Submitted to *Operations Research*
manuscript (Please, provide the manuscript number!)

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

Nishanth Mundru
Kenan-Flagler Business School, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, nmundru@unc.edu

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 $m < 50$ and is comparable to for $m \geq 50$ in relation to other state of the art methods (Importance sampling, Monte Carlo sampling and Wasserstein scenario reduction with squared Euclidean norm).

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

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

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

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

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 with a smaller distribution with scenarios $\zeta^1, \ldots, \zeta^m$ and corresponding probabilities $q_1, \ldots, q_m$, for $m << n$. To be more precise, we use knowledge of the cost function and constraints while computing this reduced distribution, 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 set 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, particularly at higher noise levels and fewer number of scenarios (degrees of freedom).

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) analyze worst case bounds of scenario reduction using the Wasserstein metric, and propose methods with worst case approximation error guarantees along with an exact mixed integer optimization formulation. These methods are based on the alternating-minimization algorithm for k-means clustering (Arya et al. 2004). In this work, we propose a similar alternating-optimization method, but with a tailored objective for the stochastic optimization problem at hand. 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.

We note a stream of research that develops scenario reduction techniques for specific objectives. Rahimian et al. (2018) propose a scenario reduction method specifically for distributionally robust stochastic optimization. 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. 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.

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 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. 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 generally, our work also belongs to the area of research demonstrating the advantages of optimization over randomization. Some related work includes Bertsimas et al. (2015), where the authors consider the application of offline experimental design with covariate matching and demonstrate that using optimization to reduce mean discrepancy between groups, rather than randomization, leads to stronger inference. In particular, they cite arguments from spin-glass theory and provide computational evidence of an exponential reduction in discrepancies for groups obtained by optimization compared to randomization. In a follow up paper, Bertsimas et al. (2019) use mixed-integer optimization techniques to allocate subjects arriving sequentially into groups for clinical trials based on balancing observed covariates (features) across groups. They demonstrate via computational experiments that such an optimization-based approach achieves statistical power at least as high as, and sometimes significantly higher than, state-of-the-art covariate-adaptive randomization approaches. 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 $\operatorname{supp}(\mathbb{P})$, 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 the optimal solution $z^*(\mathbb{Q})$ and 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 4.3. 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 $\{\lambda | W'\lambda \le g\}$ 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

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

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 note that the Wasserstein distance $\mathcal{D}_W$ can be recovered as a special case of this divergence when the cost is given by $c(z;y) = \|z-y\|_2^2$, the squared Euclidean distance between z and y, and the constraint set as $\mathcal{Z} = \mathbb{R}^d$. That is,

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

To see this, we 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,

$$\begin{aligned} c\big(z^*(\zeta);\xi\big) &= \|z^*(\zeta)-\xi\|^2, \\ &= \|\zeta-\xi\|^2, \end{aligned}$$

and

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

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

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.

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

$$\begin{aligned} \mathcal{D}_S^2(\mathbb{Q}, \mathbb{P}|c, \mathcal{Z}) = \min_{\pi \in \mathbb{R}_+^{n \times m}} \quad & \frac{1}{2}\sum_{j=1}^{m}\sum_{i=1}^{n}\pi_{ij}\left(c\big(z^*(\zeta^j);\xi^i\big) - \min_{z\in\mathcal{Z}} c(z;\xi^i) + c\big(z^*(\xi^i);\zeta^j\big) - \min_{z\in\mathcal{Z}} c(z;\zeta^j)\right) \\ \text{subject to} \quad & \sum_{j=1}^{m}\pi_{ij} = p_i, \quad \forall i \in [n], \\ & \sum_{i=1}^{n}\pi_{ij} = q_j, \quad \forall j \in [m], \end{aligned} \tag{8}$$

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

Remark 1. We emphasize the equivalence of $d_S(\cdot)$ to traditional Wasserstein distance with Euclidean metric still holds when $c(z;y) = \|z-y\|^2$, as in this case,

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

Remark 2. We also point out that the traditional (Euclidean or ℓ_1) norm can also be added to $d_S(\xi, \zeta)$ as

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

for some appropriately chosen nonnegative scaling factor $\mu \geq 0$, with the resulting divergence still satisfying both Conditions 1 and 2.

