Mean Robust Optimization

Irina Wang, Cole Becker, Bart Van Parys, and Bartolomeo Stellato

November 7, 2023

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.

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 (Roos and den Hertog, 2020). 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 for both paradigms. 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 (Ben-Tal and Nemirovski, 2000; Bandi and Bertsimas, 2012; Ben-Tal et al., 2009; Bertsimas and Sim, 2004). 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. This change 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 (Bertsimas et al., 2018). 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 $\epsilon > 0$ Esfahani and Kuhn (2018); Kuhn et al. (2019); Gao (2020); Gao and Kleywegt (2023). 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 datapoints (Esfahani and Kuhn, 2018). 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 (Jacobson et al., 2021; Emelogu et al., 2016; Beraldi and Bruni, 2014; Chen, 2015), 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 (Chen et al., 2020; Perakis et al., 2023), or enrich Wasserstein DRO formulations with partition-specific information (*e.g.*, relative weights) Esteban and Morales (2022). 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

In this work, 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 just 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 be within 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 Ben-Tal and Nemirovski (2008) and Bertsimas et al. Bertsimas et al. (2011), as well as the books by Ben-Tal et al. Ben-Tal et al. (2009) and Bertsimas and den Hertog Bertsimas and den Hertog (2022). These approaches, while powerful, may be overly-conservative, and there have been approaches that provide a tradeoff between conservatism and constraint violation (Roos and den Hertog, 2020).

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 considered existing literature on DRO can roughly be defined in two. Ambiguity sets of the first type contain all distributions that satisfy certain moment constraints (Zymler et al., 2013; Wiesemann et al., 2014; Delage and Ye, 2010; Goh and Sim, 2010). In many cases such ambiguity sets possess a tractable formulation, but have also been criticized for yielding overly conservative solutions (Wang et al., 2016). 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 (Esfahani and Kuhn, 2018; Kuhn et al., 2019; Gao, 2020; Gao and Kleywegt, 2023) and enjoy both a tractable primal as well as a tractable dual formulation. We refer to the work by Chen and Paschalidis Chen and Paschalidis (2020) for a thorough overview of DRO, and to the work by Zhen et al. Zhen et al. (2021) 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 typically chosen through time consuming cross-validation (Gao, 2020). 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 Chen et al. (2020); Esteban and Morales (2022); Perakis et al. (2023), or the discretization of the underlying distribution Liu et al. (2021), 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. Bertsimas et al. (2018) 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 Gao and Kleywegt (2023) 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, instead of increase, the number of variables and constraints to make the Wasserstein DRO problem more tractable by linking it to robust optimization.

Robust optimization as a distributionally robust optimization program. Xu et al. Xu et al. (2012) take inspiration from sample-based optimization problems to investigate probabilistic interpretations of RO. They generalize the ideas of Delage and Ye Delage and Ye (2010), that the solution to a robust optimization problem is the solution to a special Distributionally Robust Stochastic Program (DRSP), where the distributional set contains all distributions whose support is contained in the uncertainty set. In a related vein, Bertsimas et al. Bertsimas et al. (2022) 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, in view of both the primal and dual problems, that $p = \infty$ is a limiting case of $p \geq 1$. This provides a natural extension of the equivalence proved in (Bertsimas et al., 2022, Proposition 3).

Probabilistic guarantees in robust and distributionally optimization. Bertsimas et al. Bertsimas et al. (2021) 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 Esfahani and Kuhn (2018) 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 Gao (2020) 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 overly-conservative ϵ. The final ϵ values are usually chosen through empirical empirication - 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. Dupačová et al. (2003), 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. Rujeerapaiboon et al. (2022) 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 (Hartigan and Wong, 1979). Later, Bertsimas and Mundru Bertsimas and Mundru (2022) 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. (Jacobson et al., 2021), Emelogu et al. Emelogu et al. (2016), Beraldi et al. Beraldi and Bruni (2014), and Chen Chen (2015) 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 Fabiani and Goulart (2021) 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.

1.3 Layout of the paper

In Section 2, we present our approach for concave uncertainty constraints, then extend the results to maximum-of-concave functions. In Section 3, we present connections to distributionally robust optimization, and give theoretical guarantees on constraint satisfaction. In Section 4, we analyze the effect on clustering on the worst-case value of the MRO solutions for both concave and maximum-of-concave constraints. In Section 5, we give guidelines for choosing hyperparameters. In Section 6, we provide computational verification of the speedups obtained through our methodology. In Section 7, we summarize our conclusions.

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 2.1. *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 Section 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

$$g(u, x) \leq 0, \quad \forall u \in \mathcal{U},$$

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,

$$\mathbf{E}^{\mathbf{P}}(g(u, x)) \leq 0, \tag{2}$$

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

Risk measures. Expectation constraints of the form (2) can represent popular risk measures, and can imply constraints commonly used in chance-constrained programming (CCP). In CCP, the probabilistic constraint considered is

$$\mathbf{P}(g(u,x) \leq 0) \geq 1 - \alpha, \tag{3}$$

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 (Uryasev and Rockafellar, 2001). A tractable approximation of the value at risk is the *conditional value at risk* (Uryasev and Rockafellar, 2001; Rockafellar and Uryasev, 2002), 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 Sections 2.4, 6.4, and 6.5. It is well known from Uryasev and Rockafellar (2001) 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 the training dataset of these samples by $\mathcal{D}_N = \{d_i\}_{i \leq N} \subseteq S$, and note that this dataset 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 $\mathcal{U}$ we construct must imply the probabilistic guarantee

$$\mathbf{P}^N\left(\mathbf{E}^{\mathbf{P}}(g(u,\hat{x}_N)) \leq 0\right) \geq 1-\beta, \tag{4}$$

where $\beta > 0$ is the specified probability of constraint violation. From now on, when we refer to probabilistic guarantees of constraint satisfaction, it will be a reference to (4).

2.2 Our approach

To meet the probabilistic guarantees outlined above, we propose to construct $\hat{x}_N$ to satisfy particular constraints, with respect to a particular uncertainty set.

Case $p \geq 1$. In the case where $p \geq 1$, the set we consider takes the form

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

where we partition $\mathcal{D}_N$ into K disjoint subsets C_k, and $\bar{d}_k$ is the centroid of the kth 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 (4). When $p = 2$ and $S = \mathbf{R}^m$, the set can be visualized as 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, as shown in Figure 1.

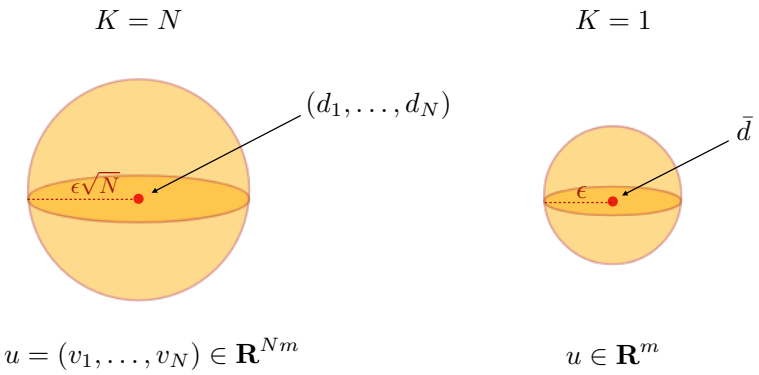

Figure 1: Visualizing the uncertainty set $\mathcal{U}(N, \epsilon)$ and $\mathcal{U}(1, \epsilon)$ as high dimension balls when $p = 2$.

Case $p = \infty$. In the case where $p = \infty$, the set we consider takes a more specific form,

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

where the constraints for individual v_k become decoupled. See Figure 2 for an example when $K = 3$ and $K = 1$. This decoupling follows the result for the Wasserstein type $p = \infty$ metric (Givens and Shortt, 1984, Equation 2), as our uncertainty set is analogous to the set of all distributions within Wasserstein-∞ distance of $\bar{d}$. We note that, if any of the decoupled constraints are violated, then $\lim_{p\to\infty} \sum_{k=1}^{K} w_k \|v_k - \bar{d}_k\|^p \geq \epsilon^p$, and the summation constraint will be violated.

For both cases, $p \geq 1$ and $p = \infty$, when $K = 1$, we have a simple uncertainty set: a ball of radius ϵ around the empirical mean of the entire dataset, $\mathcal{U}(1, \epsilon) = \left\{ v \in S \mid \|v - \bar{d}\| \leq \epsilon \right\}$. This is equivalent to the uncertainty set of traditional RO, as it is of the same dimension m as the uncertain parameter. When $K = N$ and $w_k = 1/N$, both cases closely resemble the ambiguity sets of Wasserstein-p DRO.

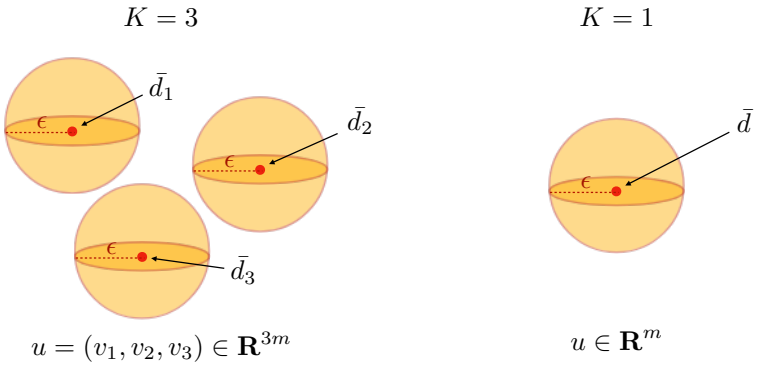

Figure 2: Visualizing the decoupled uncertainty set $\mathcal{U}(K, \epsilon)$ with $p = \infty$.

Having defined the uncertainty set, we now introduce constraints of the form

$$\bar{g}(u, x) = \sum_{k=1}^{K} w_k g(v_k, x), \tag{5}$$

where g is defined in the original constraint (1). The weights w_k correspond to the ones defined in the uncertainty set. Putting everything together, $\hat{x}_N$ 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.

Data-driven procedure. 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 (4), see Figure 3.

2.3 Solving the robust problem

We now outline two ways to solve the MRO problem, using a direct convex reformulation and using a cutting plane algorithm. The reformulation and cutting plane procedure follow usual techniques for RO problems in existing literature (Ben-Tal et al., 2009; Kuhn et al., 2019; Bertsimas and den Hertog, 2022; Bertsimas et al., 2016), with adaptations made for the MRO setup, as well as a reformulation derived for the case $p = \infty$. We include simple examples and pseudocode for completeness.

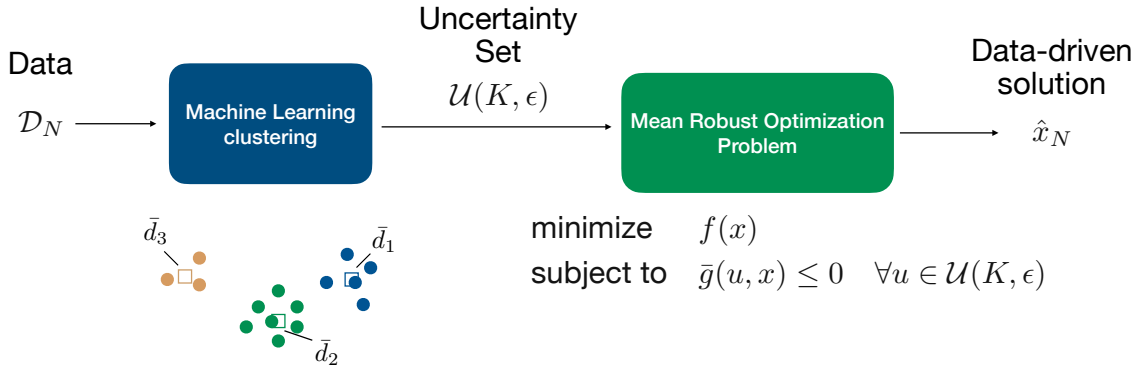

Figure 3: Mean robust optimization procedure.

2.3.1 Direct convex reformulation for $p \geq 1$

In the case where $p \geq 1$, the MRO can be rewritten as the optimization problem

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & \left\{ \begin{array}{ll} \underset{v_1,\ldots,v_K \in S}{\text{maximize}} & \sum_{k=1}^K w_k g(v_k, x) \\ \text{subject to} & \sum_{k=1}^K w_k \|v_k - \bar{d}_k\|^p \leq \epsilon^p \end{array} \right\} \leq 0, \end{array} \tag{6}$$

which, by dualizing the inner maximization problem, has the following reformulation:

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & \sum_{k=1}^K w_k s_k \leq 0 \\ & [-g]^*(z_k - y_k, x) + \sigma_S(y_k) - z_k^T \bar{d}_k + \phi(q)\lambda \|z_k/\lambda\|_*^q + \lambda \epsilon^p \leq s_k \\ & \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad k = 1, \ldots, K \\ & \lambda \geq 0, \end{array} \tag{7}$$

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 $q > 1$ (Kuhn et al., 2019, Theorem 8). Note that q satisfies $1/p + 1/q = 1$, *i.e.*, $q = p/(p-1)$. When $p = 1$ and $q = \infty$, we note the formulation in (8). 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.1. 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 (Bertsimas and den Hertog, 2022, Section 2.5) and Beck (Beck, 2017, 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 Section 3.

We note the special case when $p = 1$. We observe from (Kuhn et al., 2019, Section 2.2 Remark 1) that

$$\lim_{q\to\infty} \phi(q)\lambda \left\|z_k/\lambda\right\|_*^q = \begin{cases} 0 & \text{if } \|z_k\| \le \lambda \\ \infty & \text{otherwise.} \end{cases}$$

Therefore, the above formulation becomes

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

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

$$[-g]^*(z, x) = \sup_u z^T u + b - (a + Pu)^T x = \begin{cases} a^T x - b & \text{if } z + P^T x = 0 \\ \infty & \text{otherwise.} \end{cases} \tag{10}$$

To substitute $\sigma_S(y_k)$ and $[-g]^*(z_k - y_k, x)$ into (7), 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 (7), we arrive at the form

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

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 datapoints in $\mathcal{D}_N$ for any $K \in \{1, \ldots, N\}$, this formulation corresponds to always choosing $K = 1$.

2.3.2 Direct convex reformulation for $p = \infty$

In the case where $p = \infty$, the MRO can be rewritten as the optimization problem

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & \left\{ \begin{array}{ll} \underset{v_1,\ldots,v_K\in S}{\text{maximize}} & \sum_{k=1}^K w_k g(v_k, x) \\ \text{subject to} & \|v_k - \bar{d}_k\| \le \epsilon, \quad k = 1, \ldots, K \end{array} \right\} \le 0, \end{array} \tag{12}$$

which has a reformulation where the constraint above is dualized,

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & \sum_{k=1}^{K} w_k s_k \leq 0 \\ & [-g]^*(z_k - y_k, x) + \sigma_S(y_k) - z_k^T \bar{d}_k + \lambda_k \epsilon \leq s_k \\ & \qquad\qquad k = 1, \ldots, K \\ & \|z_k\|_* \leq \lambda_k \quad k = 1, \ldots, K, \end{array} \tag{13}$$

with new variables $s_k \in \mathbf{R}$, $z_k \in \mathbf{R}^m$, and $y_k \in \mathbf{R}^m$. The proof is delayed to Appendix A.2.

Remark 2.1 (Case $p = \infty$ is the limit of case $p \geq 1$)**.** *In terms of the primal problem,* (12) *is the limiting case of* (6) *as* $p \to \infty$*. In terms of the reformulated problem with dualized constraints, problem* (13) *is the limiting case of* (7) *as* $p \to \infty$ *. The proofs are delayed to Appendix A.5 and Appendix A.6 respectively. These proofs extend the ideas stated in (Bertsimas et al., 2022, Proposition 3).*

Example with affine constraints. Consider again the case of affine constraint as in (9) with support set $S = \mathbf{R}^m$, now with $p = \infty$. Following a similar derivation as (11), we substitute the conjugate function $[-g]^*$ (10) in problem (13), we can obtain

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & a^T x - b + \epsilon \left\| P^T x \right\|_* + (P^T x)^T \sum_{k=1}^{K} w_k \bar{d}_k \leq 0, \end{array} \tag{14}$$

where the number of constraints and variables does not depend on K. Similairy to problem (11), the term $\sum_{k=1}^{K} w_k \bar{d}_k$ is the average of the datapoints in $\mathcal{D}_N$ for any $K \in \{1, \ldots, N\}$. Therefore, the choice of K does not affect this formulation. This can be viewed as the robust counterpart when the uncertainty set is a norm ball of radius ϵ centered at $(1/N) \sum_{i=1}^{N} d_i$

Note that, 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 (Bertsimas and den Hertog, 2022, Section 2.3), (Ben-Tal et al., 2009, Chapter 2).

