

Allocation Requires Prediction Only if Inequality Is Low

Ali Shirali¹, Rediet Abebe^{*2}, and Moritz Hardt^{*3}

¹University of California, Berkeley

²Harvard Society of Fellows

³Max Planck Institute for Intelligent Systems, Tübingen and Tübingen AI Center

Abstract

Algorithmic predictions are emerging as a promising solution concept for efficiently allocating societal resources. Fueling their use is an underlying assumption that such systems are necessary to identify individuals for interventions. We propose a principled framework for assessing this assumption: Using a simple mathematical model, we evaluate the efficacy of prediction-based allocations in settings where individuals belong to larger units such as hospitals, neighborhoods, or schools. We find that prediction-based allocations outperform baseline methods using aggregate unit-level statistics only when between-unit inequality is low and the intervention budget is high. Our results hold for a wide range of settings for the price of prediction, treatment effect heterogeneity, and unit-level statistics’ learnability. Combined, we highlight the potential limits to improving the efficacy of interventions through prediction.

1 Introduction

Predictive systems are emerging as a promising solution concept for allocating societal resources. Often, such systems output individual risk scores for undesirable outcomes like eviction, poor health, and school dropout to assist decision-makers in targeting individuals with greater precision. In doing so, an underlying assumption is that individual predictions are necessary for efficient identification. See, e.g., (Bruce et al., 2011; Cuccaro-Alamin et al., 2017; Kleinberg et al., 2018).

By contrast, decision-makers can use baseline allocation schemes by proportioning their resources using aggregate welfare information such as neighborhood eviction, hospital readmission, or school dropout rates. These larger units (i.e., neighborhoods, hospitals, and schools) can then target individuals within the units crudely. While such baseline methods can be cheaper, easier to implement, and less contested, they are also feared to be wasteful compared to prediction-based allocations.¹

We propose a principled framework for assessing this fundamental assumption and evaluating the efficacy of individual prediction-based systems. We present a simple mathematical model to compare prediction-based versus unit-based allocations. The former method relies on individuals’ predicted welfare, while the latter uses aggregate unit-level statistics and coarse information within the units. We show that prediction leads to superior allocations only when between-unit inequality is low, and the allocation budget is high. (See Figure 1 for a high-level view of inequality and budget regimes covered in our results.) Our analyses cover a broad range of settings for the price of prediction, treatment effect heterogeneity, and unit-level statistics’ learnability.

*Equal contribution

¹See the discussion in (Johnson and Zhang, 2022; Moon and Guha, 2024).

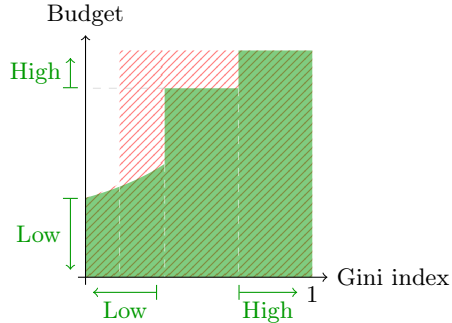


Figure 1: Sufficient conditions for a dominant unit-level allocation (green) and nondominated unit-level allocation (red). We define a high budget in relation to the cost of treating the whole population and a low budget in relation to the cost of prediction. See Figure 5 for a quantitative version of this figure.

1.1 Our Results

At a high level, we contrast two allocation mechanisms: Individual-level allocation (ILA) predicts welfare, which falls in $[0, 1]$, and treats individuals in ascending order of their predicted welfare. Unit-level allocation (ULA), roughly speaking, sorts units according to their average welfare and intervenes in increasing order of average welfare. We assume administrators within each unit target the intervention to a fraction of the individuals. They are able to avoid giving resources to the top $q \geq 0$ fraction, allowing for some errors. To start, we consider an intervention with a fixed treatment effect δ . The welfare is capped at 1 so individuals of welfare above $1 - \delta$ see diminished effects.

Individual-Level versus Unit-Level Allocations. We characterize each unit k with a parameter ρ_k that expresses the fraction of high-welfare individuals in the unit. We measure inequality in terms of the Gini index of the ρ_k -parameters. Our main results establish the fundamental role of inequality in understanding the relative effectiveness of the mechanisms. We prove that high inequality ensures the relative efficiency of ULA as long as the budget is not excessively large. This is true even when ILA uses perfect predictions (after paying some upfront cost).

What defines the high inequality regime depends on the fraction q and the average of ρ_k s, denoted by $\bar{\rho}$. The quantity $\bar{\rho}$ is also the overall proportion of individuals with high welfare (above $1 - \delta$). Roughly speaking, a Gini index below $1 - q/\bar{\rho}$ is low, and a Gini index above $1 - q/(4\bar{\rho})$ is high. When inequality is neither high nor low, ULA will still be efficient for a budget up to a constant ratio of the cost of treating everyone. In all these cases, unless the cost of prediction is minuscule, ULA *dominates* ILA (Figure 1).

Necessary conditions for ILA to dominate ULA turn out to be significantly stronger. Put differently, under mild conditions we can show that ILA does *not* dominate ULA. For example, in the natural regime of a Gini index around $1/2$, ULA remains nondominated even if $q = 0$, that is, within-unit allocation treats everyone.

Extension to Heterogeneous Treatment Effects. In Section 6, we extend our results to the case of heterogeneous treatment effects. We continue to assume that ILA uses welfare prediction as the targeting mechanism. Without further assumptions, the heterogeneity gives ILA an edge since distinguishing between individuals with close welfare through prediction can potentially yield significant gains. We limit this advantage by considering a Lipschitz continuous effect curve. Additionally, we make our analysis more realistic by restricting the concentration of welfare values. Both assumptions help avoid contrived worst-case scenarios.

Without heterogeneity, there exists a sufficient level of prediction accuracy such that more accurate predictors cannot generate more value. However, with heterogeneous treatment effects, even small changes in accuracy can be translated into value. Therefore, in the analysis of a heterogeneous treatment effect, we

incorporate prediction accuracy into our framework by assuming a model of independently noisy predictions. Under these assumptions, we obtain similar results as in the homogeneous case.

Extension to Learning Unit-Level Statistics. ULA hinges on the availability of some unit-level aggregate statistics, such as ρ_k s, to prioritize different units. We generally assume the cost of obtaining these statistics is negligible. In Section 7, we revisit this assumption and place our results in a learning-theoretic context. Although some information about individuals may be available for free, their welfare may be costly to measure. Therefore, we explore how to use readily available data to learn unit-level statistics sufficient for implementing ULA. First, we demonstrate that if an individual-level predictor can be obtained at a reasonable cost, a predictor for unit-level statistics can be obtained at a much lower cost. Next, we consider different noisy observation welfare models and discuss efficient learnability. In particular, in a general nonparametric setting, the relative cost of learning becomes negligible if the number of treated units is sufficiently large.

Summary. Our results suggest that allocation by prediction may not compare favorably to basic allocation schemes that use aggregate unit statistics in cases where inequality exists, or resources are limited. Our evaluation framework and analyses surface inequality as a fundamental mechanism linking prediction and intervention. We hope that our theoretical framework provides the tools necessary to investigate prediction as a solution concept for allocation.

1.2 Related Work

ILA, as a form of resource allocation based on prediction, is widely adopted across various domains (Kube et al., 2023a,b; Toros and Flaming, 2018; Mashiat et al., 2024; Chan et al., 2012; Mukhopadhyay and Vorobeychik, 2017; Faria et al., 2017). However, there is still minimal evidence of its efficacy (Mac Iver et al., 2019; Moon and Guha, 2024). For instance, Perdomo et al. (2023) recently demonstrated that early warning systems, when utilized to target interventions at individual students by predicting the risk of dropout, have shown little improvement despite the use of accurate predictors. Perdomo et al. (2023) and Hardt and Kim (2023) also point out how risk scores draw on unit-level information. Closely related, Perdomo (2023) asks how improvements in welfare arising from better predictions compare to those of other policy levers, such as expanding access to resources. In line with our results, it may not help investing in greater accuracy when resources are limited.

Refer to Appendix A for a complete review of related work.

2 Our Model

There are M units with N individuals each. Individual i has a welfare of $w_i \in [0, 1]$. Our goal is to improve social welfare, defined as the sum of individual welfares. To do so, we can intervene on any individual at a unit cost c subject to a total budget constraint B . We may target individuals by predicting their current welfare and allocating resources to those with lower welfare. We call this mechanism *individual-level allocation* (ILA). Alternatively, we may target units instead of individuals and delegate within-unit allocation to unit administrators, through what we call *unit-level allocation* (ULA).

The Cost of Prediction. We assume access to a welfare prediction \hat{w}_i for each individual i that satisfies the uniform error bound: $|\hat{w}_i - w_i| \leq \epsilon$. We have to pay a total price $p(\epsilon)$ to obtain such a predictor. The price p is a decreasing function of ϵ . Our results about the limitations of ILA hold even if we give ILA the advantage that after paying some fixed price, predictions are perfect. In other words, there is a fixed price of prediction $p = p(\epsilon)$ for any $\epsilon \in [0, 1]$.

Intervention Effect. Let \tilde{w}_i denote the counterfactual welfare of individual i had she received treatment. Then $\tau_i = \tilde{w}_i - w_i$ denotes the treatment effect. In our simplest model, we assume $\tilde{w}_i = \min\{w_i + \delta, 1\}$.

This is the case of an intervention that has a fixed treatment effect δ capped at the maximum welfare. Our results extend to heterogeneous treatment effects.

Individual-Level Allocation (ILA). The *individual-level allocation* gives the resources to the individuals with the lowest estimated welfare \hat{w}_i until the budget constraint is met. Formally, sort individuals ascendingly in terms of \hat{w}_i with any tie-breaking and let $s(i)$ be the individual in the i^{th} place. Given a budget of B and an ϵ -error welfare predictor, we can treat $I = \lfloor \frac{B-p(\epsilon)}{c} \rfloor$ individuals. Treating the first I individuals after sorting, the *value* of ILA is

$$V_{\text{ind}} = \sum_{i \in [I]} \tau_{s(i)}. \quad (1)$$

Unit-Level Allocation (ULA). The *unit-level allocation* targets those units that are performing worst in terms of some aggregate measurement of their welfare. More precisely, we define a statistic ρ_k for unit k that captures the well-being of individuals in unit k and allocate the resources to the units with the smallest ρ_k -values. Let \mathcal{U}_k denote the set of individuals under unit k . Define ρ_k to be the proportion of individuals in \mathcal{U}_k with a welfare more than $1 - \delta$, i.e.,

$$\rho_k = \frac{1}{N} \sum_{i \in \mathcal{U}_k} \mathbb{1}\{w_i > 1 - \delta\}. \quad (2)$$

We assume ρ_k is known and freely available. We revisit this assumption in Section 7, where we show learning ρ_k s can be done efficiently with minimal impact on the allocation value. We denote the profile of all units with a vector $\boldsymbol{\rho}$.

After choosing to intervene on a unit, the allocation of interventions within each unit will be handled by unit administrators, such as school officials. We need to assume that administrators are able to avoid a blatantly wasteful allocation of resources. Specifically, we assume administrators are able to avoid intervening on individuals of highest welfare most of the time. Quantitatively, the allocation avoids the top q fraction of individuals according to welfare, possibly misclassifying a q' fraction of all individuals, where $q' \leq q$. For example, with $N = 100$ individuals, $q = 0.25$, and $q' = 0.1$, a total of 75 individuals will receive resources. Out of these, at most 10 individuals may erroneously belong to the top 25 highest welfare individuals.

