Mathematical Programming
https://doi.org/10.1007/s10107-024-02170-4

FULL LENGTH PAPER

Series A

Mean robust optimization

Irina Wang[1] · Cole Becker[2] · Bart Van Parys[3] · Bartolomeo Stellato[1]

Received: 27 December 2023 / Accepted: 1 November 2024

Abstract

Robust optimization is a tractable and expressive technique for decision-making under uncertainty, but it can lead to overly conservative decisions when pessimistic assumptions are made on the uncertain parameters. Wasserstein distributionally robust optimization can reduce conservatism by being data-driven, but it often leads to very large problems with prohibitive solution times. We introduce mean robust optimization, a general framework that combines the best of both worlds by providing a trade-off between computational effort and conservatism. We propose uncertainty sets constructed based on clustered data rather than on observed data points directly thereby significantly reducing problem size. By varying the number of clusters, our method bridges between robust and Wasserstein distributionally robust optimization. We show finite-sample performance guarantees and explicitly control the potential additional pessimism introduced by any clustering procedure. In addition, we prove conditions for which, when the uncertainty enters linearly in the constraints, clustering does not affect the optimal solution. We illustrate the efficiency and performance preservation of our method on several numerical examples, obtaining multiple orders of magnitude speedups in solution time with little-to-no effect on the solution quality.

Keywords Robust optimization · Distributionally robust optimization · Data-driven optimization · Machine learning · Clustering · Probabilistic guarantees

Irina Wang and Bartolomeo Stellato are supported by the NSF CAREER Award ECCS 2239771.

✉ Bartolomeo Stellato
bstellato@princeton.edu

Irina Wang
iywang@princeton.edu

Cole Becker
colbeck@mit.edu

Bart Van Parys
bart.van.parys@cwi.nl

[1] Department of Operations Research and Financial Engineering, Princeton University, Princeton, USA

[2] Operations Research Center, Massachusetts Institute of Technology, Cambridge, USA

[3] Stochastics Group, Centrum Wiskunde en Informatica, Amsterdam, The Netherlands

Published online: 28 November 2024

Mathematics Subject Classification 90C17 · 90C25 · 90C46

1 Introduction

Robust optimization (RO) and distributionally robust optimization (DRO) are popular tools for decision-making under uncertainty due to their high expressiveness and versatility. The main idea of RO is to define an uncertainty set and to minimize the worst-case cost across possible uncertainty realizations in that set. However, while RO often leads to tractable formulations, it can be overly-conservative [42]. To reduce conservatism, DRO takes a probabilistic approach, by modeling the uncertainty as a random variable following a probability distribution known only to belong to an uncertainty set (also called ambiguity set) of distributions. In both RO and DRO, the choice of the uncertainty or ambiguity set can greatly influence the quality of the solution. Good-quality uncertainty sets can lead to excellent practical performance while ill chosen sets can lead to overly-conservative actions and intractable computations.

Traditional approaches design uncertainty sets based on theoretical assumptions on the uncertainty distributions [1, 3, 5, 13]. While these methods have been quite successful, they rely on a priori assumptions that are difficult to verify in practice. On the other hand, the last decade has seen an explosion in the availability of data, which has brought a shift in focus from a priori assumptions on the probability distributions to data-driven methods in operations research and decision sciences. In RO and DRO, this new paradigm has fostered data-driven methods where uncertainty sets are shaped directly from data [10]. In data-driven DRO, a popular choice of the ambiguity set is the ball of distributions whose Wasserstein distance to a nominal distribution is at most $\epsilon > 0$ [24, 28, 29, 35]. When the reference distribution is an empirical distribution, the associated Wasserstein DRO can be formulated as a convex minimization problem where the number of constraints grows linearly with the number of data-points [24]. While less conservative than RO, data-driven DRO can lead to very large formulations that are intractable, especially in mixed-integer optimization (MIO).

A common idea to reduce the dimensionality of data-driven decision-making problems is to use clustering techniques from machine learning. While clustering has recently appeared in various works within the stochastic programming literature [7, 17, 23, 34], the focus has been on the improvement of and comparisons to the sample average approximation (SAA) approach and not in a distributionally robust sense. In contrast, recent approaches in the DRO literature cluster data into partitions and either build moment-based uncertainty sets for each partition [19, 39], or enrich Wasserstein DRO formulations with partition-specific information (e.g., relative weights) [25]. While these approaches are promising, clustering is still used as a pre-processing heuristic on the data-sets in DRO, without a clear understanding of how it affects the conservatism of the optimal solutions. In particular, choosing the right clustering parameters to carefully balance computational tractability and out-of-sample performance is still an unsolved challenge.

1.1 Our contributions

We present mean robust optimization (MRO), a data-driven method that, via machine learning clustering, bridges between RO and Wasserstein DRO.

- We design the uncertainty set for RO as a ball around clustered data. Without clustering, our formulation corresponds to the finite convex reformulation in Wasserstein DRO. With one cluster, our formulation corresponds to the classical RO approach. The number of clusters is a tunable parameter that provides a tradeoff between the worst-case objective value and computational efficiency, which includes both speed and memory usage.
- We provide probabilistic guarantees of constraint satisfaction for our method, based on the quality of the clustering procedure.
- We derive bounds on the effect of clustering in case of constraints with concave and maximum-of-concave dependency on the uncertainty. In addition, we show that, when constraints are linearly affected by the uncertainty, clustering does not affect the solution nor the probabilistic guarantees.
- We show on various numerical examples that, thanks to our clustering procedure, our approach provides multiple orders of magnitude speedups over classical approaches while guaranteeing the same probability of constraint satisfaction. The code to reproduce our results is available at https://github.com/stellatogrp/mro_experiments.

1.2 Related work

Robust optimization. RO deals with decision-making problems where some of the parameters are subject to uncertainty. The idea is to restrict data perturbations to a deterministic uncertainty set, then optimize the worst-case performance across all realizations of this uncertainty. For a detailed overview of RO, we refer to the survey papers by Ben-Tal and Nemirovski [6] and Bertsimas et al. [8], as well as the books by Ben-Tal et al. [3] and Bertsimas and den Hertog [11]. These approaches, while powerful, may be overly-conservative, and there exists a tradeoff between conservatism and constraint violation [42].

Distributionally robust optimization. DRO minimizes the worst-case expected loss over a probabilistic ambiguity set characterized by certain known properties of the true data-generating distribution. Based on the type of ambiguity set, existing literature on DRO can roughly be defined in two categories. Ambiguity sets of the first type contain all distributions that satisfy certain moment constraints [21, 31, 49, 52]. In many cases such ambiguity sets possess a tractable formulation, but have also been criticized for yielding overly conservative solutions [48]. Ambiguity sets of the second type enjoy the interpretation of a ball of distributions around a nominal distribution, often the empirical distribution on the observed samples. Wasserstein uncertainty sets are one particular example [24, 28, 29, 35] and enjoy both a tractable primal as well as a tractable dual formulation. We refer to the work by Chen and Paschalidis [18] for a thorough overview of DRO, and to the work by Zhen et al. [51] for a general theory on convex dual reformulations. When the ambiguity set is well chosen, DRO formulations enjoy strong out-of-sample statistical performance guarantees. As these statistical guarantees are typically not very sharp, in practice the radius of the uncertainty set is chosen through time consuming cross-validation [28]. At the same time, DRO has the downside of being more computationally expensive than traditional robust approaches. We observe for instance that the number of constraints in Wasserstein DRO formulations scale linearly with the number of samples, which can become practically prohibitive especially when integer variables are involved. Our proposed method addresses this problem by reducing the number of constraints through clustering. While many works have recently emerged on the construction of DRO ambiguity sets through the partitioning of data [19, 25, 39], or the discretization of the underlying distribution [36], there still exists a gap in the literature. In particular, theoretical bounds on the change in problem performance as affected by the number of clusters, as well as by the quality of the cluster assignment, remain largely unexplored. In this work, we fill the gap by providing such insights.

Data-driven robust optimization. Data-driven optimization has been well-studied, with various techniques to learn the unknown data-generating distribution before formulating the uncertainty set. Bertsimas et al. [10] construct the ambiguity set as a confidence region for the unknown data-generating distribution $\mathbf{P}$ using several statistical hypothesis tests. By pairing a priori assumptions on $\mathbf{P}$ with different statistical tests, they obtain various data-driven uncertainty sets, each with its own geometric shape, computational properties, and modeling power. We, however, use machine learning in the form of clustering algorithms to preserve the geometric shape of the dataset, without explicitly learning and parametrizing the unknown distribution.

Distributionally robust optimization as a robust program. Gao and Kleywegt [29] consider a robust formulation of Wasserstein DRO similar to our mean robust optimization, but without the idea of dataset reduction. Given N samples and a positive integer K, they introduce an approximation of Wasserstein DRO by defining a new ambiguity set as a subset of the standard Wasserstein DRO set, containing all distributions supported on NK points with equal probability $1/(NK)$, as opposed to the standard set supported on N points. In this work, however, we study how to reduce the number of variables and constraints.

Robust optimization as a distributionally robust optimization program. Xu et al. [50] take inspiration from sample-based optimization problems to investigate probabilistic interpretations of RO. They generalize the ideas of Delage and Ye [21], that the solution to a robust optimization problem is the solution to a special DRO problem, where the distributional set contains all distributions whose support is contained in the uncertainty set. In a related vein, Bertsimas et al. [14] show that, under a particular construction of the uncertainty sets, multi-stage stochastic linear optimization can be interpreted as Wasserstein-∞ DRO. We establish a similar equivalence between RO and DRO, focusing especially on Wasserstein-p ambiguity sets for all p. We develop an easily interpretable construction of the primal constraints and uncertainty sets, and prove that $p = \infty$ is a limiting case of $p \geq 1$. This provides a natural extension of the equivalence proved in [14, Proposition 3].

Probabilistic guarantees in robust and distributionally optimization. Bertsimas et al. [9] propose a disciplined methodology for deriving probabilistic guarantees for solutions of robust optimization problems with specific uncertainty sets and objective functions. They derive a posteriori guarantee to compensate for the conservatism of a priori uncertainty bounds. Esfahani and Kuhn [24] obtain finite-sample guarantees for Wasserstein DRO for selecting the radius ϵ of order $N^{-1/\max\{2,m\}}$, where N is the number of samples and m is the dimension of the problem data, while Gao [28] derives finite-sample guarantees for Wasserstein DRO for selecting ϵ of order $N^{-1/2}$ under specific assumptions. We provide theoretical results of a similar vein, with a slightly increased ϵ to compensate for information lost through clustering and achieve the same probabilistic guarantees. Our theoretical guarantees hold for Wasserstein-p distance for all $p \geq 1$ and $p = \infty$, and are independent of the uncertain function to minimize. These bounds, however, following the literature, are theoretical in nature and not tight in practice, typically resulting in conservative ϵ. The final ϵ values are usually chosen through empirical experimentation - in which case, our formulation, by being lower dimensional, is overall much faster to solve.

Clustering in stochastic optimization. Clustering in stochastic optimization is closely related to the idea of *scenario reduction*. First introduced by Dupačová et al. [22], scenario reduction seeks to approximate, with respect to a probability metric, an N-point distribution with a distribution with a smaller number of points. In particular, Rujeerapaiboon et al. [43] analyze the worst-case bounds on scenario reduction the approximation error with respect to the Wasserstein metric, for initial distributions constrained to a unit ball. They provide constant-factor approximation algorithms for K-medians and K-means clustering [32]. Later, Bertsimas and Mundru [12] apply this idea to two-stage stochastic optimization problems, and provide an alternating-minimization method for finding optimal reduced scenarios under the modified objective. They also provide performance bounds on the stochastic optimization problem for different scenarios. Jacobson et al. [34], Emelogu et al. [23], Beraldi et al. [7], and Chen [17] apply a similar idea of clustering to reduce the sample/scenario size, then compare the results against the classical SAA approach where the sample size is not reduced. In MRO, we adapt and extend the scenario reduction approach to Wasserstein DRO, where upon fixing the reduced scenario points to ones found by the clustering algorithm, we allow for variation around these reduced points. We then provide performance bounds on the DRO problem depending on the number of clusters.

Data compression in data-driven problems. Fabiani and Goulart [26] compress data for robust control problems by minimizing the Wasserstein-1 distance between the original and compressed datasets, and observe a slight loss in performance in exchange for reduced computation time. While related, this is orthogonal to our approach of using machine learning clustering to reduce the dataset, where we include results and theoretical bounds for a more general set of robust optimization problems with Wasserstein-p distance, and demonstrate conditions under which no performance loss is necessary.

2 Mean robust optimization

2.1 The problem

We consider an uncertain constraint of the form,

$$g(u, x) \leq 0, \tag{1}$$

where $x \in \mathcal{X} \subseteq \mathbf{R}^n$ is the optimization variable and $\mathcal{X}$ is a compact set, $u \in \mathbf{R}^m$ is an uncertain parameter, and $-g(u, x)$ is proper, convex, and lower-semicontinuous in u for all x. Throughout this paper, we assume the support S of u to live within the domain of g for the variable u, which we will refer to as $\mathbf{dom}_u\, g$, i.e., $S \subseteq \mathbf{dom}_u\, g$. We assume $\mathbf{dom}_u\, g$ is independent of x, and that the following assumption holds.

Assumption 1 The domain $\mathbf{dom}_u\, g$ is $\mathbf{R}^m$. Otherwise, g is either element-wise monotonically increasing in u and only has a (potentially) lower-bounded domain, or element-wise monotonically decreasing in u and only has a (potentially) upper-bounded domain.

This assumption on the domain and monotonicity of g is very common in practice as it is satisfied by linear and quadratic functions, as well as other common functions (e.g., $\log(u)$, and $1/(1 + u)$). In Sect. 2.4, we extend our results for g being the maximum of of concave functions, each satisfying the aforementioned conditions.

The RO approach defines an uncertainty set $\mathcal{U} \subseteq \mathbf{R}^m$ and forms the *robust counterpart* as $g(u, x) \leq 0, \forall u \in \mathcal{U}$, where the uncertainty set is chosen so that for any solution x, the above holds with a certain probability. We define this in terms of expectation,

$$\mathbf{E}^{\mathbf{P}}(g(u, x)) \leq 0,$$

where $\mathbf{P}$ is the unknown distribution of the uncertainty u.

Risk measures. Expectation constraints can represent popular risk measures, and can imply constraints commonly used in chance-constrained programming (CCP). In CCP, the probabilistic constraint considered is $\mathbf{P}(g(u, x) \leq 0) \geq 1 - \alpha$, which corresponds to the *value at risk* being nonpositive, i.e.,

$$\mathbf{VaR}(g(u, x), \alpha) = \inf\{\gamma \mid \mathbf{P}(g(u, x) \leq \gamma) \geq 1 - \alpha\} \leq 0.$$

Unfortunately, except in very special cases, the value at risk function is intractable [46]. A tractable approximation of the value at risk is the *conditional value at risk* [40, 46], defined as $\mathbf{CVaR}(g(u, x), \alpha) = \inf_\tau\{\mathbf{E}(\tau + (1/\alpha)(g(u, x) - \tau)_+)\}$, where $(a)_+ = \max\{a, 0\}$. This expression can be modeled through our approach, by writing $\mathbf{CVaR}(g(u, x), \alpha) = \inf_\tau\{\mathbf{E}(\hat{g}(u, x, \tau))\}$, where $\hat{g}(u, x, \tau) = \tau + (1/\alpha)(g(u, x) - \tau)_+$ is the maximum of concave functions, which we study in Sects. 2.4, 6.2, and 6.3. It is well known from [46] that the relationship between these probabilistic guarantees of constraint satisfaction is

$$\mathbf{CVaR}(g(u, x), \alpha) \leq 0 \implies \mathbf{VaR}(g(u, x), \alpha) \leq 0 \iff \mathbf{P}(g(u, x) \leq 0) \geq 1 - \alpha.$$

Therefore, our expectation constraint implies common chance constraints.

Finite-sample guarantees. In data-driven optimization, while $\mathbf{P}$ is unknown, it is partially observable through a finite set of N independent samples of the random vector u. We denote this training dataset by $\mathcal{D}_N = \{d_i\}_{i \le N} \subseteq S$, and note that it is governed by $\mathbf{P}^N$, the product distribution supported on S^N. A data-driven solution of a robust optimization problem is a feasible decision $\hat{x}_N \in \mathbf{R}^n$ found using the data-driven uncertainty set $\mathcal{U}$, which in turn is constructed by the training dataset $\mathcal{D}_N$. Specifically, the feasible decision and data-driven uncertainty set $\mathcal{U}$ must imply the probabilistic guarantee

$$\mathbf{P}^N \left(\mathbf{E}^{\mathbf{P}}(g(u, \hat{x}_N)) \le 0 \right) \ge 1 - \beta, \tag{2}$$

where $\beta > 0$ is the specified probability of constraint violation. From now on, the probabilistic guarantees of constraint satisfaction refers to (2).

2.2 Our approach

To meet the probabilistic guarantees outlined above, we construct $\hat{x}_N$ to satisfy particular constraints, with respect to a particular uncertainty set.
Case $p \ge 1$. In the case where $p \ge 1$, the set we consider takes the form

$$\mathcal{U}(K, \epsilon) = \left\{ u = (v_1, \ldots, v_K) \in S^K \;\middle|\; \sum_{k=1}^{K} w_k \|v_k - \bar{d}_k\|^p \le \epsilon^p \right\},$$

where we partition $\mathcal{D}_N$ into K disjoint subsets C_k, and $\bar{d}_k$ is the centroid of the k-th subset, for $k = 1, \ldots, K$. The weight $w_k > 0$ of each subset is equivalent to the proportion of points in the subset, i.e., $w_k = |C_k|/N$. We choose p to be an integer exponent, and ϵ will be chosen depending on the other parameters to ensure satisfaction of the probability guarantee (2). When $p = 2$ and $S = \mathbf{R}^m$, the set is an ellipsoid in $\mathbf{R}^{Km}$ with the center formed by stacking together all $\bar{d}_k$ into a single vector of dimension $\mathbf{R}^{Km}$. When we additionally have $K = N$ or $K = 1$, this ellipsoid becomes a ball of dimension $\mathbf{R}^{Nm}$ or $\mathbf{R}^m$ respectively.
Case $p = \infty$. In the case where $p = \infty$, the set we consider takes becomes

$$\mathcal{U}(K, \epsilon) = \left\{ u = (v_1, \ldots, v_K) \in S^K \;\middle|\; \max_{k=1,\ldots,K} \|v_k - \bar{d}_k\| \le \epsilon \right\},$$

where the constraints for individual v_k become decoupled. This decoupling follows the result for the Wasserstein type $p = \infty$ metric [30, Equation 2], as our uncertainty set is analogous to the set of all distributions within Wasserstein-∞ distance of $\bar{d}$. Note that, if any of the decoupled constraints are violated, then $\lim_{p \to \infty} \sum_{k=1}^{K} w_k \|v_k - \bar{d}_k\|^p \ge \epsilon^p$, and the summation constraint is violated.