Remark 2.2. *When* g *is affine and* $S = \mathbf{R}^m$*, for any* ϵ *and norm, the convex reformulations for* $p = 1$ *and* $p = \infty$ *are identical. The proof appears in Appendix A.7.*

2.3.3 Cutting plane algorithm

The second approach to solve problem (MRO) is to use a cutting plane procedure, in which we consider the minimization problem where x is the variable and S a finite set of values for the uncertainty,

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & \bar{g}(u, x) \leq 0, \quad \forall u \in \hat{S}, \end{array} \tag{15}$$

and the maximization problem over u with x^k fixed,

$$\begin{array}{ll} \text{maximize} & \bar{g}(u, x^k) \\ \text{subject to} & u \in \mathcal{U}(K, \epsilon) \end{array} \tag{16}$$

The procedure works as follows. We first solve (15) with a set $\hat{S} = \{\bar{u}\}$, where $\bar{u}$ is nominal value of the uncertainty, obtaining x^k. Then, we solve (16), obtaining u^k. If $\bar{g}(u^k, x^k) > 0$, then we add u^k to the set $\hat{S}$. Otherwise, we terminate. This procedure is summarized in Algorithm 1. As demonstrated by Bertimas et al. Bertsimas et al. (2016), the cutting plane and convex reformulation methods are comparable in terms of performance, thus both are viable.

Algorithm 1 Cutting plane algorithm to solve (MRO)

given $\hat{S} = \{\bar{u}\}$
for $k = 1, \ldots, k_{\max}$ **do**
    $x^k \leftarrow$ solve minimization problem (15) over x
    $u^k \leftarrow$ solve maximization problem (16) over u
    if $\bar{g}(u^k, x^k) > 0$ **then**
        $\hat{S} \leftarrow \hat{S} \cup \{u^k\}$
    else
        return x^k

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 Section 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 Section 2.1, this can also be used to model **CVaR** constraints, which has a maximum-of-concave analytical form.

Problem parametrization. We now consider constraints of the form

$$\bar{g}(u, x) = \sum_{k=1}^{K} \sum_{j=1}^{J} \alpha_{jk} g_j(v_{jk}, x), \tag{17}$$

where $\alpha \in \Gamma$, 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 contains a set of vectors $(v_{j1}, \ldots, v_{jK})$, and a set of parameters

$(\alpha_{j1}, \ldots, \alpha_{jK})$ to denote the fraction of mass assigned to that function for each k. The total amount of mass assigned for each cluster, $\sum_{j=1}^{J} \alpha_{jk}$, is the weight of the cluster, w_k.

We use a summation over weighted pieces g_j instead of a maximum over g_j, as this is a generalization of the maximum, and has a more natural dual reformulation. We take inspiration from (Esfahani and Kuhn, 2018, Section 4.2), where α arises from the extremal distribution for Wasserstein DRO. Note that the intuitive maximization over g_j's is analogous to setting $\alpha_{jk} = w_k$ for a specific j for each k, and $\alpha_{jk} = 0$ otherwise.

The uncertainty set is given as follows.

Case $p \geq 1$. In the case where $p \geq 1$, we have

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

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

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

where we once again introduce weight parameters α.

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

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 A.4. We write the MRO problem as the optimization problem

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & \left\{ \begin{array}{ll} \underset{v_{11},\ldots,v_{JK} \in S, \alpha \in \Gamma}{\text{maximize}} & \sum_{k=1}^{K} \sum_{j=1}^{J} \alpha_{jk} g_j(v_{jk}, x) \\ \text{subject to} & \sum_{k=1}^{K} \sum_{j=1}^{J} \alpha_{jk} \|v_{jk} - \bar{d}_k\|^p \leq \epsilon^p \end{array} \right\} \leq 0, \end{array} \tag{18}$$

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

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & \sum_{k=1}^{K} w_k s_k \leq 0 \\ & [-g_j]^*(z_{jk} - y_{jk}, x) + \sigma_S(y_{jk}) - z_{jk}^T \bar{d}_k + \phi(q) \lambda \left\| z_{jk}/\lambda \right\|_*^q + \lambda \epsilon^p \leq s_k \\ & \qquad\qquad\qquad\qquad\qquad\qquad k = 1, \ldots, K, \quad j = 1, \ldots, J \\ & \lambda \geq 0, \end{array} \tag{19}$$

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.3. Again, while traditionally we take the supremum instead of maximizing, here the supremum is always achieved as we assume g_j to be upper-semicontinuous for all j.

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 Section 3.

3 Links to Wasserstein distributionally robust optimization

Distributionally robust optimization (DRO) solves the problem

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & \sup_{\mathbf{Q}\in\mathcal{P}_N} \mathbf{E}^{\mathbf{Q}}(g(u,x)) \leq 0, \end{array} \tag{20}$$

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 (4) 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 section 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 (20) is equivalent to the constraint of (MRO). In particular, for case $p \geq 1$, the expression

$$\sup_{\mathbf{Q}\in\mathbf{B}_\ell^p(\hat{\mathbf{P}}^N)} \mathbf{E}^{\mathbf{Q}}(g(u,x)), \tag{21}$$

is equivalent to the dual of the constraint of (18), when $K = N$, and $w_k = 1/N$. This is noted in (Kuhn et al., 2019; Zhen et al., 2021). We give a proof of strong duality in Appendix A.8. This is the dual of the generalized max-of-concave form, which is equivalent to the dual of the single concave form (6) when $J = 1$. By the same logic, in the case where $p = \infty$, the expression is equivalent to the dual of the constraint of (12). Given the above reductions, we can rewrite the Wasserstein DRO problem in the same form as (18), the MRO problem.

Our approach can then be viewed as a form of Wasserstein DRO, with the difference that, when $K < N$, we deal with the clustered and averaged 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

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

Case $p \geq 1$. Wasserstain DRO satisfies (4) if the data-generating distribution, supported on a convex and closed set S, satisfies a *light-tailed assumption* (Fournier and Guillin, 2015; Esfahani and Kuhn, 2018): 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 3.1 (Measure concentration (Fournier and Guillin, 2015, 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 (3.1) 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 (Esfahani and Kuhn, 2018, Theorem 3.5).

Case $p = \infty$. When $p = \infty$, Bertsimas et al. (Bertsimas et al., 2022, Section 6) note that the light-tailed assumption is no longer sufficient. Wasserstein DRO satisfies (4) under stronger assumptions, as given in the following theorem.

Theorem 3.2 (Measure concentration, $\mathbf{p} = \infty$ (Trillos and Slepčev, 2014, 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 3.3 (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* (4) *holds.*

Proof. Compared with Wasserstein DRO, MRO has to account for the additional difference between the two empirical distributions $\hat{\mathbf{P}}^N$ and $\hat{\mathbf{P}}^K$. We can write

$$\hat{\mathbf{P}}^N = \sum_{i=1}^{N} \frac{1}{N} \delta_{d_i} = \sum_{k=1}^{K} \sum_{i \in C_k} \frac{|C_k|}{N} \frac{1}{|C_k|} \delta_{d_i},$$
$$\hat{\mathbf{P}}^K = \sum_{k=1}^{K} w_k \delta_{\bar{d}_k} = \sum_{i=1}^{K} \frac{|C_k|}{N} \delta_{\bar{d}_k}.$$

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

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

where we have replaced the integral with a finite sum, as the distributions are discrete. Therefore, by Theorems 3.1, 3.2 and the triangle inequality for the Wasserstein metric Clement and Desch (2008),

$$\begin{aligned}
W_p(\mathbf{P}, \hat{\mathbf{P}}^K) &\leq W_p(\mathbf{P}, \hat{\mathbf{P}}^N) + W_p(\hat{\mathbf{P}}^K, \hat{\mathbf{P}}^N) \\
&\leq \epsilon_N(\beta) + \eta_N(K)^{1/p},
\end{aligned}$$

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

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

which implies the uncertainty set $\mathcal{U}(K, \epsilon_N(\beta) + \eta_N(K)^{1/p})$ contains all possible realizations of uncertainty with probability $1 - \beta$, so the finite sample guarantee (4) holds. ■

4 Worst-case value of the uncertain constraint

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 then cluster data-points such that the average distance of the points in each cluster to their data-center is minimized,

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

where $\bar{d}_k$ is the mean of the points in cluster C_k. A well-known algorithm is K-means (Hartigan and Wong, 1979), where we create K clusters by iteratively solving a least-squares problem. Note that once the clusters have been selected, and we assume it to be optimal (*i.e.* attain $D(K)$), then for the case $p = 2$, we have $\eta_N(K) = D(K)$ from Theorem 3.3.

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. For the latter, we also include the special case of the maximum-of-affine function.

4.1 Single concave function

For the simplest case of a single concave function, we prove that when the support is large enough,

- If g is affine in u, MRO does not increase the worst-case value, regardless of K.
- If g is concave in u and satisfies certain smoothness conditions, MRO has a higher worst-case value than Wasserstein DRO and the increase is inversely related to the number of clusters K. In other words, the smaller the K, the higher the worst-case value.

Quantifying the clustering effect. To quantify the effect of clustering, we calculate the difference between the following formulations of the worst-case value of the constraint in (MRO)

$$\begin{aligned} \bar{g}^N(x) = \underset{v_1 \ldots v_N}{\text{maximize}} \quad & \frac{1}{N} \sum_{i=1}^{N} g(v_i, x) \\ \text{subject to} \quad & \frac{1}{N} \sum_{i=1}^{N} \|v_i - d_i\|^p \leq \epsilon^p \\ & v_i \in S \quad i = 1, \ldots, N, \end{aligned} \tag{MRO-N}$$

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

and

$$\begin{aligned} \bar{g}^{N*}(x) = \underset{v_1 \ldots v_N}{\text{maximize}} \quad & \frac{1}{N} \sum_{i=1}^{N} g(v_i, x) \\ \text{subject to} \quad & \frac{1}{N} \sum_{i=1}^{N} \|v_i - d_i\|^p \leq \epsilon^p, \end{aligned} \tag{MRO-N*}$$

where (MRO-N) is the formulation of the constraint without clustering, akin to traditional Wasserstein DRO, (MRO-N*) is the same, except we drop the support constraint, and (MRO-K) is the formulation with K clusters. 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 $v_i \in S$ for $i = 1, \ldots, N$ 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

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

If the support satisfies this condition, then 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 datapoints, it is acceptable to have this dependency, as this is a condition we can check given data to potentially improve the following bounds, without having to solve the MRO problem.

With these definitions, we can construct solutions for (MRO-N), (MRO-K), and (MRO-N*) to prove the following relations.

Theorem 4.1. *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 2.1 holds, and $-g$ satisfies an L-smooth condition on its domain with respect to the ℓ_2-norm and for a given x,

$$\|\nabla g(v, x) - \nabla g(u, x)\|_2 \leq L \|u - v\|_2.$$

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 A.9. The results also hold for $p = \infty$, as we have shown in Remark 2.1 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, *i.e.*, $\Delta = \max_{x \in \mathcal{X}} \left(\bar{g}^{N*}(x) - \bar{g}^{N}(x) \right)$, subject to x being feasible for problem (MRO). As we assume Assumption 2.1 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 $\mathbf{dom}_u$ of g. Refer to a similar argument in Appendix A.9 (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 we calculate it with only $\bar{g}^{N*}(x)$ and $\bar{g}^{N}(x)$.

Remark 4.1. *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 then approximate $\Delta \approx 0$ and simply use the upper bound $(L/2)D(K)$, as this bound is often not tight. See Sections 6.5 and 6.1 for examples.*

Uncertain objective. When the uncertainty is in the objective, Theorem 4.1 quantifies the difference in optimal values.

Corollary 4.1.1. *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 4.1.2 (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)*.*

Proof. In view of the primal problem and constraints, from Theorem 4.1, if $g(u, x)$ is affine in u and the support does not affect the uncertainty set, $\bar{g}^{N}(x) = \bar{g}^{K}(x)$. So for some fixed

$\hat{x}$ we have $\bar{g}^K(\hat{x}) \leq 0 \iff \bar{g}^N(\hat{x}) \leq 0$. Therefore,

$$\hat{x} \text{ is feasible to (MRO) for } K = N \iff \hat{x} \text{ is feasible to (MRO) for } K < N.$$

The feasible region of (MRO) is identical for $K = N$ and $K < N$, and the optimal solutions will be identical so long as the optimal solution to (MRO) is unique. In view of the dual problem and constraints, if $g(u, x)$ is affine in u following (9), we observe from (11) that the only term dependent on K is $(P^T x)^T \sum_{k=1}^K w_k \bar{d}_k$, which is equivalent for all K. ■

4.2 Maximum-of-concave functions

We now consider the more general case of a maximum-of-concave constraint function, $g(u, x) = \max_{j \leq J} g_j(u, x)$, subject to a polyhedral support, $S = \{u \mid Hu \leq h\}$. We define the new primal problems

$$\begin{aligned} \bar{g}^K(x) = \underset{v_{11},\ldots,v_{JK} \in S, \alpha \in \Gamma}{\text{maximize}} \quad & \sum_{k=1}^K \sum_{j=1}^J \alpha_{jk} g_j(v_{jk}, x) \\ \text{subject to} \quad & \sum_{k=1}^K \sum_{j=1}^J \alpha_{jk} \|v_{jk} - \bar{d}_k\|^p \leq \epsilon^p \\ & v_i \in S \quad i = 1, \ldots, K, \end{aligned} \tag{MRO-K-MAX}$$

$$\begin{aligned} \bar{g}^{N*}(x) = \underset{v_{11},\ldots,v_{JN}, \alpha \in \Gamma}{\text{maximize}} \quad & \sum_{i=1}^N \sum_{j=1}^J \alpha_{ji} g_j(v_{ji}, x) \\ \text{subject to} \quad & \sum_{i=1}^N \sum_{j=1}^J \alpha_{ji} \|v_{ji} - d_i\|^p \leq \epsilon^p \end{aligned} \tag{MRO-N*-MAX}$$

We also make use of the dual versions of the optimization problems, defined as follows.

$$\begin{aligned} \bar{g}^N(x) = \underset{\lambda \geq 0, z_{ji}, y_{ji}, s_i}{\text{minimize}} \quad & (1/N) \sum_i^N s_i \\ \text{subject to} \quad & [-g_j]^*(z_{ji} - H^T \gamma_{ji}) + \gamma_{ji}^T (h - H d_i) - z_{ji}^T d_i + \phi(q) \lambda \left\| z_{ji}/\lambda \right\|_*^q \\ & \qquad + \lambda \epsilon^p \leq s_i, \quad i = 1, \ldots, N, \quad j = 1, \ldots, J, \end{aligned} \tag{MRO-N-Dual}$$

where no clustering occurs, and

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

where we have K clusters. Given these definitions, we obtain bounds on the worst-case value of the constraint function for K clusters.

Theorem 4.2. *When g is the maximum of concave functions with domain $\mathbf{dom}_u\, g = \mathbf{R}^m$ and polyhedral support $S = \{u \mid Hu \leq h\}$, and where each $-g_j$ satisfies an L-smooth condition on its domain with respect to the ℓ_2-norm at a given x, 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 variable from (MRO-K)*. The constants L_j are the L-smoothness constants for the concave functions g_j.*

The proof is delayed to Appendix A.10. We note that due to the nonconvex and nonconcave nature of maximum-of-concave functions, we can no longer directly fix the relationship between $\bar{g}^N(x)$ and $\bar{g}^K(x)$. Instead, we need to define the lower bound with the extra term $\delta(K, z, \gamma)$. However, in the special case where 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 4.2.1. *When g is the maximum of affine functions with domain $\mathbf{dom}_u = \mathbf{R}^m$ 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 4.2 and Corollary 4.2.1 quantifies the possible difference in optimal values between the MRO problem with K and N clusters. We again define $\Delta = \max_x \left(\bar{g}^{N*}(x) - \bar{g}^N(x)\right)$, subject to x being feasible for problem (MRO). Note, however, that this is only needed for the upper bound. The lower bound holds without needing to consider $\bar{g}^{N*}(x)$; we need not consider the effect of the support set on the problem.

