# Data-Driven Approximation of Contextual Chance-Constrained Stochastic Programs

Hamed Rahimian*[1] and Bernardo Pagnoncelli†[2]

[1]Department of Industrial Engineering, Clemson University, Clemson SC 29634, USA
[2]SKEMA Business School, Université Côte dAzur, Lille, France

**Abstract**

Uncertainty in classical stochastic programming models is often described solely by independent random parameters, ignoring their dependence on multidimensional features. We describe a novel contextual chance-constrained programming formulation that incorporates features, and argue that solutions that do not take them into account may not be implementable. Our formulation cannot be solved exactly in most cases, and we propose a tractable and fully data-driven approximate model that relies on weighted sums of random variables. We obtain a stochastic lower bound for the optimal value and feasibility results that include convergence to the true feasible set as the number of data points increases, as well as the minimal number of data points needed to obtain a feasible solution with high probability. We illustrate our findings in a vaccine allocation problem and compare the results with a naïve sample average approximation approach.

**Keywords:** Chance constraints, Data-driven optimization, Stochastic programming, Large deviations

## 1 Introduction

Traditional stochastic programming models, such as two-stage and chance-constrained problems, describe uncertainty using a random variable (or vector) with a known probability distribution. Such randomness is the only description of the uncertainty in those problems. By solving them, one obtains a solution with the smallest expected cost or satisfies the problem's constraints with some prescribed probability. The distribution of the random vector is often estimated or approximated by data. In this paper, we argue that auxiliary information related to such random variables, the so-called *features*, should be taken into account as well. As an example, consider a company that sells widgets. The demand on weekdays follows a different pattern than on weekends. Moreover, past data shows that inflation, unemployment, and interest rates affect demand as well. When building a model to decide how many widgets to order, it is desirable to consider the current value of each of those features if the company wants to maximize revenue or minimize stockouts.

Given the wider availability of data, and since predictive models have reached maturity, the attention of researchers in the optimization community has been shifting towards prescriptive analytics. Very recently, several publications [3, 6, 8, 9, 10, 16, 19, 23, 24, 30, 37, 39] have pointed

*hrahimi@clemson.edu
†bernardo.pagnoncelli@uai.cl

out that in the context of data science and decision-making under uncertainty, it is fundamental to incorporate the features associated to the random parameters within the optimization formulation. The basic assumption of data science is that there is value in the data: several organizations have harnessed it, and were able to increase their business value by using data-driven methodologies [1, 18, 27, 36, 40, 46]. Many applications of decision-making under uncertainty that include contextual information have recently appeared in the literature, including inventory theory [7, 35], sales team assignments [31, 54], price and revenue management [28], health care [26, 47], and load planning in rail transportation [33].

There are two main approaches in the literature to incorporate contextual information into those problems. The first is empirical risk minimization (ERM), which is discussed in [3] and is a popular approach in machine learning [4, 21]. In ERM, it is necessary to assume a functional form for the response function given the features, and a common choice is a linear one. The resulting problem is usually tractable, and once the problem is solved a course of action is immediately available for any given feature. However, as pointed out in [6], ERM cannot be applied to constrained problems, and in some situations, the linear approximation may not be a good choice.

The second approach is to use weight functions that give different importance to each past observation. There are several ways of achieving that: the first is to use kernel smoothing functions, which can measure the similarity between feature vectors. In [3], the authors build a model based on past feature vectors and demand observations, and a kernel function determines the weight of each data point according to its similarity with respect to the currently observed feature vector. The authors also consider a regularized kernel approach and compare the different models in a nurse staffing problem. In [11], the authors propose two novel contextual methods that use robust kernel formulations to guarantee better out-of-sample performance by protecting against overfitting.

The weights can also be constructed using other classical machine learning (ML) methods. For instance, in $k$ Nearest Neighbors ($k$-NN), each neighbor of the currently observed feature vector will have a weight of $1/k$, while all other data points will have zero weight. Similarly, in classification and regression trees (CART), nonzero weights will be given to data points in the tree's leaf to which the current feature vector belongs. In [6], the authors compare all those methods to solve a real-world problem of the distribution arm of a media conglomerate. They show that by including a very diverse set of features an 88% improvement can be obtained, measured by their proposed coefficient of prescriptiveness. In [20], the authors propose a stochastic quasi-gradient approach to solve contextual problems approximated via kernels or $k$-NN.

The previously cited papers mainly consider contextual extensions to two-stage stochastic programs. The focus of our paper is on chance-constrained programming (CCP). CCP was introduced in [15], and it represents a very popular class of problems that still attracts the attention of researchers in both methodological, see, e.g., [29, 41], and practical contexts, see, e.g., [51, 57].

In the unidimensional case, chance constraints are equivalent to Value-at-Risk constraints, which are widely used in economics and finance [17, 58]. Moreover, CCP is a natural formulation to describe reliability requirements in engineering problems such as water management [2, 52], energy [56], supply chain [50], among others. In the vast majority of cases, no deterministic equivalent form is available, making it impossible to solve the problem directly. Even evaluating if a given solution is feasible through an oracle is a challenging problem, which is often distribution-dependent. The work [43] established stability results for CCP problems, which allowed the decision-maker to control for perturbations to the original—sometimes not completely known—distribution of the problem. A standard method to approximate CCP is the sample average approximation (SAA), which consists of obtaining samples from the distribution of the random parameters and solving an easier approximate problem. Consistency results show that the optimal value and optimal solutions of the approximate problem converge to their optimal deterministic counterparts as the

sample size goes to infinity [34, 38]. There are several other approaches to solve CCP problems that use nonlinear programming techniques, e.g., [41, 49, 53, 55]. Numerical results are superior to SAA in some cases, although many of those methods assume the random parameters follow elliptical distributions.

Our first contribution in this paper is a novel formulation for CCP problems that includes contextual information, referred to as *contextual CCP* problems. For two-stage stochastic programs, ignoring features will generate a feasible solution that may have a significant gap with respect to the true optimal, or to another solution obtained via contextual methods. In contextual CCP problems, the situation is more dramatic: a solution that ignores features may end up being infeasible when the current state of the world is taken into account. We show through a simple portfolio example with one feature that ignoring it can lead to infeasibility when the solution is implemented.

Our second contribution is to approximate contextual CCP problems via weight functions and establish theoretical results that quantify the solution quality. We first show that the optimal value of the approximate contextual CCP problem is a lower bound to the true problem, with high probability as the number of data points increases. We also demonstrate a feasibility result, proving that a feasible solution to the approximate contextual CCP problem is feasible to the true problem, with high probability as the number of data points increases. We then provide estimates for the number of data points needed in the approximate problem to yield a lower bound and feasibility for the true problem. For all weight functions considered we have probabilistic guarantees with an exponential convergence as the number of data points increases.

We present a detailed computational study of a vaccine allocation problem with real-world features. We compare our results with the naïve SAA approach, which ignores features and only uses the samples from the random variables of the problem. We show that in several cases the naïve SAA approach does not find feasible solutions to the true problem, and that convergence is absent as the number of data points grows. For the contextual case, feasibility is quickly achieved, and we observe convergence to a feasible solution as the number of data points increases.

The rest of this paper is outlined as follows. In Section 2, we present CCPs and review some background information. In Section 3, we formally present a contextual CCP and present methods to approximate this problem in a data-driven fashion. In Section 4, we provide theoretical results on how data-driven approximations of contextual CCP problems relate to the true problem in terms of the optimal value and feasibility of resulted solutions. We then present numerical experiments in Section 5. Finally, we end with conclusions in Section 6.

*Notation:* We use $\mathbb{1}\{Z\}$ to denote the indicator function which takes value one when $Z$ holds and zero otherwise. Let $\mathcal{N}_\eta(u)$ denote the open ball with center $u$ and radius $\eta$. We let $[n]$ denote the index set $\{1,\dots,n\}$. A random variable $Z$ is said to be sub-Gaussian with variance proxy $\sigma^2$ if $\mathbb{E}[Z] = 0$ and $\mathbb{E}[\exp\{tZ\}] \leq \exp\{\frac{t^2\sigma^2}{2}\}$ for all $t \in \mathbb{R}$. Function $u \mapsto Z(u)$ is $L$-Lipschitz if there exists $L > 0$ such that $|Z(u) - Z(u')| \leq L\|u - u'\|_p \ \forall u, u'$ and for some $p \geq 0$. A family of functions $\{Z_n\}$ is equicontinuous on a set $\mathcal{U}$ if for every $\varepsilon > 0$, there exists a $\eta > 0$ such that $|Z_n(u) - Z_n(u')| < \varepsilon$ whenever $\|u - u'\| < \delta$, $u, u' \in \mathcal{U}$, and $n \in \mathbb{N}$. For a bounded set $\mathcal{U} \subseteq \mathbb{R}^m$, the diameter is defined as $\theta = \sup\{\|u - u'\| \mid u, u' \in \mathcal{U}\}$.

## 2 Background on Chance-Constrained Programming

Consider a CCP problem as

$$\begin{array}{ll} \min\limits_{u \in \mathcal{U}} & f(u) \\ \text{s.t.} & P\{G(u, Y) \leq 0\} \geq \epsilon, \end{array} \tag{CCP}$$

where the feasible set $\mathcal{U} \subset \mathbb{R}^{d_u}$ is nonempty and compact, $Y$ is a random vector defined on a probability space $(\mathcal{Y}, \mathcal{F}, P)$ with $\mathcal{Y} \subset \mathbb{R}^{d_y}$, $f : \mathbb{R}^{d_u} \to \mathbb{R}$ is a deterministic objective function, and $\epsilon \in (0, 1]$ is the reliability level associated with the chance constraint. Finally, $G : \mathbb{R}^{d_u} \times Y \to \mathbb{R}$ is a random function; that is, $G(u, \cdot)$ is measurable for $u \in \mathbb{R}^{d_u}$. Note that when there are multiple constraints $G_1(u, Y) \leq 0, \ldots, G_\kappa(u, Y) \leq 0$, our framework still applies by defining $G(\cdot, Y)$ as the maximum of $G_1(\cdot, Y), \ldots, G_\kappa(\cdot, Y)$ to convert it into an individual chance constraint.

With the exception of very particular cases, general (CCP) problems are challenging to solve since the feasible region can be nonconvex, and approximations are needed. Given a sequence of identically distributed (not necessarily independent) observations $\{y^i\}_{i\in[n]}$ from $Y$, we can construct an empirical probability distribution

$$\hat{P}_n := \frac{1}{n} \sum_{i\in[n]} \delta_{y^i}, \tag{1}$$

where $\delta_{y^i}$ is the Dirac point mass on $y^i$, $i \in [n]$. Using the empirical probability distribution, we can obtain an empirical approximation

$$p_n(u) := \hat{P}_n \{G(u, Y) \leq 0\} = \mathbb{E}_{\hat{P}^n} [\mathbb{1} \{G(u, Y) \leq 0\}] = \frac{1}{n} \sum_{i\in[n]} \mathbb{1} \left\{G(u, y^i) \leq 0\right\}. \tag{2}$$

The SAA problem is obtained by replacing the true probability

$$p(u) := P \{G(u, Y) \leq 0\} = \mathbb{E}_P [\mathbb{1} \{G(u, Y) \leq 0\}]$$

by the approximated one in (2) as

$$\begin{array}{ll} \min\limits_{u\in\mathcal{U}} & f(u) \\ \text{s.t.} & p_n(u) \geq \alpha, \end{array} \tag{SAA-CCP}$$

where $\alpha \in (0, 1]$ is the reliability level of the approximate problem, which in principle could be different from $\epsilon$ in (CCP).

Convergence results for the SAA formulation are established in [34, 38] for the general case, and in [13, 14] for $\alpha = 0$ in the formulation (SAA-CCP). Typical results show an exponential rate of convergence of the optimal value and set of optimal solutions to their deterministic counterparts.

## 3 Contextual Chance-Constrained Programming

It is often the case that the random parameters $Y$ exhibit some dependence on a vector of features. It may be beneficial to include those features in the problem formulation since they contain relevant information that helps to explain the outcomes of the random vector $Y$. Specifically, given a random observation $X = x$, the contextual chance-constrained programming problem (C-CCP) can be written as

$$z^*_\epsilon(x) = \begin{array}{ll} \min\limits_{u\in\mathcal{U}} & f(u) \\ \text{s.t.} & P \{G(u, Y) \leq 0 \mid X = x\} \geq \epsilon. \end{array} \tag{C-CCP}$$

In (C-CCP), we let $(X, Y)$ be defined on a probability space $(\mathcal{X} \times \mathcal{Y}, \mathcal{F}, P)$, where $\mathcal{X} \subset \mathbb{R}^{d_x}$ is a compact set and $\mathcal{X} \subset \mathbb{R}^{d_y}$. Moreover, the chance constraint is calculated with respect to the conditional probability of $Y$ given $X = x$. For each $X = x$, we assume that $z^*_\epsilon(x)$ exists and is finite, and a solution $u^*(x)$ to problem (C-CCP) gives the best response to the observed feature vector $x$ as measured by the objective function $f(\cdot)$.

Observe that formulation (C-CCP) is at least as hard to solve as (CCP). In order to solve (C-CCP), one needs to know the conditional probability of $Y$ given $X = x$. Even when such a distribution is known, approximation schemes need to be considered to deal with the chance constraint in (C-CCP).