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*

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

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 $|\Xi| = N$.

Theorem 1. *Under assumptions 1-4, there exist constants $\rho > 0$ and $\epsilon > 0$ such that whenever $\mathbb{Q}$ lies in the set of distributions whose support is contained in Ξ with*

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

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).

Theorem 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

$$\mathbb{C}^2(\mathbb{P}_n, m; c, \mathcal{Z}) = \min_{\mathbb{Q}} \left\{ \mathcal{D}_S^2(\mathbb{Q}, \mathbb{P}_n | c, \mathcal{Z}) \, : \mathbb{Q} \in \mathcal{P}(\mathbb{R}^d, m) \right\}. \tag{9}$$

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 (9) as a set partitioning problem.

Theorem 2. *The problem-dependent scenario reduction problem* (9) *can be written as the following problem of finding an $m-$set partition that optimizes the following problem:*

$$\begin{aligned} \mathbb{C}^2(\mathbb{P}_n, m; c, \mathcal{Z}) = \min_{I_j \in \mathcal{B}(I,m)} \quad & \frac{1}{n} \sum_{j=1}^{m} \min_{\zeta^j} \sum_{i \in I_j} \frac{1}{2} \Big(c\big(z^*(\zeta^j); \xi^i\big) - \min_{z \in \mathcal{Z}} c(z; \xi^i) + \\ & c\big(z^*(\xi^i); \zeta^j\big) - \min_{z \in \mathcal{Z}} c(z; \zeta^j) \Big). \end{aligned} \tag{10}$$

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

$$\mathbb{C}^2(\mathbb{P}_n, m; c, \mathcal{Z}) = \min_{\zeta^1, \ldots, \zeta^m} \frac{1}{n} \sum_{i=1}^{n} \min_{j \in [m]} \Big(c\big(z^*(\zeta^j); \xi^i\big) - c^*(\xi^i) + c\big(z^*(\xi^i); \zeta^j\big) - c^*(\zeta^j) \Big). \tag{11}$$

Next, we show that Problem (11) is equivalent to Problem (10). Given an optimal solution $\zeta_*^1, \ldots, \zeta_*^m$ to Problem (11), 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 (10) having the same objective as Problem (11). 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 (10), it is easy to see that these scenarios will be an optimal solution of Problem (11) with identical objective value. This completes the proof. □

Problem (10) 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 $m = 1$. Consider cost function $c(z; \xi) = \frac{1}{2} z' H z - z' v(\xi) + u(\xi)$, 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 $\mathbb{R}^d$. Also, suppose the problem is unconstrained, i.e., the constraint set $\mathcal{Z} = \mathbb{R}^{n_z}$.

Proposition 1. *When $m = 1$, solving the continuous scenario reduction problem* (9) *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 $\frac{1}{n}\sum\limits_{i=1}^{n} v(\xi^i)$ 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. □

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

$$\mathcal{Z}^*(\zeta) = \{z \in \mathcal{Z} : c(z; \zeta) \leq \min_{\hat{z} \in \mathcal{Z}} c(\hat{z}; \zeta)\}, \tag{12}$$

and modify the definition presented in (8) as

$$
\begin{aligned}
\mathcal{D}_S^2(\mathbb{Q},\mathbb{P}|c,\mathcal{Z}) = \min_{\pi\in\mathbb{R}_+^{n\times m}} \quad & \sum_{j=1}^{m}\sum_{i=1}^{n}\pi_{ij}\left[\max_{z\in\mathcal{Z}^*(\zeta^j)}\Big\{c(z;\xi^i)\Big\} - c^*(\xi^i) + c\big(z^*(\xi^i);\zeta^j\big) - c^*(\zeta^j)\right] \\
\text{subject to} \quad & \sum_{j=1}^{m}\pi_{ij} = p_i, \quad \forall i\in[n], \\
& \sum_{i=1}^{n}\pi_{ij} = q_j, \quad \forall j\in[m].
\end{aligned} \tag{13}
$$

Note that when $\mathcal{Z}^*(\zeta^j)\forall j\in[m]$ are each singleton sets, then (8) and (13) are identical. We write this equivalently as

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

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, 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 $\zeta^1,\dots,\zeta^m$. 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 $(\zeta^1,\dots,\zeta^m)$ 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 $\mathcal{D}_S^2$.