Corollary 4.2.2. *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 in optimal values.*

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 4.2 no longer directly reflect the difference in optimal values. Instead, clustering affects the feasible set for x as follows. In the case $\bar{g}^N(x) \geq \bar{g}^K(x)$, for any $\hat{x}$, $\bar{g}^K(\hat{x})$ 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 $\bar{g}^N(\hat{x})$. Since both values are constrained to be nonpositive, the feasible region of the MRO problem with K clusters may be more restricted than that of N clusters. 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 $K = 1$. We trivially cluster by averaging all data-points without using any clustering algorithm. 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 4.1. 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 4.2 also depends on $D(K)$. Notice that the value $\delta(K, z, \gamma)$ in the lower bound is also a weighted transformation of $D(K)$. The elbow method has been commonly used in machine learning problems pertaining the choice of hyper-parameters, especially for K-means, and can be traced back to Thorndike Thorndike (1953) 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 choosing different K.

Choosing ϵ. While we have outlined theoretical results in Theorem 3.3 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, we do need to choose larger ϵ, as smaller K leads 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 probabilitic 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 centoids 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 Section 6.6, 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 (MOSEK ApS, 2022) optimizer with default settings.

All numerical examples are solved through direct reformulations if not stated otherwise. 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 probablity of constraint violation is measured as the probability the average constraint value is above zero.

In Sections 6.1, 6.2, and 6.3, we demonstrate the performance of MRO when the uncertain constraint is concave. In Section 6.4, 6.5, and 6.6, we demonstrate the performance of MRO for maximum-of-affine uncertainty.

6.1 Capital budgeting

We consider the capital budgeting problem in (Ben-Tal et al., 2015, 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 $t = 0, \ldots, T$, *i.e.*, $\eta_j(u) = \sum_{t=0}^{T} F_{jt}/(1+u_j)^t$. Here, u_j is the discount rate of project j. We formulate the uncertain function to be minimized as

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

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 2.1 and Theorem 4.1 applies. The robust problem can be written as

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

where h is the vector of project weights. We refer to (7) and arrive at the convex reformulation for $p = 2$

$$\begin{array}{ll} \text{minimize} & \tau \\ \text{subject to} & \lambda\epsilon^2 + \sum_{k=1}^K w_k s_k \leq \tau \\ & -F_0^T x + \mathbf{1}^T(\delta_k a - z_k) - z_k^T \bar{d}_k + \gamma_k^T(b - C\bar{d}_k) \\ & \qquad +1/(4\lambda)\|C^T\gamma_k - z_k\|_2^2 \leq s_k, \quad k = 1, \ldots, K \\ & (-(Y_k)_{jt}, F_{jt}x_j, (\delta_k)_{jt}) \in \mathcal{K}^{t/(t+1)}, \quad j = 1, \ldots, n, \quad t = 1, \ldots, T, \\ & \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad k = 1, \ldots, K \\ & Y_k \mathbf{1} = z_k, \qquad k = 1, \ldots, K \\ & h^T x \leq \theta \\ & \lambda \geq 0, \quad \gamma_k \geq 0, \quad Y_k \leq 0, \quad \delta_k \leq 0, \quad x \in \{0,1\}, \end{array} \tag{22}$$

where $a \in \mathbf{R}^T$ with $a_t = t^{1/(t+1)} + t^{-t/(t+1)}$ for $t = 1, \ldots, T$, and $(x, y, z) \in \mathcal{K}^\alpha$ is a power cone constraint given as $x^\alpha y^{1-\alpha} \geq |z|$. The vector F_0 indicates the first column of F, and matrix C and vector b encode the support of u, which we take to be $\{u \in \mathbf{R}^m \mid 0 \leq u \leq \mathbf{1}\}$, where $m = n$. We have variables $x_j \in \mathbf{R}$, $z_k \in \mathbf{R}^n$, $Y_k \in \mathbf{R}^{n \times T}$, $\delta_k \in \mathbf{R}^{n \times T}$, $\tau \in \mathbf{R}$, $\gamma_k \in \mathbf{R}^{2n}$, $s_k \in \mathbf{R}$, for $j = 1, \ldots, n$, $k = 1, \ldots, K$, and $t = 1, \ldots, T$. The derivation of reformulation (22) is in Appendix A.11. Note that there are variables with total dimension KnT, which grows swiftly when any of the parameters are large. For each cluster k, we introduce nT new variables for y and δ, as well as nT new power cone constraints, which greatly increases the compuational complexity of the problem.

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 $\hat{x}_N$.

Choosing K Plotting the clustering value $D(K)$ over K, we note that the elbow occurs around $K = 2$, which suggests using cross-validation for K values around 2.

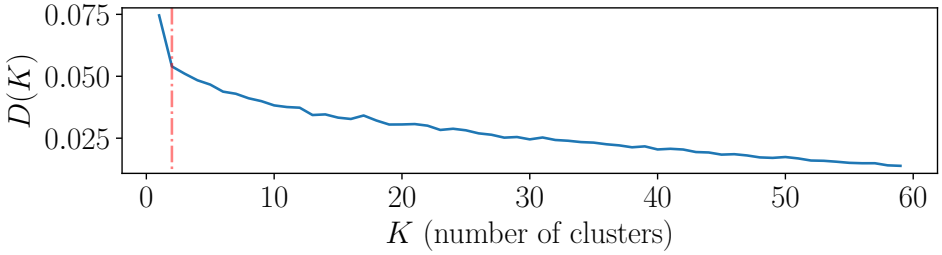

Figure 4: Capital budgeting. $D(K)$ vs. K. Dotted red line: $K = 2$.

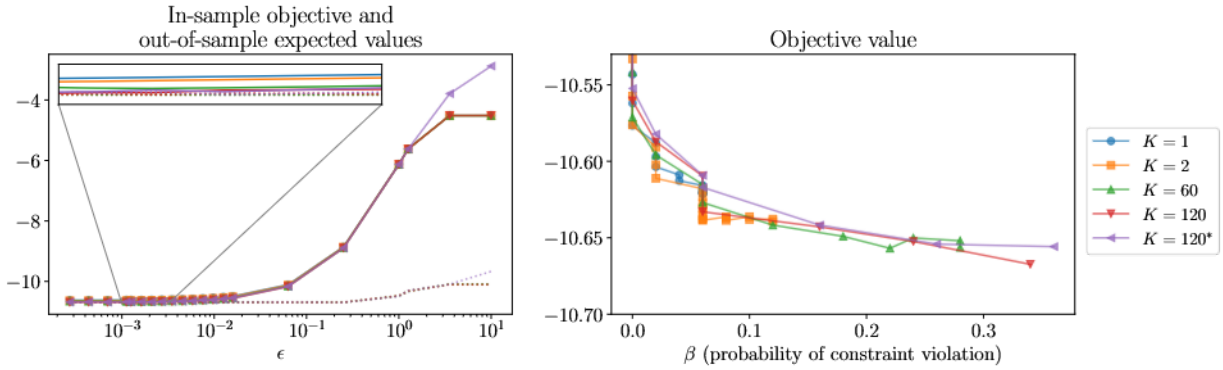

Figure 5: Capital budgeting. Left: in-sample objective values and out-of-sample expected values vs. ϵ for different K. Solid lines are the in-sample objective value, dotted lines are the out-of-sample expected value. Right: objective value vs. β for different K; each point represents the solution for the ϵ achieving the smallest objective value. $K = 120^*$ is the formulation without the support constraint.

Results. We observe in Figure 5 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 $K = 120^*$ is the formulation

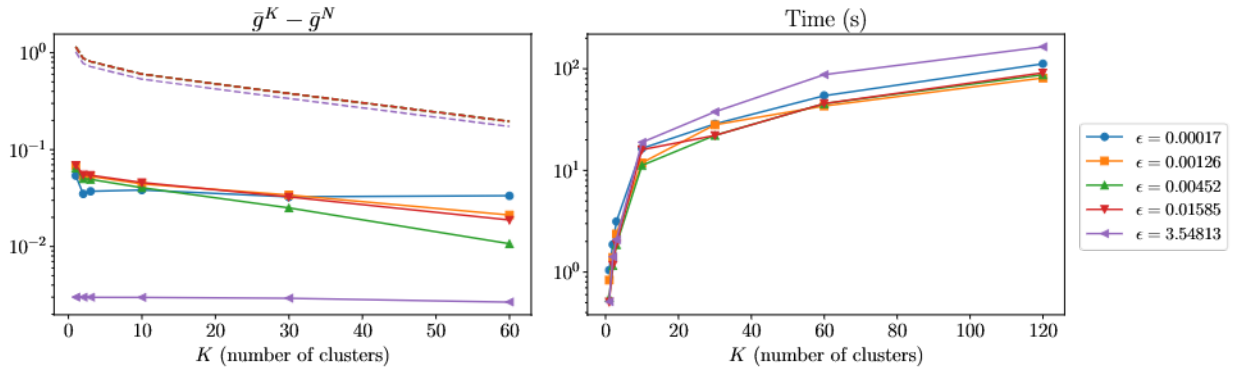

Figure 6: Capital budgeting. Left: the difference in the value of the uncertain objective between using K and N clusters, calculated as $\bar{g}^K(x) - \bar{g}^N(x)$, compared with the theoretical upper bound $(L/2)D(K)$ from Corollary 4.1.1. Solid lines are the difference, dotted lines are the upper bounds. Right: solve time.

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 $(L/2)D(K)$. In Figure 6, 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 Figure 4. Therefore, setting $K = 2$ is the optimal decision, with a time reduction of 2 orders of magnitude, and a complexity reduction from 26626 variables and 12000 power cones to 666 variables and 200 power cones.

6.2 Quadratic concave uncertainty

We refer to the example from Ben-Tal et al. (Ben-Tal et al., 2015, Section 4.2) with concave uncertainty of the form

$$g(u, x) = \sum_{i=1}^{n} h_i(u) x_i,$$

where $h_i(u) = -(1/2)u^T A_i u$, each $A_i \in \mathbf{R}^{m \times m}$ a symmetric positive definite matrix, $u \in \mathbf{R}^m$, and $x \in \mathbf{R}^n_+$. For simplicity, we also require that x sums to 1, $p = 2$, and the support of the uncertainty $S = \mathbf{R}^m$. Assuming the uncertainty is in the objective, such that the uncertain

constraint is created using epigraph form, we solve the problem

$$\begin{array}{ll} \text{minimize} & \tau \\ \text{subject to} & \lambda\epsilon^2 + \sum_{k=1}^K w_k s_k \leq \tau \\ & (1/2)\sum_{i=1}^n \left((Y_k)_i{}^T A_i^{-1} (Y_k)_i \right) / x_i - z_k^T \bar{d}_k + 1/(4\lambda) \left\| z_k \right\|_2^2 \leq s_k \\ & \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad k = 1, \ldots, K \\ & Y_k \mathbf{1} = z_k, \quad k = 1, \ldots, K \\ & \mathbf{1}^T x = 1, \quad \lambda \geq 0, \quad x \geq 0. \end{array}$$

The A_i^{-1} terms come from taking the conjugate of g, and the derivation can be found in (Ben-Tal et al., 2015, Example 24). We have variables $x \in \mathbf{R}^n$, $z_k \in \mathbf{R}^m$, $Y_k \in \mathbf{R}^{m \times n}$, $\tau \in \mathbf{R}$, $s_k \in \mathbf{R}$, for $k = 1, \ldots, K$. We let $(Y_k)_i$ indicate the ith column of Y_k.

Problem setup. We set $n = m = 10, N = 90$, and generate synthetic uncertainty as a multi-modal normal distribution with 5 modes, where $\mu_i = \gamma_j 0.03i$ for all $i = 1, \ldots, n$ for mode j, with mode scales $\gamma = (1, 5, 15, 25, 40)$. The variance is $\sigma_i = 0.02^2 + (0.025i)^2$ for all modes. We generate A_i as random positive semi-definite matrices for all $i = 1, \ldots, n$. For the upper bound, we calulate $L = \| \sum_{i=1}^n A_i (\hat{x}_N)_i \|_{2,2}$ for each data-driven solution $\hat{x}_N$.

Choosing K Plotting the clustering value $D(K)$ over K, we note that the elbow occurs at $K = 5$, which suggests a choice of $K = 5$.

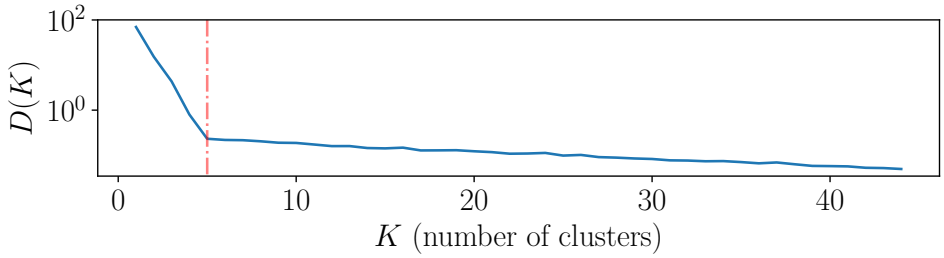

Figure 7: Quadratic concave uncertainty. $D(K)$ vs. K. Dotted red line: $K = 5$.

Results. We observe on the left of Figure 8 that using 5 clusters is enough to achieve performance almost identical to that of using 90 clusters. Indeed, in Figure 9, the elbow of the upper bound (dotted lines) on the difference in objective values is at $K = 5$, and the true difference follows the same trend, corroborating with Figure 7. Furthermore, on the left plot of Figure 9, we note for $K \geq 5$, the tradeoff between the objective value and constraint violation is the same, so we can tune ϵ to achieve the same performance and guarantees. In fact, for this particular example, using a smaller K such as 1 or 2 may allow us to tune ϵ to achieve an even better tradeoff. However, this result cannot be guaranteed in general, so the recommended action is still to choose $K = 5$.

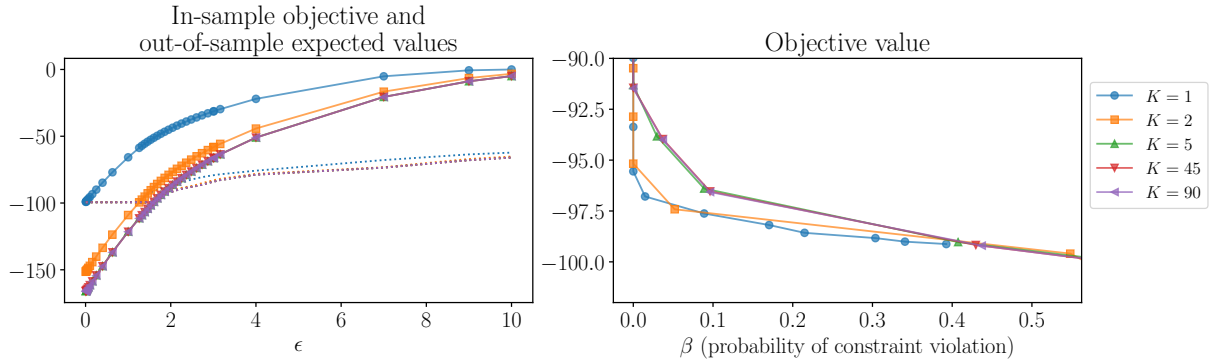

Figure 8: Quadratic concave uncertainty. Left: in-sample objective values and out-of-sample expected values of g vs. ϵ for different K. Solid lines are the in-sample objective value, dotted lines are the out-of-sample expected value. Right: objective value vs. β for different K; each point represents the solution for the ϵ achieving the smallest objective value.

6.3 Robust log-sum-exp optimization

We also consider uncertainty from Bertimas and den Hertog (Bertsimas and den Hertog, 2022, Chapter 14) of the form

$$g(u, x) = \log\left(\sum_{i=1}^{n} u_i e^{x_i}\right),$$

concave in u and convex in x. This function g is monotonically increasing in u, and we can define a domain $u \geq 0.01$ so that Assumption 2.1 and Theorem 4.1 apply. Assuming the simple case where the uncertainty is in the objective, we add some further restrictions on x and use a cutting plane procedure to solve, for $p = 2$,

$$\begin{aligned} \underset{x}{\text{minimize}} \quad \underset{u_1 \ldots u_k}{\text{maximize}} \quad & \sum_{k=1}^{K} w_k \log\left(\sum_{i=1}^{n} u_i e^{x_i}\right) \\ \text{subject to} \quad & \sum_{k=1}^{K} w_k \|u_k - \bar{d}_k\|^2 \leq \epsilon^2 \\ & \mathbf{1}^T x \geq 10, \quad x \geq 0, \quad x \leq 10 \end{aligned}$$

Problem setup. We set $n = 30, N = 90$, and observe synthetic data from 3 sets of uniform distributions, scaled respectively by $\gamma = (1, 3, 7)$. Specifically, for each set j, each d_i is generated uniformly on the intervals $0.01[\gamma_j i, \gamma_j(i+1)]$ for $i = 1, \ldots, n$. For the upper bound, we calculate $L = \|\nabla^2 g\|_{2,2} \leq \exp(\hat{x}_N)^T \exp(\hat{x}_N) \min({d_k}^T \exp(\hat{x}_N))^{-2}$, for each data-driven solution $\hat{x}_N$.