Given a sequence of identically distributed observations $\mathcal{D}_n := \{(x^i, y^i)\}_{i\in[n]}$ from $(X, Y)$ and an observation $X = x$, we construct a data-driven approximation of (C-CCP). In order to do so, we form a weight function $w_n^i(x, \mathcal{D}_n)$, $i \in [n]$, such that $\sum_{i\in[n]} w_n^i(x, \mathcal{D}_n) = 1$ and $w_n^i(x, \mathcal{D}_n) \geq 0$, $i \in [n]$. This weight function is then used to approximate the conditional distribution of $Y$ given $X = x$ as

$$\hat{P}_n := \sum_{i\in[n]} w_n^i(x, \mathcal{D}_n)\delta_{y^i}, \tag{3}$$

that measures "proximity" of each data point $i$ with respect to the observed feature $x$. Note that for simplicity in notation, we dropped the dependence of $\hat{P}_n$ to $X = x$ and $\mathcal{D}_n$. The rationale of this weight function is that data points that are "close" to the current observation $X = x$ are more valuable to accurately estimate the distribution of $Y$ given $X = x$. Using the approximated probability distribution (3), we can obtain an approximated expression for the probability as

$$\begin{aligned} m_n(u; x) &:= \hat{P}_n \{G(u, Y) \leq 0\} = \mathbb{E}_{\hat{P}^n} [\mathbb{1} \{G(u, Y) \leq 0\}] \\ &= \sum_{i\in[n]} w_n^i(x, \mathcal{D}_n)\mathbb{1} \left\{G(u, y^i) \leq 0\right\}. \end{aligned} \tag{4}$$

A data-driven contextual CCP formulation (DDC-CCP) can be obtained by replacing

$$m(u; x) := P \{G(u, Y) \leq 0 \mid X = x\} = \mathbb{E}_P [\mathbb{1} \{G(u, Y) \leq 0\} \mid X = x] \tag{5}$$

by the approximation (4) as

$$\begin{aligned} \hat{z}_{n,\alpha}(x) = \quad & \min_{u\in\mathcal{U}} \quad f(u) \\ & \text{s.t.} \quad m_n(u; x) \geq \alpha. \end{aligned} \tag{DDC-CCP}$$

We adopt the convention that if the feasible set of (DDC-CCP) is empty, we have $\hat{z}_{n,\alpha}(x) = +\infty$.

We will construct the weights using classical machine learning methods, such as CART and RF. For CART, it is assumed that the algorithm utilizes a splitting rule $\mathcal{C} : \mathcal{X} \to \{1, \ldots, c\}$, which induces a partition of the feature set $\mathcal{X} = \mathcal{C}^{-1}(1) \sqcup \ldots \sqcup \mathcal{C}^{-1}(c)$. For RF, we have rules $\mathcal{C}^1, \ldots, \mathcal{C}^T$ for each of the $T$ trees constructed. We describe the weights for a subset of nonparametric estimation approaches in Table 1 and refer the readers to [25] for a comprehensive treatment. In Section 4, we discuss further details on the construction of the trees.

Table 1: Weights obtained via nonparametric estimation approaches.

| Algorithm | $w_n^i(x, \mathcal{D}_n)$ |
|---|---|
| $k$-NN | $\frac{1}{k}\mathbb{1}\left\{x^i \text{ is a } k\text{-NN of } x\right\}$ |
| CART | $\frac{1}{\lvert\{i\in[n] \mid \mathcal{C}(x^i)=\mathcal{C}(x)\}\rvert}\mathbb{1}\left\{\mathcal{C}(x) = \mathcal{C}(x^i)\right\}$ |
| RF | $\frac{1}{T}\sum_{t\in[T]} \frac{1}{\lvert\{i\in[n] \mid \mathcal{C}^t(x^i)=\mathcal{C}^t(x)\}\rvert}\mathbb{1}\left\{\mathcal{C}^t(x) = \mathcal{C}^t(x^i)\right\}$ |

Some observations regarding problem (DDC-CCP) are in order. First, one can interpret the output of formulation (DDC-CCP) as a response function or a policy: given an observation of the

feature vector $X = x$, a solution $u_n^*(x)$ represents the best response measured by the objective function $f(\cdot)$. Second, the computational burden of solving (DDC-CCP) is comparable to solving the SAA formulation (SAA-CCP) since the computation of the weights can be done offline.

Before showing the convergence results for (DDC-CCP), in the next section we show a simple portfolio example that motivates the need to consider contextual information in CCP problems in practice.

### 3.1 Motivating Example: A Value-at-Risk Toy Problem

Consider the following two-dimensional portfolio problem, where an investor with one dollar seeks to maximize her returns by investing in stocks and bonds subject to a Value-at-Risk constraint. If short sales are not allowed, the problem can be written as

$$\begin{array}{ll} \max\limits_{u \in \mathcal{U}} & (1+\bar{y})^\top u \\ \text{s.t.} & P\left\{(1+Y)^\top u \geq \gamma \mid X = x\right\} \geq \epsilon, \end{array} \tag{6}$$

where $\bar{y} = (0.0145, 0.0083)$ represents the average monthly returns of stocks and bonds, $\gamma = 0.8$ is a minimum desired return, $\mathcal{U} = \{u \in \mathbb{R}^2 \mid u_1 + u_2 = 1, u_1 \geq 0, u_2 \geq 0\}$, and $\epsilon = 0.9$. The random vector $(1+Y)$ follows a bivariate normal distribution with mean $(1+\bar{y})$ and covariance matrix that depends on a one-dimensional feature $x \in \{0, 1\}$ as follows:

$$\Sigma(x) = \begin{cases} \begin{pmatrix} 0.02900 & 0.02051 \\ 0.02051 & 0.01819 \end{pmatrix} & \text{if } x = 0, \\ \begin{pmatrix} 0.04799 & 0.02051 \\ 0.02051 & 0.02859 \end{pmatrix} & \text{if } x = 1. \end{cases}$$

The case $x = 0$ represents normal market conditions, while $x = 1$ captures a more volatile setting expressed by higher variances for both assets. The covariance between the assets is the same in both settings.

We construct a synthetic dataset mixing returns from the normal ($x = 0$) with the volatile ($x = 1$) case. We solved the SAA problem, similar to (SAA-CCP), with 10,000 samples as described, and obtained an optimal solution

$$\hat{u} := (\hat{u}_1, \hat{u}_2) = (0.57, 0.43), \tag{7}$$

with an optimal value of 1.01186. The SAA problem is completely myopic to the presence of features, and in practice, one can often observe market conditions before investing.

To test the robustness of the SAA solution, we can compute for this simple problem the exact value of the probability in the chance constraint for the optimal solution under each market condition. When $x = 0$, the optimal value of the true problem (6) is 1.014, and the SAA solution (7) is feasible and performs well. However, when $x = 1$, the SAA solution (7) is infeasible for the true problem (6) as

$$P\left\{(1+Y)^\top \hat{u} \geq \gamma \mid X = 1\right\} = 0.885 < \epsilon = 0.9.$$

This example shows that one must aim at approximating the conditional distribution instead of using simple SAA when contextual information is available.

# 4 Finite Dataset Guarantees of (DDC-CCP)

In this section, we present theoretical results that support the use of (DDC-CCP) to approximate (C-CCP). In particular, we investigate the consistency of the optimal value and feasibility of the data-driven solution in Sections 4.2 and 4.3, respectively.

A standard technique used to prove convergence results for traditional stochastic programming problems is large deviations theory, and in particular concentration inequalities. For two-stage stochastic programs, the main reference is [32], and for CCP problems we refer the reader to [34], which serves as inspiration and point of comparison for the results we derive in this paper. Unlike the featureless case, we now have non-constant weights that are not independent of the random variable $Y$, so Hoeffding's inequality cannot be applied in this context. Our strategy is to first assume the existence of a Hoeffding-like result, and then show that some popular choices of weight functions satisfy this assumption for different choices of hyperparameters. These sufficient conditions are presented in section 4.1 for $k$-NN, CART, and RF.

Before we start the exposition, we introduce some notation. Let $\mathcal{E}_n := \{(X^1, Y^1), \ldots, (X^n, Y^n)\}$ and $(X, Y)$ be a collection of independently and identically distributed (i.i.d.) random vectors. Also, let $P^n$ denote the sampling distribution of $\mathcal{E}_n$, i.e., the n-fold product distribution of $P$, and $\mathbb{E}_{P^n}[\cdot]$ denote the corresponding expectation operator. For some $u \in \mathcal{U}$, define random variables $Z(u) := \mathbb{1}\{G(u, Y) \leq 0\}$ and $Z^i(u) := \mathbb{1}\{G(u, Y^i) \leq 0\}$, $i \in [n]$. Also, let $W_n^i(x, \mathcal{E}_n)$, $i \in [n]$, be a random weight function and define

$$M_n(u; x) := \mathbb{E}_{P^n}[Z(u) \mid X = x] = \sum_{i \in [n]} W_n^i(x, \mathcal{E}_n) Z^i(u). \tag{8}$$

By (5), recall that

$$m(u; x) = \mathbb{E}_P[Z(u) \mid X = x]. \tag{9}$$

Note that $m_n(u; x)$, defined in (4), is a realization of the random variable $M_n(u; x)$, calculated based on observations $\mathcal{D}_n$ of $\mathcal{E}_n$. Let $\mathcal{U}_\epsilon(x)$ and $\mathcal{U}_{n,\alpha}(x)$ denote the feasible region to (C-CCP) and (DDC-CCP), respectively, for $X = x$ and given $\mathcal{E}_n$. That is,

$$\mathcal{U}_\epsilon(x) := \{u \in \mathcal{U} \,|\, m(u; x) \geq \epsilon\},$$

and

$$\mathcal{U}_{n,\alpha}(x) := \{u \in \mathcal{U} \,|\, M_n(u; x) \geq \alpha\}.$$

We note that $\mathcal{U}_{n,\alpha}(x)$ is random and depends on $\mathcal{E}_n$.

## 4.1 Consistency of Estimation: Sufficient Conditions

To develop the theoretical consistency results for the optimal value and feasibility of the data-driven solution to (C-CCP), we make some assumptions on

$$R_n(x) := \sup_{u \in \mathcal{U}} |M_n(u; x) - m(u; x)|, \tag{10}$$

where $M_n(u; x)$ and $m(u; x)$ are defined in (8) and (9), respectively.

Our first assumption is that for almost every (a.e.) $x \in \mathcal{X}$, $M_n(u; x)$ converges in probability to $m(u; x)$, uniformly over $\mathcal{U}$ and with an exponential rate. That is, for a.e. $x \in \mathcal{X}$, $R_n(x)$ converges to zero in probability.

**Assumption 1.** *(Pointwise consistency) Given the weight function $W_n^i(x, \mathcal{E}_n)$, $i \in [n]$, there exist constants $A(\kappa, x)$ and $B(\kappa, x)$ such that*

$$P^n \{R_n(x) \geq \kappa\} \leq A(\kappa, x) \exp\{-a(n)B(\kappa, x)\},$$

*for any $\kappa > 0$ and a.e. $x \in \mathcal{X}$, where $R_n(x)$ is defined in (10) and $a(n) \to \infty$ as $n \to \infty$.*

In addition, we assume that $\{R_n\}$ is stochastically equicontinuous on $\mathcal{X}$ with an exponential rate, see, e.g., conditions of [42, Theorem 10.2]. We also refer to [44, Definition 7.22] for the equicontinuity of deterministic functions.

**Assumption 2.** *(Stochastic equicontinuity) Given the weight function $W_n^i(x\mathcal{E}_n)$, $i \in [n]$, there exist constants $A'(\kappa)$ and $B'(\kappa)$ such that*

$$P^n \left\{ \sup_{x \in \mathcal{X}} \sup_{x' \in \mathcal{N}_\eta(x)} |R_n(x) - R_n(x')| \geq \kappa \right\} \leq A'(\kappa) \exp\{-b(n)B'(\kappa)\},$$

*for any $\kappa > 0$ and some small enough $\eta > 0$, where $R_n(x)$ is defined in (10) and $b(n) \to \infty$ as $n \to \infty$.*

It is known that for deterministic functions, pointwise convergence and equicontinuty yield uniform convergence on a compact set, e.g., see [44, Excercise 7.16]. The following lemma is a stochastic generalization of the uniform probabilistic guarantee; it becomes possible with the pointwise probabilistic guarantee, Assumption 1, and the stochastic equicontinuity, Assumption 2, on the compact set $\mathcal{X}$.

**Lemma 1.** *(Uniform consistency) Given the weight function $W_n^i(x, \mathcal{E}_n)$, $i \in [n]$, suppose that Assumptions 1 and 2 hold. Then, there exist constants $C(\kappa)$ and $D(\kappa)$ such that*

$$P^n \left\{ \sup_{x \in \mathcal{X}} R_n(x) \geq \kappa \right\} \leq C(\kappa) \exp\{-c(n)D(\kappa)\},$$

*for any $\kappa > 0$, where $R_n(x)$ is defined in (10) and $c(n) \to \infty$ as $n \to \infty$.*

*Proof.* Consider $\kappa > 0$. Let $\{x_1, x_2, \ldots, x_r\}$ be a collection of points in $\mathcal{X}$ such that $\mathcal{X} \subset \bigcup_{k \in [r]} \mathcal{N}_\eta(x_k)$, where $\eta > 0$ makes Assumption 2 hold. We note that such a finite cover exists by the compactness of $\mathcal{X}$. Then, the triangle inequality,

$$\sup_{x \in \mathcal{X}} R_n(x) \leq \max_{k \in [r]} R_n(x_k) + \sup_{k \in [r], x \in \mathcal{N}_\eta(x_k)} |R_n(x) - R_n(x_k)|,$$

implies that

$$\{\sup_{x \in \mathcal{X}} R_n(x) < \kappa\} \supseteq \{\max_{k \in [r]} R_n(x_k) < \kappa/2\} \cap \{\sup_{k \in [r], x \in \mathcal{N}_\eta(x_k)} |R_n(x) - R_n(x_k)| < \kappa/2\}, \tag{11}$$

Hence, by an application of the union bound to (11), we have

$$\begin{aligned} P^n \left\{ \sup_{x \in \mathcal{X}} R_n(x) \geq \kappa \right\} \leq{}& P^n \{\exists\, k \in [r] \text{ such that } R_n(x_k) \geq \kappa/2\} \\ &+ P^n \left\{ \sup_{k \in [r], x \in \mathcal{N}_\eta(x_k)} |R_n(x) - R_n(x_k)| \geq \kappa/2 \right\} \end{aligned} \tag{12}$$

$$
\begin{aligned}
&\leq \sum_{k\in[r]} P^n\left\{R_n(x_k) \geq \kappa/2\right\} \\
&\quad + P^n\left\{\sup_{k\in[r], x\in\mathcal{N}_\eta(x_k)} |R_n(x) - R_n(x_k)| \geq \kappa/2\right\} \\
&\leq \sum_{k\in[r]} A(\kappa/2, x_k)\exp\{-a(n)B(\kappa/2, x_k)\} \\
&\quad + A'(\kappa/2)\exp\{-b(n)B'(\kappa/2)\},
\end{aligned}
$$

where the second inequality comes from an application of the union bound to the first term in the right-hand side of (12). Moreover, the last inequality is due to Assumptions 1 and 2. Now, the result follows by taking $c(n) = \min\{a(n), b(n)\}$, $C(\kappa) := \max\{A'(\kappa/2), \max_{k\in[r]} A(\kappa/2, x_k)\}$, and $D(\kappa) := \min\{B'(\kappa/2), \min_{k\in[r]} B(\kappa/2, x_k)\}$. □

As we shall shortly discuss, Assumptions 1 and 2 (and hence, Lemma 1) are strong in the sense that they enable us to obtain pointwise and uniform consistency results on the optimal value and feasibility of the data-driven solution over the feature space $\mathcal{X}$ (see Theorems 1, 2, 3). In what follows, we discuss sufficient conditions that either imply Assumptions 1 and 2, or the conclusion Lemma 1, for several weight functions. In particular, we discuss construction of weight functions based on $k$-NN, CART, and RF, in Definitions 1, 2, 3, respectively. We refer the reader to [25] for a comprehensive treatment of these nonparametric estimation approaches, and to [22] for $k$-NN, and to [12, 45] for CART and RF. We then discuss sufficient conditions for Assumption 1 for $k$-NN (Lemma 2 and Remark 1), CART (Lemma 3 and Remark 2), and RF (Lemma 4 and Remark 3). Finally, in Lemma 5, we provide a sufficient condition for Assumption 2.

