Scalable Distributional Robustness in a Class of Non Convex Optimization with Guarantees

Avinandan Bose
University of Washington
avibose@cs.washington.edu

Arunesh Sinha
Rutgers University
arunesh.sinha@rutgers.edu

Tien Mai
Singapore Management University
atmai@smu.edu.sg

Abstract

Distributionally robust optimization (DRO) has shown lot of promise in providing robustness in learning as well as sample based optimization problems. We endeavor to provide DRO solutions for a class of sum of fractionals, non-convex optimization which is used for decision making in prominent areas such as facility location and security games. In contrast to previous work, we find it more tractable to optimize the equivalent variance regularized form of DRO rather than the minimax form. We transform the variance regularized form to a mixed-integer second order cone program (MISOCP), which, while *guaranteeing near global optimality*, does not scale enough to solve problems with real world data-sets. We further propose two abstraction approaches based on clustering and stratified sampling to increase scalability, which we then use for real world data-sets. Importantly, we provide near global optimality guarantees for our approach and show experimentally that our solution quality is better than the locally optimal ones achieved by state-of-the-art gradient-based methods. We experimentally compare our different approaches and baselines, and reveal nuanced properties of a DRO solution.

1 Introduction

Distributionally robust optimization (DRO) is a popular approach employed in robust machine learning. Mostly, if not always, these works have focussed on the task of classification or regression. However, often in practical applications the end goal of learning is a decision output $\mathbf{z}$, which requires yet another complex optimization that uses the output $\widehat{\mathbf{x}}_1, \ldots, \widehat{\mathbf{x}}_N$ of a regressor $f(\cdot)$. For example, in facility location problem the learning of facility values is followed by an optimization using the values predicted to decide where to locate facilities and in security games adversary behavior model is learned and then an optimal defense allocation computed based on the learned model. Often the learning output is provided as public datasets with no access to the underlying private dataset used for such learning. In this set-up, we aim to provide robustness at the decision making level with access to only the non-robust learning output $\widehat{\mathbf{x}}_1, \ldots, \widehat{\mathbf{x}}_N$.

However, often the objective $F(\mathbf{z}, \mathbf{x})$ of decision optimization is *non-convex* in the learning system output $\mathbf{x}$ unlike the convex objective of classification or regression, presenting significant scalability challenges. In general, for decision making and specifically for the problem domains we consider, *global optimality is important* as sub-optimal decisions can lead to large revenue loss or loss of life; thus, the local optimality provided by gradient based methods is not sufficient. As a consequence, in this paper, we study the scenario of calculating DRO decisions using the given multi-dimensional real valued outputs $\widehat{\mathbf{x}}_1, \ldots, \widehat{\mathbf{x}}_N$ of a non-robust learned f. A first result (Theorem 1) characterizes the

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

quality of DRO decision output compared to the scenario where we know the true f^*. Our main focus is on addressing the scalability issue for the DRO decision making problem for a particular, but widely used, class of *sum of fractionals non-convex objective*. This objective arises from the well-known *discrete choice models* [Train, 2003] of human behavior, which is known to *not* have scalable globally optimal solutions [Schaible, 1995, Li et al., 2019]; we use this for tackling two different decision optimization problem in facility location and a robust version of Bayesian Stackelberg security games problem with quantal response. As far as we know, this is a first attempt to solve the aforementioned non-convex problems in a DRO setting to near global optimality.

Our *first contribution* is a modelling construct, where we reformualate the variance regularized form [Duchi and Namkoong, 2019] of our non-convex sum of fractionals objective as a mixed integer second order cone program (MISOCP). While the MISOCP form provides more scalability than the original formulation and guaranteed solution quality (Theorem 2), it still does not scale to real world sized datasets. Our *second contribution* is a pair of approaches that achieves further scalability by splitting the problem space into sub-regions and solving a smaller MISOCP over representative samples from the sub-regions. Under mild conditions, both approaches provide *global optimality guarantees* (Theorem 3, 4).

Our *final contribution* is detailed experiments validating the scalability of our approaches on a simulated security game problem as well as two variants of facility location using park and ride data-sets from New York [Holguin-Veras et al., 2012]. We compare with two gradient-based approaches [Lin et al., 2020] and show the superior solution quality achieved by our approach, which also reveals the need for global optimality. We further show a nuanced property of the DRO solution in providing better decisions for low probability scenarios over non-robust versions. Overall, our work provides desired *robustness with globally optimal solution guarantees*.

Related work: Our work is built on a recent line of research that connects the concepts of DRO and variance regularization [Duchi and Namkoong, 2019, Duchi et al., 2021, Lam, 2016, Maurer and Pontil, 2009, Staib et al., 2019]. While most the previous studies along this research line focus on convex and continuous problems or problems with submodular objectives, our work concerns a class of DRO problems with fractional structures, which are highly non-convex and requires new technical developments for globally optimal solution. Recent work Yan et al. [2020], Qi et al. [2021] has addressed non-convex objectives in DRO using gradient based methods that converge to stationary points, which is insufficient for decision making as we experimentally show that stationary points and globally optimal points can yield very different decision utilities.

The literature on DRO is vast and we refer the reader to Rahimian and Mehrotra [2019] for a review. DRO methods can be classified by different ways to define ambiguity sets of distributions, for instance, ambiguity sets based on ϕ-divergences [Ben-Tal et al., 2013, Duchi and Namkoong, 2019, Staib et al., 2019] or Wasserstein distances [Pflug and Wozabal, 2007, Esfahani and Kuhn, 2018, Shafieezadeh-Abadeh et al., 2015, Blanchet and Murthy, 2019]. In this work, we focus on ϕ-divergence based models, motivated by their interesting connections with variance regularization and the tractability of the resulting non-convex DRO models.

We show that our DRO methods can be used in some popular decision-making problems such as Stackelberg security game (SSG) with Quantal Response [Tambe, 2011, Xu, 2016, Fang et al., 2017, Sinha et al., 2018, Yang et al., 2012, Haghtalab et al., 2016] or competitive facility location under random utilities [Benati and Hansen, 2002, Freire et al., 2016, Mai and Lodi, 2020, Dam et al., 2021]. To the best of our knowledge, a DRO Bayesian model has not been studied in existing SSG works. In the context of competitive facility location under random utilities, we seem to be the first to bring DRO as a consideration. We handle a DRO version of a facility cost optimization problem, which has also never been studied in prior work.

2 Background, Preliminary Notation and Result

We use bold fonts for vectors and non-bold font for vector components and scalars, e.g., x_j is a component of $\mathbf{x}$. $[N]$ denotes $\{1, \ldots, N\}$. A d-dimensional vector is written as $\mathbf{x} = (x_j)_{j \in [d]}$ or as $(x_1, \ldots, x_d)$. The positive part of a vector $\mathbf{x}$ is $\mathbf{x}^+ = (\max(0, x_j))_{j \in [d]}$, and the negative part is $\mathbf{x}^- = (\min(0, x_j))_{j \in [d]}$. $\mathbf{0}, \mathbf{1}$ represent the all zero and all one vector.

Distributionally Robust Optimization: Consider a function F with inputs being a decision variable $\mathbf{z}$ and parameter $\mathbf{x} \in X$. Both $\mathbf{z}$ and $\mathbf{x}$ lie in an Euclidean space and both *are constrained by linear constraints*; for notational ease we skip writing the constraints in the general formulation. We seek to maximize the following objective function $\max_{\mathbf{z}} \mathbb{E}_P[F(\mathbf{z}, \mathbf{x})]$, where $\mathbf{x}$ is distributed according to P. The details of how P arises from an underlying regression problem is stated later in the text just before Theorem 1. For many classes of distributions the above is generally not tractable and one needs to sample $\mathbf{x}$ from P. Let $\widehat{\mathbf{x}}_1, \ldots, \widehat{\mathbf{x}}_N$ be N samples, we can solve the sample average approximation (SAA) problem instead $\max_{\mathbf{z}} \frac{1}{N} \sum_{n \in [N]} F(\mathbf{z}, \widehat{\mathbf{x}}_n)$. Let $\widehat{P}_N$ be the empirical distribution induced by the samples. The SAA above is same as $\max_{\mathbf{z}} \mathbb{E}_{\widehat{P}_N}[F(\mathbf{z}, \mathbf{x})]$. A distributionally robust version of the SAA problem is $\max_{\mathbf{z}} \min_{\widetilde{P} \in \mathcal{P}_{\xi,N}} \left\{\mathbb{E}_{\widetilde{P}}[F(\mathbf{z}, \mathbf{x})]\right\}$, where the ambiguity set $\mathcal{P}_{\xi,N} = \left\{\widetilde{P} \middle| \, \mathcal{D}_\phi(\widetilde{P} || \widehat{P}_N) \leq \xi/N\right\}$, and $\mathcal{D}_\phi(P||Q)$ is the χ^2 divergence: $\mathcal{D}_\phi(P||Q) = \frac{1}{2} \int (dP/dQ - 1)^2 dQ$. The above optimization can be written equivalently as ($\Delta_{\xi,N}$ defined below)

$$\max_{\mathbf{z}} \min_{\mathbf{p} \in \Delta_{\xi,n}} \left\{ \sum_{i \in [N]} p_i F(\mathbf{z}, \widehat{\mathbf{x}}_i) \right\} \tag{DRO}$$

where $\Delta_{\xi,N} = \left\{\mathbf{p} \in \mathbb{R}_+^N \middle| \, \sum_i p_i = 1; \; ||\mathbf{p} - \mathbf{1}/N||_2^2 \leq 2\frac{\xi}{N^2}\right\}$. We have earlier stated that $\widehat{\mathbf{x}}_i$ is output by a regressor, say $f \in \mathcal{F}$ for some function class $\mathcal{F}$ trained using loss $\mathcal{L}$ with N_T datapoints, but this implies that $\widehat{\mathbf{x}}_i = f(b_i)$ might not exactly same as $\mathbf{x}_i^* = f^*(b_i)$ for some underlying feature values b_i and best function $f^* \in \mathcal{F}$. We assume f^* is deterministic and has zero Bayes risk.

Let D be the probability distribution from which the feature values b_i are sampled. Then, let P^* be the true distribution on X induced by f^* acting on the feature values that are distributed according to D (i.e., the pushforward measure). Similarly, P is the distribution on X induced by f acting on the feature values that are distributed according to D. Thus, the (unknown) samples $\mathbf{x}^*$'s are obtained from P^*; hence, the true utility of any decision $\mathbf{z}$ is $\mathbb{E}_{P^*}[F(\mathbf{z}, \mathbf{x})]$. We prove an end to end guarantee about the output decision $\widehat{\mathbf{z}}^{**}$ using $\widehat{\mathbf{x}}_i$'s, which reveals that $\widehat{\mathbf{z}}^{**}$ is not much worse than the decision $\mathbf{z}^{**}$ that would be learned if $\mathbf{x}_i^*$'s would be available and used. The result shows that larger training data N_T helps.

Theorem 1. *As described above, let $\boldsymbol{x}_i^* = f^*(b_i)$ for true function f^* and let $\widehat{\boldsymbol{x}}_i = f(b_i)$ for the learned empirical risk minimizer f. Suppose the optimal decision when solving* DRO *is $\boldsymbol{z}^{**}$ using $\boldsymbol{x}_i^*$'s and $\widehat{\boldsymbol{z}}^{**}$ using $\widehat{\boldsymbol{x}}_i$'s. Also, let F be τ-Lipschitz in $\boldsymbol{x}$, X be bounded, and a scaled $\mathcal{L}$ upper bound $|| \cdot ||_2$ (i.e., $||\boldsymbol{x} - \boldsymbol{x}'||_2 \leq \max(k\mathcal{L}(\boldsymbol{x}, \boldsymbol{x}'), \epsilon)$ for constants k, ϵ) then, the following holds with probability $1 - 2\delta - 2\delta_1$: $\mathbb{E}_{P^*}[F(\widehat{\boldsymbol{z}}^{**}, \boldsymbol{x})] \geq \mathbb{E}_{P^*}[F(\boldsymbol{z}^{**}, \boldsymbol{x})] - C/\sqrt{N} - (1 + 2\sqrt{\xi})\tau\epsilon - \epsilon_N - \epsilon_{N_T}$, where $\epsilon_K = C_1 \mathcal{R}_K(\mathcal{L} \circ \mathcal{F}) + C_2/\sqrt{K}$ and $\mathcal{R}_K$ is the Rademacher complexity with K samples and C, C_1, C_2 are constants dependent on $\delta, \delta_1, \xi, k, \tau$.*

Variance Regularizer: As a large number of samples are needed for a low variance approximation of the true distribution, another proposed robust version of the SAA [Maurer and Pontil, 2009, Duchi and Namkoong, 2019] is to optimize the following variance-regularized (VR) objective function

$$\max_{\mathbf{z}} \left\{ \mathbb{E}_{\widehat{P}_N}[F(\mathbf{z}, \mathbf{x})] - C\sqrt{\frac{\mathrm{Var}_{\widehat{P}_N}(F(\mathbf{z}, \mathbf{x}))}{N}} \right\}. \tag{VR}$$

The above allows to directly optimize the trade-off between bias and variance. In a fundamental result, Duchi and Namkoong [2019, Theorem 1] show that, with high probability, problem (VR) is *equivalent* to the problem (DRO). Further, Duchi and Namkoong [2019] argue for solving the DRO version of the problem for concave F (note we are solving a maximization SAA problem) since concave F results in concavity of $\min_{\mathbf{p} \in \Delta_{\xi,n}} \sum_{i \in [N]} p_i F(\mathbf{z}, \widehat{\mathbf{x}}_i)$, thus, the overall DRO problem is a concave maximization problem. In contrast, the objective in (VR) is not concave.

In this paper, our focus is on F that is *not concave*, thus, the choice of DRO or variance regularized form is not obvious. For the class of functions F that we analyze, we argue the variance regularized version is more promising as far as scalability for global optimality is concerned. We work with the assumption that the variance regularized form is equivalent to DRO, which holds under the mild condition shown in Equation (9) in Duchi and Namkoong [2019].

3 Towards a Globally Optimal Solution

In this section, we present results for a general class of non-concave functions F that has a fractional form with non-linear numerator and denominator and that can be approximated by a linear fractional form with binary variables. Then, we show three *prominent* applications of our approach.
Notation: For ease of notation, we use shorthand to denote $F(\mathbf{z}, \widehat{\mathbf{x}}_i)$ by F_i and $2\frac{\xi}{N^2}$ by ρ.

3.1 General Recipe to Form a MISOCP

We perform a sequence of variable and constraint transformations of (VR), leading to a MISOCP.

Mixed Integer Concave Program: The variance regularized objective in shorthand notation is:

$$\mathcal{G}(\mathbf{z}) = \sum_i \frac{F_i}{N} - \sqrt{\rho \sum_i \Big(\frac{\sum_{i'} F_{i'}}{N} - F_i \Big)^2} \tag{1}$$

We substitute $l_i = \frac{\sum_{i'} F_{i'}}{N} - F_i$ and $q = \frac{\sum_{i'} F_{i'}}{N}$ for all $i \in [N]$, such that $\sum_i l_i = 0$ and $F_i = q - l_i$. The objective in Equation 1 thus becomes $q - \sqrt{\rho \sum_i l_i^2}$ which is concave in the variables q and $\{l_i\}$. We add the new constraints $\sum_i l_i = 0$ and $F_i = q - l_i$ for all $i \in [N]$. Note that, while the objective is now concave with above changes, we have pushed the non-convexity into the constraints $F_i - q + l_i = 0$ for all $i \in [N]$ that are added to the optimization.