Given a budget of B , we can treat $K = \lfloor \frac{B}{N(1-q)c} \rfloor$ units. Let T_k be the average treatment effect of unit k . Sort the units ascendingly in terms of ρ_k with any tie-breaking and let $s(k)$ be the unit in the k^{th} place. The *value* of ULA is

$$V_{\text{unit}} = \sum_{k \in [K]} N T_{s(k)}. \quad (3)$$

Dominance Notion. Given two allocation mechanisms, such as ILA and ULA, the next question is how to compare them for a fixed available budget. Ideally, this comparison should be agnostic to the specific unit profile $\boldsymbol{\rho}$, as long as $\boldsymbol{\rho}$ is within a set of unit profiles of our interest. Furthermore, we would like the comparison to be agnostic to how individual welfares are distributed within a unit k , as long as welfares are consistent with ρ_k . Here we define a dominance notion that accomplishes both desiderata.

Definition 2.1 (Dominance notion). A solution concept S_1 (e.g., ULA) dominates another solution concept S_2 (e.g., ILA) for a fixed budget B and for a set of feasible unit profiles \mathcal{P} , if for every profile $\boldsymbol{\rho} \in \mathcal{P}$, for every within-unit welfare distribution consistent with $\boldsymbol{\rho}$, and for every tie-breaking in algorithms, the value of S_1 is at least as the value of S_2 . We denote this dominance relation by $S_1 \succeq S_2$. If there exists a profile $\boldsymbol{\rho} \in \mathcal{P}$ for which S_1 has a larger value than S_2 , we say S_1 strictly dominates S_2 and denote it by $S_1 \succ S_2$. We use $\alpha \cdot S_1 \succeq S_2$ to show that the comparison is made by multiplying the value of S_1 by a constant α .

This notion of dominance is a strong test if \mathcal{P} is broad. Nonetheless, as we discuss in Section 4, surprisingly ULA dominates ILA in this strong sense for a typical \mathcal{P} . For instance, in the presence of high inequality

across units which is a common empirical observation in many settings, there will be enough units with very low ρ_k that can be targeted by ULA. On the other hand, as we discuss in Section 5, for a sufficiently broad \mathcal{P} , ULA can never be dominated. Therefore, the dominance notion can provide a useful characterization of \mathcal{P} for which one solution concept is dominant or nondominated assuming a typical \mathcal{P} .

3 Illustrating the Key Idea

We start with a visualization of the key idea behind our results, illustrating how ULA can outperform ILA in simulated and real-world settings. We will then introduce the underlying theory in a special case.

3.1 Graphical Illustration

Consider units of equal size where individual welfares are drawn from a beta distribution with unique parameters for each unit. Figure 2 depicts such units, sorted according to their median welfare. Within each unit, individuals are sorted based on their welfare and plotted with their position on the horizontal axis and their welfare on the vertical axis. We have implemented a within-unit allocation with parameters $q = 0.3$ and $q' \approx 0$ for simplicity. The top q fraction of individuals at each unit will be excluded during within-unit allocation and appear as faded in the plot.

Suppose prediction costs 20% of the budget. In the figure, we show the values realized by ULA and ILA from each unit. For simplicity, we neglect ULA’s value from high-welfare individuals with $w > 1 - \delta$. As the figure suggests, in the presence of high inequality between units, ULA with an effective within-unit allocation mostly avoids individuals with high welfare and is almost optimal. However, due to the costly prediction, ILA, despite more accurate targeting, cannot reach ULA’s value.

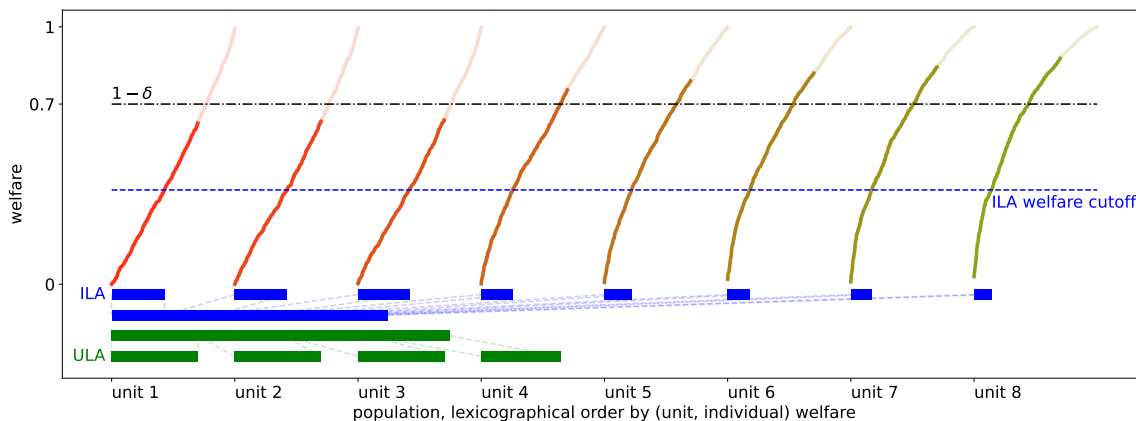


Figure 2: ULA outperforms ILA in a simulated setting with high inequality. The treatment effect is $\delta = 0.3$, and half of the units are treated, with 20% of the budget required for prediction. Within each unit, individuals have independent beta-distributed welfare. They are sorted based on their welfare and plotted according to their position and welfare. A within-unit allocation with $q = 0.3$ and $q' \approx 0$ is considered, and individuals among the top q fraction at each unit are faded. The values obtained by ILA and ULA from each unit are depicted with horizontal bars, and their total values compared. For ease of presentation, we assume ULA gets zero value from treating units above welfare $1 - \delta$.

The observation replicates on real-world data. Here, we utilize the American Community Survey Children’s Education Tabulation,² an annually updated custom data collection of demographic, economic, social, and housing characteristics about school-age children and their families, developed from the U.S. Census

²<https://nces.ed.gov/programs/edge/Demographic/ACS>

Bureau’s 2017-2021 data. We limit our analysis to certain school districts (units) and relevant students (individuals), namely those who reside within the territory of the district. As a proxy for welfare, we examine household income (inflation-adjusted), which is reported in 10 income brackets ranging from less than \$10k to more than \$200k. We consider the highest bracket as having a welfare of 0.95 and the lowest bracket as having a welfare of 0.05.

Figure 3 displays data from representative school districts of similar size in the greater Los Angeles (LA) area (excluding the LA Unified School District due to its sheer size). We consider a similar parameter setting as in the synthetic data example with $\delta = 0.3$, $q = 0.3$, $q' \approx 0$, $p(\epsilon) = 0.2 \cdot B$, and a budget B sufficient to treat half of the units. As the figure suggests, in the presence of high inequality, as is the case for the chosen 8 districts, ULA largely avoids high-welfare students. In contrast, ILA with costly prediction cannot achieve comparable value.

The superiority of ULA primarily stems from the high inequality between the different school districts. The same conclusions hold in other high-inequality regions. For example, Figure 7 illustrates the same point for school districts in New York. Conversely, in regions of low inequality, ULA may not achieve the same value as ILA. To give an example, Figure 8 shows that for Utah’s school districts, ILA comes out ahead. When considering the Gini index of the ρ_k s, the inequality in the chosen New York state school districts is almost twice that of LA and four times that of Utah.

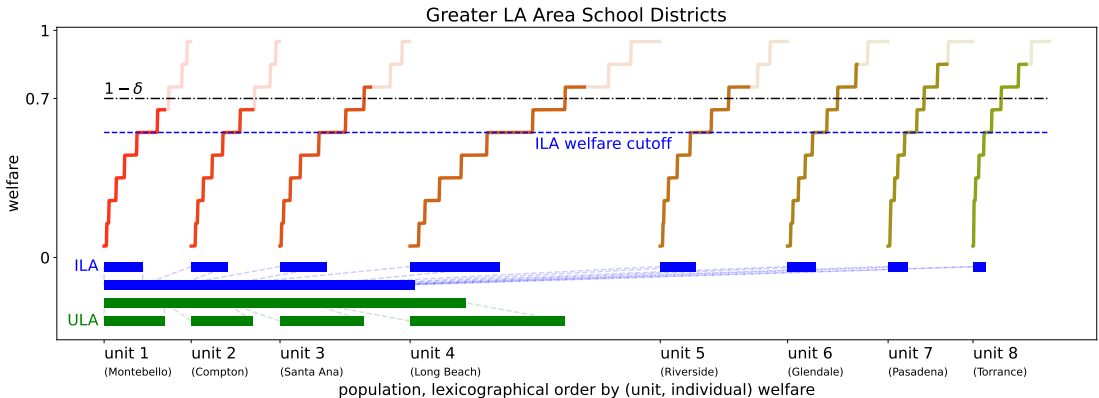


Figure 3: ULA outperforms ILA in a real-world high inequality setting. Eight school districts from the greater Los Angeles (LA) area are considered. Household income, in 10 brackets, is used as a proxy for individual welfare. A within-unit allocation with $q = 0.3$ and $q' \approx 0$ is considered, where individuals among the top q fraction at each unit appear faded. The values obtained by ILA and ULA from each unit are depicted with horizontal bars, and their overall values are compared.

3.2 Theoretical Illustration

We now present our theoretical argument in an illustrative special case as a warm-up exercise. This simple argument shows why ULA dominates ILA when inequality is high, or the budget is low. In the following, we consider cases where the cost of treating everyone $C := MNc$ is too high so that the budget B scales sublinearly or linearly with C .

Recall that we associate each unit k with the fraction of its members with a welfare above $1 - \delta$, denoted by ρ_k . In calculating ULA’s value, we lower bound the value from unit k by $(1 - \rho_k)\delta$. In other words, we neglect the value realized from high-welfare individuals. This approach obviates the need to talk about individual welfares, and the unit profile ρ characterization will be sufficient to present a lower bound on ULA’s value.

Consider M units uniformly distributed around a mean value $\bar{\rho}$: $\rho_1 = \bar{\rho} - g\bar{\rho}$ and $\rho_M = \bar{\rho} + g\bar{\rho}$, for a

constant $g \in [0, 1]$. In this case,

$$\rho_k = \bar{\rho} + 2g\bar{\rho}(k - (M + 1)/2)/(M - 1), \quad k \in [M]. \quad (4)$$

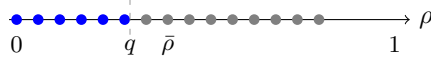
Here, the parameter g controls how spread out the units are around $\bar{\rho}$. We can interpret g as a measure of inequality across units as well. In particular, g is proportional to the Gini index of ρ_k s denoted by G_ρ . By definition,

$$G_\rho := \frac{\sum_{k,k'} |\rho_k - \rho_{k'}|}{2M^2\bar{\rho}} = 2g \cdot \frac{\sum_{k' < k} (k - k')}{(M - 1)M^2}.$$

Then a direct calculation shows $G_\rho = g(1 + 1/M)/3$.

Consider a within-unit allocation with $q > 0$ and $q' \approx 0$. If $\rho_k \leq q$, ULA avoids all individuals in unit k that have welfare above $1 - \delta$. Hence, every individual treated under ULA in unit k will experience the full δ improvement. For uniformly distributed units, a large q such as $q \geq \bar{\rho}$ ensures this happens for at least half of the units. So, if the budget is not excessively large, ULA will be optimal. In particular, so long as predictions are not free, ULA dominates ILA.