Algorithm 1 Alternating-Minimization algorithm for PDSR

```
1: procedure APPROXIMATE SOLUTION FOR PROBLEM (10).
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 $\zeta^1, \ldots, \zeta^m$, assign points $i$ to cluster $j(i)$, where
               $$j(i) \in \arg \min_{1 \leq j \leq m} \mathcal{D}_S(\xi^i, \zeta^j)$$
6:         For each $1 \leq i \leq n$, set $\pi_{i,j}^{(t)} \leftarrow 1$ if $j = j(i)$ and $\pi_{i,j}^{(t)} = 0$ else.
7:         For points in each cluster $j \in [m]$, solve for $\zeta_{(t)}^j \in \arg\min_\zeta \sum_{i=1}^n \pi_{ij}^{(t)} \mathcal{D}_S^2(\xi^i, \zeta)$.
8:         $\Delta \leftarrow \frac{1}{md} \|\zeta_{(t)} - \zeta_{(t-1)}\|$.
9:         Update $t \leftarrow t + 1$
10:    end while
11: 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

$$c(z; Y) = \max_{1 \leq t \leq k} z' A_t Y, \tag{15}$$

for known matrices $A_t, 1 \leq t \leq k$. Also, 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 (14). 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 (12) and weak duality. The result follows by considering a fixed $\hat{\alpha} \geq 0$. □

Proposition 3. *For a piecewise bilinear cost function of form* (15)*, a convex upper bound problem for estimating* ζ^j*, for a given set of points* I_j*, is given by the following problem*

$$\begin{aligned} \min_{\zeta,\lambda,\theta,\gamma} \quad & \sum_{i \in I_j} \theta_i + 2\gamma_i \\ \text{subject to} \quad & z^*(\xi^i)' A_t \zeta \leq \gamma_i \; \forall t \in [k], i \in I_j, \\ & q' \lambda^{t,i} \leq \theta_i \; \forall t \in [k], i \in I_j, \\ & P' \lambda^{t,i} \geq A_t(\xi^i - 2\zeta) \; \forall t \in [k], i \in I_j, \\ & \lambda^{t,i} \geq 0 \; \forall t \in [k], i \in I_j. \end{aligned} \tag{16}$$

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 (16) to compute the new reduced scenarios. 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 $m=1$.

