setminus N_n^{j-1}(x_0)} \exp([(L(z,\bar{y}) - \alpha) \cdot w_n(\bar{x}, x_0) + \eta_1]/\nu) \cdot D[\bar{x},\bar{y}] \\
&\quad + \textstyle\sum_{\Omega_n \setminus N_n^{j}(x_0)} D[\bar{x},\bar{y}]) \leq 0 \Big\}.
\end{aligned}$$

□

A.7 Proof Lemma 2

Proof. We will employ standard Lagrangian duality on the convex optimization characterization of the Nadaraya-Watson cost function given in Corollary 1. The Lagrangian function associated with the primal optimization problem is denoted here at the function

$$\begin{aligned}
\mathcal{L}(P, s; \alpha, \beta, \nu) := \textstyle\sum_{(\bar{x},\bar{y}) \in \Omega_n} w_n(\bar{x}, x_0) \cdot L(\bar{z},\bar{y}) \cdot P[\bar{x},\bar{y}] + \left(1 - \textstyle\sum_{(\bar{x},\bar{y}) \in \Omega_n} w_n(\bar{x}, x_0) \cdot P[\bar{x},\bar{y}] \right) \alpha + \\
\left(\textstyle\sum_{(\bar{x},\bar{y}) \in \Omega_n} P[\bar{x},\bar{y}] - s \right) \beta + \left(r \cdot s - \textstyle\sum_{(\bar{x},\bar{y}) \in \Omega_n} P[\bar{x},\bar{y}] \log \left(\tfrac{P[\bar{x},\bar{y}]}{s \cdot D[\bar{x},\bar{y}]} \right) \right) \nu
\end{aligned}$$

where P and s are the primal variables of the primal optimization problem given in Corollary 1 and α, β and ν the dual variables associated with each of its constraints. Collecting the relevant terms in the Lagrangian function results in

$$\mathcal{L}(P,s;\alpha,\beta,\nu) = \alpha + s(r\nu-\beta) + \textstyle\sum_{(\bar{x},\bar{y})\in\Omega_n}\left[P[\bar{x},\bar{y}]\left((L(\bar{z},\bar{y})-\alpha)\cdot w_n(\bar{x},x_0)+\beta\right) - \nu P[\bar{x},\bar{y}]\log\left(\frac{P[\bar{x},\bar{y}]}{s\cdot D[\bar{x},\bar{y}]}\right)\right]$$

The dual function of the primal optimization problem is identified with the concave function $g(\alpha,\beta,\nu) := \inf_{P\geq 0,\, s>0}\mathcal{L}(P,s;\alpha,\beta,\nu)$. Our dual function can be expressed alternatively as $g(\alpha,\beta,\nu) =$

$$\begin{aligned}
&\sup_{s>0}\ \alpha + s\left(r\nu-\beta\right) + \sup_{P\geq 0}\sum_{(\bar{x},\bar{y})\in\Omega_n}\left[P[\bar{x},\bar{y}]\left((L(\bar{z},\bar{y})-\alpha)\cdot w_n(\bar{x},x_0)+\beta\right) - \nu P[\bar{x},\bar{y}]\log\left(\frac{P[\bar{x},\bar{y}]}{s\cdot D[\bar{x},\bar{y}]}\right)\right]\\
=&\sup_{s>0}\ \alpha + s\left(r\nu-\beta\right) + \sum_{(\bar{x},\bar{y})\in\Omega_n}\sup_{P[\bar{x},\bar{y}]\geq 0}\left[P[\bar{x},\bar{y}]\left((L(\bar{z},\bar{y})-\alpha)\cdot w_n(\bar{x},x_0)+\beta\right) - \nu P[\bar{x},\bar{y}]\log\left(\frac{P[\bar{x},\bar{y}]}{s\cdot D[\bar{x},\bar{y}]}\right)\right]\\
=&\sup_{s>0}\ \alpha + s\left(r\nu-\beta\right) + s\textstyle\sum_{(\bar{x},\bar{y})\in\Omega_n} D[\bar{x},\bar{y}][\sup_{\lambda\geq 0}\ \lambda\left((L(\bar{z},\bar{y})-\alpha)\cdot w_n(\bar{x},x_0)+\beta\right) - \nu\lambda\log\left(\lambda\right)].
\end{aligned}$$

The inner maximization problems over λ can be dealt with using the Fenchel conjugate of the $\lambda\mapsto\lambda\cdot\log\lambda$ function as

$$\begin{aligned}
&=\sup_{s>0}\ \alpha + s\left(r\nu-\beta\right) + s\nu\textstyle\sum_{(\bar{x},\bar{y})\in\Omega_n} D[\bar{x},\bar{y}]\exp\left(\frac{(L(\bar{z},\bar{y})-\alpha)\cdot w_n(\bar{x},x_0)+\beta}{\nu}-1\right)\\
&=\left\{\alpha\ :\ r\nu+\nu\textstyle\sum_{(\bar{x},\bar{y})\in\Omega_n} D[\bar{x},\bar{y}]\exp\left(\frac{(L(\bar{z},\bar{y})-\alpha)\cdot w_n(\bar{x},x_0)+\beta}{\nu}-1\right)\leq\beta\right\}.
\end{aligned}$$

The dual optimization problem is now found as $\inf_{\alpha,\beta,\nu\geq 0}\ g(\alpha,\beta,\nu)$. As our primal optimization is convex, strong duality holds under Slater's condition which is satisfied whenever $r>0$. Using first-order optimality conditions, the optimal $\beta^\star$ must satisfy $\beta^\star = -\nu+\nu\log(\sum_{(\bar{x},\bar{y})\in\Omega_n} D[\bar{x},\bar{y}]\exp((L(\bar{z},\bar{y})-\alpha)\cdot w_n(\bar{x},x_0)/\nu))$. Substituting the optimal value of $\beta^\star$ in the back in the dual optimization problem gives

$$\begin{aligned}
\inf_{\alpha,\beta,\nu\geq 0} g(\alpha,\beta,\nu) &= \inf_{\alpha,\nu\geq 0}\ g(\alpha,\beta^\star,\nu)\\
&= \inf\left\{\alpha\in\mathrm{R}\ :\ \exists\nu\in\mathrm{R}_+,\ r\nu+\nu\log\left(\textstyle\sum_{(\bar{x},\bar{y})\in\Omega_n} D[\bar{x},\bar{y}]\exp\left(\frac{(L(\bar{z},\bar{y})-\alpha)\cdot w_n(\bar{x},x_0)}{\nu}\right)\right)\leq 0\right\}.
\end{aligned}$$

□