Even when $q < \bar{\rho}$, either a low budget or high inequality ensures the dominance of ULA (Figure 4). To see the former, observe that a minimal inequality of $g > 1 - q/\bar{\rho}$ guarantees that at least for the first unit, $\rho_1 < q$. Similarly, for a sublinear budget $B = o(C)$, a minimal inequality of $g \geq 1 - q/\bar{\rho} + o(1)$ ensures that all treated units have $\rho_k \leq q$, implying the optimality of ULA (Figure 4(b)). Even for a linear budget $B = \Theta(C)$, a simple argument shows that with a sufficiently high inequality, say, $g \geq 2(1 - q/\bar{\rho})$, as long as the budget is not excessively large, such as not surpassing the amount for treating one-fourth of the units, the optimality of ULA is ensured (Figure 4(a)).



(a) High inequality ensures for every treated unit, $\rho_k \leq q$ even when many units are treated, thus ULA remains efficient.



(b) Even when inequality is not high, so long as the budget is not large, ULA remains efficient.

Figure 4: In the special case of uniformly distributed units, each unit is represented by a dot. The first K units (the targeted units) are indicated in blue.

Inequality plays a significant role in ULA, including cases where the within-unit allocation treats everyone ($q = q' = 0$). When everyone is treated at unit k , at least a $(1 - \rho_k)$ fraction of individuals experience a δ improvement. So, for K treated units, $V_{\text{unit}} \geq \sum_{k \in [K]} N\delta(1 - \rho_k)$. Plugging ρ_k from Equation (4) into this, a direct calculation yields

$$V_{\text{unit}} \geq KN\delta \left(1 - \bar{\rho} + g\bar{\rho} \frac{M - K}{M - 1} \right).$$

Then, using $K \approx \frac{B}{Nc}$, we obtain

$$V_{\text{unit}} \geq B \frac{\delta}{c} \left(\underbrace{1 - \bar{\rho}}_{\text{value from random targeting}} + \underbrace{g\bar{\rho}(1 - B/C)}_{\text{extra value from inequality}} \right). \quad (5)$$

Equation (5) decomposes the value of ULA: For a dollar spent, the maximum realizable value is δ/c . If units were chosen randomly in ULA, on average, a unit of $\rho_k = \bar{\rho}$ had been treated, which in the worst case could decrease the rate budget is turning into value by a factor of $(1 - \bar{\rho})$. However, in the presence of inequality,

Table 1: Sufficient conditions for ULA to dominate ILA. Refer to Theorem 4.5 for accurate statements. For a similar table with quantitative values for low, medium, and high, refer to Table 3.

	High inequality	Medium inequality		Low inequality
		Medium or lower budget	Large budget	
High price of prediction	ULA \succ ILA	ULA \succ ILA	ULA \succ ILA	ULA \succ ILA
Medium price of prediction	ULA \succ ILA	ULA \succ ILA		
Low price of prediction		$(1 + O(q')) \cdot \text{ULA} \succeq \text{ILA}$?

the selected units will be those with smaller ρ_k s, and ULA performs better than random. This extra value is proportional to g and $(1 - B/C)$.

In contrast, consider ILA, where we pay a price of p for (say, perfect) prediction. Paying this price, ILA loses at least a value of $p \frac{\delta}{c}$ compared to the optimum. Therefore,

$$V_{\text{ind}} \leq B \frac{\delta}{c} (1 - p/B).$$

Comparing this with the lower bound on ULA's value in Equation (5), we have $\text{ULA} \succeq \text{ILA}$ so long as the fractional price of prediction satisfies

$$p/B \geq \bar{p} (1 - g(1 - B/C)).$$

This requirement becomes weaker as inequality grows, and as budget shrinks.

Going beyond uniformly distributed units by considering arbitrarily distributed units, we will show in Theorem 4.5 that a similar observation of $\text{ULA} \succeq \text{ILA}$ holds in general if either inequality is high or the budget is not too large in the presence of a medium inequality. We give a qualitative view of the results in Table 1.

4 Sufficient Conditions for Dominant Unit-Level Allocation

We present sufficient conditions under which ULA dominates ILA. If we restrict feasible unit profiles to those meeting these conditions, $\text{ULA} \succeq \text{ILA}$. To this end, we set upper bounds on V_{ind} and lower bounds on V_{unit} and then compare the bounds.

4.1 Upper bounding the Value of ILA

Given a predictor with ϵ error that costs $p(\epsilon)$ to be implemented, there remains a budget of $B - p(\epsilon)$ for intervention. If the predictor is sufficiently accurate, only individuals with welfare less than $(1 - \delta)$ will be targeted, and the remaining budget will be used optimally. Therefore, optimistically,

$$V_{\text{ind}} \leq \max \left\{ \frac{B - p(\epsilon)}{c}, 0 \right\} \delta. \tag{6}$$

This bound is tight if there are sufficient individuals with low welfare. Specifically, if there are $I = \lfloor \frac{B - p(\epsilon)}{c} \rfloor$ individuals with $w_i \leq 1 - \delta - 2\epsilon$, the bound is tight (Proposition E.1).

4.2 Lower bounding the Value of ULA

To set a lower bound on V_{unit} (Equation (3)), we first lower bound T_k in terms of the within-unit allocation parameters q and q' and the unit k 's fraction of high welfare individuals ρ_k .

Lemma 4.1 (Treatment effect to loss conversion). *For a (q, q') -within-unit allocation, we have*

$$T_k/\delta \geq 1 - q - \text{loss}(\rho_k),$$

where we define

$$\text{loss}(\rho_k) := \begin{cases} \min\{\rho_k, q'\}, & \rho_k \leq q, \\ \rho_k - (q - q'), & \rho_k > q. \end{cases} \quad (7)$$

See [proof](#) on page 38. The proof is straightforward using the observation that if $\rho_k \leq q$, within-unit allocation largely avoids top ρ_k individuals with $\tau_i < \delta$ by excluding the top q fraction. In the case of $\rho_k > q$, it at least avoids $(q - q')$ of high-welfare individuals with potentially small τ_i s.

Lemma 4.1 implies that instead of lower bounding V_{unit} , we may upper bound the total loss

$$\text{loss}_K(\boldsymbol{\rho}) := \sum_{k \in [K]} \text{loss}(\rho_{s(k)})$$

for any valid sorting $s(\cdot)$ of the ρ_k s. The next lemma provides such a bound on loss_K .

Lemma 4.2. *For a unit profile $\boldsymbol{\rho}$, denote the mean and the Gini index of ρ_k s by $\bar{\rho}$ and G_ρ , respectively. Assume the budget is not excessively large, so the number K of units treated by ULA does not exceed $M(1 - \bar{\rho})$. We have*

$$\text{loss}_K(\boldsymbol{\rho}) \leq \begin{cases} Kq', & q > q_c/u(K/M), \\ K(q_c - q + q'), & q \leq q_c/u(K/M). \end{cases}$$

Here, q_c is a critical value for q defined as

$$q_c := \bar{\rho} \frac{1 - \bar{\rho} - G_\rho}{1 - \bar{\rho} - \bar{\rho}G_\rho}, \quad (8)$$

and $u(K/M) := 1 - 2(K/M) \frac{1 - \bar{\rho}}{1 - \bar{\rho} - \bar{\rho}G_\rho} = 1 - O(K/M)$.

See [proof](#) on page 38. This lemma naturally gives rise to a critical value q_c for q : Neglecting $O(K/M)$ in $u(K/M)$ —as is the case for a sublinear budget, for instance— q_c defines two regimes for loss depending on whether $q > q_c$ or $q \leq q_c$. Due to its importance, we define:

Definition 4.3 (Effective within-unit allocation). We say a (q, q') -within-unit allocation is effective if $q > q_c$ and least effective otherwise. In both cases, we assume $q' \leq (1 - \bar{\rho})q$.

A direct calculation shows that $q > q_c$ in a sufficiently high inequality setting where $G_\rho/(1 - \bar{\rho}) > \frac{1 - q/\bar{\rho}}{1 - q}$. It is also straightforward to show $q_c \leq \bar{\rho}(1 - G_\rho)$. So, regardless of the level of inequality, $q > \bar{\rho}$ is sufficient to obtain $q > q_c$. As an immediate result of Lemma 4.2, we can lower bound V_{unit} . For brevity, we present this only for $B = o(C)$ next.

Corollary 4.4. *Under the conditions of Lemma 4.2, for $B = o(C)$, neglecting $o(1)$, ULA's value scales at least linearly with B . The rate at which budget converts to value has two regimes: If the within-unit allocation is effective,*

$$V_{\text{unit}} \geq B \frac{\delta}{c} \left(1 - \frac{q'}{1 - q}\right), \quad (9)$$

and if the within-unit allocation is least effective,

$$V_{\text{unit}} \geq B \frac{\delta}{c} \left(1 - \frac{q_c - (q - q')}{1 - q}\right). \quad (10)$$

Proof. Recall $T_k/\delta \geq 1 - q - \text{loss}(\rho_k)$. Then, summing the treatment effect over the first K units, we have

$$V_{\text{unit}} = \sum_{k \in [K]} N T_{s(k)} = KN\delta(1 - q) - N\delta \text{loss}_K(\boldsymbol{\rho}).$$

Plugging the bound on $\text{loss}_K(\boldsymbol{\rho})$ from Lemma 4.2 with $u(K/M) \approx 1$ into this and using $K \approx \frac{B}{N(1-q)c}$ complete the proof. \square

These bounds are insightful about the interaction of parameters. First, an effective within-unit allocation makes ULA almost optimal: Looking at Equation (9), for a small q' , ULA achieves the optimal rate of δ/c . The deficiency of ULA scales with $q'/(1 - q)$. The dependency on q' reflects the imperfection in within-unit allocation. The scaling with $1/(1 - q)$ reflects the relative number of units that should be treated compared to the case of $q = 0$. While increasing q positively influences within-unit allocation towards the effective regime, it comes at the cost of searching for more units: By refusing to allocate to the top q of the low-welfare units, we may overlook their individuals in need, risking the search over the next units that may not have enough eligible individuals.

When the within-unit allocation is least effective, ULA deviates from optimal allocation. However, the signal from inequality still makes it better than randomly targeting units. Using $q' \leq (1 - \bar{\rho})q$ and $q_c \leq \bar{\rho}(1 - G_\rho)$, we have

$$V_{\text{unit}} \geq B \frac{\delta}{c} \left(1 - \frac{q_c - (q - q')}{1 - q} \right) \geq B \frac{\delta}{c} (1 - q_c) \quad (11)$$

$$\geq B \frac{\delta}{c} \left(\underbrace{1 - \bar{\rho}}_{\text{value from random targeting}} + \underbrace{\bar{\rho} G_\rho}_{\text{extra value from inequality}} \right). \quad (12)$$

Equation (11) shows the least effective within-unit allocation is no worse than the case of $q = q' = 0$. Equation (12) shows ULA leverages inequality signal, attaining extra value proportional to $\bar{\rho} G_\rho$.

At this point, it is important to highlight the role of inequality in ULA. Examining Equation (8), we can see that greater inequality relaxes the requirement for an effective q . Therefore, inequality significantly increases the effectiveness of omitting the top q . Moreover, even when within-unit allocation is least effective, inequality still informs ULA to avoid the high welfare units. These observations underscore the significant role inequality plays in ULA.

