

# Mean Robust Optimization

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

September 21, 2022

## 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  $\epsilon > 0$ . When the reference distribution is an empirical distribution, the associated Wasserstein DRO can be formulated as a convex minimization problem where the number of constraints grows linearly with the number of data points (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).

## 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 Wasserstein DRO problem. With just one cluster, our formulation corresponds to the classical RO approach. The number of clusters is a tunable parameter that provides a trade-off 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 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](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 (2008); Bertsimas et al. (2011), and the books by Ben-Tal et al. (2009) and 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, 2016) and enjoy both a tractable primal as well as a tractable dual formulation. We refer to Chen and Paschalidis (2020) for a thorough overview of DRO, and to 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.

**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. (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 (2016) consider a robust formulation of the 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 the Wasserstein DRO by defining a new ambiguity set as a subset of the original set, containing all distributions supported on  $NK$  points with equal probability  $1/(NK)$ , as opposed to the original 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.

**Probabilistic guarantees in robust and distributionally optimization.** 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 (2018) obtain finite-sample guarantees for Wasserstein DRO for selecting the radius  $\epsilon$  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 (2020) derives finite-sample guarantees for Wasserstein DRO for selecting  $\epsilon$  of order  $N^{-1/2}$  under specific assumptions. These bounds, however, are not tight in practice, and typically result in overly-conservative  $\epsilon$ . We provide theoretical results of a similar vein, with a slightly increased  $\epsilon$  to compensate for information lost through clustering and achieve the same probabilistic guarantees. However, we note that when  $\epsilon$  is chosen through empirical experimentation, our formulation by being lower dimensional is overall much faster to solve.

**Scenario reduction.** First introduced by 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. (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 (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. In MRO, we adapt and extend the scenario reduction approach to Wasserstein DRO, where we fix the reduced scenario points to ones found by  $k$ -means clustering, then provide performance bounds on the DRO problem depending on the number of clusters.

**Data compression in data-driven problems.** 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 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 in light of both robust and distributionally robust optimization, and give theoretical guarantees on constraint satisfaction. In Section 3, we analyze the effect on clustering on the worst-case value of the MRO solutions. In Section 4, we give guidelines for choosing hyperparameters. In Section 5, we provide computational verification of the speedups obtained through our methodology. In Section 6, we summarize our conclusions.

## 2 Mean robust optimization

### 2.1 The problem

We consider an uncertain constraint of the form,

$$g(u, x) \leq 0, \quad (1)$$

where  $x \in \mathbf{R}^n$  is the optimization variable,  $u \in \mathbf{R}^m$  is an uncertain parameter, and  $g$  is a function concave in  $u$  for any  $x$ . We assume  $-g$  is proper, convex, and upper-semicontinuous in  $u$ . 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,

$$\mathbf{P}(g(u, x) \leq 0) \geq 1 - \alpha. \quad (2)$$

We define, instead, an adjusted version of the uncertain constraint (1) as

$$\mathbf{E}^{\mathbf{P}}(g(u, x)) \leq 0, \quad (3)$$

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

**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  $S^N$ . Throughout this paper, we assume  $S$  to live within the domain of  $g$  for the variable  $u$ , which we will refer to as  $\mathbf{dom}_u g$ , i.e.,  $S \subseteq \mathbf{dom}_u g$ . 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

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

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

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

where we partition  $\mathcal{D}_N$  into  $K$  disjoint subsets  $C_k$ , and  $\bar{d}_k$  is the centroid of the  $k$ th subset, for  $k = 1, \dots, K$ . 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  $\epsilon$  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.

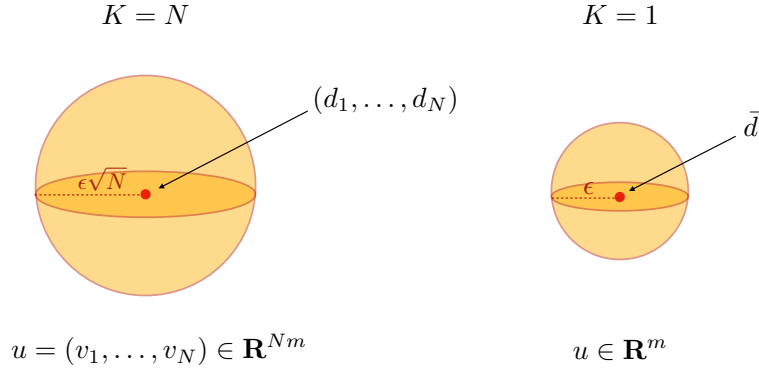


Figure 1: Visualizing the uncertainty set  $\mathcal{U}(N, \epsilon)$  and  $\mathcal{U}(1, \epsilon)$  as high dimension balls when  $p = 2$ .

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

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

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- $\infty$  distance of  $\bar{d}$ . We note that, if any of the decoupled constraints are violated, then  $\lim_{p \rightarrow \infty} \sum_{k=1}^K w_k \|v_k - \bar{d}_k\|^p \geq \epsilon^p$ , 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  $\epsilon$  around the empirical mean of the entire dataset,  $\mathcal{U}(1, \epsilon) = \{v \in S \mid \|v - \bar{d}\| \leq \epsilon\}$ . This is equivalent to the uncertainty set of traditional robust optimization, 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.

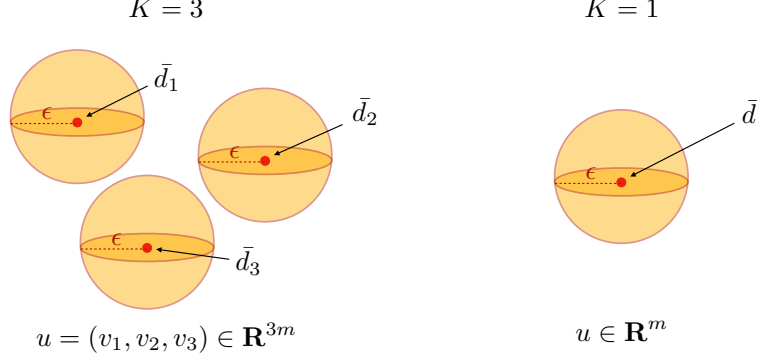


Figure 2: Visualizing the decoupled uncertainty set  $\mathcal{U}(K, \epsilon)$  with  $p = \infty$ .

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

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

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

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

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  $\epsilon$  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.

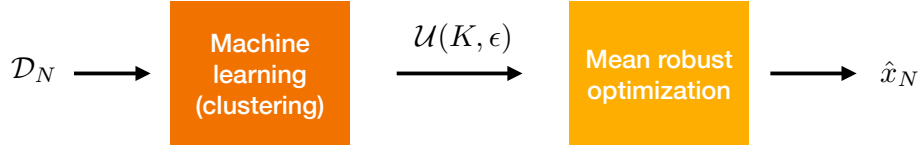


Figure 3: Mean robust optimization procedure.

## 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.

### 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

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

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

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

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,  $[-g]^*(z, x) = \sup_{u \in \text{dom}_u g} z^T u - [-g(u, x)]$  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  $q > 1$ ,  $\phi(\infty)$  decays as  $1/q$  (Kuhn et al., 2019, Theorem 8). Note that  $q$  satisfies  $1/p + 1/q = 1$ , *i.e.*,  $q = p/(p-1)$ . The support function  $\sigma_S$  is also the conjugate of  $\mathcal{I}_S$ , which is defined  $\mathcal{I}_S(u) = 0$  if  $u \in S$ , and  $\infty$  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, 2022, Section 2.5) and (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 2.4.

