Contextual Stochastic Programs with Expected-Value Constraints

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

Expected-value-constrained programming (ECP) formulations are a broad class of stochastic programming problems including integrated chance constraints, risk models, and stochastic dominance formulations. Given the wide availability of data, it is common in applications to have independent contextual information associated with the target or dependent random variables of the problem. We show how to incorporate such information to efficiently approximate ECPs, and prove that the solution set of the approximate problem approaches the true solution set exponentially fast. We illustrate our approach with a portfolio optimization problem that exemplifies the importance of taking contextual information into account in problems with expected-value constraints.

Keywords: Expected-value constraints, Data-driven optimization, Stochastic programming, Large deviations

1 Introduction

In virtually all contexts where decisions need to be made, uncertainty must be taken into account. Transportation problems can have uncertain demand, capacity, and lead times, energy problems can have both inflow and demand as random elements, while portfolio problems typically have unknown returns. While many possible alternatives exist to handle uncertainty, e.g., simply replacing them with their averages, stochastic programming (SP) is one of most the popular approaches to incorporate randomness into the decision-making process.

Classical models in SP assume a known distribution of the uncertainty and aim at finding the best course of action *before* randomness is revealed. For instance, in two-stage models, after the first-stage decision is made uncertainty becomes known, and a recourse action can be taken to ensure feasibility if needed. A significant challenge with SP problems is tractability. Computing in closed form the expected value or the risk of a random variable that depends on decisions is usually impossible; even the evaluation of feasibility for a given solution can be prohibitively time-consuming. Alternative approaches such as robust optimization [3] generate tractable problems by assuming randomness lies in an uncertainty set and the decision maker has a budget that controls the trade-off between protection and performance. Despite the challenges, the popular sample average approximation (SAA) approach converts an SP problem into a tractable optimization problem, and convergence results state that the set of optimal solutions and the optimal value of

*hrahimi@clemson.edu

†bernardo.pagnoncelli@uai.cl

those approximations converge to their true counterparts in the original formulation (see [19] for the analysis of two-stage problems, [23] for chance-constrained optimization, and [30] for problems with expected-value constraints).

The use of data in most applications of SP has been restricted to considering past realizations of uncertainty to estimate the distribution of the random parameters of the model. For example, in [27], the authors use historical data to model demand by commodity and location for a disaster relief application. In [25], the authors use inflow data from 1987 to 2006 to construct a Markov chain that is the only source of uncertainty in the model (demand is assumed to be known). In the last few years, we have witnessed a change in the use of data in SP. A series of recent publications [2, 4, 5, 6, 7, 8, 10, 11, 12, 15, 22, 24] advocates for the incorporation of contextual information into the models. Contextual information, or features, are additional measurements that are taken by surveys, sensors, or by simply tracking and adding independent variables that might be related to the random quantity of interest. The premise is that the wide availability of data, which usually comes with features, should be harnessed to give a better estimation of uncertainty, ultimately allowing for better decisions.

Suppose one is trying to model the default risk of companies in a certain sector of the economy or geographical region in order to make investment decisions. The expected losses of the portfolio are bounded by a given amount, selected by the decision-maker. The standard practice would be to estimate the joint distribution of the default risk of the companies under consideration and find an optimal allocation that maximizes expected returns while controlling the losses. It is often the case that one has access to contextual information *before* making a decision, including for instance inflation rates, unemployment, interest rates, and sentiment analysis extracted from social networks. A better model would take this observation of features into account, considering the *conditional* expected losses instead of the unconditional ones. The solution obtained by ignoring contextual information might not be feasible under all realizations of the contextual variables.

In this work, we will focus on contextual expected-value-constrained programming (ECP) problems, a broad class that includes integrated chance constraints (ICCs), risk-constrained models, and problems with stochastic dominance constraints. Contextual chance-constrained programming, which was studied in [26], is a different modeling framework because the focus is on *qualitative* behavior. The final solution satisfies the chance constraint with high probability, but nothing is said about the amount of the violation in the cases where it is not met. On the other hand, ECP problems allow for constraint violation as long as it does not exceed some pre-defined amount, which translates into a *quantitaive* behavior.

Our first contribution in this paper is a formulation for ECP problems that include contextual information. Our second contribution is to approximate those problems using k-nearest neighbors (kNN) and derive theoretical results that characterize the solution quality. We show feasibility results, proving that a feasible solution to the approximate contextual ECP problem is feasible to the true contextual ECP 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 feasibility for the true problem with high confidence. We show that a kNN-based approximate leads to probabilistic guarantees with an exponential rate of convergence as the number of data points increases.

We present a detailed computational study of a portfolio selection problem and compare our results with the naïve SAA approach, which ignores features and only uses the samples from the random parameters 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 ECPs and show that several important classes of problems in the literature can be cast as those problems. Moreover, we formally present a contextual ECP and describe a method to approximate this problem in a data-driven fashion. In Section 3, we provide theoretical results on how a data-driven approximation of contextual ECPs relates to the true problem in terms of the feasibility of resulted solutions. We then present numerical experiments in Section 4. Finally, we end with conclusions in Section 5.

Notation: We use $\mathbb{1}\{Z\}$ to denote the indicator function which takes value one when Z holds and zero otherwise. We use $\mathcal{N}_\eta(u)$ to denote the open ball with center u and radius η. We let $[n]$ denote the index set $\{1,\ldots,n\}$. A random variable Z is said to be sub-Gaussian with variance proxy σ^2 if $\mathbb{E}[Z]=0$ and $\mathbb{E}[\exp\{tZ\}] \le \exp\{\frac{t^2\sigma^2}{2}\}$ for all $t\in\mathbb{R}$. Function $u\mapsto Z(u)$ is L-Lipschitz if there exists $L>0$ such that $|Z(u)-Z(u')| \le L\|u-u'\|_p$ $\forall u,u'$ and for some $p\ge 0$. 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 Contextual Expected-Value-Constrained Programming

Consider an ECP problem as

$$\begin{aligned} \min_{u\in\mathcal{U}} \quad & f(u) \\ \text{s.t.} \quad & \mathbb{E}_P\left[G_t(u,Y)\right] \ge \epsilon_t,\ t\in\mathcal{T}, \end{aligned} \tag{ECP}$$

where $\mathcal{U}\subset\mathbb{R}^{d_u}$ is the feasible set, 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 ϵ_t is a target value, $t\in\mathcal{T}$. Moreover, $G_t:\mathbb{R}^{d_u}\times Y\to\mathbb{R}$ is a random function for $t\in\mathcal{T}$; that is, $G_t(u,\cdot)$ is measurable for $u\in\mathbb{R}^{d_u}$. Without loss of generality, we assume that $\epsilon_t=\epsilon$, $t\in\mathcal{T}$.

In Section 2.1, we present several examples that can be formulated as an ECP. Then, in Section 2.2, we present the SAA approach to approximate an ECP. Finally, in Section 2.3, we formally present a contextual ECP and describe a data-driven approximation approach.

2.1 Examples

In this section, we show several examples of fundamental problems in SP that can be written as an ECP. We start with integrated chance constraints, the quantitative counterpart of classical chance constraints. We then discuss risk measures and show three examples of popular ones that fit within our framework. Finally, we show that a class of stochastic dominance problems can also be written as ECPs.