4.3 Establishing Sufficient Conditions

ULA's value exceeds ILA's for sure if the lower bound on ULA's value surpasses the upper bound on ILA's. We can therefore combine our previous lemmas to give sufficient conditions for ULA to dominate ILA in the next theorem.

Theorem 4.5 (Sufficient conditions for a dominant ULA). *Consider ULA with a (q, q') -within-unit allocation and a budget B no more than the cost of treating $M(1 - \bar{\rho})$ units. Define the normalized Gini coefficient $\hat{G}_\rho := G_\rho/(1 - \bar{\rho}) \in [0, 1]$, and consider the inequality thresholds*

$$\hat{G}_\rho^{(1)} := 1 - \frac{q}{\bar{\rho}} \frac{1 - \bar{\rho}}{1 - q} \leq \hat{G}_\rho^{(2)} := 1 - \frac{1}{4} \frac{q}{\bar{\rho}} \frac{1 - \bar{\rho}}{1 - q}.$$

- If $\hat{G}_\rho > \hat{G}_\rho^{(2)}$, within-unit allocation is effective, and as long as the budget is not excessively large (not surpassing the cost of treating $M(1 - \bar{\rho})/2$ units), ULA achieves $(1 - q'/(1 - q))$ of the maximal value. This implies $(1 + O(q')) \cdot \text{ULA} \succeq \text{ILA}$.
If further $p(\epsilon)/B > q'/(1 - q)$, we have $\text{ULA} \succ \text{ILA}$.

- If $\widehat{G}_\rho > \widehat{G}_\rho^{(1)}$, within-unit allocation is still effective, and we obtain similar results as long as the budget meets

$$\frac{B}{C} < (1 - q) \cdot \max \left\{ 1 - \frac{\bar{\rho}}{q}, (1 - \bar{\rho}) \left(1 - \sqrt{\frac{q_c}{q}} \right) \right\}. \quad (13)$$

- In any case, if $p(\epsilon)$ consumes a small but sufficient part of the budget as $p(\epsilon)/B > q_c$, we have $\text{ULA} \succ \text{ILA}$.

See [proof](#) on page 43. Tables 1 and 3 summarize this theorem. Theorem E.2 is a stronger version for the case of sublinear budget $B = o(C)$.

5 Necessary Conditions for Dominant Individual-Level Allocation

Previously, we showed that under weak assumptions ULA dominates ILA. Now, we show that ILA dominates ULA only when strong necessary conditions are met. Put differently, under weak conditions on the set of profiles \mathcal{P} , we can show that ILA does not dominate ULA. To show that ILA does not dominate ULA, it suffices to find a profile $\rho \in \mathcal{P}$ for which ULA yields a larger value than ILA.

To approach this result we construct a specific profile ρ of small loss. The value achieved by this profile will then serve as a lower bound on ULA's value.

Lemma 5.1. *Consider ULA with a budget for treating K units that is no more than $M(1 - \bar{\rho})$. There exists a profile ρ^* with mean $\bar{\rho}$ and Gini index G_ρ or less, such that*

$$\text{loss}_K(\rho^*) = K \text{loss} \left(\max \left\{ \bar{\rho} \left(1 - \frac{M}{K} G_\rho \right), 0 \right\} \right). \quad (14)$$

See [proof](#) on page 43. The proof is constructive.

Remark 5.2. Since $\text{loss}(\cdot)$ is piecewise linear by definition, one can verify that $\text{loss}_K(\rho^*)$ in Equation (14) remains piecewise linear in K (and therefore also B).

Equation (14) already clarifies the role of inequality: Since $\text{loss}(\cdot)$ is an increasing function, $\text{loss}_K(\rho^*)$ decreases with G_ρ to the point where the loss is 0, and ULA becomes optimal. This happens for $G_\rho \geq K/M$. For a sublinear budget, this condition always holds asymptotically:

Corollary 5.3. *For a sublinear budget $B = o(C)$, if the set of feasible unit profiles \mathcal{P} does not exclude profiles with a Gini index of G_ρ or less, as long as $G_\rho \geq \frac{B}{(1-q)C} = o(1)$, there exists $\rho^* \in \mathcal{P}$ for which ULA is optimal.*

The situation is more involved in the case of a linear budget. Looking at Equation (14) and the definition of $\text{loss}(\cdot)$ in Equation (7), for ULA it is desired to have $\bar{\rho} \left(1 - \frac{M}{K} G_\rho \right) \leq q$, so the loss would not exceed Kq' . A small K corresponding to a small B and the inclusion of high-inequality profiles can make this happen. Next theorem, a counterpart of Theorem 4.5, states sufficient conditions of $\text{ILA} \not\succeq \text{ULA}$ in general:

Theorem 5.4 (Sufficient conditions for a nondominated ULA). *Consider ULA with a (q, q') -within-unit allocation and a budget B no more than the cost of treating $M(1 - \bar{\rho})$ units. Consider a scenario in which the set of feasible unit profiles \mathcal{P} does not rule out profiles with a Gini index of G_ρ or lower. Define the normalized Gini coefficient $\widehat{G}_\rho := G_\rho / (1 - \bar{\rho}) \in [0, 1]$, and consider the inequality threshold*

$$\widehat{G}_\rho^{(0)} := \frac{1}{2} - \frac{1}{2} \frac{q}{\bar{\rho}}.$$

- If $\widehat{G}_\rho \geq \widehat{G}_\rho^{(0)}$, as long as the budget is not excessively large (not exceeding the cost of treating $M(1 - \bar{\rho})/2$ units), there exists $\rho^* \in \mathcal{P}$ for which ULA attains $(1 - q'/(1 - q))$ of the maximal value, implying $\text{ILA} \not\succeq (1 + O(q')) \cdot \text{ULA}$.
If further $p(\epsilon)/B > q'/(1 - q)$, we have $\text{ILA} \not\succeq \text{ULA}$.

Table 2: Sufficient conditions for a nondominated ULA. Refer to Theorem 5.4 for accurate statements. For a similar table with quantitative values for low, medium, and high, refer to Table 4. *In general, compared to Table 1, a high price of prediction and a medium or higher inequality require weaker conditions.

	Medium or higher* inequality	Low inequality
High* price of prediction	ILA $\not\leq$ ULA	ILA $\not\leq$ ULA
Medium price of prediction	ILA $\not\leq$ ULA	?
Low price of prediction	ILA $\not\leq (1 + O(q')) \cdot \text{ULA}$	

- In any case, if $p(\epsilon)$ uses a small but sufficient part of the budget as $p(\epsilon)/B > \bar{\rho}(1 - \widehat{G}_\rho)$, we have ILA $\not\leq$ ULA.

See proof on page 44. Table 2 summarizes the theorem.

Remark 5.5. If $q > \bar{\rho}$, both $\widehat{G}_\rho^{(0)}$ and $\widehat{G}_\rho^{(1)}$ will be negative. Otherwise, $\widehat{G}_\rho^{(0)} \leq \widehat{G}_\rho^{(1)}/2 \leq \widehat{G}_\rho^{(2)}/2$.

Contrasting Theorems 5.4 and 4.5 using the above remark, a nondominated ULA requires weaker conditions than a dominant ULA as expected: Even for $q = q' = 0$, all of the cases in Theorem 5.4 are still plausible. To illustrate the results, Figure 5 gives a more precise version of Figure 1 in terms of the three different inequality thresholds that come out of Theorems 4.5 and 5.4.

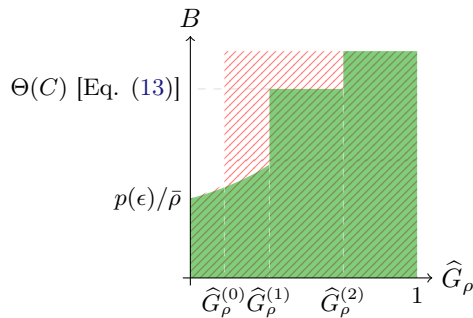


Figure 5: Sufficient conditions for dominant ULA (green) and nondominated ULA (red) for $q \leq \bar{\rho}$.

6 Heterogeneous Intervention Effects

Thus far, we assumed intervention has a fixed effect of δ capped at the maximum welfare of 1. In this section, we allow for a general intervention effect $\tau(w)$. We show that, under certain assumptions on the Lipschitzness of τ and limited concentration of welfare distribution, a sufficiently high inequality or a small budget ensures ULA outperforms ILA and, in fact, is close to the optimal allocation.

Notation and Preliminaries. We assume $\tau(\cdot)$ is a decreasing function with $\tau(0) = \delta'$ and $\tau(1) = 0$. The average treatment effect at unit k with a (q, q') -within-unit allocation is denoted by T_k . If everyone in k is treated, the average effect is denoted by T'_k . The average and Gini index of T'_k s are represented by \bar{T}' and G_T , respectively. Our focus is on a setting with inequality G_T and mean treatment effect \bar{T}' on everyone, aiming to bound the advantage that a heterogeneous τ can give to ILA. Such an approach is most effective when we make assumptions about welfare distribution. We denote the density of welfare at unit k by p_w^k and the overall welfare density by p_w . Abusing notation, we use V_{unit} and V_{ind} to denote the expected ULA and ILA's values.

6.1 Upper bounding the Value of ILA

Equation (6) shows that after paying for prediction, every remaining dollar can turn into value at a rate of δ'/c . With no assumption on welfare distribution, there would be enough individuals with a 0-welfare and no one in $[0, 2\epsilon]$, making this bound tight. We avoid such cases by imposing the following assumption.

Assumption 6.1. The overall welfare density is bounded from above and below: $\gamma \geq p_w(w) \geq \underline{\gamma}$.

Prediction accuracy was less relevant when the treatment effect was homogeneous and there were sufficient low-welfare individuals. However, in the case of a heterogeneous effect, distinguishing small differences in welfare is valuable. We propose a model for prediction error in the following. Recall ILA sorts individuals in terms of their predicted welfare \hat{w} and allocates to those with the lowest welfare. This is equivalent to allocating to any individual i with $\hat{w}_i \leq w_t$ for the largest possible choice of a threshold w_t that meets the budget constraint. We assume prediction errors are symmetric and independent. Formally, for an individual with welfare w , we have $\Pr(\hat{w} \leq w_t | w) = \sigma(w_t - w)$, where $\sigma(\Delta w)$ is 1 for $\Delta w > \epsilon$, and is 0 for $\Delta w < -\epsilon$. Denoting the derivative of σ by σ' , we assume $1/\epsilon \geq |\sigma'(\Delta w)| \geq 1/(2\epsilon)$ for $|\Delta w| \leq \epsilon$. The budget constraint requires the number of treated to not exceed $(B - p(\epsilon))/c$. We relax this constraint and assume the budget should be met in expectation.

When τ changes sharply, distinguishing individuals with a small difference in welfare gains significant value. We control this advantage by considering a Lipschitz continuous τ . By leveraging Assumption 6.1 and Lipschitzness, next, we provide a tighter bound on ILA's value.

Theorem 6.2. Consider ILA with an ϵ -accurate welfare predictor ($\epsilon \leq 1/2$) that costs $p(\epsilon)$. Define the remaining budget as $\tilde{B} := B - p(\epsilon)$. Under Assumption 6.1, suppose \tilde{B}/C is no less than $\underline{\gamma}\epsilon$ and no more than $\gamma(1 - \epsilon)$. For a concave and differentiable τ with $l \geq |\frac{d\tau}{dw}| \geq l/2$, we have

$$V_{\text{ind}} \leq \tilde{B} \frac{1}{c} \left(\delta' - \frac{l}{8\gamma} (\tilde{B}/C + \underline{\gamma}\epsilon/2) \right). \quad (15)$$

See proof on page 44. Compared to the optimal rate of δ'/c , Equation (15) reflects two deficiencies: First, a larger budget implies that more individuals will be treated, including those with treatment effects farther from δ' . Second, lower accuracy means that more individuals with high welfare values may be recognized as eligible.