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

$$(a + Pu)^T x \leq b, \quad (8)$$

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

$$[-g]^*(z, x) = \sup_u z^T u + b - (a + Pu)^T x = \begin{cases} a^T x - b & \text{if } z + P^T x = 0 \\ \infty & \text{otherwise.} \end{cases} \quad (9)$$

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

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



where the number of variables or constraints does not depend on  $K$ . Since vector  $\sum_{k=1}^K w_k \bar{d}_k$  is the average of the datapoints in  $\mathcal{D}_N$  for any  $K \in \{1, \dots, N\}$ , 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

$$\begin{aligned} & \text{minimize} && f(x) \\ & \text{subject to} && \left\{ \begin{array}{l} \text{maximize}_{v_1 \dots v_K \in S} \quad \sum_{k=1}^K w_k g(v_k, x) \\ \text{subject to} \quad \|v_k - \bar{d}_k\| \leq \epsilon, \quad k = 1, \dots, K \end{array} \right\} \leq 0, \end{aligned} \quad (11)$$

which has a reformulation where the constraint above is dualized,

$$\begin{aligned} & \text{minimize} && f(x) \\ & \text{subject to} && \sum_{k=1}^K w_k s_k \leq 0 \\ & && [-g]^*(z_k - y_k, x) + \sigma_S(y_k) - z_k^T \bar{d}_k + \epsilon \|z_k\|_* \leq s_k \quad k = 1, \dots, K, \end{aligned} \quad (12)$$

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, (11) is the limiting case of (6) as  $p \rightarrow \infty$ . In terms of the reformulated problem with dualized constraints, problem (12) is the limiting case of (7) as  $p \rightarrow \infty$ . The proofs are delayed to Appendix A.3 and Appendix A.4 respectively.*

**Example with affine constraints.** Consider again the case of affine constraint as in (8) with support set  $S = \mathbf{R}^m$ , now with  $p = \infty$ . Following a similar derivation as (10), we substitute the conjugate function  $[-g]^*$  (9) in problem (12), we can obtain

$$\begin{aligned} & \text{minimize} && f(x) \\ & \text{subject to} && a^T x - b + \epsilon \|P^T x\|_* + (P^T x)^T \sum_{k=1}^K w_k \bar{d}_k \leq 0, \end{aligned} \quad (13)$$

where the number of constraints and variables does not depend on  $K$ . Similairy to problem (10), the term  $\sum_{k=1}^K w_k \bar{d}_k$  is the average of the datapoints in  $\mathcal{D}_N$  for any  $K \in \{1, \dots, N\}$ . Therefore, the choice of  $K$  does not affect this formulation.

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, for any  $\epsilon$  and norm, the convex reformulations for  $p = 1$  and  $p = \infty$  are identical. The proof appears in Appendix A.5.*

### 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,

$$\begin{aligned} & \text{minimize} && f(x) \\ & \text{subject to} && \bar{g}(u, x) \leq 0, \quad \forall u \in \hat{S}, \end{aligned} \tag{14}$$

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

$$\begin{aligned} & \text{maximize} && \bar{g}(u, x^k) \\ & \text{subject to} && u \in \mathcal{U}(K, \epsilon) \end{aligned} \tag{15}$$

The procedure works as follows. We first solve (14) with a set  $\hat{S} = \{\bar{u}\}$ , where  $\bar{u}$  is nominal value of the uncertainty, obtaining  $x^k$ . Then, we solve (15), 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 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)

---

```

1: given  $\hat{S} = \{\bar{u}\}$ 
2: for  $k = 1, \dots, k_{\max}$  do
3:    $x^k \leftarrow$  solve minimization problem (14) over  $x$ 
4:    $u^k \leftarrow$  solve maximization problem (15) over  $u$ 
5:   if  $\bar{g}(u^k, x^k) > 0$  then
6:      $\hat{S} \leftarrow \hat{S} \cup \{u^k\}$ 
7:   else
8:     return  $x^k$ 

```

---

## 2.4 Links to Wasserstein distributionally robust optimization

Distributionally robust optimization (DRO) solves the problem

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

where the ambiguity set  $\mathcal{P}_N$  contains, with high confidence, all distributions that could have generated the training samples  $\mathcal{D}^N$ , such that the probabilistic guarantee (4) is satisfied. Wasserstein DRO constructs  $\mathcal{P}_N$  as a ball of radius  $\epsilon$  with respect to the Wasserstein metric around the empirical distribution  $\hat{\mathbf{P}}^N = \sum_{i=1}^N \delta_{d_i}/N$ , where  $\delta_{d_i}$  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 2.5), and

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

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

When  $K = N$ , the constraint of the DRO problem (16) is equivalent to the constraint of (MRO). In particular, for case  $p \geq 1$ , the expression

$$\sup_{\mathbf{Q} \in \mathbf{B}_\epsilon^p(\hat{\mathbf{P}}^N)} \mathbf{E}^{\mathbf{Q}}(g(u, x)), \quad (17)$$

is equivalent to the dual of the constraint of (6), when  $K = N$ , and  $w_k = 1/N$ . This is noted without a written-out proof in (Kuhn et al., 2019). We give a proof of strong duality in Appendix A.6. By the same logic, in the case where  $p = \infty$ , the expression is equivalent to the dual of the constraint of (11). Given the above reductions, we can rewrite the Wasserstein DRO problem in the same form as (7), 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.

## 2.5 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  $p \geq 1$ .** Wasserstein 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  $A = \mathbf{E}^{\mathbf{P}}(\exp(t\|u\|^a)) = \int_D \exp(t\|u\|^a) \mathbf{P}(du) < \infty$ . We refer to the following theorem.

**Theorem 2.1** (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  $\phi$  is an exponentially decaying function of  $N$ .

Theorem (2.1) estimates the probability that the unknown data-generating distribution  $\mathbf{P}$  lies outside the Wasserstein ball  $\mathbf{B}_\epsilon(\hat{\mathbf{P}}^N)$ , which is our ambiguity set. Thus, we can estimate the smallest radius  $\epsilon$  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  $\beta$ , 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., 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 2.2** (Measure concentration,  $p = \infty$  (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  $\rho : S \rightarrow (0, \infty)$ , 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  $\phi$  is an exponentially decaying function of  $N$ .

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

**Theorem 2.3** (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 maximum distance of any data-point in  $\mathcal{D}_N$  from its assigned cluster center, and  $\hat{x}_N$  is the optimal solution to (MRO) with uncertainty set  $\mathcal{U}(K, \epsilon_N(\beta) + \eta_N(K))$ , 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

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

If we introduce a new parameter,  $\eta_N(K)$ , defined as

$$\|\bar{d}_k - d_i\|^p \leq \eta_N(K)^p \quad \forall d_i \in C_k, \quad k = 1, \dots, K,$$

the maximum distance with respect to the norm used in the Wasserstein metric, of any

data-point in  $\mathcal{D}_N$  from its assigned cluster center  $\bar{d}_k$ , we notice that

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

where we have replaced the integral with a finite sum, as the distributions are discrete. Therefore, by Theorems 2.1, 2.2 and the triangle inequality for the Wasserstein metric,

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

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

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

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

**Multiple uncertain constraints.** Up to now we have been interested in solving problems of the form (MRO) with one uncertain constraint  $g(u, x)$ . With multiple uncertain constraints  $g_l(u, x)$ ,  $l = 1, \dots, L$ , where, again we assume each function  $-g_l(u, x)$  to be proper, convex, and upper semicontinuous in  $u$ , we treat the uncertain constraints independently to obtain a separate set of dual constraints for each. We obtain the probabilistic guarantee

$$\mathbf{P}^N \left( \max_l \mathbf{E}^{\mathbf{P}}(g_l(u, \hat{x}_N)) \leq 0 \right) \geq 1 - \beta. \quad (18)$$

### 3 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.

In this section, we then show the effects of clustering on the *worst-case value of the constraint function in* (MRO). In particular, informally, 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  $K$ .