2.1.1 Integrated Chance Constraint (ICC)

ICCs were proposed by [16] as a quantitative alternative to traditional chance constraints. The most common form is the linear one: let

$$\eta_t(u,Y) := u^\top w_t(Y) - h_t(Y),\ t\in\mathcal{T},$$

with $\mathcal{T}=\{1,\ldots,T\}$ be a list of goals we would like to satisfy, where $w_t(\cdot)$ is a d_u-dimensional vector and $h_t(\cdot)$ is a scalar, $t\in\mathcal{T}$. The ICC is written as

$$\mathbb{E}_P\left[\eta_t(u,Y)_-\right] \le \beta,\ t\in\mathcal{T},$$

where $(a)_- = \max\{0,-a\}$ and $\beta\ge 0$. In this case, choosing

$$G_t(u,Y) = -\eta_t(u,Y)_-$$

and letting $\epsilon = -\beta$ in (ECP) we can write a problem with an ICC in the ECP form, for some objective function $f(u)$. In [18] the authors propose an algorithm to solve ICC problems, and we refer the reader to [17] for an application to an asset liability management problem in the context of pension funds. The complete theory of this class of problems, including additional algorithms and representations for special cases (e.g., normality) can be found in chapter 6 of the book [14].

2.1.2 Risk Measures

Most risk measures are specified by an expected value of a given function. If the decision maker tries to bound risk by some threshold β, it is possible to accommodate such requirement in formulation (ECP). To simplify the exposition, we assume the decision function $G(\cdot, Y)$ is linear in u, as in $u^\top Y$, which is the most common form in financial problems. In this case, the decision variable u represents the allocations in each of the available financial instruments and Y are the random returns. We will also adopt the fairly standard convention in risk measures that losses are positive and gains are negative. Obviously, that can be changed depending on the problem. Let us see some examples of risk measures that can be accommodated within our framework:

1. *Conditional Value-at-Risk (CVaR)*: for a random variable Z, the CVaR_α with reliability level $\alpha \in (0,1)$ is defined as

$$\text{CVaR}_\alpha[Z] := \min_{\eta \in \mathbb{R}} \left\{ \eta + \frac{1}{1-\alpha} \mathbb{E}_P \left[(Z - \eta)_+ \right] \right\}, \tag{1}$$

where $(a)_+ = \max\{0, a\}$. By defining

$$G(\tilde{u}, Y) := -\eta - \frac{1}{1-\alpha}(u^\top Y - \eta)_+$$

and $\epsilon = -\beta$ in (ECP), where $\tilde{u} = (u, \eta)$, we have a CVaR constraint

$$\text{CVaR}_\alpha[u^\top Y] \leq \beta$$

in the form of an ECP. The minimum in (1) can be removed since we are imposing the inequality for all η. We refer the reader to [24] for a portfolio optimization application that utilizes a CVaR constraint.

2. *Upper semideviation of order p (UDp)*: For a fixed parameter $p \in [1, \infty)$ and a random variable Z with $\mathbb{E}_P [|Z|^p] < \infty$, the UDp is defined as

$$\sigma_p^+[Z] := \Big(\mathbb{E}_P \left[(Z - \mathbb{E}_P [Z])_+^p \right] \Big)^{1/p}.$$

By defining

$$G(u, Y) := -(u^\top Y - \mathbb{E}_P \left[u^\top Y \right])_+^p$$

and $\epsilon = -\beta^p$ in (ECP), we have a UDp constraint

$$\sigma_p^+[u^\top Y] \leq \beta.$$

When $p = 1$, the risk measure is referred to as absolute semideviation. We refer the reader to [29], where the authors apply the UDp in the objective function of a hydrothemal scheduling problem.

3. *Quantile semideviation (QDEV).* Using a result proven in [28], the $\mathrm{QDEV}_{\alpha_1,\alpha_2}$ for a random variable Z is defined as

$$\mathrm{QDEV}_{\alpha_1,\alpha_2}[Z] := \min_{\eta\in\mathbb{R}} \mathbb{E}_P\left[\alpha_1 \max(\eta - Z, 0) + \alpha_2 \max(Z-\eta, 0)\right].$$

By defining

$$G(u,Y) := -\alpha_1 \max(\eta - u^\top Y, 0) - \alpha_2 \max(u^\top Y - \eta, 0)$$

and $\epsilon = -\beta$ in (ECP), we have a QDEV constraint

$$\mathrm{QDEV}_{\alpha_1,\alpha_2}[u^\top Y] \leq \beta.$$

We refer the reader to [9], where the authors apply the QDEV to several problem instances available in the SP literature.

Remark 1. *The first two risk measures are coherent, according to the definition of [1], while the last one is not. However, they all satisfy convexity, which makes the formulations using those risk measures tractable.*

Remark 2. *For each of those risk measures we could have written multiple constraints by varying the risk level. For instance, for the CVaR we can have T constraints, each with reliability levels $(\alpha_1, \dots, \alpha_T)$ and right-hand sides $(\beta_1, \dots, \beta_T)$. This is useful in some contexts where we want to shape the decision-maker's preferences. See for an example in [20] for a portfolio problem with the S&P 100 stocks.*

2.1.3 Stochastic dominance

Stochastic dominance is often used when we want to compare two random variables. For simplicity, let us consider the stochastic dominance of order 2. If Z and W are random variables, we say Z dominates W in the second order, written $Z \succeq_2 W$, if

$$\mathbb{E}_P\left[\max\{\eta - Z, 0\}\right] \leq \mathbb{E}_P\left[\max\{\eta - W, 0\}\right] \quad \forall \eta \in \mathbb{R}.$$

When the random variable W has a finite distribution $\{w_1, \dots, w_T\}$, with probabilities $\{q_1, \dots, q_T\}$, it has been shown that $Z \succeq_2 W$ if and only if

$$\mathbb{E}_P\left[(w_t - Z)^+\right] \leq \mathbb{E}_P\left[(w_t - W)^+\right], \quad t \in \mathcal{T}. \tag{2}$$

In this case, we would need T expected value constraints, one for each realization of W as

$$G_t(u,Y) := (w_t - W)^+ - (w_t - u^\top Y)^+, \quad t \in \mathcal{T}.$$

and $\epsilon = 0$. A typical example in portfolio optimization is to try to construct a portfolio that outperforms some well-known benchmark. See [21] for an example with 435 stocks where the author constructs a portfolio optimization model to dominate the S&P 500 index in the first and second orders.

2.2 SAA for ECP

General (ECP) problems are challenging to solve since $\mathbb{E}_P\left[G_t(u,Y)\right]$, $t \in \mathcal{T}$, may not be calculable 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

$$\hat{P}_n := \frac{1}{n}\sum_{i\in[n]} \delta_{y^i}, \tag{3}$$

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

$$e_{t,n}(u) := \mathbb{E}_{\hat{P}_n}\left[G_t(u,Y)\right] = \frac{1}{n}\sum_{i\in[n]} G_t(u,y^i) \tag{4}$$

for $t \in \mathcal{T}$. The SAA problem is obtained by replacing the true expected value

$$e_t(u) := \mathbb{E}_P\left[G_t(u,Y)\right]$$

by the approximated one in (4) as

$$\begin{aligned} \min_{u\in\mathcal{U}} \quad & f(u) \\ \text{s.t.} \quad & e_{t,n}(u) \geq \alpha,\ t \in \mathcal{T}, \end{aligned} \tag{SAA-ECP}$$

where α is the target level of the approximate problem, which may be different from ϵ in (ECP).

