Optimization-based Scenario Reduction for Data-Driven Two-stage Stochastic Optimization

Dimitris Bertsimas
Sloan School of Management and Operations Research Center, Massachusetts Institute of Technology, Cambridge, MA 02142,
dbertsim@mit.edu,
ORCID: 0000-0002-1985-1003

Nishanth Mundru
Benefits Science Technologies, 140 Kendrick St, Building C, West Needham, MA 02494,
nishanth.mundru@gmail.com,
ORCID: 0000-0001-7177-989X

Abstract: We propose a novel, optimization-based method that takes into account the objective and problem structure for reducing the number of scenarios, m, needed for solving two-stage stochastic optimization problems. We develop a corresponding convex optimization-based algorithm, and show that as the number of scenarios increase, the proposed method recovers the SAA solution. We report computational results with both synthetic and real world data sets that show that the proposed method has significantly better performance for $m = 1 - 2\%$ of n in relation to other state of the art methods (Importance sampling, Monte Carlo sampling and Wasserstein scenario reduction with squared Euclidean norm). Additionally, we propose variants of classical scenario reduction algorithms (which rely on the Euclidean norm) and show that these variants consistently outperform their traditional versions.

Key words: scenario reduction, cost function, two-stage stochastic optimization, Wasserstein distance

1. Introduction

A wide range of decision problems that involve optimization under uncertainty can be formulated as a stochastic optimization problem. For instance, consider a production planning problem, where the decision maker wishes to make strategic decisions on plant sizing and allocating resources among plants in the first stage. Later when demand is realized, the decision maker aims to make tactical decisions about storing, processing and shipping these products to the market sources, all while ensuring minimal expected costs and satisfying relevant plant capacity constraints. The main idea in this approach is that taking this second stage decision-making into account leads to better first-stage strategic decisions.

More generally, such problems fall in the setting where a practitioner aims to select the best possible decision that satisfies certain constraints, but with the knowledge that the outcome of this decision is influenced by the realization of a random event. The quality of a decision is judged by averaging its cost over all possible realizations of this random event. These models can be applied to formulate problems in various areas such as finance, energy, fleet management, and supply chain optimization, to name a few. For a more comprehensive list of applications, we refer the reader to Wallace and Ziemba (2005).

Traditional stochastic optimization formulates this as finding an optimal decision, which among all feasible candidates in the set $\mathcal{Z}$, has the lowest average cost when averaged over all possible realizations of the uncertain parameter Y. In other words, these problems can be formulated as

$$\min_{z \in \mathcal{Z}} \mathbb{E}_Y[c(z;Y)]. \tag{1}$$

For instance, in inventory management problems, the uncertainty Y may refer to demand data, or time series of stock returns in portfolio optimization problems. More specifically, we assume that the cost function has the form

$$c(z;\xi) = f'z + \min_{y \in \mathcal{Y}(z,\xi)} g'y, \tag{2}$$

where y, a function of the first stage decision z and observed uncertainty ξ, is the recourse decision. The set $\mathcal{Y}(z,\xi)$ of feasible second stage decisions, for a fixed first-stage decision z and uncertainty ξ, is given by

$$\mathcal{Y}(z,\xi) = \Big\{ y \in \mathbb{R}_+^{d_y} : A(\xi)z + Wy \geq R\xi \Big\}.$$

We assume $\mathcal{Z}$, the set of feasible decisions, is a non-empty convex compact set, and is independent of uncertainty Y.

We provide an example of such a cost function. Consider the application of portfolio optimization, where the decision maker seeks to distribute their capital among a portfolio of assets with uncertain returns in a way that leads to high returns and low risk.

$$c((z,\beta);Y) = -\lambda z'Y + \beta + \frac{1}{\epsilon}\max\{-z'Y - \beta, 0\},$$

where Y and z are vectors of stock returns and corresponding investments (decision variable) respectively, and β is an auxiliary decision variable. Minimizing the cost $c(z;Y)$ ensures high returns $z'Y$, while at the same time controlling the risk, which here is given by CVaR (Conditional Value-at-Risk) of negative returns at level ϵ, as

$$\text{CVaR}_\epsilon(z'Y) = \inf_\beta \ \beta + \frac{1}{\epsilon}\mathbb{E}[\max\{-z'Y - \beta, 0\}].$$

The quantities $\epsilon \in (0,1)$ which parametrizes the risk measure, and $\lambda > 0$, the trade-off between risk and return, are pre-specified parameters.

While we wish to solve Problem (1), the true distribution of the uncertainty Y is typically unknown. Even if it is fully known, solving the exact optimization problem may not be tractable. In the context of data-driven stochastic optimization, where past data consisting of n samples of uncertainty $\xi^1, \ldots, \xi^n$ is assumed to be known, a popular approach to approximate Problem (1) is Sample Average Approximation (SAA) (Shapiro et al. 2009, Birge and Louveaux 2011). Under this approach, the problem we wish to solve is

$$\min_{z\in\mathcal{Z}} \frac{1}{n}\sum_{i=1}^{n} c(z;\xi^i). \tag{3}$$

It is easy to see that this approach, in effect, approximates the unknown full distribution with the empirical distribution with each data point ξ^i equally likely. In fact, Kleywegt et al. (2002) show that, under some regularity conditions, the optimal objective value and solution of Problem (3) converge to their counterparts of Problem (1) as n increases, regardless of the distribution of ξ. For more recent advances in SAA, we direct the reader towards Homem-de Mello and Bayraksan (2014), Rahimian et al. (2018) and the references therein.

In this paper, we consider the approach of scenario reduction, which approximates the empirical distribution on $\{\xi^1, \ldots, \xi^n\}$ with another distribution $\mathbb{Q}$ with (fewer) scenarios $\zeta^1, \ldots, \zeta^m$ and corresponding probabilities $q_1, \ldots, q_m$, for some positive integer m. Since ideally we wish to choose $m << n$, we shall refer to $\mathbb{Q}$ as a reduced (or smaller) distribution. In explicit terms, we consider Problem (3) as the true problem that we wish to solve, where n and the samples $\{\xi^1, \ldots, \xi^n\}$ are

fixed and available to the decision maker. Our goal is to compute a new (discrete) distribution $\mathbb{Q}$ with scenarios $\{\zeta^1, \ldots, \zeta^m\}$ and probabilities $q^1, \ldots, q^m$ such that the solution of the SAA problem on this distribution $\mathbb{Q}$ is close (in terms of decision quality) to the solution of Problem (3). To be more precise, we use knowledge of the cost function and constraints while computing this reduced distribution $\mathbb{Q}$, which typically results in better approximations. When n is very large and thus the SAA problem (3) is not computationally tractable, such an approach can substantially improve tractability while ensuring minimal loss in decision quality. Another key advantage incurred for practitioners when solutions of higher quality are computed with significantly lesser scenarios is interpretability. This can be valuable for decision makers who seek intuition on scenarios that affect the cost, which can help to understand the solution better and guide policy. In this paper, we demonstrate that using optimization to compute these smaller sets of scenarios taking into account the cost function can increase tractability, accuracy over cost-agnostic scenario reduction methods, and interpretability.

1.1. Contributions

The contributions of this work are as follows:

1. We present a novel optimization-based approach for scenario reduction for two-stage stochastic optimization problems. As part of this approach, we introduce a quantity we term as "Problem-dependent divergence" that takes into account the quality of decisions induced by two discrete distributions, and generalizes the Wasserstein distance. We prove a stability result which states that under certain regularity conditions, minimizing this quantity leads to distribution with an optimal objective value close to the SAA objective.
2. We consider scenario reduction in this setting, and present algorithms for computing these scenarios and corresponding probabilities. Our approach relies on an alternating-minimization algorithm, where we iteratively optimize for the scenarios and update cluster assignments. Since the problem of computing the scenarios is in general nonconvex, we instead solve a convex problem that is an upper bound.

3. We show that this convex upper bound, under certain conditions on the distribution of ξ and structure of cost function c, leads to scenarios that coincide with the solution to Wasserstein scenario reduction with Euclidean loss. Additionally, our result provides intuition about how our upper bound objective finds scenarios that affect the objective of the original stochastic optimization problem.
4. Finally, with the help of computational results we demonstrate the effectiveness of these methods on various data sets – both synthetic and standard test problems from the stochastic optimization literature. We compare our method to traditional scenario reduction approach based on Wasserstein distance with Euclidean norm and sampling-based methods such as Monte Carlo and Importance sampling, and demonstrate that it performs favorably compared to these other methods when m is just $1-2\%$ of n. Additionally, we propose variants of classical scenario reduction algorithms (which rely on the Euclidean norm) that instead use our Problem dependent divergence and show that these variants consistently outperform their traditional versions.

1.2. Related Work

In this section, we review related approaches for stochastic optimization problems that have been proposed in the literature. Dupačová et al. (2003) present theory and algorithms for scenario reduction using probability metrics, while Heitsch and Römisch (2003) derive bounds for forward and backward scenario selection heuristics. For a comprehensive review that extends these ideas to the multistage setting, we refer the reader to Pflug and Pichler (2014).

More recently, Rujeerapaiboon et al. (2017) establish fundamental performance guarantees for Wasserstein distance-based scenario reduction, and develop a polynomial-time constant-factor approximation algorithm based on a well-known local search algorithm in Arya et al. (2004) as well as an exact mixed integer optimization reformulation for the scenario reduction problem. Our theoretical result on worst-case error bounds builds on some of their ideas. However, a key difference from their work is that we do not use the Euclidean norm to encode the distance between two

scenarios while computing the Wasserstein distance, but use a novel divergence which we define in this paper.

Arya et al. (2004) also analyze local-search algorithms algorithms for the $k-$median clustering and facility location problems and prove a constant-factor approximation guarantee. In this work, we adapt their local search algorithm originally developed for the $k-$median clustering problem for the scenario reduction problem in our setting. A key difference is that our focus in this work is to develop a scenario reduction algorithm which relies on a novel divergence between two scenarios that takes into account the objective and constraint structure of the decision problem.

We note a recently growing stream of research that develops scenario reduction techniques for specific objectives and problem structures. Recently, Henrion and Römisch (2018) propose a problem-based scenario generation method which involves solving a generalized semi-infinite optimization problem to compute a smaller approximation of the empirical distribution. While this global approach finds the distribution that leads to the best uniform approximation of $\mathbb{E}[c(z;\xi)]$ over all feasible z, the tractability of this approach is not clear as the resulting problem is a generalized semi infinite optimization problem.

Among other more tractable approaches, Keutchayan et al. (2021) propose a mixed-integer optimization-based methodology for scenario reduction that relies on the cost-difference of scenarios rather than a norm-based difference. They develop a methodology to form clusters based on the optimal objective values of scenarios, but they do not take into account stability issues. In other recent work, Hewitt et al. (2021) propose a decision-based clustering method for scenario reduction. They define an opportunity cost between two scenarios that is identical to our symmetric problem-dependent divergence, and use graph clustering techniques to cluster points based on this cost. Using these clusters, they choose a representative scenario for each cluster. However it is not exactly clear how to choose such representative scenarios. Another key difference between our work and both these approaches is that our method allows for choosing scenarios that are not necessarily a part of the observed sample points.

For problems with a Conditional Value at Risk (CVaR) objective, Arpón et al. (2018) propose an algorithm that finds the relevant scenarios from among the original empirical scenarios. For the same problem, Pineda and Conejo (2010) redefine the distance $d(\xi, \zeta)$ between two scenarios ξ and ζ based on the CVaR objective, but this new definition violates the property that $d(\xi, \zeta) = 0 \iff \xi = \zeta$, which is important for stability purposes. Morales et al. (2009) propose a general scenario reduction technique based on difference between optimal objective values of scenarios but do not discuss stability issues. More recently, Fairbrother et al. (2015) develop the notion of a risk region, and selectively sample scenarios belonging to that region for problems optimizing tail measures. Finally, we note that our approach is not restricted to any specific problem or objective type and is designed for the general case. Rahimian et al. (2018) propose a scenario reduction method specifically for distributionally robust stochastic optimization problems.

As part of our approach, we define a novel measure, that captures the difference between two distributions by taking into account their induced decisions. Using the notion of Wasserstein distance between two discrete distributions, we propose a different measure that takes decision quality into account and analyze scenario reduction in this context. We note the recent progress in data-driven Distributionally Robust Optimization (DRO), where it has been shown that the worst-case expectation of an uncertain cost over all distributions that are within a fixed Wasserstein distance from a discrete reference distribution can often be computed efficiently via convex optimization (Esfahani and Kuhn 2018, Gao and Kleywegt 2016).

Replacing the empirical n-point distribution with a new m-scenario distribution can lead to significant computational gains in this context, as DRO problems over Wasserstein balls are harder to solve than their stochastic versions. This tractability advantage may be significant for the case of two-stage distributionally robust linear problems, which admit semidefinite optimization problems as tight approximations (Hanasusanto and Kuhn 2018). This problem of approximating distributions is closely related to the optimal quantization of probability distributions, which approximates a non-discrete initial distribution with an m-point distribution. These connections between optimal quantization and stochastic optimization are discussed in greater detail in Pflug and Pichler (2011)

Other approaches for scenario generation/reduction have been developed in the literature that determine scenarios matching a set of statistical properties, such as moment-matching (Høyland and Wallace 2001, Høyland et al. 2003). For instance, in Høyland et al. (2003) the authors attempt to find a distribution that matches the first four marginal moments and the correlations, through a least squares model. However, it is not clear beforehand which statistical properties are important for solution quality, while our approach focuses on approximating distributions that ensure high solution quality.

More broadly, other methods for solving stochastic optimization problems include Monte Carlo-based sampling and variance reduction. In a recent review on SAA (Kim et al. 2015), the authors note that a moderately large sample is likely to compute decisions of satisfactory quality for some problems. Linderoth et al. (2006) study the empirical behavior of such sampling methods for solving SAA problems, while Xiao and Zhang (2014) study variance reduction in detail. For a survey of Monte Carlo based sampling methods and variance reduction techniques for stochastic optimization, we direct the reader to Homem-de Mello and Bayraksan (2014), Bayraksan and Morton (2011) and the references therein.