If F_i can be written (or approximated) as a fraction with affine numerator and denominator, we can convert the constraint $F_i - q + l_i = 0$ into a convex constraint, giving us an overall concave program. The conversion is explained next. Suppose F_i can be written (or approximated) as $\frac{\mathbf{a}_i^T \mathbf{v} + b_i}{\mathbf{a}_i'^T \mathbf{v} + b_i'}$ where $\mathbf{v}$ represents *binary* variables after conversion ($\mathbf{v}$ completely replaces $\mathbf{z}$ and $\mathbf{a}_i, \mathbf{a}_i', b_i, b_i'$ are dependent on $\widehat{\mathbf{x}}_i$'s). Typically, such a linear fractional form is constructed by discretizing the arguments of the original non-linear functions in the numerator and denominator of F. Assume $\mathbf{v}$ is of dimension d; typically d will depend on the number of pieces. Define $\mathbf{y}_i = \mathbf{v} t_i$ where $t_i = \frac{1}{\mathbf{a}_i'^T \mathbf{v} + b_i'}$. Then, we can (re)write the fractional form for F_i as $F_i = \mathbf{a}_i^T \mathbf{y}_i + b_i t_i$. This yields the linear constraints below with the non-linearity now restricted to $\mathbf{y}_i = \mathbf{v} t_i$.

$$\sum_{i=1}^{N} l_i = 0 \tag{2}$$

$$\mathbf{a}_i^T \mathbf{y}_i + b_i t_i = q - l_i \qquad \forall i \in [N] \tag{3}$$

$$\mathbf{a}_i'^T \mathbf{y}_i + b_i' t_i - 1 = 0 \qquad \forall i \in [N] \tag{4}$$

We handle $\mathbf{y}_i = \mathbf{v} t_i$ using McCormick relaxation technique [McCormick, 1976]. Typically, McCormick relaxation is applied for bilinear terms that are the product of two continuous variables, in which case, it is an approximation. However, in our case since $\mathbf{v}$ is a binary vector variable, the McCormick relaxation yields an exact reformulation of the bilinear term. For applying McCormick technique, we need an upper and lower bound of $\mathbf{v}$ and t_i. Since $\mathbf{v}$ a vector of binary variables, we have lower bound $\mathbf{v}^L = \mathbf{0}$ and upper bound $\mathbf{v}^U = \mathbf{1}$. Similarly, $t_i^L = \frac{1}{(\mathbf{a}_i'^{+})^T \mathbf{1} + b_i'}$ and $t_i^U = \frac{1}{(\mathbf{a}_i'^{-})^T \mathbf{1} + b_i'}$ (recall superscript + and − indicate positive and negative part of a vector respectively). Further, it is assumed t_i^U and t_i^L exist. Note that these bounds are not variables but fixed constants that depend on the fixed parameters $\mathbf{a}_i, \mathbf{a}_i', b_i, b_i'$, hence these need to be computed just once. Using the upper and lower bounds of $\mathbf{v}$ and t_i in McCormick technique we get:

$$\mathbf{y}_i - \mathbf{v} t_i^U \leq 0; \qquad \forall i \in [N] \tag{5}$$

$$\mathbf{y}_i - (\mathbf{1} t_i + \mathbf{v} t_i^L - \mathbf{1} t_i^L) \leq 0; \qquad \forall i \in [N] \tag{6}$$

$$-\mathbf{y}_i + (\mathbf{1} t_i + \mathbf{v} t_i^U - \mathbf{1} t_i^U) \leq 0; \qquad \forall i \in [N] \tag{7}$$

$$-\mathbf{y}_i + \mathbf{v} t_i^L \leq 0; \qquad \forall i \in [N] \tag{8}$$

$$\mathbf{v} \in \{0,1\}^d \tag{9}$$

$$t_i^U \leq t_i \leq t_i^L; \qquad \forall i \in [N] \tag{10}$$

It is straightforward to check the above set of equations is equivalent to $\mathbf{y}_i = \mathbf{v} t_i$. With the changes, we obtain a mixed integer concave program (with all constraints linear). Next, while the above can be solved using branch and bound with general purpose convex solvers for intermediate problem, we show that a further transformation to a MISOCP is possible. Specialized SOCP's solvers provide much more scalability than a general purpose convex solvers [Bonami and Tramontani, 2015] and hence partially address the scalability challenge.

Mixed Integer SOCP: We transform further by introducing another variable s to stand for $\sqrt{\rho \sum_i l_i^2}$. We use $\mathbf{r} = (s, q, (l_i)_{i\in[N]}, \mathbf{v}, (t_i)_{i\in[N]}, \mathbf{y}_1, \ldots, \mathbf{y}_N)$ to denote all the variables of the optimization. Thus, the objective becomes the linear function $q - s$ with an additional constraint that

$$\sqrt{\rho \sum_i l_i^2} \leq s \tag{11}$$

The above is same as $||A\mathbf{r}||_2 \leq \mathbf{c}^T\mathbf{r}$ for the constant matrix A (with entries 0 or $\sqrt{\rho}$) and constant vector $\mathbf{c}$ (with 1 in the s component, rest 0's) that picks the l_i's and s respectively. This is a SOCP form of constraint, and the linear objective $q - s$ makes the problem after this transformation a MISOCP. In the above reformulation, the only approximation is introduced in writing F_i as a linear fractional term. One way of such approximation is via discretization. Suppose the fraction function $F(\mathbf{z}, \widehat{\mathbf{x}}_i)$ has a separable (in $\mathbf{z}$) numerator and denominator of the form $\frac{\sum_j n(z_j, \widehat{\mathbf{x}}_i)}{\sum_j d(z_j, \widehat{\mathbf{x}}_i)}$ where j ranges over the components of $\mathbf{z}$, and $n(z_j, \widehat{\mathbf{x}}_i)$ and $d(z_j, \widehat{\mathbf{x}}_i)$ are non-negative and Lipschitz continuous in z_j with Lipschitz constants C^n, C^d respectively. In this case, a general approximation via discretization is possible with the following guarantee:

Theorem 2. *For $F(\boldsymbol{z}, \widehat{\boldsymbol{x}}_i) = \frac{\sum_j n(z_j, \widehat{\boldsymbol{x}}_i)}{\sum_j d(z_j, \widehat{\boldsymbol{x}}_i)}$ as stated above and approximated as $\frac{\boldsymbol{a}_i^T \boldsymbol{v} + b_i}{\boldsymbol{a}_i'^T \boldsymbol{v} + b_i'}$, an approximation via discretization of z_j with K pieces yields $|\mathcal{G}(\boldsymbol{z}^*) - \mathcal{G}(\widehat{\boldsymbol{z}}^{**})| \leq O(\max\{C^n, C^d\}/K)$, where $\mathcal{G}(\boldsymbol{z}^*)$ and $\mathcal{G}(\widehat{\boldsymbol{z}}^{**})$ are the optimal objective values with approximation (MISOCP) and without the approximation respectively.*

Next, we show instantiation of the just presented general recipe for three widely studied problems.

3.2 Applications

Notation: In the SSG (facility location) application m resources (facility) are allocated to M targets (locations). $\mathbf{x}$ maps to type $\theta^a_{\mathbf{x}}$ of adversary, and type $\theta^d_{\mathbf{x}}$ of defender in SSG, and type $\theta_{\mathbf{x}}$ of clients of facility or directly $V_{\mathbf{x}}$ utility for each client type in facility location.

Bayesian Stackelberg Security Game with Quantal Response: A SSG models a Stackelberg game where a defender moves first to allocate m security resources for protecting M targets. The randomized allocation is specified by decision variables $\mathbf{z}$ of dimension M with the constraints that $\sum_{i=1}^M z_i \leq m$ ($z_i \in [0, 1]$); z_i is interpreted as the protection probability of the target i. Past works have used the model of a quantal responding adversary [Sinha et al., 2018]]. We generalize this to a Bayesian game version where there is a continuum of attackers types with the type specified by a parameter $\mathbf{x}$ and an *unknown* prior distribution over these types. The attacker's utility in attacking the target j is a function of the protection probability of target j and type: $h(z_j, \theta^a_{\mathbf{x}})$. Similarly, the defender's utility when target j is attacked is: $u(z_j, \theta^d_{\mathbf{x}})$ for some player-specific parameters θ that depend on $\mathbf{x}$. Following quantal response model (for attacker only), the attacker of type $\mathbf{x}$ attacks a target j with probability $\frac{\exp(h(z_j, \theta^a_{\mathbf{x}}))}{\sum_{j\in[M]} \exp(h(x_j, \theta^a_{\mathbf{x}}))}$ and the defender utility is $F(\mathbf{z}, \mathbf{x}) = \frac{\sum_{j\in[M]} u(z_j, \theta^d_{\mathbf{x}}) \exp(h(z_j, \theta^a_{\mathbf{x}}))}{\sum_{j\in[M]} \exp(h(x_j, \theta^a_{\mathbf{x}}))}$. Note that in case the defender's utilities $u(z_j, \theta^d_{\mathbf{x}})$ take negative values and the assumptions of Theorem 2 will be violated. This issue can be simply fixed by choosing α such that $\alpha \geq \max_{\mathbf{z},\mathbf{x}}\{-u(z_j, \theta^d_{\mathbf{x}})\}$ and replacing $F(\mathbf{z}, \mathbf{x})$ by $F(\mathbf{z}, \mathbf{x}) + \alpha = \frac{\sum_{j\in[M]} (u(z_j, \theta^d_{\mathbf{x}}) + \alpha) \exp(h(z_j, \theta^a_{\mathbf{x}}))}{\sum_{j\in[M]} \exp(h(x_j, \theta^a_{\mathbf{x}}))}$. This will make all the numerators and enumerators of the objective function non-negative, while keeping the same optimization problem. We also note that quantal response is also known as multinomial logit model in the discrete choice model literature [Train, 2003]. Our generalization here to multiple types of adversary makes the problem akin to the mixed logit model in discrete choice models, which is generally considered intractable. As a consequence, our solution addresses a basic problem in discrete choice models also.

Following our set-up, we observe N samples of the types of attackers $\widehat{\mathbf{x}}_1, \ldots, \widehat{\mathbf{x}}_N$ (which gives $\widehat{\theta}^a_1, \ldots, \widehat{\theta}^a_N, \widehat{\theta}^d_1, \ldots, \widehat{\theta}^d_N$) and we solve a robust version of the problem. Further, following our general recipe for solving the robust problem, we piecewise approximate the numerator and denominator of F using K pieces, where the dimension of $\mathbf{v}$ is $d = MK$. For this approximation, we require two additional linear constraints over the constraints in Equations (2-35). The optimization then is:

$$\max_{\mathbf{r}} \; q - s \tag{SSG}$$

$$\text{subject to Constraints (2-35)}, \sum_{j\in[M]} \sum_{k\in[K]} v_{jk} - mK \leq 0, v_{j,k} \geq v_{j,k+1}; \quad \forall k \in [K]$$

The overall additive solution bound of $O(1/K)$ can be readily inferred from Theorem 2.

Max-Capture Competitive Facility Location (MC-FLP): In this problem [Mai and Lodi, 2020], a firm has M locations ($[M]$) to set up at most $m < M$ facilities. The aim is to maximize the number of clients using this firm's facilities. The competitor(s) already have facilities running at locations $Y \subset [M]$. There are different types of clients, where types are denoted by **x**. The number of clients of type **x** is known and equal to $s_{\mathbf{x}}$. However, the distribution over types is *unknown*. The firm's decision of which location to choose is given by binary variables $z_j \in \{0, 1\}$ for $j \in [M]$. A utility of any client of type **x** for visiting location j is $V_{\mathbf{x},j}$. The *choice probability* of a client of type **x** choosing any of this firm's location is given as a quantal response model $\frac{\sum_{j\in[M]} z_j e^{V_{\mathbf{x},j}}}{\sum_{j\in[M]} z_j e^{V_{\mathbf{x},j}} + \sum_{j\in Y} e^{V_{\mathbf{x},j}}}$. For shorthand, we abuse notation and use $V_{\mathbf{x},j}$ to replace $e^{V_{\mathbf{x},j}}$ and $U_{\mathbf{x},Y}$ to replace $\sum_{j\in Y} e^{V_{\mathbf{x},j}}$. This gives $F(\mathbf{z}, \mathbf{x}) = \frac{s_{\mathbf{x}} \sum_{j\in[M]} z_j V_{\mathbf{x},j}}{\sum_{j\in[M]} z_j V_{\mathbf{x},j} + U_{\mathbf{x},Y}}$, which is interpreted as the expected number of clients of type **x** choosing this firm's facilities.

Following our set-up, we observe N samples of the types of clients samples $(\widehat{V}_{1,j})_{j\in[M]}, \ldots, (\widehat{V}_{N,j})_{j\in[M]}$ and we solve a robust version of the problem. Here, we get $F_i = \frac{s_i \sum_{j\in[M]} z_j \widehat{V}_{i,j}}{\sum_{j\in[M]} z_j \widehat{V}_{i,j} + \widehat{U}_{i,Y}}$. Next, following our general recipe for solving the robust problem, we note that F_i is already in the form $\frac{\mathbf{a}_i^T \mathbf{v} + b_i}{\mathbf{a}_i'^T \mathbf{v} + b_i'}$ where **z** plays the role of **v**. Thus, the dimension of **v** is M and no approximation is needed here for F_i; by Theorem 2, we achieve the global optimal solution by solving MISOCP optimally. The full MISOCP with an additional number of location constraint is:

$$\max_{\mathbf{r}} q - s \;\text{ subject to Constraints (2-35)}, \sum_{j\in[M]} v_j - m \leq 0.$$

Max-Capture Facility Cost Optimization (MC-FCP): In the previous **MC-FLP** problem, the budget was specified as a constraint on the number of facilities. However, often a more realistic set-up is where there is a monetary constraint and the attractiveness of a facility depends on the investment into the facility. Thus, modifying the previous problem slightly, z_j takes a different meaning of the amount of investment into facility at location j (zero investment indicates no facility). Given this investment, the attractiveness of a facility j for the client of type **x** is given as $h(z_j, \theta_{\mathbf{x},j})$ for some parameter θ dependent on **x** and j. And the *choice probability* of a client of type **x** choosing any of this firm's location is given as a quantal response model $\frac{\sum_{j\in[M]} e^{h(z_j,\theta_{\mathbf{x},j})}}{\sum_{j\in[M]} e^{h(z_j,\theta_{\mathbf{x},j})} + U_{\mathbf{x},Y}}$. This gives $F(\mathbf{z}, \mathbf{x}) = \frac{s_{\mathbf{x}} \sum_{j\in[M]} e^{h(z_j,\theta_{\mathbf{x},j})}}{\sum_{j\in[M]} e^{h(z_j,\theta_{\mathbf{x},j})} + U_{\mathbf{x},Y}}$, which is interpreted similar to **MC-FLP**. As stated, we observe N samples of the types of clients which gives $(\widehat{\theta}_{1,j})_{j\in[M]}, \ldots, (\widehat{\theta}_{N,j})_{j\in[M]}$ and we solve a robust version of the problem. Here, we get $F_i = \frac{s_i \sum_{j\in[M]} e^{h(z_j,\widehat{\theta}_{i,j})}}{\sum_{j\in[M]} e^{h(z_j,\widehat{\theta}_{i,j})} + \widehat{U}_{i,Y}}$. Next, as can be seen from the form, this is similar to the **SSG** problem and, following our general recipe, with two additional linear constraints the optimization formulation is exactly same as Equation (SSG).

4 Scaling up in Number of Samples

The transformation to a MISOCP helps in scalability over a general mixed integer concave program, but for real world dataset sizes (e.g., $80,000$ data points in our experiments) we need further scalability. We explore two related techniques towards this end: clustering and stratified sampling. For both approaches, we obtain a representative subset of S data points ($S << N$) and a modified weighted objective, which converts to a much smaller tractable MISOCP compared to the original problem.

For solution guarantees, we need mild assumptions: in particular, for the rest of this section we assume a bounded F, i.e., for some fixed ψ $\max_{\mathbf{z}}\{F(\mathbf{z},\mathbf{x})\} - \min_{\mathbf{z}}\{F(\mathbf{z},\mathbf{x})\} \leq \psi^2 \quad \forall \widehat{\mathbf{x}}_1, \ldots, \widehat{\mathbf{x}}_N$ and τ-lipschitzness of F in the argument $\mathbf{x}$: $|F(\mathbf{z},\mathbf{x}') - F(\mathbf{z},\mathbf{x})| \leq \tau ||\mathbf{x}' - \mathbf{x}||_2 \quad \forall \mathbf{z}$.

