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

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

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

where γ_{jk}^{ni} and γ_{jk}^{di} 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

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

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

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

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

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

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

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

$$\sum_i p_i^{**} F(\mathbf{z}^{**}, \mathbf{x}_i^*) \leq \sum_i \widehat{p}_i^{**} F(\mathbf{z}^{**}, \mathbf{x}_i^*) \tag{15}$$

$$= \Big(\sum_i \widehat{p}_i^{**} F(\mathbf{z}^{**}, \mathbf{x}_i^*) - \sum_i \widehat{p}_i^{**} F(\mathbf{z}^{**}, \widehat{\mathbf{x}}_i^*) \Big) + \sum_i \widehat{p}_i^{**} F(\mathbf{z}^{**}, \widehat{\mathbf{x}}_i^*) \tag{16}$$

$$\leq \psi + \sum_i \widehat{p}_i^{**} F(\mathbf{z}^{**}, \widehat{\mathbf{x}}_i^*) \tag{17}$$

$$\leq \psi + \sum_i \widehat{p}_i^{**} F(\widehat{\mathbf{z}}^{**}, \widehat{\mathbf{x}}_i^*) \tag{18}$$

$$\leq \psi + \sum_i p_i^{**} F(\widehat{\mathbf{z}}^{**}, \widehat{\mathbf{x}}_i^*) \tag{19}$$

$$= \psi + \Big(\sum_i p_i^{**} F(\widehat{\mathbf{z}}^{**}, \widehat{\mathbf{x}}_i^*) - \sum_i p_i^{**} F(\widehat{\mathbf{z}}^{**}, \mathbf{x}_i^*) \Big) + \sum_i p_i^{**} F(\widehat{\mathbf{z}}^{**}, \mathbf{x}_i^*) \tag{20}$$

$$\leq 2\psi + \sum_i p_i^{**} F(\widehat{\mathbf{z}}^{**}, \mathbf{x}_i^*) \tag{21}$$

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

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

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

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

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

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

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

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 $1 - 2\delta - 2\delta_1$:

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

□

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 $|\mathcal{G}(\boldsymbol{z}^*) - \mathcal{G}(\widehat{\boldsymbol{z}}^{**})| \leq O(\max\{C^n, C^d\}/K)$, where $\mathcal{G}(\boldsymbol{z}^*)$ and $\mathcal{G}(\widehat{\boldsymbol{z}}^{**})$ are the optimal objective values with approximation (MISOCP) and without the approximation respectively.*

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 $(q^{**} + \delta) - l_i^{**} - \widehat{F}_i(\widehat{\mathcal{T}}(\mathbf{v}^{**})) = 0$, implying that $q^{**} + \delta, \mathbf{l}^{**}, \widehat{\mathcal{T}}(\mathbf{v}^{**})$ 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 $F(\boldsymbol{z},\widehat{\boldsymbol{x}}_i) = \frac{\sum_j n(z_j,\widehat{\boldsymbol{x}}_i)}{\sum_j d(z_j,\widehat{\boldsymbol{x}}_i)}$, 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

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

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. □

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 $|\mu - \widehat{\mu}| \le \tau\epsilon, |\beta_i| \le \tau\epsilon$ we get

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

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

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

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

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

as desired. □

E Proof of Theorem 3

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

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

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

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

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

F Proof of Lemma 2

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

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

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

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

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

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

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

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

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, $\forall \mathbf{z}$ with probability $\ge 1-4\sum_t \exp^{\frac{-2N_t\epsilon^2}{4\tau^2 d_t^2 D^2}}$, $\left|\sqrt{\rho\widehat{Var}(F(\boldsymbol{z},\boldsymbol{x}))}-\sqrt{\rho\widehat{Var}^T(F(\boldsymbol{z},\boldsymbol{x}))}\right| \le \frac{2\sqrt{\xi}\epsilon}{\sqrt{\widehat{Var}(F(\boldsymbol{z},\boldsymbol{x}))}}$.*

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

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

We know from Lipschitzness assumption that

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

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

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

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

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

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

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

Also, we know from Lemma 2 that

$$P\left(\left| \widehat{\mu} - \mu \right| \leq \epsilon \right) \geq 1 - 2 \sum_t \exp^{\frac{-2N_t \epsilon^2}{\tau^2 d_t^2}} \quad \forall \mathbf{z}$$

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

$$P\left(\left| \widehat{\mu}^2 - \mu^2 \right| \leq 2\epsilon D \right) \geq 1 - 2 \sum_t \exp^{\frac{-2N_t \epsilon^2}{\tau^2 d_t^2}} \quad \forall \mathbf{z}$$

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

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

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

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

Then, note that

$$\left| \sqrt{\rho \widehat{Var}^T} - \sqrt{\rho \widehat{Var}} \right| = \sqrt{\rho} \frac{|\widehat{Var}^T - \widehat{Var}|}{\sqrt{\widehat{Var}^T} + \sqrt{\widehat{Var}}} \leq \frac{\sqrt{\xi}}{N} \frac{|\widehat{Var}^T - \widehat{Var}|}{\sqrt{\widehat{Var}}}$$

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

$$\left| \sqrt{\rho \widehat{Var}^T} - \sqrt{\rho \widehat{Var}} \right| \leq \frac{2\sqrt{\xi}\epsilon}{\sqrt{\widehat{Var}}}$$

□

H Proof of Theorem 4

We prove a more general result stated below

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

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

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

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

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 $\mathbf{x}$, 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 $\mathbf{x}$ 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 $\mathbf{x}$: $V_{\mathbf{x},j} \forall j \in [M]$, 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 $\mathbf{x}$ 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 $\mathbf{x}$: $V_{\mathbf{x},j} \forall j \in [M]$ 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.

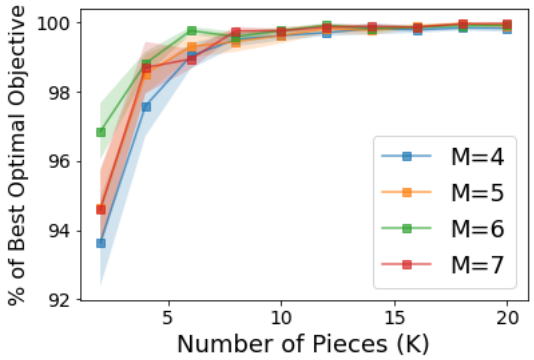

Figure 3: Optimal objective value achieved by varying number of pieces as a % of the **Best OPT** - achieved at K=20. Results are shown for varying no. of alternatives M and averaged over 10 generated **SSG** datasets with underlying parameters are $N = 500, m = 1, \xi =$1E6.

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 $\mathbf{r}$ 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 $\mathbf{c}$ (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 $\mathbf{x}$ 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 $\mathbf{x}$). Given $\mathbf{z}$ (vector of probability of defending each target), the expected utility of the defender for an attack on target j by an attacker of type $\mathbf{x}$ 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 $\mathbf{x}$ 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 $\mathbf{x}$ 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 $\mathbf{x}$ is $s_{\mathbf{x}} = \lfloor 100\mathbf{x} \rfloor$. The type $\mathbf{x}$ 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 $\mathbf{x}$. 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.