6.2 Lower bounding the Value of ULA

ULA's previous bound can be reused in this general setting if, for a threshold w_t , we lower bound τ with $\tilde{\tau}(w) = \delta \cdot \mathbb{1}\{w \leq w_t\}$, where $\delta = \tau(w_t)$. This approach underestimates the effect on the low-welfare population by an amount as large as $\delta' - \delta$ and neglects the effect on the population above w_t . So, there might be no good choice of w_t in this tradeoff. Next, we get around this by proposing ULA based on estimated unit-level average treatment effects instead of ρ_k s. Suppose ULA sorts units in terms of T'_k s and allocates to the top K units. We assume such estimates are available for free. We will show in Section 7 that under some assumptions on τ , the additional cost of learning reduces ULA's value negligibly. The following assumption helps us avoid contrived worst-case scenarios and tightly relate T'_k and T_k .

Assumption 6.3. For every unit k , assume $p_w^k(w) \leq \gamma$.

Lemma 6.4 (T'_k to T_k conversion). For a (q, q') -within-unit allocation, under Assumption 6.3, we have

$$T_k \geq Q_\tau(T'_k - q'\delta'; q, q', \gamma),$$

where

$$Q_\tau(t; q, q', \gamma) := \gamma \Gamma_{(1-q-q')/\gamma} \left(\Gamma_{(1-2q')/\gamma}^{-1}(t/\gamma) \right),$$

$$\Gamma_a(w) := \int_w^{w+a} \tau(x) dx.$$

See [proof](#) on page 46.

Remark 6.5. For every $\gamma > 0$, we have $Q_\tau(t; q, q', \gamma) \geq t \frac{1-q-q'}{1-2q'}$, where equality corresponds to $\gamma \rightarrow \infty$ (as we did not have Assumption 6.3).

Proof. For $b \geq a > 0$ and a decreasing τ , $\Gamma_a(w)/\Gamma_b(w) \geq a/b$. Let $w = \Gamma_b^{-1}(t)$. Then $\Gamma_a(\Gamma_b^{-1}(t)) \geq (a/b)t$. Defining $a = (1 - q - q')/\gamma$, $b = (1 - 2q')/\gamma$, and $t = (T'_k - q'\delta')/\gamma$ completes the proof. \square

In Lemma 6.4, Q_τ transforms T'_k into T_k for a single unit. A property of τ will enable us to utilize Q_τ for converting an average treatment effect \bar{T}'_k to V_{unit} :

Property 1. Given w_c , a , and b ($b \geq a > 0$), τ has Property 1 if $r(w) := \frac{\tau(w) - \tau(w+a)}{\tau(w) - \tau(w+b)}$ is decreasing for $w \leq w_c$.

It can be immediately observed that a strictly decreasing τ has Property 1 for any w_c if $a = b$. See Proposition E.3 for a necessary characterization of general treatment effects that have Property 1. In fact, this property holds for many of the typical τ s, for example:

Example 6.6. Consider a treatment effect of δ' at $w = 0$ decreasing linearly with w and capped at maximum welfare:

$$\tau(w) = \min \{w + (\delta' - lw), 1\} - w.$$

Here, $\delta' = \delta + (1 - \delta)l$ and $l > 0$. One can verify that for any a and b ($b \geq a > 0$), τ has Property 1 for $w_c = 1 - \delta - a$.

In Theorem E.4, we show that for a τ satisfying Property 1,

$$V_{\text{unit}} \geq KQ_\tau(\bar{T}'/(1 - G_T) - q'\delta'; q, q', \gamma). \quad (16)$$

So, effectively, ULA chooses a unit with $T' = \bar{T}'/(1 - G_T)$ — clearly showing how inequality can make ULA extremely more effective. We skip the details of the theorem here and directly utilize it to identify a dominant ULA next.

6.3 Sufficient Conditions For a Dominant ULA

Our specific interest lies in the *elbow-shaped* treatment effect (Figure 6), where, starting from $\tau(0) = \delta'$, an l -Lipschitz effect gradually decreases and reaches δ . Beyond this point, the welfare after intervention will reach its cap of 1, limiting the effect. Example 6.6 is a special case of this model that satisfies Property 1. If τ of Figure 6 similarly satisfies Property 1, Theorem E.4 enables us to bound V_{unit} as Equation (16). Then a comparison to Theorem 6.2 can further establish sufficient conditions for a dominant ULA. The next theorem proposes such conditions for a small l . Consistent with our previous observations, high inequality or a reasonably bounded budget can ensure that ULA is effective and dominant.

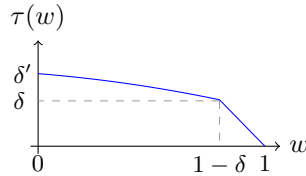


Figure 6: An elbow-shaped heterogeneous effect.

Theorem 6.7 (Sufficient conditions for a dominant ULA in case of a heterogeneous effect). *Consider ULA with a (q, q') -within-unit allocation and a budget B that is no more than the cost of treating $M\bar{T}'/\delta'$ units.*

Under Assumption 6.3 and conditions of Theorem 6.2, suppose an elbow-shaped effect with Property 1 (as in Theorem E.4). Consider the inequality thresholds

$$G_T^{(1)} := 1 - \frac{\bar{T}'}{\delta' - q^2/(2\gamma)} \leq G_T^{(2)} := 1 - \frac{1}{4} \frac{\bar{T}'}{\delta' - q^2/(2\gamma)} - \frac{3}{4} \frac{\bar{T}'}{\delta'}. \quad (17)$$

- If $G_T > G_T^{(2)}$, neglecting $O(l^2)$ and $O(q'^2)$, ULA achieves $(1 - q'/(1 - q) - l/\delta)$ of the optimal value. If further,

$$\frac{p(\epsilon)}{B} > \frac{q'}{1 - q} + \frac{l}{\delta} - \frac{l}{\delta} \frac{1}{8\gamma} \left(\frac{B}{C} + \underline{\gamma}\epsilon/2 \right), \quad (18)$$

we have $\text{ULA} \succ \text{ILA}$.

- If $G_T > G_T^{(1)}$ and the budget meets

$$\frac{B}{C} < \frac{1}{2} (1 - q)(2\gamma\delta'/q^2 - 1)(G_T - G_T^{(1)}),$$

we obtain similar results.

The proof hinges on the strong convexity of Q_τ and presents a stronger statement of the theorem.

7 Learning Unit-Level Statistics

Welfare is not directly observable. However, noisy observations of welfare may be available. For instance, grades or dropout rates can serve as informative signals of student welfare. We refer to this information as type 1, with a cost of c_1 per query per individual. On the other hand, obtaining true welfare is often expensive and may necessitate fresh data. This might involve students taking some standard exams or patients undergoing a set of tests. We call this type of data type 2 with a cost of c_2 per data point. In practice, obtaining true welfare may incur a cost comparable to the cost of intervention c , and we expect $c_2 \approx c \gg c_1$. So, we may collect type 2 only from a subset of individuals in select units to train a predictor acting on type 1.

In this section, we explore the efficient estimation of unit-level statistics, such as ρ_k or T'_k , even when individual welfare estimation poses challenges. We consider general unit-level statistics that can be expressed as the average of a function $u(\cdot)$ evaluated at w_i s. For example, ρ_k is the average of $w \rightarrow \mathbb{1}\{w > 1 - \delta\}$ at unit k , or T'_k is the average of τ . For brevity, we discuss the results when $N = \Theta(M)$; however, Appendix D provides the general results.

Starting with the simplest case, consider estimating unit-level statistics by sampling a few individual welfare from each unit. This estimation relies on type 2 data and does not involve learning. As discussed in Appendix D.1, the relative decrease in ULA's value due to inaccuracy in estimation and the cost of obtaining type 2 data asymptotically becomes negligible in expectation when there are many units and the budget is sufficient to effectively treat $K = \Omega(M^{2/3})$ units.

Moving beyond estimating unit-level statistics with sampled welfare values, we aim to learn a predictor that acts on type 1 observations. This approach potentially reduces the sampling cost by relying on fewer type 2 examples. We first demonstrate that learning unit-level statistics from type 1 observations is no harder and often significantly easier than learning individual-level welfare (Appendix D.2). Formally, suppose there exists a high-probability generalization bound on the mean squared error of the individual welfare predictor in terms of the pseudo-dimension of the hypothesis class. Then, for a monotone concave or convex u , a similarly accurate predictor of unit-level statistics can be obtained with the same or significantly better sample complexity if the predictor is additionally unbiased. Moreover, averaging within units makes the intrinsic unpredictability of individuals less relevant to unit-level statistics.

Next, we focus on methods that directly learn unit-level statistics without relying on an individual predictor. In the simplest model (Appendix D.3), we consider noise-free observations $o \in \mathcal{O}$ as an unknown

deterministic function of welfare: $o = f(w)$, assuming f is strictly monotone. Unlike individual-level prediction, which requires learning f^{-1} , learning a single parameter $f^{-1}(1 - \delta)$ suffices for simple statistics like ρ in the context of ULA. Theorem D.1 shows that under Assumption 6.1 and 6.3, with an appropriately chosen number of type 2 samples and algorithm, the relative decrease of V_{unit} due to this extra step of sampling and learning is $O(\sqrt{c/B} \ln(1/\alpha))$ with probability at least $1 - 2\alpha$. This change is negligible since we expect $B \gg c$. A weaker result holds for a general 1-Lipschitz u , implying a negligible decrease in ULA’s expected value (Theorem D.2).

In a more general setting, observation is a function of welfare and an independent noise term z : $o = f(w, z)$. In Appendix D.4, we discuss learning unit-level statistics under this model with different assumptions on f and noise. Notably, we present a result on the learnability of a ranking of the units under a nonparametric model with minimal assumptions about noise. More precisely, if the noise does not distort the observation too much, meaning there exists a function $v : \mathcal{O} \rightarrow \mathbb{R}$ such that $\mathbb{E}_{P_w^k}[u(w)] \approx \mathbb{E}_{P_o^k}[v(o)]$ while $\|v\|_\infty = O(\|u\|_\infty)$, the risk of misranking can be minimized by an SVM-like algorithm minimizing an empirical margin loss (Lemma D.6). Hinging on this result, we bound the relative decrease of V_{unit} due to ranking error and cost of learning. Asymptotically, for an appropriately chosen margin in the algorithm, when the number of treated units exceeds $M^{1/2}$, the relative decrease of V_{unit} is negligible with high probability (Theorem D.8). Although we perform a worst-case analysis in terms of where the ranking errors might occur, this result is still significant compared to the setting without learning. In sum, learning an accurate ranking of units in terms of their unit-level statistics is possible under minimal assumptions, and the cost of learning can be negligible in the analysis of ULA’s value.

8 Discussion

Our contribution is two-fold: On the technical side, we provide a mathematical framework for evaluating the efficacy of predictions—the tools presented here point to lines of inquiry to investigate prediction as a solution concept for allocation. On the conceptual front, we surface inequality as a driving mechanism limiting the utility of prediction. Our insights run counter to the intuition that prediction may curb wasteful allocations in settings where resources are scarce.