Our work also belongs to the area of research demonstrating the advantages of optimization over randomization. Some related papers include Bertsimas et al. (2015) and Bertsimas et al. (2019), where the authors consider the problems of offline and sequential experimental design respectively with covariate matching and demonstrate that using mixed integer optimization to reduce mean discrepancy between groups, rather than randomization, leads to stronger inference. While our application is different, we emphasize the key insight that selecting scenarios via optimization, in this case taking the decision quality into account, can lead to better decisions with fewer scenarios and thus, greater tractability and interpretability.

1.3. Notation

Let e be the vector of all ones, and e_i the i^{th} standard basis vector of appropriate dimensions. For any positive integer n, we define the set $[n] = \{1, \ldots, n\}$. We denote a generic norm by $\|\cdot\|$, while

$\|\cdot\|_p$ denotes the $p-$norm, for $p \geq 1$. Recall that the Euclidean norm of any vector x is defined as $\|x\|_2 = \sqrt{\sum_i x_i^2}$. For a set $\mathcal{X} \in \mathbb{R}^d$, we define $\mathcal{P}(\mathcal{X}, m)$ as the set of all probability distributions supported on at most m points belonging to $\mathcal{X}$. The support of a probability distribution $\mathbb{P}$ is denoted by supp($\mathbb{P}$), and the Dirac delta distribution at ξ denoted by $\delta(\xi)$. We define $\mathbb{P}_n(\xi^1, \ldots, \xi^n)$ as the uniform distribution supported on the n distinct points ξ^i, which we equivalently represent as

$$\mathbb{P}_n(\xi^1, \ldots, \xi^n) = \sum_{i=1}^{n} \frac{1}{n} \delta(\xi^i).$$

Cost functions are denoted by $c(z; Y)$, where $z \in \mathbb{R}^{d_z}, Y \in \mathbb{R}^d$ represent the decision variable and uncertainty respectively, and $\mathcal{Z} \subseteq \mathbb{R}^{d_z}$ represents the non empty convex set of feasible decisions. For any given ξ, we assume that $c(z; \xi)$ is a convex function of z. Finally, we denote

$$c^*(\xi) = \min_{z \in \mathcal{Z}} c(z; \xi),$$

the optimal objective value corresponding to the scenario ξ, where we assume $c^*(\xi)$ to be finite for every ξ. Next, instead of a single scenario, given a distribution $\mathbb{Q}$, we define an optimal solution $z^*(\mathbb{Q})$ and the objective value $v(\mathbb{Q})$ as

$$v(\mathbb{Q}) = \min_{z \in \mathcal{Z}} \mathbb{E}_{Y \sim \mathbb{Q}}[c(z; Y)],$$

$$z^*(\mathbb{Q}) \in \arg\min_{z \in \mathcal{Z}} \mathbb{E}_{Y \sim \mathbb{Q}}[c(z; Y)].$$

In this paper, we restrict our analysis to discrete distributions, and $\mathbb{Q}$ is defined by a finite set of scenarios and their corresponding probabilities. Thus, the optimal decision for the (discrete) distribution $\mathbb{Q}$ (with scenarios $\zeta^1, \ldots, \zeta^m$ and their corresponding probabilities $q_1, \ldots, q_m$) is given by

$$z^*(\mathbb{Q}) \in \arg\min_{z \in \mathcal{Z}} \sum_{j=1}^{m} q_j \, c(z; \zeta^j).$$

1.4. Structure of the paper

The structure of this paper is as follows. In Section 2, we present some preliminary background on the concept of Wasserstein distance and scenario reduction for stochastic optimization. In Section 3, we present more details about our approach, where we justify our approach using ideas

and results from stability theory, and formulate the scenario reduction problem in our context. Next, in Section 4 we present an algorithm for estimating these new scenarios (or distributions), and present some theoretical justification in Section 4.4. We provide computational evidence of the scenario reduction method developed in this paper by comparing it with other state-of-the-art methods on real and synthetic data in Section 5, and finally, present our conclusions in Section 6.

2. Preliminaries

In this section, we briefly review the Wasserstein distance, which defines a distance between two probability distributions, and the scenario reduction problem.

2.1. Distance between (discrete) distributions

Let $\mathbb{P}$ be a discrete probability distribution on scenarios $\xi^1, \ldots, \xi^n$ with corresponding probabilities $p_1, \ldots, p_n$, and $\mathbb{Q}$ another discrete distribution on scenarios $\zeta^1, \ldots, \zeta^m$ with corresponding probabilities $q_1, \ldots, q_m$. Next, we define the Wasserstein distance between these two discrete probability distributions.

Definition 1. The Wasserstein distance (induced by the ℓ_2 norm) between two discrete distributions $\mathbb{P}$ and $\mathbb{Q}$, which we denote as $\mathcal{D}_W(\mathbb{P}, \mathbb{Q})$, is defined as the square root of the optimal objective value of the following problem:

$$\begin{aligned} \min_{\pi \in \mathbb{R}_+^{n \times m}} \quad & \sum_{i=1}^{n} \sum_{j=1}^{m} \pi_{ij} \|\xi^i - \zeta^j\|^2 \\ \text{subject to} \quad & \sum_{j=1}^{m} \pi_{ij} = p_i, \quad \forall i \in [n], \\ & \sum_{i=1}^{n} \pi_{ij} = q_j, \quad \forall j \in [m]. \end{aligned} \tag{4}$$

The linear optimization problem (4) used to define the Wasserstein distance can be interpreted as a minimum-cost transportation problem from n sources to m destinations. Here, π_{ij} represents the amount of probability mass shipped from ξ^i to ζ^j at unit transportation cost $\|\xi^i - \zeta^j\|^2$. Note that the probabilities π_{ij} sum to one, as

$$\sum_{i=1}^{n} \sum_{j=1}^{m} \pi_{ij} = \sum_{i=1}^{n} p_i = 1 = \sum_{j=1}^{m} q_j,$$

and thus is not included in Problem (4) since it is a redundant constraint. Therefore, Problem (4) is an optimal transportation problem that aims to minimize the overall cost of moving probability mass from the initial distribution $\mathbb{P}$ to the target distribution $\mathbb{Q}$.

Next, we briefly review the idea of scenario reduction, which approximates a distribution supported on n scenarios by another distribution supported on m scenarios, with m chosen to be smaller, typically sometimes significantly, than n. As part of this approach, both the new reduced set of scenarios $\{\zeta^j\}_{j=1}^m$ and their corresponding probabilities $\{q_j\}_{j=1}^m$ are estimated. In the next section, we describe the two variants of scenario reduction – discrete and continuous.

2.2. Scenario reduction

In this section, we define the scenario reduction problem and its two variants. For notational convenience, we denote $\mathbb{P}_n(\xi^1, \ldots, \xi^n)$ as $\mathbb{P}_n$.

Definition 2. The discrete scenario reduction problem is defined as

$$\mathbb{D}_W(\mathbb{P}_n, m) = \min_{\mathbb{Q}} \left\{ \mathcal{D}_W(\mathbb{P}_n, \mathbb{Q}) \ : \mathbb{Q} \in \mathcal{P}(\{\xi^1, \ldots, \xi^n\}, m) \right\} \tag{5}$$

Definition 3. The continuous scenario reduction problem is defined as

$$\mathbb{C}_W(\mathbb{P}_n, m) = \min_{\mathbb{Q}} \left\{ \mathcal{D}_W(\mathbb{P}_n, \mathbb{Q}) \ : \mathbb{Q} \in \mathcal{P}(\mathbb{R}^d, m) \right\} \tag{6}$$

In Problem (5), the new scenarios, which are a part of the reduced distribution $\mathbb{Q}$, must be selected from among the support of the empirical distribution, given by the set $\{\xi^1, \ldots, \xi^n\}$. In contrast, the continuous scenario reduction problem (6) allows the scenarios to be chosen from outside the set of observations, and offers better flexibility and approximation to the empirical distribution.

However, in both these settings, the approximate distributions are computed without taking into account the cost function $c(z; \xi)$ and the feasible set $\mathcal{Z}$. We address this in the following section, where we first define a generalization of the Wasserstein distance between two distributions, and later use it to compute scenarios tailored for the optimization problem at hand.

3. Problem-Dependent Scenario Reduction

In this section, we present our approach of scenario reduction, which we denote as PDSR in short, where we restrict ourselves to two-stage stochastic optimization problems.

3.1. Problem Setting

Using the cost function defined in (2), we consider the general problem setting, given by

$$\min_{z\in\mathcal{Z}} \quad f'z + \mathbb{E}_{\xi}\big[\min_{y\in\mathcal{Y}(z,\xi)} g'y\big].$$

Here for any z and uncertainty value ξ, the second stage problem is given by the following linear optimization problem

$$\begin{aligned} \min_{y\geq 0} \quad & g'y \\ \text{subject to} \quad & A(\xi)z + Wy \geq R\xi. \end{aligned}$$

We assume the matrix $A(\xi)$ is a known affine function of the uncertainty, i.e.,

$$A(\xi) = A^0 + \sum_{p=1}^{d} \xi_p A^p,$$

for known matrices of appropriate dimensions $A^0, A^1, \ldots, A^d$.

Next, we introduce two assumptions that are fairly common in the stochastic optimization literature.

Assumption 1. *Relatively complete recourse: $\mathcal{Y}(z,\xi) \neq \varnothing$ for any feasible first stage decision $z \in \mathcal{Z}$ and uncertainty ξ.*

Assumption 2. *Fixed recourse: The second-stage cost vector g and recourse matrix W are not affected by uncertainty.*

Assumption 2 can also be stated as uncertainty only affects the right hand side $R\xi - A(\xi)z$. Consequently, the (non-empty) dual polyhedron of the second-stage problem $\{\lambda \geq 0 | W'\lambda \leq g\}$ is identical for all realizations of ξ. We note that the fixed recourse property has been exploited in various efficient methods to solve two-stage stochastic optimization problems (Higle and Sen 1991, Homem-de Mello and Bayraksan 2014).

Under this assumption, when the second-stage has only continuous variables which enables the use of linear optimization duality, the cost function $c(z,\xi)$ in (1) is of the form

$$c(z;\xi) = f'z + \max_{\lambda \geq 0, W'\lambda \leq g} \lambda' \Big(R\xi - \big(A^0 + \sum_{p=1}^{d} \xi_p A^p\big) z \Big).$$

3.2. Problem-dependent divergence

In this section, we present our definition of problem-dependent divergence, to quantify the difference between two scenarios that takes into account the problem structure. We note that our definition includes the Wasserstein distance with ℓ_2 norm as a special case.

First, we define $z^*(\eta)$ as an optimal decision corresponding to the scenario η, and is given by

$$z^*(\eta) \in \arg\min_{z \in \mathcal{Z}} c(z;\eta).$$

For simplicity, we assume that there exists a unique optimal solution for every possible scenario η, but we relax this assumption later.

Next, we define a variant of the Wasserstein distance metric between two probability distributions $\mathbb{P}, \mathbb{Q}$ with respect to the cost c and constraint set $\mathcal{Z}$ as $\mathcal{D}(\mathbb{Q}, \mathbb{P}|c, \mathcal{Z})$.

Definition 4. Let $\mathbb{P}$ and $\mathbb{Q}$ be two discrete probability distributions in $\mathbb{R}^d$, given by

$$\mathbb{P} = \sum_{i=1}^{n} p_i \, \delta(\xi^i), \; \mathbb{Q} = \sum_{j=1}^{m} q_j \, \delta(\zeta^j)$$

respectively. Then, $\mathcal{D}(\mathbb{Q}, \mathbb{P}|c, \mathcal{Z})$ is given by the square root of the optimal objective value of the following linear optimization problem:

$$\begin{aligned} \mathcal{D}^2(\mathbb{Q}, \mathbb{P}|c, \mathcal{Z}) = \min_{\pi \in \mathbb{R}_+^{n \times m}} \quad & \sum_{j=1}^{m} \sum_{i=1}^{n} \pi_{ij} \left(c\big(z^*(\zeta^j); \xi^i\big) - c\big(z^*(\xi^i); \xi^i\big) \right) \\ \text{subject to} \quad & \sum_{j=1}^{m} \pi_{ij} = p_i, \quad \forall i \in [n], \\ & \sum_{i=1}^{n} \pi_{ij} = q_j, \quad \forall j \in [m]. \end{aligned} \tag{7}$$

We denote this as the problem-dependent divergence between the two distributions $\mathbb{P}$ and $\mathbb{Q}$, with respect to the cost function $c(z;y)$ and constraint set $\mathcal{Z}$. It is a non-symmetric measure of the difference between two probability distributions, and hence not a metric distance. Specifically, it is a measure of the loss in decision quality when $\mathbb{Q}$ is used to approximate $\mathbb{P}$. We observe that the optimal value of the optimization problem (7) is guaranteed to be non negative, as

$$c(z;\xi^i) \geq c(z^*(\xi^i);\xi^i) = \min_{\hat{z}\in\mathcal{Z}} c(\hat{z};\xi^i) \quad \forall z \in \mathcal{Z},$$

and hence, each term is positive for any choice of $\mathbb{Q}$.

We discuss the similarities between $\mathcal{D}(\mathbb{Q},\mathbb{P}|c,\mathcal{Z})$ and $D_W(\mathbb{P},\mathbb{Q})$ in greater detail in Section 4.3.

3.3. Stability

Recall $v(\mathbb{P}) = \min_{z\in\mathcal{Z}} \mathbb{E}_{Y\sim\mathbb{P}}[c(z;Y)]$, the optimum cost assuming distribution $\mathbb{P}$ for uncertainty Y. A stability result bounds the difference between $v(\mathbb{P})$ and $v(\mathbb{Q})$, where $\mathbb{Q}$ is an approximation of the distribution $\mathbb{P}$, in terms of a distance metric between $\mathbb{P}$ and $\mathbb{Q}$. Thus, this motivates scenario reduction, where if we have a new distribution $\mathbb{Q}$ with a smaller support than the original distribution $\mathbb{P}$ but is close enough to $\mathbb{P}$ in terms of this metric, then we obtain a problem that is not only computationally easier to solve than (3) but also is guaranteed to provide a good approximation in terms of optimal values.