Clustering Approach: We cluster the N points $\mathbf{x}_1, \ldots \mathbf{x}_N$ into S groups and for each group s we have $||\mathbf{x}_i - \mathbf{x}^s|| \leq \epsilon$, where $\mathbf{x}^s$ is the cluster center of cluster s. We call ϵ the clustering radius. Let C_s be the number of points in the cluster s, hence $\sum_{s\in[S]} C_s = N$. We use a shorthand for the original objective function of the MISOCP $\mathcal{G}(\mathbf{z})$:

$$\sum_i \frac{F(\mathbf{z},\widehat{\mathbf{x}}_i)}{N} - \sqrt{\rho \sum_i \Big(\sum_i \frac{F(\mathbf{z},\widehat{\mathbf{x}}_i)}{N} - F(\mathbf{z},\widehat{\mathbf{x}}_i) \Big)^2} = \widehat{\text{Mean}}(F(\mathbf{z},\mathbf{x})) - \sqrt{\rho \widehat{\text{Var}}(F(\mathbf{z},\mathbf{x}))}$$

where $\widehat{\text{Mean}}$ is empirical mean and $\widehat{\text{Var}}$ is *unnormalized variance*. After clustering, we solve for the same problem but only with cluster centers and appropriate *weighing*, to get modified objective $\widehat{\mathcal{G}}(\mathbf{z})$:

$$\sum_s C_s \frac{F(\mathbf{z},\mathbf{x}^s)}{N} - \sqrt{\rho \sum_s C_s \Big(\sum_s C_s \frac{F(\mathbf{z},\mathbf{x}^s)}{N} - F(\mathbf{z},\mathbf{x}^s) \Big)^2} = \widehat{\text{Mean}}^S(F(\mathbf{z},\mathbf{x})) - \sqrt{\rho \widehat{\text{Var}}^S(F(\mathbf{z},\mathbf{x}))}$$

The conversion to MISOCP is exactly the same, except for F_i's being weighted as shown above; details of conversion are in the appendix. We bound the approximation incurred by the two terms above (weighted mean and unnormalized weighted variance) separately below

Lemma 1. *Under assumptions stated above, we have* $\left| \widehat{Mean}(F(z,\boldsymbol{x})) - \widehat{Mean}^S(F(z,\boldsymbol{x})) \right| \leq \tau\epsilon$ *and* $\left| \sqrt{\rho \widehat{Var}(F(z,\boldsymbol{x}))} - \sqrt{\rho \widehat{Var}^S(F(z,\boldsymbol{x}))} \right| \leq (\psi + \sqrt{2\tau\epsilon})\sqrt{\frac{2\tau\epsilon\xi}{N}}$.

The next result is obtained by using the lemma above

Theorem 3. *Given the assumptions stated above, and* $\widehat{z}$ *an optimal solution for* $\max_z \widehat{\mathcal{G}}(z)$ *and* z^* *optimal for MISOCP* $\max_z \mathcal{G}(z)$*, the following holds:* $|\mathcal{G}(\widehat{z}) - \mathcal{G}(z^*)| \leq 2(\tau\epsilon + \psi\sqrt{\frac{2\tau\epsilon\xi}{N}} + \frac{2\tau\epsilon\xi}{\sqrt{N}})$.

Stratified Sampling: Similar in spirit to clustering, the space of $\mathbf{x}$ space is divided into T strata. Each strata has C_t samples, such that $\sum_{t\in[T]} C_t = N$. Next, distinct from the clustering approach, we draw N_t samples randomly from the t^{th} stratum with a total of $\sum_t N_t = S$ samples (note same number of total samples S as in clustering). For each stratum t we have $||\mathbf{x}_i - \mathbf{x}_j|| \leq d_t$ for any $\mathbf{x}_i, \mathbf{x}_j$ in stratum t. We denote a random sample in stratum t as $\widehat{\mathbf{x}}^j$ where $j \in [N_t]$ (note superscript is to distinguish from the subscript used to index all the $\widehat{\mathbf{x}}$'s). This approach is the preferred one if the clustering approach results in cluster centers that are not allowed as parameter values (e.g., cluster center may be fractional where $\mathbf{x}$'s can only be integral).

Let $l_t = \frac{C_t}{N_t}$. Use $\widehat{Mean}^T(F(\mathbf{z},\mathbf{x}))$ to stand for $\frac{1}{N}\sum_{t\in[T]} l_t \sum_{j\in[N_t]} F(\mathbf{z},\widehat{\mathbf{x}}^j)$ and $\widehat{Var}^T(\mathbf{z},\widehat{\mathbf{x}})$ for $\sum_{t\in[T]} l_t \sum_{j\in[N_t]} \Big(\widehat{Mean}(F(\mathbf{z},\mathbf{x})) - F(\mathbf{z},\widehat{\mathbf{x}}^j)\Big)$. After stratified sampling our modified weighted objective $\widehat{G}(\mathbf{z})$ is $\widehat{Mean}^T(F(\mathbf{z},\mathbf{x})) - \sqrt{\rho \widehat{Var}^T(\mathbf{z},\widehat{\mathbf{x}})}$. Next, similar to clustering, bounds for $\widehat{Mean}^T$ and $\widehat{Var}^T$ but with high probability (Lemma 2,3 in appendix) lead to the main result:

Theorem 4. *Let* $D = \max_{z,\boldsymbol{x}} |F(z,\boldsymbol{x})|$ *for bounded function* F*. Given the assumptions stated above, and* $\widehat{z}$ *an optimal solution for* $\max_z \widehat{\mathcal{G}}(z)$ *and* z^* *optimal for MISOCP* $\max_z \mathcal{G}(z)$*, and* $N_* = \min_t N_t$*, the following statement holds with probability* $\geq 1 - 2\sum_t \exp^{\frac{-2\sqrt{N_*}\epsilon^2}{\tau^2 d_t^2}} - 4\sum_t \exp^{\frac{-2\sqrt{N_*}\epsilon^2}{4\tau^2 d_t^2 D^2}}$*:*

$$|\mathcal{G}(\widehat{z}) - \mathcal{G}(z^*)| \leq \frac{2\epsilon}{(N_*)^{1/4}}\Big(1 + 2\sqrt{\frac{\xi}{\widehat{Var}(F(z,\boldsymbol{x}))}}\Big).$$

Thus, with increasing samples in all strata, optimality gap approaches 0 with prob. approaching 1.

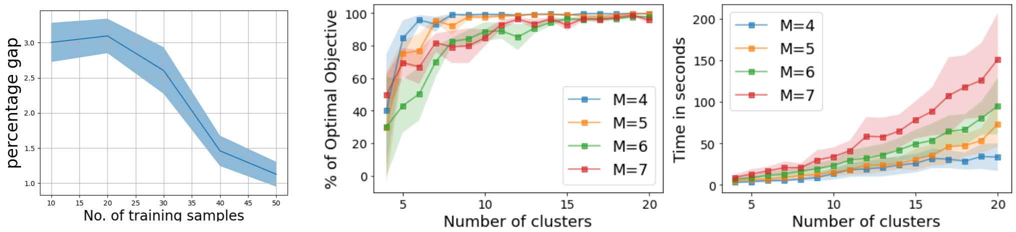


Figure 1: (Left) % gap between the utility of decisions using true and learned regressor with varying training data size N_T. (Middle) Objective value achieved using clustering approach as a % of **OPT**. (Right) Time to solve each optimization to optimality. Middle and right results are shown for varying alternatives number M. Underlying parameters are $N = 500, m = 1, \xi =$1E6.

Table 1: We use clustering/stratified sampling to approximately solve a problem (**Approx-OPT**). The table shows the mean percentage: $\frac{100\times\textbf{Approx-OPT}}{\textbf{OPT}}$ and standard deviation over 10 different synthetic SSG datasets with underlying parameters $N = 500, M = 6, m = 1, K = 10$.

Method	Total no. of samples (S)		
	8	16	24
Clustering	80.17 ± 2.14	90.74 ± 1.20	94.03 ± 0.57
1 per strata	85.49 ± 1.94	94.59 ± 0.76	99.25 ± 0.39
2 per strata	91.78 ± 1.33	94.55 ± 0.64	99.54 ± 0.25
4 per strata	78.89 ± 1.73	94.86 ± 0.94	98.77 ± 0.39
8 per strata	92.19 ± 0.67	95.85 ± 0.80	99.37 ± 0.25
Uniform sampling (no cluster/strata)	73.92 ± 33.71	78.08 ± 28.62	79.86 ± 26.81

Table 2: Objective values as a % of **OPT** across various methods repeated over 5 synthetic SSG datasets with parameters $N = 500, M = 10, m = 1$ for varying regularization (ξ, on left) and $N = 500, m = 1, \xi =$ 1E6 for varying no. of targets (M, on right). The no. of clusters/strata is 50.

Method	Regularization (ξ)				No. of Alternatives (M)		
	1E3	1E4	1E5	1E6	10	25	50
TT-GAD	99.8±0.1	99.4±0.1	92.7±0.3	82.6±0.4	82.6 ± 0.4	90.2 ± 0.5	92.2 ± 0.3
PGA	98.9±0.1	98.1±0.2	87.7±0.5	49.2±0.9	49.2 ± 0.9	90.9 ± 0.7	93.5 ± 0.4
Clustering	99.9±0.1	99.9± 0.1	99.8±0.1	99.6±0.1	99.6 ± 0.1	99.5 ± 0.2	99.4 ± 0.1
Sampling	100.0±0.0	99.9± 0.1	99.9±0.1	99.8±0.1	99.8 ± 0.1	99.6 ± 0.1	99.5 ± 0.2

5 Experiments

We evaluate our methods on (a) Stackleberg Security Games (**SSG**) with Quantal Response (synthetic data), (b) Maximum capture Facility Location Planning (**MC-FLP**) and (c) Maximum capture Facility Cost Planning (**MC-FCP**). Empirically we demonstrate (i) better solution quality of our method compared to baselines, (ii) practical scalability of our method, and (iii) improvement over non-robust optimization on those data points that contribute least to the objective (akin to rare classes in classification) while not sacrificing average performance. We fix $K = 10$ in approximation via discretization as we find that objective increase saturates for this K (see Appendix K). We use a 2.1 GHz CPU with 128GB RAM.

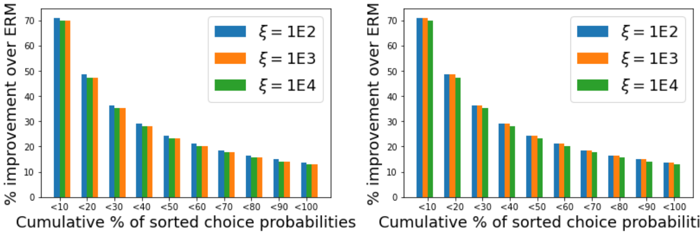
Baselines: We use the following two methods as baselines: (i) Projected Gradient Ascent (**PGA**) on the formulation (VR), (ii) Two Time Scale Gradient Ascent Descent (**TT-GAD**) [Lin et al., 2020] on the formulation (DRO) where the inner minimization is convex and the outer maximization is non-concave. The numbers reported for our baselines are the *best values* over 10 random initializations.

5.1 SSG with Quantal Response (Synthetic Data)

We generate attacker and defender utilities following Yang et al. [2012]; a complete description of data generation and choice of f^* is in Appendix I. We generate five datasets of size $N = 500$ each in order to observe the variance of every result reported in this sub-section; this is also the largest size

Table 3: Average client choice probabilities for availing the facility across various settings. H denotes average over those 5% of the clients in test data with the lowest choice probabilities, A denotes average over all the samples in the test set.

ξ	**MC-FCP**						**MC-FLP**					
	m=7		m=10		m=13		m=10		m=12		m=14	
	H	A	H	A	H	A	H	A	H	A	H	A
ERM	0.069	0.692	0.150	0.719	0.420	**0.758**	0.175	0.741	0.426	**0.769**	0.469	0.772
1E2	0.069	0.692	**0.417**	**0.751**	0.4170	0.751	**0.418**	**0.763**	0.426	**0.769**	0.469	0.772
1E3	**0.093**	**0.697**	0.416	0.747	0.417	0.757	**0.418**	**0.763**	0.425	0.768	0.533	**0.777**
1E4	**0.093**	**0.697**	0.416	0.747	**0.446**	0.750	0.417	0.759	**0.531**	0.767	**0.539**	0.769



(A) MC-FCP

(B) MC-FLP

Figure 2: The bar plots show the percentage improvement of choice probabilities of our clustering approach over ERM over cumulative buckets of choice probabilities (see text) in ascending order with varying regularization ξ with fixed m=10.

that we could solve exactly optimally within an hour using all the data points. First, we empirically validate Theorem 1 by plotting in Figure 1(left) the relative gap between the true utility $E_{P^*}[F(\cdot, \mathbf{x})]$ of the decisions output by running DRO on the output $(\mathbf{x}^*)_{i\in[N]}$ of the true f^* (assumed fixed linear function) versus on $(\widehat{\mathbf{x}})_{i\in[N]}$ from learned f, as the training data size N_T for learning f is varied.

Next, we focus on only using $(\widehat{\mathbf{x}})_{i\in[N]}$ and the optimal solution for $(\widehat{\mathbf{x}})_{i\in[N]}$ is named as **OPT**. Figure 1 (middle) demonstrates empirically that the solution of the optimization problem on cluster centers converges to **OPT** with only a few number of clusters and the time for the optimization shown in Figure 1 (right) is reasonable. Next, the results in Table 1 show a comparison of the clustering and stratified sampling approach using the metric of how close they get to **OPT**. We find stratified sampling to be better than clustering in almost all cases and a simple uniform sampling (no cluster/strata) fails to return a solution close to **OPT**. Table 2 (left) demonstrates that the baselines struggle to reach the optimal value objective as the magnitude of regularization (ξ) increases. Intuitively, as ξ increases the variance term (which is highly non-convex) contributes more to the objective and stationary points reached by the baselines are quite sub-optimal compared to the global optimal. In addition, with increasing ξ the ambiguity set becomes larger possibly containing more local optimal solutions. We also study varying the parameter M (m fixed) and Table 2 (right) shows that our approaches outperform the gradient-based baselines across different values of M.

5.2 MC-FLP and MC-FCP (Real Data)

P&R-NYC Dataset : We use a large and challenging Park-and-ride (P&R) dataset collected in New York City, which provides utilities for 82341 clients (N) for 59 park and ride locations (M), along with their incumbent utilities for competing facilities [Holguin-Veras et al., 2012]; this data was directly used for **MC-FLP**. For **MC-FCP** we additionally use generated costs, which are not present in the P&R data. A complete description of data generation is in Appendix I. Both these problems could not be solved at all with our MISOCP alone (no clustering) as the optimization did not finish in 24 hours. Hence, we use our clustering approach with 50 clusters.

We compare to a baseline solution of the non-robust empirical risk minimization (ERM) (also called the sample average approximation or SAA). We split the data (randomly) into training and test (80:20) and then obtain the decision $\widehat{\mathbf{z}}$ using the training data. Then, we obtain the choice probability (recall this as probability of a client choosing any of the firm's facility) for every client in the test data for the decision $\widehat{\mathbf{z}}$. We compare the performance of ERM and our method for clients (in test set) bucketed by choice probabilities in Figure 2 and Table 3. The buckets are made by sorting all the clients in the test set by ascending choice probabilities and then considering cumulative buckets as the first 5%, the first 10% and so on. In Fig. 2, we show that the average percentage improvement in choice probabilities of our robust approach over ERM is considerably higher for clients with lower choice probability (these clients contribute least to the objective) and the over all average over all clients (rightmost on x-axis) is slightly better than ERM. In Table 3, note the significantly increased probabilities for low choice probability clients (low choice prob. using $\widehat{\mathbf{z}}$) without compromising the average performance across all clients for varying M. Additional results are in Appendix J.

6 Conclusion

We presented an approach for a distributionally robust solution to a class of non-convex sum of fractional solutions, with guaranteed near global optimality. We presented application to three prominent practical problems and the connection to discrete choice models opens up possibilities of applying our approach to even more problems. Further investigation on how to cluster or stratify more effectively (than k-means) to achieve even more scalability is a possible future research direction. Further, we used a χ^2-divergence based ambiguity set, which only covers nearby distributions with same support as the given data samples; exploring Wasserstein ambiguity sets is a possible future research direction. We hope that our work inspires tackling robust formulation of more classes of non-convex problems, with guarantees for global optimality.

Acknowledgement

This research/project is supported by the National Research Foundation, Singapore under its AI Singapore Programme (AISG Award No: AISG2-RP-2020-017).

References

Felipe Aros-Vera, Vladimir Marianov, and John E Mitchell. p-hub approach for the optimal park-and-ride facility location problem. *European Journal of Operational Research*, 226(2):277–285, 2013.