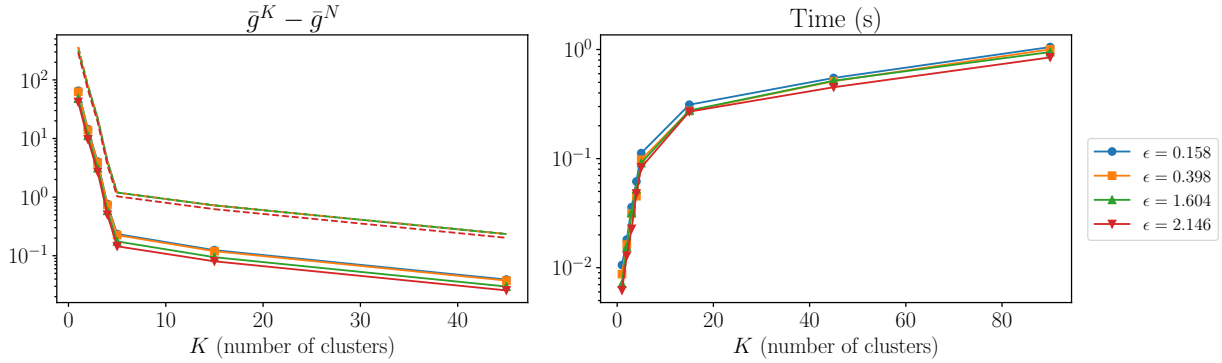

Figure 9: Quadratic concave uncertainty. Left: the difference in the value of the uncertain objective between using K and N clusters, calculated as $\bar{g}^K(x) - \bar{g}^N(x)$, compared with the theoretical upper bound $(L/2)D(K)$ from Corollary (4.1.1). Solid lines are the difference, dotted lines are the upper bounds. Right: solve time.

Choosing K Plotting the clustering value $D(K)$ over K, we note that the elbow occurs at $K = 3$, which suggests using cross-validation for K values around 3.

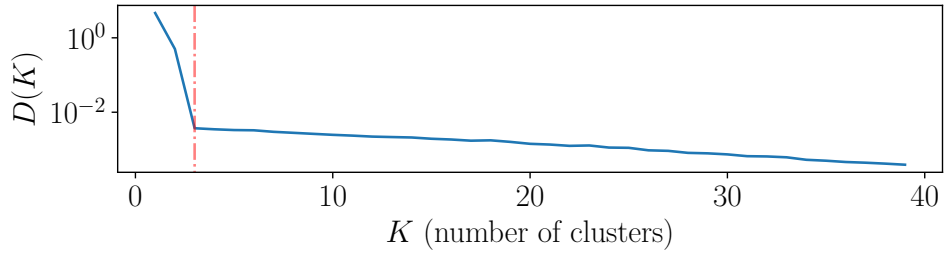

Figure 10: Log-sum-exp uncertainty. $D(K)$ vs. K. Dotted red line: $K = 3$.

Results. We observe on the left of Figure 11 that while setting K to smaller values increases the objective value, setting $K = 3$, the number of modes of the underlying distribution, already offers near identical performance to that of setting $K = 90$. On the left of Figure 12, we see that $K = 3$ is at the elbow of upper bound and actual difference, corroborating with Figure 10. Furthermore, we note that setting $K = 3$ and above give identical tradeoff curves, therefore, choosing $K = 3$ is the time-efficient solution.

6.4 Sparse portfolio optimization

We consider a market that forbids short-selling and has m assets as in (Esfahani and Kuhn, 2018). Daily returns of these assets are given by the random vector $d = (d_1, \ldots, d_m) \in \mathbf{R}^m$.

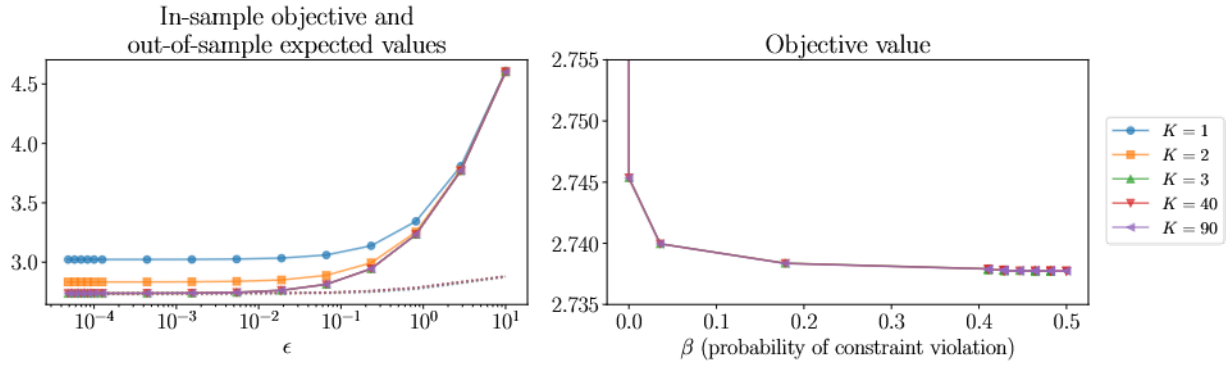

Figure 11: Log-sum-exp uncertainty. Left: in-sample objective values and out-of-sample expected values vs. ϵ for different K. Solid lines are the in-sample objective value, dotted lines are the out-of-sample expected value. Right: objective value vs. β for different K.

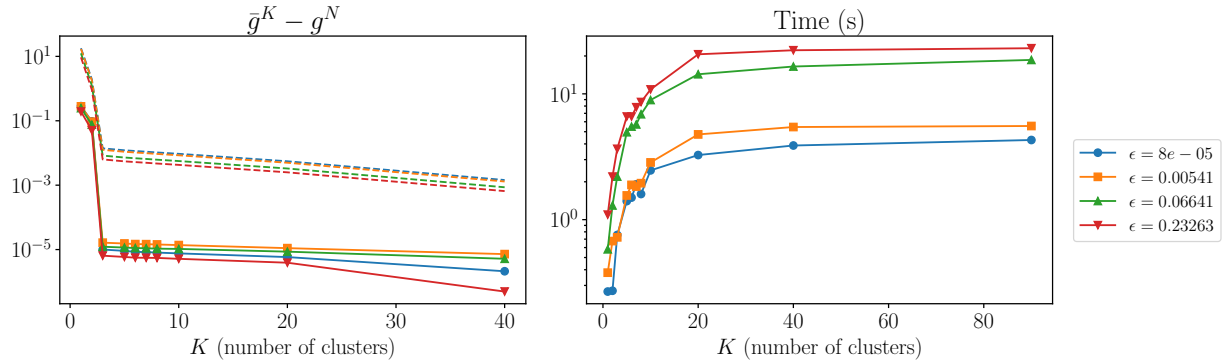

Figure 12: Log-sum-exp uncertainty. Left: the difference in the value of the uncertain objective between using K and N clusters. Solid lines are the difference, the dotted line is the upper bound. Right: solve time.

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. The underlying data-generating 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,

$$\begin{array}{ll} \text{minimize} & \mathbf{CVaR}(-u^T x, \alpha) \\ \text{subject to} & \mathbf{1}^T x = 1 \\ & x \geq 0, \quad \|x\|_1 \leq \theta, \end{array}$$

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 (Uryasev

and Rockafellar, 2001; Esfahani and Kuhn, 2018):

$$\text{minimize} \quad \mathbf{E}^{\mathbf{P}}\left(\tau + \frac{1}{\alpha}\max\{-u^T x - \tau, 0\}\right).$$

From this, we obtain g as the maximum of affine functions,

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

Using the formulation (13) with $p = \infty$, we can write a convex reformulation of the form

$$\begin{array}{ll} \text{minimize} & y \\ \text{subject to} & \epsilon \left\|(1/\alpha)x\right\|_2 + \sum_{k=1}^K w_k s_k \leq y \\ & (1-1/\alpha)\tau - (1/\alpha)x^T \bar{d}_k \leq s_k, \quad k = 1,\ldots,K \\ & s_k \geq \tau, \quad k = 1,\ldots,K \\ & x \geq 0, \quad \mathbf{1}^T x = 1 \\ & z \in \{0,1\}, \quad z - x \leq 0, \quad \mathbf{1}^T z \leq \theta. \end{array}$$

with variables $x \in \mathbf{R}^m$, $z \in \mathbf{R}^m$, $y \in \mathbf{R}$, $\tau \in \mathbf{R}$, $s_k \in \mathbf{R}$. The variables z are introduced to replace the cardinality constraint using big-M formulation (Bertsimas and Cory-Wright, 2022).

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 (Bradley and Taqqu, 2003). See the Github repository for the code, which uses the "Rsafd" R package Carmona (2020). 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 $\delta(K,z,\gamma) = \max_k\{\max_i\{(\bar{d}_k - d_i)^T x/\alpha\}\}$.

Choosing K Plotting the clustering value $D(K)$ over K, we note that the elbow occurs at $K = 5$, which suggests using cross-validation for K values around 5.

Results. In Figure 14, 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 Figures 13 and 15, 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.

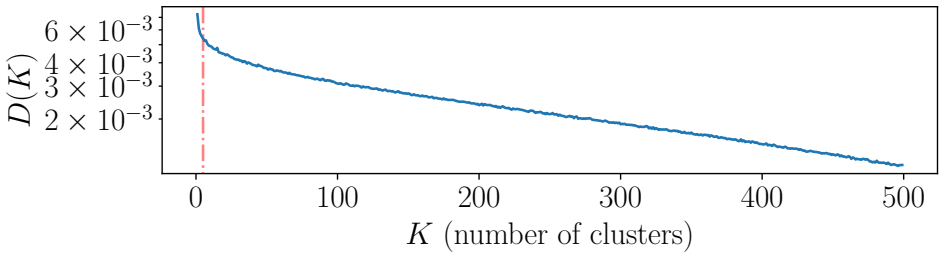

Figure 13: Sparse portfolio. $D(K)$ vs. K. Dotted red line: $K = 5$.

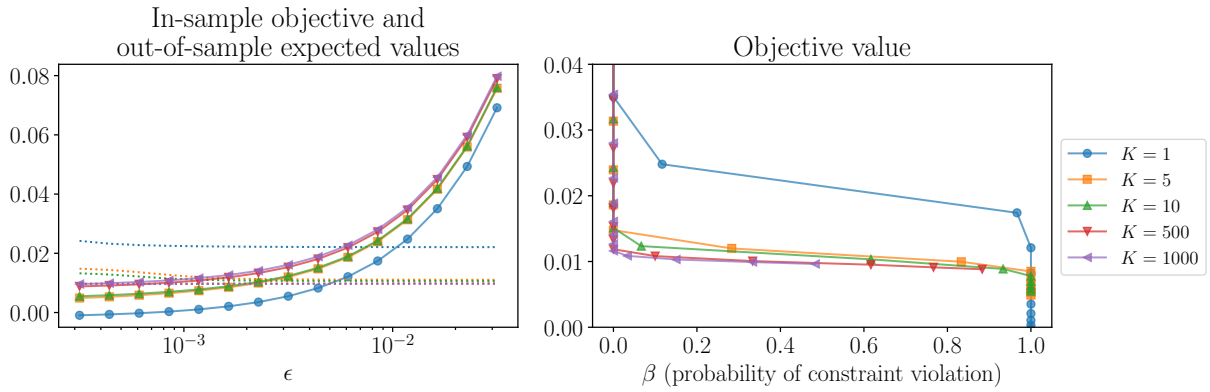

Figure 14: Sparse portfolio. Left: in-sample objective values and out-of-sample expected values vs ϵ for different K. Solid lines are the in-sample objective value, dotted lines are the out-of-sample expected value. Right: objective value vs β for different K; each point represents the solution for the ϵ achieving the smallest objective value.

6.5 Facility location

We examine the classic facility location problem (Bertsimas et al., 2021; Holmberg et al., 1999). 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 $i = 1, \ldots, n$, 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 multiple affine uncertain capacity constraints,

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

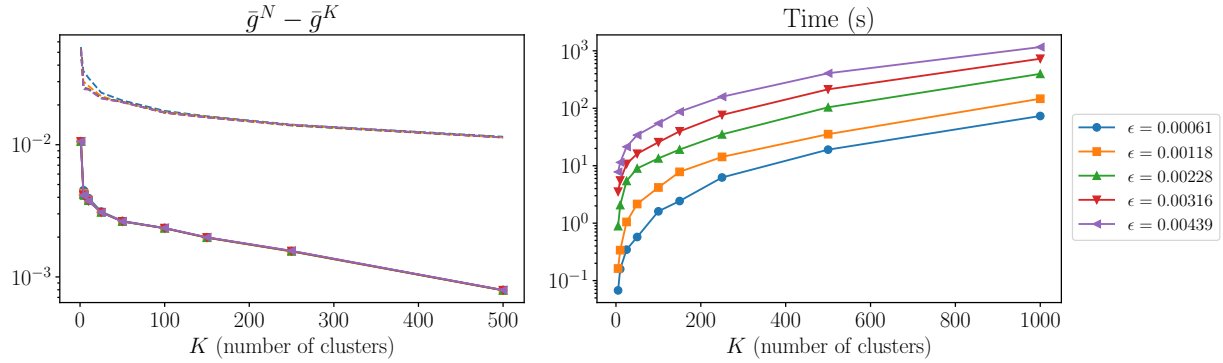

Figure 15: Sparse portfolio. Left: the difference in the value of the uncertain objective between using N and K clusters, calculated as $\bar{g}^N(x) - \bar{g}^K(x)$, compared with the theoretical upper bound $\delta(K, z, \gamma)$ from Corollary 4.2.1. Solid lines are the difference, dotted lines are the upper bounds. Right: solve time for $K \geq 5$.

which we combine to create a single maximum-of-affine constraint,

$$g(u, x) = \max_{i \leq n}((X^T)_i u - r_i x_i) \leq 0.$$

Now, to ensure a high probability of constraint satisfaction, 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 solve the problem, for $p = \infty$,

$$\begin{array}{ll}
\text{minimize} & c^T x + \mathbf{tr}(C^T X) \\
\text{subject to} & \mathbf{1}^T X_j = 1, \quad j = 1, \ldots, m \\
& \tau + \sum_{k=1}^K w_k s_k \leq 0, \\
& -(1/\alpha)\tau + \lambda_k \epsilon + (1/\alpha)((X^T)_i \bar{d}_k - r_i x_i) + \gamma_{ik}(b - H\bar{d}_k) \leq s_k, \\
& \qquad\qquad i = 1, \ldots, n, \quad k = 1, \ldots, K \\
& \lambda_k \epsilon \leq s_k, \quad k = 1, \ldots, K \\
& \left\| H^T \gamma_{ik} + (X^T)_i \right\|_2 \leq \lambda_k, \quad i = 1, \ldots, n, \quad k = 1, \ldots, K \\
& \gamma_{ik} \geq 0, \quad i = 1, \ldots, n, \quad k = 1, \ldots, K \\
& x \in \{0, 1\}^n, \quad X \in \mathbf{R}^{n \times m}.
\end{array}$$

We have variables $x \in \{0, 1\}^n$, $X \in \mathbf{R}^{n \times m}$, $s_k \in \mathbf{R}$, $\tau \in \mathbf{R}$, $\lambda_k \in \mathbf{R}$, $\gamma_{ik} \in \mathbf{R}^m$, for $i = 1, \ldots, n$ and $k = 1, \ldots, K$. The γ variables arise from enforcing the support constraints.

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, which we write as $Hu \leq b$, where $H = [-I\ I]^T$ and b is the concatenation of two vectors: 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 from the normal distribution 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 4.2.1, we have $(1/N)\sum_{k=1}^{K}\sum_{i\in C_k}\max(\max_{j\leq J}((X[i] - H^T\gamma_{jk}), 0)^T(d_i - \bar{d}_k))$. 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 a rough estimate. We cannot directly compare this upper bound against the change in constraint values, as at the optimal chonsen x, X for each K, which differ due to differences in the feasible regions, the constraint value will always be near 0 for optimality. We thus compare it against the change in objective values.