**Definition 1.** *Consider random vectors $(X, Z), (X^1, Z^1), \ldots, (X^n, Z^n) \in \mathbb{R}^{d_x} \times \mathbb{R}$ and let $k_n$ be a deterministic parameter such that $k_n \to \infty$ and $k_n/n \to 0$ as $n \to \infty$. Then, $k_n$ nearest neighbors of $(X, Z)$ are chosen as follows:*

- *Points $(X^i, Z^i)$, $i \in [n]$, are ordered in a nondecreasing sequence based on the distance between $X$ and $X^i$ using $\ell_p$-norm,*
- *$k_n$ nearest neighbors of $(X, Z)$ are chosen with ties broken randomly,*

*to construct a $k_n$-NN estimator $\mu_n(X) = \sum_{i\in[n]} \frac{1}{k}\mathbb{1}\left\{X^i \text{ is a } k_n\text{-NN of } X\right\} Z^i$.*

We state Lemma 2 from [9, Lemma 10] on the uniform consistency of $k_n$-NN.

**Lemma 2.** *Consider i.i.d. random vectors $(X, Z), (X^1, Z^1), \ldots, (X^n, Z^n) \in \mathbb{R}^{d_x} \times \mathbb{R}$. Suppose that $\mathcal{X} \subset [0, 1]^{d_x}$ and there exists $\vartheta > 0$ such that $P\{X \in \mathcal{N}_\gamma(x)\} > \vartheta\gamma^{d_x}$ for all $x \in \mathcal{X}$ and $\gamma > 0$. Furthermore, suppose that $Z - \mathbb{E}_P[Z \mid X = x]$ is conditionally sub-Gaussian given $X = x$ with variance proxy $\sigma^2$, uniformly for all $x \in \mathcal{X}$, and $\mu(x) := \mathbb{E}_P[Z \mid X = x]$ is $L$-Lipschitz. Let $\mu_n(x)$ denote a $k_n$-NN estimator, constructed based on Definition 1. Then,*

$$
\begin{aligned}
P^n\left\{\sup_{x\in\mathcal{X}} |\mu_n(x) - \mu(x)| \geq \kappa\right\} \leq \left(\frac{4\sqrt{d_x}\varphi L}{\kappa}\right)^{d_x} \exp\left\{-\frac{2}{n}\left(n\vartheta\left(\frac{\kappa}{4L}\right)^{d_x} + 1 - k_n\right)^2\right\} \\
+ 2\left(\frac{25}{d_x}\right)^{d_x} \exp\left\{-\left(\frac{k_n\kappa^2}{8\sigma^2} - 2d_x\log(n)\right)\right\},
\end{aligned}
$$

*for $\kappa \geq 4L\left(\frac{k_n - 1}{n\vartheta}\right)^{1/d_x}$ and $n \geq 2d_x$, where $\varphi > 0$ is a constant that depends on $\ell_p$-norm used in the construction of $\mu_n(x)$.*

**Remark 1.** *Suppose that the weight function $W_n^i(x, \mathcal{E}_n)$, $i \in [n]$, is formed using a $k_n$-NN estimator based on Definition 1, with $k_n = \lceil cn^\gamma \rceil$ for $\gamma \in (0,1)$ and $c > 0$ such that $k_n \leq n-1$. Then, under assumptions of Lemma 2, $C(\kappa)$ and $D(\kappa)$ in Lemma 1 can be specified as $C(\kappa) := \max\left\{ \left(\frac{\mathcal{O}(1)\sqrt{d_x}}{\kappa}\right)^{d_x}, \mathcal{O}(1)\left(\frac{\mathcal{O}(1)}{d_x}\right)^{d_x} \right\}$, and $D(\kappa) := \min\left\{ \mathcal{O}(1)\big(\mathcal{O}(1)\kappa\big)^{2d_x}, \min_{n\in\mathbb{N}}\left\{ \mathcal{O}(1)\frac{n^{\gamma-1}\kappa^2}{\sigma^2} - \mathcal{O}(1)d_x\frac{\log(n)}{n} \right\} \right\}$, where $n \geq \mathcal{O}(1)\left(\frac{\mathcal{O}(1)}{\kappa}\right)^{\frac{d_x}{1-\gamma}}$ and $\frac{n^\gamma}{\log(n)} \geq \frac{d_x\sigma^2}{\kappa^2}$. Moreover, Lemma 1 holds with $c(n) = n$. We also recall that pointwise consistency is a trivial consequence of uniform consistency. Hence, Assumption 1 holds with $A(\kappa; x) = C(\kappa)$, $B(\kappa; x) = D(\kappa)$, and $a(n) = c(n)$.*

We now discuss the construction of CART and conditions under which Assumption 1 holds.

**Definition 2.** *Let $\mathcal{X} = [0,1]^{d_x}$. Consider random vectors $(X, Z), (X^1, Z^1), \ldots, (X^n, Z^n) \in \mathbb{R}^{d_x} \times \mathbb{R}$ and let $k_n$ be a deterministic parameter such that $\log(k_n) \to \infty$ and $k_n/n \to 0$ as $n \to \infty$. Starting from a parent node $[0,1]^{d_x}$, a recursive partitioning operates by repeatedly conducting the following procedure, $\lceil \log_2(k_n) \rceil$ times:*

- *select an unsplit node $\nu$, and suppose that $\varpi_j$ is the width of coordinate $j$, $j \in [d_x]$,*
- *randomly select a splitting coordinate $j \in [d_x]$, where all coordinates are equiprobable to be selected and the selected coordinate is independent of $(X, Z), (X^1, Z^1), \ldots, (X^n, Z^n)$,*
- *split node $\nu$ into two children nodes $\nu \cap \{x : x_j \leq \varpi_j/2\}$ and $\nu \cap \{x : x_j > \varpi_j/2\}$,*

*to construct a CART estimator as $\mu_n(X) = \sum_{i\in[n]} \frac{\mathbb{1}\{X^i \in A_n(X)\}}{N_n(X)} \mathbb{1}\{N_n(X) > 0\} Z^i$, where $A_n(X)$ denotes the unique leaf node containing $X$ and $N_n(X)$ is the number of data points falling in the same leaf node as $X$, i.e., $N_n(X) = \sum_{i\in[n]} \mathbb{1}\{X^i \in A_n(X)\}$.*

Several remarks are in order about Definition 2. First, the independence of the coordinate selected to split and $(X, Z), (X^1, Z^1), \ldots, (X^n, Z^n)$ excludes any data-dependent strategy to build the tree, such as those obtained by optimizing some criterion on the dataset. Second, by construction, the tree has $2^{\lceil \log_2(k_n) \rceil} \approx k_n$ leaf nodes, and each leaf node has Lebesgue measure $2^{-\lceil \log_2(k_n) \rceil} \approx 1/k_n$. Thus, if $X$ is uniformly distributed on $[0,1]^{d_x}$, there is about $n/k_n$ observations per leaf node, on average.

**Lemma 3.** *Consider i.i.d. random vectors $(X, Z), (X^1, Z^1), \ldots, (X^n, Z^n) \in \mathbb{R}^{d_x} \times \mathbb{R}$. Suppose that $X$ is uniformly distributed on $\mathcal{X} = [0,1]^{d_x}$. Furthermore, suppose that there exists $\sigma^2 > 0$ such that $\mathrm{Var}_P[Z \mid X = x] \leq \sigma^2$, uniformly for all $x \in \mathcal{X}$, and $\mu(x) = \mathbb{E}_P[Z \mid X = x]$ is L-Lipschitz. Let $\mu_n(x)$ denote a CART estimator with $k_n = \lfloor cn^\gamma \rfloor$ for $\gamma \in (0,1)$ and $c > 0$, constructed based on Definition 2. Then,*

$$
\begin{aligned}
P^n\{|\mu_n(x) - \mu(x)| \geq \kappa\} \leq & \frac{16\sigma^2}{\kappa^2} \exp\left\{ -c(1-\gamma)\log(n) \right\} \\
& + \frac{8L^2}{\kappa^2} d_x \exp\left\{ -\frac{3\gamma c}{4d_x \log(2)} \log(n) \right\} \\
& \frac{4}{\kappa^2}\left( \sigma^2 + \sup_{x\in[0,1]^{d_x}} \big(\mu(x)\big)^2 \right) \exp\left\{ -\frac{(1-\gamma)}{2c}\log(n) \right\},
\end{aligned}
$$

*for any $\kappa > 0$ and a.e. $x \in \mathcal{X}$.*

*Proof.* Consider a fixed $X = x$. Let $N_n(x)$ be the number of data points falling in the same leaf node as $x$, i.e., $N_n(x) = \sum_{i\in[n]} \mathbb{1}\left\{X^i \in A_n(x)\right\}$. Hence, $W_n^i(x) = \frac{\mathbb{1}\left\{X^i\in A_n(x)\right\}}{N_n(x)} \mathbb{1}\left\{N_n(x)>0\right\}$ (we drop the dependence on $\mathcal{E}_n$ for simplicity). By construction, $\mu_n(x) = \sum_{i\in[n]} W_n^i(x) Z^i$. Moreover, let $\hat{\mu}_n(x) := \mathbb{E}_{P^n}\left[\mu_n(x) \mid X^1, \ldots, X^n\right] = \sum_{i\in[n]} W_n^i(x)\mathbb{E}_P\left[Z \mid X = X^i\right] = \sum_{i\in[n]} W_n^i(x)\mu(X^i)$. Note that

$$\mu_n(x) - \mu(x) = \big(\mu_n(x) - \hat{\mu}_n(x)\big) + \big(\hat{\mu}_n(x) - \mu(x)\big).$$

Given that $|\mu_n(x) - \mu(x)| \le |\mu_n(x) - \hat{\mu}_n(x)| + |\hat{\mu}_n(x) - \mu(x)|$, we have

$$\{|\mu_n(x) - \mu(x)| < \kappa\} \supseteq \{|\mu_n(x) - \hat{\mu}_n(x)| < \kappa/2\} \cap \{|\hat{\mu}_n(x) - \mu(x)| < \kappa/2\}.$$

Hence, by an application of the union bound and due to $P\left\{|X| \ge a\right\} = P\left\{|X|^2 \ge a^2\right\}$, we have

$$\begin{aligned} P^n\left\{|\mu_n(x) - \mu(x)| \ge \kappa\right\} \le\ & P^n\left\{|\mu_n(x) - \hat{\mu}_n(x)|^2 \ge \kappa^2/4\right\} \\ & + P^n\left\{|\hat{\mu}_n(x) - \mu(x)|^2 \ge \kappa^2/4\right\}. \end{aligned} \tag{13}$$

To bound the right-hand side of (13), first, we bound the bias $\mathbb{E}_{P^n}\left[|\hat{\mu}_n(x) - \mu(x)|^2\right]$ and the variance $\mathbb{E}_{P^n}\left[|\mu_n(x) - \hat{\mu}_n(x)|^2\right]$. Then, by two applications of Markov's inequality, we derive a bound on $P^n\left\{|\mu_n(x) - \mu(x)| \ge \kappa\right\}$.