In the definition of Wasserstein distance with Euclidean metric in (4), the unit transportation cost between two scenarios ξ and ζ was set to be $\|\xi-\zeta\|^2$, but generally it represents a distance between ξ and ζ, which we denote as $d(\xi,\zeta)$. Thus, the problem-dependent divergence can be viewed as a general Wasserstein-type distance with the cost function replaced by

$$d(\xi,\zeta) = c(z^*(\zeta);\xi) - c(z^*(\xi);\xi)$$

In order to derive a stability result corresponding to this d, we first note that d need not be a distance metric; it only needs to satisfy the following two conditions:

1. Symmetricity, i.e., $d(\xi',\xi) = d(\xi,\xi')$

2. $d(\xi,\xi')=0 \iff \xi=\xi'$. Note that this ensures $\mathcal{D}(\mathbb{P},\mathbb{Q})=0 \iff \mathbb{P}=\mathbb{Q}$.

Thus, we define

$$d_S(\xi,\zeta)=\frac{1}{2}\left(d(\xi,\zeta)+d(\zeta,\xi)\right).$$

Clearly, condition 1 is satisfied as this ensures that $d_S(\xi,\xi')=d_S(\xi',\xi)$. In order to satisfy the second condition, we first note that $\xi=\zeta \implies d_S(\xi,\zeta)=0$. For the other direction, we note that $d_S(\xi,\zeta)=0$ implies that both the terms $d(\xi,\zeta), d(\zeta,\xi)$ are each 0. Thus, we see that

$$d_S(\xi,\zeta)=0 \implies z^*(\xi)=z^*(\zeta).$$

Consequently, we introduce the following assumption to ensure that Condition 2 holds.

Assumption 3. *For any two scenarios ξ,ζ, we must have that*

$$z^*(\xi)=z^*(\zeta) \implies \xi=\zeta.$$

In other words, if the optimal decisions for two scenarios are the same, then the two scenarios must themselves be identical. For instance, if $c(z;Y)=\frac{1}{2}z'A(Y)z+z'b$ with constraints $\mathcal{Z}=\mathbb{R}^{n_z}$ and $A(Y)\succ 0$ for any random Y with $A(\xi)=A(\zeta) \implies \xi=\zeta$, then this condition holds.

However, this assumption can be restrictive. For instance, when the cost function c is a point wise maximum of linear functions (such as the CVaR function for instance), this condition may not hold. In this case, we propose adding a scaled norm term encoding the distance between ξ,ζ, i.e.,

$$d_S(\xi,\zeta)=\mu\frac{1}{2}\left(d(\xi,\zeta)+d(\zeta,\xi)\right)+(1-\mu)\|\xi-\zeta\|_p^q, \tag{8}$$

for some $\mu\in(0,1)$, for $p,q\in\{1,2\}$. In this case, it is easy to see that $d_S(\xi,\zeta)=0 \implies \xi=\zeta$ and hence Assumption 3 is no longer necessary. In the rest of the theoretical section of this paper, we set $\mu=1$ term for brevity but all the proofs and algorithms can be adapted for $\mu<1$.

Next, using this framework, we modify the definition in (7) to formally define $\mathcal{D}_S(\mathbb{Q},\mathbb{P}|c,\mathcal{Z})$ as

$$\begin{aligned}\mathcal{D}_S^2(\mathbb{Q},\mathbb{P}|c,\mathcal{Z}) = \min_{\pi\in\mathbb{R}_+^{n\times m}} \quad & \frac{1}{2}\sum_{j=1}^{m}\sum_{i=1}^{n}\pi_{ij}\left(c\big(z^*(\zeta^j);\xi^i\big) - \min_{z\in\mathcal{Z}} c(z;\xi^i) + c\big(z^*(\xi^i);\zeta^j\big) - \min_{z\in\mathcal{Z}} c(z;\zeta^j)\right)\\ \text{subject to} \quad & \sum_{j=1}^{m}\pi_{ij} = p_i, \quad \forall i\in[n],\\ & \sum_{i=1}^{n}\pi_{ij} = q_j, \quad \forall j\in[m],\end{aligned} \tag{9}$$

which we shall refer to as "Problem-dependent divergence" (or, PDD) from now onwards in the rest of this paper.

Next, in order to define a stability result in terms of $\mathcal{D}_S$, we introduce the following assumption.

Assumption 4. *There exists a nondecreasing function $h:\mathbb{R}_+\to\mathbb{R}_+$ with $h(0)=0$ such that*

$$|c(z;\xi)-c(z;\zeta)| \le h(\|z\|)\, d_S(\xi,\zeta).$$

Such a condition can be proven for particular classes of problems when d_S is the Euclidean norm; for instance if the set of dual solutions of the second stage problem is bounded and locally Lipschitz continuous, then this assumption follows from Proposition 3.3 in Römisch and Wets (2007).

Finally, for Theorem 1 to hold, we note that Assumption 2 can be replaced by a less stringent assumption that requires dual feasibility for any scenario ξ. Note that Assumption 2 of fixed recourse implies that the dual feasibility condition is satisfied.

Next, we present the following stability result. For ease of exposition, we present it for the case of discrete random variables, but we note that it can be replicated for the continuous case as well (Dupačová et al. 2003). We assume that ξ is a random vector which takes values in the finite set Ξ, with $|\Xi| = N$.

Corollary 1. *Under Assumptions 1, 2, 3 and 4, for a given discrete distribution $\mathbb{P}$ supported on the set Ξ of N scenarios with probabilities $p_1,\ldots,p_n$ and cost function c with constraint set $\mathcal{Z}$, there exist constants $\rho>0$ and $\epsilon>0$ (which depend on the distribution $\mathbb{P}$, c and the set $\mathcal{Z}$) such that whenever $\mathbb{Q}$ lies in the set of distributions whose support is contained in Ξ with*

$$\sup_{z\in\mathcal{Z}\cap\rho\mathbb{B}}\left|\sum_{i=1}^{N}c(z;\xi^i)p(\xi^i) - \sum_{i=1}^{N}c(z;\xi^i)q(\xi^i)\right| \le \epsilon,$$

we have

$$|v(\mathbb{P}) - v(\mathbb{Q})| \leq h(\rho)\, \mathcal{D}_S^2(\mathbb{Q}, \mathbb{P}|c, \mathcal{Z}).$$

Here the set $\mathbb{B} \in \mathbb{R}^{d_z}$ *is the ball centered in the origin with unit radius, and* $p(\xi), q(\xi)$ *are the probabilities of scenario* ξ *under distributions* $\mathbb{P}, \mathbb{Q}$ *respectively.*

Proof The result follows from Proposition 3.3 in (Römisch and Wets 2007).

REMARK 1. We note that Assumption 3 can be relaxed with the addition of a norm penalty to d_S, as described in (8).

Corollary 1 implies that if we have a new distribution $\mathbb{Q}$ that is reasonably close to the empirical distribution $\mathbb{P}$, then the absolute value of the difference between optimal objective function value corresponding to $\mathbb{Q}$ and the SAA objective will be bounded by the value of the problem-dependent divergence $\mathcal{D}_S^2$ between both distributions multiplied by a constant, $h(\rho)$. For a more detailed discussion on this general topic, we direct the reader's attention to Rachev (1991).

3.4. Problem Formulation

Analogous to Problem (6), we define the continuous problem-dependent scenario reduction problem as

$$\mathbb{C}^2(\mathbb{P}_n, m; c, \mathcal{Z}) = \min_{\mathbb{Q}} \left\{ \mathcal{D}_S^2(\mathbb{Q}, \mathbb{P}_n|c, \mathcal{Z}) \, : \mathbb{Q} \in \mathcal{P}(\mathbb{R}^d, m) \right\}. \tag{10}$$

We denote by $\mathcal{B}(I, m)$ the family of all $m-$set partitions of the set I, i.e.,

$$\mathcal{B}(I, m) = \Big\{ \{I_1, \ldots, I_m\} \, : \varnothing \neq I_1, \ldots, I_m \subseteq I, \cup_j I_j = I, I_i \cap I_j = \varnothing \; \forall i \neq j \Big\}.$$

Also, we denote a specific $m-$set partition as $\{I_j\} \in \mathcal{B}(I, m)$. Next, we present the following result, which is similar to Theorem 1 in Rujeerapaiboon et al. (2017), that reformulates the continuous problem-dependent scenario reduction problem (10) as a set partitioning problem.

THEOREM 1. *The problem-dependent scenario reduction problem* (10) *can be written as the following problem of finding an* $m-$*set partition that optimizes the following problem:*

$$\begin{aligned} \mathbb{C}^2(\mathbb{P}_n, m; c, \mathcal{Z}) = \min_{I_j \in \mathcal{B}(I,m)} \quad & \frac{1}{n} \sum_{j=1}^{m} \min_{\zeta^j} \sum_{i \in I_j} \frac{1}{2} \Big(c\big(z^*(\zeta^j); \xi^i\big) - \min_{z \in \mathcal{Z}} c(z; \xi^i) + \\ & c\big(z^*(\xi^i); \zeta^j\big) - \min_{z \in \mathcal{Z}} c(z; \zeta^j) \Big). \end{aligned} \tag{11}$$

Proof Following the argument of Theorem 2 in Dupačová et al. (2003), the optimal problem-dependent divergence (PD) between $\mathbb{P}_n$ and any distribution $\mathbb{Q}$ supported on a finite set Ψ is given by

$$\min_{\mathbb{Q}\in\mathcal{P}(\Psi,\infty)} \mathcal{D}_S^2(\mathbb{Q},\mathbb{P}_n|c,\mathcal{Z}) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{2}\min_{\zeta\in\Psi}\Big(c\big(z^*(\zeta);\xi^i\big) - c^*(\xi^i) + c\big(z^*(\xi^i);\zeta\big) - c^*(\zeta)\Big),$$

where $\mathbb{P}(\Psi,\infty)$ denotes the set of all probability distributions supported on the finite set Ψ. The continuous scenario reduction problem (6), but with PD instead of the Euclidean distance, can be written as the following problem of finding the set Ψ with m elements that leads to the smallest objective value. Letting $\Psi = \{\zeta^1,\ldots,\zeta^m\}$, we have

$$\mathbb{C}^2(\mathbb{P}_n,m;c,\mathcal{Z}) = \min_{\zeta^1,\ldots,\zeta^m}\frac{1}{n}\sum_{i=1}^{n}\min_{j\in[m]}\Big(c\big(z^*(\zeta^j);\xi^i\big) - c^*(\xi^i) + c\big(z^*(\xi^i);\zeta^j\big) - c^*(\zeta^j)\Big). \tag{12}$$

Next, we show that Problem (12) is equivalent to Problem (11). Given an optimal solution $\zeta_*^1,\ldots,\zeta_*^m$ to Problem (12), we construct a partition such that

$$I_j = \left\{i : c\big(z^*(\zeta_*^j);\xi^i\big) + c\big(z^*(\xi^i;\zeta^j)\big) - c^*(\zeta^j) = \min_{k\in[m]}\left\{c\big(z^*(\zeta_*^k);\xi^i\big) + c\big(z^*(\xi^i;\zeta^k)\big) - c^*(\zeta^k)\right\}\right\}$$

which leads to Problem (11) having the same objective as Problem (12). For the other direction, given an optimal partition $I_1,\ldots,I_m$ and corresponding inner problem-minimizing scenarios $\zeta_*^1,\ldots,\zeta_*^m$ of Problem (11), it is easy to see that these scenarios will be an optimal solution of Problem (12) with identical objective value. This completes the proof. □

We emphasize that the support information or more general additional constraints on the uncertainties can also be included in this optimization-based scenario reduction framework. In the case when uncertainties are constrained to belong to a known set $\mathcal{S}$, we modify Problem (11) by solving the following problem,

$$\mathbb{C}^2(\mathbb{P}_n,m;c,\mathcal{Z}) = \min_{I_j\in\mathcal{B}(I,m)}\ \frac{1}{n}\sum_{j=1}^{m}\min_{\zeta^j\in\mathcal{S}}\ \sum_{i\in I_j}\frac{1}{2}\Big(c\big(z^*(\zeta^j);\xi^i\big) - \min_{z\in\mathcal{Z}}c(z;\xi^i) + c\big(z^*(\xi^i);\zeta^j\big) - \min_{z\in\mathcal{Z}}c(z;\zeta^j)\Big).$$

Problem (11) can also be interpreted as a clustering problem, where the n points $\xi^1,\ldots,\xi^n$ are partitioned into m clusters with centroids $\zeta^1,\ldots,\zeta^m$. Both the cluster assignments and the centroids

within each cluster are chosen to minimize the cumulative problem-dependent divergence to the n sample points. For the j^{th} cluster comprising of points I_j, each optimal scenario ζ_*^j is chosen as the solution of the following problem:

$$\zeta_*^j \in \arg\min_{\zeta} \sum_{i \in I_j} \Big(c\big(z^*(\zeta);\xi^i\big) + c\big(z^*(\xi^i);\zeta\big) - \min_{z \in \mathcal{Z}} c(z;\zeta) \Big).$$

When $m = n$, then the scenarios $\zeta^i = \xi^i,\ \forall i \in [n]$ as $D_S(\mathbb{P}|\mathbb{P}; c, \mathcal{Z}) = 0$. Thus, the optimal decision is the same as SAA solution, which is the best that can be computed given this training data.

Next, we present a result that illustrates the flexibility of this approach. Specifically, for a class of cost functions, we show that this framework finds the SAA solution with just one scenario, i.e., when $m = 1$.