2.3 Contextual ECP and its Data-Driven Approximation

The random parameters Y may exhibit some dependency on a vector of features, which may have a predictive power to explain the outcomes of the random vector Y. Hence, it may be beneficial to include those features in the problem formulation.

For a given random observation $X = x$, a contextual ECP (C-ECP) can be formulated as

$$\begin{aligned} z^*_\epsilon(x) = \min_{u\in\mathcal{U}} \quad & f(u) \\ \text{s.t.} \quad & \mathbb{E}_P\left[G_t(u,Y) \mid X = x\right] \geq \epsilon,\ t \in \mathcal{T}. \end{aligned} \tag{C-ECP}$$

In (C-ECP), we let (X,Y) be defined on a probability space $(\mathcal{X}\times\mathcal{Y},\mathcal{F},P)$, where $\mathcal{X} \subset \mathbb{R}^{d_x}$ and $\mathcal{Y} \subset \mathbb{R}^{d_y}$. Moreover, the expected value 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-ECP) gives the best response to the observed feature vector x as measured by the objective function $f(\cdot)$.

Solving (C-ECP) requires 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 expected-value constraint in (C-ECP). 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-ECP). We first form a weight function $w^i_n(x,\mathcal{D}_n)$, $i \in [n]$, such that $\sum_{i\in[n]} w^i_n(x,\mathcal{D}_n) = 1$ and $w^i_n(x,\mathcal{D}_n) \geq 0$, $i \in [n]$, to measure "proximity" of each data point i with respect to the observed feature x. Using this weight function the conditional distribution of Y given $X = x$ can be approximated as

$$\hat{P}_n := \sum_{i\in[n]} w^i_n(x,\mathcal{D}_n)\delta_{y^i}. \tag{5}$$

For simplicity in notation, we dropped the dependence of $\hat{P}_n$ to $X = x$ and $\mathcal{D}_n$. Using the approximated probability distribution (5), we can obtain an approximated expected value as

$$m_{t,n}(u;x) := \mathbb{E}_{\hat{P}n}\left[G_t(u,Y)\right] = \sum_{i\in[n]} w_n^i(x,\mathcal{D}_n)G_t(u,y^i) \tag{6}$$

for $t \in \mathcal{T}$.

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

$$m_t(u;x) := \mathbb{E}_P\left[G_t(u,Y) \mid X = x\right] \tag{7}$$

by the approximation (6) as

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

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

Some observations regarding problem (DDC-ECP) are in order. First, one can interpret the output of formulation (DDC-ECP) as a policy: given an observed feature vector $X = x$, a solution $u_n^*(x)$ represents the best response measured by the objective function $f(\cdot)$. Second, (C-ECP) is at least as hard to solve as (SAA-ECP) since the computation of the weights can be done offline.

In this paper, we will construct the weights $w_n^i(x,\mathcal{D}_n)$, $i \in [n]$, using kNN [13], described in Definition 1:

Definition 1. *Consider a collection of independently and identically distributed (i.i.d.) 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 ℓ_p-norm,*
- *k_n nearest neighbors of (X,Z) are chosen with ties broken randomly,*

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

A kNN weight function is constructed such that data points that are "close" to the observed feature vector $X = x$ be more important to accurately estimate the conditional distribution of Y given $X = x$; and hence, the expected values $m_t(u;x)$, $t \in \mathcal{T}$.

3 Finite Dataset Guarantees of (DDC-ECP)

In this section, we present theoretical results that support the use of (DDC-ECP) to approximate (C-ECP) with kNN. In particular, we investigate the feasibility of the data-driven solution in Section 3.1 with the proofs presented in Section 3.2. To present the theoretical results, we make the following assumptions.

Assumption 1. *$\mathcal{U} \subset \mathbb{R}^{d_u}$ is a nonempty and compact set with diameter θ.*

Assumption 2. *$\mathcal{X} \subset [0,1]^{d_x}$ is a compact set, and there exists $\vartheta > 0$ such that $P\left\{X \in \mathcal{N}_\varrho(x)\right\} > \vartheta\varrho^{d_x}$ for all $x \in \mathcal{X}$ and $\varrho > 0$.*

Assumption 3. *Given a random $Z \in \mathbb{R}$, $Z - \mathbb{E}_P[Z \mid X = x]$ is conditionally sub-Gaussian given $X = x$ with variance proxy σ^2, uniformly for all $x \in \mathcal{X}$, and $\mu(x) := \mathbb{E}_P[Z \mid X = x]$ is $N(x)$-Lipschitz, with $N := \sup_{x\in\mathcal{X}} N(x) < \infty$.*

Assumption 4. *For any $u \in \mathcal{U}$ and $x \in \mathcal{X}$, $\mathbb{E}_P[G_t(u, Y) \mid X = x]$ is bounded for $t \in \mathcal{T}$.*

Assumption 5. *For any $y \in \mathcal{Y}$, function $G_t(\cdot, y)$ is $L(y)$-Lipschitz, with $L := \sup_{y\in\mathcal{Y}} L(y) < \infty$, for $t \in \mathcal{T}$.*

We note that $\mathcal{X} \subset [0, 1]^{d_x}$ is without loss of generality, and it indicates a normalization of X.

3.1 Main Results

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

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 i.i.d. random vectors. 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. Let $W_n^i(x, \mathcal{E}_n)$, $i \in [n]$, be a random weight function and define

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

Note that $m_n(u; x)$, defined in (6), 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-ECP) and (DDC-ECP), respectively, for $X = x$ and given $\mathcal{E}_n$. That is,

$$\mathcal{U}_\epsilon(x) := \{u \in \mathcal{U} \mid m_t(u; x) \geq \epsilon,\ t \in \mathcal{T}\},$$

and

$$\mathcal{U}_{n,\alpha}(x) := \{u \in \mathcal{U} \mid M_{t,n}(u; x) \geq \alpha,\ t \in \mathcal{T}\}.$$

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

Theorem 1. *Suppose that Assumptions 1–3 hold. Moreover, assume that the compact set $\mathcal{U}$ is finite and let $\alpha > \epsilon$ Then,*

$$P^n\{\mathcal{U}_{n,\alpha}(x) \subseteq \mathcal{U}_\epsilon(x)\ \forall x \in \mathcal{X}\} \geq 1 - |\mathcal{T}||\mathcal{U}|A(\alpha - \epsilon)\exp\{-nB(\alpha - \epsilon)\},$$

where constants $A(\cdot)$ and $B(\cdot)$ are defined in Corollary 1 of Lemma 1.

Theorem 2. *Suppose that Assumptions 1–5 hold. Let $\lambda > 0$ and $0 < \beta < \alpha - \lambda - \epsilon$. Then,*

$$P^n\{\mathcal{U}_{n,\alpha}(x) \subseteq \mathcal{U}_\epsilon(x)\ \forall x \in \mathcal{X}\} \geq 1 - |\mathcal{T}|\left\lceil\frac{1}{\beta}\right\rceil^{|\mathcal{T}|}\left\lceil\left(\frac{2L\theta}{\lambda}\right)^{d_u}\right\rceil A(\alpha - \lambda - \epsilon - \beta) \times \exp\{-nB(\alpha - \lambda - \epsilon - \beta)\},$$

where constants $A(\cdot)$ and $B(\cdot)$ are defined in Corollary 1 of Lemma 1.