Observe that by the bias and variance decomposition, we have

$$\mathbb{E}_{P^n}\left[|\mu_n(x) - \mu(x)|^2\right] = \mathbb{E}_{P^n}\left[|\mu_n(x) - \hat{\mu}_n(x)|^2\right] + \mathbb{E}_{P^n}\left[|\hat{\mu}_n(x) - \mu(x)|^2\right],$$

as

$$\begin{aligned} & \mathbb{E}_{P^n}\left[\big(\mu_n(x) - \hat{\mu}_n(x)\big)\big(\hat{\mu}_n(x) - \mu(x)\big)\right] \\ =& \mathbb{E}_{P^n}\left[\mathbb{E}_{P^n}\left[\big(\mu_n(x) - \hat{\mu}_n(x)\big)\big(\hat{\mu}_n(x) - \mu(x)\big) \mid X^1, \ldots, X^n\right]\right] \\ =& \mathbb{E}_{P^n}\left[\big(\hat{\mu}_n(x) - \mu(x)\big)\mathbb{E}_{P^n}\left[\mu_n(x) - \hat{\mu}_n(x) \mid X^1, \ldots, X^n\right]\right] \\ =& \mathbb{E}_{P^n}\left[\big(\hat{\mu}_n(x) - \mu(x)\big)\big(\mathbb{E}_{P^n}\left[\mu_n(x) \mid X^1, \ldots, X^n\right] - \hat{\mu}_n(x)\big)\right] \\ =& \mathbb{E}_{P^n}\left[\big(\hat{\mu}_n(x) - \mu(x)\big)\big(\hat{\mu}_n(x) - \hat{\mu}_n(x)\big)\right] \\ =& 0. \end{aligned}$$

To bound the bias term, we follow a similar argument to that of [12, Proposition 4], albeit differences in the setup, and the fact that [12, Proposition 4] bounds the expected $|\hat{\mu}_n(X) - \mu(X)|^2$ for a random $X$ and with respect to the joint distribution of $(X, Z), (X^1, Z^1), \ldots, (X^n, Z^n)$. We skip the details for brevity. We can then conclude that

$$\begin{aligned} \mathbb{E}_{P^n}\left[|\hat{\mu}_n(x) - \mu(x)|^2\right] \le & 2L^2 d_x \exp\left\{-\frac{3\gamma c}{4d_x \log(2)}\log(n)\right\} \\ & + \sup_{x\in[0,1]^{d_x}} \big(\mu(x)\big)^2 \exp\left\{-\frac{1-\gamma}{2c}\log(n)\right\}. \end{aligned} \tag{14}$$

To bound the variance term $\mathbb{E}_{P^n}\left[|\mu_n(x) - \hat{\mu}_n(x)|^2\right]$, we note that

$$\mathbb{E}_{P^n}\left[|\mu_n(x) - \hat{\mu}_n(x)|^2\right] = \mathbb{E}_{P^n}\left[\Big(\sum_{i\in[n]} W_n^i(x)\big(Z^i - \mu(X^i)\big)\Big)^2\right]$$

$$
\begin{aligned}
&= \sum_{i\in[n]}\sum_{j\in[n]} \mathbb{E}_{P^n}\left[W_n^i(x)W_n^j(x)\big(Z^i-\mu(X^i)\big)\big(Z^j-\mu(X^j)\big)\right] \\
&= \sum_{i\in[n]} \mathbb{E}_{P^n}\left[\big(W_n^i(x)\big)^2\big(Z^i-\mu(X^i)\big)^2\right] \\
&= \sum_{i\in[n]} \mathbb{E}_{P^n}\Big[\mathbb{E}_{P^n}\Big[\big(W_n^i(x)\big)^2\big(Z^i-\mu(X^i)\big)^2\big|\{X^j:\forall j\},\{Z^j:j\neq i\}\Big]\Big] \\
&\leq \sum_{i\in[n]} \mathbb{E}_{P^n}\left[\sigma^2\big(W_n^i(x)\big)^2\right] \\
&= \sigma^2\mathbb{E}_{P^n}\left[\sum_{i\in[n]}\big(W_n^i(x)\big)^2\right],
\end{aligned}
$$

where the inequality follows from the hypothesis on the conditional variance of $Z$, and the third equality is true because, for $i \neq j$, we have

$$
\begin{aligned}
&\mathbb{E}_{P^n}\left[W_n^i(x)W_n^j(x)\big(Z^i-\mu(X^i)\big)\big(Z^j-\mu(X^j)\big)\right] \\
=&\mathbb{E}_{P^n}\left[\mathbb{E}_{P^n}\left[W_n^i(x)W_n^j(x)\big(Z^i-\mu(X^i)\big)\big(Z^j-\mu(X^j)\big)\mid X^1,\ldots,X^n,Z^i\right]\right] \\
=&\mathbb{E}_{P^n}\left[W_n^i(x)W_n^j(x)\big(Z^i-\mu(X^i)\big)\mathbb{E}_{P^n}\left[\big(Z^j-\mu(X^j)\big)\mid X^1,\ldots,X^n,Z^i\right]\right] \\
=&\mathbb{E}_{P^n}\left[W_n^i(x)W_n^j(x)\big(Z^i-\mu(X^i)\big)\big(\mu(X^j)-\mu(X^j)\big)\right] \\
=&0.
\end{aligned}
$$

Now, note that $\sum_{i\in[n]}\big(W_n^i(x)\big)^2 = 0$ if $N_n(x)=0$, and $\sum_{i\in[n]}\big(W_n^i(x)\big)^2 = \frac{1}{N_n(x)}$ if $N_n(x)>0$. Let $\mathcal{A}_{n,j}$ denote the $j$-th rectangular leaf node of the tree and $N_{n,j}$ denote the number of data points, among $x, X^1, \ldots X^n$, falling in that leaf node, $j = 1, \ldots, 2^{\lceil \log_2(k_n)\rceil}$. Then, we have

$$
\begin{aligned}
&\mathbb{E}_{P^n}\left[\sum_{i\in[n]}\big(W_n^i(x)\big)^2\right] \\
\leq& P^n\{N_n(x)=0\} + \sum_{j=1}^{2^{\lceil\log_2(k_n)\rceil}} \mathbb{E}_{P^n}\left[\mathbb{1}\{x\in\mathcal{A}_{n,j}\}\frac{1}{N_{n,j}}\mathbb{1}\{N_{n,j}>0\}\right] \\
=& P^n\{N_n(x)=0\} + \sum_{j=1}^{2^{\lceil\log_2(k_n)\rceil}} \mathbb{E}_{P^n}\left[\mathbb{E}_{P^n}\left[\frac{1}{N_{n,j}}\mathbb{1}\{N_{n,j}>0\}\;\Big|\; x\in\mathcal{A}_{n,j}\right]\right] \\
\leq& P^n\{N_n(x)=0\} + \sum_{j=1}^{2^{\lceil\log_2(k_n)\rceil}} P\{x\in\mathcal{A}_{n,j}\}\frac{2}{nP\{x\in\mathcal{A}_{n,j}\}} \qquad (15)\\
=& P^n\{N_n(x)=0\} + \frac{2}{n}2^{\lceil\log_2(k_n)\rceil} \\
\leq& \left(1-2^{-\lceil\log_2(k_n)\rceil}\right)^n + \frac{2}{n}2^{\lceil\log_2(k_n)\rceil} \qquad (16)\\
\leq& \exp\left\{-\frac{n}{2k_n}\right\} + \frac{4k_n}{n} \\
\leq& \exp\Big\{-\frac{1-\gamma}{2c}\log(n)\Big\} + \frac{4k_n}{n}, \qquad (17)\\
\leq& \exp\Big\{-\frac{1-\gamma}{2c}\log(n)\Big\} + 4\exp\{-c(1-\gamma)\log(n)\}, \qquad (18)
\end{aligned}
$$

where (15) comes from the fact that conditioned on $x \in \mathcal{A}_{n,j}$, $N_{n,j}$ has a Binomial distribution with parameters $n$ and $P\{x \in \mathcal{A}_{n,j}\}$ [12, Fact 2] and the fact that $\mathbb{E}_{P^n}\left[\frac{1}{N_{n,j}}\mathbb{1}\{N_{n,j}>0\} \mid x \in \mathcal{A}_{n,j}\right] \leq \frac{2}{(n+1)P\{x\in\mathcal{A}_{n,j}\}}$ [25, Lemma 4.1]. Moreover, (16) is true because of the fact that $N_n(x)$ has a Binomial distribution with parameters $n$ and $2^{-\lceil \log_2(k_n)\rceil}$ [12, Fact 2]. Finally, (17) comes from the facts that $\exp\{-\frac{n}{2k_n}\} \leq \exp\{-\frac{n}{2cn^\gamma}\}$ and $\log(a) \leq a$, whereas (18) comes from the facts that $\frac{4k_n}{n} \leq \frac{4cn^\gamma}{n}$ and $a = \exp\{\log(a)\}$. Thus, we have

$$\mathbb{E}_{P^n}\left[|\mu_n(x) - \hat{\mu}_n(x)|^2\right] \leq \sigma^2 \exp\left\{-\frac{1-\gamma}{2c}\log(n)\right\} + 4\sigma^2 \exp\{-c(1-\gamma)\log(n)\}. \tag{19}$$

Now, putting (14) and (19) together, combined with two applications of Markov's inequality (see (13)), yield the pointwise consistency bound $P^n\{|\mu_n(x) - \mu(x)| \geq \kappa\}$ for $X = x$. □

**Remark 2.** *Suppose that the weight function $W_n^i(x, \mathcal{E}_n)$, $i \in [n]$, is formed using a CART estimator based on Definition 2, with $k_n = \lfloor cn^\gamma \rfloor$ for $\gamma \in (0,1)$ and $c > 0$. Then, under the hypothesis of Lemma 3, the quantities in Assumption 1 can be specified as $a(n) = \log(n)$, $A(\kappa, x) := \frac{1}{\kappa^2}\max\left\{\mathcal{O}(1)\sigma^2, \mathcal{O}(1)d_x, \mathcal{O}(1)\left(\sigma^2 + \sup_{x\in[0,1]^{d_x}}\left(\mu(x)\right)^2\right)\right\}$, and $B(\kappa, x) := \min\left\{\mathcal{O}(1)(1-\gamma), \frac{\mathcal{O}(1)\gamma}{d_x}, \frac{1-\gamma}{\mathcal{O}(1)}\right\}$.*

So far, we discussed the construction of $k$-NN and CART, and conditions under which Assumption 1 hold for them. We now turn our attention to RF, an ensemble of CARTs.

**Definition 3.** *Let $\mathcal{X} = [0,1]^{d_x}$. Consider random vectors $(X, Z), (X^1, Z^1), \ldots, (X^n, Z^n) \in \mathbb{R}^{d_x} \times \mathbb{R}$ and let $k_n$ be a deterministic parameter such that $\log(k_n) \to \infty$ and $k_n/n \to 0$ as $n \to \infty$. Then, a RF estimator is constructed as*

$$\mu_n(X) = \mathbb{E}_\Theta\left[\sum_{i\in[n]} \frac{\mathbb{1}\left\{X^i \in A_n(X;\Theta)\right\}}{N_n(X;\Theta)}\mathbb{1}\left\{N_n(X;\Theta) > 0\right\} Z^i\right], \tag{20}$$

*where $\Theta$ is a randomization variable to determine how CARTs are constructed, which is assumed to be independent from $(X, Z), (X^1, Z^1), \ldots, (X^n, Z^n)$. Moreover, given $\Theta$, $A_n(X;\Theta)$ denotes the unique leaf node containing $X$ and $N_n(X;\Theta)$ is the number of data points in the same leaf node as $X$, i.e., $N_{n;}(X;\Theta) = \sum_{i\in[n]}\mathbb{1}\left\{X^i \in A_n(X;\Theta)\right\}$.*

We note that the independence of $\Theta$ and $(X, Z), (X^1, Z^1), \ldots, (X^n, Z^n)$ rules out any data-dependent strategy to build the trees as well as bootstrapping or resampling steps in the training set. Moreover, in practice, the expectation with respect to $\Theta$ in the definition of $\mu_n(X)$, defined in (20), is approximated by generating a collection $\{\Theta^t\}_{t\in[T]}$ i.i.d. realizations of $\Theta$ and taking the average of the individual CART estimation. That is, $\frac{1}{T}\sum_{t\in[T]}\sum_{i\in[n]}\frac{\mathbb{1}\left\{X^i\in A_n(X;\Theta^t)\right\}}{N_n(X;\Theta^t)}\mathbb{1}\left\{N_n(X;\Theta^t) > 0\right\} Z^i$. Such an approximation is justified by the law of large numbers. We adopt the following result from [12, Theorem 5].

**Lemma 4.** *Consider i.i.d. random vectors $(X, Z), (X^1, Z^1), \ldots, (X^n, Y^n) \in \mathbb{R}^{d_x} \times \mathbb{R}$. Suppose that $X$ is uniformly distributed on $\mathcal{X} = [0,1]^{d_x}$. Furthermore, suppose that there exists $\sigma^2 > 0$ such that $\mathrm{Var}_P[Z \mid X = x] \leq \sigma^2$, uniformly for all $x \in \mathcal{X}$, and $\mu(x) = \mathbb{E}_P[Z \mid X = x]$ is L-Lipschitz. Let $\mu_n(x)$ denote a RF estimator with $k_n = \lfloor cn^\gamma \rfloor$ for $\gamma \in (0,1)$ and $c > 0$, constructed based on*

*Definition 3. Then,*

$$P^n\left\{|\mu_n(x)-\mu(x)|\geq\kappa\right\}\leq\frac{1152}{\kappa^2\pi}\left(\frac{\pi\log(2)}{16}\right)^{\frac{1}{2}}\sigma^2\left(\frac{d_x^2}{d_x-1}\right)\exp\{-c(1-\gamma)\log(n)\}$$
$$+\frac{8L^2}{\kappa^2}d_x\exp\left\{-\frac{3\gamma}{4d_x\log(2)}\log(n)\right\}$$
$$\frac{4\sup_{x\in[0,1]^{d_x}}\left(\mu(x)\right)^2}{\kappa^2}\exp\left\{-\frac{(1-\gamma)}{2c}\log(n)\right\},$$

*for any $\kappa>0$ and a.e. $x\in\mathcal{X}$.*

**Remark 3.** *Suppose that the weight function $W_n^i(x,\mathcal{E}_n)$, $i\in[n]$, is formed using a RF estimator based on Definition 3, with $k_n=\lfloor cn^\gamma\rfloor$ for $\gamma\in(0,1)$ and $c>0$. Then, under the hypothesis of Lemma 4, the quantities in Assumption 1 can be specified as $a(n)=\log(n)$, $A(\kappa,x):=\frac{1}{\kappa^2}\max\left\{\frac{\mathcal{O}(1)\sigma^2d_x^2}{d_x-1},\mathcal{O}(1)d_x,\mathcal{O}(1)\sup_{x\in[0,1]^{d_x}}\left(\mu(x)\right)^2\right\}$, and $B(\kappa,x):=\min\left\{\mathcal{O}(1)(1-\gamma),\frac{\mathcal{O}(1)\gamma}{d_x},\frac{1-\gamma}{\mathcal{O}(1)}\right\}$.*

**Remark 4.** *It is straightforward to obtain the minimum dataset size required to guarantee that Assumption 1 holds with probability at most $\rho\in(0,1)$ for different weight functions. For a $k_n$-NN estimator constructed based on Definition 1, we have*

$$n\geq\frac{1}{B(x,\kappa)}\log\left(\frac{A(x,\kappa)}{\rho}\right),$$

*where $A(x,\kappa)$ and $B(x,\kappa)$ are given in Remark 1. Similarly, for CART and RF estimators constructed based on Definitions 2 and 3, respectively, we have*

$$n\geq\exp\left\{\frac{1}{B(x,\kappa)}\log\left(\frac{A(x,\kappa)}{\rho}\right)\right\},$$

*where $A(x,\kappa)$ and $B(x,\kappa)$ are given in Remarks 2 and 3 for CART and RF, respectively.*

In Lemma 2, we presented a uniform consistency bound for a $k_n$-NN estimator constructed based on Definition 1; that is, the result in Lemma 1 without explicitly guaranteeing the stochastic equicontinuity in Assumption 2. We end this section by providing a sufficient condition for the stochastic equicontinuity. Hence, for CART and RF estimators constructed based on Definitions 2 and 3, respectively, this sufficient condition, combined with Lemmas 3 and 4, yield the result of Lemma 1. Our sufficient condition generalizes the usual sufficient condition of Lipschitz continuity with a common constant for the equicontinuity of deterministic functions.