Finally, we note that our approach shares some similarities and some key differences as compared to Wasserstein scenario reduction with the Euclidean norm. When the cost function is the squared loss with no constraints, then our divergence between the two distributions is exactly identical to the Wasserstein distance between them. In such a case, the scenario reduction problem is equivalent to traditional least squares clustering (Rujeerapaiboon et al. 2017), and our optimization algorithm (1) recovers distributions identical to those obtained by the $k-$means algorithm (Arya et al. 2004).

4.3. 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 (16). 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 $z^1,\dots,z^K$. 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 3. *For $m=1$, if Assumption* 5 *holds and for this cost function c, the solution to the upper bound problem* (16) *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 (2017). 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 atleast 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

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

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 (16). □

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., $z'\bar{\xi}\ge z^*(\bar{\xi})'\bar{\xi}>0\,\forall z\in\mathcal{Z}$. 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 is what we denote as $z^*(\xi)$ while defining the upper bound in Proposition 3.

4.4. 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

$$c(z;\xi) - \min_{z\in\mathcal{Z}} c(z;\xi) \leq \frac{M}{2}\|z - z^*(\xi)\|^2, \tag{17}$$

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 (9) as

$$\mathbb{C}^2(m,n;c,\mathcal{Z}) = \max_{\hat{\mathbb{P}}_n\in\mathcal{P}(\mathbb{R}^d,n)} \left\{\mathbb{C}^2(\hat{\mathbb{P}}_n,m;c,\mathcal{Z}) : \|\xi\|_2 \leq 1\,\forall\xi\in\mathrm{supp}(\hat{\mathbb{P}}_n)\right\} \tag{18}$$

Theorem 4. *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

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