We now present the minimum dataset size n required to guarantee with high probability the feasibility of a data-driven solution.

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

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

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 $\kappa = \alpha - \epsilon$.

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

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

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 $\kappa = \alpha - \lambda - \epsilon - \beta$.

Remark 3. *Observe from Propositions 1 and 2 that the higher the confidence level $1-\rho$ and $|\mathcal{T}|$, the larger the required dataset size n, where n grows logarithmically in $1/\rho$. Also, n is large for α close to ϵ, with a growth in a polynomial order of $1/(\epsilon-\alpha)^{2d_x}$. The larger the dimension d_x of the feature vector is, the closeness of α to ϵ has a larger impact on n. We also see in that in general, the larger d_x, the larger n. Moreover, n grows linearly in d_u. 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 β. Observe that n grows logarithmically with $\lceil\frac{1}{\beta}\rceil$. Moreover, the impact of $\lceil\frac{1}{\lambda}\rceil$ is similar to that of $\lceil\frac{1}{\beta}\rceil$.*

3.2 Proofs

A critical component for the proofs is the uniform consistency of a k_n-NN estimator. We adopt Lemma 1 from [6, Lemma 10] to state such a result and then we apply it to contextual ECP in Corollary 1.

Lemma 1. *Consider i.i.d. random vectors $(X, Z), (X^1, Z^1), \ldots, (X^n, Z^n) \in \mathbb{R}^{d_x} \times \mathbb{R}$ and let $\mu_n(x)$ denote a k_n-NN estimator, constructed based on Definition 1. Suppose that Assumptions 2 and 3 hold. Then,*

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

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

Now, for $t \in \mathcal{T}$, define

$$R_{t,n}(x) := \sup_{u \in \mathcal{U}} |M_{t,n}(u; x) - m_t(u; x)|, \tag{9}$$

where $m_t(u; x)$ and $M_t(u; x)$ are defined in (7) and (8), respectively.

Corollary 1. *Under assumptions of Lemma 1, suppose that the weight function $W_n^i(x, \mathcal{E}_n)$, $i \in [n]$, is formed with $k_n = \lceil cn^\gamma \rceil$ for $\gamma \in (0, 1)$ and $c > 0$ such that $k_n \leq n - 1$. Then, for any $\kappa > 0$, there exist constants $A(\kappa)$ and $B(\kappa)$ defined as $A(\kappa) := \max \left\{ \left(\frac{\mathcal{O}(1)\sqrt{d_x}}{\kappa} \right)^{d_x}, \mathcal{O}(1) \left(\frac{\mathcal{O}(1)}{d_x} \right)^{d_x} \right\}$ and $B(\kappa) := \min \left\{ \mathcal{O}(1) \big(\mathcal{O}(1)\kappa \big)^{2d_x}, \min_{n \in \mathbb{N}} \left\{ \mathcal{O}(1) \frac{n^{\gamma - 1}\kappa^2}{\sigma^2} - \mathcal{O}(1) d_x \frac{\log(n)}{n} \right\} \right\}$, where $n \geq \mathcal{O}(1) \left(\frac{\mathcal{O}(1)}{\kappa} \right)^{\frac{d_x}{1-\gamma}}$ and $\frac{n^\gamma}{\log(n)} \geq \frac{d_x \sigma^2}{\kappa^2}$, such that*

$$P^n \left\{ \sup_{x \in \mathcal{X}} R_{t,n}(x) \geq \kappa \right\} \leq A(\kappa) \exp\{-nB(\kappa)\}$$

for $t \in \mathcal{T}$.

We are now ready to prove Theorem 1.

Proof of Theorem 1. For $X = x$, we have

$$\begin{aligned}
&\{\mathcal{U}_{n,\alpha}(x) \not\subseteq \mathcal{U}_\epsilon(x)\} \\
&\quad = \{\exists\ u \in \mathcal{U} \text{ such that } u \in \mathcal{U}_{n,\alpha}(x) \text{ and } u \in \mathcal{U} \backslash \mathcal{U}_\epsilon(x)\} \\
&\quad = \bigcup_{u \in \mathcal{U}} \{u \in \mathcal{U}_{n,\alpha}(x) \text{ and } u \in \mathcal{U} \backslash \mathcal{U}_\epsilon(x)\} \\
&\quad = \bigcup_{u \in \mathcal{U}} \{M_{t,n}(u; x) \geq \alpha \text{ for all } t \in \mathcal{T} \text{ and } m_t(u; x) < \epsilon \text{ for some } t \in \mathcal{T}\} \\
&\quad \subseteq \bigcup_{t \in \mathcal{T}} \bigcup_{u \in \mathcal{U}} \{M_{t,n}(u; x) - m_t(u; x) \geq \alpha - \epsilon\} \\
&\quad \subseteq \bigcup_{t \in \mathcal{T}} \bigcup_{u \in \mathcal{U}} \{|M_{t,n}(u; x) - m_t(u; x)| \geq \alpha - \epsilon\} \\
&\quad \subseteq \bigcup_{t \in \mathcal{T}} \{\sup_{u \in \mathcal{U}} |M_{t,n}(u; x) - m_t(u; x)| \geq \alpha - \epsilon\}
\end{aligned} \tag{10}$$

$$= \bigcup_{t\in\mathcal{T}} \{R_{t,n}(x) \geq \alpha - \epsilon\}. \tag{11}$$

By a similar argument leading to (11) we have

$$\{\mathcal{U}_{n,\alpha}(x) \nsubseteq \mathcal{U}_{\epsilon}(x) \text{ for some } x \in \mathcal{X}\} \subseteq \bigcup_{t\in\mathcal{T}} \bigcup_{u\in\mathcal{U}} \{\sup_{x\in\mathcal{X}} R_{t,n}(x) \geq \alpha - \epsilon\}.$$

Thus, by applying $P^n\{\cdot\}$ on both sides, an application of the union bound, and finally using Corollary 1, the result follows. □

To prove Theorem 2, we recall the following fact.