Choosing K Plotting the clustering value $D(K)$ over K, we note that the elbow occurs at $K = 2$, which suggests using cross-validation for K values around 2.

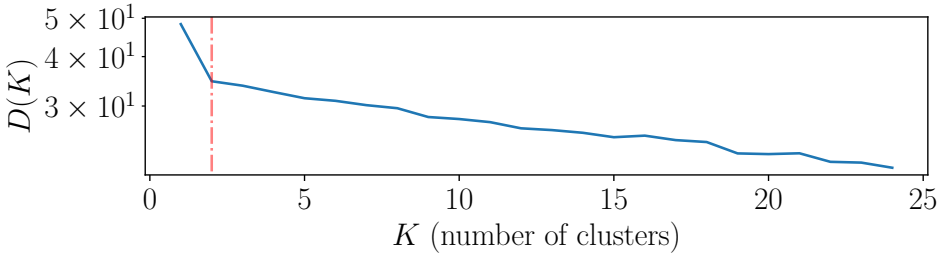

Figure 16: Facility location. $D(K)$ vs. K. Dotted red line: $K = 2$.

Results. As expected of maximum-of-affine g, we note in Figure 17 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 4.2.1 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 Figure 17. 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.

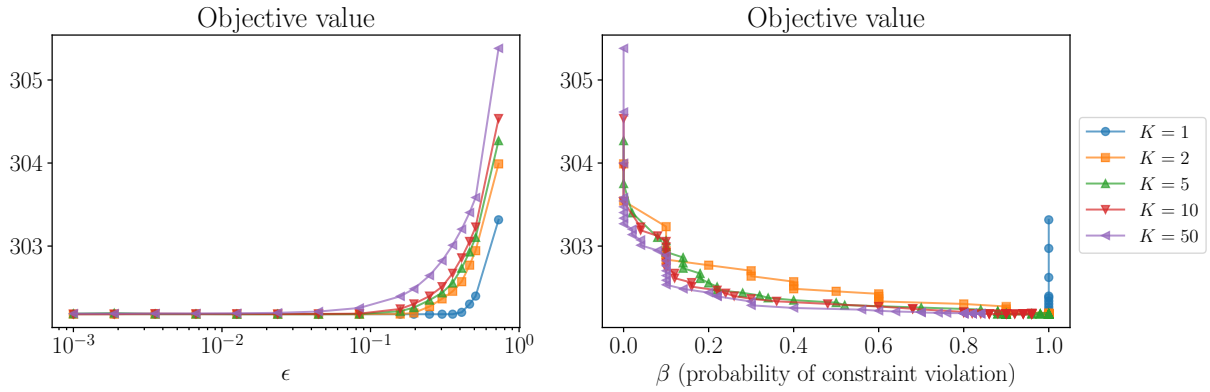

Figure 17: Facility location. Left: in-sample objective values vs ϵ for different K. Right: objective value vs β for different K; each point represents the solution for the ϵ achieving the smallest objective value.

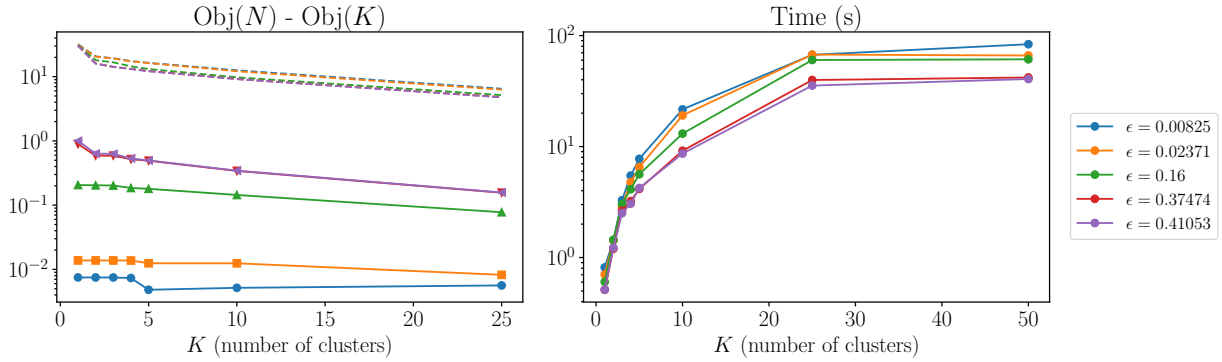

Figure 18: Facility location. Left: the difference in the value of the uncertain objective between using N and K clusters, calculated as Obj(N) - Obj(K), compared with the theoretical upper bound $\delta(K, z, \gamma)$ on the worst-case constraint value $\bar{g}^N(x) - \bar{g}^K(x)$, from Corollary 4.2.1. Solid lines are the difference, dotted lines are the upper bounds. Right: solve time.

6.6 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:

$$h^T x - c^T \min\{x, u\},$$

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 solve, with $p = 1$,

$$\begin{array}{ll}
\text{minimize} & h^T x + \lambda\epsilon + \sum_{k=1}^K w_k s_k \\
\text{subject to} & -c^T x + \gamma_{1k}^T (b - C\bar{d}_k) \leq sk, \quad k = 1, \ldots, K \\
& -c_1 x_1 - \bar{d}_k^T (c_2 e_2) + \gamma_{2k}^T (b - C\bar{d}_k) \leq s_k, \quad k = 1, \ldots, K \\
& -c_2 x_2 - \bar{d}_k^T (c_1 e_1) + \gamma_{3k}^T (b - C\bar{d}_k) \leq s_k, \quad k = 1, \ldots, K \\
& -\bar{d}_k^T (c) + \gamma_{4k}^T (b - C\bar{d}_k) \leq s_k, \quad k = 1, \ldots, K \\
& \left\| -C^T \gamma_{1k} \right\|_2 \leq \lambda_k, \quad k = 1, \ldots, K \\
& \left\| -C^T \gamma_{2k} + c_1 e_1 \right\|_2 \leq \lambda_k, \quad k = 1, \ldots, K \\
& \left\| -C^T \gamma_{3k} + c_2 e_2 \right\|_2 \leq \lambda_k, \quad k = 1, \ldots, K \\
& \left\| -C^T \gamma_{4k} + c \right\|_2 \leq \lambda_k, \quad k = 1, \ldots, K \\
& \gamma_{jk} \geq 0, \quad j = 1, \ldots, 4, \quad k = 1, \ldots, K \\
& x \geq 0
\end{array}$$

We have variables $x \in \mathbf{R}^n$,$s_k \in \mathbf{R}$, $\lambda \in \mathbf{R}$, $\gamma_{jk} \in \mathbf{R}^m$, for $j = 1, \ldots, 4$ and $k = 1, \ldots, K$. The γ variables arise from enforcing the support. We denote $e_1 = (1, 0)$ and $e_2 = (0, 1)$.

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 Figure 19, we show a set of generated data along with this outlier point.

We consider three ways to solve the problem, decribed as follows.

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 ajusted 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

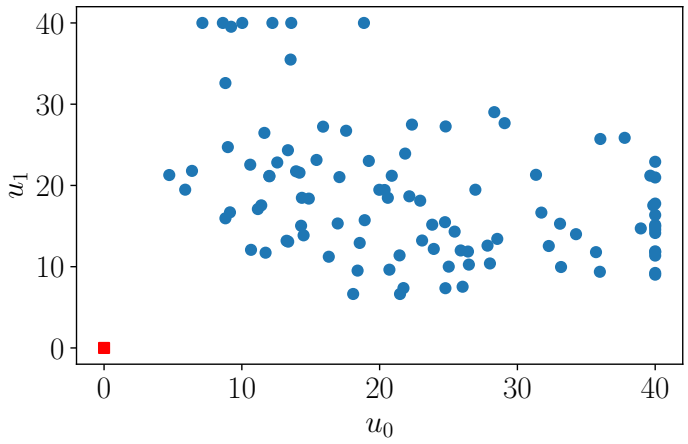

Figure 19: Newsvendor. Datapoints and the outlier at (0,0).

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 4.2.1, 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 $\tilde{c}_1 = 0,\ \tilde{c}_2 = c_1 e_1,\ \tilde{c}_3 = c_2 e_2,\ \tilde{c}_4 = c$.

Choosing K Plotting the clustering value $D(K)$ over K, for the dataset both with and without the outlier, we note an elbow at around $K = 5$, though not very prominent. The recommendation is setting K around 5, to be fine tuned through cross-validation.

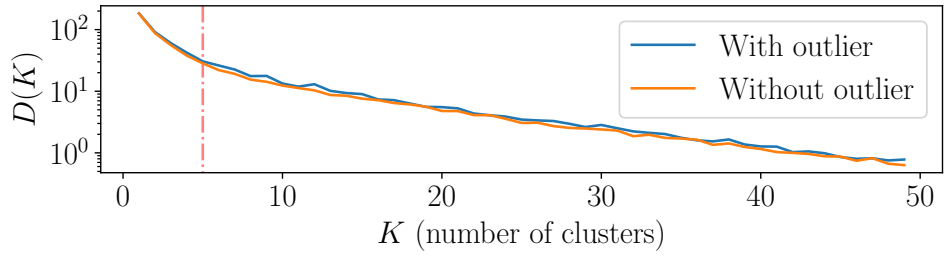

Figure 20: Newsvendor. $D(K)$ vs. K. Dotted red line: $K = 5$.

Results. In Figure 21, we compare the in and out-of-sample objective values of the three methods, and note their similar performance. While setting $K = 1$ yields suboptimal results, we note that for $K = 5$ and above, we can achieve similar performance as setting $K = 100$. To examine the effect of the outlier more closely, in Figures 22 and 23, 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 datapoints, the final in-sample objective may be improved, as the centroid moves closer to the non-outlier points. We observe this in Figure 22, where 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. Lastly, 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 Figures 22 and 23, we note that the difference between MRO and ROB-MRO for $K = 10$ is not larger than the difference for $K = N = 100$, which shows that, while removing outliers before solving the problem may be helpful, the effect of outliers will not be worse for MRO compared to classic Wasserstein DRO.

We note in Figure 24 that the upper bound on $\bar{g}^N(x) - \bar{g}^K(x)$, given in Corollary 4.2.1, 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.

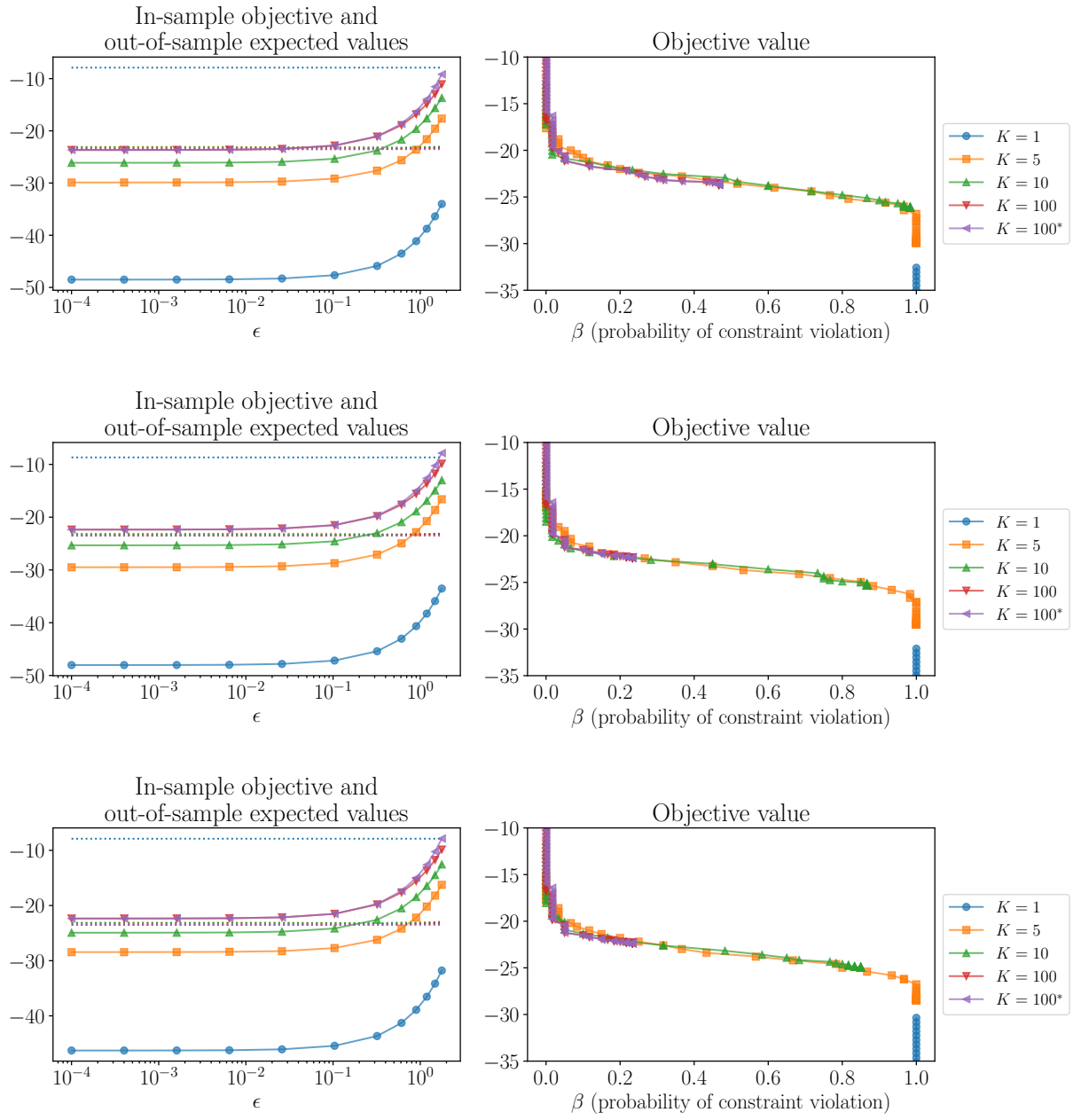

Figure 21: Newsvendor. Left: in-sample objective values vs ϵ for different K. Right: objective value vs β for different K; each point represents the solution for the ϵ achieving the smallest objective value. $K = 100^*$ is the formulation without the support constraint. Top: MRO. Middle: ROB-MRO. Bottom: AUG-MRO.

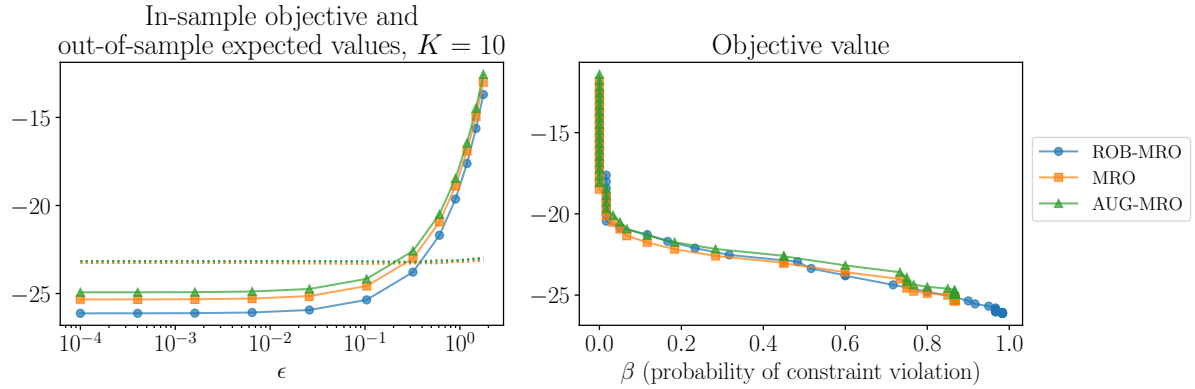

Figure 22: Newsvendor. Left: in-sample objective values vs ϵ for $K = 10$. Right: objective value vs β for $K = 10$.