**Lemma 5.** *Suppose that there exist a sequence of random variables $\{F_n\}$ and constants $M$, $A'(\kappa)$, and $B'(\kappa)$ such that $P^n\{F_n/M\geq\kappa\}\leq A'(\kappa)\exp\{-b(n)B'(\kappa)\}$, for any $\kappa>0$, where $b(n)\to\infty$ as $n\to\infty$. Moreover, suppose that for all $x,x'\in\mathcal{X}$, we have $|R_n(x)-R_n(x')|\leq F_nh(\|x-x'\|)$, where $R_n(x)$ is defined in (10) and $h:[0,\infty)\mapsto[0,\infty)$, with $h(0)=0$ and $h$ is continuous at zero. Then, Assumption 2 holds.*

*Proof.* Let us choose $\eta>0$ small enough that $h(d)\leq1/M$ for all $0\leq d\leq\eta$. Then, $\sup_{x\in\mathcal{X}}\sup_{x'\in\mathcal{N}_\eta(x)}|R_n(x)-R_n(x')|\leq F_n\sup_{0\leq d\leq\eta}h(d)\leq F_n/M$. This completes the proof as $P^n\left\{\sup_{x\in\mathcal{X}}\sup_{x'\in\mathcal{N}_\eta(x)}|R_n(x)-R_n(x')|\geq\kappa\right\}\leq P^n\{F_n/M\geq\kappa\}$ and $P^n\{F_n/M\geq\kappa\}\leq A'(\kappa)\exp\{-b(n)B'(\kappa)\}$ for any $\kappa>0$. □

We presented sufficient conditions under which estimators based on $k$-NN, CART, and RF are consistent with an exponential rate of convergence. In Sections 4.2 and 4.3, we present theoretical results on the consistency of the optimal value and feasibility of data-driven solutions to (DDC-CCP).

### 4.2 Optimal Value

In this section, we present a stochastic lower bound for the optimal value of (C-CCP). Theorem 1 states that under Assumptions 1 and 2, an optimal solution to (DDC-CCP) yields a lower bound on the optimal value of (C-CCP) with high probability. Moreover, this probability approaches one exponentially fast as the number of data points increases.

**Theorem 1.** *Suppose that Assumption 1 and Assumption 2 hold. Let $\alpha \in (0, \epsilon)$. Then, for a.e. $x \in \mathcal{X}$, we have*

$$P^n\left\{\hat{z}_{n,\alpha}(x) \leq z^*_\epsilon(x)\right\} \geq 1 - A(\epsilon-\alpha, x)\exp\{-a(n)B(\epsilon-\alpha, x)\},$$

*and*

$$P^n\left\{\hat{z}_{n,\alpha}(x) \leq z^*_\epsilon(x) \text{ for all } x \in \mathcal{X}\right\} \geq 1 - C(\epsilon-\alpha)\exp\{-c(n)D(\epsilon-\alpha)\},$$

*where $c(n)$, $C(\cdot)$, and $D(\cdot)$ are specified in Lemma 1.*

*Proof.* For $X = x$, let us consider an optimal solution $u^*(x)$ to (C-CCP). Observe that

$$\begin{aligned}
\{\hat{z}_{n,\alpha}(x) > z^*_\epsilon(x)\} &\subseteq \{u^*(x) \notin \mathcal{U}_{n,\alpha}(x)\} \\
&= \{M_n\big(u^*(x); x\big) < \alpha\} \\
&\subseteq \{M_n\big(u^*(x); x\big) \leq \alpha\} \\
&\subseteq \{M_n\big(u^*(x); x\big) - m\big(u^*(x); x\big) \leq \alpha - \epsilon\}, \\
&\subseteq \{\sup_{u\in\mathcal{U}} |M_n(u; x) - m(u; x)| \geq \epsilon - \alpha\} \\
&= \{R_n(x) \geq \epsilon - \alpha\},
\end{aligned}$$

where the third inclusion follows from $m\big(u^*(x); x\big) \geq \epsilon$ given that $u^*(x)$ is a feasible solution to (C-CCP). Now, the result in the first part follows from applying $P^n\{\cdot\}$ on both sides and using Assumption 1. To prove the second part, note that

$$\begin{aligned}
\{\hat{z}_{n,\alpha}(x) > z^*_\epsilon(x) \text{ for some } x \in \mathcal{X}\} &\subseteq \{R_n(x) \geq \epsilon - \alpha \text{ for some } x \in \mathcal{X}\} \\
&\subseteq \{\sup_{x\in\mathcal{X}} R_n(x) \geq \epsilon - \alpha\}.
\end{aligned}$$

Now, the result in the second part follows from applying $P^n\{\cdot\}$ on the both sides and using Lemma 1, which is a consequence of Assumptions 1 and 2. □

Theorem 1 offers both pointwise and uniform probabilistic guarantees. The first part of Theorem 1 can be understood as follows: when a new feature is observed, we can expect with high probability to find a lower bound for the optimal value of the true problem (C-CCP) by solving an approximate problem (DDC-CCP). The second part of Theorem 1 offers a uniform guarantee for any observed feature.

We now present the minimum dataset size $n$ required to guarantee with high probability a lower bound on the optimal value of (C-CCP) for different weight functions, based on $k$-NN, CART, and RF.

**Proposition 1.** *Under assumptions of Theorem 1, the minimum dataset size $n$ required to guarantee that $\{\hat{z}_{n,\alpha}(x) \leq z^*_\epsilon(x)\}$ for a.e. $x \in \mathcal{X}$ with probability at least $1-\rho$, $\rho \in (0,1)$, is calculated as follows:*

*(i) For a $k_n$-NN estimator constructed based on Definition 1 and under assumptions of Lemma 2, we have*

$$n \geq \frac{1}{\min\left\{\mathcal{O}(1)\big(\mathcal{O}(1)\kappa\big)^{2d_x}, \min_{n\in\mathbb{N}}\left\{\mathcal{O}(1)\frac{n^{\gamma-1}\kappa^2}{\sigma^2} - \mathcal{O}(1)d_x\frac{\log(n)}{n}\right\}\right\}} \times \left[\log\left(\frac{1}{\rho}\right) + \log\left(\max\left\{\left(\frac{\mathcal{O}(1)\sqrt{d_x}}{\kappa}\right)^{d_x}, \mathcal{O}(1)\left(\frac{\mathcal{O}(1)}{d_x}\right)^{d_x}\right\}\right)\right],$$

*where $n \geq \mathcal{O}(1)\left(\frac{\mathcal{O}(1)}{\kappa}\right)^{\frac{d_x}{1-\gamma}}$ and $\frac{n^\gamma}{\log(n)} \geq \frac{d_x\sigma^2}{\kappa^2}$, and $\kappa = \epsilon - \alpha$.*

*(ii) For a CART estimator constructed based on Definition 2 and under assumptions of Lemma 3, we have*

$$n \geq \exp\left\{\frac{1}{\min\left\{\mathcal{O}(1)(1-\gamma), \frac{\mathcal{O}(1)\gamma}{d_x}, \frac{1-\gamma}{\mathcal{O}(1)}\right\}} \times \left[\log\left(\frac{1}{\rho}\right) + \log\left(\frac{1}{\kappa^2}\max\left\{\mathcal{O}(1)\sigma^2, \mathcal{O}(1)d_x, \mathcal{O}(1)\Big(\sigma^2 + \sup_{x\in[0,1]^{d_x}}\big(\mu(x)\big)^2\Big)\right\}\right)\right]\right\},$$

*where $\kappa = \epsilon - \alpha$.*

*(iii) For a RF estimator constructed based on Definition 3 and under assumptions of Lemma 4, we have*

$$n \geq \exp\left\{\frac{1}{\min\left\{\mathcal{O}(1)(1-\gamma), \frac{\mathcal{O}(1)\gamma}{d_x}, \frac{1-\gamma}{\mathcal{O}(1)}\right\}} \times \left[\log\left(\frac{1}{\rho}\right) + \log\left(\frac{1}{\kappa^2}\max\left\{\frac{\mathcal{O}(1)\sigma^2 d_x^2}{d_x-1}, \mathcal{O}(1)d_x, \mathcal{O}(1)\sup_{x\in[0,1]^{d_x}}\big(\mu(x)\big)^2\right\}\right)\right]\right\},$$

*where $\kappa = \epsilon - \alpha$.*

*Proof.* The proof is immediate from Theorem 1, combined with Remarks 1, 2, and 3. □

**Remark 5.** *Observe from Proposition 1 that the higher the confidence level $1-\rho$, the larger the required dataset size $n$. However, $n$ grows logarithmically in $1/\rho$ for $k$-NN, whereas it grows linearly in $1/\rho$ for CART and RF (because of $\exp\{\log(1/\rho))\}$). Moreover, $n$ is large for $\alpha$ close to $\epsilon$ because of $(1/(\epsilon-\alpha))^2$. For CART and RF, $n$ grows linearly in $1/(\epsilon-\alpha)^2$ (because of $\exp\{\log(1/(\epsilon-\alpha)^2)\}$), whereas for $k$-NN, the growth of $n$ is in a polynomial order of $1/(\epsilon-\alpha)^{2d_x}$. That is, the larger the dimension of the feature vector, the closeness of $\alpha$ to $\epsilon$ has a larger impact on $n$. We also see in that in general, the larger $d_x$, the larger $n$, albeit the growth has different rates for different weight functions.*

### 4.3 Feasibility Results

In this section, we state and prove two feasibility results under Assumptions 1 and 2, one assuming that the compact set $\mathcal{U}$ is a finite set (Theorem 2) and one for a more general compact set but under some mild additional assumptions on function $G(\cdot, y)$, $y \in \mathcal{Y}$ (Theorem 3). These results state that the probability that an optimal data-driven solution to (DDC-CCP) remains feasible to (C-CCP) approaches one exponentially fast, as the number of data points increases. Similar to Theorems 1, 2 and 3 offer both pointwise and uniform probabilistic guarantees.

**Theorem 2.** *Suppose that Assumptions 1 and 2 hold. Moreover, assume that the compact set $\mathcal{U}$ is finite and let $\epsilon \in (0, \alpha)$. Then, for a.e. $x \in \mathcal{X}$, we have*

$$P^n\left\{\mathcal{U}_{n,\alpha}(x) \subseteq \mathcal{U}_\epsilon(x)\right\} \geq 1 - |\mathcal{U}| A(\alpha-\epsilon, x) \exp\{-a(n) B(\alpha-\epsilon, x)\},$$

*and*

$$P^n\left\{\mathcal{U}_{n,\alpha}(x) \subseteq \mathcal{U}_\epsilon(x) \text{ for all } x \in \mathcal{X}\right\} \geq 1 - |\mathcal{U}| C(\alpha-\epsilon) \exp\{-c(n)\} D(\alpha-\epsilon)\},$$

*where $c(n)$, $C(\cdot)$, and $D(\cdot)$ are specified in Lemma 1.*

*Proof.* For $X = x$, we have

$$\begin{aligned}
\{u \in \mathcal{U}_{n,\alpha}(x) \text{ and } u \in \mathcal{U} \backslash \mathcal{U}_\epsilon(x)\} &= \{M_n(u;x) \geq \alpha \text{ and } m(u;x) < \epsilon\} \\
&\subseteq \{M_n(u;x) - m(u;x) \geq \alpha - \epsilon\} \\
&\subseteq \{\sup_{u \in \mathcal{U}} |M_n(u;x) - m(u;x)| \geq \alpha - \epsilon\} \\
&= \{R_n(x) \geq \alpha - \epsilon\} \qquad (21)
\end{aligned}$$

Moreover, we have

$$\begin{aligned}
\{\mathcal{U}_{n,\alpha}(x) \nsubseteq \mathcal{U}_\epsilon(x)\} &= \{\exists\, u \in \mathcal{U} \text{ such that } u \in \mathcal{U}_{n,\alpha}(x) \text{ and } u \in \mathcal{U} \backslash \mathcal{U}_\epsilon(x)\} \\
&= \bigcup_{u \in \mathcal{U}} \{u \in \mathcal{U}_{n,\alpha}(x) \text{ and } u \in \mathcal{U} \backslash \mathcal{U}_\epsilon(x)\}.
\end{aligned}$$

Thus, by applying $P^n\{\cdot\}$ on the both sides, using (21), and application of the union bound, we have

$$P^n\left\{\mathcal{U}_{n,\alpha}(x) \nsubseteq \mathcal{U}_\epsilon(x)\right\} \leq \sum_{u \in \mathcal{U}} P^n\left\{R_n(x) \geq \alpha - \epsilon\right\}. \qquad (22)$$

Now, the result in the first part follows from Assumption 1. To prove the second part, note that (21) leads to

$$\{u \in \mathcal{U}_{n,\alpha}(x) \text{ and } u \in \mathcal{U} \backslash \mathcal{U}_\epsilon(x) \text{ for some } x \in \mathcal{X}\} \subseteq \{\sup_{x \in \mathcal{X}} R_n(x) \geq \alpha - \epsilon\}.$$

Now, using a similar argument as that leading to (22), we have

$$P^n\left\{\mathcal{U}_{n,\alpha}(x) \nsubseteq \mathcal{U}_\epsilon(x) \text{ for some } x \in \mathcal{X}\right\} \leq \sum_{u\in\mathcal{U}} P^n\left\{\sup_{x\in\mathcal{X}} R_n(x) \geq \alpha - \epsilon\right\}.$$

Now, the result in the second part follows from Lemma 1, which is a consequence of Assumptions 1 and 2. □

We now turn our attention to a more general compact set $\mathcal{U}$, possibly infinite. However, we need to include additional assumptions on $G(\cdot, y)$, $y \in \mathcal{Y}$.