There are several open questions of interest. For instance, in all cases, we only model individual treatment effects. An extension could consider modeling individuals on a network to examine network effects. Likewise, both ULA and ILA treat individuals, and we do not leverage the presence of units beyond identification. A separate line of inquiry may investigate comparing individual versus group interventions. That is, some interventions may only be available at the unit level, e.g., increasing the number of doctors in a hospital or teachers in a school, which would impact the entire unit.

There are known limitations to allocating by proxy, which individual risk scores often are. These limitations apply whether we are targeting individuals using predictions or unit-level information. Significant gaps exist in our understanding of the limits to prediction as a proxy, and we echo recent calls to examine this concern further.

Finally, targeting itself is a contested means for intervention, as argued by various social scientists and policy-makers (Rawls, 1971; Sen, 1980, 1981; Mkandawire, 2005). While both ULA and ILA are ultimately used for targeting individuals, the insights in our work point to the persistence of structural forces in determining outcomes.

Acknowledgements

RA was supported by a grant from the Andrew Carnegie Fellowship Program. The authors would like to thank Nathan Srebro, Jon Kleinberg, Ariel Procaccia, Evi Micha, Max Kasy, André Cruz, Ishaq Aden-Ali, and MLO Reading Group at TTIC for extensive feedback and discussions on earlier versions of this work.

References

- Unai Fischer Abaigar, Christoph Kern, Noam Barda, and Frauke Kreuter. Bridging the gap: Towards an expanded toolkit for ml-supported decision-making in the public sector, 2023.
- Rediet Abebe, Jon Kleinberg, and S Matthew Weinberg. Subsidy allocations in the presence of income shocks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7032–7039, 2020.
- Susan Athey and Stefan Wager. Policy learning with observational data. *Econometrica*, 89(1):133–161, 2021.
- Chelsea Barabas, Madars Virza, Karthik Dinakar, Joichi Ito, and Jonathan Zittrain. Interventions over predictions: Reframing the ethical debate for actuarial risk assessment. In *Proc. ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, pages 62–76, 2018.
- Dimitris Bertsimas and Nathan Kallus. From predictive to prescriptive analytics. *Management Science*, 66(3):1025–1044, 2020.
- Niclas Boehmer, Yash Nair, Sanket Shah, Lucas Janson, Aparna Taneja, and Milind Tambe. Evaluating the effectiveness of index-based treatment allocation. *arXiv preprint arXiv:2402.11771*, 2024.
- Mary Bruce, John M Bridgeland, Joanna Hornig Fox, and Robert Balfanz. On track for success: The use of early warning indicator and intervention systems to build a grad nation. *Civic Enterprises*, 2011.
- Carri W Chan, Vivek F Farias, Nicholas Bambos, and Gabriel J Escobar. Optimizing intensive care unit discharge decisions with patient readmissions. *Operations research*, 60(6):1323–1341, 2012.
- Stephanie Cuccaro-Alamin, Regan Foust, Rhema Vaithianathan, and Emily Putnam-Hornstein. Risk assessment and decision making in child protective services: Predictive risk modeling in context. *Children and Youth Services Review*, 79:291–298, 2017.
- Adam N Elmachtoub and Paul Grigas. Smart “predict, then optimize”. *Management Science*, 68(1):9–26, 2022.
- Adam N Elmachtoub, Jason Cheuk Nam Liang, and Ryan McNellis. Decision trees for decision-making under the predict-then-optimize framework. In *International Conference on Machine Learning*, pages 2858–2867. PMLR, 2020.
- Ann-Marie Faria, Nicholas Sorensen, Jessica Heppen, Jill Bowdon, Suzanne Taylor, Ryan Eisner, and Shandu Foster. Getting students on track for graduation: Impacts of the early warning intervention and monitoring system after one year. rel 2017-272. *Regional Educational Laboratory Midwest*, 2017.
- Moritz Hardt and Michael Kim. Is your model predicting the past? In *Proc. 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, 2023.
- Jake M Hofman, Duncan J Watts, Susan Athey, Filiz Garip, Thomas L Griffiths, Jon Kleinberg, Helen Margetts, Sendhil Mullainathan, Matthew J Salganik, Simine Vazire, et al. Integrating explanation and prediction in computational social science. *Nature*, 595(7866):181–188, 2021.
- Guido W Imbens and Jeffrey M Wooldridge. Recent developments in the econometrics of program evaluation. *Journal of economic literature*, 47(1):5–86, 2009.
- Rebecca Ann Johnson and Simone Zhang. What is the bureaucratic counterfactual? categorical versus algorithmic prioritization in US social policy. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1671–1682, 2022.
- Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan, and Ziad Obermeyer. Prediction policy problems. *American Economic Review*, 105(5):491–95, May 2015.

- Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. Human decisions and machine predictions. *The quarterly journal of economics*, 133(1):237–293, 2018.
- Amanda R Kube, Sanmay Das, and Patrick J Fowler. Community-and data-driven homelessness prevention and service delivery: optimizing for equity. *Journal of the American Medical Informatics Association*, 30(6):1032–1041, 2023a.
- Amanda R Kube, Sanmay Das, and Patrick J Fowler. Fair and efficient allocation of scarce resources based on predicted outcomes: implications for homeless service delivery. *Journal of Artificial Intelligence Research*, 76:1219–1245, 2023b.
- Karen Levy, Kyla E. Chasalow, and Sarah Riley. Algorithms and decision-making in the public sector. *Annual Review of Law and Social Science*, 17(1):309–334, 2021.
- Dexun Li and Pradeep Varakantham. Efficient resource allocation with fairness constraints in restless multi-armed bandits. In *Uncertainty in Artificial Intelligence*, pages 1158–1167. PMLR, 2022.
- Lydia T Liu, Solon Barocas, Jon Kleinberg, and Karen Levy. On the actionability of outcome prediction. *arXiv preprint arXiv:2309.04470*, 2023a.
- Lydia T Liu, Serena Wang, Tolani Britton, and Rediet Abebe. Reimagining the machine learning life cycle to improve educational outcomes of students. *Proceedings of the National Academy of Sciences*, 120(9):e2204781120, 2023b.
- Martha Abele Mac Iver, Marc L Stein, Marcia H Davis, Robert W Balfanz, and Joanna Hornig Fox. An efficacy study of a ninth-grade early warning indicator intervention. *Journal of Research on Educational Effectiveness*, 12(3):363–390, 2019.
- Charles F Manski. Statistical treatment rules for heterogeneous populations. *Econometrica*, 72(4):1221–1246, 2004.
- Tasfia Mashiat, Alex DiChristofano, Patrick J Fowler, and Sanmay Das. Beyond eviction prediction: Leveraging local spatiotemporal public records to inform action. *arXiv preprint arXiv:2401.16440*, 2024.
- Aditya Mate, Bryan Wilder, Aparna Taneja, and Milind Tambe. Improved policy evaluation for randomized trials of algorithmic resource allocation. *arXiv preprint arXiv:2302.02570*, 2023.
- Thandika Mkandawire. Targeting and universalism in poverty reduction. 2005.
- Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT press, 2018.
- Lucas Monteiro Paes, Carol Long, Berk Ustun, and Flavio Calmon. On the epistemic limits of personalized prediction. In *Advances in Neural Information Processing Systems*, volume 35, pages 1979–1991, 2022.
- Erina Seh-Young Moon and Shion Guha. A human-centered review of algorithms in homelessness research. *arXiv preprint arXiv:2401.13247*, 2024.
- Ayan Mukhopadhyay and Yevgeniy Vorobeychik. Prioritized allocation of emergency responders based on a continuous-time incident prediction model. In *International Conference on Autonomous Agents and MultiAgent Systems*, 2017.
- Juan Carlos Perdomo. The relative value of prediction in algorithmic decision making. *arXiv preprint arXiv:2312.08511*, 2023.
- Juan Carlos Perdomo, Tolani Britton, Moritz Hardt, and Rediet Abebe. Difficult lessons on social prediction from wisconsin public schools. *arXiv preprint arXiv:2304.06205*, 2023.

- Gaurab Pokharel, Sanmay Das, and Patrick J Fowler. Discretionary trees: Understanding street-level bureaucracy via machine learning. *arXiv preprint arXiv:2312.10694*, 2023.
- Shir Raviv. When do citizens resist the use of algorithmic decision-making in public policy? Theory and evidence. *Theory and Evidence (January 18, 2023)*, 2023.
- John Rawls. *A theory of justice*. Harvard University Press, 1971.
- Sina Rismanchian and Shayan Doroudi. Four interactions between ai and education: Broadening our perspective on what ai can offer education. In *International Conference on Artificial Intelligence in Education*, pages 1–12. Springer, 2023.
- Matthew J. Salganik, Ian Lundberg, Alexander T. Kindel, Caitlin E. Ahearn, Khaled Al-Ghoneim, Abdullah Almaatouq, Drew M. Altschul, Jennie E. Brand, Nicole Bohme Carnegie, Ryan James Compton, Debanjan Datta, Thomas Davidson, Anna Filippova, Connor Gilroy, Brian J. Goode, Eaman Jahani, Ridhi Kashyap, Antje Kirchner, Stephen McKay, Allison C. Morgan, Alex Pentland, Kivan Polimis, Louis Raes, Daniel E. Rigobon, Claudia V. Roberts, Diana M. Stanescu, Yoshihiko Suhara, Adaner Usmani, Erik H. Wang, Muna Adem, Abdulla Alhajri, Bedoor AlShebli, Redwane Amin, Ryan B. Amos, Lisa P. Argyle, Livia Baer-Bositis, Moritz Büchi, Bo-Ryehn Chung, William Eggert, Gregory Faletto, Zhilin Fan, Jeremy Freese, Tejomay Gadgil, Josh Gagné, Yue Gao, Andrew Halpern-Manners, Sonia P. Hashim, Sonia Hausen, Guanhua He, Kimberly Higuera, Bernie Hogan, Ilana M. Horwitz, Lisa M. Hummel, Naman Jain, Kun Jin, David Jurgens, Patrick Kaminski, Areg Karapetyan, E. H. Kim, Ben Leizman, Naijia Liu, Malte Möser, Andrew E. Mack, Mayank Mahajan, Noah Mandell, Helge Marahrens, Diana Mercado-Garcia, Viola Mocz, Katariina Mueller-Gastell, Ahmed Musse, Qiankun Niu, William Nowak, Hamidreza Omidvar, Andrew Or, Karen Ouyang, Katy M. Pinto, Ethan Porter, Kristin E. Porter, Crystal Qian, Tamkinat Rauf, Anahit Sargsyan, Thomas Schaffner, Landon Schnabel, Bryan Schonfeld, Ben Sender, Jonathan D. Tang, Emma Tsurkov, Austin van Loon, Onur Varol, Xiafei Wang, Zhi Wang, Julia Wang, Flora Wang, Samantha Weissman, Kirstie Whitaker, Maria K. Wolters, Wei Lee Woon, James Wu, Catherine Wu, Kengran Yang, Jingwen Yin, Bingyu Zhao, Chenyun Zhu, Jeanne Brooks-Gunn, Barbara E. Engelhardt, Moritz Hardt, Dean Knox, Karen Levy, Arvind Narayanan, Brandon M. Stewart, Duncan J. Watts, and Sara McLanahan. Measuring the predictability of life outcomes with a scientific mass collaboration. *Proceedings of the National Academy of Sciences*, 117(15):8398–8403, 2020.
- Amartya Sen. *Equality of What?* Cambridge University Press, 1980.
- Amartya Sen. Issues in the measurement of poverty. In *Measurement in public choice*, pages 144–166. Springer, 1981.
- Ali Shirali. Sequential nature of recommender systems disrupts the evaluation process. In *International Workshop on Algorithmic Bias in Search and Recommendation*, pages 21–34. Springer, 2022.
- Halil Toros and Daniel Flaming. Prioritizing homeless assistance using predictive algorithms: an evidence-based approach. *Cityscape*, 20(1):117–146, 2018.
- Shresth Verma, Aditya Mate, Kai Wang, Neha Madhiwalla, Aparna Hegde, Aparna Taneja, and Milind Tambe. Restless multi-armed bandits for maternal and child health: Results from decision-focused learning. In *AAMAS*, pages 1312–1320, 2023.
- Angelina Wang, Sayash Kapoor, Solon Barocas, and Arvind Narayanan. Against predictive optimization: On the legitimacy of decision-making algorithms that optimize predictive accuracy. *Available at SSRN*, 2022.
- Sebastian Zezulka and Konstantin Genin. From the fair distribution of predictions to the fair distribution of social goods: Evaluating the impact of fair machine learning on long-term unemployment. *arXiv preprint arXiv:2401.14438*, 2024.