The first inequality follows from (17), the following equality from our assumption as the existence of ζ that $z^*(\zeta) = \mathrm{mean}(z^*(\xi^i) : i\in I_j)$ 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). □

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. In our computational results, we compare the Prescriptive cost (for $\mathbb{Q}$ obtained by any scenario reduction method):

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

We report this metric for the test S, as the *out-of-sample* cost. For each n, we repeat this by sampling a different n training set sample, and report the average value across these experiments.

We compare Euclidean-norm based Wasserstein scenario reduction (WSR), our method of Problem-dependent scenario reduction which we denote as PDSR, Importance sampling (IS) and Monte Carlo sampling (MC). For our PDSR method, the reduced scenarios are computed using the alternating-optimization Algorithm 1 with different random restarts, similar in spirit to the k-means algorithm. We implement Wasserstein scenario reduction as the k means clustering method on the training set, using the Clustering package in Julia. 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. Finally, we implement Monte Carlo sampling where we simply sample m scenarios with replacement from the training set. We implement all algorithms in Julia (Bezanson et al. 2012), using the JuMP framework (Dunning et al. 2017) and Gurobi as the optimization solver.

5.2. 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 $n = 100, d = 5$. 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. We repeat this procedure for 100 instances (100 different training and test sets) for each m, and report the average cost incurred by each method. 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.

