Contextual Chance-Constrained Programming

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

[1]Department of Industrial Engineering, Clemson University, Clemson SC 29634, USA
[2]School of Business, Universidad Adolfo Ibañez, Santiago, Chile

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. Borrowing results from quenched large deviation theory we show the exponential convergence of our scheme as the number of data points increases. We illustrate our findings with an example from a soccer hiring problem based on the player's transfer market in the UK using real data.

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, and by solving them one obtains a solution with the smallest expected cost, or that 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 daily demand is on average 100 during the weekdays, and 1,000 during the weekend. When building a model to maximize the expected revenue, it is clearly desirable to have at least one feature next to each demand observation, denoting the day of the week when the data was obtained. If a decision maker wants to place an order to avoid stockouts for a weekday, past demands from weekends, despite being useful, might inflate the ordered quantity and generate excess inventory.

Given the wider availability of data, and since predictive models have reached a more mature stage, the attention of researchers in the optimization community has been shifting towards prescriptive analytics. Very recently, several publications (den Hertog and Postek, 2016; Elmachtoub and Grigas, 2017; Ban and Rudin, 2018; El Balghiti et al., 2019; Bertsimas and Kallus, 2020; Cohen et al., 2020; Nguyen et al., 2020; Bertsimas et al., 2019; Bertsimas and McCord, 2019, 2018; Kannan et al., 2020) have pointed out that in the context of data science and decision making

---

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

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 (Davenport and Dyché, 2013; Alahakoon and Yu, 2013; Nash, 2014; Parkins, 2017; Shamim et al., 2019; Holland et al., 2020). Many applications of decision making under uncertainty that include contextual information have recently appeared in the literature, including inventory theory (Meller et al., 2018; Bertsimas et al., 2016), sales team assignments (Kawas et al., 2013; von Bischhoffshausen et al., 2015), price and revenue management (Ito and Fujimaki, 2017), health care (Harikumar et al., 2018; Srinivas and Ravindran, 2018), and load planning in rail transportation (Larsen et al., 2018).

There are two main approaches in the literature to incorporate contextual information into those problems. The first is the empirical risk minimization (ERM), which is discussed in Ban and Rudin (2018) and is a popular approach in machine learning (Bassily et al., 2014; Donini et al., 2018). In ERM it is necessary to assume a functional form for the response function given 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 Bertsimas and Kallus (2020), 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 Ban and Rudin (2018), 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 current observed feature vector. The authors also consider a regularized kernel approach, and compare the different models in a nurse staffing problem. In Bertsimas and Van Parys (2017), the authors propose two novel contextual methods that use robust kernel formulations to guarantee better out-of-sample performance by protecting against overfitting.

Another possibility to construct the weights is to use machine learning (ML) methods, which we will refer to as *ML-weights*. The approach consists in using classical ML methods to determine the weight of each data point according to the current observed feature vector. For instance, in k-Nearest Neighbors (k-NN), each neighbor of the current 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 which the current feature vector belongs to. In Bertsimas and Kallus (2020), the authors consider both kernels and the ML-weights approach 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 improvement of 88% can be obtained, measured by their proposed coefficient of prescriptiveness. In Diao and Sen (2020), the authors propose a stochastic quasi-gradient approach to solve contextual problems approximated via kernels or ML-weights with k-NN.

The previously cited papers mainly consider contextual extensions to two-stage stochastic programming problems. The focus of our paper is on chance-constrained programming (CCP) problems. CCP was introduced in Charnes and Cooper (1959), and it represents a very popular class of problems that still attracts the attention of researchers in both methodological, see, e.g., (Jiang and Guan, 2016; Peña-Ordieres et al., 2019) and practical contexts, see, e.g., (Xie and Ahmed, 2017; van Ackooij et al., 2018). Chance constraints are equivalent to Value-at-Risk constraints, which are widely used in economics and finance (Zhao and Xiao, 2016; Cui et al., 2013). Moreover, CCP is a natural formulation to describe reliability requirements in engineering problems, see, e.g., water management (Andrieu et al., 2010; van Ackooij et al., 2014), energy (Wu et al., 2014), supply

chain (Vahdani et al., 2013), among others. In the vast majority of cases it is impossible to solve the problem directly, and even evaluating if a given solution is feasible is a challenging 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 (Luedtke and Ahmed, 2008; Pagnoncelli et al., 2009).

Our first contribution in this paper is a novel formulation for CCP problems that includes contextual information either via kernels or ML-weights. For two-stage problems, 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 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 establish theoretical results for contextual CCP problems. We use tools from large deviation theory developed specifically to understand the asymptotic behavior of *weighted sums* of random variables, which are the building blocks of our approximations. While results for classical stochastic programming problems use large deviation theory (for CCP, see (Luedtke and Ahmed, 2008), for two-stage, see (Kleywegt et al., 2002)), in the case of weighted sums of random variables results such as Hoeffding's or Chebyshev's inequality may not be applied. We state and prove consistency results for our contextual CCP formulation, providing estimates for the number of data points needed in the approximate problem to ensure feasibility for the true problem.

We conclude with a real-world application from the soccer transfer market in the UK based on Pantuso and Hvattum (2019), using real market data kindly provided by the authors of that paper. This application illustrates the advantages of using contextual CCP, focusing on the difficulty of obtaining feasible solutions when features are ignored.

2 Background on Chance-Constrained Programming

A CCP problem can be written as

$$
\begin{aligned}
\min_{u \in U} \quad & f(u) \\
\text{s.t.} \quad & p(u) := P\big(G(u, \xi) \leq 0\big) \geq 1 - \alpha,
\end{aligned} \tag{CCP}
$$

where $U \subset \mathbb{R}^n$ is the feasible set of decisions u, ξ is a random vector with probability distribution P and support in the set $\Xi \subset \mathbb{R}^d$, $f : \mathbb{R}^n \to \mathbb{R}$ is the objective function, which we will assume to be deterministic, and $\alpha \in (0, 1)$ is the reliability level associated with the chance constraint. Finally, $G : \mathbb{R}^n \times \Xi \to \mathbb{R}$ is a random function; that is, $G(u, \cdot)$ is measurable for $u \in \mathbb{R}^n$.

With the exception of very particular cases, general (CCP) are challenging to solve since the feasible region can be nonconvex, and approximations are needed. Given an independent and identically distributed (i. i. d.) sample $\{\xi_i\}_{i=1}^N$, we can construct the empirical probability distribution

$$
P^N := \frac{1}{N} \sum_{i=1}^{N} \delta_{\xi_i},
$$

where δ_{ξ_i} is the Dirac point mass on ξ_i, $i = 1, \dots, N$. Using the empirical probability distribution,

we can define

$$\hat{p}^N(u) := P^N\left(G(u,\xi) \leq 0\right) = \mathbb{E}_{P^N}\left[\mathbb{1}(G(u,\xi) \leq 0)\right] = \frac{1}{N}\sum_{i=1}^{N}\mathbb{1}(G(u,\xi_i) \leq 0), \qquad (1)$$

where $\mathbb{1}(A)$ is the indicator function of a set A. The SAA problem is obtained by replacing the true chance constraint by an empirical one as follows:

$$\begin{aligned} \min_{u \in U} \quad & f(u) \\ \text{s.t.} \quad & \hat{p}^N(u) \geq 1-\alpha, \end{aligned} \qquad \text{(SAA-CCP)}$$

Convergence results for the SAA formulation are established in Luedtke and Ahmed (2008) and Pagnoncelli et al. (2009) for the general case, and in Campi and Garatti (2008); Campi et al. (2009) for $\alpha = 0$ in formulation (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 ξ exhibits some dependence on a $p-$dimensional vector of features x. Ideally, a decision maker has access to past observations of the form $\mathcal{D}_N := \{(x_i, \xi_i)\}_{i=1}^N$, where $x_i \in \mathcal{X} \subset \mathbb{R}^p$. It may be beneficial to include the features into the problem formulation since they contain relevant information that helps to explain the outcomes of the random vector ξ. The contextual chance-constrained programming problem (C-CCP) can be written as

$$\begin{aligned} \min_{u \in \mathcal{U}} \quad & f(u) \\ \text{s.t.} \quad & p_x(u) := P_{\xi|X}\left(G(u,\xi) \leq 0 \mid X = x\right) \geq 1-\alpha, \end{aligned} \qquad \text{(C-CCP)}$$

where $P_{\xi|X}$ is the conditional probability of ξ given X. For each $X = x$, 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 ξ given X. Even when such a distribution is known, approximation schemes need to be considered. We construct a data-driven approximation of (C-CCP) as follows. Given past data $\mathcal{D}_N$ and a new observation x_{N+1} of the feature vector, we use a weight function $w_i(x_{N+1})$ that measures "proximity" of each data point i with respect to x_{N+1}. The rationale is that data points that are "close" to the current observation x_{N+1}—which represents the current state of the world—are more valuable to accurately estimate the distribution of ξ given x_{N+1}. When $X = x_{N+1}$, the empirical distribution is defined as

$$P_{\xi|X}^N := \sum_{i=1}^{N} w_i(x_{N+1})\delta_{\xi_i}, \qquad (2)$$

where as before, δ_{ξ_i} is the Dirac point mass on ξ_i, $i = 1, \ldots, N$.