Xiaolin Zheng, Menghan Wang, Renjun Xu, Jianmeng Li, and Yan Wang. Modeling dynamic missingness of implicit feedback for sequential recommendation. *IEEE Transactions on Knowledge and Data Engineering*, 34(1):405–418, 2020.

A Additional Related Work

As a form of resource allocation based on prediction, ILA is widely adopted across various domains. These include allocating homelessness funds to minimize the predicted probability of household re-entry (Kube et al., 2023a,b) or expected cost (Toros and Flaming, 2018), reducing eviction rate by targeting assistance to renters at greater predicted risk (Mashiat et al., 2024; Abebe et al., 2020), making ICU discharge decisions based on readmission probability (Chan et al., 2012), improving emergency responses using incident prediction models (Mukhopadhyay and Vorobeychik, 2017), or allocating educational resources based on early warning systems that identify students at risk of dropping out (Faria et al., 2017; Mac Iver et al., 2019; Perdomo et al., 2023). For a more complete discussion on the role of predictive tools in education refer to Liu et al. (2023b) and Rismanchian and Doroudi (2023).

Despite widespread implementation of ILA, in various domains is little evidence of its efficacy (Mac Iver et al., 2019). Reviewing existing algorithmic approaches to reducing homelessness as a resource allocation problem, Moon and Guha (2024) found broad issues when optimization is based on individual-level predictors. They suggest a more “human-centered approach” that involves key stakeholders to understand how structural and individual factors impact hard-to-quantify facets of an individual, such as one’s resilience and willingness to improve. In our model, this kind of mechanism could be seen as an instance of ULA, where within-unit allocation is delegated to the unit administrators.

Perdomo et al. (2023) recently demonstrated that early warning systems, when utilized to target interventions at individual students by predicting the risk of dropout, have shown little improvement despite the use of accurate predictors. Perdomo et al. (2023) and Hardt and Kim (2023) also point out how risk scores draw on unit-level information. Consistent observations were made by Faria et al. (2017) and Mac Iver et al. (2019), indicating no statistically significant impact of the intervention. In particular, Mac Iver et al. (2019) attribute these observations to the lack of contrast between what happened in treatment and control schools, as the control schools were effectively implementing the same mechanism internally. This supports our assumption that units can implement a reasonable within-unit allocation when directed and provided resources. Closely related to our work, Perdomo (2023) asks how improvements in welfare arising from better predictions compare to those of other policy levers, such as expanding access to resources. In line with our results, it often does not help to invest in greater accuracy when resources are limited.

In policy contexts, historically planning has relied on aggregate data. However, the promise of better resource allocation, lower costs, and more preventative intervention has increasingly motivated individual-level algorithmic solutions in public sector applications (Levy et al., 2021). Abaigar et al. (2023) present an overview of key technical challenges where discrepancies between policy goals and machine learning models commonly arise. In their resource allocation setting, policy goals can be formalized through an allocation principle that chooses the target population based on predicted individual outcomes of interest. As they highlight, such individual estimation may not always be helpful for the decision-making process, and the choice of aggregation level is of utmost importance.

There is extensive work on the use of causal inference to identify suitable policies and interventions, see, e.g., (Imbens and Wooldridge, 2009; Bertsimas and Kallus, 2020; Athey and Wager, 2021) and references therein. Zezulka and Genin (2024) argue in social contexts predictions that inform policies should be evaluated based on the change in social goods that arises after deploying the algorithm. Kleinberg et al. (2015) discuss a class of policy problems that do not necessitate causal inference but rather rely on predictive inference.

Empirical evaluation of *index-based* allocations (Mate et al., 2023), where individuals estimated to be in utmost need receive resources, such as the mechanisms we study, requires careful design and estimators (Mate et al., 2023; Boehmer et al., 2024). This is because of the interdependencies among samples, as what an individual receives under these policies depends on her prediction and predictions about others (Zheng et al., 2020; Shirali, 2022).

The extensive debate around the use of risk scores touched on the role of predictive tools as the basis for consequential decisions and interventions. For example, Wang et al. (2022) raise concerns about the legitimacy of decision-making based on predictive optimization. In the context of the criminal justice system, Barabas et al. (2018) argued that risk assessment should be intervention-driven and ideally work as a

diagnostic tool, identifying causal structures that drive crime. Hofman et al. (2021) suggest that explanatory modeling can benefit predictive modeling (and vice versa), for example, by encouraging more robust models that generalize better under interventions. Liu et al. (2023b) discuss improving outcome prediction may not necessarily help decision-makers in selecting a more effective intervention option. Related to this topic is the question of actionability of predictions as studied by Liu et al. (2023a).

Even with access to abundant data on individuals’ lives, it remains challenging to forecast outcomes at the individual level (Salganik et al., 2020). In part due to these inherent limits, some literature shows reluctance by individuals to use predictive systems to make inferences about individuals. Raviv (2023) found citizens exhibit aversion to the use of algorithmic decision systems when they are required to make inferences about individuals rather than collectives. This might root, for instance, in the disproportionate advantage individuals receive in return for sharing their personal data; in fact, it is challenging to determine whether such participation ultimately benefits or harms them (Monteiro Paes et al., 2022).

Related to our discussion of intervention-driven predictions is the predict-then-optimize paradigm in the Operations Research community. Here, a predictive tool is first acquired to maximize accuracy, followed by the optimization of allocations or decisions based on it. Our definition of ILA closely aligns with this paradigm. The disentanglement of these stages has been a subject of study for a long time; refer to Manski (2004) as one of the first attempts where a predictive model is directly optimized to maximize the social welfare resulting from a prediction-based treatment choice. Informing prediction with its downstream applications is widely known as decision-focused learning (Mukhopadhyay and Vorobeychik, 2017; Elmachtoub et al., 2020; Elmachtoub and Grigas, 2022). Although following such an approach can potentially provide ILA with a predictor informed about the follow-up decision, the main deficiency of ILA in our model comes from the cost of prediction. Our generous assumption of strong per-individual guarantees for the predictor makes a decision-focused predictor less relevant for our study of ILA. However, our proposed algorithm for learning unit-level statistics under a nonparametric model in Section 7 chooses a sampling strategy informed by downstream ranking decisions and generally operates within a decision-focused framework. Future works can also extend our framework to sequential learning under a resource-constrained setting (Li and Varakantham, 2022; Verma et al., 2023).

Social service bureaucracies neither fully give street-level bureaucrats full discretion nor fully prioritize algorithmically in targeting assistance (Johnson and Zhang, 2022; Pokharel et al., 2023). The pre-algorithmic, rule-based part typically involves a dimensionality reduction, quantization, and deliberation by the stakeholders, which overall is called categorical prioritization by Johnson and Zhang (2022). Pokharel et al. (2023) discovered that discretionary decisions in resource allocation, not easily explained by simple decision rules, often lead to improved outcomes. This suggests that caseworkers leverage the knowledge gained through human interactions and evaluations in the process. The stronger these discretionary decisions are, the stronger our within-unit allocation, thereby enhancing ULA as well.

B Additional Tables

Table 3: Sufficient conditions for ULA to dominate ILA. Refer to Theorem 4.5 for accurate statements.

	$\widehat{G}_\rho > \widehat{G}_\rho^{(2)}$	$\widehat{G}_\rho^{(2)} \geq \widehat{G}_\rho > \widehat{G}_\rho^{(1)}$	$\widehat{G}_\rho \leq \widehat{G}_\rho^{(1)}$
	B/C meets Eq. (13)		B/C doesn't meet Eq. (13)
$p(\epsilon)/B > q_c$	ULA \succ ILA	ULA \succ ILA	ULA \succ ILA
$p(\epsilon)/B > q'/(1-q)$	ULA \succ ILA	ULA \succ ILA	?
o.w.	$(1 + O(q')) \cdot \text{ULA} \succeq \text{ILA}$		

Table 4: Sufficient conditions for when ILA does not dominate ULA. Refer to Theorem 5.4 for accurate statements.

	$\widehat{G}_\rho \geq \widehat{G}_\rho^{(0)}$	$\widehat{G}_\rho < \widehat{G}_\rho^{(0)}$
$p(\epsilon)/B > \bar{\rho}(1 - \widehat{G}_\rho)$	ILA $\not\succeq$ ULA	ILA $\not\succeq$ ULA
$p(\epsilon)/B > q'/(1-q)$	ILA $\not\succeq$ ULA	?
o.w.	ILA $\not\succeq (1 + O(q')) \cdot \text{ULA}$	

C Additional Figures

Figure 7 extends our visualization to the case of school districts in New York State. These exhibit a significant level of inequality. As a result, ULA realizes a higher value than ILA.

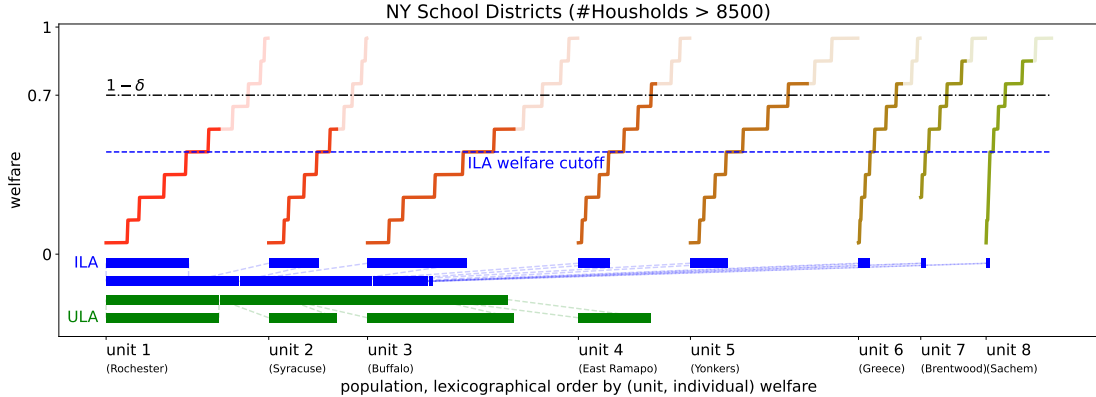


Figure 7: Similar to Figure 3, ULA outperforms ILA in a real-world high inequality setting. New York State's school districts with a household population of 8,500 or more are considered (excluding the New York City Department of Education due to its sheer size). Household income is used as a proxy for individual welfare, categorized into 10 income brackets. A within-unit allocation with $q = 0.3$ and $q' \approx 0$ is considered, and individuals among the top q fraction at each unit appear faded. The values obtained by ILA and ULA from each unit are depicted with horizontal bars, and their overall values are compared.