**Assumption 3.** *For any $y \in \mathcal{Y}$, function $G(\cdot, y)$ is $L(y)$-Lipschitz, with $L := \sup_{y\in\mathcal{Y}} L(y) < \infty$.*

Before presenting Theorem 3, we recall the following fact.

**Fact 1.** *For a bounded set $\mathcal{Z} \subseteq \mathbb{R}^m$ with diameter $\theta$, and any $v > 0$, there exists a finite set $\mathcal{Z}_v \subseteq \mathcal{Z}$ with $|\mathcal{Z}_v| \leq \lceil \left(\frac{\theta}{v}\right)^m \rceil$ such that for any $z \in \mathcal{Z}$, there exists $z' \in \mathcal{Z}_v \cap \mathcal{N}_v(z)$.*

**Theorem 3.** *Suppose that the compact set $\mathcal{U}$ has diameter $\theta$, and Assumptions 1, 2, and 3 hold. Moreover, let $0 < \beta < \alpha - \epsilon$ and $\lambda > 0$. Then, for a.e. $x \in \mathcal{X}$, we have*

$$P^n\left\{\mathcal{U}^\lambda_{n,\alpha}(x) \subseteq \mathcal{U}_\epsilon(x)\right\} \geq 1 - \left\lceil\frac{1}{\beta}\right\rceil\left\lceil\left(\frac{2L\theta}{\lambda}\right)^{d_u}\right\rceil A(\alpha-\epsilon-\beta, x)\exp\{-a(n)B(\alpha-\epsilon-\beta, x)\},$$

*where*

$$\mathcal{U}^\lambda_{n,\alpha}(x) = \left\{u \in \mathcal{U} \,\middle|\, \sum_{i\in[n]} W^i_n(x, \mathcal{E}_n)\mathbb{1}\left\{G(u, Y^i) + \lambda \leq 0\right\} \geq \alpha\right\}. \tag{23}$$

*Moreover,*

$$P^n\left\{\mathcal{U}^\lambda_{n,\alpha}(x) \subseteq \mathcal{U}_\epsilon(x) \text{ for all } x \in \mathcal{X}\right\} \geq 1 - \left\lceil\frac{1}{\beta}\right\rceil\left\lceil\left(\frac{2L\theta}{\lambda}\right)^{d_u}\right\rceil C(\alpha-\epsilon-\beta) \times \exp\{-c(n)D(\alpha-\epsilon-\beta)\}.$$

*where $c(n)$, $C(\cdot)$, and $D(\cdot)$ are specified in Lemma 1.*

To prove the first part of Theorem 3, we follow a similar argument as that of [34, Theorem 10]. One major difference is that we use results from Theorem 2 instead of inequality (12) in that paper. Moreover, to prove the second part, we use a similar argument as that in the proof of the second part of Theorem 2. For completeness, we present the proof here.

*Proof.* Consider $X = x$. Let $J = \lceil\frac{1}{\beta}\rceil$, and for $j \in [J-1]$, define

$$\mathcal{U}_j(x) := \left\{u \in \mathcal{U} \,\middle|\, \frac{j-1}{J} \leq m(u; x) < \frac{j}{J}\right\},$$

and let

$$\mathcal{U}_J(x) := \left\{u \in \mathcal{U} \,\middle|\, \frac{J-1}{J} \leq m(u; x) \leq 1\right\}.$$

We note that for some $j$, but not all, $\mathcal{U}_j(x)$ might be empty. We claim that for any $j \in [J]$ such that $\mathcal{U}_j(x) \neq \emptyset$, there exists a finite set $\mathcal{Z}^\lambda_j(x) \subseteq \mathcal{U}_j(x)$ such that $|\mathcal{Z}^\lambda_j(x)| \leq \lceil\left(\frac{2L\theta}{\lambda}\right)^{d_u}\rceil$ and for any $u \in \mathcal{U}_j(x)$, there exists $z \in \mathcal{Z}^\lambda_j(x) \cap \mathcal{N}_{\frac{\lambda}{L}}(u)$. To prove the claim, first, note that by Fact

1, there exists a finite set $\emptyset \neq \mathcal{S} \subseteq \mathcal{U}$ with $|\mathcal{S}| \leq \lceil \left(\frac{2L\theta}{\lambda}\right)^{d_u} \rceil$ such that for any $u \in \mathcal{U}$, there exists $s \in \mathcal{S} \cap \mathcal{N}_{\frac{\lambda}{2L}}(u)$. Let us define $\overline{\mathcal{S}}_j^\lambda(x) := \left\{ s \in \mathcal{S} \,\middle|\, \mathcal{U}_j(x) \cap \mathcal{N}_{\frac{\lambda}{2L}}(s) \neq \emptyset \right\}$. For any $s \in \overline{\mathcal{S}}_j^\lambda(x)$, choose an arbitrary $u_s \in \mathcal{U}_j(x) \cap \mathcal{N}_{\frac{\lambda}{2L}}(s)$. Let $\widetilde{\mathcal{Z}}_j^\lambda(x) := \bigcup_{s \in \overline{\mathcal{S}}_j^\lambda(x)} u_s$. Note that by construction, $\widetilde{\mathcal{Z}}_j^\lambda(x) \subseteq \mathcal{U}_j(x)$ and $|\widetilde{\mathcal{Z}}_j^\lambda(x)| = |\overline{\mathcal{S}}_j^\lambda(x)| \leq |\mathcal{S}| \leq \lceil \left(\frac{2L\theta}{\lambda}\right)^{d_u} \rceil$. Moreover, for any $u \in \mathcal{U}_j(x) \subseteq \mathcal{U}$, there exists $s \in \mathcal{S} \cap \mathcal{N}_{\frac{\lambda}{2L}}(u)$, for which $\mathcal{U}_j(x) \cap \mathcal{N}_{\frac{\lambda}{2L}}(s) \neq \emptyset$ (because $u \in \mathcal{U}_j(x) \cap \mathcal{N}_{\frac{\lambda}{2L}}(s)$). Consequently, this $s$ belongs to $\overline{\mathcal{S}}_j^\lambda(x)$, and hence, there exists $u_s \in \widetilde{\mathcal{Z}}_j^\lambda(x)$. Note that by the definition of $\widetilde{\mathcal{Z}}_j^\lambda(x)$, this $u_s$ belongs to $\mathcal{N}_{\frac{\lambda}{2L}}(s)$ as well, i.e., $u_s \in \widetilde{\mathcal{Z}}_j^\lambda(x) \cap \mathcal{N}_{\frac{\lambda}{2L}}(s)$. Now, by the triangle inequality we have

$$\|u - u_s\| \leq \|u - s\| + \|s - u_s\| \leq \frac{\lambda}{2L} + \frac{\lambda}{2L} = \frac{\lambda}{L},$$

that is, $u_s \in \widetilde{\mathcal{Z}}_j^\lambda(x) \cap \mathcal{N}_{\frac{\lambda}{L}}(u)$. Now, by taking $\widetilde{\mathcal{Z}}_j^\lambda(x)$ as $\mathcal{Z}_j^\lambda(x)$, the claim is proved.

Now, let us define the finite set $\mathcal{Z}^\lambda(x) := \bigcup_{j \in [J]} \mathcal{Z}_j^\lambda(x)$, where $|\mathcal{Z}^\lambda(x)| \leq J\lceil \left(\frac{2L\theta}{\lambda}\right)^{d_u} \rceil$ (again, we note that $\mathcal{Z}_j^\lambda(x)$ might be empty for some $j$, but not all. Hence, $\mathcal{Z}^\lambda(x)$ is nonempty). We also define

$$\overline{\mathcal{Z}}_{\epsilon+\beta}^\lambda(x) = \left\{ u \in \mathcal{Z}^\lambda(x) \,\middle|\, m(u;x) \geq \epsilon + \beta \right\},$$

and

$$\overline{\mathcal{Z}}_{n,\alpha}^\lambda(x) = \left\{ u \in \mathcal{Z}^\lambda(x) \,\middle|\, M_n(u;x) \geq \alpha \right\}.$$

Because $\mathcal{Z}^\lambda(x)$ is finite and $\alpha > \epsilon + \beta$, by Theorem 2, we have

$$\begin{aligned} P^n \left\{ \overline{\mathcal{Z}}_{n,\alpha}^\lambda(x) \subseteq \overline{\mathcal{Z}}_{\epsilon+\beta}^\lambda(x) \right\} \geq 1 - \left\lceil \frac{1}{\beta} \right\rceil \left\lceil \left(\frac{2L\theta}{\lambda}\right)^{d_u} \right\rceil A(\alpha - \epsilon - \beta, x) \\ \times \exp\{-a(n) B(\alpha - \epsilon - \beta, x)\}. \end{aligned} \tag{24}$$

Now, to complete the proof of the result in the first part, consider $u \in \mathcal{U}_{n,\alpha}^\lambda(x)$. Let $j \in [J]$ be such that $u \in \mathcal{U}_j(x)$. Using the claim, for this $u$, there exists $z \in \mathcal{Z}_j^\lambda(x) \cap \mathcal{N}_{\frac{\lambda}{L}}(u)$. In particular, this $z$ belongs to $\mathcal{U}_j(x) \supseteq \mathcal{Z}_j^\lambda(x)$ and $|m(u;x) - m(z;x)| \leq \frac{1}{J} \leq \beta$. Moreover, by Assumption 3, we have $|G(u, Y^i) - G(z, Y^i)| \leq L\|u - z\| \leq \lambda$ for $i \in [n]$. In particular, $G(z, Y^i) \leq G(u, Y^i) + \lambda$ for all $i \in [n]$, implying that if $G(u, Y^i) + \lambda \leq 0$, we have $G(z, Y^i) \leq 0$. That is, $\mathbb{1}\left\{G(z, Y^i) \leq 0\right\} \geq \mathbb{1}\left\{G(u, Y^i) + \lambda \leq 0\right\}$. Consequently, given that $u \in \mathcal{U}_{n,\alpha}^\lambda(x)$, we have $M_n(z;x) \geq \alpha$. In addition, as $z \in \mathcal{Z}_j^\lambda(x) \subseteq \mathcal{Z}^\lambda(x)$, we conclude that $z \in \overline{\mathcal{Z}}_{n,\alpha}^\lambda(x)$. Now, if $\overline{\mathcal{Z}}_{n,\alpha}^\lambda(x) \subseteq \overline{\mathcal{Z}}_{\epsilon+\beta}^\lambda(x)$, then we have $m(z;x) \geq \epsilon + \beta$, which combined with $m(u;x) \geq m(z;x) - \beta$, imply that $m(u;x) \geq \epsilon$, i.e., $u \in \mathcal{U}_\epsilon(x)$. Consequently, we showed that $\{\mathcal{U}_{n,\alpha}^\lambda(x) \subseteq \mathcal{U}_\epsilon(x)\} \supseteq \{\overline{\mathcal{Z}}_{n,\alpha}^\lambda(x) \subseteq \overline{\mathcal{Z}}_{\epsilon+\beta}^\lambda(x)\}$. This, in turn, imply that

$$P^n \left\{ \mathcal{U}_{n,\alpha}^\lambda(x) \subseteq \mathcal{U}_\epsilon(x) \right\} \geq P^n \left\{ \overline{\mathcal{Z}}_{n,\alpha}^\lambda(x) \subseteq \overline{\mathcal{Z}}_{\epsilon+\beta}^\lambda(x) \right\},$$

and thus, the result in the first part follows from (24). The second part can be proved using a similar argument as that in the proof of the second part of Theorem 2, and by simply using the uniform consistency result (i.e., Lemma 1) in the right-hand side of (24). □

Note that the modified feasible set (23) requires strict satisfaction of $G(u, Y^i) \leq 0$ for a sufficient number of data points. For instance, if $W_n^i(x, \mathcal{E}_n) = \frac{1}{n}$, any feasible solution to $\mathcal{U}_{n,\alpha}^\lambda(x)$ strictly satisfies $G(u, Y^i) \leq 0$ for at least $n\alpha$ data points.

We now present the minimum dataset size $n$ required to guarantee with high probability the feasibility of a data-driven solution, for different weight functions, based on $k$-NN, CART, and RF.

**Proposition 2.** *Suppose that* $|\mathcal{U}| \leq U^{d_u}$. *Under assumptions of Theorem 2, the minimum dataset size* $n$ *required to guarantee that* $\{\mathcal{U}_{n,\alpha}(x) \subseteq \mathcal{U}_\epsilon(x)\}$ *for a.e.* $x \in \mathcal{X}$ *with probability at least* $1-\rho$, $\rho \in (0,1)$, *is calculated as follows:*

*(i) For a* $k_n$*-NN estimator constructed based on Definition 1 and under assumptions of Lemma 2, we have*

$$n \geq \frac{1}{\min\left\{\mathcal{O}(1)\big(\mathcal{O}(1)\kappa\big)^{2d_x}, \min_{n\in\mathbb{N}}\left\{\mathcal{O}(1)\frac{n^{\gamma-1}\kappa^2}{\sigma^2} - \mathcal{O}(1)d_x\frac{\log(n)}{n}\right\}\right\}} \times \left[\log\left(\frac{1}{\rho}\right) + d_u\log(U) + \log\left(\max\left\{\left(\frac{\mathcal{O}(1)\sqrt{d_x}}{\kappa}\right)^{d_x}, \mathcal{O}(1)\left(\frac{\mathcal{O}(1)}{d_x}\right)^{d_x}\right\}\right)\right],$$

*where* $n \geq \mathcal{O}(1)\left(\frac{\mathcal{O}(1)}{\kappa}\right)^{\frac{d_x}{1-\gamma}}$ *and* $\frac{n^\gamma}{\log(n)} \geq \frac{d_x\sigma^2}{\kappa^2}$, *and* $\kappa = \alpha - \epsilon$.