**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)

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

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

and

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

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 uncertainty set*, we refer to situations where the constraints  $v_i \in S$  for  $i = 1, \dots, N$  are binding. We also require the following assumption to hold.

**Assumption 3.1.** *The domain  $\text{dom}_u g$  is  $\mathbf{R}^m$ . Otherwise,  $g$  is either monotonically increasing or monotonically decreasing in  $u$ .*

With this assumption on the domain and curvature of  $g$ , we can then construct solutions for (MRO-N), (MRO-K), and (MRO-N\*) to prove the following relations.

**Theorem 3.1.** *Suppose that Assumption 3.1 holds, and  $-g$  satisfies an  $L$ -smooth condition on its domain with respect to the  $L_2$ -norm, given as*

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

*Then, with the same  $x$  and  $\epsilon$ , and for any integer  $p \geq 1$ , we always have*

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

*If in addition  $S$  does not affect the uncertainty set  $\mathcal{U}(K, \epsilon)$ , we have*

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

The proof is delayed to Appendix A.7. 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  $\Delta$  be the maximum difference in constraint value resultant from relaxing the support constraint on the MRO uncertainty sets, *i.e.*,  $\Delta = \max_x (\bar{g}^{N^*}(x) - \bar{g}^N(x))$ , subject to  $x$  being feasible for problem (MRO). 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  $\Delta$  is independent of  $K$ .

**Remark 3.1.** *While  $\Delta$  could be constructed to be arbitrarily bad, in practice, we expect our relevant range of  $\epsilon$  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 5.1 and 5.2 for examples.*

**Uncertain objective.** When the uncertainty is in the objective, Theorem 3.1 quantifies the difference in optimal values.

**Corollary 3.1.1.** *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.*

*Proof.* The feasible region of the MRO problem with  $K$  and  $N$  does not change with clustering, as there are no further constraints that involve the uncertain parameter. From Theorem 3.1, for any fixed  $\hat{x}$ , we have  $\bar{g}^K(\hat{x}) - \bar{g}^N(\hat{x}) \leq (L/2)D(K) + \Delta$ . Therefore, no matter what  $\hat{x}$  the MRO problem with  $N$  clusters selects, the MRO problem with  $K$  clusters can always choose the same  $\hat{x}$  to be at most  $(L/2)D(K) + \Delta$  higher in value. ■

**Uncertain constraints.** When the uncertainty is in the constraints, the difference between  $\bar{g}^K(x)$  and  $\bar{g}^N(x)$  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 3.1.2** (Clustering with affine dependence on the uncertainty). *If  $g(u, x)$  is affine in  $u$  and the uncertainty set 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 3.1, 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 (8), we observe from (10) 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  $K$ . ■

This result can be extended to the formulation with multiple uncertain constraints.

**Corollary 3.1.3.** *If the support does not affect the uncertainty set and we have multiple uncertain functions  $g_l(u, x)$ , each of which is affine in  $u$  for all  $l \in \{1, \dots, L\}$ , and we treat them independently as in (18), then clustering makes no difference to the optimal value and optimal solution.*

On the other hand, if  $S$  affects the uncertainty set, clustering may result in a difference of at most  $\Delta$  in the objective value.

## 4 Parameter selection

**Choosing  $K$ .** When the uncertain constraint is affine and  $S$  does not affect the uncertainty set, 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 uncertainty set, there is a difference of at most  $\Delta$  between setting  $K = 1$  and  $K = N$ , which can often be approximated  $\approx 0$  for small  $\epsilon$ . 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 3.1. 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. No matter if the uncertainty lies in the objective or the constraints, this bound will inform us of the potential difference between choosing different  $K$ . Note that, by directly returning  $D(K)$ , this procedure can be completed in the clustering step without having to solve the downstream optimization problem.



**Choosing  $\epsilon$ .** While we have outlined theoretical results in Theorem 2.3 for choosing  $\epsilon$ , in practice, we experimentally select  $\epsilon$  through cross validation to arrive at the desired guarantee. Therefore, while the theoretical bounds suggest to choose a larger  $\epsilon$  when we cluster, this may not be the case experimentally. In fact, we show a powerful result in the upcoming concave numerical examples: although for the same  $\epsilon$ , MRO with  $K$  clusters is more conservative than Wasserstein DRO ( $N$  clusters), there are cases where we can tune  $\epsilon$  such that MRO and DRO provide almost identical tradeoffs between objective values and probabilistic guarantees, such that no loss in performance results from choosing a smaller cluster number  $K$ , see Sections 5.2, 5.3 and 5.4.

## 5 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 at

[https://github.com/stellatogrp/mro\\_experiments](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.

### 5.1 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  $i = 1, \dots, n$ , with cost  $c_i$ . In addition, we would like to satisfy the uncertain demand  $u \in \mathbf{R}^m$  at minimal cost. We define variable  $X \in \mathbf{R}^{n \times m}$  where  $X_{ij}$  corresponding to the portion of the demand of customer  $j$  shipped from facility  $i$  with corresponding cost  $C_{ij}$ . Furthermore,  $r \in \mathbf{R}^n$  represents the production capacity for each facility, and  $u \in \mathbf{R}^m$  represents the uncertain demand from each customer. For each customer  $j$ ,  $X_j$  represents the proportion of goods shipped from any facility to that customer, which sums to 1. For each facility  $i$ ,  $(X^T)_i$  represents the proportion of goods shipped to any customer. Putting this all together, we obtain multiple uncertain capacity constraints for each facility, which are affine in  $u$ ,

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

Since demands are nonnegative, we assume nonnegative support and solve the problem, for  $p = \infty$ ,

$$\begin{aligned}
& \text{minimize} && c^T x + \text{tr}(C^T X) \\
& \text{subject to} && \mathbf{1}^T X_j = 1, \quad j = 1, \dots, m \\
& && \sum_{k=1}^K w_k s_{ik} \leq 0, \quad i = 1, \dots, n \\
& && \lambda_{ik} \epsilon + (X^T)_i \bar{d}_k - r_i x_i + \gamma_{ik} \bar{d}_k \leq s_{ik}, \quad i = 1, \dots, n, \quad k = 1, \dots, K \\
& && \|\gamma_{ik} + (X^T)_i\|_2 \leq \lambda_{ik} \\
& && \gamma_{ik} \geq 0 \quad i = 1, \dots, n, \quad k = 1, \dots, K \\
& && x \in \{0, 1\}^n, \quad X \in \mathbf{R}^{n \times m}.
\end{aligned}$$

We have variables  $x \in \{0, 1\}^n$ ,  $X \in \mathbf{R}^{n \times m}$ ,  $s_{ik} \in \mathbf{R}$ ,  $\gamma_{ik} \in \mathbf{R}^m$ , for  $i = 1, \dots, n$  and  $k = 1, \dots, K$ . The  $\gamma$  variables arise from enforcing nonnegative support  $S = \mathbf{R}_+^m$ .