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

Proof of Theorem 2. Consider an arbitrary $X = x$ and a fixed $t \in \mathcal{T}$. Let $J = \lceil \frac{1}{\beta} \rceil$, and for $j \in [J-1]$, define

$$\mathcal{U}_{t,j}(x) := \left\{ u \in \mathcal{U} \,\middle|\, a_t + (b_t - a_t)\frac{j-1}{J} \leq m_t(u;x) < a_t + (b_t - a_t)\frac{j}{J} \right\},$$

and let

$$\mathcal{U}_{t,J}(x) := \left\{ u \in \mathcal{U} \,\middle|\, a_t + (b_t - a_t)\frac{J-1}{J} \leq m_t(u;x) \leq b_t \right\},$$

where $m_t(u;x) \in [a_t, b_t]$ for any $u \in \mathcal{U}$, i.e., $a_t = \inf_{u\in\mathcal{U}} m_t(u;x) > -\infty$, $b_t = \sup_{u\in\mathcal{U}} m_t(u;x) < \infty$, and $a_t \leq b_t$. We note that for some $j_t \in [J]$, but not all, $\mathcal{U}_{t,j_t}(x)$ might be empty, $t \in \mathcal{T}$. We start the proof by claiming that for any $j := [j_1, \ldots, j_T]$ with $\mathcal{U}_{t,j_t}(x) \neq \emptyset$ for all $t \in \mathcal{T}$, there exists a finite set $\mathcal{Z}_j^\lambda(x) \subseteq \bigcup_{t\in\mathcal{T}} \mathcal{U}_{t,j_t}(x)$ such that $|\mathcal{Z}_j^\lambda(x)| \leq |\mathcal{T}| \lceil \big(\frac{2L\theta}{\lambda}\big)^{d_u} \rceil$ and for any $u \in \bigcup_{t\in\mathcal{T}} \mathcal{U}_{t,j_t}(x)$, there exists $z \in \mathcal{Z}_j^\lambda(x) \cap \mathcal{N}_{\frac{\lambda}{L}}(u)$. To prove the claim, note that by Fact 1, there exists a finite set $\emptyset \neq \mathcal{S} \subseteq \mathcal{U}$ with $|\mathcal{S}| \leq \lceil \big(\frac{2L\theta}{\lambda}\big)^{d_u} \rceil$ such that for any $u \in \mathcal{U}$, there exists $s \in \mathcal{S} \cap \mathcal{N}_{\frac{\lambda}{2L}}(u)$. Let us define $\overline{\mathcal{S}}_{t,j_t}^\lambda(x) := \left\{ s \in \mathcal{S} \,\middle|\, \mathcal{U}_{t,j_t}(x) \cap \mathcal{N}_{\frac{\lambda}{2L}}(s) \neq \emptyset \right\}$ and $\widetilde{\mathcal{Z}}_j^\lambda(x) := \bigcup_{t\in\mathcal{T}} \bigcup_{s\in\overline{\mathcal{S}}_{t,j_t}^\lambda(x)} u_s$, where u_s is an arbitrary element $\in \mathcal{U}_{t,j_t}(x) \cap \mathcal{N}_{\frac{\lambda}{2L}}(s)$, $t \in \mathcal{T}$. Note that by construction, $\widetilde{\mathcal{Z}}_j^\lambda(x) \subseteq \bigcup_{t\in\mathcal{T}} \mathcal{U}_{t,j_t}(x)$ and $|\widetilde{\mathcal{Z}}_j^\lambda(x)| \leq \sum_{t\in\mathcal{T}} |\overline{\mathcal{S}}_{t,j_t}^\lambda(x)| \leq |\mathcal{T}||\mathcal{S}| \leq |\mathcal{T}| \lceil \big(\frac{2L\theta}{\lambda}\big)^{d_u} \rceil$. Moreover, for any $u \in \bigcup_{t\in\mathcal{T}} \mathcal{U}_{t,j_t}(x) \subseteq \mathcal{U}$, there exists $s \in \mathcal{S} \cap \mathcal{N}_{\frac{\lambda}{2L}}(u)$, for which $\mathcal{U}_{t,j_t}(x) \cap \mathcal{N}_{\frac{\lambda}{2L}}(s) \neq \emptyset$ for some $t \in \mathcal{T}$ (because $u \in \mathcal{U}_{t,j_t}(x) \cap \mathcal{N}_{\frac{\lambda}{2L}}(s)$ for some $t \in \mathcal{T}$). Consequently, this s belongs to $\overline{\mathcal{S}}_{t,j_t}^\lambda(x)$ for some $t \in \mathcal{T}$, 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 $\mathcal{Z}^\lambda(x) := \bigcup_{j\in\mathcal{J}} \mathcal{Z}_j^\lambda(x)$, where $\mathcal{J} := \bigtimes_{t\in\mathcal{T}} [J]$ and $|\mathcal{Z}^\lambda(x)| \leq |\mathcal{J}||\mathcal{T}| \lceil \big(\frac{2L\theta}{\lambda}\big)^{d_u} \rceil$. We note that $\mathcal{Z}_j^\lambda(x)$ might be empty for some $j \in \mathcal{J}$, but not all. Hence, $\mathcal{Z}^\lambda(x)$ is a nonempty finite set. We also define

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

and

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

Now, because $\alpha - \lambda > \epsilon + \beta$ and $\mathcal{Z}^{\lambda}(x)$ is finite for $x \in \mathcal{X}$, by Theorem 1, we have

$$P^n \left\{ \overline{\mathcal{Z}}^{\lambda}_{n,\alpha}(x) \subseteq \overline{\mathcal{Z}}^{\lambda}_{\epsilon+\beta}(x)\ \forall x \in \mathcal{X} \right\} \geq 1 - |\mathcal{T}| \left\lceil \frac{1}{\beta} \right\rceil^{|\mathcal{T}|} \left\lceil \left(\frac{2L\theta}{\lambda} \right)^{d_u} \right\rceil A(\kappa) \exp\{-nB(\kappa)\}, \tag{12}$$

where $\kappa = \alpha - \lambda - \epsilon - \beta$. To complete the proof, consider $u \in \mathcal{U}_{n,\alpha}(x)$. Let $j = [j_1, \ldots, j_T]$ with $j_t \in [J]$, be such that $u \in \mathcal{U}_{t,j_t}(x)$, $t \in \mathcal{T}$. Using the claim, for this u, there exists $z \in \mathcal{Z}^{\lambda}_{j}(x) \cap \mathcal{N}_{\frac{\lambda}{L}}(u)$. This z belongs to $\mathcal{Z}^{\lambda}_{j}(x) \subseteq \bigcup_{t \in \mathcal{T}} \mathcal{U}_{t,j_t}(x)$ and $|m_t(u;x) - m_t(z;x)| \leq \frac{1}{J} \leq \beta$ for all $t \in \mathcal{T}$. Moreover, by Assumption 5, we have $|G_t(u, Y^i) - G_t(z, Y^i)| \leq L\|u - z\| \leq \lambda$ for $i \in [n]$ and $t \in \mathcal{T}$. In particular, $G_t(z, Y^i) + \lambda \geq G_t(u, Y^i)$ for all $i \in [n]$ and $t \in \mathcal{T}$. Because $\sum_{i \in [n]} W^i_n(x, \mathcal{E}_n) = 1$, we have $\sum_{i \in [n]} W^i_n(x, \mathcal{E}_n) G_t(z, Y^i) + \lambda \geq \sum_{i \in [n]} W^i_n(x, \mathcal{E}_n) G_t(u, Y^i)$, $t \in \mathcal{T}$. Consequently, given that $u \in \mathcal{U}_{n,\alpha}(x)$, we have $M_{t,n}(z;x) \geq \alpha - \lambda$ for all $t \in \mathcal{T}$. In addition, as $z \in \mathcal{Z}^{\lambda}_{j}(x) \subseteq \mathcal{Z}^{\lambda}(x)$, we conclude that $z \in \overline{\mathcal{Z}}^{\lambda}_{n,\alpha}(x)$. Now, if $\overline{\mathcal{Z}}^{\lambda}_{n,\alpha}(x) \subseteq \overline{\mathcal{Z}}^{\lambda}_{\epsilon+\beta}(x)$, then we have $m_t(z;x) \geq \epsilon + \beta$ for all $t \in \mathcal{T}$, which combined with $m_t(u;x) \geq m_t(z;x) - \beta$, imply that $m_t(u;x) \geq \epsilon$ for all $t \in \mathcal{T}$, i.e., $u \in \mathcal{U}_{\epsilon}(x)$. Consequently, we showed that $\{\mathcal{U}_{n,\alpha}(x) \subseteq \mathcal{U}_{\epsilon}(x)\} \supseteq \{\overline{\mathcal{Z}}^{\lambda}_{n,\alpha}(x) \subseteq \overline{\mathcal{Z}}^{\lambda}_{\epsilon+\beta}(x)\}$. This, in turn, imply that