For both cases, when $K = 1$, we have a simple uncertainty set: $\mathcal{U}(1, \epsilon) = \left\{ v \in S \mid \|v - \bar{d}\| \le \epsilon \right\}$, a ball of radius ϵ around the empirical mean of the dataset. This is equivalent to the uncertainty set of traditional RO, as it is of the same dimension

m as the uncertain parameter. In addition, when $K = N$ and $w_k = 1/N$, both cases resemble the ambiguity sets of Wasserstein-p DRO.

Having defined the uncertainty set, we now introduce the constraints

$$\bar{g}(u, x) = \sum_{k=1}^{K} w_k g(v_k, x), \tag{3}$$

where g is defined in the original constraint (1) and w_k are the weights from above. Subsequently, $\hat{x}_N$ is the solution to the robust optimization problem

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & \bar{g}(u, x) \leq 0 \quad \forall u \in \mathcal{U}(K, \epsilon), \end{array} \tag{MRO}$$

where f is the objective function. We call this problem the mean robust optimization (MRO) problem. Given the problem data, we formulate the uncertainty set from clustered data using machine learning, with the choice of K and ϵ chosen experimentally. Then, we solve the MRO problem to arrive at a data-driven solution $\hat{x}_N$ which satisfies the probabilistic guarantee (2).

2.3 Solving the robust problem

We solve the MRO problem using a direct convex reformulation, following usual techniques for RO problems in existing literature [3, 11, 35], with adaptations made for the MRO setup, as well as a reformulation derived for the case $p = \infty$. We include simple examples for completeness.

Case $p \geq 1$. In the case where $p \geq 1$, the MRO can be rewritten as

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & \left\{ \begin{array}{ll} \underset{v_1,\ldots,v_K \in S}{\text{maximize}} & \sum_{k=1}^{K} w_k g(v_k, x) \\ \text{subject to} & \sum_{k=1}^{K} w_k \|v_k - \bar{d}_k\|^p \leq \epsilon^p \end{array} \right\} \leq 0, \end{array} \tag{4}$$

which, by dualizing the inner maximization problem, is reformulated as:

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & \sum_{k=1}^{K} w_k s_k \leq 0 \\ & [-g]^*(z_k - y_k, x) + \sigma_S(y_k) - z_k^T \bar{d}_k + \phi(q)\lambda \, \|z_k/\lambda\|_*^q + \lambda\epsilon^p \leq s_k \\ & \qquad\qquad\qquad\qquad\qquad\qquad\qquad k = 1, \ldots, K \\ & \lambda \geq 0, \end{array} \tag{5}$$

with variables $\lambda \in \mathbf{R}$, $s_k \in \mathbf{R}$, $z_k \in \mathbf{R}^m$, and $y_k \in \mathbf{R}^m$. Here, $[-g]^*(z, x) = \sup_{u \in \mathbf{dom}_u g} z^T u - [-g(u, x)]$ is the conjugate of $-g$, $\sigma_S(z) = \sup_{u \in S} z^T u$ is the support function of $S \subseteq \mathbf{R}^m$, $\| \cdot \|_*$ is the dual norm of $\| \cdot \|$, and $\phi(q) = (q-1)^{(q-1)}/q^q$ for $q > 1$ [35, Theorem 8]. Note that q satisfies $1/p + 1/q = 1$, i.e., $q = p/(p-1)$. The support function σ_S is also the conjugate of χ_S, which is defined $\chi_S(u) = 0$ if $u \in S$, and ∞ otherwise. The proof of the derivation and strong duality of the constraint is delayed to Appendix A. Since the dual of the constraint becomes a minimization problem, any feasible solution that with objective less than or equal to 0 will satisfy the constraint, so we can remove the minimization to arrive at the above form. While traditionally we take the supremum instead of maximizing, here the supremum is always achieved as we assume g to be upper-semicontinuous. For specific examples of the conjugate forms of different g, see Bertsimas and den Hertog [11, Section 2.5] and Beck [2, Chapter 4].

When K is set to be N, w_k is $1/N$, and this is of an analogous form to the convex reduction of the worst case problem for Wasserstein DRO, which we will introduce in Sect. 3.

We observe from [35, Section 2.2 Remark 1] that $\lim_{q\to\infty}\phi(q)\lambda\left\|z_k/\lambda\right\|_*^q = 0$ if $\|z_k\| \le \lambda$ and $=\infty$ otherwise. Therefore, when $p=1$ the reformulation becomes

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & \sum_{k=1}^K w_k s_k \le 0 \\ & [-g]^*(z_k - y_k, x) + \sigma_S(y_k) - z_k^T \bar{d}_k + \lambda\epsilon \le s_k \quad k=1,\dots,K \\ & \lambda \ge 0, \quad \|z_k\| \le \lambda, \quad k=1,\dots,K. \end{array} \tag{6}$$

Case $p=\infty$. In the case where $p=\infty$, the MRO can be rewritten as

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & \left\{ \begin{array}{ll} \underset{v_1,\dots,v_K \in S}{\text{maximize}} & \sum_{k=1}^K w_k g(v_k, x) \\ \text{subject to} & \|v_k - \bar{d}_k\| \le \epsilon, \quad k=1,\dots,K \end{array} \right\} \le 0, \end{array} \tag{7}$$

which has a reformulation where the constraint above is dualized,

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & \sum_{k=1}^K w_k s_k \le 0 \\ & [-g]^*(z_k - y_k, x) + \sigma_S(y_k) - z_k^T \bar{d}_k + \lambda_k\epsilon \le s_k \quad k=1,\dots,K \\ & \|z_k\|_* \le \lambda_k \quad k=1,\dots,K, \end{array} \tag{8}$$

with $s_k \in \mathbf{R}$, $z_k \in \mathbf{R}^m$, and $y_k \in \mathbf{R}^m$. The proof is delayed to Appendix B.

Remark 1 *(Case $p=\infty$ is the limit of case $p \ge 1$)* In terms of the primal problem, (7) is the limiting case of (4) as $p \to \infty$. In terms of the reformulated problem with dualized constraints, problem (8) is the limiting case of (5). The proof, delayed to Appendix C, extends the ideas stated in [14, Proposition 3].

Example with affine constraints. Consider a single affine constraint of the form

$$(a + Pu)^T x \le b, \tag{9}$$

where $a \in \mathbf{R}^n$, $P \in \mathbf{R}^{n\times m}$, and $b \in \mathbf{R}$. In other words, $g(u,x) = (a+Pu)^T x - b$, and the support set is $S = \mathbf{R}^m$. Note that, in this case, y_k must be 0 for the support function $\sigma_S(y_k)$ to be finite. We compute the conjugate as $[-g]^*(z,x) = \sup_u z^T u +$

$b-(a+Pu)^T x = a^T x - b$ if $z + P^T x = 0$, and ∞ otherwise. To substitute $\sigma_S(y_k)$ and $[-g]^*(z_k - y_k, x)$ into (5), we note that $y_k = 0$ and $z_k = -P^T x$, i.e., z_k is independent from k. By combining the K constraints in (5), we arrive at the form for general $p \geq 1$

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & a^T x - b + \phi(q)\lambda \left\| P^T x/\lambda \right\|_*^q + \lambda\epsilon^p + (P^T x)^T \sum_{k=1}^K w_k \bar{d}_k \leq 0 \\ & \lambda \geq 0, \end{array} \tag{10}$$

where the number of variables or constraints does not depend on K. Since vector $\sum_{k=1}^K w_k \bar{d}_k$ is the average of the data-points in $\mathcal{D}_N$ for any $K \in \{1, \ldots, N\}$, this formulation corresponds to always choosing $K = 1$. We also note that, for both $p = 1$ and $p = \infty$, the subsequent reformulation (11) can be viewed as the robust counterpart when the uncertainty set is a norm ball of radius ϵ centered at $(1/N)\sum_{i=1}^N d_i$. If $\bar{d} = 0$, the constraint can be simplified even further, obtaining $a^T x + \epsilon \| P^T x \|_* \leq b$, which corresponds to the robust counterpart in RO with norm uncertainty sets [11, Section 2.3], [3, Chapter 2].

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & a^T x - b + \epsilon \left\| P^T x \right\|_* + (P^T x)^T \sum_{k=1}^K w_k \bar{d}_k \leq 0. \end{array} \tag{11}$$

2.4 Maximum-of-concave constraint function

We now consider a more general maximum-of-concave function

$$g(u, x) = \max_{j \leq J} g_j(u, x),$$

with each $-g_j$ being proper, convex, and lower-semicontinuous in u for all x. When we take $J = 1$, we arrive back at the formulations given in Sect. 2. Note that any problem with multiple uncertain constraints $g_j(u, x),\ j = 1, \ldots, J$, where we assume the usual conditions on g_j, can be combined to create a joint constraint of this maximum-of-concave form. As mentioned in Sect. 2.1, this can also be used to model **CVaR** constraints, which has a maximum-of-concave analytical form. The intuitive constraint to formulate is then

$$\bar{g}(u, x) = \sum_{k=1}^K w_k \max_{j \leq J} g_j(v_k, x) = \max_{(j_1, \ldots, j_K) \in \mathcal{G}} \sum_{k=1}^K w_k g_{j_k}(v_k, x), \tag{12}$$

where in the last expression we brought the maximum outside the summation by defining the set $\mathcal{G}$ of all possible choices of $(j_1, \ldots, j_K)$. This set has size J^K, as each cluster k has J possible pieces in the maximization function. We perform this switch in order to attain a reformulation akin to (4), where we have a single inner maximization problem for the MRO constraint. As these indices are hard to express, we seek an alternative formulation. We turn to Section 4.2 of Esfahani and Kuhn [24], on the attainment of the worst-case distribution of Wasserstein DRO for a maximum-of-concave function; Wasserstein DRO is closely related to our formulations, as will be explored in Sect. 3. Adapting the ideas from Theorem 4.4 of Esfahani and Kuhn [24], we note that the worst-case constraint value for any x can in fact be attained by maximizing the function

$$\bar{g}(u, \alpha, x) = \sum_{k=1}^{K} \sum_{j=1}^{J} \alpha_{jk} g_j(v_{jk}, x) \tag{13}$$

over u in the uncertainty set, which is to be defined, and $\alpha \in \Gamma$, with $\Gamma = \{\alpha \mid \sum_{j=1}^{J} \alpha_{jk} = w_k, \alpha_{jk} \geq 0 \ \forall k, j\}$. For each constituent function g_j, the uncertainty set then contains a set of vectors $(v_{j1}, \ldots, v_{jK})$, where there exists a set of parameters $(\alpha_{j1}, \ldots, \alpha_{jK})$ to denote the fraction of mass assigned to that function and those vectors. The total amount of mass assigned for each cluster remains the weight of the cluster, i.e., $w_k = \sum_{j=1}^{J} \alpha_{jk}$. Note that choosing the j-th function as the maximum function for cluster k, i.e., the index jk in (12), is equivalent to setting $\alpha_{jk} = w_k$ and $\alpha_{j'k} = 0$ for all $j' \neq j$ in (13). We adopt this formulation instead of (12), as it easily allows us to apply the Von Neumann-Fan minimax theorem [38] in the dual reformulation (see Appendices A and B), and the existence of α instead of a maximization over g_j is useful for a proof in Sect. 4.2.

The uncertainty set is then as follows. *Case* $p \geq 1$. In the case where $p \geq 1$, we have

$$\mathcal{U}(K, \epsilon) = \left\{ u = (v_{11}, \ldots, v_{JK}) \in S^{J \times K} \;\middle|\; \exists \alpha \in \Gamma, \sum_{k=1}^{K} \sum_{j=1}^{J} \alpha_{jk} \|v_{jk} - \bar{d}_k\|^p \leq \epsilon^p \right\}.$$

Note that the single concave case given previously follows when we take $J = 1$. All parameters are defined as in the single concave case.

Case $p = \infty$. In the case where $p = \infty$, the set we consider becomes

$$\mathcal{U}(K, \epsilon) = \left\{ u = (v_{11}, \ldots, v_{JK}) \in S^{J \times K} \;\middle|\; \exists \alpha \in \Gamma, \max_{k=1,\ldots,K} \sum_{j=1}^{J} \frac{\alpha_{jk}}{w_k} \|v_{jk} - \bar{d}_k\| \leq \epsilon \right\}.$$

Following these changes, $\hat{x}_N$ is again the solution to the robust optimization problem (MRO), defined now with the generalized uncertainty set.

Solving the robust problem. We give the direct reformulation approach for solving the generalized problem for $p \geq 1$. The case $p = \infty$ is delayed to Appendix B. We write the MRO problem as the optimization problem

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & \left\{ \begin{array}{ll} \underset{v_{11},\ldots,v_{JK} \in S, \alpha \in \Gamma}{\text{maximize}} & \sum_{k=1}^{K} \sum_{j=1}^{J} \alpha_{jk} g_j(v_{jk}, x) \\ \text{subject to} & \sum_{k=1}^{K} \sum_{j=1}^{J} \alpha_{jk} \|v_{jk} - \bar{d}_k\|^p \leq \epsilon^p \end{array} \right\} \leq 0, \end{array} \tag{14}$$

and, by dualizing the inner maximization problem, arrive at the reformulation:

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & \sum_{k=1}^{K} w_k s_k \leq 0 \\ & [-g_j]^*(z_{jk} - y_{jk}, x) + \sigma_S(y_{jk}) - z_{jk}^T \bar{d}_k + \phi(q)\lambda \left\| z_{jk}/\lambda \right\|_*^q + \lambda \epsilon^p \leq s_k \\ & \qquad k = 1, \dots, K, \quad j = 1, \dots, J \\ & \lambda \geq 0, \end{array} \tag{15}$$

with variables $\lambda \in \mathbf{R}$, $s_k \in \mathbf{R}$, $z_{jk} \in \mathbf{R}^m$, and $y_{jk} \in \mathbf{R}^m$. The proof is delayed to Appendix A. In addition, when K is set to be N, and w_k's are $1/N$, this is also of an analogous form to the convex reduction of the worst case problem for Wasserstein DRO, given in Sect. 3.

3 Links to Wasserstein distributionally robust optimization

Distributionally robust optimization (DRO) solves the problem

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & \sup_{\mathbf{Q} \in \mathcal{P}_N} \mathbf{E}^{\mathbf{Q}}(g(u, x)) \leq 0, \end{array} \tag{16}$$

where the ambiguity set $\mathcal{P}_N$ contains, with high confidence, all distributions that could have generated the training samples $\mathcal{D}^N$, such that the probabilistic guarantee (2) is satisfied. Wasserstein DRO constructs $\mathcal{P}_N$ as a ball of radius ϵ with respect to the Wasserstein metric around the empirical distribution $\hat{\mathbf{P}}^N = \sum_{i=1}^{N} \delta_{d_i}/N$, where δ_{d_i} denotes the Dirac distribution concentrating unit mass at $d_i \in \mathbf{R}^m$. Specifically, we write $\mathcal{P}_N = \mathbf{B}_\epsilon^p(\hat{\mathbf{P}}^N) = \{\mathbf{Q} \in \mathcal{M}(S) \mid W_p(\hat{\mathbf{P}}^N, \mathbf{Q}) \leq \epsilon\}$, where $\mathcal{M}(S)$ is the set of probability distributions supported on S satisfying a light-tailed assumption (more details in Sect. 3.1), and

$$W_p(\mathbf{Q}, \mathbf{Q}') = \inf \left\{ \left(\int_S \|u - u'\|^p \Pi(\mathrm{d}u, \mathrm{d}u') \right)^{1/p} \right\}.$$

Here, p is any integer greater than 1, and Π is any joint distribution of u and u' with marginals $\mathbf{Q}$ and $\mathbf{Q}'$.

When $K = N$, the constraint of the DRO problem (16) is equivalent to the constraint of (MRO). In particular, for case $p \geq 1$, the dual of the constraint of (16) is equivalent to the dual of the constraint of (14), with $w_k = 1/N$ [35, 51]. Similarly, in the case where $p = \infty$, the dual of the constraint of (16) is equivalent to the dual of the constraint of (17). We can then rewrite the Wasserstein DRO problem as (14), the MRO problem, when $K = N$.

Our approach can be viewed as a form of Wasserstein DRO, with the difference that, when $K < N$, we deal with the clustered dataset. We form $\mathcal{P}_N$ as a ball around the empirical distribution $\hat{\mathbf{P}}^K$ of the centroids of our clustered data $\hat{\mathbf{P}}^K = \sum_{k=1}^{K} w_k \delta_{\bar{d}_k}$, where w_k is the proportion of data in cluster k. This formulation allows for the reduction of the sample size while preserving key properties of the sample, which translates directly to a reduction in the number of constraints and variables, while maintaining high quality solutions.

3.1 Satisfying the probabilistic guarantees

Following the parallels between MRO and Wasserstein DRO, we now show that the conditions for satisfying the probabilistic guarantees are also analogous.

Case $p \geq 1$. Wasserstein DRO satisfies (2) if the data-generating distribution, supported on a convex and closed set S, satisfies a *light-tailed assumption* [24, 27]: there exists an exponent $a > 0$ and $t > 0$ such that $A = \mathbf{E}^{\mathbf{P}}(\exp(t\|u\|^a)) = \int_S \exp(t\|u\|^a)\mathbf{P}(du) < \infty$. We refer to the following theorem.