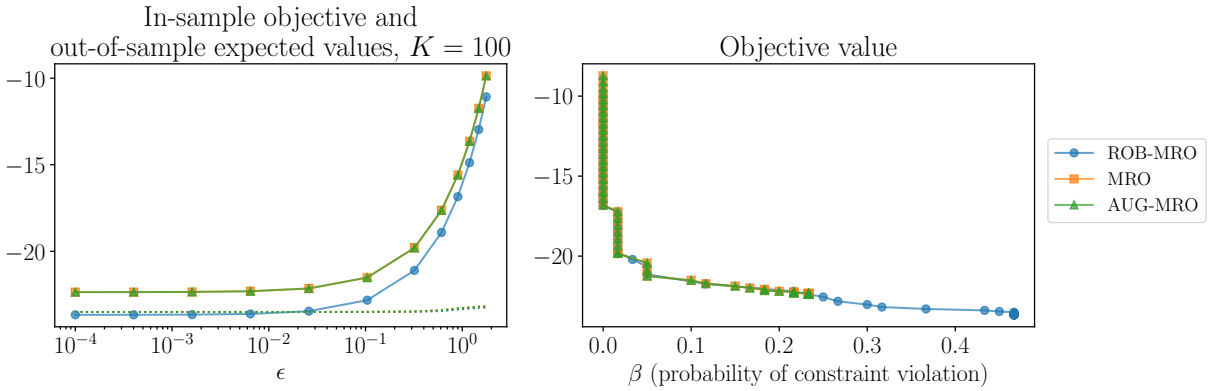

Figure 23: Newsvendor. Left: in-sample objective values vs ϵ for $K = 100$. Right: objective value vs β for $K = 100$.

Acknowledgements

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

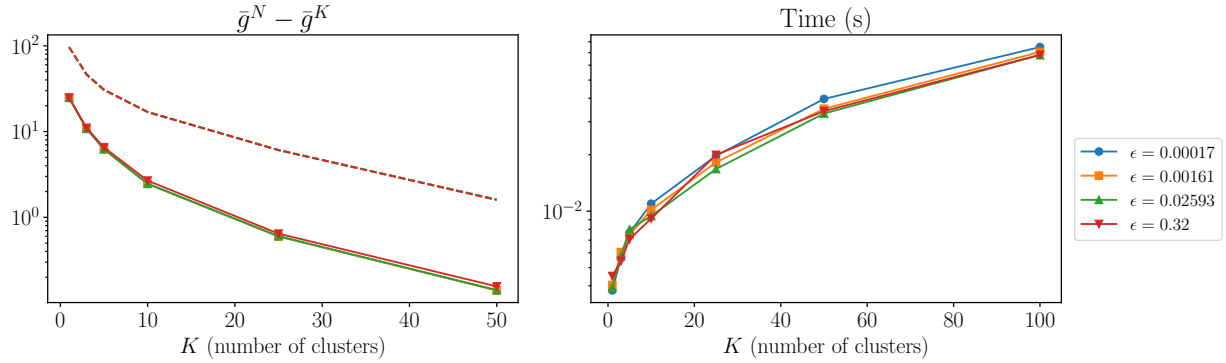

Figure 24: Newsvendor. Left: the difference in the value of the uncertain objective between using N and K clusters, calculated as $\bar{g}^N(x) - \bar{g}^K(x)$, compared with the theoretical upper bound $\delta(K, z, \gamma)$ from Corollary 4.2.1. Solid lines are the difference, dotted lines are the upper bounds. Right: solve time.

References

C. Bandi and D. Bertsimas. Tractable stochastic analysis in high dimensions via robust optimization. *Mathematical Programming*, 134(1):23–70, Aug. 2012.

A. Beck. *First-Order Methods in Optimization*. SIAM-Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2017.

A. Ben-Tal and A. Nemirovski. Robust solutions of Linear Programming problems contaminated with uncertain data. *Mathematical Programming*, 88(3):411–424, Sept. 2000.

A. Ben-Tal and A. Nemirovski. Selected topics in robust convex optimization. *Math. Program.*, 112:125–158, 03 2008.

A. Ben-Tal, L. El Ghaoui, and A. Nemirovski. *Robust Optimization*. Princeton University Press, Princeton, NJ, 2009.

A. Ben-Tal, D. den Hertog, and J. P. Vial. Deriving robust counterparts of nonlinear uncertain inequalities. *Mathematical Programming*, 149(1-2):265–299, Feb. 2015.

P. Beraldi and M. E. Bruni. A clustering approach for scenario tree reduction: an application to a stochastic programming portfolio optimization problem. *TOP*, 22(3):934–949, 2014.

D. Bertsimas and R. Cory-Wright. A scalable algorithm for sparse portfolio selection. *INFORMS Journal on Computing*, 2022.

D. Bertsimas and D. den Hertog. *Robust and Adaptive Optimization*. Dynamic Ideas, Belmont, MA, 2022.

D. Bertsimas and N. Mundru. Optimization-based scenario reduction for data-driven two-stage stochastic optimization. *Operations Research*, 04 2022.

D. Bertsimas and M. Sim. The Price of Robustness. *Operations Research*, 52(1):35–53, Feb. 2004.

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

D. Bertsimas, I. Dunning, and M. Lubin. Reformulation versus cutting-planes for robust optimization. *Computational Management Science*, 13(2):195–217, 2016.

D. Bertsimas, V. Gupta, and N. Kallus. Data-driven robust optimization. *Mathematical Programming*, 167(2):235–292, Feb. 2018.

D. Bertsimas, den Hertog, D., and Pauphilet, J. Probabilistic guarantees in robust optimization. *SIAM Journal on Optimization*, 31(4):2893–2920, 2021.

D. Bertsimas, B. Sturt, and S. Shtern. A data-driven approach to multistage stochastic linear optimization. *Management Science*, 03 2022.

B. O. Bradley and M. S. Taqqu. *Financial Risk and Heavy Tails*, volume 1 of *Handbooks in Finance*. North-Holland, Amsterdam, 2003.

R. A. Carmona. *Rsafd: Statistical Analysis of Financial Data in R*, 2020. R package version 1.2.

L. Chen. Clustering of sample average approximation for stochastic program. 2015.

R. Chen and I. Paschalidis. Distributionally robust learning. *Foundations and Trends in Optimization*, 4(1-2):1–243, 2020.

Z. Chen, M. Sim, and P. Xiong. Robust stochastic optimization made easy with RSOME. *Management Science*, 66(8):3329–3339, 2020.

P. Clement and W. Desch. An elementary proof of the triangle inequality for the wasserstein metric. *Proceedings of the American Mathematical Society*, 136(1):333–339, 2008.

E. Delage and Y. Ye. Distributionally robust optimization under moment uncertainty with application to data-driven problems. *Operations Research*, 58(3):595–612, 2010.

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

A. Emelogu, S. Chowdhury, M. Marufuzzaman, L. Bian, and B. Eksioglu. An enhanced sample average approximation method for stochastic optimization. *International Journal of Production Economics*, 182:230–252, 2016.

P. M. Esfahani and D. Kuhn. Data-driven distributionally robust optimization using the Wasserstein metric: performance guarantees and tractable reformulations. *Mathematical Programming*, 171:115–166, Sept. 2018.

P. A. Esteban and J. M. Morales. Partition-based distributionally robust optimization via optimal transport with order cone constraints. *4OR*, 20(3):465–49,, 2022.

F. Fabiani and P. Goulart. The optimal transport paradigm enables data compression in data-driven robust control. In *2021 American Control Conference (ACC)*, pages 2412–2417, 2021.

N. Fournier and A. Guillin. On the rate of convergence in Wasserstein distance of the empirical measure. *Probability Theory and Related Fields*, 162(3):707–738, 2015.

R. Gao. Finite-sample guarantees for Wasserstein distributionally robust optimization: Breaking the curse of dimensionality. *CoRR*, 2020.

R. Gao and A. Kleywegt. Distributionally robust stochastic optimization with wasserstein distance. *Mathematics of Operations Research*, 48, 04 2023.

C. R. Givens and R. M. Shortt. A class of wasserstein metrics for probability distributions. *Michigan Mathematical Journal*, 31(2), jan 1984.

J. Goh and M. Sim. Distributionally robust optimization and its tractable approximations. *Operations Research*, 58(4-part-1):902–917, 2010.

J. A. Hartigan and M. A. Wong. Algorithm AS 136: A k-means clustering algorithm. *Journal of the royal statistical society. series c (applied statistics)*, 28(1):100–108, 1979.

K. Holmberg, M. Rönnqvist, and D. Yuan. An exact algorithm for the capacitated facility location problems with single sourcing. *European Journal of Operational Research*, 113 (3):544–559, 1999.

D. Jacobson, M. Hassan, and Z. S. Dong. Exploring the effect of clustering algorithms on sample average approximation. *2021 Institute of Industrial and Systems Engineers (IISE) Annual Conference & Expo*, 2021.

D. Kuhn, P. M. Esfahani, V. Nguyen, and S. Shafieezadeh-Abadeh. *Wasserstein Distributionally Robust Optimization: Theory and Applications in Machine Learning*, pages 130–166. 10 2019. ISBN 978-0-9906153-3-0.

Y. Liu, X. Yuan, and J. Zhang. Discrete approximation scheme in distributionally robust optimization. *Numerical Mathematics: Theory, Methods and Applications*, 14(2):285–320, 2021.

MOSEK ApS. *The MOSEK optimization toolbox. Version 9.3.*, 2022.

G. Perakis, M. Sim, Q. Tang, and P. Xiong. Robust pricing and production with information partitioning and adaptation. *Management Science*, 2023.

R. T. Rockafellar and S. Uryasev. Conditional value-at-risk for general loss distributions. *Journal of banking & finance*, 26(7):1443–1471, 2002.

R. T. Rockafellar and R. J. Wets. Variational analysis. *Grundlehren der mathematischen Wissenschaften*, 1998.

E. Roos and D. den Hertog. Reducing conservatism in robust optimization. *INFORMS Journal on Computing*, 32(4):1109–1127, 2020.

N. Rujeerapaiboon, K. Schindler, D. Kuhn, and W. Wiesemann. Scenario reduction revisited: fundamental limits and guarantees. *Mathematical Programming*, 191(1):207–242, 2022.

R. Thorndike. Who belongs in the family? *Psychometrika*, 18(4):267–276, December 1953.

N. Trillos and D. Slepčev. On the rate of convergence of empirical measures in ∞-transportation distance. *Canadian Journal of Mathematics*, 67, 07 2014.

S. Uryasev and R. T. Rockafellar. *Conditional value-at-risk: optimization approach*. Springer, New York, NY, 2001.

Z. Wang, P. Wang, and Y. Ye. Likelihood robust optimization for data-driven problems. *Computational Management Science*, 13(2):241–261, Apr. 2016.

W. Wiesemann, D. Kuhn, and M. Sim. Distributionally robust convex optimization. *Operations Research*, 62(6):1358–1376, dec 2014.

H. Xu, C. Caramanis, and S. Mannor. A distributional interpretation of robust optimization. *Math. Oper. Res.*, 37(1):95–110, feb 2012. ISSN 0364-765X.

J. Zhen, D. Kuhn, and W. Wiesemann. A unified theory of robust and distributionally robust optimization via the primal-worst-equals-dual-best principle, 2021.

S. Zymler, D. Kuhn, and B. Rustem. Distributionally robust joint chance constraints with second-order moment information. *Mathematical Programming*, 137(1):167–198, 2013.

A Appendices

A.1 Proof of the constraint reformulation in (7)

To simplify notation, we define $c_k(v_k) = \|v_k - \bar{d}_k\|^p - \epsilon^p$. Then, starting from the inner optimization problem of (6):

$$
\begin{aligned}
&\begin{cases} \sup\limits_{v_1 \dots v_K \in S} & \sum_{k=1}^K w_k g(v_k, x) \\ \text{subject to} & \sum_{k=1}^K w_k c_k(v_k) \leq 0 \end{cases} \\
&= \begin{cases} \sup\limits_{v_1 \dots v_K \in S} & \inf\limits_{\lambda \geq 0} \sum_{k=1}^K w_k g(v_k, x) - \lambda \sum_{k=1}^K w_k c_k(v_k) \end{cases} \\
&= \begin{cases} \inf\limits_{\lambda \geq 0} & \sup\limits_{v_1 \dots v_K \in S} \sum_{k=1}^K w_k g(v_k, x) - \lambda \sum_{k=1}^K w_k c_k(v_k) \end{cases} \\
&= \begin{cases} \inf\limits_{\lambda \geq 0} & \sum_{k=1}^K w_k s_k \\ \text{subject to} & \sup\limits_{v_k \in S} g(v_k, x) - \lambda c_k(v_k) \leq s_k \quad k = 1, \dots, K \end{cases} \\
&= \begin{cases} \inf\limits_{\lambda \geq 0} & \sum_{k=1}^K w_k s_k \\ \text{subject to} & [-g + \chi_S + \lambda c_k]^*(0) \leq s_k \quad k = 1, \dots, K, \end{cases}
\end{aligned}
$$

where we used the Lagrangian in the first equality, the Von Neumann-Fan minimax theorem for the second equality, where we applied the assumption that g is upper-semicontinuous and concave in v, and the convexity of the supports. For the last equality, we used the definition of conjugate functions. Now, borrowing results from Esfahani et al. (Esfahani and Kuhn, 2018, Theorem 4.2), Rockafellar and Wets (Rockafellar and Wets, 1998, Theorem 11.23(a), p. 493), and Zhen et al. (Zhen et al., 2021, Lemma B.8), with regards to the conjugate functions of infimal convolutions and p-norm balls, we note that:

$$
[-g + \chi_S + \lambda c_k]^*(0) = \inf_{y_k, z_k} ([-g]^*(z_k - y_k, x) + \sigma_S(y_k) + [\lambda c_k]^*(-z_k)]),
$$

and

$$
[\lambda c_k]^*(-z_k) = \sup_{v_k} (-z_k^T v_k - \lambda \|v_k - \bar{d}_k\|^p + \lambda \epsilon^p) = -z_k^T \bar{d}_k + \phi(q) \lambda \|z_k / \lambda\|_*^q + \lambda \epsilon^p.
$$

Substituting these in, we arrive at:

$$
\begin{cases} \inf\limits_{\lambda \geq 0, z_k, y_k, s_k} & \sum_k^K w_k s_k \\ \text{subject to} & [-g]^*(z_k - y_k, x) + \sigma_S(y_k) - z_k^T \bar{d}_k + \phi(q) \lambda \left\| z_k / \lambda \right\|_*^q + \lambda \epsilon^p \leq s_k \\ & \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad k = 1, \dots, K. \end{cases}
$$

A.2 Proof of the constraint reformulation in (13)

Starting from the inner optimization problem of (12):

$$
\begin{aligned}
&\begin{cases} \sup\limits_{v_1 \dots v_K \in S} & \sum_{k=1}^K w_k g(v_k, x) \\ \text{subject to} & \|v_k - \bar{d}_k\| \le \epsilon, \quad k = 1, \dots, K \end{cases} \\
&= \begin{cases} \sup\limits_{v_1 \dots v_K \in S} & \inf\limits_{\lambda_k \ge 0} \sum_{k=1}^K w_k (g(v_k, x) + (1/w_k)\lambda_k(\epsilon - \|v_k - \bar{d}_k\|)) \end{cases} \\
&= \begin{cases} \inf\limits_{\lambda_k \ge 0} & \sup\limits_{v_1 \dots v_K \in S} \sum_{k=1}^K w_k (g(v_k, x) + (1/w_k)\lambda_k(\epsilon - \|v_k - \bar{d}_k\|)) \end{cases} \\
&= \begin{cases} \inf\limits_{\lambda_k \ge 0} & \sum_{k=1}^K w_k s_k \\ \text{subject to} & (\lambda_k/w_k)\epsilon + \sup\limits_{v_k \in S} g(v_k, x) - (\lambda_k/w_k)\|v_k - \bar{d}_k\| \le s_k \quad k = 1, \dots, K \end{cases} \\
&= \begin{cases} \inf\limits_{\lambda_k \ge 0} & \sum_{k=1}^K w_k s_k \\ \text{subject to} & (\lambda_k/w_k)\epsilon + \sup\limits_{v_k \in S} g(v_k, x) - \max\limits_{\|z_k\|_* \le \lambda_k/w_k} z_k^T(v_k - \bar{d}_k) \le s_k \\ & \qquad\qquad\qquad\qquad\qquad\qquad k = 1, \dots, K \end{cases} \\
&= \begin{cases} \inf\limits_{\lambda_k \ge 0} & \sum_{k=1}^K w_k s_k \\ \text{subject to} & (\lambda_k/w_k)\epsilon + \min\limits_{\|z_k\|_* \le \lambda_k/w_k} \sup\limits_{v_k \in S} g(v_k, x) - z_k^T(v_k - \bar{d}_k) \le s_k \\ & \qquad\qquad\qquad\qquad\qquad\qquad k = 1, \dots, K \end{cases} \\
&= \begin{cases} \inf\limits_{\lambda_k \ge 0, z_k} & \sum_{k=1}^K w_k s_k \\ \text{subject to} & (\lambda_k/w_k)\epsilon + [-g + \chi_S]^*(-z_k) + z_k^T \bar{d}_k \le s_k \quad k = 1, \dots, K \\ & \|z_k\|_* \le \lambda_k/w_k, \quad k = 1, \dots, K \end{cases} \\
&= \begin{cases} \inf\limits_{z_k} & \sum_{k=1}^K w_k s_k \\ \text{subject to} & [-g]^*(z_k - y_k, x) + \sigma_S(y_k) - z_k^T \bar{d}_k + \epsilon\|z_k\|_* \le s_k \quad k = 1, \dots, K. \end{cases}
\end{aligned}
$$

We have again used the Lagragian, the minmax theorem, and the definition of conjugate function. In particular, in the fourth equality, we refer to the proof of (Esfahani and Kuhn, 2018, Theorem 4.2) where we use the definition of the dual norm. In the final equality, we make the substition $z_k = -z_k$ and $\lambda_k/w_k = \|z_k\|_*$, because the coefficient multiplying λ_k, ϵ/w_k, is nonnegative and λ_k/w_k achieves its lower bound, $\|z_k\|_*$, for all k.