PROPOSITION 1. *Assume the cost function has the form $c(z;\xi) = \frac{1}{2} z'Hz - z'v(\xi) + u(\xi)$, with positive definite matrix H and a full rank transformation $v(\cdot)$, i.e., $v(\xi) = v(\zeta) \implies \xi = \zeta$ that spans the entire space $\mathbb{R}^d$ (Note that $u(\xi)$ can be any arbitrary function of the uncertainty ζ). Also, suppose the problem is unconstrained, i.e., the constraint set $\mathcal{Z} = \mathbb{R}^{n_z}$. Under these assumptions and when $m = 1$, solving the continuous scenario reduction problem* (10) *leads directly to the SAA solution.*

Proof First, we begin by noting that the (unique) optimal solution corresponding to uncertainty value ξ is given by

$$z^*(\xi) = H^{-1} v(\xi),$$

for any ξ. Next, with some algebra, we see that $d_S(\xi, \zeta)$ is given by the following expression,

$$d_S(\xi,\zeta) = \big(v(\xi) - v(\zeta)\big)' H^{-1} \big(v(\xi) - v(\zeta)\big).$$

The scenario reduction problem reduces to computing ζ_* by solving

$$\min_{\zeta} \sum_{i=1}^{n} \big(v(\xi^i) - v(\zeta)\big)' H^{-1} \big(v(\xi^i) - v(\zeta)\big).$$

If $\frac{1}{n} \sum_{i=1}^{n} v(\xi^i)$ lies in the range of the mapping $v(\cdot)$, which is satisfied as $v(\cdot)$ spans the entire space $\mathbb{R}^d$ by our assumption, then the optimal ζ_* will indeed satisfy

$$v(\zeta_*) = \frac{1}{n} \sum_{i=1}^{n} v(\xi^i).$$

Finally, we note that the optimal decision under this new point distribution is given by

$$z^*(\zeta_*) = H^{-1}v(\zeta_*) = \frac{1}{n}\sum_{i=1}^{n} H^{-1}v(\xi^i),$$

which is the SAA solution. □

Before we proceed, we note that the cost function in Proposition 1, while restrictive, serves as an illustrative example to motivate our approach. Problems with such structure appear in response surface analysis for stochastic programming problems, where a second order local approximation of the cost function is studied (see e.g., Bailey et al. (1999)).

We point out that in this result, solving the scenario reduction problem in this setting with $m = 1$ still requires solving the full SAA problem with n scenarios. While this means there is no computational advantage by solving with a single scenario in this setting, it does indicate the flexibility and modeling power of our approach. We emphasize that while traditional scenario reduction aims to compute $\mathbb{Q}$ "close" to $\mathbb{P}$, problem-dependent scenario reduction takes into account the quality of decisions induced.

We note that in the presence of constraints or for other objective functions the estimation of $z^*(\eta)$ may not, in general, be given by a closed form expression or even be unique. To address this issue, we introduce a variant of PDD, where we consider the worst case over the set of optimal solutions $\mathcal{Z}^*(\zeta)$, for every ζ.

To be precise, we define

$$\mathcal{Z}^*(\zeta) = \{z \in \mathcal{Z} : c(z;\zeta) \leq \min_{\hat{z}\in\mathcal{Z}} c(\hat{z};\zeta)\}, \tag{13}$$

and modify the definition presented in (9) as

$$\begin{aligned} \mathcal{D}_S^2(\mathbb{Q},\mathbb{P}|c,\mathcal{Z}) = \min_{\pi\in\mathbb{R}_+^{n\times m}} \quad & \sum_{j=1}^{m}\sum_{i=1}^{n} \pi_{ij}\left[\max_{z\in\mathcal{Z}^*(\zeta^j)} \left\{c(z;\xi^i)\right\} - c^*(\xi^i) + c\big(z^*(\xi^i);\zeta^j\big) - c^*(\zeta^j)\right] \\ \text{subject to} \quad & \sum_{j=1}^{m} \pi_{ij} = p_i, \quad \forall i \in [n], \\ & \sum_{i=1}^{n} \pi_{ij} = q_j, \quad \forall j \in [m]. \end{aligned} \tag{14}$$

Note that when $\mathcal{Z}^*(\zeta^j)\forall j \in [m]$ are each singleton sets, then (9) and (14) are identical. We write this equivalently as

$$\begin{aligned} \mathcal{D}_S^2(\mathbb{Q},\mathbb{P}|c,\mathcal{Z}) = \min_{\pi\in\mathbb{R}_+^{n\times m}} \quad & \sum_{j=1}^{m}\sum_{i=1}^{n}\pi_{ij}\left[\max_{z\in\mathcal{Z}^*(\zeta^j)}\left\{c(z;\xi^i)-c(z;\zeta^j)\right\}-c^*(\xi^i)+c\big(z^*(\xi^i);\zeta^j\big)\right] \\ \text{subject to} \quad & \sum_{j=1}^{m}\pi_{ij}=p_i, \quad \forall i\in[n], \\ & \sum_{i=1}^{n}\pi_{ij}=q_j, \quad \forall j\in[m]. \end{aligned} \tag{15}$$

Next, we present our approach of scenario reduction in this framework.

4. Optimization Approach

In this section, we present our optimization-based approach for computing these scenarios. First, we propose an algorithm motivated by Lloyd's algorithm for k-means clustering (Lloyd 1982), where it alternate between computing m scenarios and updating assignments of points to the m clusters represented by these scenarios.

4.1. Alternating-optimization framework

We assume a given initial solution of assignments π and scenarios $\zeta^1,\ldots,\zeta^m$. The algorithm proceeds in an iterative manner, where given the m scenarios, the assignments of n points to m clusters (or π variables) are updated in order to minimize $\mathcal{D}_S^2$. Once the assignments are fixed, the scenarios $(\zeta^1,\ldots,\zeta^m)$ are updated by solving an appropriate optimization problem within each cluster. Once the change in the scenario values between successive iterations falls below a threshold or a maximum number of iterations is reached, we terminate the algorithm. We repeat this procedure with multiple random initial starts of assignments $\pi^{(0)}$ and scenarios $\zeta_{(0)}$, and choose the solution with lowest in-sample $\mathcal{D}_S^2$. We keep track of the best solution so far at each iteration, and stop when there is no change in the objective.

Algorithm 1 Alternating-Minimization algorithm for PDSR

1: **procedure** APPROXIMATE SOLUTION FOR PROBLEM (11).

2: Start with random assignments $\pi^{(0}$ and $\zeta_{(0)} = (\zeta^1, \ldots, \zeta^m)$.

3: Initialize $t \leftarrow 1$

4: **while** $\Delta > TOL$ and $t < MAX\_ITER$ **do**

5: For fixed $\zeta^1_{(t)}, \ldots, \zeta^m_{(t)}$, assign points i to cluster $j(i)$, where

$$j(i) \in \arg\min_{1 \le j \le m} \mathcal{D}_S(\xi^i, \zeta^j_{(t)})$$

6: For each $1 \le i \le n$, set $\pi^{(t)}_{i,j} \leftarrow 1$ if $j = j(i)$ and $\pi^{(t)}_{i,j} = 0$ else.

7: For points in each cluster $j \in [m]$, solve for $\zeta^j_{(t+1)} \in \arg\min_\zeta \sum_{i=1}^{n} \pi^{(t)}_{ij} \mathcal{D}^2_S(\xi^i, \zeta_{(t)})$.

8: Compute objective $\theta_{(t)} = \sum_{i \in [n]} \pi^t_{i,j(i)} \mathcal{D}^2_S(\xi^i, \zeta^{j(i)}_{(t)})$.

9: Keep track of best objective so far, with $t^* = \arg\min_{i \in [t]} \theta_{(i)}$.

10: Update $\zeta_{(t)} \leftarrow \zeta_{(t^*)}$.

11: $\Delta \leftarrow \frac{1}{md} \|\zeta_{(t)} - \zeta_{(t-1)}\|$.

12: Update $t \leftarrow t + 1$

13: **end while**

14: **end procedure**

A feature of this algorithm is that it is parallelizable, as each of the m scenarios, in Step 7, can be computed in parallel. Next, we discuss methods to estimate the m reduced scenarios required in Step 7 of Algorithm 1.

4.2. Optimization approach for an upper bound

In this section, we present our approach for computing the scenarios. We restrict our analysis for cost functions of the form

$$c(z; \xi) = \max_{1 \le t \le k} z' A_t \xi, \tag{16}$$

for known matrices $A_t, 1 \leq t \leq k$. We emphasize that in such a representation, the number of k could be exponentially large, as each piece represents an extreme point of the set $\{\lambda \geq 0, W'\lambda \leq g\}$. For smaller values of k, this is an approximate representation of the full cost function, but which can be made arbitrarily close by sampling additional dual points λ. Finally, we let the constraint set $\mathcal{Z}$ be a polytope given by

$$\mathcal{Z} = \{z \in \mathbb{R}^{n_z}_+ : Pz \leq q\}.$$

Our strategy relies on solving a convex upper bound of the objective in (15). First, we note the following result that provides an upper bound.

PROPOSITION 2. *We have*

$$\max_{z \in \mathcal{Z}^*(\zeta)} c(z;\xi) \leq \max_{z \in \mathcal{Z}} \Big\{ c(z;\xi) - \hat{\alpha} c(z;\zeta) \Big\} + \hat{\alpha} \min_{\hat{z} \in \mathcal{Z}} c(\hat{z};\zeta).$$

for any $\hat{\alpha} \geq 0$.

Proof We begin the proof by noting that

$$\begin{aligned}
&\max_{z \in \mathcal{Z}^*(\zeta)} && c(z;\xi) \\
=&\max_{z \in \mathcal{Z}} \inf_{\alpha \geq 0} && c(z;\xi) + \alpha\big(- c(z;\zeta) + \min_{\hat{z} \in \mathcal{Z}} c(\hat{z};\zeta)\big) \\
\leq &\inf_{\alpha \geq 0} \max_{z \in \mathcal{Z}} && c(z;\xi) + \alpha\big(- c(z;\zeta) + \min_{\hat{z} \in \mathcal{Z}} c(\hat{z};\zeta)\big),
\end{aligned}$$

by using the definition of $\mathcal{Z}^*(\zeta)$ in (13) and weak duality. The result follows by considering a fixed $\hat{\alpha} \geq 0$. $\square$

PROPOSITION 3. *For a piecewise bilinear cost function of form* (16)*, a convex upper bound problem for estimating* ζ^j*, for a given set of points* I_j*, is given by the following problem*

$$\begin{aligned}
\min_{\zeta,\lambda,\theta,\gamma} \quad & \sum_{i \in I_j} \theta_i + 2\gamma_i \\
\text{subject to} \quad & z^*(\xi^i)' A_t \zeta \leq \gamma_i \; \forall t \in [k], i \in I_j, \\
& q'\lambda^{t,i} \leq \theta_i \; \forall t \in [k], i \in I_j, \\
& P'\lambda^{t,i} \geq A_t(\xi^i - 2\zeta) \; \forall t \in [k], i \in I_j, \\
& \lambda^{t,i} \geq 0 \; \forall t \in [k], i \in I_j.
\end{aligned} \tag{17}$$

Proof For any two scenarios ξ, ζ, we bound $d_S(\xi, \zeta)$ by

$$\begin{aligned}
d_S(\xi,\zeta) &\leq \max_{z\in\mathcal{Z}} \big(c(z;\xi) - c(z;\zeta) - \hat{\alpha}c(z;\zeta)\big) + c(z^*(\xi);\zeta) + \hat{\alpha}\min_{\eta\in\mathcal{Z}} c(\eta;\zeta) - c(z^*(\xi);\xi),\\
&= \max_{z\in\mathcal{Z}} \big(c(z;\xi) - (\hat{\alpha}+1)c(z;\zeta)\big) + \hat{\alpha}\min_{\eta\in\mathcal{Z}} c(\eta;\zeta) + c(z^*(\xi);\zeta) - c(z^*(\xi);\xi),\\
&\leq \max_{z\in\mathcal{Z}} \big(c(z;\xi) - 2c(z;\zeta)\big) + \min_{\eta\in\mathcal{Z}} c(\eta;\zeta) + c(z^*(\xi);\zeta) - c(z^*(\xi);\xi),\\
&\leq \max_{z\in\mathcal{Z}} \big(\max_{t\in[k]}\{z'A_t\xi\} - 2\max_{t\in[k]}\{z'A_t\zeta\}\big) + 2\max_{t\in[k]}\{z^*(\xi)'A_t\zeta\} - c(z^*(\xi);\xi),\\
&\leq \max_{z\in\mathcal{Z}}\max_{t\in[k]}\{z'A_t(\xi-2\zeta)\} + 2\max_{t\in[k]}\{z^*(\xi)'A_t\zeta\} - c(z^*(\xi);\xi)\\
&= \max_{t\in[k]}\max_{z\in\mathcal{Z}} z'A_t(\xi-2\zeta) + 2\max_{t\in[k]}\{z^*(\xi)'A_t\zeta\} - c(z^*(\xi);\xi).
\end{aligned}$$

The first inequality follows from the definition of $d_S(\cdot)$ and Proposition 2, the third inequality follows from setting $\hat{\alpha} = 1$, and the fourth from the definition of c and replacing $\min_{\eta\in\mathcal{Z}} c(\eta;\zeta)$ with $c(z^*(\xi);\zeta)$, which is a further upper bound. Thus, summing these terms over points in I_j and omitting the constant terms $c(z^*(\xi^i);\xi^i)$, we solve the following approximate convex problem for ζ^j

$$\min_{\zeta}\sum_{i\in I_j}\big(\max_{t\in[k]}\{\max_{z\in\mathcal{Z}} z'A_t(\xi^i-2\zeta)\} + 2\max_{t\in[k]}\{z^*(\xi^i)'A_t\zeta\}\big).$$

Reformulating this objective using linear optimization duality gives us the desired result. □

In Step 7 of Algorithm 1, we solve Problem (17) to compute the new reduced scenarios.

Before we present our theoretical results, we discuss the key points of how our method is connected to the Euclidean-norm based Wasserstein scenario reduction.

4.3. Connection with Euclidean norm-based Wasserstein scenario reduction

Finally, we note that our approach shares some similarities and some key differences as compared to Wasserstein scenario reduction with the Euclidean norm.

1. We emphasize the equivalence of $d_S(\cdot,\cdot)$ to traditional Wasserstein distance with Euclidean metric when $c(z;y) = \|z-y\|^2$ and the constraint set as $\mathcal{Z} = \mathbb{R}^d$. To see this, we first note that the optimal decision for any scenario η is given by

$$z^*(\eta) \in \arg\min_{z\in\mathbb{R}^d} \|z-\eta\|^2 = \eta.$$

Hence,

$$c\big(z^*(\zeta);\xi\big) = \|z^*(\zeta)-\xi\|^2,$$
$$= \|\zeta-\xi\|^2,$$

and

$$\min_z\ c(z;\xi) = \|z^*(\xi)-\xi\|^2 = 0,$$

Also, noting that

$$c\big(z^*(\zeta);\xi\big) = \|\zeta-\xi\|^2 = c\big(z^*(\xi);\zeta\big),$$

this leads to the conclusion that

$$d_S(\xi,\zeta) = \|\xi-\zeta\|^2.$$

Note that adding the term $\mu\|\xi-\zeta\|^2$ will simply lead to a scaling factor of $(\mu+1)$.