Aharon Ben-Tal, Dick Den Hertog, Anja De Waegenaere, Bertrand Melenberg, and Gijs Rennen. Robust solutions of optimization problems affected by uncertain probabilities. *Management Science*, 59(2):341–357, 2013.

Stefano Benati and Pierre Hansen. The maximum capture problem with random utilities: Problem formulation and algorithms. *European Journal of Operational Research*, 143(3), 2002.

Jose Blanchet and Karthyek Murthy. Quantifying distributional model risk via optimal transport. *Mathematics of Operations Research*, 44(2):565–600, 2019.

Pierre Bonami and A Tramontani. Recent improvement to misocp in cplex. *INFORMS, Philadelphia, PA, USA, Tech. Rep*, 2015.

Tien Thanh Dam, Thuy Anh Ta, and Tien Mai. Submodularity and local search approaches for maximum capture problems under generalized extreme value models. *European Journal of Operational Research*, 2021. ISSN 0377-2217.

John Duchi and Hongseok Namkoong. Variance-based regularization with convex objectives. *The Journal of Machine Learning Research*, 20(1):2450–2504, 2019.

John C Duchi, Peter W Glynn, and Hongseok Namkoong. Statistics of robust optimization: A generalized empirical likelihood approach. *Mathematics of Operations Research*, 2021.

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

Fei Fang, Thanh H Nguyen, Rob Pickles, Wai Y Lam, Gopalasamy R Clements, Bo An, Amandeep Singh, Brian C Schwedock, Milin Tambe, and Andrew Lemieux. PAWS—A deployed game-theoretic application to combat poaching. *AI Magazine*, 38(1), 2017.

Alexandre S Freire, Eduardo Moreno, and Wilfredo F Yushimito. A branch-and-bound algorithm for the maximum capture problem with random utilities. *European journal of operational research*, 252(1), 2016.

Nika Haghtalab, Fei Fang, Thanh H. Nguyen, Arunesh Sinha, Ariel D. Procaccia, and Milind Tambe. Three strategies to success: Learning adversary models in security games. In *25th International Joint Conference on Artificial Intelligence (IJCAI)*, 2016.

Jose Holguin-Veras, Jack Reilly, Felipe Aros-Vera, et al. New york city park and ride study. Technical report, University Transportation Research Center, 2012.

Henry Lam. Robust sensitivity analysis for stochastic systems. *Mathematics of Operations Research*, 41(4):1248–1275, 2016.

Hongmin Li, Scott Webster, Nicholas Mason, and Karl Kempf. Product-line pricing under discrete mixed multinomial logit demand: winner—2017 m&som practice-based research competition. *Manufacturing & Service Operations Management*, 21(1):14–28, 2019.

Tianyi Lin, Chi Jin, and Michael Jordan. On gradient descent ascent for nonconvex-concave minimax problems. In *International Conference on Machine Learning*, pages 6083–6093. PMLR, 2020.

Tien Mai and Andrea Lodi. A multicut outer-approximation approach for competitive facility location under random utilities. *European Journal of Operational Research*, 284(3), 2020.

Andreas Maurer and Massimiliano Pontil. Empirical bernstein bounds and sample-variance penalization. In *COLT*, 2009.

Garth P McCormick. Computability of global solutions to factorable nonconvex programs: Part i—convex underestimating problems. *Mathematical programming*, 10(1):147–175, 1976.

Georg Pflug and David Wozabal. Ambiguity in portfolio selection. *Quantitative Finance*, 7(4): 435–442, 2007.

Qi Qi, Zhishuai Guo, Yi Xu, Rong Jin, and Tianbao Yang. An online method for a class of distributionally robust optimization with non-convex objectives. *Advances in Neural Information Processing Systems*, 34, 2021.

Hamed Rahimian and Sanjay Mehrotra. Distributionally robust optimization: A review. *arXiv preprint arXiv:1908.05659*, 2019.

Henry WJ Reeve and Ata Kaban. Optimistic bounds for multi-output prediction. In *37th International Conference on Machine Learning (ICML 2020)*, 2020.

Siegfried Schaible. Fractional programming. In *Handbook of global optimization*, pages 495–608. Springer, 1995.

Soroosh Shafieezadeh-Abadeh, Peyman Mohajerin Esfahani, and Daniel Kuhn. Distributionally robust logistic regression. *arXiv preprint arXiv:1509.09259*, 2015.

Arunesh Sinha, Fei Fang, Bo An, Christopher Kiekintveld, and Milind Tambe. Stackelberg security games: Looking beyond a decade of success. In *27th International Joint Conference on Artificial Intelligence (IJCAI)*, 2018.

Matthew Staib, Bryan Wilder, and Stefanie Jegelka. Distributionally robust submodular maximization. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 506–516. PMLR, 2019.

Milind Tambe. *Security and game theory: algorithms, deployed systems, lessons learned.* Cambridge university press, 2011.

Kenneth Train. *Discrete Choice Methods with Simulation*. Cambridge University Press, 2003.

Haifeng Xu. The mysteries of security games: Equilibrium computation becomes combinatorial algorithm design. In *Proceedings of the 2016 ACM Conference on Economics and Computation*, pages 497–514, 2016.

Yan Yan, Yi Xu, Lijun Zhang, Wang Xiaoyu, and Tianbao Yang. Stochastic optimization for non-convex inf-projection problems. In *International Conference on Machine Learning*, pages 10660–10669. PMLR, 2020.

Rong Yang, Fernando Ordonez, and Milind Tambe. Computing optimal strategy against quantal response in security games. In *AAMAS*, pages 847–854, 2012.

Rong Yang, Benjamin J Ford, Milind Tambe, and Andrew Lemieux. Adaptive resource allocation for wildlife protection against illegal poachers. In *Aamas*, pages 453–460, 2014.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [No]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [N/A] It is public use
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A Approximation via Discretization (AD)

Recall the general functional form

$$F(\mathbf{z}, \widehat{\mathbf{x}}_i) = \frac{\sum_j n(z_j, \widehat{\mathbf{x}}_i)}{\sum_j d(z_j, \widehat{\mathbf{x}}_i)}$$

and note that both the numerator and denominator are separable in the components of the decision variables $\mathbf{z}$. Let assume each variable z_j can vary in the interval $[L_j, U_j]$, the idea here is to divide each interval $[L_j, U_j]$ into K equal sub-intervals of size $(U_j - L_j)/K$ and approximate z_j by K binary variables $v_{jk} \in \{0, 1\}$ as

$$z_j = L_j \frac{U_j - L_j}{K} \sum_{k \in [K]} v_{jk},$$

where $v_{jk} \in \{0, 1\}$ satisfying $v_{ik} \geq v_{j,k+1}$ for $k = 1, 2, \ldots, K-1$. We then approximate $n(z_j, \widehat{\mathbf{x}}_i)$ and $d(z_j, \widehat{\mathbf{x}}_i)$ as

$$n(z_j, \widehat{\mathbf{x}}_i) \approx \widehat{n}(z_j, \widehat{\mathbf{x}}_i) = n\left(L_j + \lfloor z_j K/(U_j - L_j) \rfloor \frac{U_j - L_j}{K}, \widehat{\mathbf{x}}_i\right) = n(L_j, \widehat{\mathbf{x}}_i) + \frac{U_j - L_j}{K} \sum_{k \in [K]} \gamma^{ni}_{jk} v_{jk},$$

$$d(z_j, \widehat{\mathbf{x}}_i) \approx \widehat{d}(z_j, \widehat{\mathbf{x}}_i) = d\left(L_j + \lfloor z_j K/(U_j - L_j) \rfloor \frac{U_j - L_j}{K}, \widehat{\mathbf{x}}_i\right) = d(L_j, \widehat{\mathbf{x}}_i) + \frac{U_j - L_j}{K} \sum_{k \in [K]} \gamma^{di}_{jk} v_{jk}$$

where γ^{ni}_{jk} and γ^{di}_{jk} are the slopes of the approximate linear functions in $[L_j + (U_j - L_j)(k-1)/K; L_j + (U_j - L_j)(k)/K]$, $\forall k = 1, \ldots, K$, computed as

$$\gamma^{ni}_{j,k+1} = \frac{K}{U_j - L_j}\left(n\left(L_j + \frac{(U_j - L_j)(k+1)}{K}, \widehat{\mathbf{x}}_i\right) - n\left(L_j + \frac{(U_j - L_j)(k)}{K}, \widehat{\mathbf{x}}_i\right)\right),\ k = 0, \ldots, K-1$$

$$\gamma^{di}_{j,k+1} = \frac{K}{U_j - L_j}\left(d\left(L_j + \frac{(U_j - L_j)(k+1)}{K}, \widehat{\mathbf{x}}_i\right) - d\left(L_j + \frac{(U_j - L_j)(k)}{K}, \widehat{\mathbf{x}}_i\right)\right),\ k = 0, \ldots, K-1.$$

We can then approximate $F(\mathbf{z}, \widehat{\mathbf{x}}_i)$ as

$$F(\mathbf{z}, \widehat{\mathbf{x}}_i) \approx \frac{\sum_j (n(L_j, \widehat{\mathbf{x}}_i) + \frac{U_j - L_j}{K} \sum_{k \in [K]} \gamma^{ni}_{jk} v_{jk})}{\sum_j (d(L_j, \widehat{\mathbf{x}}_i) + \frac{U_j - L_j}{K} \sum_{k \in [K]} \gamma^{di}_{jk} v_{jk})}.$$

The transformed/approximated problem will have the following parameters and variables

- $\mathbf{a}_i = \left[\gamma^{ni}_{jk} \middle| \ j \in [M], k \in [K]\right]$
- $\mathbf{a}'_i = \left[\gamma^{di}_{jk} \middle| \ j \in [M], k \in [K]\right]$
- $b_i = \sum_{j \in [M]} n(L_j, \widehat{\mathbf{x}}_i)$
- $b'_i = \sum_{j \in [M]} d(L_j, \widehat{\mathbf{x}}_i)$
- $\mathbf{v} \in \mathcal{V} \overset{\text{def}}{=} \left\{v_{jk} \middle| \ v_{jk} \in \{0, 1\}, v_{jk} \geq v_{j,k+1}, j \in [M], k \in [K]\right\}.$

B Proof of Theorem 1

Theorem. *As described above, let $\boldsymbol{x}^*_i = f^*(b_i)$ for true function f^* and let $\widehat{\boldsymbol{x}}_i = f(b_i)$ for the learned empirical risk minimizer f. Suppose the optimal decision when solving* DRO *is $\boldsymbol{z}^{**}$ using $\boldsymbol{x}^*_i$'s and $\widehat{\boldsymbol{z}}^{**}$ using $\widehat{\boldsymbol{x}}_i$'s. Also, let F be τ-Lipschitz in $\boldsymbol{x}$, X be bounded, and a scaled $\mathcal{L}$ upper bound $||\cdot||_2$ (i.e., $||\boldsymbol{x} - \boldsymbol{x}'||_2 \leq \max(k\mathcal{L}(\boldsymbol{x}, \boldsymbol{x}'), \epsilon)$ for constants k, ϵ) then, the following holds with probability $1 - 2\delta - 2\delta_1$: $\mathbb{E}_{P^*}[F(\widehat{\boldsymbol{z}}^{**}, \boldsymbol{x})] \geq \mathbb{E}_{P^*}[F(\boldsymbol{z}^{**}, \boldsymbol{x})] - C/\sqrt{N} - (1 + 2\sqrt{\xi})\tau\epsilon - \epsilon_N - \epsilon_{N_T}$, where $\epsilon_K = C_1 \mathcal{R}_K(\mathcal{L} \circ \mathcal{F}) + C_2/\sqrt{K}$ and $\mathcal{R}_K$ is the Rademacher complexity with K samples and C, C_1, C_2 are constants dependent on $\delta, \delta_1, \xi, k, \tau$.*

Proof. We first list the mild assumptions: (1) $||\mathbf{x}' - \mathbf{x}||_2 \leq \max(k\mathcal{L}(\mathbf{x}', \mathbf{x}), \epsilon)$ for some constant k and a small constant ϵ and (2) space X (that contains $\widehat{\mathbf{x}}, \mathbf{x}^*$) is bounded with a diameter d_X. The k in the first assumption can be found since the space X is bounded, and for close $\mathbf{x}', \mathbf{x}$, if needed, ϵ provides an upper bound. With this, it is easy to check that

$$(1/N)\sum_i ||\mathbf{x}_i^* - \widehat{\mathbf{x}}_i||_2 \leq (1/N)\sum_i \max(k\mathcal{L}(\mathbf{x}_i^*, \widehat{\mathbf{x}}_i), \epsilon) \leq \epsilon + (1/N)\sum_i k\mathcal{L}(\mathbf{x}_i^*, \widehat{\mathbf{x}}_i)).$$

We also have $(1/N)\sum_i \mathcal{L}(\widehat{\mathbf{x}}_i, \mathbf{x}_i^*) \leq \mathbb{E}[\mathcal{L}_f] + \epsilon_N$, where ϵ_N is of the form $C_1\mathcal{R}_N(\mathcal{L} \circ \mathcal{F}) + \frac{C_2}{\sqrt{N}}$ for constants C_1, C_2 that depend on the probability term δ, $\mathbb{E}[\mathcal{L}_f]$ is the expected risk of function f, $\mathcal{R}_N$ is the Rademacher complexity with N samples, and $\mathcal{R}_N(\mathcal{L} \circ \mathcal{F})$ is well-defined for vector valued output of functions in $\mathcal{F}$ using each component of the output (see Proposition 1 in Reeve and Kaban [2020]). Next, the true risk of f^* is assumed zero (text above the theorem in main paper): $\mathbb{E}[\mathcal{L}_{f^*}] = 0$. Then, the ERM training using N_T training data provides a high probability $1 - \delta$ guarantee which can be stated as $\mathbb{E}[\mathcal{L}_f] \leq \epsilon_{N_T}$, where ϵ_{N_T} is defined is same way as ϵ_N with N_T replacing N.

With this, we further have with probability $1 - 2\delta$

$$(1/N)\sum_i ||\mathbf{x}_i^* - \widehat{\mathbf{x}}_i||_2 \leq \epsilon + k(\epsilon_N + \epsilon_{N_T}).$$

Let $\mathbf{z}^{**}$ and $p_1^{**}, \ldots, p_N^{**}$ be the optimal solution found for $\mathbf{x}_1^*, \ldots, \mathbf{x}_N^*$. Note that $\mathbf{z}^{**}$ and $p_1^{**}, \ldots, p_N^{**}$ is also a feasible point for the optimization with $\widehat{\mathbf{x}}_1, \ldots, \widehat{\mathbf{x}}_N$. We know that

$$|\sum_i p_i^{**} F(\mathbf{z}^{**}, \widehat{\mathbf{x}}_i) - \sum_i p_i^{**} F(\mathbf{z}^{**}, \mathbf{x}_i^*)| = |\sum_i p_i^{**}(F(\mathbf{z}^{**}, \widehat{\mathbf{x}}_i) - F(\mathbf{z}^{**}, \mathbf{x}_i^*))|$$
$$= |\sum_i (p_i^{**} - 1/N)(F(\mathbf{z}^{**}, \widehat{\mathbf{x}}_i) - F(\mathbf{z}^{**}, \mathbf{x}_i^*)) + \sum_i (1/N)(F(\mathbf{z}^{**}, \widehat{\mathbf{x}}_i) - F(\mathbf{z}^{**}, \mathbf{x}_i^*))|$$

We know that for any feasible $\mathbf{p}, \mathbf{z}$

$$|\sum_i p_i F(\mathbf{z}, \widehat{\mathbf{x}}_i) - \sum_i p_i F(\mathbf{z}, \mathbf{x}_i^*)| = |\sum_i p_i(F(\mathbf{z}, \widehat{\mathbf{x}}_i) - F(\mathbf{z}, \mathbf{x}_i^*))|$$
$$= |\sum_i (p_i - 1/N)(F(\mathbf{z}, \widehat{\mathbf{x}}_i) - F(\mathbf{z}, \mathbf{x}_i^*)) + \sum_i (1/N)(F(\mathbf{z}, \widehat{\mathbf{x}}_i) - F(\mathbf{z}, \mathbf{x}_i^*))|$$

Note that by Lispschitzness,

$$|\sum_i (1/N)(F(\mathbf{z}, \widehat{\mathbf{x}}_i) - F(\mathbf{z}, \mathbf{x}_i^*))| \leq \sum_i (1/N)\tau||\widehat{\mathbf{x}}_i - \mathbf{x}_i^*||_2 \leq \tau(\epsilon + k(\epsilon_N + \epsilon_{N_T})). \tag{12}$$