Even though we obtain a probabilistic guarantee of the form (18), which can be interpreted as a separate probabilistic guarantee for each facility, we can still tune  $\epsilon$  to obtain the stricter guarantee

$$\mathbf{P}^N \left( \mathbf{E}^{\mathbf{P}} \left( \max_l g_l(u, \hat{x}_N) \right) \leq 0 \right) \geq 1 - \beta, \quad (19)$$

which guarantees that the uncertain constraints for all facilities are satisfied simultaneously.

**Problem setup.** To generate data, we set  $n = 5$  facilities,  $m = 25$  customers, and  $N = 50$  data samples. We generate  $c$  from a uniform distribution on  $[30, 70]$ , 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 generate  $r$  from a uniform distribution on  $[10, 50]$ , and generate demands  $d$  from a uniform distribution on  $[1, 6]$ . Since the underlying distribution is uniform, it satisfies the assumptions of Theorem 2.2, so theoretical guarantees also exist for  $p = \infty$ .

**Results.** As expected for linear  $g$ , we see in Figure 4 that the optimal value and probability of constraint satisfaction is unaffected by the number of clusters  $K$ . Even though there is a nonnegativity constraint for the support, the support  $S$  is large enough to not affect the MRO uncertainty sets for this range of  $\epsilon$ , so clustering makes no difference to the MRO solution. Choosing  $K = 1$  leads to a time reduction of up to 2 orders of magnitude while preserving the same optimal value. In fact, we see that the plots for  $K = 1^*$ , which is the problem solved without including the support constraint, overlaps the other plots. Therefore, for problems with linear uncertainty, and for which the support constraint only minimally affects the uncertainty set, we can use the simplification (13) where we remove the support constraint and trivially cluster all points. This leads to another slight time reduction, as we marked out using black squares on the bottom right of Figure 4.

Lastly, we note that the probabilistic guarantee (18) implied by our multiple uncertainty constraints formulation can be satisfied by very small  $\epsilon$  and a subsequently lower range of optimal values, while guarantee (19) requires a higher spectrum. No matter which guarantee is desired, however, we can always set  $K = 1$  with no performance loss and tune  $\epsilon$  to achieve the required value of  $\beta$ .

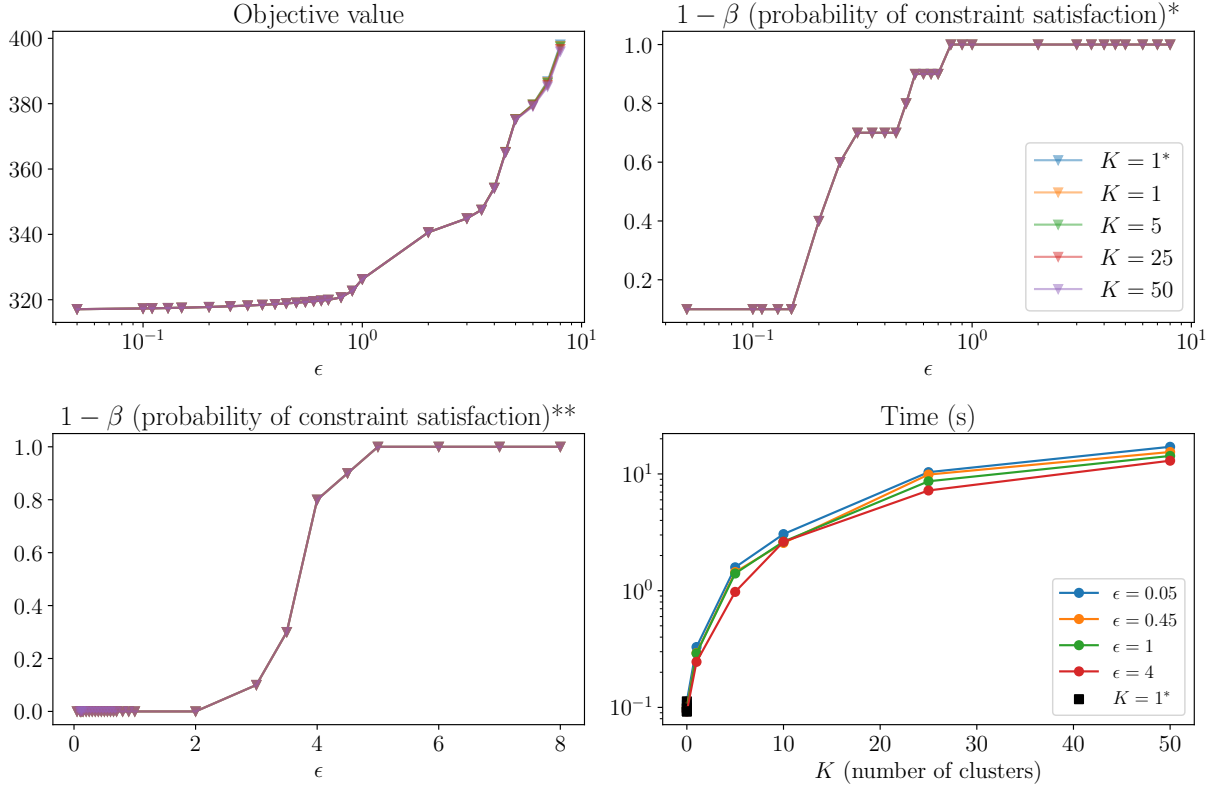


Figure 4: Facility location. Top left: in-sample objective values vs.  $\epsilon$  for different  $K$ . Top right\*:  $\epsilon$  vs. the original probabilistic guarantee (18) for different  $K$ . Bottom left\*\*:  $\epsilon$  vs. the stricter probabilistic guarantee (19). Bottom right: solve time.  $K = 1^*$  is the formulation without the support constraint.

## 5.2 Capital budgeting

We consider the capital budgeting problem in (Ben-Tal et al., 2015a, 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  $\theta$ . The NPV for project  $j$  is the sum of discounted cash flows  $F_{jt}$  over the years  $t = 0, \dots, T$ . We then formulate its negation as the uncertain function to be minimized