*(ii) For a CART estimator constructed based on Definition 2 and under assumptions of Lemma 3, we have*

$$n \geq \exp\left\{\frac{1}{\min\left\{\mathcal{O}(1)(1-\gamma), \frac{\mathcal{O}(1)\gamma}{d_x}, \frac{1-\gamma}{\mathcal{O}(1)}\right\}} \times \left[\log\left(\frac{1}{\rho}\right) + d_u\log(U) + \log\left(\frac{1}{\kappa^2}\max\left\{\mathcal{O}(1)\sigma^2, \mathcal{O}(1)d_x, \mathcal{O}(1)\Big(\sigma^2 + \sup_{x\in[0,1]^{d_x}}\big(\mu(x)\big)^2\Big)\right\}\right)\right]\right\},$$

*where* $\kappa = \alpha - \epsilon$.

*(iii) For a RF estimator constructed based on Definition 3 and under assumptions of Lemma 4, we have*

$$n \geq \exp\left\{\frac{1}{\min\left\{\mathcal{O}(1)(1-\gamma), \frac{\mathcal{O}(1)\gamma}{d_x}, \frac{1-\gamma}{\mathcal{O}(1)}\right\}} \times \left[\log\left(\frac{1}{\rho}\right) + d_u\log(U) + \log\left(\frac{1}{\kappa^2}\max\left\{\frac{\mathcal{O}(1)\sigma^2 d_x^2}{d_x-1}, \mathcal{O}(1)d_x, \mathcal{O}(1)\sup_{x\in[0,1]^{d_x}}\big(\mu(x)\big)^2\right\}\right)\right]\right\},$$

*where* $\kappa = \alpha - \epsilon$.

*Proof.* The proof is immediate from Theorem 2, combined with Remarks 1, 2, and 3. □

In addition to the observations in Proposition 2, note that the dimension of the feasible region, $d_u$, also impacts the required dataset size $n$. For $k$-NN, $n$ grows linearly in $d_u$, whereas it grows exponentially in $d_u$ for CART and RF.

**Proposition 3.** *Under assumptions of Theorem 3, the minimum dataset size $n$ required to guarantee that $\{\mathcal{U}^{\lambda}_{n,\alpha}(x) \subseteq \mathcal{U}_{\epsilon}(x)\}$ for a.e. $x \in \mathcal{X}$ with probability at least $1-\rho$, $\rho \in (0,1)$, is calculated as follows:*

*(i) For a $k_n$-NN estimator constructed based on Definition 1 and under assumptions of Lemma 2, we have*

$$n \geq \frac{1}{\min\left\{\mathcal{O}(1)\big(\mathcal{O}(1)\kappa\big)^{2d_x}, \min_{n\in\mathbb{N}}\left\{\mathcal{O}(1)\frac{n^{\gamma-1}\kappa^2}{\sigma^2} - \mathcal{O}(1)d_x\frac{\log(n)}{n}\right\}\right\}} \times \left[\log\left(\frac{1}{\rho}\right) + d_u \log\left\lceil\frac{2L\theta}{\lambda}\right\rceil + \log\left\lceil\frac{1}{\beta}\right\rceil + \log\left(\max\left\{\left(\frac{\mathcal{O}(1)\sqrt{d_x}}{\kappa}\right)^{d_x}, \mathcal{O}(1)\left(\frac{\mathcal{O}(1)}{d_x}\right)^{d_x}\right\}\right)\right],$$

*where $n \geq \mathcal{O}(1)\left(\frac{\mathcal{O}(1)}{\kappa}\right)^{\frac{d_x}{1-\gamma}}$ and $\frac{n^{\gamma}}{\log(n)} \geq \frac{d_x\sigma^2}{\kappa^2}$, and $\kappa = \alpha - \epsilon - \beta$.*

*(ii) For a CART estimator constructed based on Definition 2 and under assumptions of Lemma 3, we have*

$$n \geq \exp\left\{\frac{1}{\min\left\{\mathcal{O}(1)(1-\gamma), \frac{\mathcal{O}(1)\gamma}{d_x}, \frac{1-\gamma}{\mathcal{O}(1)}\right\}} \times \left[\log\left(\frac{1}{\rho}\right) + d_u \log\left\lceil\frac{2L\theta}{\lambda}\right\rceil + \log\left\lceil\frac{1}{\beta}\right\rceil + \log\left(\frac{1}{\kappa^2}\max\left\{\mathcal{O}(1)\sigma^2, \mathcal{O}(1)d_x, \mathcal{O}(1)\Big(\sigma^2 + \sup_{x\in[0,1]^{d_x}}\big(\mu(x)\big)^2\Big)\right\}\right)\right]\right\},$$

*where $\kappa = \alpha - \epsilon - \beta$.*

*(iii) For a RF estimator constructed based on Definition 3 and under assumptions of Lemma 4, we have*

$$n \geq \exp\left\{\frac{1}{\min\left\{\mathcal{O}(1)(1-\gamma), \frac{\mathcal{O}(1)\gamma}{d_x}, \frac{1-\gamma}{\mathcal{O}(1)}\right\}} \times \left[\log\left(\frac{1}{\rho}\right) + d_u \log\left\lceil\frac{2L\theta}{\lambda}\right\rceil + \log\left\lceil\frac{1}{\beta}\right\rceil + \log\left(\frac{1}{\kappa^2}\max\left\{\frac{\mathcal{O}(1)\sigma^2 d_x^2}{d_x - 1}, \mathcal{O}(1)d_x, \mathcal{O}(1)\sup_{x\in[0,1]^{d_x}}\big(\mu(x)\big)^2\right\}\right)\right]\right\},$$

*where $\kappa = \alpha - \epsilon - \beta$.*

*Proof.* The proof is immediate from Theorem 3, combined with Remarks 1, 2, and 3. □

**Remark 6.** *In addition to the observations we made for Propositions 1 and 2, the required dataset size $n$ to guarantee the feasibility of a data-driven solution when the feasible region is infinite is also impacted by the choice of the parameter $\beta$. Observe that $n$ grows logarithmically with $\lceil \frac{1}{\beta} \rceil$ for k-NN, whereas as it grows linearly in $\lceil \frac{1}{\beta} \rceil$ for CART and RF. Moreover, the impact of $\lceil \frac{1}{\lambda} \rceil$ is similar to the impact of $\lceil \frac{1}{\beta} \rceil$. That is, the requirement of strict satisfaction of $G(u, Y^i) \leq 0$ for a sufficient number of data points with a slack of at least $\lambda$ leads to a growth in $n$ logarithmically proportional to $\lceil \frac{1}{\lambda} \rceil$ for k-NN, whereas for CART and RF, the impact is linear in $\lceil \frac{1}{\lambda} \rceil$. This suggests that when CART and RF are used the modified feasible region (23) is more impacted than when k-NN is used for approximation.*

## 5 Numerical Experiments

In this section, we present numerical experiments for a chance-constrained vaccine allocation problem with a continuous, infinite feasible region. We investigate the *feasibility* of a data-driven solution to the true problem and analyze the rate of convergence by increasing the required number of data points in the approximation.

The problem was originally proposed in [48], and we adapt it to include features. The decision-maker aims to minimize the number of vaccinated people (i.e., the cost of vaccination) while maintaining the post-vaccination reproductive number $R^*$ below one with high probability. If that goal is achieved, the disease tends to die out and the vaccination policy is deemed successful. We consider a simplified population formed by families of two adults and two children. Hence, there are nine different vaccination policies consisting of the combinations of selecting 0, 1, or 2 adults, and 0, 1, or 2 children. The decision-maker decides which proportion of a given combination should be vaccinated. Let $\mathcal{V}$ be the set of vaccination policies and $\mathcal{T}$ be the set for the type of people (children or adults). Defining decision variable $u_v$ as the proportion of families with two children and two adults vaccinated under policy $v \in \mathcal{V}$, the problem is formulated as follows:

$$\begin{aligned} \min_{u_v} \quad & \sum_{v\in\mathcal{V}} \sum_{t\in\mathcal{T}} v_t u_v \\ \text{s.t.} \quad & P\left\{ \sum_{v\in\mathcal{V}} Y_v u_v \leq 1 \mid X = x \right\} \geq \epsilon, \\ & 0 \leq u_v \leq 1, \quad v \in \mathcal{V}, \end{aligned} \tag{25}$$

where $v_t$ is the number of people of type $t$ vaccinated under policy $v \in \mathcal{V}$. Moreover, we set $\epsilon = \alpha - \chi$, where $\chi$ is some positive number, implying that $\epsilon < \alpha$ (see below for more details).

The randomness is given by $Y_v$, which accounts for the impact of immunization, under policy $v \in \mathcal{V}$. We assume that there are features that may be used to predict $Y_v$, $v \in \mathcal{V}$. The feature vector $X = \left(E, C, S, \{\Phi_t, \Psi_t\}_{t\in\mathcal{T}}\right)$ of nine independent random variables that are necessary to predict $Y_v$ is given in Table 2. The full expression for $Y_v$, which is adopted from [5], is

$$\begin{aligned} Y_v = & \frac{C}{4} \left( \sum_{t\in\mathcal{T}} \Phi_t \Psi_t [(1-S)(f_t - v_t E) + S v_t E(1-E)] \right. \\ & \left. + S \sum_{t\in\mathcal{T}} \sum_{r\in\mathcal{T}} \Phi_r \Psi_t (f_t - v_t E)(f_r - v_r E) \right) + \varepsilon_v, \end{aligned}$$

where $f_t$ denotes the number of people of type $t$ in the family and $\varepsilon_v$ follows a folded normal distribution $\mathrm{N}(0, (0.01)^2)$.

| Parameter name | Symbol | Distribution |
|---|---|---|
| Vaccine efficiency | $E$ | Truncated normal $(0.85, 0.32)$ in $[0, 1]$ |
| Inter-household contact rate | $C$ | Truncated normal $(1, 0.5)$ in $[0, \infty)$ |
| Intra-household spread rate | $S$ | Truncated normal $(0.6, 0.32)$ in $[0, 1]$ |
| Relative infectivity, person type $t$ | $\Phi_t$ | 0.7 w.p. 0.5 and 1.3 w.p. 0.5 |
| Relative susceptibility, person type $t$ | $\Psi_t$ | 0.7 w.p. 0.5 and 1.3 w.p. 0.5 |

Table 2: Features $X$ of the vaccine allocation problem.

To conduct experiments, we consider an instance with $\alpha \in \{0.75, 0.80, 0.85, 0.90\}$. We simulate i.i.d. data $\mathcal{D}_n = \{(x^i, y^i)\}_{i\in[n]}$ following the distributions of $X$ and $Y$. We then build a data-driven approximation to (25) as in (DDC-CCP), with a reliability level $\alpha$ (recall that our feasibility guarantees require that $\alpha - \chi = \epsilon < \alpha$, which is ensured by any choice of $\chi > 0$). We perform 50 microsimulations with a fixed $X = x$, i.e., 50 sets of training data $\mathcal{D}_n$ are generated. To calculate the out-of-sample probability of maintaining the post-vaccination reproductive number $R^*$ below one, we used 30 test sets of size 1000. We then calculated the lower bound of a 95% confidence interval (LCB) on the chance constraint. We report the results of 50 microsimulations in boxplots of these LCBs. We also let $n$ vary in the set $\{25, 50, 100, 200, 300\}$ to understand the effect on the out-of-sample probability as the dataset size grows. We consider two data-driven approaches to solve (25):

`kNN`: The constraint in (25) is approximated by a $k_n$-NN, where $k_n = \lceil n^{0.5} \rceil$,

`nSAA`: The "naïve" SAA approach without contextual information.

We present the results in Figure 1. Each plot in this figure depicts the performance of the `kNN` and `nSAA` approaches over varying training datasets of size $n$. Several trends can be seen from Figure 1. Observe that the median value of the LCBs on the probability of maintaining the post-vaccination reproductive number $R^*$ below one for the `kNN` approach consistently outperformed that of the `nSAA` approach. More importantly, `nSAA` did not yield solutions with increasing probability levels as a function of the dataset size (except for when $\alpha$ is large, i.e., $\alpha = 0.90$), whereas `kNN` always yielded consistent solutions. The case of $\alpha = 0.85$ is particularly interesting. While `kNN` is generating solutions with higher median and smaller variability as the dataset size grows, `nSAA` has a very erratic pattern. For smaller dataset sizes most solutions have zero probability of being feasible. When $n = 200$ the median moves up but the variability is so large that the possible values of the probability are essentially the interval $[0, 1]$. At $n = 300$ we still do not see any indication of convergence, with a significant proportion of solutions having probabilities below the $\alpha = 0.85$ level.