Also, $|\sum_i (p_i - 1/N)(F(\mathbf{z}, \widehat{\mathbf{x}}_i) - F(\mathbf{z}, \mathbf{x}_i^*))| \leq \sum_i |(p_i - 1/N)(F(\mathbf{z}, \widehat{\mathbf{x}}_i) - F(\mathbf{z}, \mathbf{x}_i^*))|$ and by Holder's inequality with $\infty, 1$ norm we get

$$\sum_i |(p_i - 1/N)(F(\mathbf{z}, \widehat{\mathbf{x}}_i) - F(\mathbf{z}, \mathbf{x}_i^*))| \leq \Big(\max_i |(p_i - 1/N)|\Big)\sum_i |F(\mathbf{z}, \widehat{\mathbf{x}}_i) - F(\mathbf{z}, \mathbf{x}_i^*)|$$

Since, $||\mathbf{p} - \mathbf{1}/N||_2^2 \leq \xi/N^2$, thus, $\max_i |(p_i - 1/N)| \leq \sqrt{\xi}/N$. Hence, we get

$$\sum_i |(p_i - 1/N)(F(\mathbf{z}, \widehat{\mathbf{x}}_i) - F(\mathbf{z}, \mathbf{x}_i^*))| \leq \sqrt{\xi}(1/N)\sum_i |F(\mathbf{z}, \widehat{\mathbf{x}}_i) - F(\mathbf{z}, \mathbf{x}_i^*)| \leq \sqrt{\xi}\tau(\epsilon + k(\epsilon_N + \epsilon_{N_T})) \tag{13}$$

With this, overall we get for any feasible $\mathbf{p}, \mathbf{z}$

$$|\sum_i p_i F(\mathbf{z}, \widehat{\mathbf{x}}_i) - \sum_i p_i F(\mathbf{z}, \mathbf{x}_i^*)| \leq (1 + \sqrt{\xi})\tau(\epsilon + k(\epsilon_N + \epsilon_{N_T})) = \psi \tag{14}$$

Note the following inequalities

$$
\begin{aligned}
\sum_i p_i^{**} F(\mathbf{z}^{**}, \mathbf{x}_i^*) \le & \sum_i \widehat{p}_i^{**} F(\mathbf{z}^{**}, \mathbf{x}_i^*) && (15)\\
& = \Big(\sum_i \widehat{p}_i^{**} F(\mathbf{z}^{**}, \mathbf{x}_i^*) - \sum_i \widehat{p}_i^{**} F(\mathbf{z}^{**}, \widehat{\mathbf{x}}_i^*) \Big) + \sum_i \widehat{p}_i^{**} F(\mathbf{z}^{**}, \widehat{\mathbf{x}}_i^*) && (16)\\
& \le \psi + \sum_i \widehat{p}_i^{**} F(\mathbf{z}^{**}, \widehat{\mathbf{x}}_i^*) && (17)\\
& \le \psi + \sum_i \widehat{p}_i^{**} F(\widehat{\mathbf{z}}^{**}, \widehat{\mathbf{x}}_i^*) && (18)\\
& \le \psi + \sum_i p_i^{**} F(\widehat{\mathbf{z}}^{**}, \widehat{\mathbf{x}}_i^*) && (19)\\
& = \psi + \Big(\sum_i p_i^{**} F(\widehat{\mathbf{z}}^{**}, \widehat{\mathbf{x}}_i^*) - \sum_i p_i^{**} F(\widehat{\mathbf{z}}^{**}, \mathbf{x}_i^*) \Big) + \sum_i p_i^{**} F(\widehat{\mathbf{z}}^{**}, \mathbf{x}_i^*) && (20)\\
& \le 2\psi + \sum_i p_i^{**} F(\widehat{\mathbf{z}}^{**}, \mathbf{x}_i^*) && (21)
\end{aligned}
$$

where the first inequality is since p_i^{**} is minimizer, Eq. 17 is from Eq. 14, Eq. 18 is since $\widehat{\mathbf{z}}^{**}$ is maximizer, Eq. 19 is since $\widehat{p}_i^{**}$ is minimizer, and the last inequality is from Eq. 14.

Next, by writing p_i^{**} as $(p_i^{**} - 1/N) + 1/N$, we get from the above that

$$
1/N \sum_i (F(\mathbf{z}^{**}, \mathbf{x}_i^*) \le 2\psi + \sum_i (p_i^{**} - 1/N)(F(\widehat{\mathbf{z}}^{**}, \mathbf{x}_i^*) - F(\mathbf{z}^{**}, \mathbf{x}_i^*)) + 1/N \sum_i F(\widehat{\mathbf{z}}^{**}, \mathbf{x}_i^*)
$$

By, Eq. 13 and that $\psi = (1 + \sqrt{\xi})\tau(\epsilon + k(\epsilon_N + \epsilon_{N_T}))$, we get

$$
1/N \sum_i (F(\mathbf{z}^{**}, \mathbf{x}_i^*) \le (1 + 2\sqrt{\xi})\tau(\epsilon + k(\epsilon_N + \epsilon_{N_T})) + 1/N \sum_i F(\widehat{\mathbf{z}}^{**}, \mathbf{x}_i^*)
$$

Also we absorb all constants in $(1 + 2\sqrt{\xi})\tau\epsilon$ to call it just ϵ and likewise, $(1 + 2\sqrt{\xi})\tau k(\epsilon_N + \epsilon_{N_T})$ is just $\epsilon_N + \epsilon_{N_T}$. Further, a standard concentration inequality for τ-Lipschitz $F(\mathbf{z}, \cdot)$ and bounded diameter d_X of space X can be invoked with the two decisions to get

$$
P\left(\frac{1}{N} \sum_{i \in [N]} F(\mathbf{z}^{**}, \mathbf{x}_i^*) \ge \mathbb{E}_{\mathbf{x} \sim P^*}[F(\mathbf{z}^{**}, \mathbf{x})] - t \right) \ge 1 - \exp^{\frac{-2Nt^2}{\tau^2 d_X^2}}
$$

$$
P\left(\frac{1}{N} \sum_{i \in [N]} F(\widehat{\mathbf{z}}^{**}, \mathbf{x}_i^*) \le t + \mathbb{E}_{\mathbf{x} \sim P^*}[F(\widehat{\mathbf{z}}^{**}, \mathbf{x})] \right) \ge 1 - \exp^{\frac{-2Nt^2}{\tau^2 d_X^2}}
$$

Putting $\exp^{\frac{-2Nt^2}{\tau^2 d_X^2}}$ as δ_1, we get t of the form $C/\sqrt{N}$. Put all these together with a union bound yields, with probability $1 - 2\delta - 2\delta_1$:

$$
\mathbb{E}_{\mathbf{x} \sim P^*}[F(\mathbf{z}^{**}, \mathbf{x})] - C/\sqrt{N} - (1 + 2\sqrt{\xi})\tau\epsilon - \epsilon_N - \epsilon_{N_T} \le \mathbb{E}_{\mathbf{x} \sim P^*}[F(\widehat{\mathbf{z}}^{**}, \mathbf{x})]
$$

□

C Proof of Theorem 2

Theorem. *For $F(\boldsymbol{z}, \widehat{\boldsymbol{x}}_i) = \frac{\sum_j n(z_j, \widehat{\boldsymbol{x}}_i)}{\sum_j d(z_j, \widehat{\boldsymbol{x}}_i)}$ as stated above and approximated as $\frac{\boldsymbol{a}_i^T \boldsymbol{v} + b_i}{\boldsymbol{a}_i'^T \boldsymbol{v} + b_i'}$, an approximation via discretization of z_j with K pieces yields $|\mathcal{G}(\boldsymbol{z}^*) - \mathcal{G}(\widehat{\boldsymbol{z}}^{**})| \le O(\max\{C^n, C^d\}/K)$, where $\mathcal{G}(\boldsymbol{z}^*)$ and $\mathcal{G}(\widehat{\boldsymbol{z}}^{**})$ are the optimal objective values with approximation (MISOCP) and without the approximation respectively.*

Proof. The proof essentially follows by combining the results of the two lemmas below. We first prove the following two lemmas.

Lemma. *If*

$$\left|F(\boldsymbol{z}, \widehat{\boldsymbol{x}}_i) - \frac{\boldsymbol{a}_i^T \boldsymbol{v} + b_i}{\boldsymbol{a}_i'^T \boldsymbol{v} + b_i'}\right| \leq \epsilon_i$$

for some ϵ_i that is independent of $\widehat{\boldsymbol{z}}$, then $|L^* - \widehat{L}^*| \leq \max_i\{\epsilon_i\}$, where L^* and $\widehat{L}^*$ are the optimal objective values with and without the approximation.

Proof. After the transformation, the decision variable $\mathbf{z}$ changes from a continuous domain to $\mathbf{v}$ in a discrete domain. Thus the original function $F_i(\mathbf{z}) = F(\mathbf{z}, \widehat{\mathbf{x}}_i) : \mathbf{z} \longrightarrow \mathbb{R}$ and the approximate function $\widehat{F}_i(\mathbf{v}) = \frac{\mathbf{a}_i^T \mathbf{v} + b_i}{\mathbf{a}_i'^T \mathbf{v} + b_i'} : \mathbf{v} \longrightarrow \mathbb{R}$. For ease of notation, given any $\mathbf{z}$, let $\mathbf{v} = \mathcal{T}(\mathbf{z})$ be the binary transformation of the continuous variables $\mathbf{z}$, and $\mathbf{z} = \widetilde{\mathcal{T}}(\mathbf{v})$ be the backward transformation from the binary variables $\mathbf{v}$ to $\mathbf{z}$. From our assumption, we have $|F_i(\mathbf{z}) - \widehat{F}_i(\mathcal{T}(\mathbf{z}))| \leq \epsilon_i$ for any $\mathbf{z}$

Let us define (**OPT**) as the original optimization problem with continuous decision variable $\mathbf{z}$ and (**Approx-OPT**) as the approximated problem with binary variable $\mathbf{v}$. Let $q^*, \mathbf{l}^*, \mathbf{z}^*$ be an optimal solution to (**OPT**) and $q^{**}, \mathbf{l}^{**}, \mathbf{v}^{**}$ be an optimal solution to (**Approx-OPT**). Denote $\epsilon = \max_i\{\epsilon_i\}$, then we have $|F_i(\mathbf{z}) - \widehat{F}_i(\mathcal{T}(\mathbf{z}))| \leq \epsilon,\ \forall i \in [N]$, for any $\mathbf{z}$, which leads to (i) $F_i(\mathbf{z}) \leq \widehat{F}_i(\mathcal{T}(\mathbf{z})) + \epsilon$ and (ii) $\widehat{F}_i(\mathcal{T}(\mathbf{z})) \leq F_i(\mathbf{z}) + \epsilon,\ \forall i \in [N]$. We also have (iii) $F_i(\widetilde{\mathcal{T}}(\mathbf{v})) \leq \widehat{F}_i(\mathbf{v}) + \epsilon$ and (iv) $\widehat{F}_i(\mathbf{v}) \leq F_i(\widetilde{\mathcal{T}}(\mathbf{v})) + \epsilon,\ \forall i \in [N]$. We consider the following two cases: $L^* \geq \widehat{L}^*$ or $L^* \leq \widehat{L}^*$ as follows

- If $L^* \geq \widehat{L}^*$, we first see that
$$q^* - l_i^* - F_i(\mathbf{z}^*) = 0; \quad \forall i \in [N]$$
From Inequalities (i) and (ii) above, we will have
$$(q^* - \epsilon) - l_i^* - \widehat{F}_i(\mathcal{T}(\mathbf{z}^*)) \leq 0 \leq (q^* + \epsilon) - l_i^* - \widehat{F}_i(\mathcal{T}(\mathbf{z}^*)).$$
Thus, there exists $\delta \in [-\epsilon, \epsilon]$ such that $(q^* + \delta) - l_i^* - \widehat{F}_i(\mathcal{T}(\mathbf{z}^*)) = 0$, implying that $q^* + \delta, \mathbf{l}^*, \mathcal{T}(\mathbf{z}^*)$ is feasible to (**Approx-OPT**), leading to $\widehat{L}^* \geq q^* + \delta - \sqrt{\rho \sum_i (l^*)_i^2}$. Thus,
$$\begin{aligned} |L^* - \widehat{L}^*| &\leq \left| L^* - \left(q^* + \delta - \sqrt{\rho \sum_i (l^*)_i^2} \right) \right| \\ &= |\delta| \leq \epsilon. \end{aligned} \tag{22}$$

- If $L^* < \widehat{L}^*$, in analogy to the first case, we also see that
$$q^{**} - l_i^{**} - F_i(\mathbf{v}^{**}) = 0; \quad \forall i \in [N].$$
From the above inequalities (iii) and (iv), it can also be seen that there is $\delta \in [-\epsilon, \epsilon]$ such that $(q^{**} + \delta) - l_i^{**} - \widehat{F}_i(\widetilde{\mathcal{T}}(\mathbf{v}^{**})) = 0$, implying that $q^{**} + \delta, \mathbf{l}^{**}, \widetilde{\mathcal{T}}(\mathbf{v}^{**})$ is feasible to (**OPT**), leading to $L^* \geq q^{**} + \delta - \sqrt{\rho \sum_i (l^{**})_i^2}$. We then have the following inequalities
$$\begin{aligned} |\widehat{L}^* - L^*| &\leq \left| \widehat{L}^* - \left(q^{**} + \delta - \sqrt{\rho \sum_i (l^{**})_i^2} \right) \right| \\ &= |\delta| \leq \epsilon. \end{aligned} \tag{23}$$

Putting the two cases together, we have $|L^* - \widehat{L}^*| \leq \epsilon$, as desired.

□

Lemma. *For $F(\boldsymbol{z}, \widehat{\boldsymbol{x}}_i) = \frac{\sum_j n(z_j, \widehat{x}_i)}{\sum_j d(z_j, \widehat{x}_i)}$, an approximation via discretization with K pieces yields*

$$\left|F(\boldsymbol{z}, \widehat{\boldsymbol{x}}_i) - \frac{\boldsymbol{a}_i^T \boldsymbol{v} + b_i}{\boldsymbol{a}_i'^T \boldsymbol{v} + b_i'}\right| \leq \frac{C \max\{C^n, C^d\}}{K}$$

with constant C independent of $\boldsymbol{z}$.