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

where  $u_j$  is the discount rate of project  $j$ , and  $x_j$  is the indicator for selecting project  $j$ , for  $j = 1, \dots, n$ . 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 3.1 and Theorem 3.1

applies. The robust problem can be written as

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

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

$$\begin{aligned}
& \underset{\tau}{\text{minimize}} && \tau \\
& \text{subject to} && \lambda \epsilon^2 + \sum_{k=1}^K w_k s_k \leq \tau \\
& && -F_0^T x + \mathbf{1}^T (\delta_k a - z_k) - z_k^T \bar{d}_k + \gamma_k^T (b - C \bar{d}_k) + 1/(4\lambda) \|C^T \gamma_k - z_k\|_2^2 \leq s_k, \\
& && k = 1, \dots, K \\
& && (-(Y_k)_{jt}, F_{jt} x_j, (\delta_k)_{jt}) \in \mathcal{K}^{t/(t+1)}, \quad j = 1, \dots, n, \quad t = 1, \dots, T, \quad k = 1, \dots, K \\
& && Y_k \mathbf{1} = z_k, \quad k = 1, \dots, K \\
& && h^T x \leq \theta \\
& && \lambda \geq 0, \quad \gamma_k \geq 0, \quad Y_k \leq 0, \quad \delta_k \leq 0, \quad x \in \{0, 1\},
\end{aligned} \tag{20}$$

where  $a \in \mathbf{R}^T$  with  $a_t = t^{1/(t+1)} + t^{-t/(t+1)}$  for  $t = 1, \dots, T$ , 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  $j = 1, \dots, n$ ,  $k = 1, \dots, K$ , and  $t = 1, \dots, T$ . The derivation of reformulation (20) is in Appendix A.8. 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  $\delta$ , as well as  $nT$  new power cone constraints, which greatly increases the computational 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  $j = 1, \dots, n$ ,  $t = 0, \dots, T$ . For all  $j$ ,  $h_j$  is generated from a uniform distribution on  $[1, 3 - 0.5j]$ , and the total budget  $\theta$  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  $j[0.005, 0.02]$ , and the other half on  $j[0.01, 0.025]$ , for all  $j$ . We calculate an upper bound on the  $L$ -smooth parameter,  $L = \|\nabla^2 \sum_{j=1}^n \sum_{t=0}^T F_{jt}(\hat{x}_N)_j (1 + u_j)^{-t}\|_{2,2} \leq \|\sum_{j=1}^n \sum_{t=0}^T t(t+1) F_{jt}(\hat{x}_N)_j\|_{2,2}$  for each data-driven solution  $\hat{x}_N$ .

**Results.** We observe in 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  $\epsilon$  to achieve the same performance and guarantees. Notice that the results for

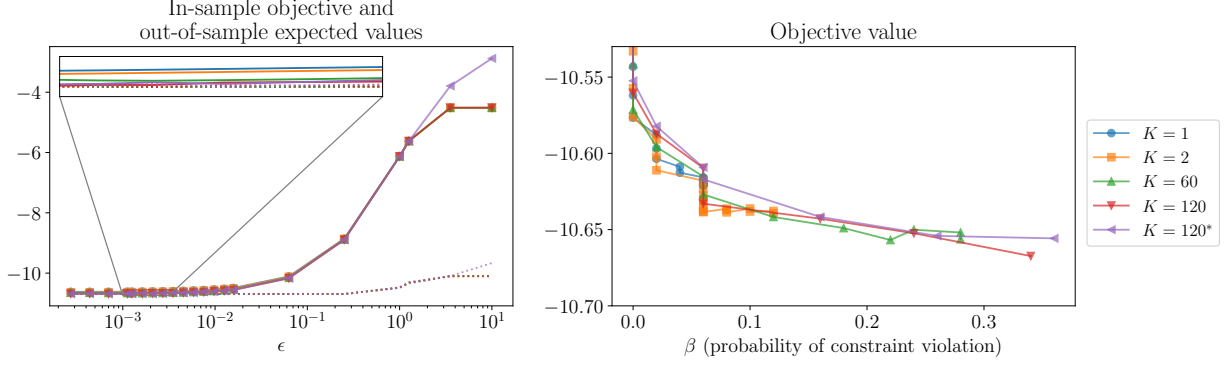


Figure 5: Capital budgeting. Left: in-sample objective values and out-of-sample expected values vs.  $\epsilon$  for different  $K$ . Solid lines are the in-sample objective value, dotted lines are the out-of-sample expected value. Right: objective value vs.  $\beta$  for different  $K$ ; each point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

$K = 120$  and  $K = 120^*$  are near identical for small  $\epsilon$ , where  $K = 120^*$  is the formulation without the support constraint. Therefore, while  $\bar{g}^{N^*}(x)$  slightly upper bounds  $\bar{g}^N(x)$ , we can approximate their difference  $\Delta \approx 0$  for small enough  $\epsilon$ , for which the upper bound  $(L/2)D(K)$  thus hold. In fact in this example, even for larger  $\epsilon$  where we observe  $\Delta > 0$ , the actual difference between  $\bar{g}^K$  and  $\bar{g}^N$  is still less than  $(L/2)D(K)$ . In Figure 6, we see that the elbow of the upper bound is at  $K = 2$ , and the true difference follows the same trend. 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.

### 5.3 Quadratic concave uncertainty

We refer to the example in (Ben-Tal et al., 2015a, 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

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

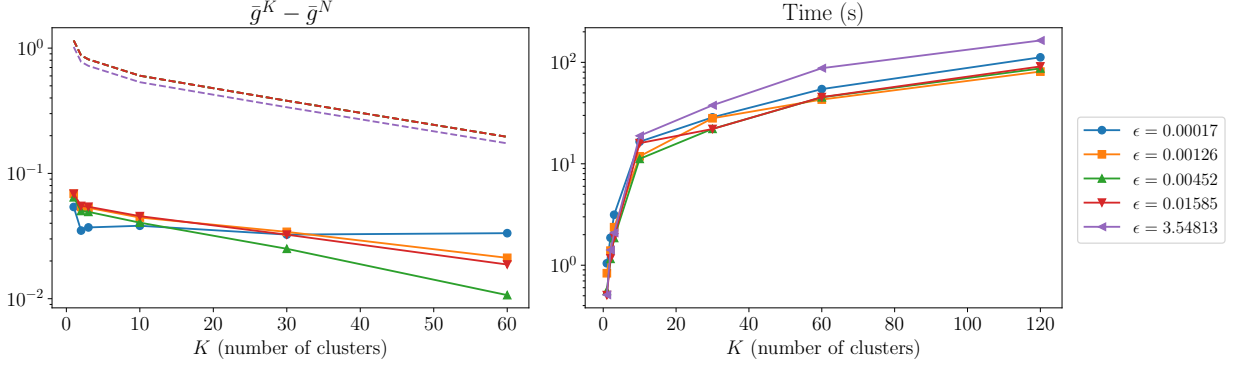


Figure 6: Capital budgeting. Left: the difference in the value of the uncertain objective between using  $K$  and  $N$  clusters, calculated as  $\bar{g}^K(x) - \bar{g}^N(x)$ , compared with the theoretical upper bound  $(L/2)D(K)$  from Corollary 3.1.1. Solid lines are the difference, dotted lines are the upper bounds. Right: solve time.

The  $A_i^{-1}$  terms come from taking the conjugate of  $g$ , and the derivation can be found in (Ben-Tal et al., 2015b, 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  $k = 1, \dots, K$ . We let  $(Y_k)_i$  indicate the  $i$ th 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  $i = 1, \dots, n$  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  $i = 1, \dots, n$ . For the upper bound, we calculate  $L = \|\sum_{i=1}^n A_i(\hat{x}_N)_i\|_{2,2}$  for each data-driven solution  $\hat{x}_N$ .