## 6 Conclusions

Chance-constrained programming is characterized by stating that a solution is feasible when it satisfies a given constraint with a probability higher than a prescribed threshold. As opposed to two-stage stochastic programming formulations, uncertainty is present in the constraints. In this paper, we propose a contextual chance-constrained programming formulation that accommodates features in addition to the dependent random variables. We first illustrate our framework in a

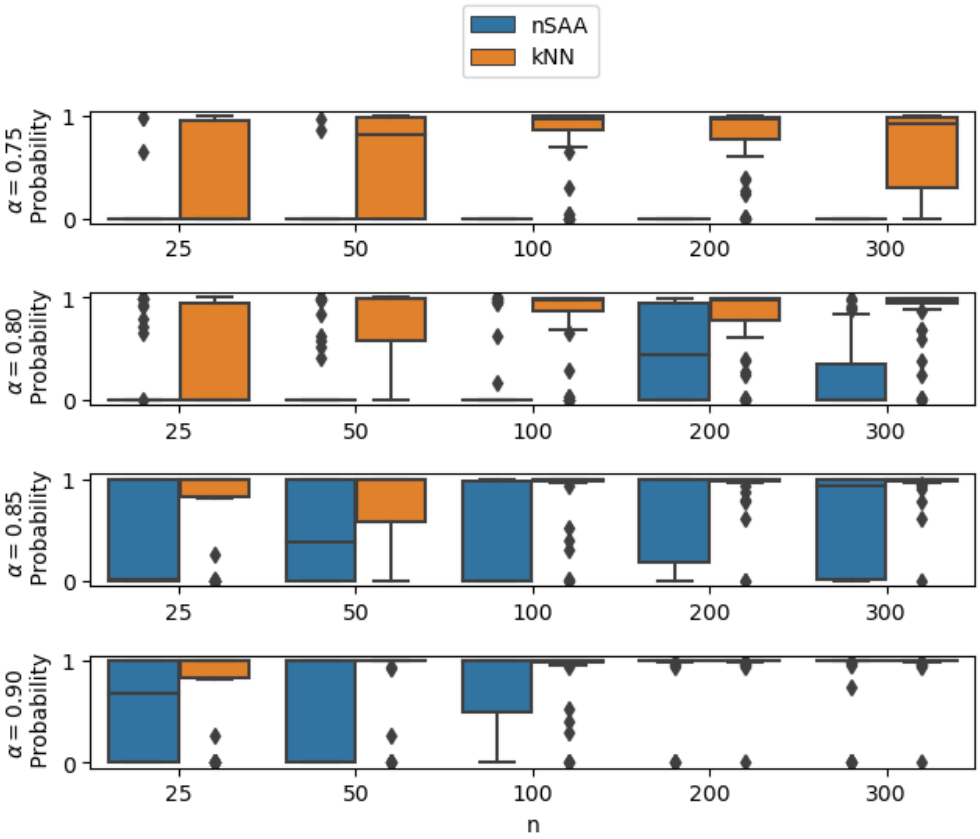


Figure 1: Comparison of `nSAA` and `kNN` for the vaccine allocation problem.

small portfolio selection problem, showing that by ignoring features the decision-maker may end up with an infeasible—and therefore, non-implementable—solution. We then show theoretical results stating that by solving an approximate, data-driven formulation one can obtain an asymptotic lower bound to the true optimal value, as well as a feasible solution to the true problem. Our results accommodate weight functions obtained by popular machine learning methods such as $k$-NN, CART, and random forest.

We test our methodology on a vaccine allocation problem. Our results show that by ignoring features a feasible solution may never be found and that our feature-based data-driven solution converges to a feasible solution as the dataset size increases.

Future work includes the study of contextual risk-averse problems and extensions to the dynamic case. It would also be interesting to investigate if our approach can accommodate decision-dependent uncertainty problems. On the application side, we plan to explore problems in energy systems (optimal power flow, unit commitment) and transportation (urban mobility, air-cargo transportation).

## References

[1] Alahakoon, D. and X. Yu (2013). Advanced analytics for harnessing the power of smart meter big data. In *2013 IEEE International Workshop on Inteligent Energy Systems (IWIES)*, pp. 40–45. IEEE.

[2] Andrieu, L., R. Henrion, and W. Römisch (2010). A model for dynamic chance constraints in hydro power reservoir management. *Eur. J. Oper. Res. 207*(2), 579–589.

[3] Ban, G.-Y. and C. Rudin (2019). The big data newsvendor: Practical insights from machine learning. *Oper. Res. 67*(1), 90–108.

[4] Bassily, R., A. Smith, and A. Thakurta (2014). Private empirical risk minimization: Efficient algorithms and tight error bounds. In *2014 IEEE 55th Annual Symposium on Foundations of Computer Science*, pp. 464–473. IEEE.

[5] Becker, N. G. and D. N. Starczak (1997). Optimal vaccination strategies for a community of households. *Mathematical Biosciences 139*(2), 117–132.

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

[7] Bertsimas, D., N. Kallus, and A. Hussain (2016). Inventory management in the era of big data. *Prod. Oper. Management 25*(12), 2006–2009.

[8] Bertsimas, D. and C. McCord (2018). Optimization over continuous and multi-dimensional decisions with observational data. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (Eds.), *Advances in Neural Information Processing Systems*. Curran Associates, Inc.

[9] Bertsimas, D. and C. McCord (2019). From predictions to prescriptions in multistage optimization problems. `arXiv preprint arXiv:1904.11637 [stat.ML]`.

[10] Bertsimas, D., C. McCord, and B. Sturt (2023). Dynamic optimization with side information. *Eur. J. Oper. Res. 304*(2), 634–651.

[11] Bertsimas, D. and B. Van Parys (2021). Bootstrap robust prescriptive analytics. *Math. Program.*, 1–40.

[12] Biau, G. (2012). Analysis of a random forests model. *J. Mach. Learn. Res. 13*(1), 1063–1095.

[13] Campi, M. C. and S. Garatti (2008). The exact feasibility of randomized solutions of uncertain convex programs. *SIAM J. Optim. 19*(3), 1211–1230.

[14] Campi, M. C., S. Garatti, and M. Prandini (2009). The scenario approach for systems and control design. *Ann. Rev. Control 33*(2), 149–157.

[15] Charnes, A. and W. W. Cooper (1959). Chance-constrained programming. *Management Sci. 6*(1), 73–79.

[16] Cohen, M. C., I. Lobel, and R. Paes Leme (2020). Feature-based dynamic pricing. *Management Sci.*.

[17] Cui, X., S. Zhu, X. Sun, and D. Li (2013). Nonlinear portfolio selection using approximate parametric value-at-risk. *J. Bank. Financ. 37*(6), 2124–2139.

[18] Davenport, T. H. and J. Dyché (2013). Big data in big companies. *International Institute for Analytics 3*, 1–31.

[19] den Hertog, D. and K. Postek (2016). Bridging the gap between predictive and prescriptive analytics- new optimization methodology needed. Technical report, Tilburg University, Netherlands. Optimization Online https://optimization-online.org/2016/12/5779.

[20] Diao, S. and S. Sen (2020). Distribution-free algorithms for learning enabled predictive stochastic programming. Optimization Online http://www.optimization-online.org/DB_HTML/2020/03/7661.html.

[21] Donini, M., L. Oneto, S. Ben-David, J. S. Shawe-Taylor, and M. Pontil (2018). Empirical risk minimization under fairness constraints. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (Eds.), *Advances in Neural Information Processing Systems*, pp. 2791–2801. Curran Associates, Inc.

[22] Döring, M., L. Györfi, and H. Walk (2017). Rate of convergence of k-nearest-neighbor classification rule. *J. Mach. Learn. Res. 18*(1), 8485–8500.

[23] El Balghiti, O., A. N. Elmachtoub, P. Grigas, and A. Tewari (2019). Generalization bounds in the predict-then-optimize framework. In H. Wallach, H. Larochelle, A. Beygelzimer, F. Alché-Buc, E. Fox, and R. Garnett (Eds.), *Advances in Neural Information Processing Systems*, pp. 14412–14421. Curran Associates, Inc.

[24] Elmachtoub, A. N. and P. Grigas (2022). Smart "predict, then optimize". *Management Sci. 68*(1), 9–26.

[25] Györfi, L., M. Kohler, A. Krzyżak, and H. Walk (2002). *A distribution-free theory of nonparametric regression*. Springer.

[26] Harikumar, H., S. Rana, S. Gupta, T. Nguyen, R. Kaimal, and S. Venkatesh (2018). Prescriptive analytics through constrained Bayesian optimization. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pp. 335–347. Springer.

[27] Holland, C. P., S. C. Thornton, and P. Naudé (2020). B2B analytics in the airline market: Harnessing the power of consumer big data. *Ind. Mark. Management 86*, 52–64.

[28] Ito, S. and R. Fujimaki (2017). Optimization beyond prediction: Prescriptive price optimization. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1833–1841. ACM.

[29] Jiang, R. and Y. Guan (2016). Data-driven chance constrained stochastic program. *Math. Program. 158*(1-2), 291–327.

[30] Kannan, R., G. Bayraksan, and J. R. Luedtke (2020). Data-driven sample average approximation with covariate information. Optimization Online http://www.optimization-online.org/DB_HTML/2020/07/7932.html.

[31] Kawas, B., M. S. Squillante, D. Subramanian, and K. R. Varshney (2013). Prescriptive analytics for allocating sales teams to opportunities. In *2013 IEEE 13th International Conference on Data Mining Workshops*, pp. 211–218. IEEE.

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

[33] Larsen, E., S. Lachapelle, Y. Bengio, E. Frejinger, S. Lacoste-Julien, and A. Lodi (2018). Predicting tactical solutions to operational planning problems under imperfect information. `arXiv preprint arXiv:1807.11876 [cs.LG]`.

[34] Luedtke, J. and S. Ahmed (2008). A sample approximation approach for optimization with probabilistic constraints. *SIAM J. Optim. 19*(2), 674–699.

[35] Meller, J., F. Taigel, and R. Pibernik (2018). Prescriptive analytics for inventory management: A comparison of new approaches. Available at SSRN http://dx.doi.org/10.2139/ssrn.3229105.

[36] Nash, D. B. (2014). Harnessing the power of big data in healthcare. *Am. Health Drug Benefits 7*(2), 69.

[37] Nguyen, V. A., F. Zhang, J. Blanchet, E. Delage, and Y. Ye (2020). Distributionally robust local non-parametric conditional estimation. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin (Eds.), *Advances in Neural Information Processing Systems*, pp. 15232–15242. Curran Associates, Inc.

[38] Pagnoncelli, B. K., S. Ahmed, and A. Shapiro (2009). Sample average approximation method for chance constrained programming: theory and applications. *J. Optim. Theory App. 142*(2), 399–416.

[39] Pagnoncelli, B. K., D. Ramírez, H. Rahimian, and A. Cifuentes (2022). A synthetic data-plus-features driven approach for portfolio optimization. *Comput. Econ.*. https://doi.org/10.1007/s10614-022-10274-2.

[40] Parkins, D. (2017). The worlds most valuable resource is no longer oil, but data. *The Economist*. https://www.economist.com/leaders/2017/05/06/the-worlds-most-valuable-resource-is-no-longer-oil-but-data [Last accessed in October 2022].

[41] Peña-Ordieres, A., J. R. Luedtke, and A. Wachter (2020). Solving chance-constrained problems via a smooth sample-based nonlinear approximation. *SIAM J. Optim. 30*(3), 2221–2250.

[42] Pollard, D. (1989). Empirical processes: Theory and applications. Volume 2 of *NSF-CBMS Regional Conference Series in Probability and Statistics*. Institute of Mathematical Statistics.

[43] Römisch, W. and R. Schultz (1991). Distribution sensitivity for certain classes of chance-constrained models with application to power dispatch. *J. Optim. Theory App. 71*(3), 569–588.

[44] Rudin, W. (1976). *Principles of Mathematical Analysis*. McGraw-Hill, New York.

[45] Scornet, E., G. Biau, and J.-P. Vert (2015). Consistency of random forests. *The Annals of Statistics 43*(4), 1716–1741.

[46] Shamim, S., J. Zeng, S. M. Shariq, and Z. Khan (2019). Role of big data management in enhancing big data decision-making capability and quality among chinese firms: A dynamic capabilities view. *Inf. Management 56*(6), 103135.

[47] Srinivas, S. and A. R. Ravindran (2018). Optimizing outpatient appointment system using machine learning algorithms and scheduling rules: a prescriptive analytics framework. *Expert Syst. Appl 102*, 245–261.

[48] Tanner, M. W. and L. Ntaimo (2010). Iis branch-and-cut for joint chance-constrained stochastic programs and application to optimal vaccine allocation. *European Journal of Operational Research 207*(1), 290–296.

[49] Tong, S., A. Subramanyam, and V. Rao (2022). Optimization under rare chance constraints. *SIAM J. Optim. 32*(2), 930–958.

[50] Vahdani, B., R. Tavakkoli-Moghaddam, F. Jolai, and A. Baboli (2013). Reliable design of a closed loop supply chain network under uncertainty: An interval fuzzy possibilistic chance-constrained model. *Engineering Optimization 45*(6), 745–765.

[51] van Ackooij, W., E. C. Finardi, and G. M. Ramalho (2018). An exact solution method for the hydrothermal unit commitment under wind power uncertainty with joint probability constraints. *IEEE Trans. Power Syst. 33*(6), 6487–6500.

[52] van Ackooij, W., R. Henrion, A. Möller, and R. Zorgati (2014). Joint chance constrained programming for hydro reservoir management. *Optim. Eng. 15*(2), 509–531.

[53] van Ackooij, W. and P. Pérez-Aros (2021). Gradient formulae for probability functions depending on a heterogenous family of constraints. *Open J. Math. Optim. 2*. article no. 7, 29 p. https://doi.org/10.5802/ojmo.9.

[54] von Bischhoffshausen, J. K., M. Paatsch, M. Reuter, G. Satzger, and H. Fromm (2015). An information system for sales team assignments utilizing predictive and prescriptive analytics. In *2015 IEEE 17th Conference on Business Informatics*, pp. 68–76. IEEE.

[55] Wang, P., R. Jiang, Q. Kong, and L. Balzano (2023). Difference-of-convex reformulation for chance constrained programs. `arXiv preprint arXiv:2301.00423 [math.OC]`.

[56] Wu, H., M. Shahidehpour, Z. Li, and W. Tian (2014). Chance-constrained day-ahead scheduling in stochastic power system operation. *IEEE Trans. Power Syst. 29*(4), 1583–1591.

[57] Xie, W. and S. Ahmed (2017). Distributionally robust chance constrained optimal power flow with renewables: A conic reformulation. *IEEE Trans. Power Syst. 33*(2), 1860–1867.

[58] Zhao, P. and Q. Xiao (2016). Portfolio selection problem with value-at-risk constraints under non-extensive statistical mechanics. *J. Comput. Appl. Math 298*, 64–71.