Proof. Let C^n, C^d be the Lipschitz constant of $n(z_j, \widehat{\mathbf{x}}_i)$ and $d(z_j, \widehat{\mathbf{x}}_i)$, respectively. We use the Lipschitz continuity of these functions to get the following

$$|n(z_j, \widehat{\mathbf{x}}_i) - \widehat{n}(z_j, \widehat{\mathbf{x}}_i)| \leq \frac{U_j - L_j}{K} C^n$$
$$|d(z_j, \widehat{\mathbf{x}}_i) - \widehat{d}(z_j, \widehat{\mathbf{x}}_i)| \leq \frac{U_j - L_j}{K} C^d$$

Then, by the above we have

$$\left|\sum_j n(z_j, \widehat{\mathbf{x}}_i) - \sum_j \widehat{n}(z_j, \widehat{\mathbf{x}}_i)\right| \leq \frac{\sum_j U_j - \sum_j L_j}{K} C^n \overset{\text{def}}{=} \epsilon^n$$
$$\left|\sum_j d(z_j, \widehat{\mathbf{x}}_i) - \sum_j \widehat{d}(z_j, \widehat{\mathbf{x}}_i)\right| \leq \frac{\sum_j U_j - \sum_j L_j}{K} C^d \overset{\text{def}}{=} \epsilon^d.$$

Now, we write

$$\begin{aligned}|\widehat{F}_i(\mathbf{v}) - F_i(\mathbf{z})| &= \left|\frac{\sum_j n(z_j, \widehat{\mathbf{x}}_i)}{\sum_j d(z_j, \widehat{\mathbf{x}}_i)} - \frac{\sum_j \widehat{n}(z_j, \widehat{\mathbf{x}}_i)}{\sum_j \widehat{d}(z_j, \widehat{\mathbf{x}}_i)}\right| \\ &= \left|\frac{\sum_j n(z_j, \widehat{\mathbf{x}}_i) \sum_j \widehat{d}(z_j, \widehat{\mathbf{x}}_i) - \sum_j \widehat{n}(z_j, \widehat{\mathbf{x}}_i) \sum_j d(z_j, \widehat{\mathbf{x}}_i)}{\sum_j d(z_j, \widehat{\mathbf{x}}_i)}\right|,\end{aligned}$$

We handle the absolute value by considering the following two cases

- If $\sum_j n(z_j, \widehat{\mathbf{x}}_i) \sum_j \widehat{d}(z_j, \widehat{\mathbf{x}}_i) \geq \sum_j \widehat{n}(z_j, \widehat{\mathbf{x}}_i) \sum_j d(z_j, \widehat{\mathbf{x}}_i)$, then

$$\begin{aligned}|\widehat{F}_i(\mathbf{v}) - F_i(\mathbf{z})| &\leq \left|\frac{\sum_j n(z_j, \widehat{\mathbf{x}}_i)(\sum_j d(z_j, \widehat{\mathbf{x}}_i) + \epsilon^d) - (\sum_j n(z_j, \widehat{\mathbf{x}}_i) - \epsilon^n) \sum_j d(z_j, \widehat{\mathbf{x}}_i)}{\sum_j d(z_j, \widehat{\mathbf{x}}_i)}\right| \\ &= \left|\frac{\epsilon^d \sum_j n(z_j, \widehat{\mathbf{x}}_i) + \epsilon^n \sum_j d(z_j, \widehat{\mathbf{x}}_i)}{\sum_j d(z_j, \widehat{\mathbf{x}}_i)}\right| \\ &\leq \max\{\epsilon^n, \epsilon^d\} \max_{\mathbf{z}} \left\{\left|\frac{\sum_j n(z_j, \widehat{\mathbf{x}}_i) + \sum_j d(z_j, \widehat{\mathbf{x}}_i)}{\sum_j d(z_j, \widehat{\mathbf{x}}_i)}\right|\right\} \\ &= \frac{\sum_j U_j - \sum_j L_j}{K} \max\{C^n, C^d\} \max_{\mathbf{z}} \left\{\left|\frac{\sum_j n(z_j, \widehat{\mathbf{x}}_i) + \sum_j d(z_j, \widehat{\mathbf{x}}_i)}{\sum_j d(z_j, \widehat{\mathbf{x}}_i)}\right|\right\}.\end{aligned}$$

- If $\sum_j n(z_j, \widehat{\mathbf{x}}_i) \sum_j \widehat{d}(z_j, \widehat{\mathbf{x}}_i) \leq \sum_j \widehat{n}(z_j, \widehat{\mathbf{x}}_i) \sum_j d(z_j, \widehat{\mathbf{x}}_i)$, similarly we have

$$\begin{aligned}|\widehat{F}_i(\mathbf{v}) - F_i(\mathbf{z})| &\leq \left|\frac{(\sum_j n(z_j, \widehat{\mathbf{x}}_i) + \epsilon^n) \sum_j d(z_j, \widehat{\mathbf{x}}_i) - \sum_j n(z_j, \widehat{\mathbf{x}}_i)(\sum_j d(z_j, \widehat{\mathbf{x}}_i) - \epsilon^d)}{\sum_j d(z_j, \widehat{\mathbf{x}}_i)}\right| \\ &= \left|\frac{\epsilon^d \sum_j n(z_j, \widehat{\mathbf{x}}_i) + \epsilon^n \sum_j d(z_j, \widehat{\mathbf{x}}_i)}{\sum_j d(z_j, \widehat{\mathbf{x}}_i)}\right| \\ &\leq \frac{\sum_j U_j - \sum_j L_j}{K} \max\{C^n, C^d\} \max_{\mathbf{z}} \left\{\left|\frac{\sum_j n(z_j, \widehat{\mathbf{x}}_i) + \sum_j d(z_j, \widehat{\mathbf{x}}_i)}{\sum_j d(z_j, \widehat{\mathbf{x}}_i)}\right|\right\}.\end{aligned}$$

Therefore, if we let

$$C = \left(\sum_j U_j - \sum_j L_j\right) \max_{\mathbf{z}} \left\{\left|\frac{\sum_j n(z_j, \widehat{\mathbf{x}}_i) + \sum_j d(z_j, \widehat{\mathbf{x}}_i)}{\sum_j d(z_j, \widehat{\mathbf{x}}_i)}\right|\right\},$$

which is independent of $\mathbf{z}$, then we obtain the desired inequality $|\widehat{F}_i(\mathbf{v}) - F_i(\mathbf{z})| \leq C \max\{C^n, C^d\}/K$. □

and

Taking $\mathcal{G}(\mathbf{z}^*)$ as L^*, $\mathcal{G}(\widehat{\mathbf{z}}^{**})$ as $\widehat{L}^*$, and ϵ_i as $C\max\{C^n, C^d\}/K$ we get the desired result for the theorem. $\square$

D Proof of Lemma 1

Lemma. *We have* $\left|\widehat{Mean}(F(\boldsymbol{z},\boldsymbol{x})) - \widehat{Mean}^S(F(\boldsymbol{z},\boldsymbol{x}))\right| \leq \tau\epsilon$ *and* $\left|\sqrt{\rho\widehat{Var}(F(\boldsymbol{z},\boldsymbol{x}))} - \sqrt{\rho\widehat{Var}^S(F(\boldsymbol{z},\boldsymbol{x}))}\right| \leq (\psi + \sqrt{2\tau\epsilon})\sqrt{\frac{2\tau\epsilon\xi}{N}}$.

Proof. For the first result, By Lipschitzness,

$$|F(\mathbf{z}, \widehat{\mathbf{x}}_i) - F(\mathbf{z}, \mathbf{x}^s)| \leq \tau\epsilon, \ \forall \mathbf{z}, \forall \widehat{\mathbf{x}}_i \text{ in cluster } s$$

The result follows by summing over $\widehat{\mathbf{x}}^i$ and averaging.

To get error bound for variance term, let I_s be the set of indices that belong to cluster s, thus, $\{I_s\}_{s\in[S]}$ is a partition of $[N]$ and $C_s = |I_s|$. Let use define

$$\mu = \frac{1}{N}\sum_{i\in[N]} F(\mathbf{z}, \widehat{\mathbf{x}}_i)$$

$$\widehat{\mu} = \frac{1}{N}\sum_{s} C_s F(\mathbf{z}, \widehat{\mathbf{x}}^s) \tag{24}$$

Let $\alpha_i = \mu - F(\mathbf{z}, \widehat{\mathbf{x}}_i)$ (or $F(\mathbf{z}, \widehat{\mathbf{x}}_i) = \mu + \alpha_i$), thus, $\sum_i \alpha_i = 0$. As we know from Lipschitzness assumption that

$$|F(\mathbf{z}, \widehat{\mathbf{x}}_i) - F(\mathbf{z}, \widehat{\mathbf{x}}^s)| \leq \tau\epsilon, \ \forall \mathbf{z}, i \in I_s, \tag{25}$$

we always can write $F(\mathbf{z}, \widehat{\mathbf{x}}^s)$ as $F(\mathbf{z}, \widehat{\mathbf{x}}_i) + \beta_i = \mu + \alpha_i + \beta_i$ for any $\widehat{\mathbf{x}}_i$ in cluster s, where β_i are constants chosen such that

$$-\tau\epsilon \leq \beta_i \leq \tau\epsilon, \tag{26}$$

$$\frac{1}{N}\sum_{i\in[N]} \beta_i = \widehat{\mu} - \mu. \tag{27}$$

Then, we note that

$$\sqrt{\sum_{i\in[N]}\left(\frac{1}{N}\sum_{i\in[N]} F(\mathbf{z}, \widehat{\mathbf{x}}_i) - F(\mathbf{z}, \widehat{\mathbf{x}}_i)\right)^2} = \sqrt{\sum_{i\in[N]} \alpha_i^2}.$$

Also, we have

$$\sqrt{\sum_s C_s\left(\frac{1}{N}\sum_s C_s F(\mathbf{z}, \widehat{\mathbf{x}}^s) - F(\mathbf{z}, \widehat{\mathbf{x}}^s)\right)^2} = \sqrt{\sum_{i\in[N]} (\widehat{\mu} - \mu - \alpha_i - \beta_i)^2},$$

as C_s is the number of points in cluster s and $\widehat{\mu} = \frac{1}{N}\sum_s C_s F(\mathbf{z}, \widehat{\mathbf{x}}^s)$ and $F(\mathbf{z}, \widehat{\mathbf{x}}^s) = \mu + \alpha_i + \beta_i$ for all $i \in I_S$. Now, let us assume that

$$\sqrt{\rho\sum_s C_s\left(\frac{1}{N}\sum_s C_s F(\mathbf{z}, \widehat{\mathbf{x}}^s) - F(\mathbf{z}, \widehat{\mathbf{x}}^s)\right)^2} \geq \sqrt{\rho\sum_i\left(\frac{1}{N}\sum_i F(\mathbf{z}, \widehat{\mathbf{x}}_i) - F(\mathbf{z}, \widehat{\mathbf{x}}_i)\right)^2},$$

noting that the other case

$$\sqrt{\rho\sum_s C_s\left(\frac{1}{N}\sum_s C_s F(\mathbf{z}, \widehat{\mathbf{x}}^s) - F(\mathbf{z}, \widehat{\mathbf{x}}^s)\right)^2} < \sqrt{\rho\sum_i\left(\frac{1}{N}\sum_i F(\mathbf{z}, \widehat{\mathbf{x}}_i) - F(\mathbf{z}, \widehat{\mathbf{x}}_i)\right)^2}$$

can be handled similarly by rewriting $F(\mathbf{z}, \widehat{\mathbf{x}}_i)$ as $\mu + \alpha_i + \beta_i$ and $F(\mathbf{z}, \widehat{\mathbf{x}}^s)$ as $\mu + \alpha_i$. We write

$$\begin{aligned}
\sqrt{\sum_s C_s \left(\frac{1}{N}\sum_s C_s F(\mathbf{z}, \widehat{\mathbf{x}}^s) - F(\mathbf{z}, \widehat{\mathbf{x}}^s)\right)^2} &= \sqrt{\sum_{i\in[N]} (\widehat{\mu} - \mu - \alpha_i - \beta_i)^2} \\
&= \sqrt{\sum_{i\in[N]} (\alpha_i + (\beta_i + \mu - \widehat{\mu}))^2} \\
&= \sqrt{\sum_{i\in[N]} \alpha_i^2 + 2\sum_{i\in[N]} \alpha_i(\beta_i + \mu - \widehat{\mu}) + \sum_{i\in[N]} (\beta_i + \mu - \widehat{\mu})^2}
\end{aligned}$$

Using $\sum_i \alpha_i = 0$ and $|\mu - \widehat{\mu}| \leq \tau\epsilon, |\beta_i| \leq \tau\epsilon$ we get

$$\begin{aligned}
\sqrt{\sum_s C_s \left(\frac{1}{N}\sum_s C_s F(\mathbf{z}, \widehat{\mathbf{x}}^s) - F(\mathbf{z}, \widehat{\mathbf{x}}^s)\right)^2} &= \sqrt{\sum_{i\in[N]} \alpha_i^2 + 2\sum_{i\in[N]} \alpha_i\beta_i + \sum_{i\in[N]} (\beta_i + \mu - \widehat{\mu})^2} \\
&\leq \sqrt{\sum_{i\in[N]} \alpha_i^2 + 2\sum_{i\in[N]} \alpha_i\beta_i + N(2\tau\epsilon)^2}.
\end{aligned}$$

Using F_U, F_L are upper and lower limits of F and let $\psi = \sqrt{(}F_U - F_L)$, we further expand the inequalities

$$\begin{aligned}
\sqrt{\sum_s C_s \left(\frac{1}{N}\sum_s C_s F(\mathbf{z}, \widehat{\mathbf{x}}^s) - F(\mathbf{z}, \widehat{\mathbf{x}}^s)\right)^2} &\leq \sqrt{\sum_i \alpha_i^2 + 2N(F_U - F_L)\tau\epsilon + N(2\tau\epsilon)^2} \\
&\overset{(a)}{\leq} \sqrt{\sum_i \alpha_i^2} + \sqrt{2N(F_U - F_L)\tau\epsilon + N(2\tau\epsilon)^2} \\
&\leq \sqrt{\sum_i \alpha_i^2} + (\psi + \sqrt{2\tau\epsilon})\sqrt{2N\tau\epsilon}
\end{aligned}$$

where (a) is due to the fact that $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ for any non-negative numbers a, b. Thus, after plugging in $\rho = \frac{\xi}{N^2}$ we get

$$\begin{aligned}
&\left|\sqrt{\rho\sum_{i\in[N]} \Big(\frac{1}{N}\sum_{i\in[N]} F(\mathbf{z}, \widehat{\mathbf{x}}_i) - F(\mathbf{z}, \widehat{\mathbf{x}}_i)\Big)^2} - \sqrt{\rho\sum_s C_s \left(\frac{1}{N}\sum_s C_s F(\mathbf{z}, \widehat{\mathbf{x}}^s) - F(\mathbf{z}, \widehat{\mathbf{x}}^s)\right)^2}\right| \\
&\leq (\psi + \sqrt{2\tau\epsilon})\sqrt{\rho}\sqrt{2N\tau\epsilon} \\
&= (\psi + \sqrt{2\tau\epsilon})\sqrt{\frac{2\tau\epsilon\xi}{N}},
\end{aligned}$$

as desired. □

E Proof of Theorem 3

Theorem. *Given the assumptions stated above, and $\widehat{z}$ an optimal solution for* $\max_z \widehat{\mathcal{G}}(z)$ *and z^* optimal for* $\max_z \mathcal{G}(z)$*, the following holds:*

$$|\mathcal{G}(\widehat{z}) - \mathcal{G}(z^*)| \leq 2(\tau\epsilon + \psi\sqrt{\frac{2\tau\epsilon\xi}{N}} + \frac{2\tau\epsilon\xi}{\sqrt{N}}).$$

Proof. Let $\widehat{\mathbf{z}}$ be an optimal solution to $\max_{\mathbf{z}} \widehat{\mathcal{G}}(\mathbf{z})$ and $\mathbf{z}^*$ be optimal to $\max_{\mathbf{z}} \mathcal{G}(\mathbf{z})$, we have

$$|\mathcal{G}(\widehat{\mathbf{z}}) - \mathcal{G}(\mathbf{z}^*)| \leq |\mathcal{G}(\widehat{\mathbf{z}}) - \widehat{\mathcal{G}}(\widehat{\mathbf{z}})| + |\widehat{\mathcal{G}}(\widehat{\mathbf{z}}) - \mathcal{G}(\mathbf{z}^*)|$$

The later can be further evaluated by considering two cases, $\widehat{\mathcal{G}}(\widehat{\mathbf{z}}) \geq \mathcal{G}(\mathbf{z}^*)$ and $\widehat{\mathcal{G}}(\widehat{\mathbf{z}}) < \mathcal{G}(\mathbf{z}^*)$. If $\widehat{\mathcal{G}}(\widehat{\mathbf{z}}) \geq \mathcal{G}(\mathbf{z}^*)$, then $|\widehat{\mathcal{G}}(\widehat{\mathbf{z}}) - \mathcal{G}(\mathbf{z}^*)| = \widehat{\mathcal{G}}(\widehat{\mathbf{z}}) - \mathcal{G}(\mathbf{z}^*) \leq \widehat{\mathcal{G}}(\widehat{\mathbf{z}}) - \mathcal{G}(\widehat{\mathbf{z}})$. The other case can be done similarly to have

$$\begin{aligned}|\mathcal{G}(\widehat{\mathbf{z}}) - \mathcal{G}(\mathbf{z}^*)| \leq 2|\mathcal{G}(\widehat{\mathbf{z}}) - \widehat{\mathcal{G}}(\widehat{\mathbf{z}})\Big| &\leq 2|\widehat{\text{Mean}}^S(F(\widehat{\mathbf{z}}, \mathbf{x})) - \widehat{\text{Mean}}(F(\widehat{\mathbf{z}}, \mathbf{x}))\Big| \\ &+ 2\Big|\sqrt{\rho\widehat{\text{Var}}(F(\widehat{\mathbf{z}}, \mathbf{x}))} - \sqrt{\rho\widehat{\text{Var}}^S(F(\widehat{\mathbf{z}}, \mathbf{x}))}\Big|\end{aligned}$$