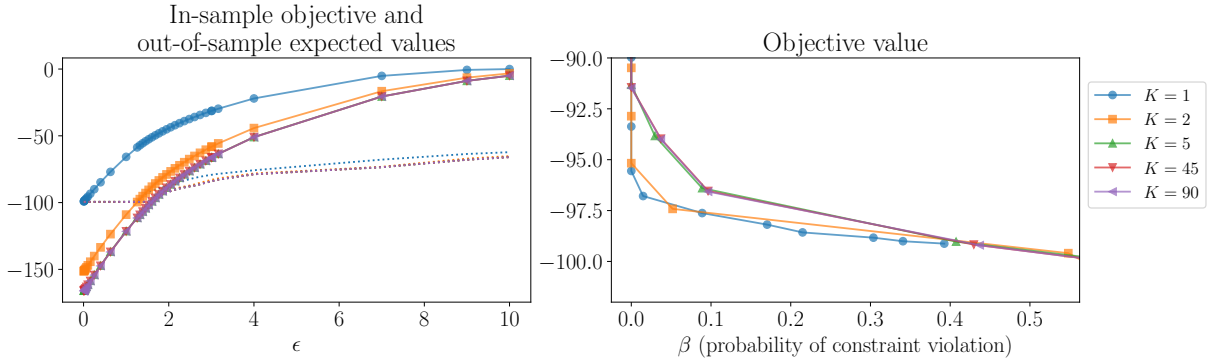


Figure 7: Quadratic concave uncertainty. Left: in-sample objective values and out-of-sample expected values of  $g$  vs.  $\epsilon$  for different  $K$ . Solid lines are the in-sample objective value, dotted lines are the out-of-sample expected value. Right: objective value vs.  $\beta$  for different  $K$ ; each point represents the solution for the  $\epsilon$  achieving the smallest objective value.

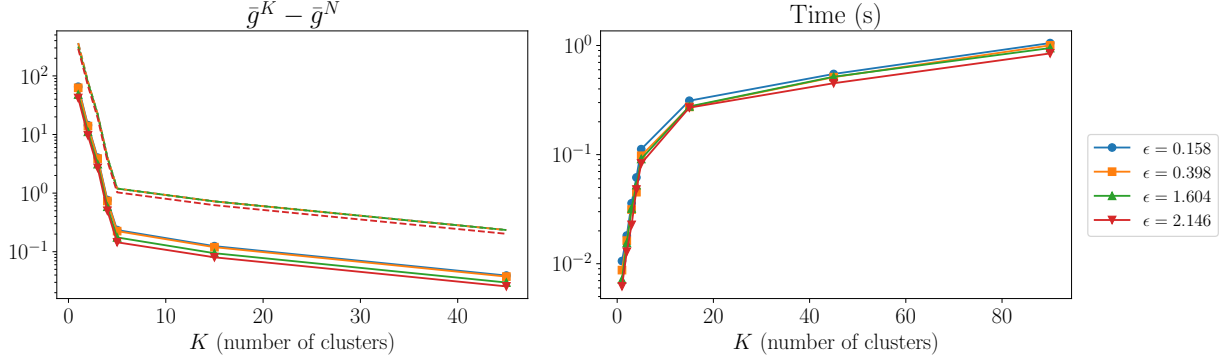


Figure 8: Quadratic concave uncertainty. Left: the difference in the value of the uncertain objective between using  $K$  and  $N$  clusters, calculated as  $\bar{g}^K(x) - \bar{g}^N(x)$ , compared with the theoretical upper bound  $(L/2)D(K)$  from Corollary (3.1.1). Solid lines are the difference, dotted lines are the upper bounds. Right: solve time.

**Results.** We observe on the left of Figure 7 that using 5 clusters is enough to achieve performance almost identical to that of using 60 clusters. Indeed, in Figure 8, 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. Furthermore, on the left plot of Figure 8, we note for  $K \geq 5$ , the tradeoff between the objective value and constraint violation is the same, so we can tune  $\epsilon$  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  $\epsilon$  to achieve an even better tradeoff. However, this result cannot be guaranteed in general, so the recommended action is still to choose  $K = 5$ .

## 5.4 Robust log-sum-exp optimization

We also consider uncertainty from (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 3.1 and Theorem 3.1 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  $p = 2$ ,

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

**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  $i = 1, \dots, n$ . 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  $\hat{x}_N$ .

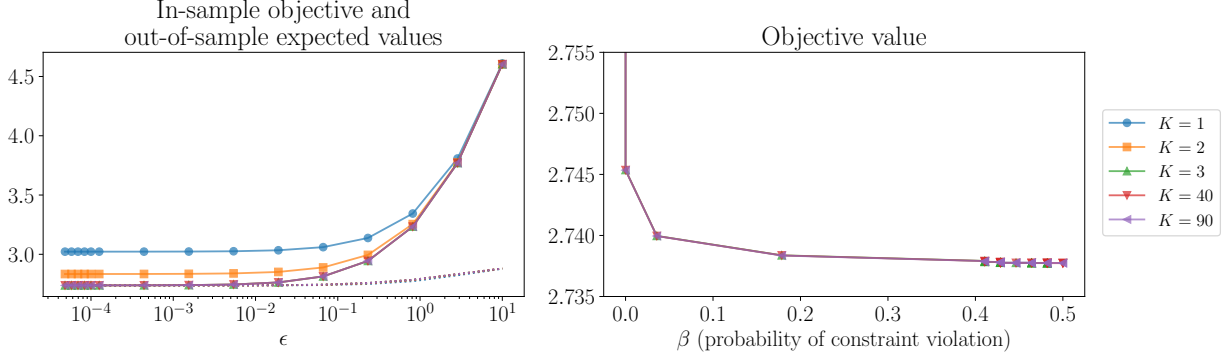


Figure 9: Log-sum-exp uncertainty. Left: in-sample objective values and out-of-sample expected values vs.  $\epsilon$  for different  $K$ . Solid lines are the in-sample objective value, dotted lines are the out-of-sample expected value. Right: objective value vs.  $\beta$  for different  $K$ .

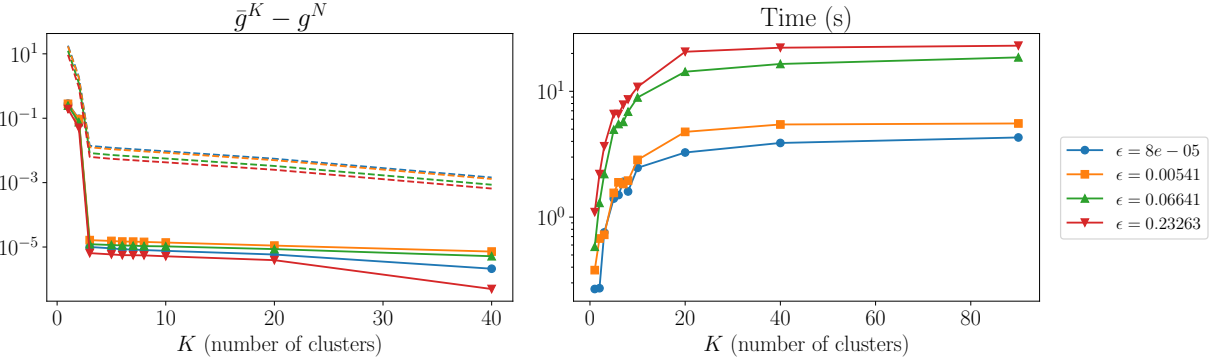


Figure 10: Log-sum-exp uncertainty. Left: the difference in the value of the uncertain objective between using  $K$  and  $N$  clusters. Solid lines are the difference, the dotted line is the upper bound. Right: solve time.

**Results.** We observe on the left of Figure 9 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 10, we see that  $K = 3$  is at the elbow of upper bound and actual difference, and is thus recommended. Furthermore, we note that setting  $K = 3$  and above give identical tradeoff curves, therefore, choosing  $K = 3$  is the time-efficient solution.



## 6 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 in the uncertainty, we directly quantified the increase in worst-case value that is 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.