Similar to (1), we can write

$$\hat{p}_{x_{N+1}}^N(u) := P_{\xi|X}^N\left(G(u,\xi) \leq 0 \mid X = x_{N+1}\right) = \qquad (3)$$

$$\mathbb{E}_{P_{\xi|X}^N}\left[\mathbb{1}(G(u,\xi) \leq 0) \mid X = x_{N+1}\right] = \sum_{i=1}^{N} w_i(x_{N+1})\mathbb{1}(G(u,\xi^i) \leq 0). \qquad (4)$$

4

The data-driven contextual CCP formulation (DDC-CCP) can be written as follows:

$$\min_{u \in \mathbb{R}^n} \quad f(u)$$
$$\text{s.t.} \quad \hat{p}^N_{x_{N+1}}(u) \geq 1 - \alpha. \tag{DDC-CCP}$$

Formulation (DDC-CCP) is equivalent to

$$\min_{u \in \mathbb{R}^n} \quad f(u)$$
$$\text{s.t.} \quad G(u, \xi^i) - M z_i \leq 0, \qquad i = 1, \ldots, N,$$
$$\sum_{i=1}^{N} w_i(x_{N+1}) z_i \leq \alpha, \tag{5}$$
$$z_i \in \{0, 1\}, \qquad\qquad i = 1, \ldots N,$$

where the constant M ensures feasibility when $G(u, \xi^i)$ is positive.

As we mentioned before, there are several ways to choose the weight functions, and hence, to approximate the conditional distribution of ξ given $X = x_{N+1}$. When

$$w_i(x_{N+1}) = \frac{K_b(x_{N+1} - x_j)}{\sum_{i=1}^{N} K_b(x_{N+1} - x_i)}, \quad i = 1, \ldots N, \tag{6}$$

we have the Nadaraya-Watson estimation, where $K_b(\cdot) = K(\cdot/b)/b$ is a kernel function with bandwidth $b \in \mathbb{R}$. A popular choice is the Gaussian Kernel, but other alternatives also exist such as the Epanechnokov, or the quartic kernel (Diao and Sen (2020)). A key reference to kernels for discrete features is Aitchison and Aitken (1976). In problems that combine continuous and categorical variables—which is the case in most applications—the resulting estimation will be a product of all kernels, which can be numerically unstable. Another potential problem is that discrete kernels have additional parameters that need to be estimated, and if discrete and continuous features are present, it may be difficult to identify in an application which component needs to better calibrated to improve estimation accuracy. Discrete kernels have been implemented in the `np` package in R, and more details can be found in Hayfield and Racine (2008).

In the ML-weights case, the weight function depends on the particular algorithm chosen. For CART and random forest (RF) we assume the algorithm generates a decision 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 each ML algorithm on Table 1.

Algorithm	$w_i(x_{N+1})$
k-NN	$\frac{1}{k} \mathbb{1}(x_i \text{ is a } k\text{NN of } x_{N+1})$
CART	$\frac{\mathbb{1}\left(\mathcal{C}(x_{N+1}) = \mathcal{C}(x_i)\right)}{\lvert \{i \mid \mathcal{C}(x_i) = \mathcal{C}(x_{N+1})\} \rvert}$
RF	$\frac{1}{T} \sum_{t=1}^{T} \frac{\mathbb{1}\left(\mathcal{C}^t(x_{N+1}) = \mathcal{C}^t(x_i)\right)}{\lvert \{i \mid \mathcal{C}^t(x_i) = \mathcal{C}^t(x_{N+1})\} \rvert}$

Table 1: ML-weights.

Some observations regarding (DDC-CCP) are in order. First, one can interpret formulation (5) as a response function, or a policy: given any observation of the features x_{N+1}, a solution $u^*(x_{N+1})$ represents the best response measured by the objective function $f(\cdot)$. Second, the computational burden of solving (5) is comparable to solving the SAA formulation (SAA-CCP) since computation of the weights via kernels or ML-weights can be done offline.

Before showing the consistency 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

$$\max_{u \in U} \quad \bar{r}^T u$$
$$\text{s.t.} \quad P_{r|X}\left(r^T u \geq v \mid X = x\right) \geq 1 - \alpha, \tag{7}$$

where $\bar{r} = (0.0145, 0.0083)$ represents the average monthly returns of stocks and bonds, $v = 0.8$ is a minimum desired return, and $U = \{u \in \mathbb{R}^2 \mid u_1 + u_2 = 1, u_1 \geq 0, u_2 \geq 0\}$. We consider $\alpha = 0.1$ throughout the example. The random vector r follows a bivariate normal distribution with mean $1 + \bar{r}$ 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 the artificial data in the following way. We sample a uniform $[0,1]$ random variable; if the value is smaller than 0.5 then we take a sample using the covariance matrix corresponding to the $x = 0$ case, and if it is greater than 0.5 we sample from the $x = 1$ case. We solved the SAA problem, similar to (SAA-CCP), with 10,000 samples as described, and obtained as optimal solution

$$(u_1^*, u_2^*) = (0.57113, 0.42887), \tag{8}$$

with an optimal value of 1.01186. The SAA problem is completely myopic to the presence of covariates, and its formulation simply takes all samples and assigns them equal weights as described in approximation (SAA-CCP).

In practice one can often observe market conditions before investing, so we want to investigate how the solution described in (8) would perform under each case. Since we assume normality for the asset returns we can compute the optimal solution for the problem under each market condition, and the exact value of the probability in the chance constraint for a given candidate solution (u_1, u_2). When $x = 0$, the optimal value of the true problem (7) is 1.01417, and the SAA solution (8) is feasible and performs well. However, when $x = 1$ the solution obtained for the SAA problem is infeasible for the true problem (7):

$$P_{r|X}\left(r^T u \geq v \mid X = 1\right) = P_{r|X}\left(r^T(0.57113, 0.42887) \geq v \mid X = 1\right) = 1 - \Phi(-1.20409) = 0.88572.$$

Our simple example shows that by ignoring contextual information one may end up with a solution that is not implementable in practice.

4 Convergence Results

In this section, we will show some theoretical results that support the use of (DDC-CCP) to approximate (C-CCP). A standard technique used to prove convergence results for traditional stochastic programming problem is the large deviation theory. For the two-stage case the main reference is Kleywegt et al. (2002), for CCP problems we refer the reader to Luedtke and Ahmed (2008), and for problems with expected value constraints see Wang and Ahmed (2008).

4.1 Background

Our goal is to establish convergence of (DDC-CCP) to (C-CCP) as the number of data points goes to infinity, and offer finite sample size estimates that guarantee the feasibility of the solutions to (DDC-CCP) in (C-CCP). The techniques and results used to prove convergence results for classical stochastic programming problems cannot be directly applied in this context mainly for two reasons. First, (C-CCP) has a conditional probability on the feature vector x. The results obtained for the classical CCP formulation—model (CCP)—rely on results for annealed stochastic processes, while for our purposes we need extensions that study the behavior of quenched processes. Second, and most importantly, we do not have uniform weights in (DDC-CCP). In (SAA-CCP), each sample has a weight of $1/N$, and classical large deviation results rely heavily upon this fact. To establish convergence rates, and to estimate adequate data sizes that guarantee feasibility to the original problem we need large deviation results that take into account weighted sums of random variables.

A starting point for the study of weighted sums of random variables is Book (1972). The author uses a special form of Cramér's theorem to derive asymptotic representations for the probability that the weighted sum exceeds a threshold. In Wolf (1980), the author derives asymptotic expressions for the weighted sum when Cramér's condition is not satisfied. The work of Kiesel and Stadtmüller (2000) assumes that if the limit of the series of the weights (and of the absolute value of arbitrary powers of the weights) converge, then versions of the Erdös-Rényi law can be derived.

In Gantert et al. (2014) the authors generalize Kiesel and Stadtmüller (2000) and Nagaev (1969) and prove annealed and quenched large deviation results for weighted sums of random variables with subexponential decay. Their main contribution is to derive results without the assumption of finite exponential moments, but under the hypothesis of nonnegative weights, which cannot be relaxed. Fortunately, formulation (DDC-CCP) will always have nonnegative weights, so we can make use of those results to derive our convergence results.

Random variables with subexponetial decay (between fat tails and exponential decays) have several applications in a wide array of problems, such as the study of radio and light emissions from galaxies, of country population sizes, estimation of oil reserve sizes, among many others (see Laherrere and Sornette (1998) and the references therein). In Bonin (2002) the author applies large deviation results with weighted sums to a problem of determining the accumulated errors on travel times when using geographical databases.

Large deviation theory for a weighted sum of random variables continues to attract the attention of researchers, and in the last 10 years several other publications extended the previous results. In Gao (2003), the author derives a new version of the Erdös-Rényi-Shepp law the with different hypothesis for the limiting behavior of the coefficients. In Giuliano and Macci (2011), the authors consider logarithmic weighted means, and in Giuliano et al. (2014) they derive large deviation results for weighted means of Gaussian processes. Finally, the very recent work of Aurzada (2020) states results similar to the ones in Gantert et al. (2014) under even milder conditions on the coefficients of the weighted sums.