2. Next, under the same assumption that $c(z;y) = \|z-y\|^2$ and the constraint set as $\mathcal{Z} = \mathbb{R}^d$, then our divergence between the two distributions is exactly identical to the Wasserstein distance between them. To be more precise, $\mathcal{D}_W$ can be recovered as a special case of this divergence, i.e.,

$$\mathcal{D}_S(\mathbb{Q},\mathbb{P}|\|z-y\|^2;\mathbb{R}^d) = \mathcal{D}_W(\mathbb{P},\mathbb{Q}).$$

and we conclude that Problem (7) is equivalent to Problem (4).

3. Also in this case, the scenario reduction problem is equivalent to traditional least squares clustering (Rujeerapaiboon et al. 2017).
4. Notice that in Algorithm 1, at each iteration computing the scenarios reduces to computing the mean of all data points in that cluster. In this case, the objectives $\theta_{(t)}$ are monotonically decreasing with t, and hence the best solution so far just reduces to the current iterate. Consequently in this special case, our optimization algorithm (1) recovers distributions identical to those obtained by the classical $k-$means clustering algorithm.

Next, we discuss statistical properties of our approach where we present conditions under which the solution to the scenario reduction problem in this context reduces to the population mean, which is the point-distribution solution obtained by Wasserstein scenario reduction when $m=1$.

4.4. Justification of the upper bound

In this section, we characterize the solution obtained when optimizing the convex upper bound that we present in Proposition 3. Specifically, we show that under some conditions on the uncertainty distribution and on the cost function, the optimal scenario obtained for $m = 1$ is simply the distribution mean. We let the cost function $c(z;\xi) = \max\{z'\xi, 0\}$, and denote $z^*(\xi) \in \arg\min_{z\in\mathcal{Z}} z'\xi$. We emphasize that this is result is mainly for conceptual understanding of this method, and provides a justification of the upper bound problem (17). First, we present the assumptions on the interplay between the distribution of ξ and the cost c required for our result.

Assumption 5.

a. ξ has a continuous distribution on $\mathcal{U} = \{\xi : \min_{z\in\mathcal{Z}} z'\xi < 0\}$, and a density of 0, elsewhere.

b. The mean $\mathbb{E}[\xi] = \bar{\xi}$ is finite and satisfies $z^*(\bar{\xi})'\bar{\xi} > 0$.

c. The distribution of ξ is symmetric about its mean $\bar{\xi}$.

Regarding assumption 5a, suppose $\mathcal{Z}$ has K extreme points, with $z^1, \ldots, z^K$. Let $\mathcal{U}_-^i = \{\xi : \xi' z^i < 0\}$, and not all of these sets are empty. It requires ξ to have a non-zero density on $\mathcal{U} = \cup_{i=1}^K \mathcal{U}_-^i$, which is not necessarily a convex set. Thus, the mean $\bar{\xi}$, does not necessarily belong to the set $\mathcal{U}$, which is, partly, what assumption 5b requires. Finally, we point out that Assumption 5a does not mean that the optimal cost of $\mathbb{E}[c(z;\xi)]$ is necessarily 0; indeed there still could be certain scenarios ξ for which $c(z;\xi) \geq 0$ at the optimum z. The following result provides intuition on how our upper bound objective focuses on such scenarios.

Theorem 2. *For $m = 1$, if Assumption 5 holds and for this cost function c, the solution to the upper bound problem (17) is simply the mean $\bar{\xi}$, which is also the corresponding solution for Wasserstein scenario reduction.*

Proof The proof technique follows ideas presented in Theorem 1 in Elmachtoub and Grigas (2021). The objective to be minimized can be written as

$$\min_{\zeta} \mathbb{E}[L(\xi,\zeta)] = \mathbb{E}[\max\{\max_{z\in\mathcal{Z}} z'(\xi - 2\zeta), 0\}] + 2\mathbb{E}[\max\{z^*(\xi)'\zeta, 0\}]$$

We restrict our expectation on ξ to be conditional on $\xi \in \mathcal{U}$. Next, we restrict our analysis to ζ that satisfy both $\min_{z\in\mathcal{Z}} z'(2\zeta-\xi) < 0$ and $z^*(\xi)'\zeta > 0$. We will prove that $\bar{\xi}$ satisfies the first condition, with the second condition satisfied by assumption 5.2, which guarantees the existence of at least one feasible ζ. Now, since L is the sum of two functions, where each is a point-wise maximum of a convex function, and hence is itself convex.

Next, we show that $L(\xi,\zeta)$ is finite for any such ζ, as

$$\begin{aligned} L(\xi,\zeta) &\leq |\max_{z\in\mathcal{Z}} z'(\xi-2\zeta)| + 2|\zeta' z^*(\xi)|, \\ &\leq \big(\|\xi-2\zeta\|_1 + 2\|\zeta\|_1\big) \max_{z\in\mathcal{Z}} \|z\|_\infty, \\ &\leq \big(\|\xi\|_1 + 4\|\zeta\|_1\big) \max_{z\in\mathcal{Z}} \|z\|_\infty. \end{aligned}$$

Thus, $\mathbb{E}[L(\xi,\zeta)]$ is bounded, as $\mathbb{E}[\xi]$ is finite and hence $\mathbb{E}[\|\xi\|_1]$ is finite, and $\mathcal{Z}$ is bounded. Finiteness of L implies that the partial derivative can be taken inside the expectation.

Using linearity of expectation and the fact that $\max_{z\in\mathcal{Z}} z'(\xi-2\zeta) = -\min_{z\in\mathcal{Z}} z'(2\zeta-\xi)$, the subdifferentials of $\mathbb{E}[L]$ is given by the sum of $-2\mathbb{E}[\mathcal{Z}^*(2\zeta-\xi)]$ and $2\mathbb{E}[z^*(\xi)]$. Since the distribution of ξ is continuous on $\mathcal{U}$, restricting $2\zeta-\xi$ to belong to $\mathcal{U}$ implies that $\mathcal{Z}^*(2\zeta-\xi)$ is a singleton with probability one.

Under assumption 5c, ξ and $2\bar{\xi}-\xi$ are equal in distribution, which implies that $\mathbb{E}[z^*(2\bar{\xi}-\xi)]$ and $\mathbb{E}[z^*(\xi)]$ are equal, and hence, $0 \in \partial\mathbb{E}[L(\xi,\zeta)]$. Also, $\zeta = \bar{\xi}$ satisfies both $\min_{z\in\mathcal{Z}} z'(2\zeta-\xi) < 0$ and $z^*(\xi)'\zeta > 0$ from Assumptions 5.1 and 5.2 respectively. Thus, we conclude that $\bar{\xi}$ is an optimal solution of the convex Problem (17). $\square$

We emphasize the key intuition that the upper bound, by combining over existing data points, seeks to find scenarios that affect the cost. This follows from the fact that $z^*(\bar{\xi})'\bar{\xi} > 0$ implies that any feasible z will also have a positive cost, i.e., $z'\bar{\xi} \geq z^*(\bar{\xi})'\bar{\xi} > 0 \,\forall z \in \mathcal{Z}$. Finally, we point out that as we consider uncertainties ξ that have $\min_{z\in\mathcal{Z}} z'\xi < 0$, we use $z^*(\xi) \in \arg\min_{z\in\mathcal{Z}} z'\xi$. Thus, this $z^*(\xi)$ is also a solution to $\min_{z\in\mathcal{Z}} \max\{z'\xi, 0\}$, which we denote as $z^*(\xi)$ while defining the upper bound in Proposition 3.

4.5. Performance Bound

In this section, we present a performance bound of our approach. First, we assume that the cost function $c(\cdot,\cdot)$, for some known $M>0$, satisfies

$$c(z;\xi)-\min_{z\in\mathcal{Z}} c(z;\xi)\leq \frac{M}{2}\|z-z^*(\xi)\|^2, \tag{18}$$

with a unique optimal solution $z^*(\xi)$ for all values of uncertainty ξ. Next, we assume that $z^*(\xi)$, as a function of ξ, spans $\mathbb{R}^{d_z}$. Finally, we assume that the set $\mathcal{Z}$ lies within the Euclidean unit norm ball $\|z\|_2\leq 1$. Next, we define the worst-case value of (10) as

$$\mathbb{C}^2(m,n;c,\mathcal{Z})=\max_{\hat{\mathbb{P}}_n\in\mathcal{P}(\mathbb{R}^d,n)}\left\{\mathbb{C}^2(\hat{\mathbb{P}}_n,m;c,\mathcal{Z}):\|\xi\|_2\leq 1\,\forall\xi\in\text{supp}(\hat{\mathbb{P}}_n)\right\} \tag{19}$$

THEOREM 3. *The worst-case quantity $\mathbb{C}(m,n;c,\mathcal{Z})$ is bounded above by $\sqrt{M\frac{n-m}{n-1}}$.*

Proof We note that

$$\begin{aligned}
\mathbb{C}^2(m,n;c,\mathcal{Z}) &= \max_{\mathbb{P}_n\in\mathcal{P}(\mathbb{R}^d,n)}\left\{\min_{\mathbb{Q}} \quad \left\{\mathcal{D}_S^2(\mathbb{Q},\mathbb{P}_n|c,\mathcal{Z})\ :\mathbb{Q}\in\mathcal{P}(\mathbb{R}^d,m)\right\}:\|\xi\|_2\leq 1\,\forall\xi\in\text{supp}(\mathbb{P}_n)\right\},\\
&= \max_{\|\xi^i\|_2\leq 1\forall i\in[n]}\ \min_{I_j\in\mathcal{B}(I,m)} \quad \frac{1}{n}\sum_{j=1}^m \min_{\zeta^j}\sum_{i\in I_j}\frac{1}{2}\Big(c\big(z^*(\zeta^j);\xi^i\big)-\min_{z\in\mathcal{Z}}c(z;\xi^i)+\\
&\qquad\qquad c\big(z^*(\xi^i);\zeta^j\big)-\min_{z\in\mathcal{Z}}c(z;\zeta^j)\Big),\\
&\leq \max_{\|\xi^i\|_2\leq 1\forall i\in[n]}\ \min_{I_j\in\mathcal{B}(I,m)} \quad \frac{1}{n}\sum_{j=1}^m \min_{\zeta^j}\sum_{i\in I_j}\frac{M}{2}\|z^*(\xi^i)-z^*(\zeta^j)\|^2,\\
&= \max_{\|\xi^i\|_2\leq 1\forall i\in[n]}\ \min_{I_j\in\mathcal{B}(I,m)} \quad \frac{1}{n}\sum_{j=1}^m\sum_{i\in I_j}\frac{M}{2}\|z^*(\xi^i)-\text{mean}(z^*(\xi^i):i\in I_j)\|^2,\\
&\leq \max_{\|z^i\|_2\leq 1\forall i\in[n]}\ \min_{I_j\in\mathcal{B}(I,m)} \quad \frac{M}{n}\sum_{j=1}^m\sum_{i\in I_j}\|z^i-\text{mean}(z^i:i\in I_j)\|^2,\\
&= \qquad\qquad M\,\frac{n-m}{n-1}.
\end{aligned}$$

The first inequality follows from (18), the following equality from our assumption as the existence of ζ that $z^*(\zeta)=\text{mean}(z^*(\xi^i):i\in I_j)$ for any I_j is guaranteed, the next inequality from optimizing directly over the decisions z^i and relaxing them to lie in the unit norm ball, and the final equality from Theorem 2 in Rujeerapaiboon et al. (2017). $\square$

Suppose the cost function is not strongly convex, but has the form

$$c(z;\xi) = f'z + \max_{1\le t\le k} z'A_t\xi.$$

In this case, for each ξ the function $c(z;\xi)$ has Lipschitz constant $\hat{L}(\xi) = \|f\|_2 + \max_{1\le t\le k}\|A_t\xi\|_2$ (Nesterov 2003). Since we have $\|\xi\|_2 \le 1$, taking the supremum of $\hat{L}(\xi)$ over this unit ball of ξ, we have for any $L \ge \tilde{L} = \|f\|_2 + \max_{1\le t\le k}\|A_t\|_2$ (here the matrix norm $\|A\|_2 = \sqrt{\lambda_{\max}(A^TA)}$, the square root of the largest eigen value of A^TA) and for any $z, \hat{z} \in \mathcal{Z}$

$$|c(z;\xi) - c(\hat{z};\xi)| \le L\|z-\hat{z}\|_2 \quad \forall \|\xi\|_2 \le 1.$$

In this case, we prove that this worst-case cost is bounded above by $\sqrt{\tilde{L}\frac{n-m}{n-1}}$.

THEOREM 4. *The worst-case quantity* $\mathbb{C}^2(m,n;c,\mathcal{Z})$ *is bounded above by* $\sqrt{\tilde{L}\frac{n-m}{n-1}}$.

Proof The proof follows along the lines of Theorem 3, where we rely on the bound of Theorem 3 in Rujeerapaiboon et al. (2017) for the final step, and hence omit it for the sake of brevity. □

5. Computational Examples

In this section, we compare our method of scenario reduction with other approaches from the literature. We consider a synthetic example of portfolio optimization, followed by some test examples from the stochastic optimization literature, including airlift allocation (gbd) and electric power expansion planning (APL1P). First, we present some details on our computational experiments.

5.1. Experiment Details

For each problem, we first fix a value of n, the size of training set. We generate n training samples of uncertainty $\xi^1, \ldots, \xi^n$ from the population distribution, and run the scenario reduction methods for different values of m. Once each of these methods outputs a final distribution, we compute the optimal first-stage decision under each, and expose it to a test set, which we generate from the same population distribution. We denote the Prescriptive cost (for an output distribution $\mathbb{Q}$ obtained by any scenario reduction method) on a test set S as:

$$v_{\mathbb{Q}} = \frac{1}{|S|}\sum_{i\in S} c(z^*(\mathbb{Q});\xi^i).$$

We compare this cost with the out-of-sample cost obtained by using the empirical distribution $\mathbb{P}_n$, i.e., the SAA solution on the full training set.