We present results when the noise level, or standard deviation of the noise term, $\sigma = 1.5$. We note that the difference in performance between PDSR and the other methods is statistically significant at the 0.05 significance level, by performing a Wilcoxon signed-rank test. In Figure 1, we compare the expected out-of-sample performances of the distributions produced by various scenario reduction algorithms. We see that the PDSR method outperforms all the other methods for smaller values of m, while the gap narrows as m increases to 50. This indicates the value in incorporating the cost function in the scenario computation procedure.

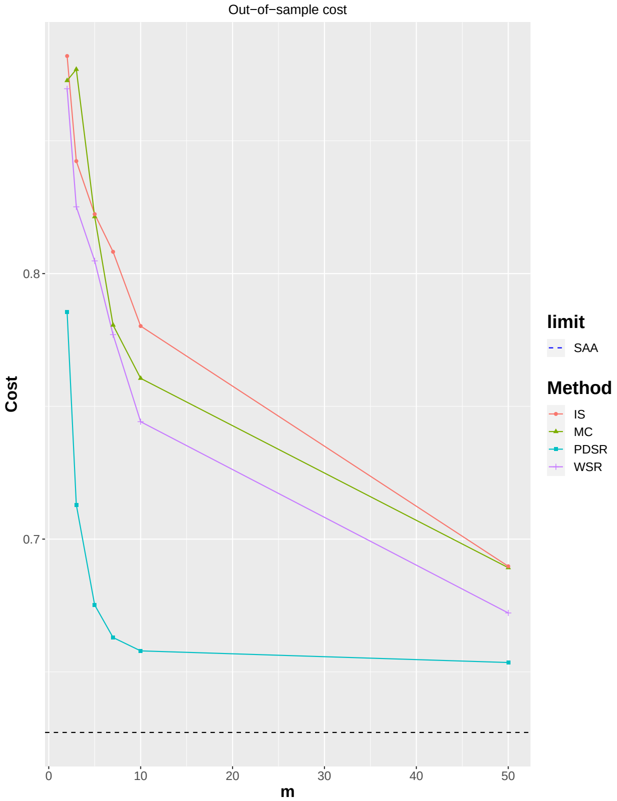
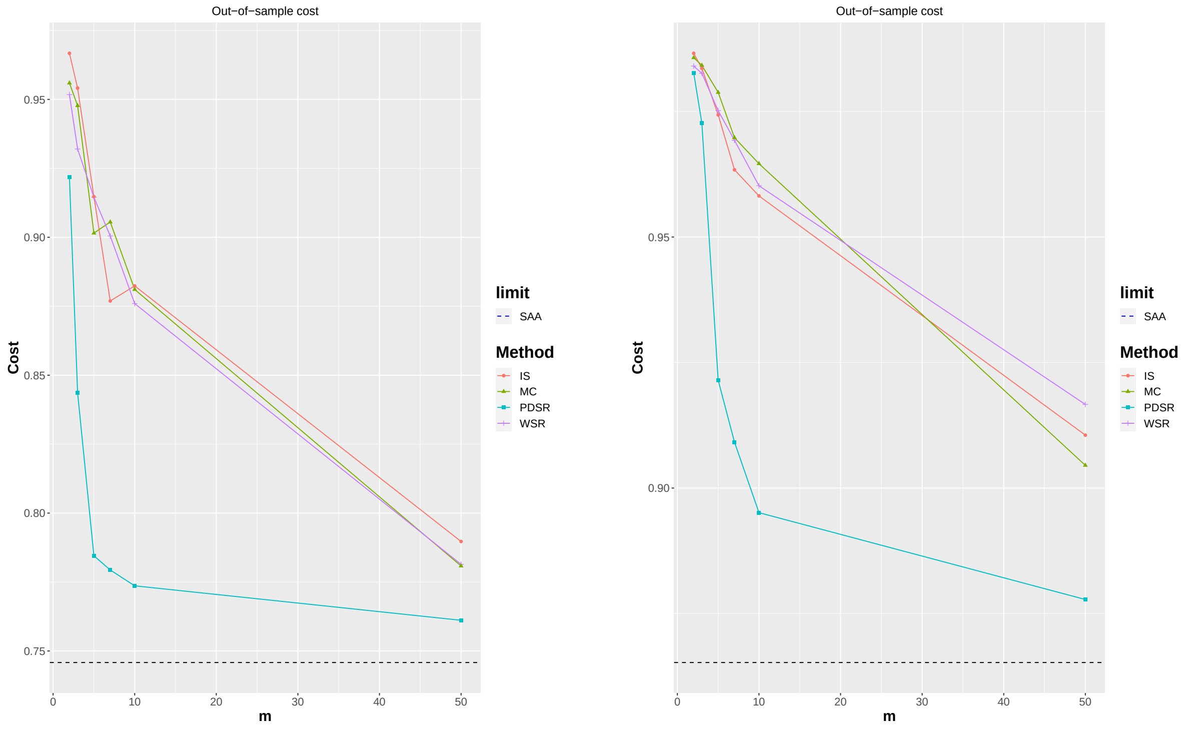


Figure 1 Average out-of-sample costs for reduced distributions generated by WSR, PDSR, MC, and IS as a function of m, the number of reduced scenarios ($\sigma = 1.5$)

As we increase the noise to $\sigma = 2.0$ and $\sigma = 4.0$ in Figures 2a and 2b respectively, we observe that PDSR is still able to maintain its performance at large m ($m = 50$), while the other methods converge slower. At the same time, note that the gap in performance is still maintained for smaller m values.



(a) $\sigma = 2.0$

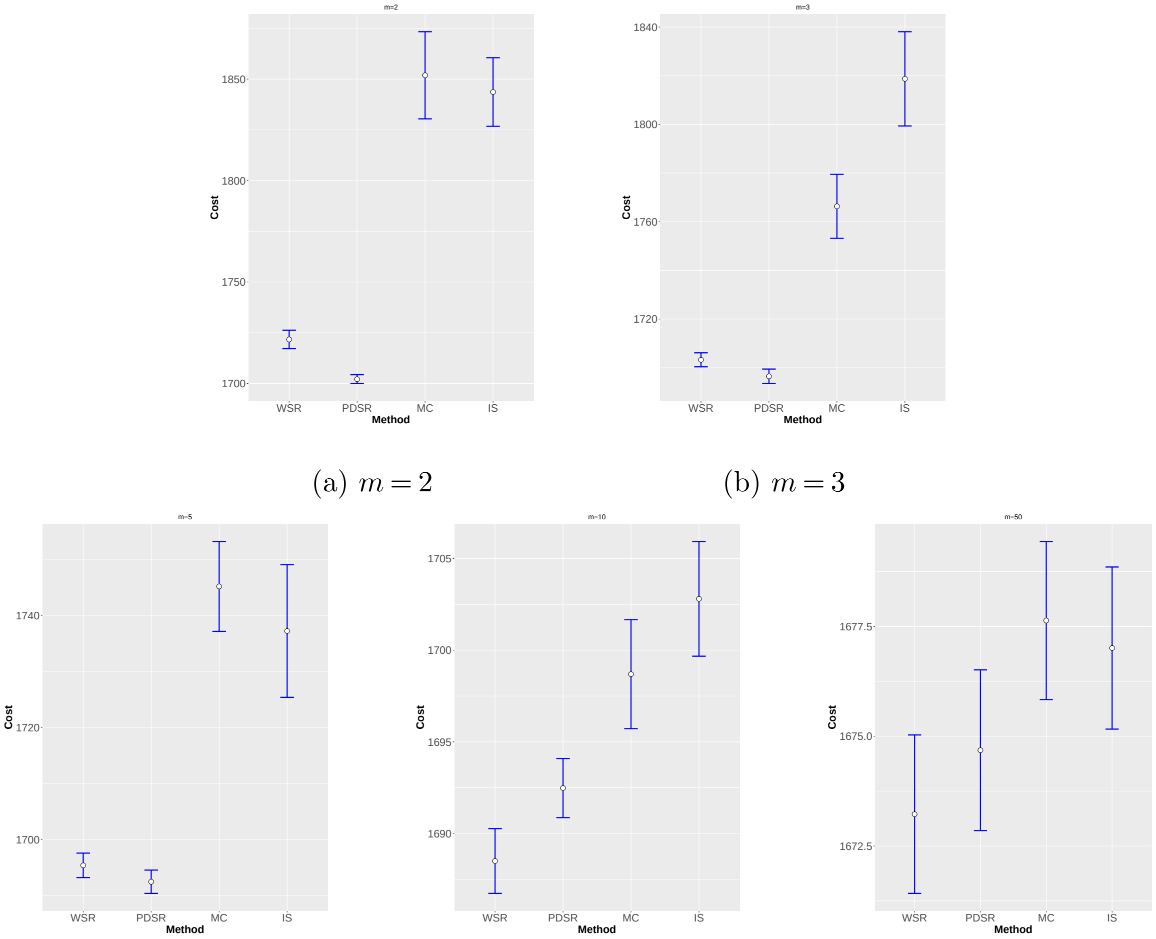
(b) $\sigma = 4.0$

Figure 2 Average out-of-sample costs for reduced distributions generated by WSR, PDSR, MC, and IS as a function of m for higher noise levels

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.

5.3. 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/`.



(a) $m = 2$

(b) $m = 3$

(c) $m = 5$

(d) $m = 10$

(e) $m = 50$

Figure 3 Average out-of-sample costs for WSR, PDSR, MC, and IS for each of $m = 2, 3, 5, 10, 50$

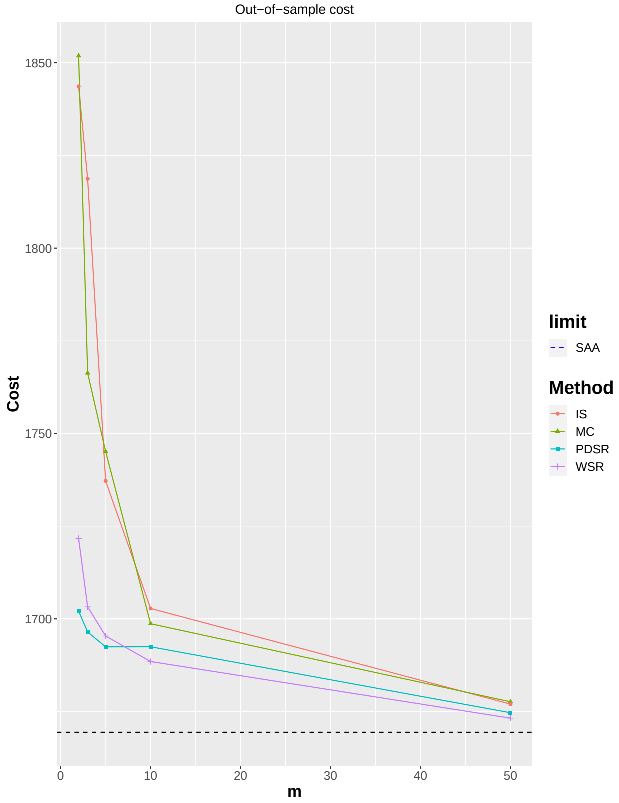
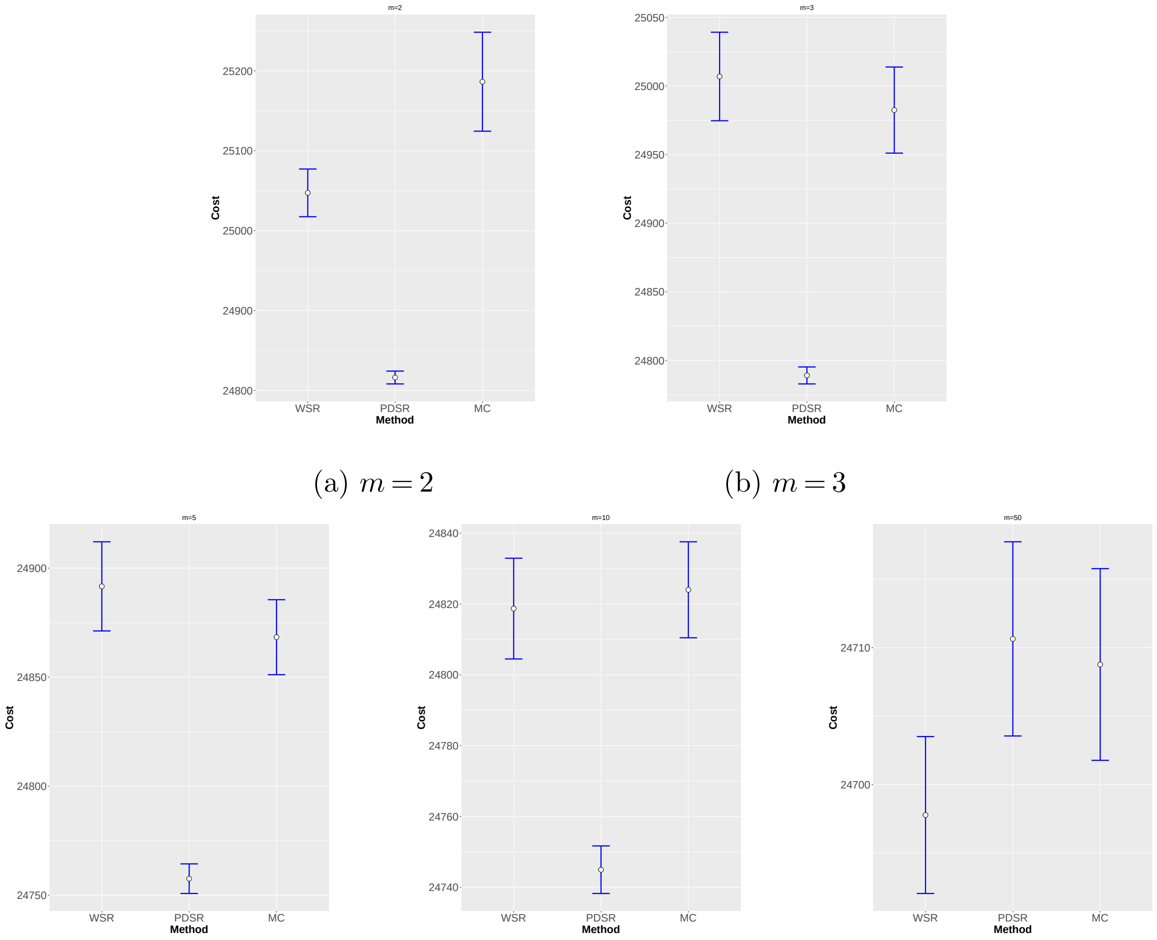


Figure 4 Average out-of-sample costs for reduced distributions generated by WSR, PDSR, MC, IS, and SAA as a function of m, the number of reduced scenarios