Theorem 1 (Measure concentration [27, Theorem 2]) *If the light-tailed assumption holds, we have* $\mathbf{P}^N(W_p(\mathbf{P}, \hat{\mathbf{P}}^N) \geq \epsilon) \leq \phi(p, N, \epsilon)$*, where* ϕ *is an exponentially decaying function of* N.

Theorem (1) estimates the probability that the unknown data-generating distribution $\mathbf{P}$ lies outside the Wasserstein ball $\mathbf{B}_\epsilon^p(\hat{\mathbf{P}}^N)$, which is our ambiguity set. Thus, we can estimate the smallest radius ϵ such that the Wasserstein ball contains the true distribution with probability $1 - \beta$, for some target $\beta \in (0, 1)$. We equate the right-hand-side to β, and solve for $\epsilon_N(\beta)$ that provides us the desired guarantees for Wasserstein DRO [24, Theorem 3.5].

Case $p = \infty$. When $p = \infty$, Bertsimas et al. [14, Section 6] note that the light-tailed assumption is no longer sufficient. Wasserstein DRO satisfies (2) under stronger assumptions, as given in the following theorem.

Theorem 2 (Measure concentration, $\mathbf{p} = \infty$ [45, Theorem 1.1]) *Let the support* $S \subset \mathbf{R}^m$ *of the data-generating distribution be a bounded, connected, open set with Lipschitz boundary. Let* $\mathbf{P}$ *be a probability measure on* S *with density* $\rho : S \to (0, \infty)$*, such that there exists* $\lambda \geq 1$ *for which* $1/\lambda \leq \rho(x) \leq \lambda, \ \forall x \in S$*. Then,* $\mathbf{P}^N(W_\infty(\mathbf{P}, \hat{\mathbf{P}}^N) \geq \epsilon) \leq \phi(N, \epsilon)$*, where* ϕ *is an exponentially decaying function of* N.

We can again equate the right-hand-side to β and find $\epsilon_N(\beta)$. We extend this result to the clustered set in MRO.

Theorem 3 (MRO finite sample guarantee) *Assume the light-tailed assumption holds when* $p \geq 1$*, and the corresponding assumptions hold when* $p = \infty$*. If* $\beta \in (0, 1)$*,* $\eta_N(K)$ *is the average* p*-th powered distance of data-points in* $\mathcal{D}_N$ *from their assigned cluster centers, and* $\hat{x}_N$ *is the optimal solution to* (MRO) *with uncertainty set* $\mathcal{U}(K, \epsilon_N(\beta) + \eta_N(K)^{1/p})$*, then the finite sample guarantee* (2) *holds.*

Proof Compared with Wasserstein DRO, MRO has to account for the additional difference between the two empirical distributions $\hat{\mathbf{P}}^N$ and $\hat{\mathbf{P}}^K$. If we introduce a new parameter, $\eta_N(K)$, defined as

$$\eta_N(K) = \frac{1}{N}\sum_{i=1}^{K}\sum_{i\in C_k} \|d_i - \bar{d}_k\|^p$$

the average p-powered distance with respect to the norm used in the Wasserstein metric, of all data-points in $\mathcal{D}_N$ from their assigned cluster centers $\bar{d}_k$, we notice that

$$\begin{aligned} W_p(\hat{\mathbf{P}}^K, \hat{\mathbf{P}}^N)^p &= \inf_{\Pi}\left\{\int_S \|u-u'\|^p\,\Pi(\mathrm{d}u,\mathrm{d}u')\right\} \quad (\Pi \text{ any joint dist. of } \hat{\mathbf{P}}^K, \hat{\mathbf{P}}^N\ )\\ &\le \sum_{i=1}^{K}\frac{|C_k|}{N}\int_S \|u-\bar{d}_k\|^p\,\hat{\mathbf{P}}^N(u|u'=\bar{d}_k)(\mathrm{d}u)\\ &\le \sum_{i=1}^{K}\frac{|C_k|}{N}\frac{1}{|C_k|}\sum_{i\in C_k}\|d_i-\bar{d}_k\|^p = \eta_N(K), \end{aligned}$$

where we have replaced the integral with a finite sum, as the distributions are discrete. Therefore, by Theorems 1, 2 and the triangle inequality [20],

$$W_p(\mathbf{P}, \hat{\mathbf{P}}^K) \le W_p(\mathbf{P}, \hat{\mathbf{P}}^N) + W_p(\hat{\mathbf{P}}^K, \hat{\mathbf{P}}^N) \le \epsilon_N(\beta) + \eta_N(K)^{1/p},$$

with probability at least $1-\beta$. We thus have

$$\mathbf{P}(\mathbf{P} \in \mathbf{B}^p_{\epsilon_N(\beta)+\eta_N(K)^{1/p}}(\hat{\mathbf{P}}^K)) \ge 1-\beta,$$

which implies $\mathcal{U}(K, \epsilon_N(\beta) + \eta_N(K)^{1/p})$ contains all possible realizations of uncertainty with probability $1-\beta$, so the finite sample guarantee (2) holds. □

4 Worst-case value of the uncertain constraint

In the previous section, we proposed a theoretical increase in ϵ to maintain the same finite sample guarantee before and after clustering. However, a question remains: what is the extent of the effects of clustering if we *don't* increase ϵ? In this section, we thus approach the analysis in a different manner: keeping ϵ constant, we quantify the change in the *worst-case value of the constraint function* that arises from clustering. In fact, for select cases, our results suggest there is no need to increase ϵ after clustering; under specific curvature conditions on the constraint function g, we may obtain a more conservative, or even unchanged solution after clustering, in which case the original finite sample guarantee is retained.

We begin with a remark on the clustering value attained. The MRO approach is closely centered around the concept of clustering to reduce sample size while maintaining sample diversity. We wish to cluster points that are close together, such that the objective is only minimally affected. With this goal, we would like to minimize the average distance of the points in each cluster to their data-center,

$$D(K)^\star = \text{minimize } D(K) = \text{minimize } \frac{1}{N}\sum_{k=1}^{K}\sum_{d_i \in C_k} \|d_i - \bar{d}_k\|_2^2,$$

where $\bar{d}_k$ is the mean of the points in cluster C_k. While the best performance is attained with $D(K)^\star$, in practice we work with the approximation $D(K)$, where C_k is decided by a clustering algorithm. This value upper bounds $D(K)^\star$. A well-known algorithm is K-means [32], where we create K clusters by iteratively solving a least-squares problem. From here on we use only $D(K)$, and note that for the case $p = 2$, we have $D(K) = \eta_N(K)$ from Theorem 3.

In this section, we then show the effects of clustering on the worst-case value of the constraint function in (MRO). We prove two sets of results, corresponding to g given as a single concave function, and as a more general maximum-of-concave function, which includes the maximum-of-affine function.

4.1 Single concave function

Quantifying the clustering effect. We calculate the difference between the worst-case value of the constraint in (MRO), for different K,

$$\begin{aligned} \bar{g}^K(x) = \underset{u_1,\dots,u_K \in S}{\text{maximize}} \quad & \sum_{k=1}^{K} \frac{|C_k|}{N} g(u_k, x) \\ \text{subject to} \quad & \sum_{k=1}^{K} \frac{|C_k|}{N} \|u_k - \bar{d}_k\|^p \le \epsilon^p. \end{aligned} \tag{MRO-K}$$

When $K = N$, $\bar{g}^N(x)$ is akin to the constraint value for traditional Wasserstein DRO. We also denote by $\bar{g}^{N*}(x)$ the value of the constraint value without the support constraints $u_1, \dots, u_N \in S$. From here on, when we mention that the support *affects the worst-case constraint value*, we refer to situations where at least one of the constraints $u_i \in S$ for $i = 1, \dots, N$ is binding. Formally, the definition is $\bar{g}^N(x) \neq \bar{g}^{N*}(x)$ for any x feasible for the DRO problem. We note a sufficient but not necessary condition for the support to not affect the worst-case constraint value: the situation in which the support doesn't affect the uncertainty set, which is defined as

$$\begin{aligned} &\left\{u \in \mathbf{R}^{N\times m} : (1/N)\sum_{i=1}^{N} \|u_i - d_i\|^p \le \epsilon^p\right\} \\ = &\left\{u \in S^{N\times m} : (1/N)\sum_{i=1}^{N} \|u_i - d_i\|^p \le \epsilon^p\right\}. \end{aligned}$$

If the support satisfies this condition, we can conclude that $\bar{g}^N(x) = \bar{g}^{N*}(x)$ for any x feasible for the DRO problem, and obtain improved bounds below. While the condition depends on the location of the data-points, it is acceptable, as this is a condition we can check given data to potentially improve the following bounds, without having to solve the MRO problem. We also define the L-smooth condition, which is needed in the subsequent theorems.

Definition 1 *(L-smoothness)* A differentiable function $g(u, x)$ is L-smooth on its domain, with constant L, with respect to the ℓ_2-norm and for a given x, if

$$\|\nabla g(v, x) - \nabla g(u, x)\|_2 \leq L\|u - v\|_2, \quad \forall u, v \in \mathbf{dom}_u\, g.$$

With these definitions, we can prove the following relations.

Theorem 4 *With the same x and ϵ, and for any integer $p \geq 1$, we always have*

$$\bar{g}^N(x) \leq \bar{g}^K(x).$$

Suppose that Assumption 1 holds, and $-g$ is L-smooth according to Definition 1. Then, with the same x and ϵ, and for any integer $p \geq 1$, we always have

$$\bar{g}^K(x) \leq \bar{g}^{N*}(x) + (L/2)D(K).$$

The proof is delayed to Appendix D. The results also hold for $p = \infty$, as we have shown in Remark 1 that the case $p = \infty$ is the limit of the case $p \geq 1$, and these results hold under the limit.

Let Δ be the maximum difference in constraint value resultant from relaxing the support constraint on the MRO uncertainty sets, subject to x being feasible for problem (MRO), $\Delta = \max_{x \in \mathcal{X}}(\bar{g}^{N*}(x) - \bar{g}^N(x))$. As we assume Assumption 1 to hold, combined with the smoothness of g, we note that when solving for $\bar{g}^{N*}(x)$, the chosen v_i values without the support constraint will still remain in the domain $\mathbf{dom}_u\, g$. Refer to a similar argument in Appendix D (ii) for details. The function $\bar{g}^{N*}(x) - \bar{g}^N(x)$ is then continuous in x and everywhere defined for $x \in \mathcal{X}$, thus maximizing with respect to $\mathcal{X}$, a compact set, the value Δ is finite. Then, we observe that $\bar{g}^K(x) - \bar{g}^N(x) \leq \Delta + (L/2)D(K)$ for all such x, so the smaller the $D(K)$, (i.e., higher-quality clustering procedure), the smaller the increase in the worst-case constraint value. In addition, the value Δ is independent of K, as it only depends on $\bar{g}^{N*}(x)$ and $\bar{g}^N(x)$.

Remark 2 While Δ could be constructed to be arbitrarily bad, in practice, we expect our relevant range of ϵ to be small enough such that the difference is insignificant. We can approximate $\Delta \approx 0$ and use the upper bound $(L/2)D(K)$, as this bound is often not tight. See Sects. 6.3 and 6.1 for examples.

Uncertain objective When the uncertainty is in the objective, Theorem 4 quantifies the difference in optimal values.

Corollary 1 *Consider the problem where g is itself the objective function we would like to minimize and $X \subseteq \mathbf{R}^n$ represents the constraints, which are deterministic. Then, $(L/2)D(K) + \Delta$ upper bounds the difference in optimal values of the MRO problem with K and N clusters.*

Uncertain constraints. When the uncertainty is in the constraints, the difference between $\bar{g}^K(x)$ and $\bar{g}^K(x)$ no longer directly reflects the difference in optimal values. Instead, clustering creates a restriction on the feasible set for x as follows. For the same $\hat{x}$, $\bar{g}^K(\hat{x})$ takes a greater value than $\bar{g}^N(\hat{x})$. Since both of them are constrained to be nonpositive from (MRO), the feasible region with K clusters is smaller.
Affine dependence on uncertainty. As a special case, when g is affine in u, $L = 0$, so we observe the following corollary.

Corollary 2 (Clustering with affine dependence on the uncertainty) *If $g(u, x)$ is affine in u and the worst-case constraint value is not affected by the support constraint, then clustering makes no difference to the optimal value and optimal solution to* (MRO).

4.2 Maximum-of-concave functions

We now consider the more general case of a maximum-of-concave constraint function, $g(u, x) = \max_{j \le J} g_j(u, x)$, subject to a polyhedral support, $S = \{u \mid Hu \le h\}$. For $p \ge 1$, we make use of the dual of the optimization problem in the constraint of (14), defined for various K,

$$\begin{array}{rll} \bar{g}^K(x) = & \underset{\lambda \ge 0, \gamma \ge 0, z, s}{\text{minimize}} & \sum_k^K (|C_k|/N) s_k \\ & \text{subject to} & [-g_j]^*(z_{jk} - H^T\gamma_{jk}) + \gamma_{jk}^T(h - H\bar{d}_k) - z_{jk}^T\bar{d}_k + \lambda\epsilon^p \\ & & + \phi(q)\lambda \left\| z_{jk}/\lambda \right\|_*^q \le s_k, \quad k = 1, \dots, K, \quad j = 1, \dots, J, \end{array} \tag{MRO-K-Dual}$$

where the variables y_{jk} from (15) are replaced by $H^T\gamma_{jk}$, with $\gamma_{jk} \ge 0$, due to the specific form of the polyhedral support. Similarly, for $p = \infty$, we define

$$\begin{array}{rll} \bar{g}^K(x) = & \underset{\lambda \ge 0, \gamma \ge 0, z, s}{\text{minimize}} & \sum_k^K (|C_k|/N) s_k \\ & \text{subject to} & [-g_j]^*(z_{jk} - H^T\gamma_{jk}) + \gamma_{jk}^T(h - H\bar{d}_k) - z_{jk}^T\bar{d}_k \\ & & + \lambda_k\epsilon \le s_k, k = 1, \dots, K, \quad j = 1, \dots, J, \\ & & \left\| z_{jk} \right\|_* \le \lambda_k, \quad k = 1, \dots, K, \quad j = 1, \dots, J. \end{array} \tag{MRO-K-Dual-∞}$$

The following theorems hold for both $p \ge 1$ and $p = \infty$.

Theorem 5 *When g is the maximum of concave functions with domain $\mathbf{dom}_u\, g_j = \mathbf{R}^m$ and polyhedral support $S = \{u \mid Hu \le h\}$, and where each $-g_j$ is L_j-smooth* *according to Definition 1, we have, for the same x and ϵ,*

$$\bar{g}^N(x) - \delta(K, z, \gamma) \leq \bar{g}^K(x) \leq \bar{g}^{N*}(x) + \max_{j \leq J}(L_j/2)D(K),$$

where $\delta(K, z, \gamma) = (1/N)\sum_{k=1}^{K}\sum_{i \in C_k} \max_{j \leq J}((-z_{jk} - H^T\gamma_{jk})^T(d_i - \bar{d}_k))$, and z, γ are the dual variables. $\bar{g}^{N*}(x)$ is the problem without support constraints.

The proof is delayed to Appendix E. Due to the nonconvex and nonconcave nature of maximum-of-concave functions, the lower bound now involves an extra term $\delta(K, z, \gamma)$. However, when g is a maximum-of-affine function, which is convex, we know $\bar{g}^{N*}(x)$ to be an upper bound on $\bar{g}^K(x)$.

Corollary 3 *When g is the maximum of affine functions with domain $\mathbf{dom}_u\, g_j = \mathbf{R}^m$ and polyhedral support $S = \{u \mid Hu \leq h\}$, for the same x and ϵ,*

$$\bar{g}^N(x) - \delta(K, z, \gamma) \leq \bar{g}^K(x) \leq \bar{g}^{N*}(x)$$

This follows from the fact that $L_j = 0$ for all affine functions g_j.

Uncertain objective When the uncertainty is in the objective, Theorem 5 and Corollary 3 quantifies the possible difference in optimal values between K and N clusters. We let $\Delta = \max_x \left(\bar{g}^{N*}(x) - \bar{g}^N(x)\right)$, subject to x being feasible for problem (MRO). Note that this is only needed for the upper bound.

Corollary 4 *Consider the problem where g is itself the objective function we would like to minimize and $X \subseteq \mathbf{R}^n$ represents the constraints, which are deterministic. Then, $\delta(K, z, \gamma)$ upper bounds the possible decrease in optimal values of the MRO problem with K clusters compared with that of N clusters. Similarly, $\max_{j \leq J}(L_j/2)D(K) + \Delta$ upper bounds the possible increase.*

Uncertain constraints When the uncertainty is in the constraints, the difference between $\bar{g}^N(x)$ and $\bar{g}^K(x)$ as given in Theorem 5 no longer directly reflect the difference in optimal values. Instead, clustering affects the feasible set for x as follows. For any $\hat{x}$, in the case $\bar{g}^N(\hat{x}) \geq \bar{g}^K(\hat{x})$, $\bar{g}^K(\hat{x})$ can be at most $\delta(K, z, \gamma)$ lower in value than $\bar{g}^N(\hat{x})$. Since both values are constrained to be nonpositive from (MRO), the feasible region of the MRO problem with K clusters may be less restricted than that of N clusters. This indirectly allows MRO with K clusters to obtain a smaller optimal value. On the other hand, in the case $\bar{g}^N(\hat{x}) \leq \bar{g}^K(\hat{x})$, $\bar{g}^K(\hat{x})$ can be at most $\max_{j \leq J}(L_j/2)D(K) + \Delta$ higher in value than $\bar{g}^N(\hat{x})$. This indirectly lets MRO with K clusters to obtain a larger optimal value.