4.2 Theoretical Preliminaries

To establish our results we first need some assumptions for the weights $w_i(x_{N+1})$ in (3). We state these assumptions in a generic form. Let $\{w_i\}_{i \in \mathbb{N}}$ be a sequence of nonnegative real numbers. Suppose that this sequence satisfies the following assumptions (Gantert et al., 2014)

Assumption 1. *There exists a real number $s_1 \neq 0$ such that the sequence $\{R(1, N)\}_{N \in \mathbb{N}}$ of real*

numbers defined by

$$\sum_{j=1}^{N} w_i = s_1 R(1, N), \ \forall N \in \mathbb{N},$$

satisfies $R(1, N) \to 1$ as $N \to \infty$.

Assumption 2.

There exists a real number s such that for $w_{max} := \max\limits_{i=1,\dots N} w_i$,

$$\lim_{N \to \infty} N \cdot w_{max} = s.$$

Moreover, we need to define a class of functions intimately connected to random variables with subexponential decay.

Definition 1. *A function $l : (0, \infty) \to (0, \infty)$ is called slow varying (at infinity) if for every $a > 0$,*

$$\lim_{x \to \infty} \frac{l(ax)}{l(x)} = 1.$$

Slowly varying functions are extensively used in extreme value theory (examples can be found in Embrechts et al. (2013)). Typical examples of slow varying functions are positive constants, functions that converge to a positive constant, the logarithm, and iterated logarithms.

In traditional CCP, Hoeffding's inequality was essential to prove convergence results. The following theorem, adapted from Gantert et al. (2014), is the key piece to prove convergence results for (DDC-CCP).

Theorem 1. *Consider a sequence of pairs of i.i.d. random vectors $\{(X_i, Y_i)\}_{i \in \mathbb{N}}$ on a probability space $(\mathcal{X} \times \mathcal{Y}, \mathcal{F}_\mathcal{X} \times \mathcal{F}_Y, P)$, where $X_i \in \mathbb{R}^p$ and $Y_i \in \mathbb{R}$. Suppose that*

$$\mathbb{E}\left[|Y|^k \mid X = x\right] < \infty, \ \forall k \in \mathbb{N},$$

and let $m := \mathbb{E}[Y \mid X = x]$. Moreover, suppose that there exists a constant $r \in (0, 1)$ and slowly varying functions $b, c_1, c_2 : (0, \infty) \to (0, \infty)$ and a constant t^* such that for $t \geq t^*$,

$$c_1(t) \exp(-b(t)t^r) \leq P(Y \geq t \mid X = x) \leq c_2(t) \exp(-b(t)t^r).$$

Additionally, assume that $\{X_i\}_{i \in \mathbb{N}}$ is almost surely (a.s.) uniformly bounded by M^* as follows:

$$M^* := \inf\left\{a \in \mathbb{R} \ : \ P\left(Q_x(X) > a\right) = 0\right\} < \infty,$$

where $Q_x : \mathbb{R}^p \to \mathbb{R}$ is a continuous function. Define the triangular array of weights $w_i(x)$, $i = 1, \dots, N$, as

$$w_i(x) := \frac{Q_x(X_i)}{\sum_{i=1}^{N} Q_x(X_i)}, \ i = 1, \dots, N, \ N \in \mathbb{N}, \tag{9}$$

and construct the sequence of weighted sums $\{\bar{S}_N\}_{N \in \mathbb{N}}$ as follows:

$$\bar{S}_N := \sum_{i=1}^{N} w_i(x) Y_i.$$

If the triangular array of weights satisfy Assumptions 1 and 2, then for $y > m$, we have

$$\lim_{N \to \infty} \frac{1}{b(N)N^r} \log P_{\xi|X}(\bar{S}_N \geq y \mid X = x) = -\left(\frac{\mathbb{E}[Q_x(X)]}{M^*}(y - m)\right)^r, \ P\text{-a.s.}$$

Proof. Since $\sum_{i=1}^{N} w_i(x) = 1$ a.s., we can choose $s_1 = 1$ and satisfy Assumption 1. For Assumption 2, let us define $w_{max}(x) := \max_{i=1,\ldots,N} w_i(x)$. Note that

$$N \cdot w_{max}(x) = \frac{N \cdot \max\{Q_x(X_i) : 1 \le i \le N\}}{\sum_{i=1}^{N} Q_x(X_i)} = \frac{\max\{Q_x(X_i) : 1 \le i \le N\}}{\frac{1}{N}\sum_{i=1}^{N} Q_x(X_i)}. \tag{10}$$

By assumption, we have that $\max\{Q_x(X_i) : 1 \le i \le N\} \to M^*$ as $N \to \infty$. By the strong law of the large numbers, we have that the limit of (10) is equal to $s := M^*/\mathbb{E}[Q_x(X)]$, which is the requirement of Assumption 2. The rest of the proof follows from Theorem 1 of Gantert et al. (2014). $\qquad\square$

From Theorem 1 we have

$$P_{\xi|X}(\bar{S}_N \ge y | X = x) \le e^{-I(y)b(N)N^r}, \tag{11}$$

where

$$I(y) = \left(\frac{\mathbb{E}\left[Q_x(X)\right]}{M^*}(y - m) \right)^r. \tag{12}$$

4.3 Main Results

The similarity of (11) with Hoeffding and Chebyshev's inequality is apparent. We will make extensive use of the inequality (11) to prove our consistency results, which state that as the number of data points increases, the probability that the feasible set of the approximate problem (DDC-CCP) is contained in the feasible of problem (C-CCP) approaches one exponentially fast. We first state our result for the case that U is a finite set in Theorem 2. Then, we state our convergence result for the case that U is an infinite set in Theorem 3.

Theorem 2. *Suppose that the assumptions in Theorem 1 hold. Assume that $U \subset \mathbb{R}^n$ is a finite set and let $\gamma \in [0, \alpha)$. For a covariate vector $x_{N+1} \in \mathbb{R}^p$, we have*

$$P_{\xi|X}(U_\gamma^N \subseteq U_\alpha \mid X = x_{N+1}) \ge 1 - |U \backslash U_\alpha| e^{-I(\alpha - \gamma)b(N)N^r},$$

where

$$U_\gamma^N = \{u \in U \mid \hat{p}_{x_{N+1}}^N(u) \ge 1 - \gamma\}$$

is the feasible set of (DDC-CCP),

$$U_\alpha = \{u \in U \mid p(u) \ge 1 - \alpha\}$$

is the feasible set of (C-CCP), *and $I(\cdot)$ is of form* (12).

Proof. Consider $u \in U \backslash U_\alpha$; that is, $p(u) < 1 - \alpha$. We have that

$$P_{\xi|X}(u \in U_\gamma^N \mid X = x_{N+1}) = P_{\xi|X}\left(\hat{p}_{x_{N+1}}^N(u) \ge 1 - \gamma \mid X = x_{N+1} \right)$$
$$\le e^{-I(1-\gamma)b(N)N^r}$$
$$\le e^{-I(1-\gamma-(1-\alpha))b(N)N^r}$$
$$= e^{-I(\alpha - \gamma)b(N)N^r},$$

where the first inequality follows from Theorem 1 by defining

$$\bar{S}_N = \sum_{i=1}^{N} w_i(x_{N+1})\mathbb{1}(G(u, \xi_i) \le 0).$$

9

Since $\alpha < \gamma$, we have

$$\mathbb{E}[\mathbb{1}(G(u,\xi) \leq 0) \mid X = x_{N+1}] = P_{\xi|X}\big(G(u,\xi) \leq 0\big) < 1 - \alpha < 1 - \gamma, \tag{13}$$

satisfying the hypothesis of $y = 1 - \gamma > m$ in Theorem 1. We then have

$$P_{\xi|X}(U_\gamma^N \nsubseteq U_\alpha \mid X = x_{N+1}) = P_{\xi|X}\big(\exists\, u \in U_\gamma^N \mid p(u) < 1 - \alpha, X = x_{N+1}\big)$$
$$\leq \sum_{u \in U \setminus U_\alpha} P_{\xi|X}(u \in U_\gamma^N \mid X = x_{N+1}) \leq |U \setminus U_\alpha| e^{-I(\alpha - \gamma) b(N) N^r}. \quad \square$$

Remark 1. *Theorem 2 establishes that as N grows to infinity one can expect that any feasible solution to* (DDC-CCP) *will also be feasible to* (C-CCP) *as long as $\gamma \in [0, \alpha)$. If the slowly varying function $b(N)$ is a positive constant C, we have that if*

$$N \geq \frac{1}{\frac{\mathbb{E}[Q_x(X)]}{M^*}(\alpha - \gamma)} \left(\frac{1}{C} \ln \left(\frac{|U \setminus U_\gamma|}{\delta} \right) \right)^{1/r}, \tag{14}$$

then with probability at least $1 - \delta$ we can guarantee that a feasible solution to (DDC-CCP) *will be feasible for* (C-CCP).