## Acknowledgements

The simulations presented in this article were performed on computational resources managed and supported by Princeton Research Computing, a consortium of groups including the Princeton Institute for Computational Science and Engineering (PICSciE) and the Office of Information Technology’s High Performance Computing Center and Visualization Laboratory at Princeton University.

We would like to thank Daniel Kuhn for the useful feedback and for pointing us to the literature on scenario reduction techniques.

## 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, 2009.
- A. Ben-Tal, D. den Hertog, and J. Vial. Deriving robust counterparts of nonlinear uncertain inequalities. *Mathematical Programming*, 149(1):265–299, 2015a.
- 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. 2015b.

- D. Bertsimas and D. den Hertog. *Robust and Adaptive Optimization*. Dynamic Ideas, 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.
- R. Chen and I. Paschalidis. Distributionally robust learning. *Foundations and Trends in Optimization*, 4(1-2):1–243, 2020.
- 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.
- 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.
- 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*, abs/2009.04382, 2020.
- R. Gao and A. Kleywegt. Distributionally robust stochastic optimization with Wasserstein distance, 2016.

- 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. Kuhn, P. M. Esfahani, V. Nguyen, and S. Shafieezadeh-Abadeh. Wasserstein distributionally robust optimization: Theory and applications in machine learning, 2019.
- MOSEK ApS. *The MOSEK optimization toolbox. Version 9.3.*, 2022.
- 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.
- N. Trillos and D. Slepčev. On the rate of convergence of empirical measures in  $\infty$ -transportation distance. *Canadian Journal of Mathematics*, 67, 07 2014.
- 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.
- J. Zhen, D. Kuhn, and W. Wiesemann. Mathematical foundations of robust and distributionally robust optimization, 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):

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

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 in  $v$ , and the convexity of the supports. For the last equality, we used the definition of conjugate functions. Now, borrowing results from (Esfahani and Kuhn, 2018, Theorem 4.2), (Rockafellar and Wets, 1998, Theorem 11.23(a), p. 493), and (Zhen et al., 2021, Lemma B.8), with regards to the conjugate functions of infimal convolutions and  $p$ -norm balls, we note that:

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

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:

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

## A.2 Proof of the constraint reformulation in (12)

Starting from the inner optimization problem of (11):

$$\begin{aligned}
& \begin{cases} \sup_{v_1 \dots v_K \in S} & \sum_{k=1}^K w_k g(v_k, x) \\ \text{subject to} & \|v_k - \bar{d}_k\| \leq \epsilon, \quad k = 1, \dots, K \end{cases} \\
&= \begin{cases} \sup_{v_1 \dots v_K \in S} & \inf_{\lambda_k \geq 0} \sum_{k=1}^K w_k (g(v_k, x) + (1/w_k) \lambda_k (\epsilon - \|v_k - \bar{d}_k\|)) \end{cases} \\
&= \begin{cases} \inf_{\lambda_k \geq 0} & \sup_{v_1 \dots v_K \in S} \sum_{k=1}^K w_k (g(v_k, x) + (1/w_k) \lambda_k (\epsilon - \|v_k - \bar{d}_k\|)) \end{cases} \\
&= \begin{cases} \inf_{\lambda_k \geq 0} & \sum_{k=1}^K w_k s_k \\ \text{subject to} & (\lambda_k/w_k) \epsilon + \sup_{v_k \in S} g(v_k, x) - (\lambda_k/w_k) \|v_k - \bar{d}_k\| \leq s_k \quad k = 1, \dots, K \end{cases} \\
&= \begin{cases} \inf_{\lambda_k \geq 0} & \sum_{k=1}^K w_k s_k \\ \text{subject to} & (\lambda_k/w_k) \epsilon + \sup_{v_k \in S} g(v_k, x) - \max_{\|z_k\|_* \leq \lambda_k/w_k} z_k^T (v_k - \bar{d}_k) \leq s_k \quad k = 1, \dots, K \end{cases} \\
&= \begin{cases} \inf_{\lambda_k \geq 0} & \sum_{k=1}^K w_k s_k \\ \text{subject to} & (\lambda_k/w_k) \epsilon + \min_{\|z_k\|_* \leq \lambda_k/w_k} \sup_{v_k \in S} g(v_k, x) - z_k^T (v_k - \bar{d}_k) \leq s_k \quad k = 1, \dots, K \end{cases} \\
&= \begin{cases} \inf_{\lambda_k \geq 0, z_k} & \sum_{k=1}^K w_k s_k \\ \text{subject to} & (\lambda_k/w_k) \epsilon + [-g + \mathcal{I}_S]^*(-z_k) + z_k^T \bar{d}_k \leq s_k \quad k = 1, \dots, K \\ & \|z_k\|_* \leq \lambda_k/w_k, \quad k = 1, \dots, K \end{cases} \\
&= \begin{cases} \inf_{\lambda_k \geq 0, z_k} & \sum_{k=1}^K w_k s_k \\ \text{subject to} & [-g]^*(z_k - y_k) + \sigma_S(y_k) - z_k^T \bar{d}_k + \epsilon \| -z_k \|_* \leq s_k \quad k = 1, \dots, K. \end{cases}
\end{aligned}$$

We have again used the Lagrangian, 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 substitutions  $z_k = -z_k$  and  $\lambda_k/w_k = \|z_k\|_*$ .

## A.3 Proof of the primal problem reformulation as $p \rightarrow \infty$

Consider again the function  $\bar{g}^K$  discussed in Section 3 and defined as

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

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

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

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

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

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

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

and

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

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  $k \in 1, \dots, K$ . As we hence have for all  $k = 1, \dots, K$  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  $\bar{g}_l(x; p) \leq \bar{g}(x; p) \leq \bar{g}_u(x; p)$  for any  $p \geq 1$ .

Furthermore, observe that when  $t_k \geq 0$  for all  $k = 1, \dots, K$  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

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

Similarly, observe that when  $t_k \geq 0$  for all  $k = 1, \dots, K$  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

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

Finally, chaining all the inequalities together we obtain

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

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

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

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

#### A.4 Proof of the dual problem reformulation as $p \rightarrow \infty$

**Theorem A.2.** *Let the function  $[-g^*]$  be uniformly continuous. Define here*

$$\bar{g}^K(x; \infty) = \begin{cases} \text{minimize} & \sum_{k=1}^K w_k s_k \\ \text{subject to} & \lambda \geq 0, z_k \in \mathbf{R}^m, y_k \in \mathbf{R}^m, s_k \in \mathbf{R}^m \quad k = 1, \dots, K, \\ & [-g]^*(z_k - y_k) + \sigma_S(y_k) - z_k^T d_k + \epsilon \|z_k\|_* \leq s_k \quad k = 1, \dots, K \end{cases} \quad (21)$$

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

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

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

where the final inequality follows from Lemma A.1. Hence, considering the limit for  $p$  tending to infinity gives us now  $\lim_{p \rightarrow \infty} \bar{g}^K(x; p) \geq \bar{g}^K(x; \infty)$ . It remains to prove the reverse  $\lim_{p \rightarrow \infty} \bar{g}^K(x; p) \leq \bar{g}^K(x; \infty)$ .