5 Parameter selection and outliers

Choosing K. When the uncertain constraint is affine and S does not affect the worst-case constraint value, the number of clusters K does not affect the final solution, so it is always best to choose $K = 1$. When S affects the worst-case constraint value, there is a difference of at most Δ between setting $K = 1$ and $K = N$, which can often be approximated ≈ 0 for small ϵ. Therefore, setting $K = 1$ remains the recommendation. When the constraint is concave, we choose K to obtain a reasonable upper bound on $\bar{g}^K(x)$, as described in Theorem 4. This upper bound depends linearly on $D(K)$, the clustering value, so by choosing the *elbow* of the plot of $D(K)$, we choose a cluster number that, while being a reasonably low value, best conforms to the shape of the underlying distribution. When the constraint is maximum-of-concave, the bounds given in Theorem 5 are also related to $D(K)$. The upper bound is related in the same manner as above. For the lower bound, note that we wish for $\delta(K, z, \gamma)$, which is nonnegative from Appendix E, to take lower values. By definition, $\delta(K, z, \gamma)$ is linearly dependent on the differences $d_i - \bar{d}_k$, which are smaller when $D(K)$ takes lower values. Therefore, while γ and z are unknown, we can attempt to lower the possible value of $\delta(K, z, \gamma)$ by choosing a reasonably low $D(K)$. The elbow method has been commonly used in machine learning problems pertaining the choice of hyperparameters, especially for K-means, and can be traced back to Thorndike [44] in 1953. Note that, by directly returning $D(K)$ and examining the elbow as an initial step, this procedure can be completed in the clustering step without having to solve the downstream optimization problem. To further improve the choice of K, or if the elbow is unclear, cross-validation may be used for low K values or K values around the elbow. No matter if the uncertainty lies in the objective or the constraints, this bound will inform us of the potential difference between different K.

Choosing ϵ While we have outlined theoretical results in Theorem 3 for choosing ϵ, in practice, we experimentally select ϵ through cross validation to arrive at the desired guarantee. Therefore, while the theoretical bounds suggest to choose a larger ϵ when we cluster, this may not be the case experimentally. In fact, for concave g, we may even choose a smaller ϵ, due to the increase in the level of conservatism for small K. On the other hand, for maximum-of-concave g, there is the possibility of needing a larger ϵ, as smaller K may lead to less conservative solutions. However, for both cases, we show a powerful result in the upcoming numerical examples: although for the same ϵ, MRO with K clusters differs in conservatism from Wasserstein DRO (N clusters), there are cases where we can tune ϵ such that MRO and DRO provide almost identical tradeoffs between objective values and probabilistic guarantees, such that no loss in performance results from choosing a smaller cluster number K.

Data with outliers When the provided dataset contains outliers, one might imagine that the centroids created by the clustering algorithm will be biased towards the outliers. While this is true, the weights of the outliers will not increase through clustering, thus the effect of outliers on these clustered Wasserstein balls is not worse than their effect on the original Wasserstein balls, which include the Wasserstein ball around the outlier point. In fact, by clustering the outlier point with other points, MRO offers protection against the outlier. We demonstrate this in on the numerical experiment in Sect. 6.4, where we compare three methods: MRO, MRO with outlier removal, and MRO with the outlier considered as its own cluster.

6 Numerical examples

We now illustrate the computational performance and robustness of the proposed method on various numerical examples. All the code to reproduce our experiments is available, in Python, at

https://github.com/stellatogrp/mro_experiments.

We run the experiments on the Princeton Institute for Computational Science and Engineering (PICSciE) facility with 20 parallel 2.4 GHz Skylake cores. We solve all optimization problems with MOSEK [37] optimizer with default settings. In Sect. 6.1, we demonstrate the performance of MRO when the uncertain constraint is concave. In Sect. 6.2, 6.3, and 6.4, we demonstrate the performance of MRO for maximum-of-affine uncertainty.

The calculated in-sample objective value and out-of-sample expected values, as well as the out-of-sample probability of constraint violation, are averaged over 50 independent runs of each experiment. For each run, we generate evaluation data of the same size N as the training dataset. For numerical examples with an uncertain objective, the probability of constraint violation is measured as the probability the average out-of-sample value is above the in-sample value. For numerical examples with an uncertain constraint, the probability of constraint violation is measured as the probability the average constraint value is above zero. For all experiments, we plot the following.

1. In-sample objective values and out-of-sample expected values vs. ϵ for different K. We use solid lines for the in-sample objective value, and dotted lines for the out-of-sample expected value.
2. Objective value vs. β for different K; each point represents the solution for the ϵ achieving the smallest objective value. Starred K values indicate the formulation without support constraints.
3. The difference in the value of the uncertain objective between using K and N clusters, compared with the theoretical upper bound from Theorems/Corollaries 4, 5, 3. We use solid lines for the actual difference, and dotted lines for the upper bounds.
4. Solve time for select K and ϵ values.

We also plot the clustering value $D(K)$ over K, and use the elbow method to suggest an a priori K value to perform cross-validation around.

6.1 Capital budgeting

We consider the capital budgeting problem in [4, Section 4.2], where we select a portfolio of investment projects maximizing the total net present value (NPV) of the portfolio, while the weighted sum of the projects is less than a total budget θ. The NPV for all projects is $\eta(u) \in \mathbf{R}^n$, where for each project j, $\eta_j(u)$ is the sum of discounted cash flows F_{jt} over the years $t = 0, \dots, T$, i.e., $\eta_j(u) = \sum_{t=0}^{T} F_{jt}/(1+u_j)^t$. Here, $u_j \in \mathbf{R}_+$ is the discount rate of project j. We formulate the uncertain function to be minimized as

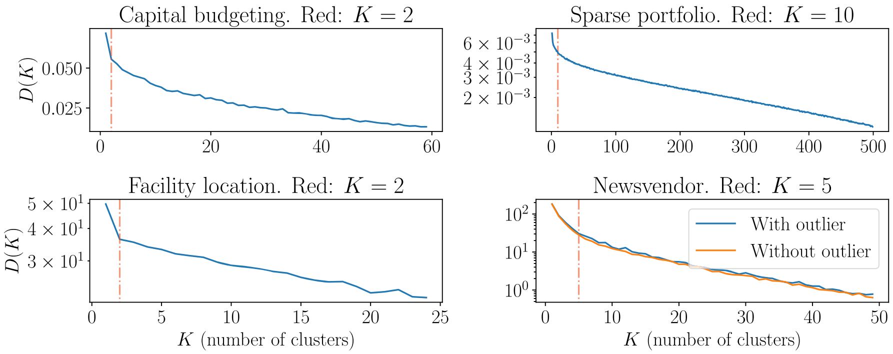


Fig. 1 $D(K)$ vs. K. for all experiments. The red lines are the suggested values

$$g(u, x) = -\eta(u)^T x,$$

where $x = (x_1, \ldots, x_n) \in \{0, 1\}^n$ is the indicator for selecting each project. The discount rate u_j is subject to uncertainty, as it depends on several factors, such as the interest rate of the country where project j is located and the level of return the decision-maker wants to compensate the risk. The function g is concave and monotonically increasing in u, and we can define a domain $u \geq 0$ so that Assumption 1 and Theorem 4 applies. The robust problem becomes

$$\begin{array}{ll} \underset{x,t}{\text{minimize}} & \tau \\ \text{subject to} & \bar{g}(u, x) \leq \tau, \quad u \in \mathcal{U}(K, \epsilon) \\ & h^T x \leq \theta, \quad x \in \{0, 1\}, \end{array}$$

where $h \in \mathbf{R}^n$ is the vector of project weights. We solve the convex reformulation obtained through applying (5), for $p = 2$, which gives rise to a number of power-cone constraints proportional to the number of clusters.

Problem setup. We set $n = 20$, $N = 120$, $T = 5$. We generate F_{jt} from a uniform distribution on $[0.1, 0.5 + 0.004t]$ for $j = 1, \ldots, n, \quad t = 0, \ldots, T$. For all j, h_j is generated from a uniform distribution on $[1, 3 - 0.5j]$, and the total budget θ is set to be 12. We generate uncertain data from two slightly different uniform distributions, to simulate two different sets of predictions on the discount rates. The first half is generated on $[0.005j, 0.02j]$, and the other half on $[0.01j, 0.025j]$, for all j. We calculate an upper bound on the L-smooth parameter, $L = \|\nabla^2 \sum_{j=1}^{n} \sum_{t=0}^{T} F_{jt}(\hat{x}_N)_j (1+u_j)^{-t}\|_{2,2} \leq \| \sum_{j=1}^{n} \sum_{t=0}^{T} t(t+1) F_{jt}(\hat{x}_N)_j \|_{2,2}$ for each data-driven solution $\hat{x}_N$.

Results We observe in Fig. 2 that using two clusters is enough to achieve performance almost identical to that of using 120 clusters. Although from the left image, we see that $K = 2$ slightly upper bounds $K = 120$, from the right, their tradeoffs between the objective value and relevant constraint violation probability ($\beta \leq 0.2$) are largely the same, so we can always tune ϵ to achieve the same performance and guarantees. Notice that the results for $K = 120$ and $K = 120^*$ are near identical for small ϵ, where

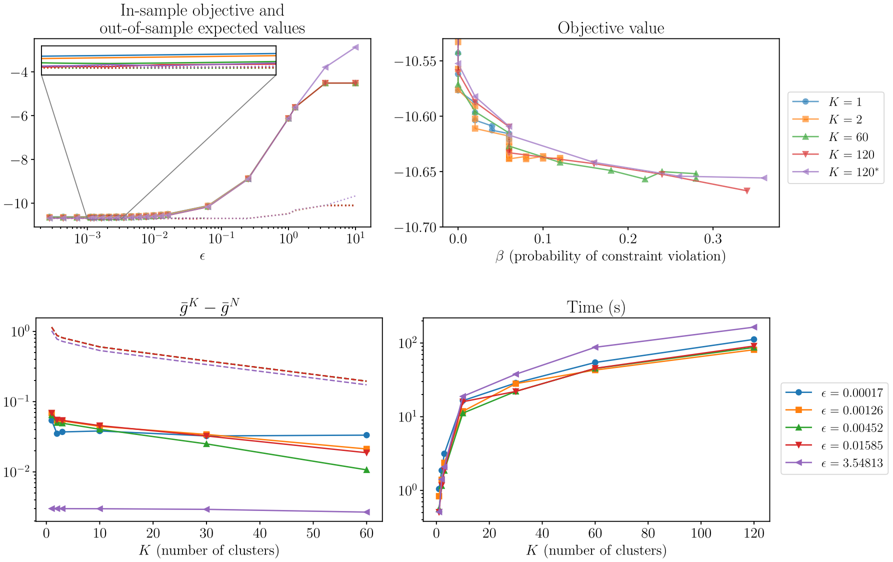


Fig. 2 Capital budgeting. Descriptions are given in the beginning of Sect. 6. The difference in objective values (bottom left) is calculated as $\bar{g}^K(x) - \bar{g}^N(x)$, and the theoretical bound is $(L/2)D(K)$ from Corollary 1

$K = 120^*$ is the formulation without the support constraint. Therefore, while $\bar{g}^{N*}(x)$ slightly upper bounds $\bar{g}^N(x)$, we can approximate their difference $\Delta \approx 0$ for small enough ϵ, for which the upper bound $(L/2)D(K)$ thus hold. In fact in this example, even for larger ϵ where we observe $\Delta > 0$, the actual difference between $\bar{g}^K$ and $\bar{g}^N$ is bounded by $(L/2)D(K)$. We see that the elbow of the upper bound is at $K = 2$, and the true difference follows the same trend, matching the suggestion from Fig. 1. Therefore, setting $K = 2$ is the optimal decision, with a time reduction of 2 orders of magnitude, and a complexity reduction from 26,626 variables and 12,000 power cones to 666 variables and 200 power cones.

6.2 Sparse portfolio optimization

We consider a market that forbids short-selling and has m assets as in [24]. Daily returns of these assets are given by the random vector $d = (d_1, \ldots, d_m) \in \mathbf{R}^m$. The percentage weights (of the total capital) invested in each asset are given by the decision vector $x = (x_1, \ldots, x_n) \in \mathbf{R}^n$. We restrict our selection to at most θ assets, given by the 0-th norm cardinality constraint below. The distribution $\mathbf{P}$ is unknown, but we have observed a historical dataset $\mathcal{D}_N$. Our objective is to minimize the CVaR with respect to variable x,

$$\begin{array}{ll} \text{minimize} & \mathbf{CVaR}(-u^T x, \alpha) \\ \text{subject to} & \mathbf{1}^T x = 1, \quad x \geq 0, \quad \|x\|_0 \leq \theta, \end{array}$$

which represents the average of the α largest portfolio losses that occur. In other words, the **CVaR** term seeks to ensure that the expected magnitude of portfolio losses, when they occur, is low. The objective has an analytical form with an extra variable τ given as [24, 46]: $\mathbf{E}^{\mathbf{P}}\left(\tau + \frac{1}{\alpha}\max\{-u^T x - \tau, 0\}\right)$. From this, we obtain g as the maximum of affine functions,

$$g(u, x) = \max\{(-1/\alpha)x^T u + (1 - 1/\alpha)\tau, \tau\}.$$

We then apply the convex reformulation given in Appendix B, for $p = \infty$.

Problem setup. We take stock data from the past 10 years of S&P500, and generate synthetic data from their fitted general Pareto distributions. We choose a generalized Pareto fit over a normal distribution as it better models the heavy tails of the returns [15]. See the Github repository for the code, which uses the "Rsafd" R package [16]. We let $\alpha = 20\%$, $m = 50$ stocks, and generate a dataset size of $N = 1000$. Our portfolio can include at most $\theta = 5$ stocks. For the upper bound $\delta(K, z, \gamma)$ on $\bar{g}^N(x) - \bar{g}^K(x)$, we note the special structure of this problem, where one of the affine pieces is independent of u, to arrive at a bound $\delta(K, z, \gamma) = \max_k\{\max_i\{(\bar{d}_k - d_i)^T x/\alpha\}\}$.

Results In Fig. 3, while setting K to smaller values lead to a decrease in the optimal value across ϵ, we note that for $K = 5$ and above, we can already achieve a tradeoff curve between the optimal value and probability of constraint satisfaction that is similar to that of $K = 1000$, and setting $K = 10$ brings it slightly closer. In the plots of $D(K)$ and of the upper bound on the difference, we also note that the elbow is around $K = 5$. We thus recommend choosing K through cross validation around 5, as tuning ϵ for these small K gives 1–3 orders of magnitude time reduction.

6.3 Facility location

We examine the classic facility location problem [9, 33]. Consider a set of n potential facilities, and m customers. Variable $x \in \{0, 1\}^n$ describes whether or not we construct each facility i for $i = 1, \dots, n$, with cost c_i. In addition, we would like to satisfy the uncertain demand $u \in \mathbf{R}^m$ at minimal cost. We define variable $X \in \mathbf{R}^{n \times m}$ where X_{ij} corresponding to the portion of the demand of customer j shipped from facility i with corresponding cost C_{ij}. Furthermore, $r \in \mathbf{R}^n$ represents the production capacity for each facility, and $u \in \mathbf{R}^m$ represents the uncertain demand from each customer. For each customer j, X_j represents the proportion of goods shipped from any facility to that customer, which sums to 1. For each facility i, $(X^T)_i$ represents the proportion of goods shipped to any customer. Putting this all together, we obtain the objective to minimize, $c^T x + \mathbf{tr}(C^T X)$, subject to constraints $\mathbf{1}^T X_j = 1, \;\; j = 1, \dots, m$, as well as multiple affine uncertain capacity constraints,

$$g_i(u, x) = (X^T)_i u - r_i x_i \le 0 \quad i = 1, \dots, n,$$

which we combine to create a single maximum-of-affine constraint, $g(u, x) = \max_{i \le n}((X^T)_i u - r_i x_i) \le 0$. Now, to ensure a high probability of constraint sat-

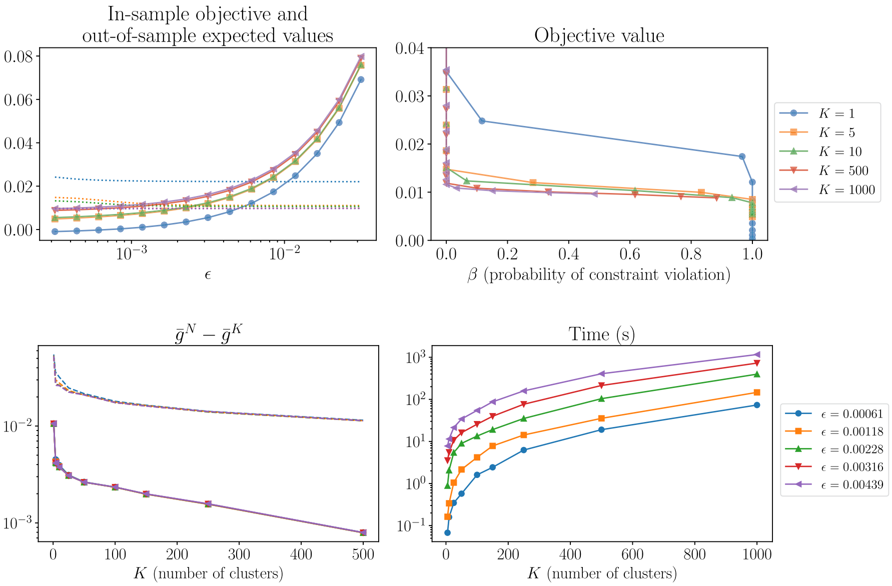


Fig. 3 Sparse portfolio. Descriptions are given in the beginning of Sect. 6. The difference in objective values (bottom left) is calculated as $\bar{g}^N(x) - \bar{g}^K(x)$, and the theoretical bound is $\delta(K, z, \gamma)$ from Corollary 3

isfaction, we use the **CVaR** reformulation,

$$g(u, x, \tau) = \tau + (1/\alpha) \max \left(\max_{i \leq n} ((X^T)_i u - r_i x_i - \tau), 0 \right) \leq 0,$$

where we add the auxiliary variable τ. We assume a polyhedral support $S = \{u \mid Hu \leq b\}$ for the demand, and apply the convex reformulation given in Appendix B, for $p = \infty$.