Remark 2. *Expression (14) is similar to sample size estimates for* (CCP) *that ignores covariates (expression (6) in Luedtke and Ahmed (2008)). However, the $(\alpha - \gamma)$ term in the denominator has a square term for classical CCP problems, implying that the sample size needed to achieve feasibility (or number of data points needed in the contextual case) grows faster for values of α close to γ than in the contextual case. The dependence on the confidence δ is similar in both cases, although in the contextual framework the value of δ could be even smaller due to the r-th root term. The interpretation is that the knowledge of the current value of the feature vector allows for a smaller value of N when compared to the featureless case.*

We now turn our attention to the infinite case. For this case, we will follow Luedtke and Ahmed (2008) and make the following assumptions:

Assumption 3. *There exists $L > 0$ such that*

$$|G(u,\xi) - G(u',\xi)| \leq L\|u - u'\|_\infty \quad \forall u, u' \in U \text{ and } \forall \xi \in \Xi. \tag{15}$$

Assumption 4. *For a fixed $\eta > 0$, the feasible region of* (DDC-CCP) *is redefined as*

$$U_{\alpha,\eta}^N = \left\{ u \in U : \sum_{i=1}^N w_i(x_{N+1}) \mathbb{1}\big(G(u,\xi^i) + \eta \leq 0\big) \geq 1 - \alpha \right\}. \tag{16}$$

In other words, we need to have strict inequality for function $G(\cdot, \cdot)$ across all data points.

Assumption 5. *We assume the set U is bounded, and define*

$$D = \sup\{\|u - u'\|_\infty \; : u, u' \in U\}.$$

Theorem 3. *Suppose that the assumptions in Theorem 1 hold. Moreover, suppose that Assumptions 3, 4 and 5 hold, and let $\gamma \in [0, \alpha), \beta \in (0, \alpha - \gamma)$ and $\eta > 0$. For a covariate vector $x_{N+1} \in \mathbb{R}^p$, we have*

$$P_{\xi|X}(U_{\alpha,\eta}^N \subseteq U_\alpha \mid X = x_{N+1}) \geq 1 - \lceil 1/\beta \rceil \lceil 2LD/\gamma \rceil^n e^{-I(\alpha - \gamma) b(N) N^r}$$

where $U_{\alpha,\eta}^N$ is defined in (16), U_α is defined in Theorem (2), and $I(\cdot)$ is of form (12).

Proof. The proof follows the same steps as the proof of Theorem 10 in Luedtke and Ahmed (2008). The only difference is that equation (12) in that paper is obtained by applying our Theorem 2. □

Remark 3. *Similarly as done for the finite case, fixing α, γ such that $\gamma < \alpha$ and $\beta \in (0, \gamma)$, e.g., $\beta = (\alpha - \gamma)/2$, we have that a feasible solution to (DDC-CCP) is feasible to (C-CCP) with confidence at least $1 - \delta$ if we have*

$$N \geq \frac{1}{\frac{\mathbb{E}[Q_x(X)]}{M^*}(\alpha - \gamma)} \left[\frac{1}{C} \left(\log \frac{1}{\delta} + n \log \left\lceil \frac{2LD}{\gamma} \right\rceil + \log \left\lceil \frac{2}{\alpha - \gamma} \right\rceil \right) \right]^{1/r}. \tag{17}$$

As in (14), the value of δ can be extremely small without significantly affecting the amount of data needed to guarantee feasibility of solutions of (DDC-CCP) to (C-CCP). Similar to the finite case, in (17) the number of data points N depends on the inverse of the difference between α and γ, and not on the square of the difference as shown in Luedtke and Ahmed (2008).

Remark 4. *The function $Q(\cdot)$ in Theorem 1 allows for kernel functions and k-NN as weights, among others. Unfortunately, results from Theorems 2 and 3 do not apply to CART and RF because they require the values of the dependent variable ξ in addition to the features as arguments of function $Q(\cdot)$. However, in our numerical experiments we will work with both methods and show empirically that feasibility is observed in practice.*

5   Case Study: Soccer Hiring Problem

We consider a soccer hiring problem, inspired by the model developed in Pantuso and Hvattum (2019). At the beginning of a season a team has a roster formed by players with active contracts, that is, those belong to the club or have been borrowed from other teams. In addition to the players currently in the roster, which can be sold or returned to their teams, the team has a series of target players that it considers buying or borrowing from other teams. The team wants to define next season's roster, and the *universe of interest* is formed by *current roster* plus *target players*.

The objective is to maximize the team's performance, given by the players' ratings. Ratings are based on the plus-minus system, see Hvattum (2019) for a discussion on plus-minus rating of individual players in sports team, and Sæbø and Hvattum (2019) for a regularized version of the method applied to soccer. We build on the model developed in Pantuso and Hvattum (2019), with data provided by the authors at `https://github.com/GioPan/instancesFTCP` and through personal communications. Their novel approach includes home field advantage, an improved method to take into account red cards, and age-dependent ratings. Additionally, there is a budget constraint and deterministic constraints that define minimum and maximum number of players per technical role/position, as well as the total number of players in the roster.

Randomness in the model comes from the future market value of players. Each player has a known market value, and the team wants to select the next season's roster keeping the market value above a given threshold, with a certain probability. In Pantuso and Hvattum (2019), the authors define a team's market value, and propose a chance constraint that imposes that the sum of the market values of all players in next season's roster has to be greater or equal to a given value, with a certain probability. Scenarios for future market values are constructed parametrically, assuming a functional relationship—a quartic equation—between the future market value (dependent variable) and current value, age, and (technical) role in the team (features, or independent variables).

We adopt a different approach. Following the development of our C-CCP methodology, we avoid any parametric construction and work directly with data. We propose a player-centered

chance constraint formulation that defines desirable market values for each player in the universe of interest of a team. Our model will have one chance constraint per player, and the approximate conditional distribution of the future market value given features will be obtained using the data. For each player of interest we compare his features with those of all players in the data base, and the weight given to each observation is determined by how close the current player's features are to the features of each player in the data base. By repeating this procedure for each player in the universe of interest, we can build weighted estimations of future market value that leverage the information contained in the data without assuming any explicit functional relationship between features and the dependent variable.

5.1 Chance-Constrained Formulation

Before we present our chance-constrained formulation, let us define some notation. Let $\mathcal{P}$ represent the universe of interest of players for a team, where the subscript $p \in \mathcal{P}$ denotes a player. Let the constant Y_p be equal to one if the player belongs to the team at the beginning of the planning season, and zero otherwise. For each $p \in \mathcal{P}$, we let R_p represent the rating of the player p, calculated by an on-field performance measure. Let $\mathcal{R}$ be the set of technical roles, and r be a generic element of this set. For each $r \in \mathcal{R}$, let $\underline{n}_r$ and $\overline{n}_r$ denote a lower and upper bound for the number of players in position r. The team must have a total of n players in the roster. For each target player p, we let V_p^P denote the current purchase price, and we let V_p^B denote the current loan fee (i.e., borrowing fee from another team). Similarly, for each player p currently in the team, we let V_p^S denote the current selling price, and we let V_p^L denote the current loan fee (i.e., loaning out to another team). The random future market value of a player is denoted by the the random variable $\tilde{V}_p$. In addition to these parameters, we define the following decision variables. Binary decision variable y_p represent players that belong to the team, that is, y_p is equal to one if the player $p \in \mathcal{P}$ belongs to the club, and zero otherwise. Binary decision variables v_p^P and v_p^S represent players that are purchased and sold, respectively, and similarly, u_p^B and u_p^L represent players arriving and leaving on loan agreements, respectively.

Given the above notation, our player-centered chance-constrained formulation is as follows:

$$\max \sum_{p \in \mathcal{P}} R_p(y_p + u_p^B - u_p^L) \tag{18}$$

$$\text{s.t.} \quad y_p - v_p^P + v_p^S = Y_p, \qquad\qquad p \in \mathcal{P} \tag{19}$$

$$\sum_{p \in \mathcal{P}} (y_p + u_p^B - u_p^L) = n, \tag{20}$$

$$\sum_{p \in \mathcal{P}} (y_p + u_p^B - u_p^L) \geq \underline{n}_r, \qquad\qquad r \in \mathcal{R} \tag{21}$$

$$\sum_{p \in \mathcal{P}} (y_p + u_p^B - u_p^L) \leq \overline{n}_r, \qquad\qquad r \in \mathcal{R} \tag{22}$$

$$u_p^B + v_p^P \leq 1 - Y_p, \qquad\qquad p \in \mathcal{P} \tag{23}$$

$$u_p^L + v_p^S \leq Y_p, \qquad\qquad p \in \mathcal{P} \tag{24}$$

$$\sum_{p \in \mathcal{P}} (V_p^P v_p^P + V_p^B u_p^B - V_p^S v_p^S - V_p^L u_p^L) \leq B, \tag{25}$$

$$P\left(y_p \tilde{V}_p \geq F_p \mid X_p = x_p\right) \geq y_p(1 - \alpha_p), \qquad\qquad p \in \mathcal{P} \tag{26}$$

$$y_p, u_p^B, u_p^S, v_p^P, v_p^S \in \{0, 1\}. \tag{27}$$

The objective function (18) maximizes the team's total ratings. Constraint (19) ensures that players that belong to the team that remain in the roster are not sold, and it allows players not currently on the roster to be purchased. Constraint (20) forces that the team ends up with exactly n players in the roster for the following season, and constraints (21) and (22) enforce upper and lower bounds for the number of players in each position, $\underline{n}_r$ and $\overline{n}_r$ for $r \in \mathcal{R}$, respectively. Constraint (23) allows buying or arriving on loan for players that are not currently part of the team, while it forbids those decisions for players currently belong to the team. Similarly, constraint (24) allows selling or loaning players currently belonging to the team, while it forbids those decisions for players that are not in the team's roster. The budget constraint (25) limits expenditures on transfers of any type, and the chance constraints in (26) requires that a player joins the team (either by purchase or loan) only if his market value is above a threshold F_p in the next season with probability over $1 - \alpha_p$, where $\alpha_p \in [0, 1]$. Note that the distribution of the future market value $\tilde{V}_p$ is conditional on the current feature vector X_p of the player $p \in \mathcal{P}$ under consideration. Also, note that the chance constraint (26) is active if $y_p = 1$. Finally, the last set of constraints (27) states that all decision variables are binary.