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

and thus, the result follows from (12). □

4 Numerical Experiments

In this section, we present numerical experiments for an expected-value-constrained optimization problem with a continuous, infinite feasible region. We investigate the feasibility of a data-driven solution to the true problem and analyze the impact of various parameters, including the dimension of the feature vector, on the rate of convergence and the required number of data points to achieve feasibility in the original problem. To this end, we consider a portfolio optimization problem with synthetic data and features. A version of the problem was proposed in [12], and we adapt the formulation for an expected-valued-constrained optimization problem.

We assume that there are features that may be used to predict the return of d_u assets, while $\Sigma \in \mathbb{R}^{d_u \times d_u}$, the covariance matrix of the asset returns, does not depend on the features. We assume that the decision maker aims to minimize the variance of the portfolio's return while keeping the expected return of the portfolio above a desired target ϵ as follows:

$$\begin{aligned} \min_{u \in \mathcal{U}} \quad & u^\top \Sigma u \\ \text{s.t.} \quad & \mathbb{E}\left[u^\top Y \mid X = x \right] \geq \epsilon. \end{aligned} \tag{13}$$

Vector $u \in \mathbb{R}^{d_u}$ defines the portfolio positions, and $Y \in \mathbb{R}^{d_u}$ is the random vector of asset returns. We have $\mathcal{U} = \left\{ u \in \mathbb{R}^{d_u} \,\middle|\, e^\top u = 1,\ u \geq 0 \right\}$, where $e \in \mathbb{R}^{d_u}$ is a vector of all ones. Moreover, we set $\epsilon = \alpha - \chi$, where χ is some positive number, implying that $\epsilon < \alpha$ (recall that our feasibility guarantees require that $\alpha - \chi = \epsilon < \alpha$, which is ensured by any choice of $\chi > 0$).

We assume that the feature vector $X \in \mathbb{R}^{d_x}$ follows a normal distribution $\mathrm{N}(0, \mathbb{I}_{d_x})$, i.e., entries X_l, $l \in [d_x]$, of X are drawn independently from a standard normal distribution. We also assume that Y is formed by a linear factor model of the form $Y = \bar{Y} + Ef + \varepsilon$. Here, $E \in \mathbb{R}^{d_u \times 4}$ is the loading factor matrix such that each entry of E is drawn independently from a uniform distribution on $[-0.0025\tau, 0.0025\tau]$, where $\tau \geq 0$ is a noise level parameter. Moreover, $f \sim \mathrm{N}(0, \mathbb{I}_4)$ and $\varepsilon \sim \mathrm{N}(0, (0.01\tau)^2 \mathbb{I}_{d_u})$. We assume that only a subset $\mathcal{L}^* \subseteq [d_x]$ has predictive power and we form the vector of conditional mean returns $\bar{Y}$ as

$$\bar{Y}_j = \left(\frac{0.05}{\sqrt{5}} \sum_{l \in \mathcal{L}^*} B^*_{jl} X_l + 0.1^{\frac{1}{p}} \right)^p, \quad j \in [d_u],$$

where p is a fixed positive integer number and $B^* \in \mathbb{R}^{d_u \times d_x}$ is a random matrix that contains the parameters of the true linear factor model. Each entry of B^* is drawn independently from a Bernoulli distribution with parameter 0.5. Given the setup described, conditioned on $X = x$, we have $Y|(X = x) \sim \mathrm{N}(\bar{Y}, EE^\top + (0.01\tau)^2 \mathbb{I}_{d_u})$. As it can be seen, the covariance of $Y|(X = x)$ does not depend on x by design, and we set $\Sigma = EE^\top + (0.01\tau)^2 \mathbb{I}_{d_u}$.

To conduct experiments, we consider an instance with $d_u = 10$ assets, $p \in \{1, 4, 8, 16\}$, $d_x \in \{5, 10, 100\}$, $\tau \in \{1, 2\}$, and $\alpha = 0.10$. For each value of τ, we first generate matrix E and subsequently matrix Σ (note that they only depend on τ) to have the deterministic parameters to form an instance of (13). Then, we generate B^* and $X = x$, in a nested way, such that an instance with a higher d_x would contain all information from an instance with a smaller d_x, i.e., the first five elements of x in an instance with $d_x = 10$ are the same five elements of x in an instance with $d_x = 5$. For B^*, the first five columns of B^* in an instance with $d_x = 10$ are the same five columns of x in an instance with $d_x = 5$. Throughout, we assume that $|\mathcal{L}^*| = 5$, and hence, without loss of generality, we can assume that columns of B^* that do not belong to the index set $\mathcal{L}^*$ are effectively zero. These imply that for a fixed $X = x$, regardless of the dimension d_x, we always have the same $\bar{Y}$.

Now, we have a full instance of (13), and we can simulate i.i.d. data $\mathcal{D}_n = \{(x^i, y^i)\}_{i \in [n]}$ following the distributions of X and Y. Again, we emphasize that by construction, there is no difference between the y-component of the data point i in datasets with different dimension d_x, $i \in [n]$. We can then form a data-driven approximation to (13) as (DDC-ECP). We refer to an optimal data-driven solution to this problem as $\hat{u}_{n,\alpha}(x)$. We perform 50 microsimulations per instance with a fixed B^* and $X = x$, i.e., 50 sets of training data $\mathcal{D}_n$ are generated. Solutions obtained from different data-driven approaches are compared by calculating their expected out-of-sample returns, i.e., $\bar{Y}^\top \hat{u}_{n,\alpha}(x)$. We report the results of 50 microsimultations per instance in boxplots of these expected returns. We also let $n \in \{2d_x, 5d_x, 10d_x, 100d_x\}$. We consider two data-driven approaches to solve (13):

`kNN`: The constraint in (13) 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 Figures 1 and 2. Each of the presented figures depicts the performance of the `kNN` and `nSAA` approaches for a specific configuration of ϵ and τ, over varying the feature dimension d_x, the model degree p, and the training dataset of size n.

Several trends can be seen in Figures 1 and 2. Observe that the median value of the expected return for the `kNN` approach always outperformed that of the `nSAA` approach. More importantly, `nSAA` did not yield consistent solutions, whereas `kNN` always yielded consistent solutions. We note varying convergence rates, and we will analyze them by varying the model degree p, the feature dimension d_x, the target return ϵ, and the noise level τ.