Problem setup. To generate data, we set $n = 5$ facilities, $m = 25$ customers, and $N = 50$ data samples. For the **CVaR** reformulation, we set $\alpha = 20\%$. We set costs $c = (46.68, 58.81, 30, 42.09, 35.87)$, and generate the two coordinates of each customer's location from a uniform distribution on $[0, 15]$. We then calculate C as the ℓ_2 distance between each pair of customers. We set production capacities $r = (33, 26, 41, 26, 22)$. We assume the demand d is supported between 1 and 6, written as $Hu \leq b$, where $H = [-I \;\; I]^T$ and b is the concatenation of a vector of -1's of length m and a vector of 6's of length m. We generate demands as the combination of two normal distributions. Half of the data is generated with mean $\mu_1 = 3$ and variance $\sigma_1 = 0.9$, the second half has mean $\mu_1 = 4$ and variance $\sigma_1 = 0.8$. We then project the demands onto $(1, 6)$. For the upper bound $\delta(K, z, \gamma)$ on $\bar{g}^N(x) - \bar{g}^K(x)$ from Corollary 3, we have $(1/N)\sum_{k=1}^{K}\sum_{i \in C_k} \max(\max_{j \leq J}(((1/\alpha)X[i] - H^T\gamma_{jk}), 0)^T(d_i - \bar{d}_k))$. Note that this upper bounds the difference in constraint values, and only indirectly affects the objective values through restrictions on the feasible region. Therefore, it is not an upper bound on the difference in objective value, merely an estimate. We cannot directly compare this upper bound against the change in constraint values, as the

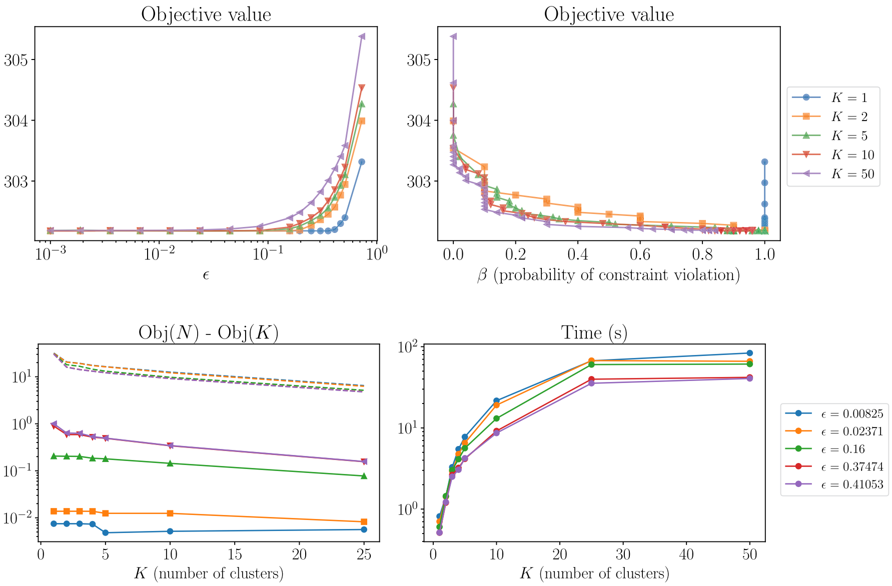


Fig. 4 Facility location. Descriptions are given in the beginning of Sect. 6. The difference in objective values (bottom left) is calculated as Obj(N) - Obj(K), and we compare it to the theoretical upper bound $\delta(K, z, \gamma)$ on the worst-case constraint value $\bar{g}^N(x) - \bar{g}^K(x)$, from Corollary 3

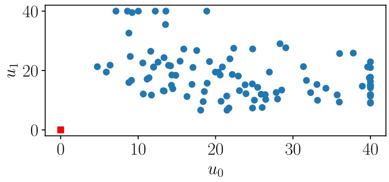
constraint value will always be near 0 for optimality. We thus compare it against the change in objective values.

Results As expected of maximum-of-affine g, we note in Fig. 4 that setting K to smaller values lead to a decrease in the optimal value across different ϵ values. While $K = 1$ yields poor performance in terms of the probability of constraint violation, we observe that $K = 2$ already yields a tradeoff between the objective and probability of constraint violation close to that of $K = 50$. Through cross-validation with different K, we select $K = 5$, which provides a tradeoff curve closer to optimality. As this problem has uncertainty in the constraints and not the objective, the bounds given in Corollary 3 do not directly reflect the difference in the objective values. However, they do give a reference value and inform us of the general trend of the difference. In this case, they still upper bound the actual difference, as shown in Fig. 4. We note that the bounds we use do not depend on $\bar{g}^{N*}(x)$, so it is irrelevant whether or not the support has an affect on the worst-case constraint value. Overall, choosing $K = 5$ leads to a time reduction of an order of magnitude while achieving near-optimal performance.

6.4 Newsvendor problem

We consider a 2-item newsvendor problem where, at the beginning of each day, the vendor orders $x \in \mathbf{R}^2_+$ products at price $h = (4, 5)$. These products will be sold at the prices $c = (5, 6.5)$, until either the uncertain demand u or inventory x is exhausted. The objective function to minimize is the sum of the ordering cost minus the revenue,

Fig. 5 Newsvendor. data-points and the outlier at (0,0)



$h^T x - c^T \min\{x, u\}$, from which we obtain the maximum-of-affine uncertain function g to minimize,

$$g(u, x) = h^T x + \max(-c_1 x_1 - c_2 x_2, -c_1 x_1 - c_2 u_2, -c_1 u_1 - c_2 x_2, -c_1 u_1 - c_2 u_2).$$

We assume a polyhedral support $S = \{u \mid Cu \leq b\}$, and obtain the convex reformulation, for $p = 1$, by applying (15).

For this problem, we consider the effects of outliers on the performance of MRO. Therefore, we consider the data to have an outlier at (0, 0), the worst-case value of the support set. In Fig. 5, we show a set of generated data along with this outlier point.

We consider three ways to solve the problem.

1. MRO, where we directly apply MRO to the dataset with the outlier.
2. ROB-MRO, where we perform preliminary analysis on the dataset to remove the outlier point, then apply MRO to the cleaned dataset.
3. AUG-MRO, where we perform the clustering step on data without the outlier, then define an augmented distribution supported on $K + 1$ points, where the extra point is the outlier point (0,0), with weight $1/N$. The weights of the other clusters are adjusted accordingly.

Problem setup. To generate data, we set $N = 100$ data samples. We assume demand is supported between 0 and 40, which we write as $Cu \leq b$, where $C = [-I \;\; I]^T$ and $b = (0, 0, 0, 0, 40, 40, 40, 40)$. We allow non-integer demand to allow for more variance in the data. We generate the demand from a log-normal distribution, where the underlying normal distribution has parameters

$$\mu = \begin{bmatrix} 3.0 \\ 2.8 \end{bmatrix}, \quad \Sigma = \begin{bmatrix} 0.3 & -0.1 \\ -0.1 & 0.2 \end{bmatrix},$$

and take the minimum between the generated values and 40. For the upper bound $\delta(K, z, \gamma)$ on $\bar{g}^N(x) - \bar{g}^K(x)$ from Corollary 3, we have $(1/N)\sum_{k=1}^{K} \sum_{i \in C_k} \max_{j \leq 4}((-\tilde{c}_j - C^T \gamma_{jk})^T (d_i - \bar{d}_k))$, where $\tilde{c}_1 = 0,\ \tilde{c}_2 = c_1 e_1,\ \tilde{c}_3 = c_2 e_2,\ \tilde{c}_4 = c$.

Results To examine the effect of the outlier, in Fig. 6, we compare, for $K = 10$ and $K = 100$, the objectives and tradeoff curves for the three methods. We note that, when the outlier is averaged with other data-points, the final in-sample objective may be improved, as the centroid moves closer to the non-outlier points. We observe that MRO, in which the outlier may be clustered with other points, offers a lower in-sample objective than AUG-MRO, in which the outlier is considered its own cluster. MRO has in fact offered protection against the outlier. And, as expected, ROB-MRO, where the outlier point is removed, yields the best in-sample results. Regardless of the method, we note that the final out-of-sample tradeoff curves are near-identical. Comparing the plots for $K = 10$ and $K = N = 100$, the difference between MRO and ROB-MRO for $K = 10$ is not larger than the difference for $K = 100$, which shows that, while removing outliers a priori may be helpful, the effect of outliers will not be worse for MRO compared to classic Wasserstein DRO.

In Fig. 7, we compare the in and out-of-sample objective values for MRO. While setting $K = 1$ yields suboptimal results, we note that for $K = 5$ and above, we can achieve similar performance as setting $K = 100$. We note that the upper bound on $\bar{g}^N(x) - \bar{g}^K(x)$, given in Corollary 3, holds for MRO. We again note that bounds we observe do not depend on $\bar{g}^{N*}(x)$, so it is irrelevant whether or not the support has an affect on the worst-case constraint value. Regardless, we see that the support only minimally affects the worst-case constraint value, at only at higher values of ϵ. Overall, choosing $K = 5$, we obtain an order of magnitude computational speed-up.

7 Conclusions

We have presented mean robust optimization (MRO), a new data-driven methodology for decision-making under uncertainty that bridges robust and distributionally robust optimization while preserving rigorous probabilistic guarantees. By clustering the dataset before performing MRO, we solve an efficient and computationally tractable formulation with limited performance degradation. In particular, we showed that when the constraints are affine in the uncertainty, clustering does not affect the optimal value of the objective. When the constraint is concave or maximum-of-concave in the uncertainty, we directly quantified the change in worst-case constraint value that is caused by clustering. For problems with objective uncertainty, this directly bounds the change in the optimal value caused by clustering. We demonstrated this result through a set of numerical examples, where we observed the possibility of tuning the size of the uncertainty set such that using a small number of clusters achieves near-identical performance of traditional DRO, with much higher computational efficiency. In the final example, we also demonstrated that MRO offers protection against outliers compared to Wasserstein DRO.

A Proof of the constraint reformulation in (15)

We give a prove for the general case of maximum-of-concave functions. When $J = 1$, we take $\alpha_{1k} = w_k$, for all k. For a simpler proof for the case of single-concave functions, refer to [47, Appendix A].

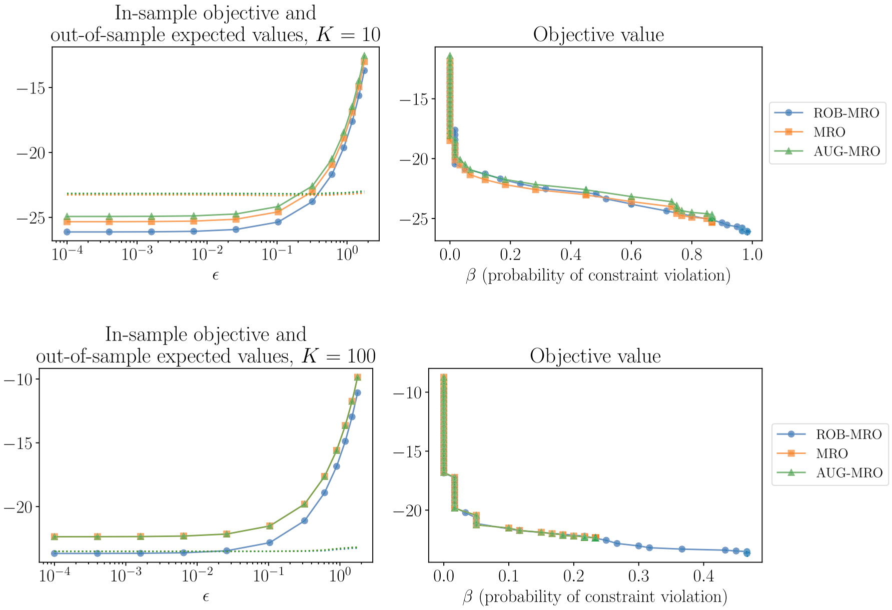


Fig. 6 Newsvendor. Comparing the three methods for $K = 10$ and $K = 100$. Left: in-sample objective values vs ϵ. Right: objective value vs β

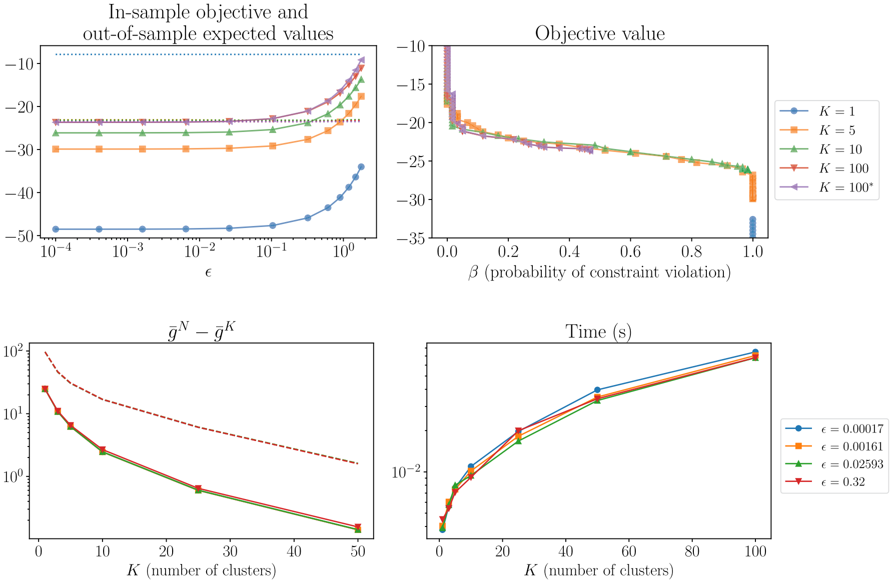


Fig. 7 Newsvendor, MRO. Descriptions are given in the beginning of Sect. 6. The difference in objective values (bottom left) is calculated as $\bar{g}^N(x) - \bar{g}^K(x)$, and the theoretical bound is $\delta(K, z, \gamma)$ from Corollary 3

To simplify notation, we define $c_k(v_{jk}) = \|v_{jk} - \bar{d}_k\|^p - \epsilon^p$. Then, starting from the inner optimization problem of (14):

$$\begin{cases} \sup\limits_{v_{11},\dots,v_{JK}\in S,\alpha\in\Gamma} & \sum_{k=1}^K \sum_{j=1}^J \alpha_{jk} g_j(v_{jk},x) \\ \text{subject to} & \sum_{k=1}^K \sum_{j=1}^J \alpha_{jk} c_k(v_{jk}) \le 0 \end{cases}$$

$$= \begin{cases} \sup\limits_{v_{11},\dots,v_{JK}\in S,\alpha\in\Gamma} & \inf\limits_{\lambda\ge 0} \sum_{k=1}^K \sum_{j=1}^J \alpha_{jk} g_j(v_{jk},x) - \lambda \sum_{k=1}^K \sum_{j=1}^J \alpha_{jk} c_k(v_{jk}) \end{cases}$$

$$= \begin{cases} \sup\limits_{\alpha\in\Gamma} \inf\limits_{\lambda\ge 0} \sup\limits_{v_{11},\dots,v_{JK}\in S} & \sum_{k=1}^K \sum_{j=1}^J \alpha_{jk} g_j(v_{jk},x) - \lambda \sum_{k=1}^K \sum_{j=1}^J \alpha_{jk} c_k(v_{jk}), \end{cases}$$

We applied the Lagrangian in the first equality. Then, as the summation is over upper-semicontinuous functions $g_j(v_{jk}, x)$ concave in v_{jk}, we applied the Von Neumann-Fan minimax theorem [38] to interchange the inf and the sup. Next, we rewrite the formulation using an epigraph trick, and make a change of variables.

$$= \begin{cases} \sup\limits_{\alpha\in\Gamma} \inf\limits_{\lambda\ge 0,s} & & \sum_{k=1}^K s_k \\ \text{subject to} & \sup\limits_{v_{11},\dots,v_{JK}\in S} & \sum_{j=1}^J \alpha_{jk}(g_j(v_{jk},x) - \lambda c_k(v_{jk})) \le s_k \quad k = 1,\dots,K, \end{cases}$$

$$= \begin{cases} \sup\limits_{\alpha\in\Gamma} \inf\limits_{\lambda\ge 0,s} & & \sum_{k=1}^K s_k \\ \text{subject to} & \sup\limits_{\alpha_{11}v_{11}\in\alpha_{11}S,\dots,\alpha_{JK}v_{JK}\in\alpha_{JK}S} & \sum_{j=1}^J \alpha_{jk}(g_j((\alpha_{jk}v_{jk})/\alpha_{jk},x) \\ & & -\lambda c_k((\alpha_{jk}v_{jk})/\alpha_{jk})) \le s_k \quad k = 1,\dots,K. \end{cases}$$

In the last step, we rewrote $v_{jk} = (\alpha_{jk}v_{jk})/\alpha_{jk}$, and maximized over $\alpha_{jk}v_{jk} \in \alpha_{jk}S$. In the case $\alpha_{ij} > 0$, the terms in the summation are unchanged, and maximizing over v_{jk} is equivalent to maximizing over $\alpha_{jk}v_{jk}$. In the case $\alpha_{jk} = 0$, we have in the transformed formulation $(\alpha_{jk}v_{jk})/\alpha_{jk} = 0/0 = 0$, and $\alpha_{jk}(g_j(0,x) - \lambda c_k(0)) = 0(g_j(0,x) - \lambda c_k(0)) = 0$. In the original formulation, the term $\alpha_{jk}(g_j(v_{jk},x) - \lambda c_k(v_{jk}))$ is also equivalent to 0. As the terms in the summation are 0 regardless of the value of v_{jk}, maximizing over v_{jk} is equivalent to maximizing over 0. Therefore, we note that the optimal value remains unchanged. Next, we make substitutions $h_{jk} = \alpha_{jk}v_{jk}$, and define functions $g'_j(h_{jk},x) = \alpha_{jk}g_j(h_{jk}/\alpha_{jk},x)$, $c'_k(h_{jk}) = \alpha_{jk}c_k(h_{jk}/\alpha_{jk})$.

$$= \begin{cases} \sup\limits_{\alpha\in\Gamma} \inf\limits_{\lambda\ge 0,s} & & \sum_{k=1}^K s_k \\ \text{subject to} & \sup\limits_{h_{11}\in\alpha_{11}S,\dots,h_{JK}\in\alpha_{JK}S} & \sum_{j=1}^J g'_j(h_{jk},x) - \lambda c'_k(h_{jk}) \le s_k \quad k = 1,\dots,K, \end{cases}$$

$$= \begin{cases} \sup\limits_{\alpha\in\Gamma} \inf\limits_{\lambda\ge 0,s} & \sum_{k=1}^K s_k \\ \text{subject to} & \sum_{j=1}^J [-g'_j + \chi_{\alpha_{jk}S} + \lambda c'_k]^*(0) \le s_k \quad k = 1,\dots,K. \end{cases}$$

For the new functions defined, we applied the definition of conjugate functions. We also define the characteristic function $\chi_S(v)$ with $\chi_S(v) = 0$ if $v \in S$; $= \infty$ otherwise.