5.2   Data-Driven Prescriptive Formulation

For each chance constraint described in equations (26) in formulation (18)–(27) we do not have an explicit expression for the conditional probability of the future market value given a player's current features. In the classical SAA approach, the features would be ignored, and sampling N data points $\{V_p^i\}_{i=1}^N$ from the random variable $\tilde{V}_p$, the chance-constraint would be approximated by a set of constraints as follows:

$$y_p V_p^i + M z_p^i \geq F_p, \qquad\qquad i = 1, \ldots, N, \qquad\qquad (28)$$

$$\frac{1}{N} \sum_{i=1}^N z_p^i \leq \alpha_p y_p, \qquad\qquad\qquad (29)$$

$$z_p^i \in \{0, 1\}, \qquad\qquad\qquad i = 1, \ldots, N, \qquad\qquad (30)$$

where the binary decision variable z_p^i is equal to zero if the player's future market value V_p^i exceeds the threshold F_p, and M is a conveniently chosen upper bound for $F_p - y_p V_p^i$, e.g., $M = F_p$.

The contextual approach implicitly assumes there is value in the data, and that players with features similar to the current player in consideration should have a weight higher than $1/N$ in (29). We first considered using kernels as in (6) to define those weights, comparing the current features with the features of each player in our database to obtain the weights. However, such approach would be more involved because two of the features are categorical (age and player's position), while only the current market value is numeric.

We decided to take a simpler approach and opted for ML-weights, as described in Table 1. We ran experiments with regression trees (CART) and random forests (RF), which can accommodate categorical variables without additional parameters. For these methods, we tuned the tree depth and minimum leaf size. We ran a cross-validation scheme and picked the values that generated a higher numbers of feasible problems for different choices of α_p and F_p. A similar procedure was done for the RF. The DDC-CCP approximation of our chance-constraint problem will be written exactly as (5), repeating the equations for each chance-constraint.

13

5.3 Experimental Setup

We considered the Premier League, the first tier soccer tournament in the UK, to test our proposed methodology. There were 20 teams competing during the 2013-2014 season, and we considered they were planning for the 2014-2015 season. We assume each team is making decisions that are best for them, irrespective of other teams. In other words, we are not casting the problem as a game where all teams have to compete to form their rosters. A list of teams is given in Table 2.

Table 2: Teams considered for the soccer hiring problem.

Arsenal FC	Aston Villa	Cardiff City	Chelsea FC
Crystal Palace	Everton	Fullham	Hull City
Liverpool	Manchester City	Manchester United	Newcastle United
Norwich City	Southampton	Stoke City	Sunderland
Swansea City	Tottenham Hotspur	West Bromwich Albion	West Ham United

Our data set contains the market value at the beginning of the season, age, role, and market value at the end of the season for each player for the 2011-2012, 2012-2013, and 2013-2014 seasons. Additionally, for each team, we have (i) the current roster, (ii) a specification whether each player in the roster belongs to the team or is under a loan agreement, (iii) a specification whether each player that belongs to the team can be sold or loaned to another team, (iv) the set of target players that the team considers for the forthcoming season, and (v) a specification whether a target player can be borrowed from another team for the forthcoming season.

All the other parameters, including the transfer prices, minimum/maximum number of players in each role, total number of players, and budget are also based on the data in Pantuso and Hvattum (2019).

Finally, for all players $p \in \mathcal{P}$ we assume $1 - \alpha_p = 1 - \alpha$ and $F_p = \rho V_p^C$, where V_p^C is the current market value of player p and ρ is a constant. To run the experiments, we used the following parameters: $1 - \alpha \in \{0.5, 0.6, 0.7, 0.8\}$ and $\rho \in \{0.5, 0.6, 0.7, 0.75\}$, for a total of 16 parameter settings.

To conduct the experiments, we used four approaches to approximate the chance constraints and form a roster for the next season

- **CART:** Each chance constraint (26) in formulation (18)–(27) is conditional on the features of those players, and we approximate conditional probability in each case by using a regression tree.

- **RF:** Similar to CART, but we used RF to approximate the conditional probabilities.

- **SAA:** The SAA method proposed in Pantuso and Hvattum (2019), with a parametric relationship between the dependent and independent variables.

- **nSAA:** The "naïve" SAA approach without considering the contextual information.

As it can be understood, CART, RF, and SAA, as defined above, exploit the contextual information (i.e., current market value, age, and role), while nSAA totally ignores context and uniformly samples future market values from the entries in our data set. Although SAA considers the context, it does it in a different way from CART and RF. For completeness, we briefly explain the SAA approach, presented in Pantuso and Hvattum (2019), below. For each age group $a \in \mathcal{A} := \{(0, 20], (20, 22], (22, 24], (24, 26], (26, 28], (28, 30], (30, 32], (32, \infty)\}$, a (parametric) regression model is trained to capture the relationship between the independent and dependent variables.

Once those models are trained, for each player $p \in \mathcal{P}$ and data point i, $i = 1, \ldots, N$, the future market value V_p^i is obtained as

$$V_p^i = \left(\alpha_a \sqrt[4]{V_p^C} + \sum_{r \in \mathcal{R}} \beta_{ar} \mathbb{1}_r(p) \right)^4 \times (1 + \epsilon_a^i),$$

where $\mathbb{1}_r(p)$ is equal to one if player p belongs to the set of players in position $r \in \mathcal{R}$, and zero otherwise, and ϵ_a^i is an i. i.d. sample from the empirical prediction error distribution of the regression model for the age group $a \in \mathcal{A}$. Parameters α_a and β_{ar} are estimated from the regression model for each age group $a \in \mathcal{A}$.

To train all models using the above four approaches we used the entire data set. To train CART- and RF-based C-CCP models (i.e., approaches `CART` and `RF`), we used trees of depths 8 and 10, guided by some preliminary experiments.

We implemented all models in Python 3.7. All computations were performed using GUROBI 9.0, on a Linux Ubuntu environment, with an Intel Core i7-2640M 2.80 GHz processor and 8.00 GB of RAM. To train `CART` and `RF`, we used scikit-learn 0.23.2 library (Pedregosa et al., 2011).

5.4 Numerical Results

To draw a parallel with the portfolio problem described in Section 3.1, given a team, our features are the union of features of all players in the team's universe of players (roster plus targeted players). Unlike the portfolio problem (7), we are not interested in testing the performance of the `SAA` and `nSAA` solutions (when they exist) under different features. The main question here is whether `SAA` and `nSAA`—which ignore the current features observed for each player in a team's current roster— may yield a feasible solution for a problem with several *conditional* chance-constraints. More precisely, we are interested in the feasibility of `SAA` and `nSAA` solutions when computed as in (3), which we will refer to as CART chance constraint (CART-CC).

We report the results for different teams and under each of the 16 parameter choices of the pairs $1 - \alpha$ and ρ, and two tree depths 8 and 10. First, for a selected number of teams, we investigate the chance-constrained programming model using all approaches, `nSAA`, `SAA`, `CART`, `RF`. If `nSAA` and `SAA` yield a feasible solution, we investigate the *conditional* probability of satisfying chance constraints among all players who ended up being in the roster for the next season, evaluated with respect to the CART-CC. Finally, we investigate the number of feasible solutions obtained for all teams and approaches.

Tables 3–5 report detailed results for three selected teams, Crystal Palace, Chelsea, and Tottenham Hotspur. The column "Opt. Val" in these tables reports the ratio of the optimal value of a model with respect to the maximum optimal value over all models. For instance, for a fixed $1 - \alpha$ and ρ, the number under column "Opt. Val" for model `nSAA` reports the ratio of the optimal value of `nSAA` with respect to the maximum optimal value over all models. The column "Chance" reports the minimum probability of satisfying chance constraints among all players in the roster for the next season.

Observe from Table 3 that `nSAA`—the model that completely ignores the contextual information— yields a feasible solution for the corresponding model in all cases. In some cases, this solution is even comparable to those of `CART` and `RF` in terms of the optimal value. However, `nSAA`'s solution is often infeasible with respect to the CART-CC. In particular, using a tree of depth 10 resulted in a solution for the case $1 - \alpha = 0.5$ and $\rho = 0.6$ that has the highest optimal value among all models (tie with `RF`), but such a solution is not feasible when contextual information is taken into account.

15

Table 3: Optimal values and feasibility for different models for Crystal Palace.

(a) Depth=8.