Then using the two results in Lemma 1, we get the required result. □

F Proof of Lemma 2

Lemma 2. *$\forall z$ with probability $\geq 1 - 2\sum_t \exp^{\frac{-2N_t\epsilon^2}{\tau^2 d_t^2}}$, $\big|\widehat{Mean}(F(z, \boldsymbol{x})) - \widehat{Mean}^T(F(z, \boldsymbol{x}))\big| \leq \epsilon$. In other words,*

$$P\left(\left|\frac{1}{N}\sum_{j\in[M]} lF(z, \widehat{\boldsymbol{x}}^j) - \frac{1}{N}\sum_{j\in[N]}[F(z, \boldsymbol{x}^j)]\right| \leq \epsilon\right) \geq \prod_t \left(1 - 2\exp^{\frac{-2N_t\epsilon^2}{\tau^2 d_t^2}}\right)$$

Proof. We utilize the concentration of Lipchitz functions. In particular, we have $\widehat{\mathbf{x}}^1, \ldots, \widehat{\mathbf{x}}^{N_t}$ which are sampled uniformly and independently from strata t and bounded (has diameter d_t). Let U_t denote the uniform probability distribution over the C_t points in strata t. Let I_t denote a set of indexes that lie in the strata t. Then, for our function $F(\mathbf{z}, \mathbf{x})$ with Lipchitz constant τ we have :

$$P\left(\Big|\frac{1}{N_t}\sum_{j\in[N_t]} F(\mathbf{z}, \widehat{\mathbf{x}}^j) - \mathbb{E}_{\mathbf{x}\sim U_t}[F(\mathbf{z}, \mathbf{x})]\Big| \leq \epsilon\right) \geq 1 - 2\exp^{\frac{-2N_t\epsilon^2}{\tau^2 d_t^2}} \quad \forall t, \mathbf{z}.$$

Observe that by definition

$$\mathbb{E}_{\mathbf{x}\sim U_t}[F(\mathbf{z}, \mathbf{x})] = \frac{1}{C_t}\sum_{j\in I_t}[F(\mathbf{z}, \widehat{\mathbf{x}}_j)].$$

Hence,

$$\begin{aligned}&P\left(\Big|\frac{1}{N_t}\sum_{j\in[N_t]} F(\mathbf{z}, \widehat{\mathbf{x}}^j) - \frac{1}{C_t}\sum_{j\in I_t}[F(\mathbf{z}, \widehat{\mathbf{x}}_j)]\Big| \geq \epsilon\right) \leq 2\exp^{\frac{-2N_t\epsilon^2}{\tau^2 d_t^2}} \\ \Rightarrow\ &P\left(\Big|\sum_{j\in[N_t]} l_t F(\mathbf{z}, \widehat{\mathbf{x}}^j) - \sum_{j\in I_t}[F(\mathbf{z}, \mathbf{x}_j)]\Big| \geq C_t\epsilon\right) \leq 2\exp^{\frac{-2N_t\epsilon^2}{\tau^2 d_t^2}}.\end{aligned}$$

Call the event in the probability above as E_t. It is obvious that E_t is independent over all different strata t's due to the independent sampling of points across strata. Hence $\neg E_t$ are also independent. Next, using product of independent events over all strata we get

$$P\Big(\cap_t \neg E_t\Big) \geq \prod_t \Big(1 - 2\exp^{\frac{-2N_t\epsilon^2}{\tau^2 d_t^2}}\Big).$$

Note that $\cap_t \neg E_t$ implies

$$\sum_t \Big|\sum_{j\in[N_t]} l_t F(\mathbf{z}, \widehat{\mathbf{x}}^j) - \sum_{j\in I_t}[F(\mathbf{z}, \mathbf{x}_j)]\Big| \leq \sum_t C_t\epsilon.$$

Noting that $|a + b| \leq |a| + |b|$ and the fact that $\{I_t\}_{t\in[T]}$ is a partition of $[N]$, the above implies that

$$\Big|\sum_t l_t \sum_{j\in[N_t]} F(\mathbf{z}, \widehat{\mathbf{x}}^j) - \sum_{j\in[N]}[F(\mathbf{z}, \mathbf{x}^j)]\Big| \leq \sum_t C_t\epsilon$$

This gives

$$P\left(\Big|\sum_t l_t \sum_{j\in[N_t]} F(\mathbf{z},\widehat{\mathbf{x}}^j) - \sum_{j\in[N]}[F(\mathbf{z},\mathbf{x}^j)]\Big| \leq \sum_t C_t\epsilon\right) \geq \prod_t \left(1 - 2\exp^{\frac{-2N_t\epsilon^2}{\tau^2 d_t^2}}\right)$$

(Then, since $N = \sum_t C_t$)

$$\Rightarrow P\left(\Big|\frac{1}{N}\sum_t l_t \sum_{j\in[N_t]} F(\mathbf{z},\widehat{\mathbf{x}}^j) - \frac{1}{N}\sum_{j\in[N]}[F(\mathbf{z},\mathbf{x}_j)]\Big| \leq \epsilon\right) \geq \prod_t \left(1 - 2\exp^{\frac{-2N_t\epsilon^2}{\tau^2 d_t^2}}\right)$$

(Then, since $(1-a)(1-b) \geq 1 - a - b)$)

$$\Rightarrow P\left(\Big|\frac{1}{N}\sum_t l_t \sum_{j\in[N_t]} F(\mathbf{z},\widehat{\mathbf{x}}^j) - \frac{1}{N}\sum_{j\in[N]}[F(\mathbf{z},\mathbf{x}_j)]\Big| \leq \epsilon\right) \geq 1 - 2\sum_t \exp^{\frac{-2N_t\epsilon^2}{\tau^2 d_t^2}},$$

which is the desired inequality. □

G Proof of Lemma 3

Lemma 3. *Define* $D = \max_{\boldsymbol{z},\boldsymbol{x}} |F(\boldsymbol{z},\boldsymbol{x})|$ *for bounded function* F. *Then,* $\forall \boldsymbol{z}$ *with probability* $\geq 1 - 4\sum_t \exp^{\frac{-2N_t\epsilon^2}{4\tau^2 d_t^2 D^2}}$, $\left|\sqrt{\rho\widehat{Var}(F(\boldsymbol{z},\boldsymbol{x}))} - \sqrt{\rho\widehat{Var}^T(F(\boldsymbol{z},\boldsymbol{x}))}\right| \leq \frac{2\sqrt{\xi}\epsilon}{\sqrt{\widehat{Var}(F(\boldsymbol{z},\boldsymbol{x}))}}$.

Proof. Fix $\mathbf{z}$. Recall I_t be the set of index that belong to strata t, thus, $\{I_t\}_{t\in[T]}$ is a partition of $[N]$ and $C_t = |I_t|$. For sake of simplicity, we use the shorthand for the sample/random variable $Y^j = F(\mathbf{z},\widehat{\mathbf{x}}^j)$. Note that the samples are independent. We use the following notations :

$$\mu = \frac{1}{N}\sum_{i\in[N]} F(\mathbf{z},\widehat{\mathbf{x}}_i)$$

$$\widehat{\mu} = \frac{1}{N}\sum_t \sum_{j\in[N_t]} l_t Y^j$$

Note that $\sum_t \sum_{j\in[N_t]} l_t = N$.

The unnormalized weighted variance is

$$\begin{aligned}
\widehat{Var}^T &= \sum_t \sum_{j\in[N_t]} l_t\left(\widehat{\mu} - Y^j\right)^2 \\
&= \sum_t \sum_{j\in[N_t]} l_t\left(\widehat{\mu}^2 - 2\widehat{\mu}Y^j + (Y^j)^2\right) \\
&= N\widehat{\mu}^2 - 2\widehat{\mu}\sum_t \sum_{j\in[N_t]} l_t Y^j + \sum_t \sum_{j\in[N_t]} l_t (Y^j)^2 \\
&= N\widehat{\mu}^2 - 2N\widehat{\mu}^2 + \sum_t \sum_{j\in[N_t]} l_t (Y^j)^2 \\
&= \sum_t \sum_{j\in[N_t]} l_t (Y^j)^2 - N\widehat{\mu}^2
\end{aligned}$$

We wish to compare this to

$$\widehat{Var} = \sum_{j\in[N]} F(\mathbf{z},\widehat{\mathbf{x}}_j)^2 - N\mu^2$$

Towards this end, we have

$$|\widehat{Var}^T - \widehat{Var}| \leq |\sum_t \sum_{j\in[N_t]} l_t (Y^j)^2 - \sum_{j\in[N]} F(\mathbf{z},\widehat{\mathbf{x}}_j)^2| + N|\widehat{\mu}^2 - \mu^2| \tag{28}$$

We know from Lipschitzness assumption that

$$|F(\mathbf{z},\widehat{\mathbf{x}}_i) - F(\mathbf{z},\widehat{\mathbf{x}}_j)| \le \tau d_t,\ \forall \mathbf{z}, i, j \in I_s \tag{29}$$

Multiplying both sides by $|F(\mathbf{z},\widehat{\mathbf{x}}_i) + F(\mathbf{z},\widehat{\mathbf{x}}_j)|$ (which is $\le 2D$), we get

$$|F(\mathbf{z},\widehat{\mathbf{x}}_i)^2 - F(\mathbf{z},\widehat{\mathbf{x}}_j)^2| \le 2\tau d_t D,\ \forall \mathbf{z}, i, j \in I_s \tag{30}$$

Let U_t denote the uniform probability distribution over the C_t points in strata t. Observe that by definition

$$\mathbb{E}_{\mathbf{x}\sim U_t}[(Y^j)^2] = \frac{1}{C_t}\sum_{j\in I_t}[F(\mathbf{z},\widehat{\mathbf{x}}_j)^2]$$

Then, by Hoeffding inequality and Equation 30

$$P\left(\Big|\frac{1}{N_t}\sum_{j\in[N_t]}(Y^j)^2 - \mathbb{E}_{\mathbf{x}\sim U_t}[F(\mathbf{z},\mathbf{x})^2]\Big| \le \epsilon\right) \ge 1 - 2\exp^{\frac{-2N_t\epsilon^2}{4\tau^2 d_t^2 D^2}} \qquad \forall t, \mathbf{z}$$

Then, using the same sequence of steps as for Lemma 2, we get

$$P\left(\Big|\frac{1}{N}\sum_t l_t \sum_{j\in[N_t]}(Y^j)^2 - \frac{1}{N}\sum_{j\in[N]}F(\mathbf{z},\mathbf{x}_j)^2\Big| \le \epsilon\right) \ge 1 - 2\sum_t \exp^{\frac{-2N_t\epsilon^2}{4\tau^2 d_t^2 D^2}} \qquad \forall \mathbf{z} \tag{31}$$

Also, we know from Lemma 2 that

$$P\left(\Big|\widehat{\mu} - \mu\Big| \le \epsilon\right) \ge 1 - 2\sum_t \exp^{\frac{-2N_t\epsilon^2}{\tau^2 d_t^2}} \qquad \forall \mathbf{z}$$

Multiplying both sides of the term inside the probability by $|\widehat{\mu} + \mu|$ (which is $\le 2D$), we get

$$P\left(\Big|\widehat{\mu}^2 - \mu^2\Big| \le 2\epsilon D\right) \ge 1 - 2\sum_t \exp^{\frac{-2N_t\epsilon^2}{\tau^2 d_t^2}} \qquad \forall \mathbf{z}$$

Replacing $2\epsilon D$ by ϵ (slight abuse of notation)

$$P\left(\Big|\widehat{\mu}^2 - \mu^2\Big| \le \epsilon\right) \ge 1 - 2\sum_t \exp^{\frac{-2N_t\epsilon^2}{4\tau^2 d_t^2 D^2}} \qquad \forall \mathbf{z} \tag{32}$$

Denote the event in Equation 31 as A and Equation 32 as B, using union bound we get $P(\neg A \vee \neg B) \le 4\sum_t \exp^{\frac{-2N_t\epsilon^2}{4\tau^2 d_t^2 D^2}}$, or by taking negation $P(A \wedge B) \ge 1 - 4\sum_t \exp^{\frac{-2N_t\epsilon^2}{4\tau^2 d_t^2 D^2}}$. $A \wedge B$ together with Equation 28 implies that with probability $1 - 4\sum_t \exp^{\frac{-2N_t\epsilon^2}{4\tau^2 d_t^2 D^2}}$

$$|\widehat{Var}^T - \widehat{Var}| \le 2N\epsilon \tag{33}$$

Then, note that

$$\Big|\sqrt{\rho\widehat{Var}^T} - \sqrt{\rho\widehat{Var}}\Big| = \sqrt{\rho}\frac{|\widehat{Var}^T - \widehat{Var}|}{\sqrt{\widehat{Var}^T} + \sqrt{\widehat{Var}}} \le \frac{\sqrt{\xi}}{N}\frac{|\widehat{Var}^T - \widehat{Var}|}{\sqrt{\widehat{Var}}}$$

Then, using Equation 33, we get with probability $1 - 4\sum_t \exp^{\frac{-2N_t\epsilon^2}{4\tau^2 d_t^2 D^2}}$

$$\Big|\sqrt{\rho\widehat{Var}^T} - \sqrt{\rho\widehat{Var}}\Big| \le \frac{2\sqrt{\xi}\epsilon}{\sqrt{\widehat{Var}}}$$

□

H Proof of Theorem 4

We prove a more general result stated below

Theorem. *Given the assumptions stated above, and* $\widehat{z}$ *an optimal solution for* $\max_z \widehat{\mathcal{G}}(z)$ *and* z^* *optimal for* $\max_z \mathcal{G}(z)$*, the following statement holds with probability* $\geq 1 - 2\sum_t \exp^{\frac{-2N_t\epsilon^2}{\tau^2 d_t^2}} - 4\sum_t \exp^{\frac{-2N_t\epsilon^2}{4\tau^2 d_t^2 D^2}}$*:*

$$|\mathcal{G}(\widehat{z}) - \mathcal{G}(z^*)| \leq 2\epsilon\left(1 + 2\sqrt{\frac{\xi}{\widehat{Var}}}\right).$$

For $N_* = \min_t N_t$*, then the above can be written as with probability* $\geq 1 - 2\sum_t \exp^{\frac{-2\sqrt{N_*}\epsilon^2}{\tau^2 d_t^2}} - 4\sum_t \exp^{\frac{-2\sqrt{N_*}\epsilon^2}{4\tau^2 d_t^2 D^2}}$*:*

$$|\mathcal{G}(\widehat{z}) - \mathcal{G}(z^*)| \leq \frac{2\epsilon}{(N_*)^{1/4}}\left(1 + 2\sqrt{\frac{\xi}{\widehat{Var}(F(z, \boldsymbol{x}))}}\right).$$

Proof. Following style of proof of Theorem 3 using union bound with lemmas 2 and 3 we get the first claim above. For the second claim set $\epsilon' = \sqrt{N_*}^{1/4}\epsilon$ and replace ϵ by ϵ' (note that $\sqrt{N_t} \geq \sqrt{N_*}$). □

I Data Generation Details

(Synthetic) SSG: Following standard terminology and set-up in SSG, for every target j, under a type specified by parameters **x**, if the adversary attacks j and the target is protected then the defender obtains reward $r^d_{\mathbf{x},j}$ and the adversary obtains $l^a_{\mathbf{x},j}$. Conversely, if the defender is not protecting target j, then the defender obtains $l^d_{\mathbf{x},j}$ ($r^d_{\mathbf{x},j} > l^d_{\mathbf{x},j}$) and the adversary gets $r^a_{\mathbf{x},j}$ ($r^a_{\mathbf{x},j} > l^a_{\mathbf{x},j}$). Given z_j as the marginal probability of defending target j, the expected utility of the defender and attacker of type **x** for an attack on target j is formulated as follows: $u(z_j, \theta^d_{\mathbf{x}}) = z_j r^d_{\mathbf{x},j} + (1 - z_j) l^d_{\mathbf{x},j}$ and $h(z_j, \theta^a_{\mathbf{x}}) = \lambda_{\mathbf{x}}(z_j l^a_{\mathbf{x},j} + (1 - z_j) r^a_{\mathbf{x},j})$, where parameter $\lambda_{\mathbf{x}} \geq 0$ governs rationality. $\lambda_{\mathbf{x}} \to 0$ means least rational, as the adversary chooses its attack uniformly at random and $\lambda_{\mathbf{x}} \to \infty$ means fully rational (i.e., attacks a target with highest utility). We compactly rewrite $u(z_j, \theta^d_{\mathbf{x}}) = z_j a^d_{\mathbf{x},j} + l^d_{\mathbf{x},j}$ and $h(z_j, \theta^d_{\mathbf{x}}) = -z_j c^a_{\mathbf{x},j} + l^a_{\mathbf{x},j}$. We add two layers of randomness to our *parameters* $\{a^d_{\mathbf{x},j}, l^d_{\mathbf{x},j}, c^a_{\mathbf{x},j}, l^a_{\mathbf{x},j} | \forall j \in [M], \forall \mathbf{x}\}$ by (1) generating i.i.d. samples from a mean shifted beta-distribution : low + (high − low)**Beta**(α, β), and (2) then using these samples as means for the Gaussian distribution : $\mathcal{N}(., \sigma^2)$ to i.i.d. generate the final *parameters*. In our experiments we chose : low = 5, high=8, $\alpha = 3$, $\beta = 3$, $\sigma^2 = 3$.