In our computational results, we present the following three metrics for each method:

1. Out-of-sample performance gap for each method, defined as, for an output probability distribution $\mathbb{Q}$,

$$\delta_{\mathbb{Q}} = \frac{v_{\mathbb{Q}} - v_{\mathbb{P}_n}}{|v_{\mathbb{P}_n}|} \times 100\%.$$

   The smaller this value of δ, the better the distribution $\mathbb{Q}$ in terms of decision quality.
2. Fraction of instances where each method has the lowest out-of-sample cost (also expressed as percentage). If multiple methods tie for the lowest out-of-sample cost, we count it as a success for all those methods.
3. Run times for each method.

For each n, we sample a different training set sample of size n from a fixed reference distribution and run various methods on this set. We compute the three metrics described above for each instance separately for various values of m. We repeat this twenty times for different training sets of size n and report the average values of these three metrics across these twenty experiments as a function of m.

5.2. Method Details

In this section, we present details on the methods we use in our experiments. We compare the following methods:

1. Euclidean-norm based Wasserstein scenario reduction (WSR),
2. our method of Problem-dependent scenario reduction, which we denote as PDSR,
3. Dupačová's algorithm (Dupačová et al. 2003), which we denote as DPCV,
4. Dupačová's algorithm with d_S, which we denote as DPCV-pdsr,
5. an adaptation of a popular local search algorithm for $k-$median clustering (Arya et al. 2004) as described in Rujeerapaiboon et al. (2017), which we denote as LS,

6. the local search algorithm with d_S, which we denote as LS-pdsr,
7. an exact MILP approach, which we denote as MILP,
8. the MILP approach with d_S, which we denote as MILP-pdsr,
9. Importance sampling (IS), and
10. Monte Carlo sampling (MC).

We implement Euclidean norm-based Wasserstein scenario reduction (WSR) via $k-$means clustering on the training set, using the Clustering package in Julia. For our PDSR method, the reduced scenarios are computed using the alternating-optimization Algorithm 1 with ten different random restarts, similar in spirit to the classical k-means clustering algorithm. We set the maximum iterations parameter MAX_ITER to 20 and the convergence tolerance TOL to 10^{-3}. Since our Algorithm 1 is parallelizable, we present the maximum run time over these random restarts. To ensure all methods are at an equal footing, we do not include any support information in our scenario reduction approach, potentially weakening our method.

For the sake of completeness, we next describe DPCV, LS, and MILP algorithms and each of their modifications (with d_S instead of the Euclidean norm which we denote as DPCV-pdsr, LS-pdsr, and MILP-pdsr respectively) in detail.

The following three methods (and their modifications) are all designed for the discrete scenario reduction problem (5); these methods can only pick scenarios from the training set.

Dupačová's algorithm We outline Dupačová's algorithm for Problem (5) below.

Algorithm 2 Dupačová's algorithm for Problem (5).

1: **procedure** Approximate solution for Problem (5).

2: Initialize the reduced set $R \leftarrow \emptyset$.

3: Select the next scenario ζ to be added to R as

$$\zeta \in \arg\min_{\eta \in \{\xi_1, \ldots, \xi_n\}} D_W(\mathbb{P}_n, R \cup \{\eta\})$$

and update $R \leftarrow R \cup \{\zeta\}$.

4: Repeat Step 3 till $|R| = m$.

5: **end procedure**

Here, D_W denotes the Wasserstein distance with squared Euclidean norm (4). Given an empirical distribution $\mathbb{P}_n$, this algorithm iteratively adds scenarios to the reduced set R, where each scenario is selected greedily. Note that at each stage the reduced distribution $\mathbb{Q}$ on the reduced set R can be recovered as $\sum_{\zeta \in R} q_\zeta \mathbb{1}(\zeta)$ with the probabilities $q_\zeta = \frac{|I_\zeta|}{n}$. The sets $\{I_\zeta\}$ constitute a partition of $\{\xi_1, \ldots, \xi_n\}$ such that I_ζ contains all elements of $\{\xi^1, \ldots, \xi^n\}$ closest to ζ in terms of the squared Euclidean norm, with ties broken arbitrarily. We refer to the solution of this method as DPCV in our experiments.

Using the same Algorithm (2), we replace $D_W(\mathbb{P}_n, R)$ by $\mathcal{D}_S(\mathbb{P}_n, R|c, \mathcal{Z})$, i.e., by simply replacing $\|\xi^i - \xi^j\|_2^2$ with $d_S(\xi^i, \xi^j)$ and report the performance of this method as DPCV-pdsr.

Local Search Algorithm Here, we describe an adaptation of a local search algorithm for $k-$median clustering (Arya et al. 2004) as described in Rujeerapaiboon et al. (2017).

Algorithm 3 Local Search algorithm for Problem (5).

1: **procedure** Approximate solution for Problem (5).

2: Initialize the reduced set $R \subseteq \{\xi_1, \ldots, \xi^n\}$ with $|R| = m$, arbitrarily.

3: Select the next exchange to be applied to R as

$$(\zeta, \zeta') \in \arg\min \left\{ D_W\big(\mathbb{P}_n, R \cup \{\eta\} n \{\eta'\}\big) : (\eta, \eta') \in \big(\{\xi^1, \ldots, \xi^n\} n R\big) \times R \right\},$$

and update $R \leftarrow R \cup \{\zeta\} n \{\zeta'\}$ if $D_W\big(\mathbb{P}_n, R \cup \{\zeta\} n \{\zeta'\}\big) < D_W\big(\mathbb{P}_n, R\big)$.

4: Repeat Step 3 till no further improvement is possible.

5: **end procedure**

For a given empirical distribution $\mathbb{P}_n$, this Algorithm (3) constructs a sequence of reduced sets (or scenario sets) R by swapping points ζ' from R and ζ from $\mathbb{P}_n n R$ that maximally reduces the Euclidean norm-based Wasserstein distance $D_W\big(\mathbb{P}_n, R \cup \{\zeta\} n \{\zeta'\}\big)$. As Rujeerapaiboon et al. (2017) point out, for performance reasons, this 'best fit' strategy can also be replaced with a 'first fit' strategy that performs a swap the first time a reduction in $D_W(\mathbb{P}_n, R)$ is found. We report the performance of this method as LS.

Using the same Algorithm (3), we replace $D_W(P_n, R)$ by $\mathcal{D}_S(\mathbb{P}_n, R|c, \mathcal{Z})$, i.e., by simply replacing $\|\xi^i - \xi^j\|_2^2$ with $d_S(\xi^i, \xi^j)$ and report the performance of this method as LS-pdsr.

Exact MILP approach We present the exact MILP reformulation (for e.g., see Rujeerapaiboon et al. (2017), Heitsch and Römisch (2003)) as follows, with the matrix $D \in \mathbb{R}_+^{n \times n}$ given by $D_{ij} = \|\xi_i - \xi_j\|_2^2$.

$$
\begin{aligned}
\min_{\pi,\lambda} \quad & \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\pi_{ij}D_{ij} \\
\text{subject to} \quad & \sum_{j=1}^{n}\pi_{ij} = 1 \quad \forall i \in [n], \\
& \pi_{ij} \leq \lambda_j \quad \forall i \in [n], j \in [n], \\
& \sum_{j=1}^{n}\lambda_j = m, \\
& \pi \in \mathbb{R}_{+}^{n\times n}, \lambda \in \{0,1\}^n.
\end{aligned} \tag{20}
$$

We refer to the performance of the solution obtained by this method in our experiments as MILP. Next, we use the same formulation (20) but replace D_{ij} with $\tilde{D}_{ij} = d_S(\xi, \xi_j)$, and refer to it as MILP-pdsr. We warm start both the MILP formulations with the solutions from the respective local search solutions (LS and LS-pdsr), and set a time limit of 1800 seconds. Finally, we add a norm-penalty term (as described in(8)) as the ℓ_1 norm scaled by μ, which we choose via three fold cross validation.

Finally, we note that Rujeerapaiboon et al. (2017) also include a MISOCP approach for exact continuous scenario reduction (with the squared Euclidean norm) as well, but they present results from the formulation of (20) in their computational experiments.

While importance sampling was originally introduced by Infanger (1992), we implement it as presented in Papavasiliou and Oren (2013), where we sample m scenarios based on the relative importances that are obtained by solving the static problem (same recourse value for all uncertainties) over the entire training set. We report the performance of this method as IS. Next, we implement Monte Carlo sampling where we simply sample m scenarios with replacement from the training set,and denote it as MC.

Finally, we compare the performance of all these methods to the performance of the solution obtained by solving the full SAA problem (3) with the expectation taken over all n samples. Clearly this does not depend on m for any given training set of size n and use it to compute the performance gap $\delta_{\mathbb{Q}}$ between the various approximations $\mathbb{Q}$ and the full distribution $\mathbb{P}_n$. We

implement all algorithms in Julia (Bezanson et al. 2012), using the JuMP framework (Dunning et al. 2017) and Gurobi as the optimization solver. All the following experiments were performed on the MIT Sloan's Engaging EOSloan HPC cluster located at the Massachusetts Green High Performance Computing Center (MGHPCC) data center, with each instance allocated 20 GB of memory.

5.3. Portfolio optimization

First, we consider a portfolio optimization problem, for a distribution $\mathbb{Q}$, where the problem is given by

$$\begin{aligned}\big(z^*(\mathbb{Q}),\beta^*(\mathbb{Q})\big) \in \arg\min_{z\in\mathbb{R}^d_+,\beta\in\mathbb{R}} \quad & \mathbb{E}_{Y\sim\mathbb{Q}}\left[\beta+\frac{1}{\epsilon}\max\{-z'Y-\beta,0\}-\lambda z'Y\right]\\ \text{subject to} \quad & e'z=1.\end{aligned}$$

We generate the returns Y sampled as

$$Y=\mu+\Sigma^{\frac{1}{2}}\epsilon,$$

where $\mu\sim N(0,I_{d\times d})$, the noise $\epsilon\sim N(0,\sigma^2 I_{d\times d})$ and the covariance matrix Σ with entries given by

$$\Sigma_{ij}=\rho^{|i-j|}\;\forall 1\le i,j\le d.$$

We sample n points from this distribution, with $d=10$. We fix parameters ϵ,λ as $\epsilon=0.05,\lambda=0.01$, and the correlation parameter $\rho=0.1$. The parameter ρ controls the correlation levels of the stock returns, with $\rho=0$ implying no correlation, while ρ closer to $+1$ (-1) results in more positively (negatively) correlated returns. We threshold the returns data from below to ensure they do not fall below -1.0, and set the noise term $\sigma=2.0$.

To compute the out-of-sample cost, we expose the decision computed by each method to a test set, of size $100,000$ points, generated from the same distribution as the training set and average the cost over these points. Next, we present and discuss the results for $n=1000$ and $n=10,000$. In all the tables that follow, the values represent the median value of that metric and the number in parentheses represents the standard deviation.

m	WSR	PDSR	DPCV	DPCV-pdsr	LS	LS-pdsr	MILP	MILP-pdsr	MC	IS
5	0.050 (0.05)	0.300 (0.11)	0.000 (0.00)	0.300 (0.11)	0.050 (0.05)	0.050 (0.05)	0.050 (0.05)	0.100 (0.07)	0.100 (0.07)	0.100 (0.07)
10	0.000 (0.00)	0.200 (0.09)	0.100 (0.07)	0.200 (0.09)	0.000 (0.00)	0.050 (0.05)	0.050 (0.05)	0.150 (0.08)	0.050 (0.05)	0.200 (0.09)
20	0.000 (0.00)	0.450 (0.11)	0.050 (0.05)	0.000 (0.00)	0.000 (0.00)	0.400 (0.11)	0.000 (0.00)	0.050 (0.05)	0.050 (0.05)	0.000 (0.00)

Table 1 Fraction of instances where each method has lowest out-of-sample prescriptive cost for $n = 1000$ as a function of m, the number of reduced scenarios.

m	WSR	PDSR	DPCV	DPCV-pdsr	LS	LS-pdsr	MILP	MILP-pdsr	MC	IS
5	136.242 (56.66)	122.044 (35.33)	116.388 (70.01)	105.115 (52.43)	113.089 (57.22)	104.526 (57.93)	121.334 (49.19)	116.186 (55.95)	126.811 (40.34)	112.762 (60.29)
10	113.152 (74.17)	95.203 (17.96)	82.755 (40.53)	57.958 (44.81)	74.771 (41.90)	75.116 (35.02)	84.904 (59.75)	64.023 (35.97)	91.767 (41.23)	71.477 (30.61)
20	98.676 (38.24)	36.300 (23.36)	70.487 (32.52)	57.151 (27.18)	65.289 (18.24)	39.955 (19.91)	61.081 (38.00)	45.244 (37.53)	52.990 (56.06)	59.231 (29.04)

Table 2 Median out-of-sample performance gap δ (in %) for various methods as a function of m for $n = 1000$.

m	WSR	PDSR	DPCV	DPCV-pdsr	LS	LS-pdsr	MILP	MILP-pdsr	MC	IS
5	0.90 (0.0)	2.03 (0.3)	1.05 (0.0)	1.58 (0.0)	2.63 (0.3)	5.96 (0.3)	1112.21 (89.1)	821.73 (111.0)	0.00 (0.0)	0.95 (0.0)
10	1.00 (0.0)	1.8 (0.2)	1.93 (0.0)	2.54 (0.0)	9.66 (0.7)	26.60 (1.5)	1803.43 (0.2)	266.83 (38.8)	0.00 (0.0)	1.05 (0.0)
20	1.03 (0.0)	1.1 (0.2)	4.20 (0.1)	5.04 (0.2)	46.99 (4.0)	128.49 (3.9)	1803.35 (0.1)	88.51 (16.8)	0.00 (0.0)	1.10 (0.0)

Table 3 Median run times (in seconds) for various methods as a function of m for $n = 1,000$.