		Opt. Val				Chance	
$1-\alpha$	ρ	nSAA	SAA	CART	RF	nSAA	SAA
0.5	0.50	1.000	0.862	1.000	1.000	0.71	0.50
	0.60	1.000	0.819	1.000	1.000	0.60	0.90
	0.70	0.990	0.798	1.000	1.000	†0.15	0.60
	0.75	0.993	0.801	1.000	1.000	†0.10	0.60
0.6	0.50	1.000	0.738	1.000	1.000	0.71	0.98
	0.60	0.962	0.775	1.000	1.000	†0.27	0.88
	0.70	0.993	0.630	1.000	1.000	†0.15	0.60
	0.75	0.970	-	1.000	0.964	†0.10	-
0.7	0.50	0.916	-	1.000	1.000	†0.50	-
	0.60	0.852	-	1.000	1.000	†0.27	-
	0.70	0.858	-	0.984	1.000	†0.15	-
	0.75	0.763	-	0.984	1.000	†0.10	-
0.8	0.50	0.828	-	1.000	0.972	†0.50	-
	0.60	0.733	-	1.000	0.997	†0.27	-
	0.70	0.718	-	1.000	0.932	†0.60	-
	0.75	0.925	-	1.000	0.951	†0.60	-

- : indicates infeasibility in corresponding model;
† : indicates infeasibility for the CART-CC.

(b) Depth=10.

		Opt. Val				Chance	
$1-\alpha$	ρ	nSAA	SAA	CART	RF	nSAA	SAA
0.5	0.50	1.000	0.862	1.000	1.000	0.55	†0.38
	0.60	1.000	0.819	0.983	1.000	†0.36	0.84
	0.70	0.990	0.798	1.000	1.000	†0.06	0.60
	0.75	0.990	0.798	1.000	0.997	†0.00	0.60
0.6	0.50	1.000	0.738	0.983	1.000	0.55	0.98
	0.60	0.990	0.798	1.000	1.000	†0.12	0.84
	0.70	0.990	0.628	1.000	0.997	†0.06	0.60
	0.75	0.970	-	1.000	1.000	†0.00	-
0.7	0.50	0.962	-	0.972	1.000	†0.38	-
	0.60	0.852	-	1.000	1.000	†0.12	-
	0.70	0.858	-	0.984	1.000	†0.06	-
	0.75	0.763	-	0.944	1.000	†0.00	-
0.8	0.50	0.852	-	1.000	1.000	†0.38	-
	0.60	0.733	-	1.000	0.997	†0.12	-
	0.70	0.754	-	1.000	0.980	†0.60	-
	0.75	0.862	-	1.000	0.937	†0.60	-

- : indicates infeasibility in corresponding model;
† : indicates infeasibility for the CART-CC.

In Table 3, the SAA model only yields a feasible solution in half of the cases, and that solution was mostly (except for one case) feasible to CART-CC. Given that SAA considers the contextual information, albeit via a parametric model, it seems to be the case that if SAA yields a feasible solution, that solution is be likely to remain feasible with respect to the CART-CC. However, Table 4 reveals that this is not always the case. While SAA yields a feasible solution for around half of the cases, that solution was mostly (except for one case) infeasible with respect to the CART-CC.

Table 4: Optimal values and feasibility for different models for Chelsea FC.

(a) Depth=8.

		Opt. Val				Chance	
$1-\alpha$	ρ	nSAA	SAA	CART	RF	nSAA	SAA
0.5	0.50	-	0.995	1.000	0.974	-	0.50
	0.60	-	1.000	0.946	0.946	-	†0.27
	0.70	-	1.000	0.856	0.919	-	†0.15
	0.75	-	1.000	0.915	0.915	-	†0.10
0.6	0.50	-	1.000	0.991	0.992	-	†0.50
	0.60	-	0.977	1.000	0.948	-	†0.27
	0.70	-	-	1.000	1.000	-	-
	0.75	-	-	1.000	1.000	-	-
0.7	0.50	-	-	1.000	1.000	-	-
	0.60	-	-	0.932	1.000	-	-
	0.70	-	-	0.957	1.000	-	-
	0.75	-	-	0.957	1.000	-	-
0.8	0.50	-	-	1.000	0.986	-	-
	0.60	-	-	0.957	1.000	-	-
	0.70	-	-	0.974	1.000	-	-
	0.75	-	-	-	-	-	-

‾ : indicates infeasibility in corresponding model;
† : indicates infeasibility for the CART-CC.

(b) Depth=10.

		Opt. Val				Chance	
$1-\alpha$	ρ	nSAA	SAA	CART	RF	nSAA	SAA
0.5	0.50	-	1.000	0.979	0.980	-	†0.38
	0.60	-	1.000	0.946	0.946	-	†0.12
	0.70	-	1.000	0.919	0.919	-	†0.06
	0.75	-	1.000	0.982	0.915	-	†0.00
0.6	0.50	-	1.000	0.991	0.992	-	†0.38
	0.60	-	1.000	0.983	0.983	-	†0.12
	0.70	-	-	1.000	0.932	-	-
	0.75	-	-	1.000	0.932	-	-
0.7	0.50	-	-	1.000	1.000	-	-
	0.60	-	-	1.000	1.000	-	-
	0.70	-	-	1.000	0.971	-	-
	0.75	-	-	1.000	0.971	-	-
0.8	0.50	-	-	1.000	0.948	-	-
	0.60	-	-	1.000	0.971	-	-
	0.70	-	-	1.000	0.909	-	-
	0.75	-	-	-	-	-	-

‾ : indicates infeasibility in corresponding model;
† : indicates infeasibility for the CART-CC.

Interestingly, such behavior occurs despite the fact that the SAA's solution had the highest optimal value among all models in most cases. In the single case where SAA yields a feasible solution to the CART-CC ($1-\alpha = 0.5$, $\rho = 0.5$ and the tree depth is 8, see Table 4a), the objective function of the SAA's solution is within 0.5% of the optimal value. Nevertheless, the results clearly show that SAA is not a reliable method to generate feasible solutions for C-CCP problems.

Another noteworthy observation is that nSAA did not generate a feasible solution to the cor-

Table 5: Optimal values and feasibility for different models for Tottenham Hotspur.

(a) Depth=8.

		Opt. Val				Chance	
$1-\alpha$	ρ	nSAA	SAA	CART	RF	nSAA	SAA
0.5	0.50	0.740	0.951	1.000	0.951	0.50	0.50
	0.60	-	0.994	1.000	1.000	-	0.60
	0.70	-	1.000	0.911	0.942	-	†0.43
	0.75	-	0.991	0.967	1.000	-	†0.43
0.6	0.50	-	0.916	1.000	1.000	-	0.80
	0.60	-	0.897	1.000	1.000	-	0.62
	0.70	-	0.743	1.000	1.000	-	0.60
	0.75	-	-	1.000	1.000	-	-
0.7	0.50	-	-	1.000	0.977	-	-
	0.60	-	-	0.967	1.000	-	-
	0.70	-	-	0.952	1.000	-	-
	0.75	-	-	0.952	1.000	-	-
0.8	0.50	-	-	1.000	0.946	-	-
	0.60	-	-	0.975	1.000	-	-
	0.70	-	-	1.000	-	-	-
	0.75	-	-	-	-	-	-

$^{-}$: indicates infeasibility in corresponding model;
† : indicates infeasibility for the CART-CC.

(b) Depth=10.

		Opt. Val				Chance	
$1-\alpha$	ρ	nSAA	SAA	CART	RF	nSAA	SAA
0.5	0.50	0.778	1.000	1.000	1.000	†0.38	†0.38
	0.60	-	0.994	0.945	1.000	-	†0.36
	0.70	-	1.000	0.942	0.942	-	†0.36
	0.75	-	0.991	1.000	1.000	-	0.60
0.6	0.50	-	0.916	0.945	1.000	-	0.94
	0.60	-	0.948	1.000	1.000	-	0.75
	0.70	-	0.718	1.000	0.967	-	0.60
	0.75	-	-	1.000	0.967	-	-
0.7	0.50	-	-	0.946	1.000	-	-
	0.60	-	-	1.000	1.000	-	-
	0.70	-	-	0.986	1.000	-	-
	0.75	-	-	0.986	1.000	-	-
0.8	0.50	-	-	1.000	1.000	-	-
	0.60	-	-	1.000	0.991	-	-
	0.70	-	-	1.000	-	-	-
	0.75	-	-	-	-	-	-

$^{-}$: indicates infeasibility in corresponding model;
† : indicates infeasibility for the CART-CC.

responding model in almost all cases (except for two) for Chelsea FC (Table 4) and Tottenham Hotspur (Table 5). As it can be seen from Tables 3–5, while nSAA and SAA approaches rarely generate a feasible solution for the corresponding model and for the CART-CC, CART and RF only failed to yield a feasible solution around 5% of the time (10 out of 192 total experiments with the 3 teams for CART and RF).