Now, using the conjugate form $f^*(y) = \alpha g^*(y)$ of a right-scalar-multiplied function $f(x) = \alpha g(x/\alpha)$, and noting that $\chi_{\alpha_{jk}S}(h)$ takes the same value as $\alpha_{jk}\chi_S(h/\alpha_{jk})$, we can rewrite the constraint as

$$\sum_{j=1}^{J} \alpha_{jk}[-g_j + \chi_S + \lambda c_k]^*(0) \le s_k \quad k = 1, \dots, K.$$

Next, borrowing results from Esfahani et al. [24, Theorem 4.2], Rockafellar and Wets [41, Theorem 11.23(a), p. 493], and Zhen et al. [51, Lemma B.8], with regards to the conjugate functions of infimal convolutions and p-norm balls, we note that:

$$\begin{aligned}\alpha_{jk}[(-g_j + \chi_S + \lambda c_k)]^*(0) = \alpha_{jk} \inf_{y_{jk}, z_{jk}} &([-g_j]^*(z_{jk} - y_{jk}, x) \\ &+ \sigma_S(y_{jk}) + [\lambda c_k]^*(-z_{jk})),\end{aligned}$$

$$\begin{aligned}[\lambda c_k]^*(-z_{jk}) &= \sup_{v_{jk}}(-z_{jk}^T v_{jk} - \lambda \|v_{jk} - \bar{d}_k\|^p + \lambda \epsilon^p) \\ &= -z_{jk}^T \bar{d}_k + \phi(q)\lambda \|z_{jk}/\lambda\|_*^q + \lambda \epsilon^p.\end{aligned}$$

Substituting this in, we have

$$\begin{cases} \sup\limits_{\alpha \in \Gamma} \inf\limits_{\lambda \ge 0, s, z, y} & \sum_k^K s_k \\ \text{subject to} & \sum_{j=1}^J \alpha_{jk}([-g_j]^*(z_{jk} - y_{jk}, x) \\ & +\sigma_S(y_{jk}) - z_{jk}^T \bar{d}_k + \phi(q)\lambda \left\|z_{jk}/\lambda\right\|_*^q + \lambda\epsilon^p) \le s_k \\ & \qquad k = 1, \dots, K. \end{cases}$$

Taking the supremum over α, noting that $\sum_{j=1}^J \alpha_{jk} = w_k$ for all k, we arrive at

$$\begin{cases} \inf\limits_{\lambda \ge 0, s, z, y} & \sum_k^K s_k \\ \text{subject to} & w_k([-g_j]^*(z_{jk} - y_{jk}, x) + \sigma_S(y_{jk}) - z_{jk}^T \bar{d}_k \\ & \quad + \phi(q)\lambda \left\|z_{jk}/\lambda\right\|_*^q + \lambda\epsilon^p) \le s_k \quad k = 1, \dots, K, \quad j = 1, \dots, J, \end{cases}$$

which is equivalent to (15).

B Reformulation of the maximum-of-concave case for $p = \infty$

We again give the general proof for $J \ge 1$. For the case $J = 1$, refer to the simpler proof [47, Appendix B]. When $p = \infty$, we have

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & \left\{ \begin{array}{ll} \underset{v_{11},\dots,v_{JK}\in S,\alpha\in\Gamma}{\text{maximize}} & \sum_{k=1}^{K}\sum_{j=1}^{J}\alpha_{jk}g(v_{jk},x) \\ \text{subject to} & \sum_{j=1}^{J}(\alpha_{jk}/w_k)\|v_{jk}-\bar{d}_k\| \le \epsilon, \quad k=1,\dots,K \end{array} \right\} \le 0, \end{array} \tag{17}$$

which has a reformulation where the constraint above is dualized,

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & \sum_{k=1}^{K} w_k s_k \le 0 \\ & [-g_j]^*(z_{jk}-y_{jk},x)+\sigma_S(y_{jk})-z_{jk}^T\bar{d}_k+\lambda_k\epsilon \le s_k \\ & \qquad\qquad\qquad\qquad k=1,\dots,K, \quad j=1,\dots,J \\ & \left\|z_{jk}\right\|_* \le \lambda_k \quad k=1,\dots,K, \quad j=1,\dots,J, \end{array} \tag{18}$$

with new variables $s_k\in\mathbf{R}$, $z_{jk}\in\mathbf{R}^m$, and $y_{jk}\in\mathbf{R}^m$. We prove this by starting from the inner optimization problem of (17):

$$\begin{aligned} &\left\{ \begin{array}{ll} \underset{v_{11},\dots,v_{JK}\in S,\alpha\in\Gamma}{\sup} & \sum_{k=1}^{K}\sum_{j=1}^{J}\alpha_{jk}g_j(v_{jk},x) \\ \text{subject to} & \sum_{j=1}^{J}(\alpha_{jk}/w_k)\|v_{jk}-\bar{d}_k\| \le \epsilon, \quad k=1,\dots,K \end{array} \right. \\ &= \left\{ \begin{array}{ll} \underset{v_{11},\dots,v_{JK}\in S,\alpha\in\Gamma}{\sup} & \underset{\lambda\ge 0}{\inf} \ \sum_{k=1}^{K}(\sum_{j=1}^{J}\alpha_{jk}g_j(v_{jk},x) \\ & \qquad +\lambda_k(\epsilon-\sum_{j=1}^{J}(\alpha_{jk}/w_k)\|v_{jk}-\bar{d}_k\|)) \end{array} \right. \\ &= \left\{ \begin{array}{ll} \underset{\alpha\in\Gamma}{\sup}\ \underset{\lambda\ge 0}{\inf}\ \underset{v_{11},\dots,v_{JK}\in S}{\sup} & \sum_{k=1}^{K}(\sum_{j=1}^{J}\alpha_{jk}g_j(v_{jk},x) \\ & \qquad +\lambda_k(\epsilon-\sum_{j=1}^{J}(\alpha_{jk}/w_k)\|v_{jk}-\bar{d}_k\|)) \end{array} \right. \end{aligned}$$

We have again formulated the Lagrangian and applied the minmax theorem to the sum of concave functions in v_{jk}. Now, we can rewrite this in epigraph form,

$$= \left\{ \begin{array}{lll} \underset{\alpha\in\Gamma}{\sup}\ \underset{\lambda\ge 0,s}{\inf} & \sum_{k=1}^{K} s_k & \\ \text{subject to} & \underset{v_{11},\dots,v_{JK}\in S}{\sup} \ \lambda_k\epsilon+\sum_{j=1}^{J}\alpha_{jk}g_j(v_{jk},x) & \\ & \qquad -\lambda_k(\alpha_{jk}/w_k)\|v_{jk}-\bar{d}_k\| \le s_k & k=1,\dots,K, \end{array} \right.$$

then use the definition of the dual norm to rewrite the constraints,

$$\begin{aligned} &\left\{ \begin{array}{l} \underset{v_{11},\dots,v_{JK}\in S}{\sup} \ \lambda_k\epsilon+\sum_{j=1}^{J}\underset{\|z_{jk}\|_*\le\lambda_k/w_k}{\min}\alpha_{jk}g_j(v_{jk},x) \\ \qquad -\alpha_{jk}z_{jk}^T(v_{jk}-\bar{d}_k) \le s_k \quad k=1,\dots,K, \end{array} \right. \\ &= \left\{ \begin{array}{l} \underset{\alpha_{11}v_{11}\in\alpha_{11}S,\dots,\alpha_{JK}v_{JK}\in\alpha_{JK}S}{\sup} \ \lambda_k\epsilon+\sum_{j=1}^{J}\underset{\|z_{jk}\|_*\le\lambda_k/w_k}{\min}\alpha_{jk}g_j \\ \qquad ((\alpha_{jk}v_{jk})/\alpha_{jk},x)-\alpha_{jk}z_{jk}^T(v_{jk}-\bar{d}_k) \le s_k \qquad k=1,\dots,K. \end{array} \right. \end{aligned}$$

In the last step, we again rewrote $v_{jk} = (\alpha_{jk} v_{jk})/\alpha_{jk}$ inside functions g_j. Using similar logic as in Appendix A, we note that this does not change the optimal value for both $\alpha_{jk} > 0$ and $\alpha_{jk} = 0$, as taking the supremum over the transformed variables is equivalent to taking the supremum over the original variables. For conciseness, we omit the steps of creating the right-scalar-multiplied function and transformations. For details, please refer to Appendix A. We apply the definition of conjugate functions and arrive at the optimization problem

$$\begin{cases} \sup\limits_{\alpha \in \Gamma} \inf\limits_{\lambda \geq 0, s, z} & \sum_{k=1}^K s_k \\ \text{subject to} & \lambda_k \epsilon + \sum_{j=1}^J \alpha_{jk} [-g_j + \chi_S]^*(-z_{jk}, x) + \alpha_{jk} z_{jk}^T \bar{d}_k \leq s_k \\ & \hfill k = 1, \dots, K \\ & \|z_{jk}\|_* \leq \lambda_k / w_k \quad k = 1, \dots, K, \quad j = 1, \dots, J. \end{cases}$$

Now, substituting $\lambda_k = \lambda_k w_k, z_{jk} = -z_{jk}$, and substituting in the conjugate functions derived in Appendix A, we have

$$= \begin{cases} \sup\limits_{\alpha \in \Gamma} \inf\limits_{\lambda \geq 0, s, z} & \sum_{k=1}^K s_k \\ \text{subject to} & \lambda_k w_k \epsilon + \sum_{j=1}^J \alpha_{jk} ([-g_j]^*(z_{jk} - y_{jk}, x) + \sigma_S(y_{jk}) \\ & \qquad -z_{jk}^T \bar{d}_k) \leq s_k \quad k = 1, \dots, K \\ & \|z_{jk}\|_* \leq \lambda_k, \quad k = 1, \dots, K, \quad j = 1, \dots, J. \end{cases}$$

Note that rescaling λ_k did not affect value of the problem, as minimizing λ_k is equivalent to minimizing $\lambda_k w_k$, as $w_k > 0$. Lastly, taking the supremum over α, we arrive at

$$= \begin{cases} \inf\limits_{\lambda \geq 0, s, z} & \sum_{k=1}^K s_k \\ \text{subject to} & w_k (\lambda_k \epsilon + [-g_j]^*(z_{jk} - y_{jk}, x) + \sigma_S(y_{jk}) - z_{jk}^T \bar{d}_k) \leq s_k \\ & \hfill k = 1, \dots, K, \quad j = 1, \dots, J \\ & \|z_{jk}\|_* \leq \lambda_k, \quad k = 1, \dots, K, \quad j = 1, \dots, J. \end{cases}$$

C Proof of the dual problem reformulation as p → ∞

We prove the dual equivalence. For a proof of primal equivalence, see the appendix of [47].

Theorem 6 *Let S be a bounded set. Define here*

$$\bar{g}^K(x; \infty) = \begin{cases} \text{minimize} & \sum_{k=1}^K w_k s_k \\ \text{subject to} & \lambda \geq 0, \; z_k \in \mathbf{R}^m, \; y_k \in \mathbf{R}^m, \; s_k \in \mathbf{R}^m \quad k = 1, \dots, K, \\ & [-g]^*(z_k - y_k) + \sigma_S(y_k) - z_k^T d_k + \epsilon \, \|z_k\|_* \leq s_k \\ & \hfill k = 1, \dots, K. \end{cases} \tag{19}$$

Then, $\lim_{p\to\infty} \bar{g}^K(x;p) = \bar{g}^K(x;\infty)$ *for any* $x \in X$.

Proof First, from Equation (5) we have for any $p > 1$ that

$$\begin{aligned}
&\bar{g}^K(x;p)\\
&=\left\{\begin{array}{ll}
\text{minimize} & \sum_{k=1}^K w_k s_k\\
\text{subject to} & \lambda \geq 0,\ z_k \in \mathbf{R}^m,\ y_k \in \mathbf{R}^m,\ s_k \in \mathbf{R}^m \quad k = 1,\dots,K,\\
& [-g]^*(z_k - y_k) + \sigma_S(y_k) - z_k^T d_k\\
& \qquad\qquad +\phi(q)\lambda \left\|z_k/\lambda\right\|_*^q + \lambda\epsilon^p \leq s_k \qquad k = 1,\dots,K
\end{array}\right.\\
&\geq\left\{\begin{array}{ll}
\text{minimize} & \sum_{k=1}^K w_k s_k\\
\text{subject to} & \lambda_k \geq 0,\ z_k \in \mathbf{R}^m,\ y_k \in \mathbf{R}^m,\ s_k \in \mathbf{R}^m \quad k = 1,\dots,K,\\
& [-g]^*(z_k - y_k) + \sigma_S(y_k) - z_k^T d_k\\
& \qquad\qquad +\phi(q)\lambda_k \left\|z_k/\lambda_k\right\|_*^q + \lambda_k\epsilon^p \leq s_k \qquad k = 1,\dots,K
\end{array}\right.\\
&=\left\{\begin{array}{ll}
\text{minimize} & \sum_{k=1}^K w_k s_k\\
\text{subject to} & z_k \in \mathbf{R}^m,\ y_k \in \mathbf{R}^m,\ s_k \in \mathbf{R}^m \quad k = 1,\dots,K,\\
& [-g]^*(z_k - y_k) + \sigma_S(y_k) - z_k^T d_k\\
& \qquad\qquad +\epsilon \left\|z_k\right\|_* \leq s_k \qquad k = 1,\dots,K
\end{array}\right.
\end{aligned}$$

where the first equality is established in Appendix A and the second equality follows from Lemma 1. Remark that the inequality in the second step simply follows as we introduce λ_k and do not impose that $\lambda_k = \lambda$ for all $k = 1,\dots,K$. Hence, considering the limit for p tending to infinity gives us now $\liminf_{p\to\infty}\bar{g}^K(x;p) \geq \bar{g}^K(x;\infty)$. It remains to prove the reverse $\limsup_{p\to\infty}\bar{g}^K(x;p) \leq \bar{g}^K(x;\infty)$.

Second, we have for any $p > 1$ with $1/p + 1/q = 1$ that

$$\begin{aligned}
&\bar{g}^K(x;p)\\
&\leq\left\{\begin{array}{ll}
\text{minimize} & \sum_{k=1}^K w_k s_k\\
\text{subject to} & z_k \in \mathbf{R}^m,\ y_k \in \mathbf{R}^m,\ s_k \in \mathbf{R}^m \quad k = 1,\dots,K,\\
& [-g]^*(z_k - y_k) + \sigma_S(y_k) - z_k^T d_k + \phi(q)\\
& \qquad\qquad \left[\frac{q-1}{q}\epsilon^{\frac{1}{1-q}} \max_{k'=1}^K \left\|z_{k'}\right\|_*\right]^{1-q} \quad \left\|z_k\right\|_*^q\\
& \qquad\qquad +\left[\frac{q-1}{q}\epsilon^{\frac{1}{1-q}} \max_{k'=1}^K \left\|z_{k'}\right\|_*\right]\epsilon^p \leq s_k \quad k = 1,\dots,K\\
& (q-1)^{1/4} \leq \left\|z_k\right\|_* \leq (q-1)^{-1/4} \quad k = 1,\dots,K,
\end{array}\right.
\end{aligned}$$

which follows from the choice $\lambda_k = \left[\frac{q-1}{q}\epsilon^{\frac{1}{1-q}} \max_{k'=1}^K \left\|z_{k'}\right\|_*\right]$ and by imposing the restrictions $(q-1)^{1/4} \leq \left\|z_k\right\|_* \leq (q-1)^{-1/4}$ for all $k = 1,\dots,K$. Next, using the identities $\phi(q)\left[\frac{q-1}{q}\right]^{1-q} = (q-1)^{(q-1)}/q^q\left[\frac{q-1}{q}\right]^{1-q} = 1/q$ and $p + \frac{1}{1-q} = \frac{1}{\frac{1}{p}} + \frac{1}{1-q} = \frac{1}{1-\frac{1}{q}} + \frac{1}{1-q} = \frac{-q}{-q+1} + \frac{1}{1-q} = \frac{1-q}{1-q} = 1$, as well as pulling $\epsilon\left\|z_k\right\|_* > 0$ out of the last two terms, we can rewrite the first constraints as

$$\left\{\begin{array}{l}
[-g]^*(z_k - y_k) + \sigma_S(y_k) - z_k^T d_k\\
\qquad +\epsilon\left\|z_k\right\|_*\left[\frac{1}{q}\left[\max_{k'=1}^K \frac{\left\|z_{k'}\right\|_*}{\left\|z_k\right\|_*}\right]^{1-q} + \frac{q-1}{q}\max_{k'=1}^K \frac{\left\|z_{k'}\right\|_*}{\left\|z_k\right\|_*}\right] \leq s_k \quad k = 1,\dots,K.
\end{array}\right.$$