A.3 Proof of the constraint reformulation in (19)

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 (18):

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

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 to interchange the inf and the sup. Next, we rewrite the formulation using an epigraph trick, and make a change of variables.

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

In the last step, we rewrote $v_{jk} = (\alpha_{jk}v_{jk})/\alpha_{jk}$, and maximized over $\alpha_{jk}v_{jk} \in S'$, where S' is the support trasformed by α_{jk}. 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 $h_{ij} = \alpha_{jk}v_{jk}$, and define functions $g'_j(h_{jk}, x) = \alpha_{jk}g_j(h_{jk}/\alpha_{jk})$, $c'_k(h_{jk}) = \alpha_{jk}c'_k(h_{jk}/\alpha_{jk})$.

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

For the new functions defined, we applied the definition of conjugate functions. Now, using the conjugate form $f^*(y) = \alpha g^*(y)$ of a right-scalar-multiplied function $f(x) = \alpha g(x/\alpha)$, we rewrite the above as

$$= \begin{cases} \sup\limits_{\alpha\in\Gamma} \inf\limits_{\lambda\geq 0} & \sum_{k=1}^K s_k \\ \text{subject to} & \sum_{j=1}^J \alpha_{jk}[-g_j + \chi_S + \lambda c_k]^*(0) \leq s_k \quad k = 1,\ldots,K. \end{cases}$$

We note that while the support of h_{jk} is S', the support of h_{jk}/α_{jk}, which is the input of g_j, is still S. Now, again using properties of conjuate functions as given in Appendix A.1, we note that:

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

Substituting in the conjugate functions derived in Appendix A.1, we have

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

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

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

which is equivalent to (19).

A.4 Reformulation of the maximum-of-concave case for $p = \infty$

In the case where $p = \infty$, we have

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & \left\{ \begin{array}{ll} \underset{v_{11},\ldots,v_{JK}\in S, \alpha\in\Gamma}{\text{maximize}} & \sum_{k=1}^K \sum_{j=1}^J \alpha_{jk} g(v_{jk}, x) \\ \text{subject to} & \sum_{j=1}^J (\alpha_{jk}/w_k)|v_{jk} - \bar{d}_k\|^p \leq \epsilon, \quad k = 1,\ldots,K \end{array} \right\} \leq 0, \end{array} \tag{23}$$

which has a reformulation where the constraint above is dualized,

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & \sum_{k=1}^K w_k s_k \leq 0 \\ & [-g_j]^*(z_{jk} - y_{jk}, x) + \sigma_S(y_{jk}) - z_{jk}^T \bar{d}_k + \lambda_k\epsilon \leq s_k \\ & \qquad\qquad k = 1,\ldots,K, \quad j = 1,\ldots,J \\ & \left\| z_{jk} \right\|_* \leq \lambda_k \quad k = 1,\ldots,K, \quad j = 1,\ldots,J, \end{array} \tag{24}$$

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 (23):

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

We have again formulated the Lagragian and applied the minmax theorem to the sum of concave functions in v_{jk}. Now, using the same procedure as in Appendix A.1, we can rewrite this in epigraph form, and use the dual norm definition,

$$\begin{aligned}
&= \begin{cases} \sup\limits_{\alpha\in\Gamma} \inf\limits_{\lambda\ge 0} & \sum_{k=1}^K s_k \\ \text{subject to} & \sup\limits_{v_{11},\ldots,v_{JK}\in S} \lambda_k\epsilon + \sum_{j=1}^J \alpha_{jk} g_j(v_{jk}, x) - \lambda_k(\alpha_{jk}/w_k)\|v_{jk} - \bar{d}_k\| \le s_k \\ & \qquad\qquad k = 1, \ldots, K, \end{cases} \\
&= \begin{cases} \sup\limits_{\alpha\in\Gamma} \inf\limits_{\lambda\ge 0} & \sum_{k=1}^K s_k \\ \text{subject to} & \sup\limits_{v_{11},\ldots,v_{JK}\in S} \lambda_k\epsilon + \sum_{j=1}^J \min\limits_{\|z_{jk}\|_*\le\lambda_k/w_k} \alpha_{jk} g_j(v_{jk}, x) \\ & \qquad\qquad -\alpha_{jk} z_{jk}^T(v_{jk} - \bar{d}_k) \le s_k \quad k = 1, \ldots, K. \end{cases} \\
&= \begin{cases} \sup\limits_{\alpha\in\Gamma} \inf\limits_{\lambda\ge 0} & \sum_{k=1}^K s_k \\ \text{subject to} & \sup\limits_{\alpha_{11}v_{11},\ldots,\alpha_{JK}v_{JK}\in S'} \lambda_k\epsilon + \sum_{j=1}^J \min\limits_{\|z_{jk}\|_*\le\lambda_k/w_k} \alpha_{jk} g_j((\alpha_{jk}v_{jk})/\alpha_{jk}, x) \\ & \qquad\qquad -\alpha_{jk} z_{jk}^T(v_{jk} - \bar{d}_k) \le s_k \quad k = 1, \ldots, K. \end{cases}
\end{aligned}$$

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.3, 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.3. We directly apply the definition of conjugate functions,

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

Now, substituting $\lambda_k = \lambda_k w_k$, $z_{jk} = -z_{jk}$, and substituting in the conjugate functions derived in Appendix A.3, we have

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

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

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

A.5 Proof of the primal problem reformulation as $\mathbf{p} \to \infty$

Consider again the function $\bar{g}^K$ discussed in Section 4 and defined as

$$\bar{g}^K(x; p) = \begin{cases} \text{maximize} & \sum_{k=1}^K w_k g(v_k, x) \\ \text{subject to} & \sum_{k=1}^K w_k \|v_k - d_k\|^p \leq \epsilon^p \\ & v_k \in S \quad k = 1, \ldots, K \end{cases}$$

where we make its dependence on p explicit. We have that $1/M < w_k = |C_k|/N < M$ for all $k = 1, \ldots, K$ for some large $M \geq 1$.

Theorem A.1. *Let the functions $\epsilon \mapsto g^\epsilon(d_k, x)$ be continuous for all $k = 1, \ldots, K$ where $g^\epsilon(d, x) = \max\{g(v, x) \mid v \in S, \|v - d\| \leq \epsilon\}$. We have that $\bar{g}^K(x; \infty) = \lim_{p\to\infty} \bar{g}^K(x; p) = \sum_{k=1}^K w_k g^\epsilon(d_k, x)$.*

Proof. Using the auxiliary variables $t_k \geq 0$ for $k = 1, \ldots, K$ we have that

$$\bar{g}^K(x; p) = \begin{cases} \text{maximize} & \sum_{k=1}^K w_k g(v_k, x) \\ \text{subject to} & v_k \in S, \ t_k \geq 0 \quad k = 1, \ldots, K \\ & \sum_{k=1}^K t_k^p \leq \epsilon^p \\ & t_k \geq \|v_k - d_k\| w_k^{1/p} \quad k = 1, \ldots, K. \end{cases}$$

The function $\bar{g}^K(x;p)$ is hard to study directly. Hence, let us first introduce two auxiliary functions

$$\bar{g}_u^K(x;p) = \begin{cases} \text{maximize} & \sum_{k=1}^K w_k g(v_k, x) \\ \text{subject to} & v_k \in S,\, t_k \geq 0 \quad k = 1, \ldots, K \\ & \sum_{k=1}^K t_k^p \leq \epsilon^p \\ & t_k \geq \|v_k - d_k\| M^{-1/p} \quad k = 1, \ldots, K \end{cases}$$

and

$$\bar{g}_l^K(x;p) = \begin{cases} \text{maximize} & \sum_{k=1}^K w_k g(v_k, x) \\ \text{subject to} & v_k \in S,\, t_k \geq 0 \quad k = 1, \ldots, K \\ & \sum_{k=1}^K t_k^p \leq \epsilon^p, \\ & t_k \geq \|v_k - d_k\| M^{1/p} \quad k = 1, \ldots, K. \end{cases}$$

Observe that for $p \geq 1$ we have $1/M < w_k < M \implies M^{-1/p} < w_k^{1/p} < M^{1/p}$ for any $k \in 1, \ldots, K$. As we hence have for all $k = 1, \ldots, K$ that $M^{-1/p}\|v_k - d_k\| \leq w_k^{1/p}\|v_k - d_k\| \leq M^{1/p}\|v_k - d_k\|$ we obtain the sandwich inequality $\bar{g}_l^K(x;p) \leq \bar{g}^K(x;p) \leq \bar{g}_u^K(x;p)$ for any $p \geq 1$.

Furthermore, observe that when $t_k \geq 0$ for all $k = 1, \ldots, K$ then we have the implication $\sum_{k=1}^K t_k^p \leq \epsilon^p \implies \max_{k=1}^K t_k \leq \epsilon$. Hence, we have that

$$\begin{aligned} \bar{g}_u(x;p) \leq & \begin{cases} \text{maximize} & \sum_{k=1}^K w_k g(v_k, x) \\ \text{subject to} & v_k \in S,\, t_k \geq 0 \quad k = 1, \ldots, K, \\ & \max_{k=1}^K t_k \leq \epsilon, \\ & t_k \geq \|v_k - d_k\| M^{-1/p} \quad k = 1, \ldots, K \end{cases} \\ = & \begin{cases} \text{maximize} & \sum_{k=1}^K w_k g(v_k, x) \\ \text{subject to} & v_k \in S \quad k = 1, \ldots, K \\ & \max_{k=1}^K \|v_k - d_k\| M^{-1/p} \leq \epsilon \end{cases} \\ = & \sum_{k=1}^K w_k \left[\max_{v \in S,\, \|v - d_k\| \leq \epsilon M^{1/p}} g(v, x) \right]. \end{aligned}$$

Similarly, observe that when $t_k \geq 0$ for all $k = 1, \ldots, K$ we also have the inequality

$\sum_{k=1}^K t_k^p \leq K(\max_{k=1}^K t_k)^p$. Hence, we have that

$$\begin{aligned}
\bar{g}_l^K(x;p) \geq & \begin{cases} \text{maximize} & \sum_{k=1}^K w_k g(v_k, x) \\ \text{subject to} & v_k \in S,\, t_k \geq 0 \quad k = 1, \ldots, K \\ & K(\max_{k=1}^K t_k)^p \leq \epsilon^p \\ & t_k \geq \|v_k - d_k\| M^{1/p} \quad k = 1, \ldots, K \end{cases} \\
= & \begin{cases} \text{maximize} & \sum_{k=1}^K w_k g(v_k, x) \\ \text{subject to} & v_k \in S,\, t_k \geq 0 \quad k = 1, \ldots, K, \\ & \max_{k=1}^K t_k \leq K^{-1/p}\epsilon, \\ & t_k \geq \|v_k - d_k\| M^{1/p} \quad k = 1, \ldots, K \end{cases} \\
= & \begin{cases} \text{maximize} & \sum_{k=1}^K w_k g(v_k, x) \\ \text{subject to} & v_k \in S \;\; \forall k \in [1, \ldots, K], \\ & \max_{k=1}^K \|v_k - d_k\| M^{1/p} \leq \epsilon K^{-1/p} \end{cases} \\
= & \sum_{k=1}^K w_k \left[\max_{v \in S,\, \|v - d_k\| \leq \epsilon (MK)^{-1/p}} g(v, x) \right].
\end{aligned}$$

Finally, chaining all the inequalities together we obtain

$$\sum_{k=1}^K w_k \left[\max_{v \in S,\, \|v - d_k\| \leq \epsilon (MK)^{-1/p}} g(v, x) \right] \leq \bar{g}^K(x;p) \leq \sum_{k=1}^K w_k \left[\max_{v \in S,\, \|v - d_k\| \leq \epsilon M^{1/p}} g(v, x) \right]$$

for any $p \geq 1$. Considering now the limit for p to infinity

$$\begin{aligned}
& \lim_{p \to \infty} \sum_{k=1}^K w_k g^{\epsilon (MK)^{-1/p}}(d_k, x) \leq \lim_{p \to \infty} \bar{g}^K(x;p) \leq \lim_{p \to \infty} \sum_{k=1}^K w_k g^{\epsilon M^{1/p}}(d_k, x) \\
\implies & \sum_{k=1}^K w_k \lim_{p \to \infty} g^{\epsilon (MK)^{-1/p}}(d_k, x) \leq \lim_{p \to \infty} \bar{g}^K(x;p) \leq \sum_{k=1}^K w_k \lim_{p \to \infty} g^{\epsilon M^{1/p}}(d_k, x) \\
\implies & \sum_{k=1}^K w_k g^{\epsilon}(d_k, x) \leq \lim_{p \to \infty} \bar{g}^K(x;p) \leq \sum_{k=1}^K w_k g^{\epsilon}(d_k, x)
\end{aligned}$$

establishes the claim. The first implication follows from the fact that the finite sums and limits can be exchanged. The final implication follows from $\lim_{p \to \infty} (MK)^{-1/p} = \lim_{p \to \infty} M^{1/p} = 1$ and the fact that the functions $\epsilon \mapsto g^{\epsilon}(d_k, x)$ are continuous for all $k = 1, \ldots, K$. ■

A.6 Proof of the dual problem reformulation as $\mathbf{p} \to \infty$

Theorem A.2. *Let S be a bounded set. Define here*

$$\bar{g}^K(x;\infty) = \left\{ \begin{array}{ll} \text{minimize} & \sum_{k=1}^K w_k s_k \\ \text{subject to} & \lambda \geq 0,\ z_k \in \mathbf{R}^m,\ y_k \in \mathbf{R}^m,\ s_k \in \mathbf{R}^m \quad k = 1,\ldots,K, \\ & [-g]^*(z_k - y_k) + \sigma_S(y_k) - z_k^T d_k + \epsilon \left\| z_k \right\|_* \leq s_k \\ & \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad k = 1,\ldots,K. \end{array} \right. \tag{25}$$

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

Proof. First, from Equation (7) we have for any $p > 1$ that

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

where the first equality is established in Appendix A.1 and the second equality follows from Lemma A.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 $k = 1,\ldots,K$. 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, \ldots, K, \\
& [-g]^*(z_k - y_k) + \sigma_S(y_k) - z_k^T d_k \\
& \qquad +\phi(q) \left[\frac{q-1}{q} \epsilon^{\frac{1}{1-q}} \max_{k'=1}^K \|z_{k'}\|_* \right]^{1-q} \|z_k\|_*^q \\
& \qquad + \left[\frac{q-1}{q} \epsilon^{\frac{1}{1-q}} \max_{k'=1}^K \|z_{k'}\|_* \right] \epsilon^p \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{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, \ldots, K, \\
& [-g]^*(z_k - y_k) + \sigma_S(y_k) - z_k^T d_k + \frac{1}{q} \epsilon \left[\max_{k'=1}^K \|z_{k'}\|_* \right]^{1-q} \|z_k\|_*^q \\
& \qquad + \frac{q-1}{q} \epsilon \max_{k'=1}^K \|z_{k'}\|_* \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{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, \ldots, K, \\
& [-g]^*(z_k - y_k) + \sigma_S(y_k) - z_k^T d_k \\
& \qquad +\epsilon \|z_k\|_* \left[\frac{1}{q} \left[\max_{k'=1}^K \frac{\|z_{k'}\|_*}{\|z_k\|_*} \right]^{1-q} + \frac{q-1}{q} \max_{k'=1}^K \frac{\|z_{k'}\|_*}{\|z_k\|_*} \right] \leq s_k \\
& \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad k = 1, \ldots, K \\
& (q-1)^{1/4} \leq \|z_k\|_* \leq (q-1)^{-1/4} \quad k = 1, \ldots, K
\end{array} \right. \\
&\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, \ldots, K, \\
& [-g]^*(z_k - y_k) + \sigma_S(y_k) - z_k^T d_k + \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{array} \right.
\end{aligned}
$$