Effects of varying model degree p and feature dimension d_x. The rate of convergence of the `kNN` approach highly depends on the model degree p and the feature dimension d_x. On the one hand, for a fixed (d_x, ϵ, τ), increasing p led to a lower rate of convergence. Recall that when $p = 1$,

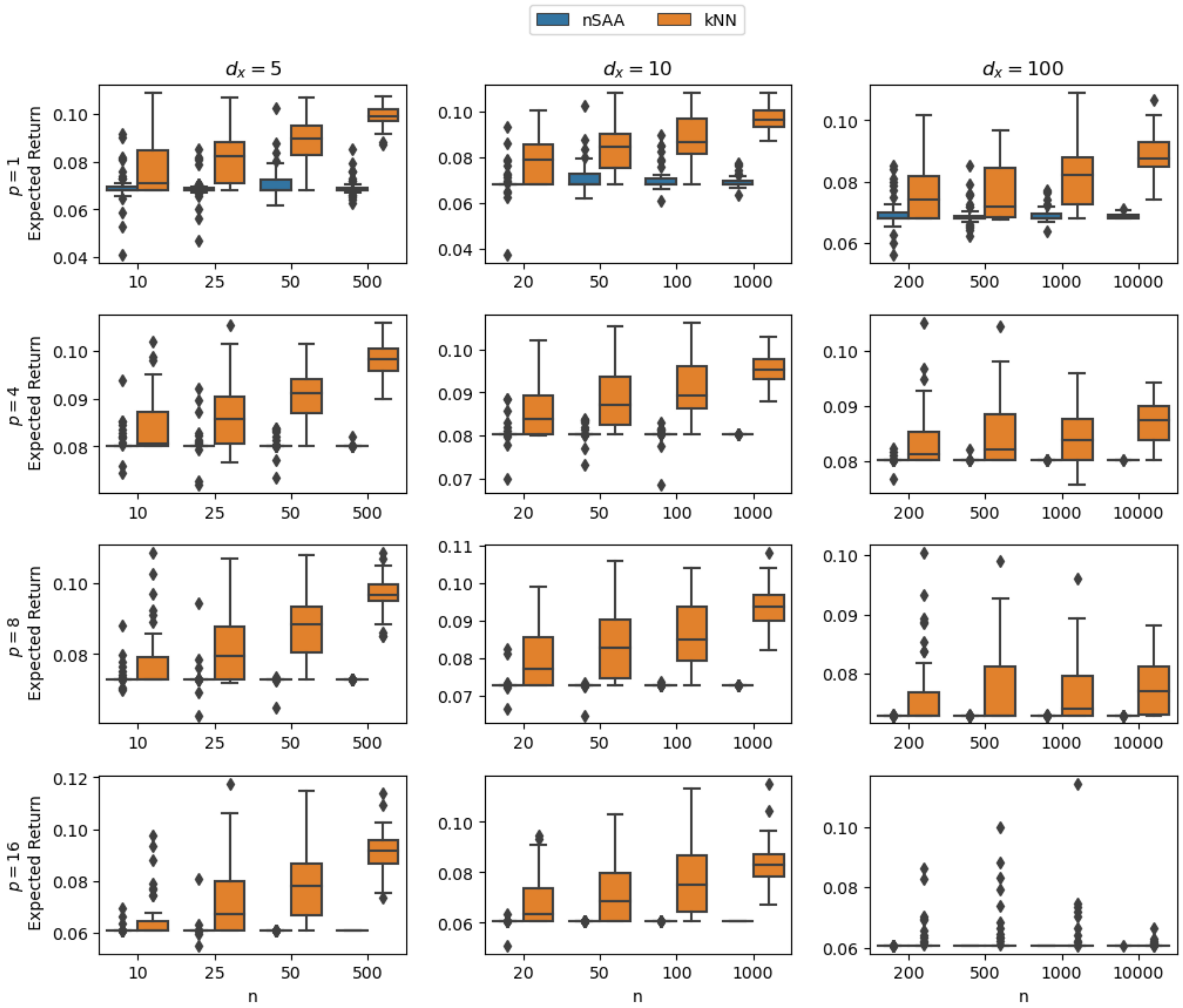

Figure 1: $d_u = 10$, $\alpha = 0.10$, and $\tau = 1$.

there is a linear relationship between X and Y, whereas $p > 1$ leads to a nonlinear relationship. Our results indicate that while the form of this relationship is unknown to the kNN approach, a higher degree of nonlinearity leads to a lower rate of convergence. Moreover, for a fixed (ϵ, τ), the impact of increasing p on the convergence rate was more pronounced for a higher d_x. On the other hand, as expected, for a fixed (p, ϵ, τ), a higher d_x results in a lower rate of convergence for the kNN approach (see Remark 3). In other words, for a fixed dataset of size n, increasing d_x leads to a smaller coverage probability on the feasibility of the data-driven solution.

Effects of varying target return ϵ and noise level τ. To analyze the impact of the target return ϵ on the rate of convergence, let us fix (p, d_x, τ). Observe that the closer ϵ is to $\alpha = 0.10$, the lower the rate of convergence, as expected from Remark 3. We know that an estimation of the required number of data points grows quite large for values of α close to ϵ, in the order of $1/(\epsilon - \alpha)^{2d_x}$. The impact of such closeness on the convergence rate is more pronounced for a larger d_x. Moreover, we see that for a fixed (p, d_x, ϵ), a higher noise level τ led to a higher variability in the expected return per its role in the covariance of $Y|(X = x)$. However, the median value of the expected return does not seem to be impacted significantly.