(Synthetic) Regressor for SSG utilities: To validate Theorem 1, we first fix a linear regressor $f^* = \langle s^*_{a^d_j}, b^*_{a^d_j}, s^*_{l^d_j}, b^*_{l^d_j}, s^*_{c^a_j}, b^*_{c^a_j}, s^*_{l^a_j}, b^*_{l^a_j} | \forall j \in [M] \rangle$ and sample $\{V^{*,a^d_{\mathbf{x}}}, V^{*,l^d_{\mathbf{x}}}, V^{*,c^a_{\mathbf{x}}}, V^{*,l^a_{\mathbf{x}}} | \forall \mathbf{x} \in [N_T]\}$ to generate $\{a^{*,d_{\mathbf{x},j}}, l^{*,d_{\mathbf{x},j}}, c^{*,a_{\mathbf{x},j}}, l^{*,a_{\mathbf{x},j}} | \forall j \in [M], \forall \mathbf{x} \in [N_T]\}$ such that $a^{*,d_{\mathbf{x},j}} = s^*_{a^d_j} * V^{*,a^d_{\mathbf{x}}} + b^*_{a^d_j}$, $l^{*,d_{\mathbf{x},j}} = s^*_{l^d_j} * V^{*,l^d_{\mathbf{x}}} + b^*_{l^d_j}$, $c^{*,a_{\mathbf{x},j}} = s^*_{c^a_j} * V^{*,c^a_{\mathbf{x}}} + b^*_{c^a_j}$, $l^{*,a_{\mathbf{x},j}} = s^*_{l^a_j} * V^{*,l^a_{\mathbf{x}}} + b^*_{l^a_j}$. Now a linear regressor $\widehat{f}$ is learnt on the given dataset of N_T samples by minimizing the L-2 loss between outputs of $\widehat{f}$: $\{\widehat{a}^{d_{\mathbf{x},j}}, \widehat{l}^{d_{\mathbf{x},j}}, \widehat{c}^{a_{\mathbf{x},j}}, \widehat{l}^{a_{\mathbf{x},j}} | \forall j \in [M], \forall \mathbf{x} \in [N_T]\}$ and actual utilities : $\{a^{*,d_{\mathbf{x},j}}, l^{*,d_{\mathbf{x},j}}, c^{*,a_{\mathbf{x},j}}, l^{*,a_{\mathbf{x},j}} | \forall j \in [M], \forall \mathbf{x} \in [N_T]\}$. DRO is performed on both true and learnt utilities to get decisions and then evaluated on held out test set of true utilities.

(Semi-Synthetic) Maximum Capture Facility Cost Planning Problem (MC-FCP): The P&R Aros-Vera et al. [2013] dataset provides fixed utilities for different facility locations which is useful when considering **MC-FCP**, where the utilities of each facility is a function of the budget allocated to it and our goal is to optimally distribute a limited budget across these facilities. Given the utilities of client **x** : $V_{\mathbf{x},j} \forall j \in [M]$, we solve for parameters $\{a_{\mathbf{x},j} | j \in [M]\}$ governed by $V_{\mathbf{x},j} = a_{\mathbf{x},j} + b_{\mathbf{x}}$, where $b_{\mathbf{x}}$ is chosen as $\min_j V_{\mathbf{x},j}$, so that all $a_{\mathbf{x},j}$ are non negative, and utilities increase on allocating more budget. Once we have the parameters, we can write the utility function : $h(z_j, \theta_{\mathbf{x},j}) = a_{\mathbf{x},j} z_j + b_{\mathbf{x}}$. Intuitively $b_{\mathbf{x}}$ is the bias of the client **x** and $a_{\mathbf{x},j} \geq 0$ is the rate at which the client's utility can be raised by allocating more budget to the j^{th} facility.

Table 4: Objective values of the baselines as a % of the objective obtained by our approach across on **MC-FCP** across various settings.

	TTGA			**PGA**		
ξ	m=7	m=10	m=13	m=7	m=10	m=13
1E2	51.6	50.3	61.2	38.7	45.7	55.3
1E3	49.2	46.2	60.0	18.3	26.2	27.8
1E4	48.2	45.0	30.4	15.0	18.2	19.1

Table 5: Training time (seconds) using our MISOCP formulation across various settings.

	MC-FCP			**MC-FLP**		
ξ	m=7	m=10	m=13	m=10	m=12	m=14
ERM	62.16	182.47	128.54	11.62	28.20	11.98
1E2	271.51	267.50	80.24	11.57	33.42	30.92
1E3	80.94	297.64	900.84	11.66	33.07	33.56
1E4	263.53	558.82	820.54	32.84	33.76	42.28

The **MC-FLP** problem directly uses the utilities of client **x** : $V_{\mathbf{x},j} \forall j \in [M]$ from the P&R [Aros-Vera et al., 2013] dataset, so **MC-FLP** is based completely on real data.

J Additional Results for Real Data

Baseline Performance on Real Data: Gradient based approaches failed to attain decent performance on this dataset on **MC-FCP** as the choice probabilities F_i are near zero almost everywhere in the space of decisions C, and since the derivative of the objective w.r.t. the decision, ie. $\frac{\partial F_i}{\partial z} = F_i \times g_i(z)$, the baselines run into a vanishing gradient problem and fail to move from the initial point. This also demonstrates the advantage of an MISOCP solver which can locate good solutions despite the above issue. Nonetheless we use gradient clipping (clipped away from zero) to train our baselines on the dataset and the results are reported in Table 4.

Training time (in secs) for our approach: We present the times for convergence for our proposed method as well as the baselines in Tables 5, 6, 7. As demonstrated in Table 5, even in the worst case our algorithm takes only about 15 minutes thus reflecting its scalability.

Need for speed up and one time cost: The problem at hand scales exponentially both in memory and time, so solving on real world datasets such as the Max Capture Facility dataset of 80,000 datapoints is simply infeasible on regular computers as the program does not even load on a machine with 128GB RAM. It is known that for SSG decisions change monthly as new attack data is received (Fang et.al) and the tool runs on resource constrained computers. Similarly, facility cost optimization decisions can also change with changing profile of customers and/or change in type of services or promotions offered (revealed in newly collected data). Thus, the DRO optimization can run repeatedly at given frequencies and needs to be be efficient in practice.

K On Choosing Optimal Number of Pieces

We proved in Appendix C that approximation via discretization guarantees improve with increasing K. To choose a suitable K for our experiments, we varied the number of pieces (K) from 2 to 20 in steps of 2, and report the relevant statistics in Figure 3. We note that across various settings, the results have saturated by $K = 10$, and thus use $K = 10$ for all our experiments.

L Converting Weighted Objective to MISOCP

Let

$$\widehat{\mu} = \frac{1}{N} \sum_{t} \sum_{j \in [N_t]} l_t F(\mathbf{z}, \mathbf{x}^j).$$

Table 6: Training time (seconds) using PGA formulation across various settings.

	MC-FCP			MC-FLP		
ξ	m=7	m=10	m=13	m=10	m=12	m=14
ERM	82.16	142.47	228.54	71.62	158.27	211.48
1E2	91.24	147.56	280.44	81.37	143.44	230.92
1E3	90.11	197.63	250.54	73.16	153.17	233.56
1E4	83.45	178.12	320.56	82.84	167.66	242.28

Table 7: Training time (seconds) using TTGD formulation across various settings.

	MC-FCP			MC-FLP		
ξ	m=7	m=10	m=13	m=10	m=12	m=14
ERM	44.17	50.31	62.13	40.11	45.64	60.63
1E2	45.12	52.12	60.01	42.34	43.11	63.18
1E3	50.11	43.17	64.32	45.77	46.23	63.66
1E4	43.43	55.82	62.54	42.14	47.76	60.28

Further, let $Y^j = F(\mathbf{z}, \widehat{\mathbf{x}}^j)$. Consider the stratified sampling objective

$$\widehat{\mu} - \sqrt{\rho \sum_t l_t \sum_{j\in[N_t]} (\widehat{\mu} - Y^j)^2} \tag{34}$$

It is enough to show the conversion for the above as the clustering is a special case with $N_t = 1$ for all t. As before we substitute $l_{t,j} = \frac{1}{N}\sum_t \sum_{j\in[N_t]} l_t Y^j - Y^j$ (notation l is abused, but the constant l_t subscript is t and the variable subscript is t, j) for all $i \in [N]$ and $q = \frac{1}{N}\sum_t \sum_{j\in[N_t]} l_t Y^j$. Note that $\sum_{j\in N_t} l_{t,j} = \frac{N_t}{N}\sum_t l_t \sum_{j\in[N_t]} Y^j - \sum_{j\in[N_t]} Y^j$, and since $l_t = \frac{C_t}{N_t}$, we have $\sum_{j\in N_t} l_{t,j} = \frac{1}{N}\sum_t C_t \sum_{j\in[N_t]} Y^j - \sum_{j\in[N_t]} Y^j$. Also, since $\sum_t C_t = N$ then

$$\sum_t C_t \sum_{j\in[N_t]} l_{t,j} = \frac{\sum_t C_t}{N} \sum_t C_t \sum_{j\in[N_t]} Y^j - \sum_t C_t \sum_{j\in[N_t]} Y^j = 0$$

Also, $Y^j = F(\mathbf{z}, \widehat{\mathbf{x}}^j) = q - l_{t,j}$. The objective becomes $q - \sqrt{\rho \sum_t l_t \sum_{j\in[N_t]} l_{t,j}^2}$. Thus, like the original (non-clustered) problem the objective is concave, and the only non-convexity is in the constraint $F(\mathbf{z}, \widehat{\mathbf{x}}^j) = q - l_{t,j}$, which can be approximated as earlier.

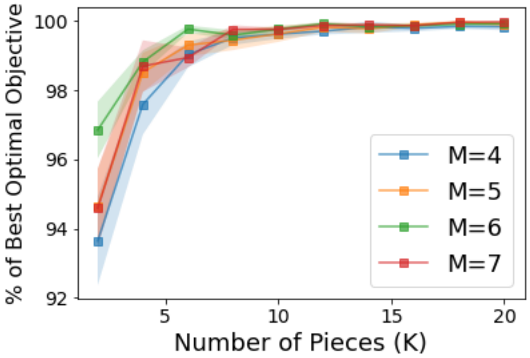


Figure 3: Optimal objective value achieved by varying number of pieces as a % of the **Best OPT** - achieved at K=20. Results are shown for varying no. of alternatives M and averaged over 10 generated **SSG** datasets with underlying parameters are $N = 500, m = 1, \xi =$1E6.

For MISOCP, we move the part of the objective becomes the linear function $q - s$ with an additional constraint that

$$\sqrt{\rho \sum_t l_t \sum_{j \in [N_t]} l_{t,j}^2} \leq s \tag{35}$$

(Recall **r** is the vector of all variables). The above is same as $||A\mathbf{r}||_2 \leq \mathbf{c}^T\mathbf{r}$ for the constant matrix A (with entries 0 or $\sqrt{\rho l_t}$ at appropriate entries) and constant vector **c** (with 1 in the s component, rest 0's) that picks the l_i's and s respectively.

M Illustrative Examples for SSG and Facility Location

We first describe the quantal response or multinomial logit model that has been used in SSG and many other applications. Briefly, given K choices with utility u_k for choice k, the quantal response model states that human choose choice j with probability $\propto \exp(\lambda u_k)$, where λ is a rationality parameter. $\lambda = 0$ means the choice is uniformly random and $\lambda = \infty$ means the highest utility choice is chosen.

SSG: Next, consider a small example SSG where the attacker type denotes its rationality. Let there be three targets to be protected, only one defender resource, and attacker types **x** is a scalar given by a real number in $[0, 10]$ (intuitively higher number type is more rational as explained next). Following typical SSG style, each target has a reward or penalty for defender and adversary when that target is attacked and is defended or undefended respectively. The defender has $r^d_{\mathbf{x},1} = 0.5, r^d_{\mathbf{x},2} = 1, r^d_{\mathbf{x},3} = 1.5$ and $l^d_{\mathbf{x},1} = -0.5, l^d_{\mathbf{x},2} = -1, l^d_{\mathbf{x},2} = -1.5$. Similarly, the attacker has $r^d_{\mathbf{x},1} = 1, r^d_{\mathbf{x},2} = 2, r^d_{\mathbf{x},3} = 3$ and $l^d_{\mathbf{x},1} = -1, l^d_{\mathbf{x},2} = -2, l^d_{\mathbf{x},2} = -3$ (for simplicity, these have been chosen independent of **x**). Given **z** (vector of probability of defending each target), the expected utility of the defender for an attack on target j by an attacker of type **x** is formulated as follows: $u(z_j, \theta^d_{\mathbf{x}}) = z_j r^d_{\mathbf{x},j} + (1 - z_j) l^d_{\mathbf{x},j}$, similarly for the attacker of type **x** its expected utility is $u^a(z_j, \theta^a_{\mathbf{x}}) = z_j l^a_{\mathbf{x},j} + (1 - z_j) r^a_{\mathbf{x},j}$. Consider a quantal responding adversary according to Yang et al. [2014] who attacks a target according to probability proportional to e to the power a λ-scaled version of the utility. Hence $h(z_j, \theta^a_{\mathbf{x}}) = \lambda_{\mathbf{x}} u^a(z_j, \theta^a_{\mathbf{x}})$, where $\lambda_{\mathbf{x}} = \mathbf{x}$ is the rationality parameter and it shows more rationality for higher type (recall h notation from main paper). Then, the adversary of type **x** chooses target j with probability $\propto \exp(h(z_j, \theta^a_{\mathbf{x}}))$. The distribution over types is not known but multiple attacks by the same type of adversary can be used to infer that type of attacker's $\lambda_{\mathbf{x}}$ using maximum likelihood techniques from Yang et al. [2014]. This gives us the N observations of the types of attackers: $\widehat{x}_1, \widehat{x}_2, \ldots, \widehat{x}_N$. The DRO formulation from this point on follows the same style as shown in Equation (SSG). Note that this can be made general by considering a vector $\lambda_{\mathbf{x}}$, but we stick with the simpler model as the quantal response and the Bayesian version we describe aligns with the well-known discrete choice models.

Facility location: Consider a small example facility location problem where there are 5 locations on a *straight line* possible to set-up 2 facilities. The competitors already runs two facilities at location 3 and 5. There are types of clients given by $[0, 5]$ which roughly indicates their position on the straight line. The number of clients of type **x** is $s_{\mathbf{x}} = \lfloor 100\mathbf{x} \rfloor$. The type **x** client has utility $V_{\mathbf{x},j} = \exp(-|x - j|)$ of visiting location j, i.e., clients have more utility from visiting location nearer to their position **x**. The user will have four total locations after the new facilities open (including two from competitor). Then, using binary variable z_j to denotes if location j is chosen for a facility, the rest of the problem is set-up as described in the main paper.

In the cost version of the above problem, the variable z_j is continuous between $[0, 1]$. The utility of an user for a facility run by our firm can then be given as $h(z_j, \mathbf{x}) = z_j \exp(-|x - j|)$. Here every location has a facility (by this firm) but a very low z_j can be treated as no facility. Then, again the user has 7 facilities to choose from where the investment z_j of the opponent for its facility at location 3 and 5 is known and fixed. The rest of the problem is set-up as described in the main paper.