We now focus on the frequency of infeasibility for all different approaches for all teams, and

investigate whether approaches `CART` and `RF` yield feasible solutions more often than `SAA` and `nSAA`. The results are reported in Table 6 and 7 for depth 8 and depth 10, respectively. In these tables, the results are categorized based on the value of $1 - \alpha$, in addition to the approach and team. For each value of $1 - \alpha$, there are four experiments per approach and team, corresponding to the four choices of $\rho \in \{0.5, 0.6, 0.7, 0.75\}$. Each cell in these tables show the number of times (out of four) that the approach yields either an infeasible solution for the corresponding approach, or with respect to the CART-CC. Observe that `nSAA` and `SAA` generally resulted in infeasibility, much more often than `CART` and `RF`. The only exception is for the team Everton FC, where the total number of infeasibility for `CART` and `RF` is comparable to that for `nSAA` and `SAA`, albeit still smaller. A summary of these results is reported in Table 8, categorized by tree depth and approaches. We can observe that `nSAA` and `SAA` suffer from infeasibility at least 7 times more than `CART` and `RF`.

5.5   Discussion

As mentioned earlier, our focus in this study is on feasibility of solutions obtained via approximations, similar to most of the literature in chance-constrained programming. The question is whether approaches that ignore features (i.e., `nSAA` and `SAA`) can generate feasible solutions in a frequency comparable to those of contextual approaches (i.e., `CART` and `RF`). In the case where all approximations yielded feasible solutions, it is unclear how solutions can be compared because none of the teams formed by the optimization model actually existed. We believe the only way to compare those solutions is to simulate the teams created in a fantasy sport environment, and measure the performance of each synthetic team after a predefined number of rounds. In our case, when `nSAA` or `SAA` obtained solutions, and those solutions were feasible for the CART-CC, we evaluate their quality by comparing it to the prescriptive solution. If the objective function is, say 5% away from the prescriptive optimal value, then the value of considering a contextual model is reduced. Examples of this situation occurred with `SAA` for Chelsea, where $1 - \alpha = 0.5$, $\rho = 0.5$, and tree depth is 8 (see Table 4a), and for Tottenham Hotspur, where $1 - \alpha = 0.5$, $\rho = 0.5$, and tree depth is 8 (see Table 5a). In any case, our experiments show that infeasibility is a great hurdle for `nSAA` and `SAA` when contextual information is present.

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. In this paper, we propose a contextual chance-constrained formulation that accommodates features in addition to the dependent random variable. We illustrate our framework in a portfolio selection problem, showing that by ignoring features the decision-maker may end up with an infeasible—and therefore, non-implementable—solution.

   We propose and discuss the advantages and disadvantages of different data-driven approximations for the contextual formulation, based on kernel functions and on machine learning. In both cases, we have tractable formulations that contain weighted sums of random variables, which require the same computational effort as classical SAA approximations since all weight calculations can be done offline. Building on large deviation results for those sums, we derive convergence results that guarantee that the feasible set of the approximate problem is contained in that of the true problem with probability approaching one exponentially fast. We also provide a data-size estimate based on the theoretical results.

   We apply our methodology to a soccer hiring problem, using data from the Premier League, UK's top-tier tournament. The problem consists in defining a team's roster for next year's season

Table 6: Frequency of infeasibility for different approaches and teams, and using a tree with depth 8.

$1-\alpha$	Arsenal FC				Aston Villa				Cardiff City				Chelsea FC			
	nSAA	SAA	CART	RF	nSAA	SAA	CART	RF	nSAA	SAA	CART	RF	nSAA	SAA	CART	RF
0.5	4	2	0	0	3	2	0	0	3	0	0	0	4	3	0	0
0.6	4	3	0	0	4	3	0	0	4	2	0	0	4	4	0	0
0.7	4	4	0	0	4	4	0	0	4	4	0	0	4	4	0	0
0.8	4	4	1	1	4	4	1	2	4	4	0	1	4	4	1	1
Total	16	13	1	1	15	13	1	2	15	10	0	1	16	15	1	1

$1-\alpha$	Crystal Palace				Everton FC				Fulham FC				Hull City			
	nSAA	SAA	CART	RF	nSAA	SAA	CART	RF	nSAA	SAA	CART	RF	nSAA	SAA	CART	RF
0.5	2	0	0	0	3	3	2	2	3	3	0	0	3	3	0	0
0.6	3	1	0	0	4	4	2	3	4	4	0	0	4	4	0	0
0.7	4	4	0	0	4	4	3	3	4	4	0	0	4	4	0	0
0.8	4	4	0	0	4	4	3	4	4	4	1	0	4	4	0	1
Total	13	9	0	0	15	15	10	12	15	15	1	0	15	15	0	1

$1-\alpha$	Liverpool FC				Manchester City				Manchester United				Newcastle United			
	nSAA	SAA	CART	RF	nSAA	SAA	CART	RF	nSAA	SAA	CART	RF	nSAA	SAA	CART	RF
0.5	3	3	0	0	4	3	0	0	4	3	0	0	3	3	0	0
0.6	4	4	0	2	4	3	0	0	4	3	0	0	4	4	0	0
0.7	4	4	2	3	4	4	0	0	4	4	2	0	4	4	0	0
0.8	4	4	2	3	4	4	1	2	4	4	2	1	4	4	1	1
Total	15	15	4	8	16	14	1	2	16	14	4	1	15	15	1	1

$1-\alpha$	Norwich City				Southampton FC				Stoke City				Sunderland AFC			
	nSAA	SAA	CART	RF	nSAA	SAA	CART	RF	nSAA	SAA	CART	RF	nSAA	SAA	CART	RF
0.5	3	2	0	0	3	2	0	0	3	3	0	0	3	3	0	0
0.6	4	3	0	0	4	3	0	0	4	4	0	0	4	4	0	0
0.7	4	4	0	0	4	4	0	0	4	4	0	0	4	4	0	0
0.8	4	4	1	1	4	4	1	1	4	4	0	0	4	4	0	0
Total	15	13	1	1	15	13	1	1	15	15	0	0	15	15	0	0

$1-\alpha$	Swansea City				Tottenham Hotspur				West Bromwich Albion				West Ham United			
	nSAA	SAA	CART	RF	nSAA	SAA	CART	RF	nSAA	SAA	CART	RF	nSAA	SAA	CART	RF
0.5	2	2	0	0	3	2	0	0	2	2	0	0	3	3	0	0
0.6	4	3	0	0	4	1	0	0	3	3	0	0	4	4	0	0
0.7	4	4	0	0	4	4	0	0	4	4	0	0	4	4	0	0
0.8	4	4	1	2	4	4	1	2	4	4	1	2	4	4	0	0
Total	14	13	1	2	15	11	1	2	13	13	1	2	15	15	0	0

by maximizing performance based on data on player's ratings. Aside from deterministic constraints related to the team's formation, the decision-maker wants to ensure that for each player his market value does not decrease beyond a certain threshold, with high probability. We perform experiments with all teams that participated in the 2013-2014 season, and compare results for nSAA, SAA, and the contextual chance-constrained approach using classification and regression trees, and random forest. The results show that approaches that ignore contextual information usually generate infeasible solutions, and when candidate solutions are obtained they do not satisfy the chance constraint

Table 7: Frequency of infeasibility for different approaches and teams, and using a tree with depth 10.

$1-\alpha$	Arsenal FC				Aston Villa				Cardiff City				Chelsea FC			
	nSAA	SAA	CART	RF	nSAA	SAA	CART	RF	nSAA	SAA	CART	RF	nSAA	SAA	CART	RF
0.5	4	0	0	0	3	0	0	0	3	0	0	0	4	4	0	0
0.6	4	3	0	0	3	3	0	0	4	2	0	0	4	4	0	0
0.7	4	4	0	0	4	4	0	0	4	4	0	0	4	4	0	0
0.8	4	4	2	1	4	4	1	2	4	4	0	0	4	4	1	1
Total	16	11	2	1	14	11	1	2	15	10	0	0	16	16	1	1

$1-\alpha$	Crystal Palace				Everton FC				Fulham FC				Hull City			
	nSAA	SAA	CART	RF	nSAA	SAA	CART	RF	nSAA	SAA	CART	RF	nSAA	SAA	CART	RF
0.5	3	1	0	0	4	4	3	2	4	4	0	0	3	4	0	0
0.6	4	1	0	0	4	4	4	3	4	4	0	0	3	3	0	0
0.7	4	4	0	0	4	4	4	3	4	4	0	0	4	4	0	0
0.8	4	4	0	0	4	4	4	4	4	4	1	0	4	4	0	0
Total	15	10	0	0	16	16	15	12	16	16	1	0	14	15	0	0

$1-\alpha$	Liverpool FC				Manchester City				Manchester United				Newcastle United			
	nSAA	SAA	CART	RF	nSAA	SAA	CART	RF	nSAA	SAA	CART	RF	nSAA	SAA	CART	RF
0.5	4	4	0	0	4	3	0	0	4	4	0	0	4	4	0	0
0.6	4	4	0	2	4	3	0	0	4	3	0	0	4	4	0	0
0.7	4	4	2	3	4	4	0	0	4	4	2	0	4	4	0	0
0.8	4	4	2	3	4	4	1	2	4	4	2	1	4	4	1	1
Total	16	16	4	8	16	14	1	2	16	15	4	1	16	16	1	1

$1-\alpha$	Norwich City				Southampton FC				Stoke City				Sunderland AFC			
	nSAA	SAA	CART	RF	nSAA	SAA	CART	RF	nSAA	SAA	CART	RF	nSAA	SAA	CART	RF
0.5	3	0	0	0	3	0	0	0	4	4	0	0	4	4	0	0
0.6	3	2	0	0	4	3	0	0	4	4	0	0	4	3	0	0
0.7	4	4	0	0	4	4	0	0	4	4	0	0	4	4	0	0
0.8	4	4	0	1	4	4	0	1	4	4	0	0	4	4	0	0
Total	14	10	0	1	15	11	0	1	16	16	0	0	16	15	0	0

$1-\alpha$	Swansea City				Tottenham Hotspur				West Bromwich Albion				West Ham-United			
	nSAA	SAA	CART	RF	nSAA	SAA	CART	RF	nSAA	SAA	CART	RF	nSAA	SAA	CART	RF
0.5	3	3	0	0	4	3	0	0	3	3	0	0	4	4	0	0
0.6	4	4	0	0	4	1	0	0	4	3	0	0	4	4	0	0
0.7	4	4	0	0	4	4	0	0	4	4	1	0	4	4	0	0
0.8	4	4	1	2	4	4	1	2	4	4	2	1	4	4	0	0
Total	15	15	1	2	16	12	1	2	15	14	3	1	16	16	0	0

Table 8: Percentage of the frequency of infeasibility for different approaches and tree depths.