The first inequality follows from the choice $\lambda_k = \left[\frac{q-1}{q} \epsilon^{\frac{1}{1-q}} \max_{k'=1}^K \|z_{k'}\|_* \right]$ and by imposing the restrictions $(q-1)^{1/4} \leq \|z_k\|_* \leq (q-1)^{-1/4}$ for all $k = 1, \ldots, K$. The first equality follows from 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

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

The second equality follows by pulling $\epsilon \|z_k\|_* > 0$ out of the last two terms in the constraints. To establish the last inequality 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 A.2. Let

$$\bar{g}_u^K(x;p) = \left\{ \begin{array}{ll} \text{minimize} & \sum_{k=1}^K w_k s_k \\ \text{subject to} & z_k \in \mathbf{R}^m,\ y_k \in \mathbf{R}^m,\ s_k \in \mathbf{R}^m \quad k = 1, \ldots, K, \\ & [-g]^*(z_k - y_k) + \sigma_S(y_k) - z_k^T d_k + \epsilon \|z_k\|_* D\left(\frac{p}{p-1}\right) \leq s_k \\ & \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad k = 1, \ldots, K \\ & (p-1)^{-1/4} \leq \|z_k\|_* \leq (p-1)^{1/4} \quad k = 1, \ldots, K. \end{array} \right. \tag{26}$$

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 $\bar{g}^K(x;\infty) = \lim_{p\to\infty} \bar{g}^K(x;p) \leq \liminf_{p\to\infty} \bar{g}_u^K(x;p)$. In fact, as the function $D\left(\frac{p}{p-1}\right)$ defined in Lemma A.2 is nonincreasing for all p sufficiently large this implies that $\bar{g}^K(x;p)$ is nonincreasing for p sufficiently large and hence we have $\bar{g}^K(x;\infty) = \lim_{p\to\infty} \bar{g}^K(x;p) \leq \liminf_{p\to\infty} \bar{g}_u^K(x;p) = \lim_{p\to\infty} \bar{g}_u^K(x;p)$. 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 $\{(z_k^t,\, y_k^t,\, s_k^t = [-g]^*(z_k^t - y_k^t) + \sigma_S(y_k^t) - (z_k^t)^T d_k + \epsilon \|z_k^t\|_*)\}_{t\geq 1}$ in the optimization problem characterizing $\bar{g}^K(x;\infty)$ in Equation (25) 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 $\{(\tilde{z}_k^t,\, \tilde{y}_k^t = y_k^t + (\tilde{z}_k^t - z_k^t),\, \tilde{s}_k^t = [-g]^*(\tilde{z}_k^t - \tilde{y}_k^t) + \sigma_S(\tilde{y}_k^t) - (\tilde{z}_k^t)^T d_k + \epsilon \|\tilde{z}_k^t\|_* D\left(p_t/(p_t - 1)\right))\}_{t\geq 1}$ is by construction feasible in the minimization problem characterizing the function $\bar{g}_u^K(x;p_t)$

in Equation (26). Hence, finally, we have

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

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

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

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

Lemma A.1. *We have*

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

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

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

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

$$\left[(q-1)\phi(q)\right]^{1/q} = \left[(q-1)^q/q^q\right]^{1/q} = (q-1)/q$$

and

$$-\frac{p}{q} = -\frac{1}{\frac{1}{p}q} = -\frac{1}{(1-1/q)q} = -\frac{1}{q-1} = \frac{1}{1-q}.$$

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 A.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 $\lim_{q\to 1}(q-1)^{1/4}\left(D(q)-1\right) = 0$.

Proof. Observe that 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. ■

A.7 Proof of the equivalence between $\mathbf{p = 1}$ and $\mathbf{p = \infty}$ for affine uncertainty

We again consider the single affine constraint (9). In formulation (11), when $p = 1$, we observe from Kuhn et al. (Kuhn et al., 2019, Section 2.2 Remark 1) that

$$\lim_{q\to\infty} \phi(q)\lambda \left\|z_k/\lambda\right\|_*^q = \begin{cases} 0 & \text{if } \|z_k\| \le \lambda \\ \infty & \text{otherwise,} \end{cases}$$

so when the support is $S = \mathbf{R}^m$, the reformulation becomes

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & a^T x - b + \lambda\epsilon + (P^T x)^T \sum_{k=1}^K w_k \bar{d}_k \le 0 \\ & \|P^T x\|_* \le \lambda \\ & \lambda \ge 0, \end{array} \tag{27}$$

where we can make the substitution $\lambda = \|P^T x\|_*$, in which case this becomes equivalent to (14).

A.8 Proof of convex reduction of the worst-case problem (21)

We assume all preconditions given in Section 2.2. Referencing the proof for the case where $p = 1$ in Esfahani and Kuhn (2018), we first expand-out the definition of the expected value and the Wasserstein-ball constraint. Then, we replace the joint distribution by a conditional one, since one of the distributions is the known empirical distribution, given by data. We use K instead of N and w_i instead of $1/N$ such that this generalizes to ambiguitity sets defined

as the Wasserstein-ball around the weighted empirical distribution $\mathbf{P}^K$ of the clustered and averaged dataset.

$$\sup_{\mathbf{Q}\in\mathbf{B}_\epsilon^p(\hat{\mathbf{P}}^K)} \mathbf{E}^{\mathbf{Q}}[g(u,x)] = \begin{cases} \sup\limits_{\Pi,\mathbf{Q}} & \int_S g(u,x)\mathbf{Q}(\mathrm{d}u) \\ \text{s.t.} & \int_S \|u-u'\|^p \Pi(\mathrm{d}u,\mathrm{d}u') \le \epsilon^p \end{cases}$$
$$= \begin{cases} \sup\limits_{\mathbf{Q_k}\in\mathcal{M}(S)} & \sum_{k=1}^K w_k \int_S g(u,x)\mathbf{Q_k}(\mathrm{d}u) \\ \text{s.t.} & \sum_{k=1}^K w_k \int_S \|u-\bar{d}_k\|^p \mathbf{Q_k}(\mathrm{d}u) \le \epsilon^p. \end{cases}$$

Next, we take the Lagrangian and utilize the definition of conjugacy.

$$= \begin{cases} \sup\limits_{\mathbf{Q_k}\in\mathcal{M}(S)} \inf\limits_{\lambda\ge 0} & \sum_{k=1}^K w_k \int_S g(u,x)\mathbf{Q_k}(\mathrm{d}u) + \lambda(\epsilon^p - \sum_{k=1}^K w_k \int_S \|u-\bar{d}_k\|^p \mathbf{Q_k}(\mathrm{d}u)) \end{cases}$$
$$= \begin{cases} \inf\limits_{\lambda\ge 0} \sup\limits_{\mathbf{Q_k}\in\mathcal{M}(S)} & \lambda\epsilon^p + \sum_{k=1}^K w_k \int_S g(u,x) - \lambda\|u-\bar{d}_k\|^p \mathbf{Q_k}(\mathrm{d}u) \end{cases}$$
$$= \begin{cases} \inf\limits_{\lambda\ge 0} & \lambda\epsilon^p + \sup\limits_{u=(v_1,\ldots,v_K)\in S} \sum_{k=1}^K w_k (g(v_k,x) - \lambda\|v_k-\bar{d}_k\|^p), \end{cases}$$

where the second equality is due to a well-known strong duality result for moment problems (Esfahani and Kuhn, 2018). Now, separating g into its constituent functions following (Esfahani and Kuhn, 2018, Theorem 4.2),

$$= \begin{cases} \inf\limits_{\lambda\ge 0} & \lambda\epsilon^p + \sum_{k=1}^K w_k s_k \\ \text{subject to} & \sup\limits_{v_k\in S} g_j(v_k,x) - \lambda\|v_k-\bar{d}_k\|^p \le s_k \quad k=1,\ldots,K, \quad j=1,\ldots,J \end{cases}$$
$$= \begin{cases} \inf\limits_{\lambda\ge 0} & \sum_{k=1}^K w_k s_k \\ \text{subject to} & [-g_j + \mathcal{I}_S + \lambda c_k]^*(0) \le s_k \quad k=1,\ldots,K, \quad j=1,\ldots,J, \end{cases}$$

where $c_k(v_k) = \|v_k - \bar{d}_k\|^p$. The last expression is identical to the form in Appendix A.1, except now with multiple j, so the final dual is equivalent to the dualized constraint in (19).

A.9 Proof of Theorem 4.1

We prove (i) $\bar{g}^N(x) \le \bar{g}^K(x)$, (ii) $\bar{g}^K(x) \le \bar{g}^{N*}(x) + (L/2)D(K)$, and (iii) when the support constraint does not affect the uncertainty set, $\bar{g}^K(x) \le \bar{g}^N(x) + (L/2)D(K)$.

Proof of (i). We begin with a feasible solution $v_1,\ldots,v_N$ of (MRO-N), and set $u_k = \sum_{i\in C_k} v_i/|C_k|$ for each of the K clusters. We see u_k with $k=1,\ldots,K$ satisfies the constraints

of (MRO-K), as

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

where we have applied triangle inequality, Jensen's inequality for the convex function $f(x) = \|x\|^p$, and the constraint of (MRO-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), it is a feasible solution. We now show that for this pair of feasible solutions, in terms of the objective value, $\bar{g}^K(x) \ge \bar{g}^N(x)$. By assumption, g is concave in the uncertain parameter, so by Jensen's inequality,

$$\begin{aligned}\sum_{k=1}^{K}\frac{|C_k|}{N}g\left(\frac{1}{|C_k|}\sum_{i\in C_k}v_i, x\right) &\ge \sum_{k=1}^{K}\frac{|C_k|}{N}\frac{1}{|C_k|}\sum_{i\in C_k}g(v_i)\\ \sum_{k=1}^{K}\frac{|C_k|}{N}g(u_k,x) &\ge \frac{1}{N}\sum_{i\in N}g(v_i).\end{aligned}$$

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

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

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

where the inequality makes use of the constraint of (MRO-K). Since the constraints are satisfied, the constructed $v_i \dots v_N$ are a valid solution for (MRO-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 $\Delta_k = u_k - \bar{d}_k \geq 0$ in the solution of (MRO-K), as the maximization of g over u_k wil lead to $u_k \geq \bar{d}_k$. 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 (Beck, 2017, Theorem 5.8) on $-g$: $\forall v_1, v_2 \in S, \lambda \in [0,1]$,

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

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

$$g\left(\frac{1}{|C_k|}v_1 + \frac{|C_k|-1}{|C_k|}\bar{v}_2\right) \leq \frac{1}{|C_k|}g(v_1) + \frac{|C_k|-1}{|C_k|}g(\bar{v}_2) + \frac{L}{2}\frac{1}{|C_k|}\frac{|C_k|-1}{|C_k|}\|v_1 - \bar{v}_2\|_2^2,$$

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:

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

where we used the equivalence

$$\sum_{i=2}^{|C_k|}\frac{i-1}{i}\left\|d_i - \frac{\sum_{j=1}^{i-1} d_j}{i-1}\right\|_2^2 = \sum_{i\in C_k}\|d_i - \bar{d}_k\|_2^2.$$

Now, summing over all clusters, we have

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

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

A.10 Proof of Theorem 4.2

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

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

we have obtained a valid solution for (MRO-N-Dual). The increase in the objective value from $\bar{g}^K(x)$ to that of the constructed solution of (MRO-N-Dual) 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 $\delta(K, z, \gamma) \geq 0$, as

$$\begin{aligned} \delta(K, z, \gamma) &= (1/N) \sum_{k=1}^{K} \sum_{i \in C_k} \max_j \{(-z_{jk} - H^T \gamma_{jk})^T (d_i - \bar{d}_k)\} \\ &\geq (1/N) \sum_{k=1}^{K} \sum_{i \in C_k} (-z_{1k}^T - \gamma_{1k}^T H)(d_i - \bar{d}_k) \\ &= (|C_k|/N) \sum_{k=1}^{K} (-z_{1k} - H^T \gamma_{1k})^T (1/|C_k|) \sum_{i \in C_k} (d_i - \bar{d}_k) \\ &= (|C_k|/N) \sum_{k=1}^{K} (-z_{1k} - H^T \gamma_{1k})^T 0 \\ &= 0 \end{aligned}$$

The constructed feasible solution for (MRO-N-Dual) is an upper bound for its optimal solution, since we solve a minimization problem. As this holds true for any solution of (MRO-K-Dual), we must have $\bar{g}^K(x) + \delta(K, z, \gamma) \geq \bar{g}^N(x)$, which translates to $\bar{g}^N(x) - \delta(K, z, \gamma) \leq \bar{g}^K(x)$.

Proof of the upper bound. We use the primal formulations of the MRO constraints. We first solve (MRO-K) to obtain a feasible solution $u_{11}, \ldots, u_{JK}$. 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 (MRO-N*), as

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

The constructed solutions remain in $\mathbf{dom}_u$, following the arguments in the proof of (ii) in Appendix A.9. Now, for each cluster k and function g_j, we note using the L-smooth condition on $-g_j$,

$$\alpha_{jk}g_j\left(\frac{1}{|C_k|}\sum_{i\in C_k}v_{ji},x\right) \leq \alpha_{jk}\left(\sum_{i\in C_k}\frac{1}{|C_k|}g_j(v_{ji},x)+\frac{L_j}{2|C_k|}\sum_{i=2}^{|C_k|}\frac{i-1}{i}\left\|d_i-\frac{\sum_{j=1}^{i-1}d_j}{i-1}\right\|_2^2\right)$$

$$\alpha_{jk}g_j(\bar{d}_k+\Delta_k,x) \leq \sum_{i\in C_k}\alpha_{ji}g_j(v_{ji},x)+\frac{\alpha_{ji}L_j}{2}\sum_{i\in C_k}\|d_i-\bar{d}_k\|_2^2$$

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

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

Since this holds for all feasible solutions of (MRO-K), we conclude that

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

A.11 Conjugate derivation for the Capital Budgeting problem (22)

We begin with

$$g(u,x) = -\eta(u)^T x = -\sum_{j=1}^{n}\sum_{t=0}^{T}F_{jt}x_j/(1+u_j)^t,$$

and take its conjugate $[-g]^*(z)$ in the uncertain parameter u. We use the theorem on infimal convolutions (Rockafellar and Wets, 1998, Theorem 11.23(a), p. 493) to arrive at

$$[-g]^*(z) = \sum_{t=0}^{T}\sup_u\left(u^T y_t-\left[\sum_{j=1}^{n}F_{jt}x_j(1+u_j)^t\right]\right), \quad \sum_{t=0}^{T}y_t = z.$$

We calculate the inner conjugates using the the first order optimality condition,

$$\begin{aligned}
\nabla\left(y_t^T u - \sum_{j=1}^{n} F_{jt}x_j(1+u_j)^{-t}\right) &= 0 \\
y_{jt} = -tF_{jt}x_j(1+u_j^\star)^{-(t+1)}, \quad j &= 1,\ldots,n \\
u_j^\star = (tF_{jt}x_j(-y_{jt})^{-1})^{1/(t+1)} - 1, \quad j &= 1,\ldots,n.
\end{aligned}$$

Substituting this back into the expression for the conjugate, we have, for each j and t,

$$\begin{aligned}
y_{jt}u_j^\star - F_{jt}x_j(1+u_j)^{-t} &= y_{jt}(tF_{jt}x_j(-y_{jt}^{-1})^{(1/(t+1)} - 1)) \\
&\quad - F_{jt}(tF_{jt}x_j(-yjt)^{-1})^{-t/(t+1)} \\
&= -y_{jt} - ((-y_{jt})^{t/(t+1)}(F_{jt}x_j)^{1/(t+1)})(t^{1/(t+1)} + t^{-1/(t+1)}),
\end{aligned}$$

after combining terms. Note that $-(-y_{jt})^{t/(t+1)}(F_{jt}x_j)^{1/(t+1)}$ can be replaced by an auxiliary variable $\delta_{jtk} \leq 0$, by introducing the power cone constraint $(-y_{jt})^{t/(t+1)}(F_{jt}x_j)^{1/(t+1)} \geq |\delta_{jtk}|$. By substituting these results into (7) and further vectoring some constraints, we arrive at the desired formulation.