Then, we note that $\max_{k'=1}^{K} \|z_{k'}\|_* / \|z_k\|_* \geq \|z_k\|_* / \|z_k\|_* = 1$ and $\max_{k'=1}^{K} \|z_{k'}\|_* / \|z_k\|_* \leq (q-1)^{-1/2}$, and hence we can apply Lemma 2 to obtain

$$\bar{g}^K(x; p) \leq \begin{cases} \text{minimize} & \sum_{k=1}^K w_k s_k \\ \text{subject to} & z_k \in \mathbf{R}^m,\ y_k \in \mathbf{R}^m,\ s_k \in \mathbf{R}^m \quad k = 1, \ldots, K, \\ & [-g]^*(z_k - y_k) + \sigma_S(y_k) - z_k^T d_k \\ & \qquad +\epsilon \|z_k\|_* D(q) \leq s_k \quad k = 1, \ldots, K, \\ & (q-1)^{1/4} \leq \|z_k\|_* \leq (q-1)^{-1/4} \quad k = 1, \ldots, K. \end{cases}$$

Let

$$\bar{g}_u^K(x; p) = \begin{cases} \text{minimize } \sum_{k=1}^K w_k s_k \\ \text{subject to } z_k \in \mathbf{R}^m,\ y_k \in \mathbf{R}^m,\ s_k \in \mathbf{R}^m \quad k = 1, \ldots, K, \\ \qquad [-g]^*(z_k - y_k) + \sigma_S(y_k) - z_k^T d_k \\ \qquad\qquad +\epsilon \|z_k\|_* D\left(\frac{p}{p-1}\right) \leq s_k \quad k = 1, \ldots, K \\ \qquad (p-1)^{-1/4} \leq \|z_k\|_* \leq (p-1)^{1/4} \quad k = 1, \ldots, K. \end{cases} \tag{20}$$

Hence, as $q = p/(p-1)$ and $q - 1 = 1/(p-1)$ we have $\bar{g}^K(x; p) \leq \bar{g}_u^K(x; p)$ for all $p > 1$. Hence, taking the limit $p \to \infty$ we have $\bar{g}^K(x; \infty) \leq \limsup_{p\to\infty} \bar{g}^K(x; p)$. In fact, as the function $D\left(\frac{p}{p-1}\right)$ defined in Lemma 2 is nonincreasing for all p sufficiently large this implies that $\bar{g}_u^K(x; p)$ is nonincreasing for p sufficiently large and hence we have $\bar{g}^K(x; \infty) \leq \limsup_{p\to\infty} \bar{g}^K(x; p) \leq \liminf_{p\to\infty} \bar{g}_u^K(x; p) = \lim_{p\to\infty} \bar{g}_u^K(x; p)$. We now prove here that $\lim_{p\to\infty} \bar{g}_u^K(x; p) = \liminf_{p\to\infty} \bar{g}_u^K(x; p) \leq \bar{g}^K(x; \infty)$. Consider any feasible sequence $\{(z_k^t,\ y_k^t,\ s_k^t = [-g]^*(z_k^t - y_k^t) + \sigma_S(y_k^t) - (z_k^t)^T d_k + \epsilon \left\|z_k^t\right\|_*)\}_{t\geq 1}$ in the optimization problem characterizing $\bar{g}^K(x; \infty)$ in Equation (19) so that $\lim_{t\to\infty} \sum_{k=1}^K w_k s_k^t = \bar{g}^K(x; \infty)$. Let $\tilde{z}_k^t \in \arg\max\{\|z\|_* \mid z \in \mathbf{R}^m,\ \|z - z_k^t\|_* \leq 1/t\}$ for all $t \geq 1$ and $k = 1, \ldots, K$ and observe that $\|\tilde{z}_k^t\|_* = 1/t + \|z_k^t\|_* \geq 1/t$. Consider now an increasing sequence $\{p_t\}_{t\geq 1}$ so that $(p_t - 1)^{1/4} \geq \max_{k=1}^K \|\tilde{z}_k^t\|_*$ and $(p_t - 1)^{-1/4} \leq 1/t$. Finally observe that the auxiliary sequence $\{(\tilde{z}_k^t,\ \tilde{y}_k^t = y_k^t + (\tilde{z}_k^t - z_k^t),\ \tilde{s}_k^t = [-g]^*(\tilde{z}_k^t - \tilde{y}_k^t) + \sigma_S(\tilde{y}_k^t) - (\tilde{z}_k^t)^T d_k + \epsilon \left\|\tilde{z}_k^t\right\|_* D\left(p_t/(p_t - 1)\right))\}_{t\geq 1}$ is by construction feasible in the minimization problem characterizing the function $\bar{g}_u^K(x; p_t)$ in Equation (20). Hence, finally, we have

$$\begin{aligned} \lim_{p\to\infty} g_u^K(x; p) &= \lim_{t\to\infty} g_u^K(x; p_t) = \lim_{t\to\infty} \sum_{k=1}^K w_k \tilde{s}_k^t \\ &= \lim_{t\to\infty} \sum_{k=1}^K w_k \left([-g]^*(\tilde{z}_k^t - \tilde{y}_k^t) + \sigma_S(\tilde{y}_k^t) - (\tilde{z}_k^t)^T d_k + \epsilon \left\|\tilde{z}_k^t\right\|_* D\left(p_t/(p_t - 1)\right) \right) \\ &\leq \lim_{t\to\infty} \sum_{k=1}^K w_k \left([-g]^*(z_k^t - y_k^t) + \sigma_S(y_k^t) - (z_k^t)^T d_k + \epsilon \left\|z_k^t\right\|_* D\left(p_t/(p_t - 1)\right) \right) \end{aligned}$$

$$
\begin{aligned}
&+\sum_{k=1}^{K} w_k \left(\max_{s \in S} \|s\| + \|d_k\| + \epsilon D\left(p_t/(p_t-1)\right) \right)/t \\
\leq &\lim_{t\to\infty} \sum_{k=1}^{K} w_k \left([-g]^*(z_k^t - y_k^t) + \sigma_S(y_k^t) - (z_k^t)^T d_k + \epsilon \left\| z_k^t \right\|_* D\left(p_t/(p_t-1)\right) \right) \\
= &\lim_{t\to\infty} \sum_{k=1}^{K} w_k([-g]^*(z_k^t - y_k^t) + \sigma_S(y_k^t) - (z_k^t)^T d_k \\
&+ \epsilon \left\| z_k^t \right\|_* + \epsilon \left\| z_k^t \right\|_* \left(D\left(p_t/(p_t-1)\right) - 1\right)) \\
\leq &\lim_{t\to\infty} \sum_{k=1}^{K} w_k([-g]^*(z_k^t - y_k^t) + \sigma_S(y_k^t) - (z_k^t)^T d_k \\
&+ \epsilon \left\| z_k^t \right\|_* + \epsilon (p_t-1)^{1/4} \left(D\left(p_t/(p_t-1)\right) - 1\right)) \\
\leq &\lim_{t\to\infty} \sum_{k=1}^{K} w_k s_k = \bar{g}^K(x;\infty).
\end{aligned}
$$

To establish the third inequality observe first that $-(\tilde{z}_k^t)^T d_k = -(z_k^t)^T d_k - (\tilde{z}_k^t - z_k^t)^T d_k \leq -(z_k^t)^T d_k + \|\tilde{z}_k^t - z_k^t\|_* \|d_k\| \leq -(z_k^t)^T d_k + \|d_k\|/t$. Second, remark that we have

$$
\begin{aligned}
\sigma_S(\tilde{y}_k^t) = \sigma_S(y_k^t + (\tilde{z}_k^t - z_k^t)) &\leq \max_{s\in S} s^T (y_k^t + (\tilde{z}_k^t - z_k^t)) \\
&\leq \max_{s\in S} s^T y_k^t + \max_{s\in S} s^T (\tilde{z}_k^t - z_k^t) \\
&\leq \max_{s\in S} s^T y_k^t + \|s\| \|\tilde{z}_k^t - z_k^t\|_* \leq \max_{s\in S} \; s^T y_k^t + 1/t \max_{s\in S} \|s\|,
\end{aligned}
$$

as $\|\tilde{z}_t - z_t\| \leq 1/t$. Lemma 2 guarantees that $\lim_{t\to\infty} D\left(p_t/(p_t-1)\right) = 1$. Finally, $\left\| z_k^t \right\|_* \leq \left\| \tilde{z}_k^t \right\|_* \leq (p_t-1)^{1/4}$ and

$$
\begin{aligned}
\lim_{t\to\infty} (p_t-1)^{1/4} \left(D\left(p_t/(p_t-1)\right) - 1\right) &= \lim_{p\to\infty} (p-1)^{1/4} \left(D\left(p/(p-1)\right) - 1\right) \\
&= \lim_{q\to 1} (q-1)^{-1/4} \left(D(q) - 1\right) = 0
\end{aligned}
$$

with $1/p + 1/q = 1$ using again Lemma 2. □

Lemma 1 *We have*

$$
\min_{\lambda\geq 0} \; \phi(q)\lambda \left\| z/\lambda \right\|_*^q + \lambda\epsilon^p = \|z\|_* \epsilon
$$

for any $p > 1$ *and* $q > 1$ *for which* $1/p + 1/q = 1$, $\phi(q) = (q-1)^{q-1}/q^q$ *and* $\epsilon > 0$.

Proof Remark that as the objective function $\lambda \mapsto \phi(q)\lambda \left\| z/\lambda \right\|_*^q + \lambda\epsilon^p$ is continuous and we have $\lim_{\lambda\to 0} \phi(q)\lambda \left\| z/\lambda \right\|_*^q + \lambda\epsilon^p = \lim_{\lambda\to\infty} \phi(q)\lambda \left\| z/\lambda \right\|_*^q + \lambda\epsilon^p = \infty$ as

$\epsilon > 0$ there must exist a minimizer $\lambda^\star \in \min_{\lambda \geq 0}\ \phi(q)\lambda\ \|z/\lambda\|_*^q + \lambda\epsilon^p$ with $\lambda_\star > 0$. The necessary and sufficient first-order convex optimality conditions of the minimization problem guarantee

$$\begin{aligned}
&\lambda^\star \in \min_{\lambda \geq 0}\ \phi(q)\lambda\ \|z/\lambda\|_*^q + \lambda\epsilon^p \iff (1-q)\phi(q)\lambda_\star^{-q}\|z\|_\star^q + \epsilon^p = 0\\
&\iff \epsilon^p = (q-1)\phi(q)\lambda_\star^{-q}\|z\|_\star^q \iff \lambda_\star = [(q-1)\phi(q)]^{1/q}\ \|z\|_\star\epsilon^{-p/q}\\
&\iff \lambda_\star = \frac{q-1}{q}\epsilon^{\frac{1}{1-q}}\|z\|_\star
\end{aligned}$$

where we exploit that $1/p + 1/q = 1$ and $\phi(q) = (q-1)^{q-1}/q^q$. Indeed, we have

$$\begin{aligned}
&[(q-1)\phi(q)]^{1/q} = \left[(q-1)^q/q^q\right]^{1/q} = (q-1)/q,\\
&-\frac{p}{q} = -\frac{1}{\frac{1}{p}q} = -\frac{1}{(1-1/q)q} = -\frac{1}{q-1} = \frac{1}{1-q}.
\end{aligned}$$

Hence, we have

$$\begin{aligned}
&\min_{\lambda \geq 0}\ \phi(q)\lambda^{1-q}\|z\|_\star^q + \lambda\epsilon^p = \ \phi(q)\lambda_\star^{1-q}\|z\|_\star^q + \lambda_\star\epsilon^p\\
&= \ \phi(q)\left[\frac{(q-1)^{1-q}}{q^{1-q}}\epsilon\|z\|_\star^{1-q}\right]\|z\|_\star^q + \left[\frac{q-1}{q}\epsilon^{\frac{1}{1-q}}\|z\|_\star\right]\epsilon^p\\
&= \ \phi(q)\frac{(q-1)^{1-q}}{q^{1-q}}\epsilon\|z\|_\star + \frac{q-1}{q}\epsilon^{p+\frac{1}{1-q}}\|z\|_\star\\
&= \phi(q)\frac{(q-1)^{1-q}}{q^{1-q}}\epsilon\|z\|_\star + \frac{q-1}{q}\epsilon\|z\|_\star\\
&= \ \frac{(q-1)^{q-1}}{q^q}\frac{(q-1)^{1-q}}{q^{1-q}}\epsilon\|z\|_\star + \frac{q-1}{q}\epsilon\|z\|_\star = \frac{1}{q}\epsilon\|z\|_\star + \frac{q-1}{q}\epsilon\|z\|_\star\\
&= \left[\frac{1}{q} + \frac{q-1}{q}\right]\epsilon\|z\|_\star = \epsilon\|z\|_\star
\end{aligned}$$

where we exploit that $1/p + 1/q = 1$ and $\phi(q) = (q-1)^{q-1}/q^q$. Indeed, we have

$$p + \frac{1}{1-q} = \frac{1}{\frac{1}{p}} + \frac{1}{1-q} = \frac{1}{1-\frac{1}{q}} + \frac{1}{1-q} = \frac{-q}{-q+1} + \frac{1}{1-q} = \frac{1-q}{1-q} = 1$$

establishing the claim. □

Lemma 2 *Let* $q > 1$ *then*

$$\max_{t \in [1, 1/\sqrt{q-1}]}\ \frac{1}{q}t^{1-q} + \frac{q-1}{q}t = D(q) = \max\left(1, \frac{1}{q}\frac{1}{(q-1)^{(1-q)/2}} + \frac{\sqrt{q-1}}{q}\right)$$

with $\lim_{q\to 1} D(q) = 1$ *and* $\lim_{q\to 1}(q-1)^{1/4}\,(D(q) - 1) = 0$.

Proof The objective function is convex in t. Convex functions attain their maximum on the extreme points of their domain. The limits can be verified using standard manipulations. □

D Proof of Theorem 4

We prove (i) $\bar{g}^N(x) \leq \bar{g}^K(x)$, (ii) $\bar{g}^K(x) \leq \bar{g}^{N*}(x) + (L/2)D(K)$, and (iii) when the support constraint does not affect the worst-case value, $\bar{g}^K(x) \leq \bar{g}^N(x) + (L/2)D(K)$. *Proof of (i).* We begin with a feasible solution $v_1, \dots, v_N$ of (MRO-K) with $K = N$. Then for $K < N$, we set $u_k = \sum_{i \in C_k} v_i / |C_k|$ for each of the K clusters. We see u_k with $k = 1, \dots, K$ satisfies the constraints of (MRO-K) for $K < N$, as

$$\begin{aligned}
\sum_{k=1}^{K} \frac{|C_k|}{N} \|u_k - \bar{d}_k\|^p &= \sum_{k=1}^{K} \frac{|C_k|}{N} \left\| \frac{\sum_{i \in C_k} v_i}{|C_k|} - \frac{\sum_{i \in C_k} d_i}{|C_k|} \right\|^p \\
&\leq \sum_{k=1}^{K} \frac{|C_k|}{N} \sum_{i \in C_k} \frac{1}{|C_k|} \|v_i - d_i\|^p \\
&= \sum_{k=1}^{K} \frac{1}{N} \sum_{i \in C_k} \|v_i - d_i\|^p \leq \epsilon^p,
\end{aligned}$$

where we have applied triangle inequality, Jensen's inequality for the convex function $f(x) = \|x\|^p$, and the constraint of (MRO-K) with $K = N$. In addition, since the support S is convex, for every k our constructed u_k, as the average of select points v_i $\in S$, must also be within S. The same applies with respect to the domain of g.

Since we have shown that the u_k's satisfies the constraints for (MRO-K) with $K < N$, it is a feasible solution. We now show that for this pair of feasible solutions, in terms of the objective value, $\bar{g}^K(x) \geq \bar{g}^N(x)$. By assumption, g is concave in the uncertain parameter, so by Jensen's inequality,

$$\sum_{k=1}^{K} \frac{|C_k|}{N} g\left(\frac{1}{|C_k|} \sum_{i \in C_k} v_i, x \right) \geq \sum_{k=1}^{K} \frac{|C_k|}{N} \frac{1}{|C_k|} \sum_{i \in C_k} g(v_i)$$

$$\sum_{k=1}^{K} \frac{|C_k|}{N} g(u_k, x) \geq \frac{1}{N} \sum_{i \in N} g(v_i).$$

Since this holds true for u_k's constructed from any feasible solution $v_i, \dots, v_N$, we must have $\bar{g}^K(x) \geq \bar{g}^N(x)$.