In Figures 3 and 4, we observe a similar pattern where the PDSR method outperforms the other methods when the number of scenarios is smaller ($m = 2, 3, 5$), while this performance gap diminishes as more scenarios are included ($m = 10, 50$).

5.4. Electric Power Expansion Planning

This test problem, denoted as APLIP in the literature, is a model of a simple power network with one demand region. There are two 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 2 first-stage decision variables denote the capacities which can be built and operated to meet demands from the three load levels. The uncertain quantities are the three demand variables, and availabilities of the two generators. There are 8 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. We use the problem data, formulation, and uncertainty distribution as presented in Infanger (1992).



(a) $m = 2$

(b) $m = 3$

(c) $m = 5$

(d) $m = 10$

(e) $m = 50$

Figure 5 Average out-of-sample costs for WSR, PDSR, and MC for each of $m = 2, 3, 5, 10, 50$

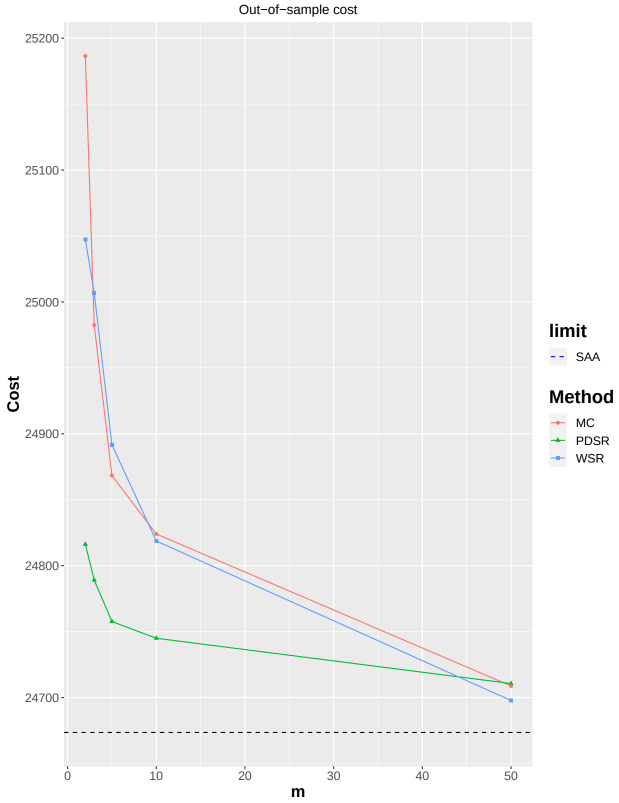


Figure 6 Average out-of-sample costs for reduced distributions generated by WSR, PDSR, and MC as a function of m, the number of reduced scenarios

In this example, we omit displaying Importance sampling (IS) in Figures 5 and 6 as it was uniformly outperformed by all the other methods. Interestingly, Wasserstein scenario reduction and Monte Carlo sampling perform very similarly, but both are outperformed by PDSR, particularly when m is small. The previously observed pattern repeats here as well, as PDSR outperforms other methods for smaller values of m ($m \in \{2, 3, 5, 10\}$), while all methods perform at a similar level at large value of m ($m = 50$).

5.5. 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 small, compared to n. It outperforms state-of-the-art standard methods for scenario reduction at the same m. However, this edge decreases with increasing m, and some times for higher m PDSR is not the best (which is the case for the APL1P example).