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

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

$$\bar{g}_u^K(x; p) = \left\{ \begin{array}{l} \text{minimize} \quad \sum_{k=1}^K w_k s_k \\ \text{subject to} \quad z_k \in \mathbf{R}^m, y_k \in \mathbf{R}^m, s_k \in \mathbf{R}^m \quad k = 1, \dots, K, \\ \quad [-g]^*(z_k - y_k) + \sigma_S(y_k) - z_k^T d_k + \epsilon \|z_k\|_* D\left(\frac{p}{p-1}\right) \leq s_k \quad k = 1, \dots, K \\ \quad (p-1)^{-1/4} \leq \|z_k\|_* \leq (p-1)^{1/4} \quad k = 1, \dots, K. \end{array} \right. \quad (22)$$

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  $p \rightarrow \infty$  we have  $\lim_{p \rightarrow \infty} \bar{g}^K(x; p) \leq \lim_{p \rightarrow \infty} \bar{g}_u^K(x; p)$ . We now prove here that  $\lim_{p \rightarrow \infty} \bar{g}_u^K(x; p) \leq \bar{g}^K(x; \infty)$ . 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 (21) so that  $\lim_{t \rightarrow \infty} \sum_{k=1}^K w_k s_k^t = \bar{g}^K(x; \infty)$ . 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  $k = 1, \dots, K$  and observe that by construction  $\|\tilde{z}_k^t\|_* \geq 1/t$ . 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  $\{(z_k^t, y_k^t, \tilde{s}_k^t = [-g]^*(z_k^t - y_k^t) + \sigma_S(y_k^t) - (\tilde{z}_k^t)^T d_k + \epsilon \|\tilde{z}_k^t\|_* D(p_t/(p_t - 1)))\}_{t \geq 1}$  is by construction feasible in the minimization problem characterizing the function  $\bar{g}_u^K(x; p_t)$  in Equation (22). Hence, finally, we have

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

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,  $\|\tilde{z}_k^t\|_* = \|z_k^t + (\tilde{z}_k^t - z_k^t)\|_* \leq \|z_k^t\|_* + \|\tilde{z}_k^t - z_k^t\|_* \leq \|z_k^t\|_* + 1/t$ . Third, we let  $\delta(a) = \sup_{v \in \mathbf{R}^m, \|\epsilon\|_* \leq a} [-g]^*(v + \epsilon) - g(v)$ . Hence,  $[-g]^*(\tilde{z}_k^t - y_k^t) \leq [-g]^*(z_k^t - y_k^t) + \delta(1/t)$  for all  $t \geq 1$  and  $k = 1, \dots, K$  as  $\|\tilde{z}_k^t - z_k^t\| \leq 1/t$ . The absolute continuity of  $[-g]^*$  guarantees that  $\lim_{t \rightarrow \infty} \delta(1/t) = \lim_{a \rightarrow 0} \delta(a) = 0$  while Lemma A.2 guarantees that  $\lim_{t \rightarrow \infty} D(p_t/(p_t - 1)) = 1$ . Finally,  $\|z_k^t\|_* \leq \|\tilde{z}_k^t\|_* \leq (p_t - 1)^{1/4}$  and

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

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

**Lemma A.1.** *We have*

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

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

*Proof.* Remark that as the objective function  $\lambda \mapsto \phi(q)\lambda \|z/\lambda\|_*^q + \lambda\epsilon^p$  is continuous and we have  $\lim_{\lambda \rightarrow 0} \phi(q)\lambda \|z/\lambda\|_*^q + \lambda\epsilon^p = \lim_{\lambda \rightarrow \infty} \phi(q)\lambda \|z/\lambda\|_*^q + \lambda\epsilon^p = \infty$  as  $\epsilon > 0$  there must exist a minimizer  $\lambda^* \in \min_{\lambda \geq 0} \phi(q)\lambda \|z/\lambda\|_*^q + \lambda\epsilon^p$  with  $\lambda_* > 0$ . The necessary and sufficient first-order convex optimality conditions of the minimization problem guarantee

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

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

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

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

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

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  $\lim_{q \rightarrow 1} D(q) = 1$  and  $\lim_{q \rightarrow 1} (q-1)^{1/4} (D(q) - 1) = 0$ .

*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.5 Proof of the equivalence between $p = 1$ and $p = \infty$ for affine uncertainty

We again consider the single affine constraint (8). In formulation (10), when  $p = 1$ , we observe that (Kuhn et al., 2019, Section 2.2 Remark 1)

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

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

$$\begin{aligned} & \text{minimize} && f(x) \\ & \text{subject to} && a^T x - b + \lambda \epsilon + (P^T x)^T \sum_{k=1}^K w_k \bar{d}_k \leq 0 \\ & && \|P^T x\|_* \leq \lambda \\ & && \lambda \geq 0, \end{aligned} \tag{23}$$

where we can make the substitution  $\lambda = \|P^T x\|_*$ , in which case this becomes equivalent to (13).

## A.6 Proof of convex reduction of the worst-case problem (17)

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 ambiguity sets defined

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

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

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

$$\begin{aligned} &= \left\{ \sup_{\mathbf{Q}_{\mathbf{k}} \in \mathcal{M}(S)} \inf_{\lambda \geq 0} \sum_{k=1}^K w_k \int g(u, x) \mathbf{Q}_{\mathbf{k}}(du) + \lambda(\epsilon^p - \sum_{k=1}^K w_k \int \|u - \bar{d}_k\|^p \mathbf{Q}_{\mathbf{k}}(du)) \right\} \\ &= \left\{ \inf_{\lambda \geq 0} \sup_{\mathbf{Q}_{\mathbf{k}} \in \mathcal{M}(S)} \lambda \epsilon^p + \sum_{k=1}^K w_k \int g(u, x) - \lambda \|u - \bar{d}_k\|^p \mathbf{Q}_{\mathbf{k}}(du) \right\} \\ &= \left\{ \inf_{\lambda \geq 0} \lambda \epsilon^p + \sup_{u=(v_1, \dots, v_K) \in \mathcal{U}} \sum_{k=1}^K w_k (g(v_k, x) - \lambda \|v_k - \bar{d}_k\|^p), \right\} \end{aligned}$$

where the second equality is due to a well-known strong duality result for moment problems (Esfahani and Kuhn, 2018). The last expression is identical to the form in Appendix A.1, so the final dual is equivalent to the dualized constraint in (7).

## A.7 Proof of Theorem 3.1

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

**Proof of (i).** We begin with a feasible solution  $v_1, \dots, v_N$  of (MRO-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  $k = 1, \dots, K$  satisfies the constraints of (MRO-K), as

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

where we have utilized 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, our constructed  $u_k$ 's, as the average of select  $v_i$ 's  $\in S$ , must also be within  $S$ . The same applies with 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,  $\bar{g}^K(x) \geq \bar{g}^N(x)$ . By assumption,  $g$  is concave in the uncertain parameter, so by Jensen's inequality,

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

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

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

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

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  $L$ -smooth 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). Therefore,  $u_i = d_i + \Delta_k$  is also in the domain, as the function is also concave. For monotonically decreasing functions  $g$ , the same logic applies with a nonpositive  $\Delta_k$ . 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, when applied iteratively, results in:

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

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

**Proof of (iii).** When the support constraint does not affect the uncertainty set, we note that the constructed  $v_i$ 's in part (ii) are then also in  $S$ . Therefore,  $\bar{g}^N(x) = \bar{g}^{N*}(x)$ , and we arrive at the desired conclusion.

## A.8 Conjugate derivation for the Capital Budgeting problem (20)

We begin with

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

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,

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

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

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

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.