In Tables 1-3, we compare the performance of various scenario reduction methods for various m. For $m = 20$, the PDSR and Local search-pdsr methods dominate the instances in Table 1, along with the lowest average optimality gap in Table 2. For smaller values of m, the MILP-based methods perform better out-of-sample compared to the rest. Another interesting trend from Table 2 is that the PDSR versions of DPCV, LS, and MILP outperform their respective standard versions. Finally, we see that the PDSR method outperforms the Wasserstein scenario reduction method consistently across different values of m. This indicates the value of incorporating the cost function in scenario computation algorithms.

m	WSR	PDSR	DPCV	DPCV-pdsr	LS	LS-pdsr	MC	IS
10	0.000 (0.00)	0.300 (0.15)	0.100 (0.10)	0.000 (0.00)	0.100 (0.10)	0.300 (0.15)	0.100 (0.10)	0.100 (0.10)
20	0.000 (0.00)	0.200 (0.13)	0.00 (0.00)	0.300 (0.15)	0.000 (0.00)	0.300 (0.15)	0.000 (0.00)	0.200 (0.13)

Table 4 Fraction of instances where each method has lowest out-of-sample prescriptive cost for $n = 10,000$.

m	WSR	PDSR	DPCV	DPCV-pdsr	LS	LS-pdsr	MC	IS
10	205.64 (656.6)	89.24 (273.2)	176.58 (224.4)	110.77 (614.5)	177.70 (203.5)	98.62 (484.3)	94.67 (377.9)	108.90 (375.4)
20	196.03 (178.5)	67.18 (103.0)	146.34 (233.0)	50.10 (109.2)	135.92 (130.2)	52.10 (80.3)	89.22 (158.9)	56.97 (385.1)

Table 5 Median out-of-sample performance gap δ (in %) for various methods as a function of m for $n = 10,000$

m	WSR	PDSR	DPCV	DPCV-pdsr	LS	LS-pdsr	MC	IS
10	1.03 (0.0)	154.37 (6.6)	130.91 (5.2)	242.02 (9.2)	920.28 (151.1)	3137.12 (285.7)	0.00 (0.0)	1.14 (0.0)
20	1.11 (0.02)	18.12 (11.20)	430.75 (20.49)	548.02 (21.36)	4606.08 (522.91)	12285.70 (687.74)	0.00 (0.00)	1.17 (0.01)

Table 6 Median computation time (in seconds) for various methods as a function of m for $n = 10,000$.

In tables 4-6, we present the results for $n = 10,000$, where we average the results over 10 instances. We ran out of memory while solving the MILP formulation (20) (as it defines $O(n^2)$ variables) and hence we do not include MILP and MILP-pdsr in our computations. We see similar trends as before, where the PDSR versions of DPCV and LS outperform not only their standard versions, but also the other methods for $m = 20$. For this problem as well, we observe that the PDSR method outperforms the standard WSR method. Since the standard deviations are high compared to the median values, we report the full instance-level data of the out-of-sample gap results in the Appendix EC.3.

Next, we present two test problems (gbd, APL1P) from the stochastic optimization literature, and compare different methods on these problems. The following descriptions of gbd and APL1P are taken from Linderoth et al. (2006) and Infanger (1992) respectively. These problems have piecewise separately linear costs as objectives, with $c(z;\xi) = f'z + \max_{\lambda\in\Lambda}\{\lambda'(Az + B\xi)\}$. We modify our experiments slightly, where we now restrict both the Euclidean-norm based Wasserstein and our method of scenario reduction to choose scenarios from the training set at each iteration, while keeping everything else the same. We run Algorithm 1, with same parameters (number of initial random re-starts, maximum iterations, convergence tolerance), for both Euclidean-norm based Wasserstein scenario reduction and our method of scenario reduction, with the only difference being the objective being minimized. Also, for our method, since estimating the cost function repeatedly is

computationally expensive, we approximate the cost as $\hat{c}(z;\xi) = f'z + \max_{\lambda \in \{\lambda^1,\ldots,\lambda^n\}}\{\lambda'(Az+B\xi)\}$, where $\lambda^1,\ldots,\lambda^n$ are the dual solutions that we pre-compute for each of the n training data points.

5.4. Electric Power Expansion Planning

We consider a test problem, denoted as APLIP in the literature (Infanger 1992), which is a model of a simple power network with one demand region. We consider a larger scale variant of this problem, with 10 generators with different investment and operating costs, and the demand is given by a load duration curve with three load levels: base, medium, and peak. The 10 first-stage decision variables denote the capacities which can be built and operated to meet demands from the three load levels. The dimension of the uncertainty is 13, where the uncertain quantities are the three demand variables, and availabilities of the ten generators. There are 33 second-stage decision variables, which denote the operating levels of generators in each load level and slack variables, which allow unserved demand to be purchased with some penalty and also ensure complete recourse. Finally, we evaluate the out-of-sample performance of all the methods on a test set of size $50,000$ for each instance. We present the full formulation, problem data, and uncertainty distribution in the Appendix EC.1.

m	WSR	PDSR	DPCV	DPCV-pdsr	LS	LS-pdsr	MILP	MILP-pdsr	MC	IS
10	0.000 (0.00)	0.000 (0.00)	0.150 (0.08)	0.100 (0.07)	0.250 (0.10)	0.150 (0.08)	0.200 (0.09)	0.250 (0.10)	0.200 (0.09)	0.000 (0.00)
20	0.000 (0.00)	0.000 (0.00)	0.200 (0.09)	0.050 (0.05)	0.050 (0.05)	0.250 (0.10)	0.150 (0.08)	0.250 (0.10)	0.100 (0.07)	0.000 (0.00)

Table 7 Fraction of instances where each method has lowest out-of-sample prescriptive cost as a function of m.

m	WSR	PDSR	DPCV	DPCV-pdsr	LS	LS-pdsr	MILP	MILP-pdsr	MC	IS
10	17.782 (0.90)	9.697 (0.66)	6.762 (0.74)	7.091 (0.80)	5.561 (0.60)	5.698 (0.64)	5.877 (0.71)	4.726 (0.74)	4.722 (0.85)	29.397 (4.05)
20	12.077 (1.18)	8.147 (0.60)	4.049 (0.63)	4.495 (0.63)	4.793 (0.50)	3.748 (0.44)	4.069 (0.55)	3.275 (0.68)	4.106 (0.69)	38.783 (4.26)

Table 8 Median out-of-sample gap δ (in %) for various methods as a function of m.

m	WSR	PDSR	DPCV	DPCV-pdsr	LS	LS-pdsr	MILP	MILP-pdsr	MC	IS
10	1.40 (0.0)	336.43 (14.9)	1.90 (0.0)	484.32 (14.2)	18.31 (1.4)	494.51 (14.2)	1803.20 (65.7)	111.1 (10.1)	0.00 (0.0)	4.82 (0.2)
20	1.40 (0.0)	217.62 (7.6)	4.27 (0.0)	464.43 (14.5)	80.74 (6.9)	571.55 (12.3)	1803.13 (95.0)	92.4 (13.0)	0.00 (0.0)	4.80 (0.1)

Table 9 Median computation time (in seconds) for various methods as a function of m.

From the results in Table 7, we see that the PDSR variants of LS and MILP outperform the other methods most frequently. This is evident in Table 8, where these two methods have the lowest out-of-sample optimality gap for both $m = 10, 20$. Also, similar to pattern observed in the previous portfolio example, the WSR method is again outperformed by the PDSR method. However, as Table 9 shows, this superior performance comes with a cost as evaluating the optimal costs repeatedly is computationally expensive. Another interesting observation is that the MILP-pdsr method terminates in much lesser time than its allocated time limit of 1800 seconds, which is potentially because of the higher quality of the warm start solution used (recall we use the solution obtained by LS-pdsr).

5.5. Aircraft allocation

This problem, also denoted as gbd in the literature, is derived from the aircraft allocation problem described by Dantzig (1963). Here, aircraft of different types are to be allocated to routes in a way that maximizes profit under uncertain demand. In addition to the cost of operating the aircraft, there are costs associated with bumping passengers when the demand for seats outstrips the capacity. In this model, there are four types of aircraft flying on five routes, and the first-stage variables are the number of aircraft of each type allocated to each route. (Since three of the type-route pairs are infeasible, there are 17 first-stage variables in all.) The first-stage constraints are bounds on the available number of aircraft of each type. The second-stage variables indicate the number of carried passengers and the number of bumped passengers on each of the five routes, and the five second-stage constraints are demand balance equations for the five routes. Each of the five demands is assumed to follow a discrete distribution, as detailed in Linderoth et al. (2006). We use problem data, formulation, and uncertainty distribution from `http://pages.cs.wisc.edu/~swright/stochastic/sampling/`.

Finally, we evaluate the out-of-sample performance of all the methods on a test set of size $50,000$ for each instance. We present the full formulation, problem data, and uncertainty distribution in the Appendix EC.2.

m	WSR	PDSR	DPCV	DPCV-pdsr	LS	LS-pdsr	MILP	MILP-pdsr	MC	IS
5	0.150 (0.08)	0.300 (0.11)	0.000 (0.00)	0.050 (0.05)	0.100 (0.07)	0.200 (0.09)	0.100 0.07)	0.150 (0.08)	0.150 (0.08)	0.150 (0.08)
10	0.250 (0.10)	0.000 (0.00)	0.100 (0.07)	0.150 (0.08)	0.100 (0.07)	0.200 (0.09)	0.150 (0.08)	0.100 (0.07)	0.050 (0.05)	0.200 (0.09)
20	0.250 (0.10)	0.100 (0.07)	0.100 (0.07)	0.100 (0.07)	0.100 (0.07)	0.200 (0.09)	0.150 (0.08)	0.150 (0.08)	0.150 (0.08)	0.100 (0.07)

Table 10 Fraction of instances where each method has lowest out-of-sample prescriptive cost.

m	WSR	PDSR	DPCV	DPCV-pdsr	LS	LS-pdsr	MILP	MILP-pdsr	MC	IS
5	0.427 (0.53)	0.606 (0.33)	1.887 (0.24)	1.764 (0.29)	1.344 (0.31)	1.010 (0.30)	1.357 (0.35)	1.010 (0.30)	2.438 (1.33)	2.806 (1.53)
10	0.305 (0.26)	0.718 (0.33)	0.312 (0.27)	0.389 (0.21)	0.482 (0.23)	0.201 (0.28)	0.482 (0.24)	0.218 (0.28)	1.560 (0.36)	1.243 (0.39)
20	0.066 (0.18)	0.162 (0.23)	0.224 (0.23)	0.137 (0.25)	0.181 (0.16)	0.155 (0.18)	0.273 (0.20)	0.151 (0.20)	0.804 (0.29)	1.182 (0.51)

Table 11 Median out-of-sample performance gap δ (in %) for various methods as a function of m.

m	WSR	PDSR	DPCV	DPCV-pdsr	LS	LS-pdsr	MILP	MILP-pdsr	MC	IS
5	1.29 (0.0)	2.44 (0.0)	0.28 (0.0)	0.09 (0.0)	0.42 (0.0)	0.12 (0.0)	0.57 (0.0)	0.48 (0.0)	0.00 (0.0)	1.21 (0.0)
10	1.30 (0.0)	3.77 (0.1)	0.28 (0.0)	0.10 (0.0)	0.43 (0.0)	0.25 (0.0)	0.52 (0.0)	0.47 (0.1)	0.00 (0.0)	1.23 (0.0)
20	1.32 (0.0)	7.09 (0.1)	0.37 (0.0)	0.12 (0.0)	0.74 (0.0)	0.81 (0.0)	0.61 (0.0)	0.43 (0.0)	0.00 (0.0)	1.22 (0.0)

Table 12 Median computation time (in seconds) for various methods as a function of m.

In this problem, the uncertainties are generated from a discrete distribution with a total of 1024 possibilities. Since choosing $n = 1000$ will lead to insufficient variation between the different instances, we choose $n = 100$ for this problem. From Table 11, we observe that all the methods have an out-of-sample performance gap less than 1% at $m = 10$. Also, from table 10, we see that WSR followed by LS-pdsr have the highest fraction of instances where they are the best performing methods for $m = 10, 20$. The other trend is that the PDSR variants of DPCV, LS, and MILP largely improve over their respective standard versions.

5.6. Observations and Discussion

1. In each of the preceding examples, the performance of all the methods, to a large extent, improves monotonically as m increases. This is expected, as a higher m means higher degrees of freedom for the final reduced distribution.

2. In each of the preceding examples, the recurring pattern we observe is that the PDSR method performs very strongly when m, or the degrees of freedom, is around 1% of n. It consistently outperforms the Euclidean norm-based Wasserstein scenario reduction at the same m for the portfolio and electric power expansion problems. However, this edge decreases with increasing m.
3. Another trend we observe is that the PDSR variants of the state-of-the-art methods such as DPCV, LS, and MILP consistently outperform their standard Euclidean norm versions with lower δ values. This clearly indicates the advantage of incorporating the problem information in the scenario reduction method.
4. Comparing the run times of Euclidean norm-based (DPCV, LS) and PDSR-based methods (DPCV-pdsr, LS-pdsr), the PDSR-based methods typically take more computational effort as they need repeated estimation of the cost function for various candidate scenarios.
5. Finally, we observe in our experiments that the randomization-based scenario reduction methods such as MC and IS are very competitive but are generally outperformed by optimization-based scenario reduction methods in terms of out-of-sample performance gap, particularly when m is very small compared to n. However, these methods are much faster than some of the scenario reduction methods, and thus have a computational advantage.

6. Conclusion

In this paper, we introduced a novel optimization-based framework that combines ideas from scenario reduction and convex optimization to compute scenarios that lead to improved decisions. Unlike most existing approaches, our approach is general and applies in a wide range of settings. We propose a new quantity that generalizes the traditional Wasserstein distance with the Euclidean metric by taking into account the problem structure, i.e., cost and constraints. Under some assumptions, we demonstrate a stability result that establishes that minimizing this quantity leads to good solutions, and propose an algorithm for estimating scenarios in this context. With the help of computational examples on real and synthetic data, we provide evidence that our approach

consistently outperforms other state-of-the-art standard methods such as cost-agnostic standard Wasserstein-based scenario reduction, and random sampling based approaches such as Monte Carlo and Importance sampling. These improvements are significant when m is just $1-2\%$ of n. Additionally, we propose variants of classical scenario reduction algorithms (which traditionally rely on the Euclidean norm) that instead use our Problem dependent divergence and show that these variants consistently outperform their traditional versions. These results demonstrate the effectiveness of incorporating problem information in the scenario reduction methodology.

References

Arpón S, Homem-de Mello T, Pagnoncelli B (2018) Scenario reduction for stochastic programs with conditional value-at-risk. *Mathematical Programming* 170(1):327–356.