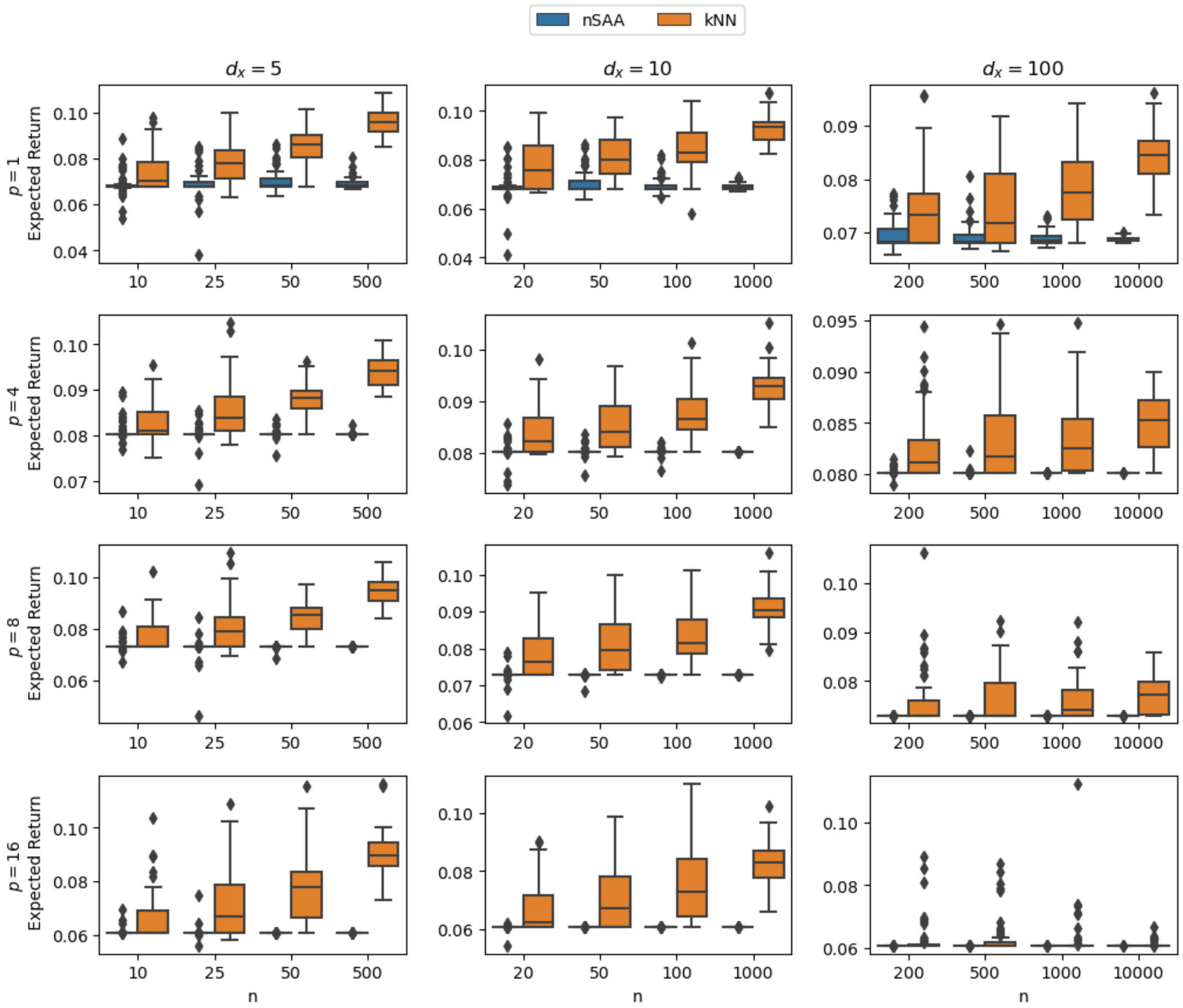

Figure 2: $d_u = 10$, $\alpha = 0.10$, and $\tau = 2$.

We end this section with several remarks. First, notice that for any pair of (ϵ, τ) and for a specific configuration of p, the performance of `nSAA` remains the same by varying d_x but with a fixed n. For instance, when $p = 1$ and $n = 50$, `nSAA` results in the same solution for both $d_x = 5$ and $d_x = 10$. This is because not only all these $n = 1000$ samples are equally likely but also they have the same y-component by construction. Thus, the performance of `nSAA` is the same in both cases. Second, we observed similar trends hold for the case that $d_u = 50$. Additionally, we observed that for a fixed (p, d_x, ϵ, τ), increasing d_u leads to a slower rate of convergence. This result is expected, as the estimated required dataset size grows linearly in the dimension of the feasible region, d_u (see Remark 3).

5 Conclusions

Expected-value-constrained programming problems represent a broad class of models that include integrated chance constraints, problems with a risk constraint, and stochastic dominance formulations. In this paper, we propose a contextual expected-value-constrained programming formulation that accommodates features in addition to the dependent random variables. We show how the important aforementioned classes of problems can be written in this format, and also how the resulting problems can be approximated using data. The contextual information is incorporated using kNN, which averages the past observations that are closest to the currently observed feature vector.

We then show theoretical results stating that by solving the data-driven formulation one can obtain a feasible solution to the true problem with high probability. Such probability approaches one exponentially fast as the dataset size grows. While our results accommodate approximations obtained by popular machine learning methods such as classification and regression trees (CARTs) and random forest (RF), we presented theoretical results for kNN.

We test our methodology on a synthetic portfolio selection problem. We vary several parameters of the problem, including the dimensions of the feature vector. 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 number of data points increases. The convergence is slower for higher-dimensional problems, for which case more data points are needed.

Future work includes extensions to dynamic problems. In sequential settings, the way to incorporate new features is not unique and it makes sense to mix offline and online components, in an adaptive fashion. 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), where there is the availability of data with contextual information.

References

[1] Artzner, P., F. Delbaen, J.-M. Eber, and D. Heath (1999). Coherent measures of risk. *Math. Financ. 9*(3), 203–228.

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

[3] Bertsimas, D., D. B. Brown, and C. Caramanis (2011). Theory and applications of robust optimization. *SIAM Rev. 53*(3), 464–501.

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

[5] 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.

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

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

[8] Cohen, M. C., I. Lobel, and R. Paes Leme (2020). Feature-based dynamic pricing. *Management Sci. 66*(11), 4921–4943.

[9] Cotton, T. G. and L. Ntaimo (2015). Computational study of decomposition algorithms for mean-risk stochastic linear programs. *Math. Program. Computation 7*(4), 471–499.

[10] 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`.

[11] 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.

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

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

[14] Haneveld, W. K. K., M. H. Van der Vlerk, and W. Romeijnders (2019). *Stochastic programming: Modeling decision problems under uncertainty.* Springer Nature.

[15] 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`.

[16] Klein Haneveld, W. (1986). On integrated chance constraints. In *Stochastic Programming.*

[17] Klein Haneveld, W. K., M. H. Streutker, and M. H. Van Der Vlerk (2010). An alm model for pension funds using integrated chance constraints. *Ann. Oper. Res. 177*, 47–62.

[18] Klein Haneveld, W. K. and M. H. Van Der Vlerk (2006). Integrated chance constraints: reduced forms and an algorithm. *Comput. Management Sci. 3*(4), 245–269.

[19] 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.

[20] Krokhmal, P., J. Palmquist, and S. Uryasev (2002). Portfolio optimization with conditional value-at-risk objective and constraints. *J. Risk 4*, 43–68.

[21] Luedtke, J. (2008). New formulations for optimization under stochastic dominance constraints. *SIAM J. Optim. 19*(3), 1433–1450.

[22] 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.

[23] 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.

[24] Pagnoncelli, B. K., D. Ramírez, H. Rahimian, and A. Cifuentes (2023). A synthetic data-plus-features driven approach for portfolio optimization. *Comput. Econ. 62*(1), 187–204.

[25] Philpott, A. B. and V. L. De Matos (2012). Dynamic sampling algorithms for multi-stage stochastic programs with risk aversion. *Eur. J. Oper. Res. 218*(2), 470–483.

[26] Rahimian, H. and B. Pagnoncelli (2023). Data-driven approximation of contextual chance-constrained stochastic programs. *SIAM J. Optim. 33*(3), 2248–2274.

[27] Rawls, C. G. and M. A. Turnquist (2010). Pre-positioning of emergency supplies for disaster response. *Transp. Res. B: Methodol. 44*(4), 521–534.

[28] Ruszczyński, A. and A. Shapiro (2006). Optimization of convex risk functions. *Math. Oper. Res. 31*(3), 433–452.

[29] Shapiro, A., W. Tekaya, J. P. da Costa, and M. P. Soares (2013). Risk neutral and risk averse stochastic dual dynamic programming method. *Eur. J. Oper. Res. 224*(2), 375–391.

[30] Wang, W. and S. Ahmed (2008). Sample average approximation of expected value constrained stochastic programs. *Oper. Res. Letters 36*(5), 515–519.