Proof of (ii). Next, we prove $\bar{g}^K(x) \leq \bar{g}^{N*}(x) + (L/2)D(K)$ by making use of the L-smooth condition on $-g$. We first solve (MRO-K) with $K < N$ to obtain a feasible solution $u_1, \dots, u_K$. We then set $\Delta_k = u_k - \bar{d}_k$ for each $k \leq K$, and set $v_i = d_i + \Delta_k \quad \forall i \in C_k, k = 1, \dots, K$. These satisfy the constraint of (MRO-K) with

$K = N$ and no support constraints (which we will refer to as $K = N^*$ from here on), as

$$\frac{1}{N}\sum_{i=1}^{N}\|v_i - d_i\|^p = \frac{1}{N}\sum_{k=1}^{K}\sum_{i\in C_k}\|\Delta_k\|^p = \sum_{k=1}^{K}\frac{|C_k|}{N}\|u_k - \bar{d}_k\|^p \le \epsilon^p.$$

Since the constraints are satisfied, the constructed $v_i \dots v_N$ are a valid solution for (MRO-K) with $K = N^*$. We note that these v_i's are also in the domain of g, given that the uncertain data $\mathcal{D}_N$ is in the domain of g. For monotonically increasing functions g, (e.g., $\log(u)$, $1/(1+u)$), we must have $\Delta_k = u_k - \bar{d}_k \ge 0$ in the solution of (MRO-K), as the maximization of g over u_k will lead to $u_k \ge \bar{d}_k$. Therefore, $v_i = d_i + \Delta_k$ is also in the domain, as the L-smooth and concave function g with only a potential lower bound will not have holes in its domain above the lower bound. For monotonically decreasing functions g, the same logic applies with a nonpositive Δ_k. We now make use of the convex and L-smooth conditions [2, Theorem 5.8] on $-g : \forall v_1, v_2 \in S, \lambda \in [0,1]$,

$$g(\lambda v_1 + (1-\lambda)v_2) \le \lambda g(v_1) + (1-\lambda)g(v_2) + \frac{L}{2}\lambda(1-\lambda)\|v_1 - v_2\|_2^2,$$

which, we can apply iteratively, with the first iteration being

$$\begin{aligned} g\left(\frac{1}{|C_k|}v_1 + \frac{|C_k|-1}{|C_k|}\bar{v}_2\right) \le{}& \frac{1}{|C_k|}g(v_1) + \frac{|C_k|-1}{|C_k|}g(\bar{v}_2) \\ &+\frac{L}{2}\frac{1}{|C_k|}\frac{|C_k|-1}{|C_k|}\|v_1 - \bar{v}_2\|_2^2, \end{aligned}$$

where $\bar{v}_2 = \frac{1}{|C_k|-1}\sum_{i\in C_k, i\neq 1} v_i$. Note that $v_1 - \bar{v}_2 = d_1 - \frac{1}{|C_k|-1}\sum_{i\in C_k, i\neq 1} d_i$, as they share the same Δ_k. The next iteration will be applied to $g(\bar{v}_2)$, and so on. For each cluster k, this results in:

$$\begin{aligned} g\left(\frac{1}{|C_k|}\sum_{i\in C_k} v_i, x\right) &\le \frac{1}{|C_k|}\sum_{i\in C_k} g(v_i, x) + \frac{L}{2|C_k|}\sum_{i=2}^{|C_k|}\frac{i-1}{i}\left\|d_i - \frac{\sum_{j=1}^{i-1} d_j}{i-1}\right\|_2^2 \\ g(\bar{d}_k + \Delta_k, x) &\le \frac{1}{|C_k|}\sum_{i\in C_k} g(v_i, x) + \frac{L}{2|C_k|}\sum_{i\in C_k}\|d_i - \bar{d}_k\|_2^2 \\ g(u_k, x) &\le \frac{1}{|C_k|}\sum_{i\in C_k} g(v_i, x) + \frac{L}{2|C_k|}\sum_{i\in C_k}\|d_i - \bar{d}_k\|_2^2, \end{aligned}$$

where we used the equivalence $\sum_{i=2}^{|C_k|}((i-1)/i)\left\|d_i - \sum_{j=1}^{i-1} d_j/(i-1)\right\|_2^2 = \sum_{i \in C_k} \|d_i - \bar{d}_k\|_2^2$. Now, summing over all clusters, we have

$$\sum_{k=1}^{K}(|C_k|/N)g(u_k, x) \leq 1/N \sum_{i=1}^{N} g(v_i, x) + (L/2)D(K).$$

Since this holds for any feasible solution of (MRO-K) with $K < N$, we must have $\bar{g}^K(x) \leq \bar{g}^{N*}(x) + (L/2)D(K)$.

E Proof of Theorem 5

Proof of the lower bound. We use the dual formulations of the MRO constraints. For the case $p \geq 1$, we first solve (MRO-K-Dual) with $K < N$ to obtain dual variables z_{jk}, γ_{jk}. For each data label i in cluster C_k, for all clusters $k = 1, \dots, K$, and for all pieces $j = 1, \dots, J$, if we set

$$\lambda = \lambda, \quad z_{ji} = z_{jk}, \quad \gamma_{ji} = \gamma_{jk}, \quad s_i = s_k + \max_j\{(-z_{jk} - H^T\gamma_{jk})^T(d_i - \bar{d}_k)\},$$

we have obtained a valid solution for (MRO-K-Dual) with $K = N$. The increase in the objective value from $\bar{g}^K(x)$ to that of $\bar{g}^N(x)$, *i.e.* $\bar{g}^N(x) - \bar{g}^K(x)$, is

$$\begin{aligned}\delta(K, z, \gamma) &= (1/N)\sum_{i=1}^{N} s_i - \sum_{k=1}^{K}(|C_k|/N)s_k \\ &= (1/N)\sum_{k=1}^{K}\sum_{i \in C_k}\max_j\{(-z_{jk} - H^T\gamma_{jk})^T(d_i - \bar{d}_k)\}.\end{aligned}$$

We note that $\delta(K, z, \gamma) \geq (|C_k|/N)\sum_{k=1}^{K}(-z_{1k} - H^T\gamma_{1k})^T(1/|C_k|)\sum_{i \in C_k}(d_i - \bar{d}_k) = (|C_k|/N)\sum_{k=1}^{K}(-z_{1k} - H^T\gamma_{1k})^T 0 = 0$. The constructed feasible solution for (MRO-K-Dual) with $K = N$ is an upper bound for its optimal solution, since it is a minimization problem. We then have $\bar{g}^N(x) - \bar{g}^K(x) \leq \delta(K, z, \gamma)$, which translates to $\bar{g}^N(x) - \delta(K, z, \gamma) \leq \bar{g}^K(x)$.

Now, for $p = \infty$, the same procedure can be applied; we obtain a solution to (MRO-K-Dual-∞) with $K < N$, and construct a feasible solution for (MRO-K-Dual-∞) with $K = N$. We modify the variables in the same manner, with the exception of λ, which is now set by $\lambda_i = \lambda_k$. The definition of $\delta(K, z, \gamma)$ remains unchanged, so the same result follows.

Proof of the upper bound. We use the primal formulations of the MRO constraints. Similar to the single-concave proof, for $p \geq 1$, we first solve the MRO problem with K clusters to obtain a feasible solution $u_{11}, \dots, u_{JK}, \alpha_{11}, \dots, \alpha_{JK}$. Note that the existence of α removes the need to analyze which function g_j attains the maximum for each cluster. We then set $\Delta_{jk} = u_{jk} - \bar{d}_k$ for each $k \leq K$, and set $v_{ji} = d_i + \Delta_{jk}$, $\alpha_{ji} = \alpha_{jk}/|C_k| \quad \forall i \in C_k, k = 1, \ldots, K$. These satisfy the constraint of the problem with N clusters and without support constraints, as

$$\sum_{i=1}^{N}\sum_{j=1}^{J} \alpha_{ji} \|v_{ji} - d_i\|^p = \sum_{k=1}^{K}\sum_{j=1}^{J}\sum_{i \in C_k} \alpha_{ji} \|\Delta_{jk}\|^p = \sum_{k=1}^{K}\sum_{j=1}^{J} \alpha_{jk} \|u_{jk} - \bar{d}_k\|^p \leq \epsilon^p.$$

For $p = \infty$, we repeat the process of obtaining and modifying a feasible solution, and observe, for $i = 1, \ldots, N$,

$$\begin{aligned}\sum_{j=1}^{J} (\alpha_{ji}/(1/N)) \|v_{ji} - d_i\| &= \sum_{j=1}^{J} (\alpha_{jk}/(|C_k|/N)) \|\Delta_{jk}\| \\ &= \sum_{j=1}^{J} (\alpha_{jk}/w_k) \|u_{jk} - \bar{d}_k\| \leq \epsilon.\end{aligned}$$

Therefore, we also have constraint satisfaction for $p = \infty$. The constructed solutions for both cases remain in $\mathbf{dom}_u\, g$, following the arguments in the proof of (ii) in Appendix D.

Next, the objective functions for $p \geq 1$ and $p = \infty$ are identical, so the same analysis applies. For each cluster k and function g_j, using the L-smooth condition on $-g_j$, we observe

$$\begin{aligned}\alpha_{jk} g_j \left(\frac{1}{|C_k|} \sum_{i \in C_k} v_{ji}, x \right) &\leq \alpha_{jk} \left(\sum_{i \in C_k} \frac{1}{|C_k|} g_j(v_{ji}, x) \right. \\ &\quad \left. + \frac{L_j}{2|C_k|} \sum_{i=2}^{|C_k|} \frac{i-1}{i} \left\| d_i - \frac{\sum_{j=1}^{i-1} d_j}{i-1} \right\|_2^2 \right) \\ \alpha_{jk} g_j(\bar{d}_k + \Delta_k, x) &\leq \sum_{i \in C_k} \alpha_{ji} g_j(v_{ji}, x) + \frac{\alpha_{ji} L_j}{2} \sum_{i \in C_k} \|d_i - \bar{d}_k\|_2^2.\end{aligned}$$

Then, summing over all the clusters and functions, we have

$$\begin{aligned}\sum_{k=1}^{K}\sum_{j=1}^{J} \alpha_{jk} g_j(u_{jk}, x) &\leq \sum_{k=1}^{K}\sum_{j=1}^{J}\sum_{i \in C_k} \alpha_{ji} g_j(v_{ji}, x) \\ &\quad + \sum_{k=1}^{K}\sum_{j=1}^{J} \frac{\alpha_{ji} \max_{j \leq J} L_j}{2} \sum_{i \in C_k} \|d_i - \bar{d}_k\|_2^2 \\ &\leq \sum_{i=1}^{N}\sum_{j=1}^{J} \alpha_{ji} g(v_{ji}, x) + \max_{j \leq J}(L_j/2N) \sum_{i=1}^{N} \|d_i - \bar{d}_k\|_2^2.\end{aligned}$$

Since this holds for all feasible solutions of the problem with K clusters, we conclude that

$$\bar{g}^K(x) \leq \bar{g}^{N*}(x) + \max_{j \leq J}(L_j/2)D(K).$$

Acknowledgements The simulations presented in this article were performed on computational resources managed and supported by Princeton Research Computing, a consortium of groups including the Princeton Institute for Computational Science and Engineering (PICSciE) and the Office of Information Technology's High Performance Computing Center and Visualization Laboratory at Princeton University. We would like to thank Daniel Kuhn for the useful feedback and for pointing us to the literature on scenario reduction techniques.

Declarations

Conflict of interest The authors have no relevant financial or non-financial interests to disclose.

References

1. Bandi, C., Bertsimas, D.: Tractable stochastic analysis in high dimensions via robust optimization. Math. Program. **134**(1), 23–70 (2012)
2. Beck, A.: First-Order Methods in Optimization. SIAM-Society for Industrial and Applied Mathematics, Philadelphia (2017)
3. Ben-Tal, A., El Ghaoui, L., Nemirovski, A.: Robust Optimization. Princeton University Press, Princeton (2009)
4. Ben-Tal, A., den Hertog, D., Vial, J.P.: Deriving robust counterparts of nonlinear uncertain inequalities. Math. Program. **149**(1–2), 265–299 (2015)
5. Ben-Tal, A., Nemirovski, A.: Robust solutions of Linear Programming problems contaminated with uncertain data. Math. Program. **88**(3), 411–424 (2000)
6. Ben-Tal, A., Nemirovski, A.: Selected topics in robust convex optimization. Math. Program. **112**, 125–158 (2008)
7. Beraldi, P., Bruni, M.E.: A clustering approach for scenario tree reduction: an application to a stochastic programming portfolio optimization problem. TOP **22**(3), 934–949 (2014)
8. Bertsimas, D., Brown, D.B., Caramanis, C.: Theory and applications of robust optimization. SIAM Rev. **53**(3), 464–501 (2011)
9. Bertsimas, D., den Hertog, D., Pauphilet, J.: Probabilistic guarantees in robust optimization. SIAM J. Optim. **31**(4), 2893–2920 (2021)
10. Bertsimas, D., Gupta, V., Kallus, N.: Data-driven robust optimization. Math. Program. **167**(2), 235–292 (2018)
11. Bertsimas, D., den Hertog, D.: Robust and Adaptive Optimization. Dynamic Ideas, Belmont (2022)
12. Bertsimas, D., Mundru, N.: Optimization-based scenario reduction for data-driven two-stage stochastic optimization. Oper. Res. (2022)
13. Bertsimas, D., Sim, M.: The price of robustness. Oper. Res. **52**(1), 35–53 (2004)
14. Bertsimas, D., Sturt, B., Shtern, S.: A data-driven approach to multistage stochastic linear optimization. Manag. Sci. (2022)
15. Bradley, B.O., Taqqu, M.S.: Financial Risk and Heavy Tails, Handbooks in Finance, vol. 1. North-Holland, Amsterdam (2003)
16. Carmona, R.A.: Rsafd: Statistical Analysis of Financial Data in R (2020). R package version 1.2
17. Chen, L.: Clustering of sample average approximation for stochastic program (2015)
18. Chen, R., Paschalidis, I.: Distributionally robust learning. Found. Trends Optim. **4**(1–2), 1–243 (2020)
19. Chen, Z., Sim, M., Xiong, P.: Robust stochastic optimization made easy with RSOME. Manag. Sci. **66**(8), 3329–3339 (2020)

20. Clement, P., Desch, W.: An elementary proof of the triangle inequality for the Wasserstein metric. Proc. Am. Math. Soc. **136**(1), 333–339 (2008)
21. Delage, E., Ye, Y.: Distributionally robust optimization under moment uncertainty with application to data-driven problems. Oper. Res. **58**(3), 595–612 (2010)
22. Dupačová, J., Gröwe-Kuska, N., Römisch, W.: Scenario reduction in stochastic programming. Math. Program. **95**(3), 493–511 (2003)
23. Emelogu, A., Chowdhury, S., Marufuzzaman, M., Bian, L., Eksioglu, B.: An enhanced sample average approximation method for stochastic optimization. Int. J. Prod. Econ. **182**, 230–252 (2016)
24. Esfahani, P.M., Kuhn, D.: Data-driven distributionally robust optimization using the Wasserstein metric: performance guarantees and tractable reformulations. Math. Program. **171**, 115–166 (2018)
25. Esteban, P.A., Morales, J.M.: Partition-based distributionally robust optimization via optimal transport with order cone constraints. 4OR **20**(3), 465–497 (2022)
26. Fabiani, F., Goulart, P.: The optimal transport paradigm enables data compression in data-driven robust control. In: 2021 American Control Conference (ACC), pp. 2412–2417 (2021)
27. Fournier, N., Guillin, A.: On the rate of convergence in Wasserstein distance of the empirical measure. Probab. Theory Relat. Fields **162**(3), 707–738 (2015)
28. Gao, R.: Finite-sample guarantees for Wasserstein distributionally robust optimization: Breaking the curse of dimensionality. CoRR (2020)
29. Gao, R., Kleywegt, A.: Distributionally robust stochastic optimization with Wasserstein distance. Math. Oper. Res. **48**, 603–655 (2023)
30. Givens, C.R., Shortt, R.M.: A class of Wasserstein metrics for probability distributions. Mich. Math. J. **31**(2), 231–240 (1984)
31. Goh, J., Sim, M.: Distributionally robust optimization and its tractable approximations. Oper. Res. **58**(4–part–1), 902–917 (2010)
32. Hartigan, J.A., Wong, M.A.: Algorithm AS 136: A k-means clustering algorithm. J. R. Stat. Soc. Ser. C (Appl. Stat.) **28**(1), 100–108 (1979)
33. Holmberg, K., Rönnqvist, M., Yuan, D.: An exact algorithm for the capacitated facility location problems with single sourcing. Eur. J. Oper. Res. **113**(3), 544–559 (1999)
34. Jacobson, D., Hassan, M., Dong, Z.S.: Exploring the effect of clustering algorithms on sample average approximation. In: 2021 Institute of Industrial and Systems Engineers (IISE) Annual Conference & Expo (2021)
35. Kuhn, D., Esfahani, P.M., Nguyen, V., Shafieezadeh-Abadeh, S.: Wasserstein Distributionally Robust Optimization: Theory and Applications in Machine Learning, pp. 130–166 (2019)
36. Liu, Y., Yuan, X., Zhang, J.: Discrete approximation scheme in distributionally robust optimization. Numer. Math. Theory Methods Appl. **14**(2), 285–320 (2021)
37. MOSEK ApS: The MOSEK optimization toolbox. Version 9.3. (2022)
38. Neumann, J.V.: Zur theorie der gesellschaftsspiele. Math. Ann. **100**, 295–320 (1928)
39. Perakis, G., Sim, M., Tang, Q., Xiong, P.: Robust pricing and production with information partitioning and adaptation. Manag. Sci. (2023)
40. Rockafellar, R.T., Uryasev, S.: Conditional value-at-risk for general loss distributions. J. Bank. Finance **26**(7), 1443–1471 (2002)
41. Rockafellar, R.T., Wets, R.J.: Variational analysis. Grundlehren der mathematischen Wissenschaften (1998)
42. Roos, E., den Hertog, D.: Reducing conservatism in robust optimization. INFORMS J. Comput. **32**(4), 1109–1127 (2020)
43. Rujeerapaiboon, N., Schindler, K., Kuhn, D., Wiesemann, W.: Scenario reduction revisited: fundamental limits and guarantees. Math. Program. **191**(1), 207–242 (2022)
44. Thorndike, R.: Who belongs in the family? Psychometrika **18**(4), 267–276 (1953)
45. Trillos, N., Slepčev, D.: On the rate of convergence of empirical measures in ∞-transportation distance. Can. J. Math. **67**, 1358–1383 (2014)
46. Uryasev, S., Rockafellar, R.T.: Conditional Value-at-Risk: Optimization Approach. Springer, New York (2001)
47. Wang, I., Becker, C., Van Parys, B., Stellato, B.: Mean robust optimization. arXiv (2023)
48. Wang, Z., Wang, P., Ye, Y.: Likelihood robust optimization for data-driven problems. CMS **13**(2), 241–261 (2016)
49. Wiesemann, W., Kuhn, D., Sim, M.: Distributionally robust convex optimization. Oper. Res. **62**(6), 1358–1376 (2014)

50. Xu, H., Caramanis, C., Mannor, S.: A distributional interpretation of robust optimization. Math. Oper. Res. **37**(1), 95–110 (2012)
51. Zhen, J., Kuhn, D., Wiesemann, W.: A unified theory of robust and distributionally robust optimization via the primal-worst-equals-dual-best principle (2021). https://doi.org/10.1287/opre.2021.0268
52. Zymler, S., Kuhn, D., Rustem, B.: Distributionally robust joint chance constraints with second-order moment information. Math. Program. **137**(1), 167–198 (2013)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.