Tree depth	nSAA	SAA	CART	RF
8	95%	85%	9%	12%
10	97%	86%	11%	11%

when features are taken into account. By using `CART` and `RF` we obtained feasible solutions for all teams, and for most parameter choices, validating our methodology.

Future work includes the study of contextual risk aversion problems, and extensions to the dynamic case. It would also be interesting to investigate if our approach can accommodate decision-dependent uncertainty problems. In terms of applications, we plan to explore problems in energy systems, transportation and finance. For instance portfolio selection problems are natural candidates for contextual formulations due to the wide availability of feature data, such as unemployment rates, GDP growth, among others.

Acknowledgments

References

Aitchison, J. and C. G. Aitken (1976). Multivariate binary discrimination by the kernel method. *Biometrika 63*(3), 413–420.

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.

Andrieu, L., R. Henrion, and W. Römisch (2010). A model for dynamic chance constraints in hydro power reservoir management. *European Journal of Operational Research 207*(2), 579–589.

Aurzada, F. (2020). Large deviations for infinite weighted sums of stretched exponential random variables. *Journal of Mathematical Analysis and Applications*, 123814.

Ban, G.-Y. and C. Rudin (2018). The big data newsvendor: Practical insights from machine learning. *Operations Research*.

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.

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

Bertsimas, D., N. Kallus, and A. Hussain (2016). Inventory management in the era of big data. *Production and Operations Management 25*(12), 2006–2009.

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*, Volume 31, pp. 2962–2970. Curran Associates, Inc.

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

Bertsimas, D., C. McCord, and B. Sturt (2019). Dynamic optimization with side information. `arXiv preprint arXiv:1907.07307[math.OC]`.

Bertsimas, D. and B. Van Parys (2017). Bootstrap robust prescriptive analytics. `arXiv preprint arXiv:1711.09974 [math.OC]`.

Bonin, O. (2002). Large deviation theorems for weighted sums applied to a geographical problem. *Journal of applied probability 39*(2), 251–260.

Book, S. A. (1972). Large deviation probabilities for weighted sums. *The Annals of Mathematical Statistics*, 1221–1234.

Campi, M. C. and S. Garatti (2008). The exact feasibility of randomized solutions of uncertain convex programs. *SIAM Journal on Optimization 19*(3), 1211–1230.

Campi, M. C., S. Garatti, and M. Prandini (2009). The scenario approach for systems and control design. *Annual Reviews in Control 33*(2), 149–157.

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

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

Cui, X., S. Zhu, X. Sun, and D. Li (2013). Nonlinear portfolio selection using approximate parametric value-at-risk. *Journal of Banking & Finance 37*(6), 2124–2139.

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

den Hertog, D. and K. Postek (2016). Bridging the gap between predictive and prescriptive analytics-new optimization methodology needed. Technical report, Technical report, Tilburg University, Netherlands, 2016. Available at: http . . . .

Diao, S. and S. Sen (2020). Distribution-free algorithms for learning enabled predictive stochastic programming. Available at `http://www.optimization-online.org/DB_HTML/2020/03/7661.html`.

Donini, M., L. Oneto, S. Ben-David, J. S. Shawe-Taylor, and M. Pontil (2018). Empirical risk minimization under fairness constraints. In *Advances in Neural Information Processing Systems*, pp. 2791–2801.

El Balghiti, O., A. N. Elmachtoub, P. Grigas, and A. Tewari (2019). Generalization bounds in the predict-then-optimize framework. In *Advances in Neural Information Processing Systems*, pp. 14412–14421.

Elmachtoub, A. N. and P. Grigas (2017). Smart" predict, then optimize". *arXiv preprint arXiv:1710.08005*.

Embrechts, P., C. Klüppelberg, and T. Mikosch (2013). *Modelling extremal events: for insurance and finance*, Volume 33. Springer Science & Business Media.

Gantert, N., K. Ramanan, F. Rembart, et al. (2014). Large deviations for weighted sums of stretched exponential random variables. *Electron. Commun. Probab 19*(41), 1–14.

Gao, F. (2003). Moderate deviations and Erdös–Rényi–Shepp laws for weighted sums. *Journal of mathematical analysis and applications 285*(1), 1–7.

Giuliano, R. and C. Macci (2011). Large deviation principles for sequences of logarithmically weighted means. *Journal of mathematical analysis and applications 378*(2), 555–570.

Giuliano, R., C. Macci, and B. Pacchiarotti (2014). On large deviations for some sequences of weighted means of gaussian processes. *Journal of Mathematical Analysis and Applications 414*(2), 612–631.

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.

Hayfield, T. and J. Racine (2008). Nonparametric econometrics: The np package. *Journal of statistical software 27*(5), 1–32.

Holland, C. P., S. C. Thornton, and P. Naudé (2020). B2b analytics in the airline market: Harnessing the power of consumer big data. *Industrial Marketing Management 86*, 52–64.

Hvattum, L. M. (2019). A comprehensive review of plus-minus ratings for evaluating individual players in team sports. *International Journal of Computer Science in Sport 18*(1), 1–23.

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.

Jiang, R. and Y. Guan (2016). Data-driven chance constrained stochastic program. *Mathematical Programming 158*(1-2), 291–327.

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`.

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.

Kiesel, R. and U. Stadtmüller (2000). A large deviation principle for weighted sums of independent identically distributed random variables. *Journal of mathematical analysis and applications 251*(2), 929–939.

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

Laherrere, J. and D. Sornette (1998). Stretched exponential distributions in nature and economy:"fat tails" with characteristic scales. *The European Physical Journal B-Condensed Matter and Complex Systems 2*(4), 525–539.

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*.

Luedtke, J. and S. Ahmed (2008). A sample approximation approach for optimization with probabilistic constraints. *SIAM Journal on Optimization 19*(2), 674–699.

Meller, J., F. Taigel, and R. Pibernik (2018). Prescriptive analytics for inventory management: A comparison of new approaches. *Available at SSRN 3229105*.

Nagaev, A. V. (1969). Integral limit theorems taking large deviations into account when cramér's condition does not hold. i. *Theory of Probability & Its Applications 14*(1), 51–64.

Nash, D. B. (2014). Harnessing the power of big data in healthcare. *American health & drug benefits 7*(2), 69.

Nguyen, V. A., F. Zhang, J. Blanchet, E. Delage, and Y. Ye (2020). Distributionally robust local non-parametric conditional estimation. *Advances in Neural Information Processing Systems 33*.

Pagnoncelli, B. K., S. Ahmed, and A. Shapiro (2009). Sample average approximation method for chance constrained programming: theory and applications. *Journal of optimization theory and applications 142*(2), 399–416.

Pantuso, G. and L. M. Hvattum (2019). Maximizing performance with an eye on the finances: a chance-constrained model for football transfer market decisions. *arXiv preprint arXiv:1911.04689*.

Parkins, D. (2017). The world's most valuable resource is no longer oil, but data. *The Economist*.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research 12*, 2825–2830.

Peña-Ordieres, A., J. R. Luedtke, and A. Wächter (2019). Solving chance-constrained problems via a smooth sample-based nonlinear approximation. *arXiv preprint arXiv:1905.07377*.

Sæbø, O. D. and L. M. Hvattum (2019). Modelling the financial contribution of soccer players to their clubs. *Journal of Sports Analytics 5*(1), 23–34.

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. *Information & Management 56*(6), 103135.

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

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.

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 Transactions on Power Systems 33*(6), 6487–6500.

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

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*, Volume 1, pp. 68–76. IEEE.

Wang, W. and S. Ahmed (2008). Sample average approximation of expected value constrained stochastic programs. *Operations Research Letters 36*(5), 515–519.

Wolf, W. (1980). Some remarks on large deviations for weighted sums if cramér's condition is not satisfied. *Mathematical Statistics 6*, 347–352.

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

Xie, W. and S. Ahmed (2017). Distributionally robust chance constrained optimal power flow with renewables: A conic reformulation. *IEEE Transactions on Power Systems 33*(2), 1860–1867.

Zhao, P. and Q. Xiao (2016). Portfolio selection problem with value-at-risk constraints under non-extensive statistical mechanics. *Journal of computational and applied mathematics 298*, 64–71.