3. A consequence of the previous observation is that, at times, the PDSR method flattens earlier, and improvement is gradual as m becomes closer to n. However, in both the test examples (gbd, APL1P), we observe substantial improvement from $m = 10$ to $m = 50$ for all the methods. Note that this improvement for PDSR is still lesser (in absolute terms) than the improvement of the other methods from $m = 10$ to 50, simply because they have more ground to cover in order to catch up as the performances of MC, WSR, PDSR are all identical or very close to that of the SAA solution at $m = n$.
4. For the synthetic portfolio example, the increase in noise variance, σ, reduced the gap in performance between WSR and PDSR at higher values of m. The increase in noise affects the scenarios computed by the PDSR method via the cost function, as opposed to WSR where it affects the scenarios via their Euclidean norm. This could result in performance gains when the cost function is relatively stable to perturbations in the uncertainties Y.
5. For the same portfolio example, an interesting trend with the increase in noise is that the performance of all the methods becomes similar for $m = 2$, indicating that even PDSR needs at least a certain minimum degrees of freedom to shine at higher noise levels.

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. While the improvement in performance holds across different choices of the number of scenarios m, it is particularly enhanced when m is much smaller than n.

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.

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 (2017) *Smart* "predict, then optimize". *arXiv preprint arXiv:1710.08005* .

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`.

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.

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.

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.