Arya V, Garg N, Khandekar R, Meyerson A, Munagala K, Pandit V (2004) Local search heuristics for k-median and facility location problems. *SIAM Journal on computing* 33(3):544–562.

Bailey TG, Jensen PA, Morton DP (1999) Response surface analysis of two-stage stochastic linear programming with recourse. *Naval Research Logistics (NRL)* 46(7):753–776.

Bayraksan G, Morton DP (2011) A sequential sampling procedure for stochastic programming. *Operations Research* 59(4):898–913.

Bertsimas D, Johnson M, Kallus N (2015) The power of optimization over randomization in designing experiments involving small samples. *Operations Research* 63(4):868–876.

Bertsimas D, Korolko N, Weinstein AM (2019) Covariate-adaptive optimization in online clinical trials. *Operations Research* 67(4):1150–1161.

Bezanson J, Karpinski S, Shah VB, Edelman A (2012) Julia: A fast dynamic language for technical computing. *arXiv preprint arXiv:1209.5145* .

Birge JR, Louveaux F (2011) *Introduction to stochastic programming* (Springer Science & Business Media).

Dantzig GB (1963) *Linear programming and extensions* (Princeton university press).

Dunning I, Huchette J, Lubin M (2017) JuMP: A modeling language for mathematical optimization. *SIAM Review* 59(2):295–320.

Dupačová J, Gröwe-Kuska N, Römisch W (2003) Scenario reduction in stochastic programming. *Mathematical Programming* 95(3):493–511.

Elmachtoub AN, Grigas P (2021) Smart "predict, then optimize". *Forthcoming at Management Science* (ePub ahead of print), URL `https://doi.org/10.1287/mnsc.2020.3922`.

Esfahani PM, Kuhn D (2018) Data-driven distributionally robust optimization using the wasserstein metric: Performance guarantees and tractable reformulations. *Mathematical Programming* 171(1-2):115–166.

Fairbrother J, Turner A, Wallace S (2015) Problem-driven scenario generation: an analytical approach for stochastic programs with tail risk measure. *arXiv preprint arXiv:1511.03074* .

Gao R, Kleywegt AJ (2016) Distributionally robust stochastic optimization with wasserstein distance. *arXiv preprint arXiv:1604.02199* .

Hanasusanto GA, Kuhn D (2018) Conic programming reformulations of two-stage distributionally robust linear programs over wasserstein balls. *Operations Research* 66(3):849–869.

Heitsch H, Römisch W (2003) Scenario reduction algorithms in stochastic programming. *Computational optimization and applications* 24(2-3):187–206.

Henrion R, Römisch W (2018) Problem-based optimal scenario generation and reduction in stochastic programming. *Mathematical Programming* URL `http://dx.doi.org/10.1007/s10107-018-1337-6`.

Hewitt M, Ortmann J, Rei W (2021) Decision-based scenario clustering for decision-making under uncertainty. *Annals of Operations Research* 1–25.

Higle JL, Sen S (1991) Stochastic decomposition: An algorithm for two-stage linear programs with recourse. *Mathematics of operations research* 16(3):650–669.

Homem-de Mello T, Bayraksan G (2014) Monte carlo sampling-based methods for stochastic optimization. *Surveys in Operations Research and Management Science* 19(1):56–85.

Høyland K, Kaut M, Wallace SW (2003) A heuristic for moment-matching scenario generation. *Computational optimization and applications* 24(2-3):169–185.

Høyland K, Wallace SW (2001) Generating scenario trees for multistage decision problems. *Management science* 47(2):295–307.

Infanger G (1992) Monte carlo (importance) sampling within a benders decomposition algorithm for stochastic linear programs. *Annals of Operations Research* 39(1):69–95.

Keutchayan J, Ortmann J, Rei W (2021) Problem-driven scenario clustering in stochastic optimization. *arXiv preprint arXiv:2106.11717* .

Kim S, Pasupathy R, Henderson SG (2015) A guide to sample average approximation. *Handbook of simulation optimization*, 207–243 (Springer).

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

Linderoth J, Shapiro A, Wright S (2006) The empirical behavior of sampling methods for stochastic programming. *Annals of Operations Research* 142(1):215–241.

Lloyd S (1982) Least squares quantization in pcm. *IEEE transactions on information theory* 28(2):129–137.

Morales JM, Pineda S, Conejo AJ, Carrion M (2009) Scenario reduction for futures market trading in electricity markets. *IEEE Transactions on Power Systems* 24(2):878–888.

Nesterov Y (2003) *Introductory lectures on convex optimization: A basic course*, volume 87 (Springer Science & Business Media).

Papavasiliou A, Oren SS (2013) Multiarea stochastic unit commitment for high wind penetration in a transmission constrained network. *Operations Research* 61(3):578–592.

Pflug GC, Pichler A (2011) Approximations for probability distributions and stochastic optimization problems. *Stochastic optimization methods in finance and energy*, 343–387 (Springer).

Pflug GC, Pichler A (2014) *Multistage stochastic optimization* (Springer).

Pineda S, Conejo A (2010) Scenario reduction for risk-averse electricity trading. *IET generation, transmission & distribution* 4(6):694–705.

Rachev ST (1991) *Probability metrics and the stability of stochastic models*, volume 269 (John Wiley & Son Ltd).

Rahimian H, Bayraksan G, Homem-de Mello T (2018) Identifying effective scenarios in distributionally robust stochastic programs with total variation distance. *Mathematical Programming* 1–38.

Römisch W, Wets RB (2007) Stability of ε-approximate solutions to convex stochastic programs. *SIAM Journal on Optimization* 18(3):961–979.

Rujeerapaiboon N, Schindler K, Kuhn D, Wiesemann W (2017) Scenario reduction revisited: Fundamental limits and guarantees. *Mathematical Programming* 1–36.

Shapiro A, Dentcheva D, Ruszczyński A (2009) *Lectures on stochastic programming: modeling and theory* (SIAM).

Wallace SW, Ziemba WT (2005) *Applications of stochastic programming* (SIAM).

Xiao L, Zhang T (2014) A proximal stochastic gradient method with progressive variance reduction. *SIAM Journal on Optimization* 24(4):2057–2075.

EC.1. Problem details for (modified) electric power expansion problem (apl1p)

We present the formulation of the electric power expansion problem we use in our experiments. Here, $n_z = 10, d = 13$ and the number of second stage variables is 33.

$$z^*(\mathbb{Q}) \in \arg\min_{z \in \mathbb{R}^{10}_+} \quad \mathbb{E}_{Y\sim\mathbb{Q}}[c(z;Y)]$$

$$\text{subject to} \quad z \geq e.$$

Here, the cost function, for a given uncertainty ξ, is given by

$$c(z;\xi) = c'z + \min_{y,s \in \mathcal{Y}(z,\xi)} \sum_{i=1}^{3}\sum_{j=1}^{10} f_{ij}y_{ij} + p\sum_{i=1}^{3} s_i,$$

where the set $\mathcal{Y}(z,\xi)$ is specified by (y,s) that satisfy the following constraints:

$$\sum_{j=1}^{10} y_{ij} + s_i \geq \xi_i, \quad \forall i \in [3],$$
$$\sum_{i=1}^{3} y_{ij} \leq \xi_{(3+j)} z_j, \quad \forall j \in [10],$$
$$y \in \mathbb{R}^{3\times 10}_+, s \in \mathbb{R}^3_+.$$

Here, e refers to a vector of ones. The coefficients p, c, f are given by $p = 100$, $c_i = 4$ if i is odd and $c_i = 2.5$ if i is even, and

$$f_{1,:} = [4.3, 2.0, 0.5]$$

$$f_{2,:} = [8.7, 4.0, 1.0]$$

$$f_{i,:} = f_{1,:} \quad \text{if } i \geq 3 \text{ and } i \text{ even},$$

$$f_{i,:} = 2f_{1,:} \quad \text{if } i \geq 3 \text{ and } i \text{ odd}.$$

Finally, the uncertainty $\xi \in \mathbb{R}^{13}$ is distributed as

$$\xi_i \sim 6 + 2N(0,1), \quad \forall i \in [3],$$

$$\xi_i \sim U[0, 0.1], \quad \forall 4 \leq i \leq 13.$$

EC.2. Problem details for aircraft allocation problem (gbd)

We present the formulation of the aircraft allocation problem we use in our experiments. Here, $n_z = 17, d = 5$ and the number of second stage variables is 5.

$$\begin{aligned} z^*(\mathbb{Q}) \in \arg\min_z \quad & \mathbb{E}_{Y\sim\mathbb{Q}}\big[c(z;Y)\big] \\ \text{subject to} \quad & \sum_{i=1}^{5} z_i \le 10, \\ & \sum_{i=6}^{9} z_i \le 19, \\ & \sum_{i=10}^{12} z_i \le 25, \\ & \sum_{i=13}^{17} z_i \le 15, \\ & z \in \mathbb{R}^{17}_+. \end{aligned}$$

Here, the cost function, for a given uncertainty ξ, is given by

$$c(z;\xi) = c'z + \min_{y\in\mathcal{Y}(z,\xi)} f'y$$

where the set $\mathcal{Y}(z,\xi)$ is specified by $y \in \mathbb{R}^5_+$ that satisfy the following constraints:

$$\begin{aligned} & 16z_1 + 9z_{13} + y_1 \ge \xi_1, \\ & 15z_2 + 10z_6 + 5z_{10} + 11z_{14} + y_2 \ge \xi_2, \\ & 28z_3 + 14z_7 + 22z_{15} + y_3 \ge \xi_3, \\ & 23z_4 + 15z_8 + 7z_{11} + 17z_{16} + y_4 \ge \xi_4, \\ & 81z_5 + 57z_9 + 29z_{12} + 55z_{17} + y_5 \ge \xi_5, \\ & y \in \mathbb{R}^5_+. \end{aligned}$$

The vectors c, f are given as

$$\begin{aligned} c &= [18, 21, 18, 16, 10, 15, 16, 14, 9, 10, 9, 6, 17, 16, 17, 15, 10], \\ f &= [13, 13, 7, 7, 1]. \end{aligned}$$

Finally, the uncertainty ξ is generated as $\xi_i \sim \mathbb{P}_{(i)}$, where $\mathbb{P}_{(i)}$ is given by the following discrete distributions (with scenarios and their corresponding probabilities) for $i = 1, \ldots, 5$ in order:

$$[175, 185, 195, 200, 210, 220, 250, 270, 280, 290, 300, 305, 310, 312, 314],$$

$$[0.04, 0.04, 0.04, 0.04, 0.04, 0.05, 0.35, 0.1, 0.05, 0.05, 0.04, 0.04, 0.04, 0.04, 0.04],$$

$$[40, 45, 50, 55, 134, 138, 142, 146, 150, 154, 158, 160, 162],$$

$$[0.05, 0.05, 0.05, 0.05, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.05, 0.05],$$

$$[138, 140, 156, 158, 160, 162, 164, 170, 175, 180, 185, 188, 200, 205, 210, 220, 222],$$

$$[0.05, 0.05, 0.02, 0.04, 0.1, 0.02, 0.02, 0.1, 0.1, 0.1, 0.1, 0.06, 0.06, 0.04, 0.04, 0.07, 0.03],$$

$$[5, 10, 30, 37, 50, 57, 80, 85, 100, 110, 300, 320, 340, 360, 380],$$

$$[0.1, 0.1, 0.05, 0.05, 0.05, 0.05, 0.15, 0.15, 0.1, 0.1, 0.02, 0.02, 0.02, 0.02, 0.02],$$

$$[570, 580, 590, 600, 602, 604, 606, 610, 612, 614, 616, 618, 620],$$

$$[0.03, 0.04, 0.03, 0.05, 0.05, 0.1, 0.1, 0.2, 0.1, 0.1, 0.1, 0.05, 0.05].$$

EC.3. Instance-wise results for Portfolio example with $n = 10,000$

In this section, we present the instance-level results for the portfolio example with $n = 10,000$. Note that for a given instance number, we use the same training data and hence, we notice the out-of-sample gap (largely) decreases for each individual instance and method as m increases from 10 to 20.

Instance	WSR	PDSR	DPCV	DPCV-pdsr	LS	LS-pdsr	MC	IS
1	232.74	90.27	201.36	99.16	201.36	176.72	156.59	129.45
2	62.99	50.87	63.91	57.05	62.99	27.82	45.39	38.33
3	245.08	59.11	199.58	182.38	222.86	117.43	70.29	149.18
4	6784.08	2828.40	2382.58	6265.86	2187.59	4947.59	3894.22	3874.57
5	178.54	88.21	153.58	143.65	154.04	59.95	77.41	95.83
6	116.63	74.02	40.55	45.06	68.47	79.82	60.26	103.95
7	99.22	99.22	81.59	78.54	96.77	64.82	82.02	79.39
8	621.77	348.66	402.80	318.95	416.57	230.94	196.83	574.18
9	462.76	131.73	266.96	122.37	388.02	161.04	326.60	113.85
10	152.66	53.98	83.03	82.95	106.40	58.89	107.33	81.55

Table EC.1 Out-of-sample gap (in %) of various methods for $m = 10$ across instances

Instance	WSR	PDSR	DPCV	DPCV-pdsr	LS	LS-pdsr	MC	IS
1	213.51	30.79	212.13	158.23	212.13	37.26	96.08	95.44
2	47.94	25.83	35.07	40.57	43.67	24.48	27.60	20.57
3	244.89	97.94	174.22	27.84	161.84	89.12	125.85	55.65
4	1941.35	1091.88	2461.40	1150.80	1407.92	858.42	1677.66	3922.91
5	178.54	92.76	84.28	59.62	160.38	40.19	53.01	58.30
6	115.23	31.21	81.64	39.81	82.97	40.65	52.19	37.84
7	87.41	38.76	69.02	25.56	26.04	57.54	49.89	38.49
8	283.34	212.71	287.69	223.17	93.93	87.60	329.18	316.34
9	462.76	157.61	257.55	96.45	184.41	131.58	290.80	43.21
10	131.03	41.59	118.45	36.05	111.47	46.67	82.35	64.03

Table EC.2 Out-of-sample gap (in %) of various methods for $m = 20$ across instances