Figure 8 presents an example with school districts from Utah. The school districts have lower inequality compared with those in New York and Los Angeles. As a result, the comparison between ILA and ULA is closer. We see a slight advantage for ILA. However, recall that for the ease of visualization, we assume that ULA gets no value from individuals above the welfare cutoff $1 - \delta$. In other words, we round any

treatment effect strictly smaller than δ to zero. If we took such individuals into account, then the green bar corresponding to unit 3 would be significantly wider.

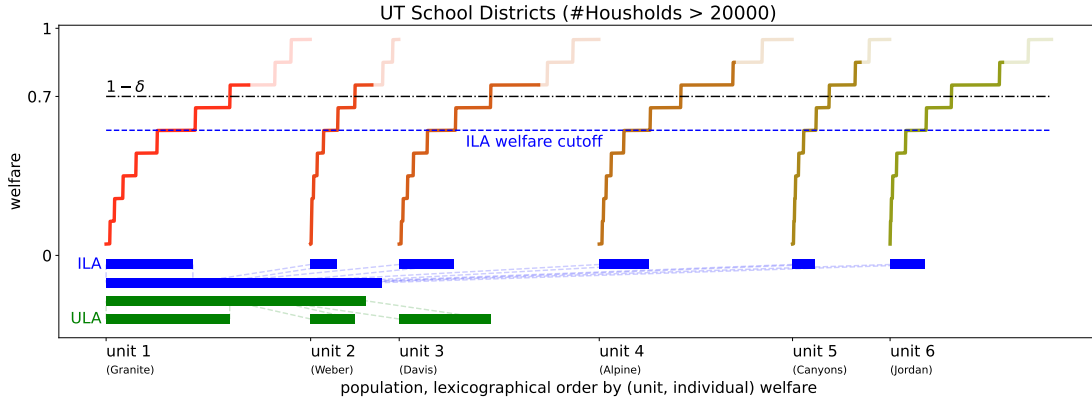


Figure 8: Unlike Figure 3, ILA outperforms ULA in a real-world low inequality setting. Utah’s school districts with a household population of 20,000 or more are considered. Household income is used as a proxy for individual welfare, categorized into 10 income brackets. A within-unit allocation with $q = 0.3$ and $q' \approx 0$ is considered, and individuals among the top q fraction at each unit appear faded. The values obtained by ILA and ULA from each unit are depicted with horizontal bars, and their overall values are compared.

D Further on Learning Unit-Level Statistics

D.1 Estimating Unit-Level Statistics by Sampling From All Units

Suppose we estimate unit-level statistics at each unit by sampling the welfare of n individuals at each unit. These samples are classified as type 2 because they contain true welfare values. Let \mathcal{S}_k denote the set of sampled individuals from unit k . Our estimate of the true unit-level statistic $U_k := \frac{1}{N} \sum_{i \in \mathcal{U}_k} u(w_i)$ will be $\hat{U}_k := \frac{1}{n} \sum_{i \in \mathcal{S}_k} u(w_i)$. Given that the samples are randomly drawn, we can approximate \hat{U}_k with $\mathcal{N}(U_k, \sigma^2/n)$. Since ULA only considers the ranking of these statistics, we can normalize any u to the interval $[0, 1]$, implying $\sigma^2 \leq 1/4$.

Suppose ULA's value is $(N\delta')$ -Lipschitz in unit-level statistics, which holds for $u(w) = \tau(w)/\delta'$ for example. Here, consistent with Section 6, δ' denotes the maximum treatment effect an individual may experience. This property allows us to bound the potential decrease in V_{unit} due to inaccurate estimations by

$$|\Delta V_{\text{unit}}| \leq N\delta' \sum_{k \in [M]} |\hat{U}_k - U_k|.$$

Hence, in expectation with respect to the randomness in sampling, we have

$$\mathbb{E}[|\Delta V_{\text{unit}}|] \leq MN\delta' \mathbb{E}[|\mathcal{N}(0, 1/(4n))|] = 2MN\delta' \int_0^\infty \sqrt{\frac{2n}{\pi}} x \exp(-2nx^2) dx = \frac{MN\delta'}{\sqrt{2\pi n}}.$$

On the other hand, collecting n type 2 samples from each unit reduces the treatment budget by $Mn c_2$. Assuming $c_2 \approx c$, this implies Mn fewer individuals will receive treatment. Consequently, we can bound the reduction of V_{unit} due to estimation cost by $Mn\delta'$. Therefore, we can bound the overall reduction of ULA's value, accounting for both inaccurate estimation and the additional cost of estimation, by

$$\mathbb{E}[|\Delta V_{\text{unit}}|] \leq Mn\delta' + \frac{MN\delta'}{\sqrt{2\pi n}}.$$

Optimally, choosing $n = \Theta(N^{2/3})$ implies $\mathbb{E}[|\Delta V_{\text{unit}}|] = O(MN^{2/3}\delta')$. Since for an efficient allocation we expect $V_{\text{unit}} = \Theta(KN\delta')$, the relative decrease of ULA's value is bounded by $O(\frac{M}{KN^{1/3}})$. For this effect to become negligible asymptotically, the number of treated units should exceed $\Omega(M/N^{1/3})$.

D.2 Learning Unit-Level Statistics Is No Harder Than Learning Individual Welfare

An individual-level predictor uses a set of features, denoted by x , to make a prediction about welfare w . The features include observations about the individual as well as any relevant contextual information accessible to the predictor. Consider a joint distribution \mathcal{D} over $\mathcal{X} \times \mathcal{W}$, where \mathcal{X} is the feature space and $\mathcal{W} = [0, 1]$. Assume n_2 independent samples \mathcal{S} are drawn from \mathcal{D} . These samples go under type 2 data because they contain the true welfare values. For a choice of loss function $l : \mathcal{W} \times \mathcal{W} \rightarrow [0, 1]$, the risk and the empirical risk of h are defined as

$$\begin{aligned} R(h) &:= \mathbb{E}_{(x,w) \sim \mathcal{D}} [l(h(x), w)], \\ \hat{R}(h) &:= \frac{1}{|\mathcal{S}|} \sum_{(x,w) \in \mathcal{S}} l(h(x), w). \end{aligned}$$

To learn a predictor with low risk, the standard approach is to minimize the empirical risk. If empirical risk minimization can be performed perfectly over a hypothesis class \mathcal{H} consisting of functions with range in \mathcal{W} , then based on a well-known result (see Mohri et al. (2018), Theorem 11.8), the risk of the empirical risk minimizer \hat{h} can be bounded with probability at least $1 - 2\alpha$ by

$$R(\hat{h}) \leq \min_{h \in \mathcal{H}} R(h) + \sqrt{\frac{8d \log(en_2/d)}{n_2}} + \sqrt{\frac{2 \log(1/\alpha)}{n_2}} = \min_{h \in \mathcal{H}} R(h) + \tilde{O}\left(\sqrt{\frac{d}{n_2}} + \sqrt{\frac{\log(1/\alpha)}{n_2}}\right). \quad (19)$$

Here d is the pseudo-dimension (Mohri et al. (2018), Definition 11.5) of $\mathcal{G} = \{(x, w) \rightarrow l(h(x), w) : h \in \mathcal{H}\}$. Let $l(\hat{w}, w) = (\hat{w} - w)^2$. Collecting $n_2 = \Omega(d/\epsilon^2)$ samples ensures the second term of the bounded risk will be $\tilde{O}(\epsilon)$. Then \hat{h} could have a mean squared error of $\tilde{O}(\epsilon)$ if $\min_h R(h)$ is sufficiently small. The best we can hope for this term is the Bayes risk from the Bayes predictor $h_B(x) = \mathbb{E}[w|x]$. Unfortunately, in many practical settings, limited relevant observations from individuals and their intrinsic unpredictability can lead to a large Bayes risk, making accurate prediction challenging. Next, we discuss why this concern is less relevant for estimating unit-level statistics.

Consider a unit-level statistic that can be expressed as the average of a strictly monotone bounded function $u(\cdot)$ evaluated at w_i s. Note that since only the ranking of such statistics matters in ULA, we can normalize any u to be in $[0, 1]$. Redefine the loss as $l(\hat{w}, w) := (u(\hat{w}) - u(w))^2$. For strictly monotone convex or concave u , one can verify that the pseudo-dimension based on the new l is similar to the pseudo-dimension based on the $(\hat{w} - w)^2$ loss. We may assume that minimizing the empirical risk based on the new loss function is computationally as hard as minimizing the empirical risk based on the original $(\hat{w} - w)^2$ loss. Given this assumption, a similar argument demonstrates that the risk of the empirical risk minimizer \hat{h} can be bounded by Equation (19). Using \hat{h} , we can estimate the unit-level statistic at unit k by calculating $\frac{1}{N} \sum_{i \in \mathcal{U}_k} u(\hat{h}(x_i))$. Our primary interest lies in the L_2 risk of the estimated unit-level statistics. Using the Jensen's inequality, we can bound this with the risk of \hat{h} :

$$\mathbb{E} \left[\frac{1}{M} \sum_{k \in [M]} \left(\frac{1}{N} \sum_{i \in \mathcal{U}_k} u(\hat{h}(x_i)) - \frac{1}{N} \sum_{i \in \mathcal{U}_k} u(w_i) \right)^2 \right] \leq \mathbb{E} \left[\frac{1}{MN} \sum_{k \in [M]} \sum_{i \in \mathcal{U}_k} (u(\hat{h}(x_i)) - u(w_i))^2 \right] = R(\hat{h}). \quad (20)$$

This shows learning unit-level statistics is no harder than learning individual-level welfare.

Suppose \hat{h} is additionally unbiased: $\mathbb{E}[u(\hat{h}(x)) | w] = u(w)$. Then a straightforward calculation shows that the upper bound of Equation (20) is loose and the L_2 risk of the estimated unit-level statistics can be bounded by $\frac{1}{N} R(\hat{h})$. Moreover, $R(\hat{h}) \leq R(h_B) + \tilde{O}(\sqrt{d/n_2})$. Note that the Bayes predictor in this case outputs $h_B(x) = u^{-1}(\mathbb{E}[u(w)|x])$. One can see for a large number of individuals per unit the Bayes risk is less relevant here. When this is the case, $n_2 = \Omega(d/(N\epsilon)^2)$ samples are sufficient to get a mean-squared error of $\tilde{O}(\epsilon)$ which is significantly less number of samples compared to learning an individual-level predictor with a similar error.

In sum, assuming the efficient learnability of individual welfare, learning unit-level statistics can be achieved using similar tools as for predicting individual welfare but in a more efficient way possibly requiring fewer assumptions about the intrinsic unpredictability of individuals. In the following, we show learning unit-level statistics can be done directly based on type 1 observations and without an individual welfare predictor under the hood.

D.3 Learning Unit-Level Statistics Under Noise-Free Observations $o = f(w)$

In the simplest case, observation o is an unknown deterministic function of welfare w . We assume f is a strictly increasing function of w , however, similar results hold true for a strictly decreasing function. Learning a function from observation to individual welfare requires learning f^{-1} . However, this is unnecessary for estimating unit-level statistics. For instance, ρ_k is an average of threshold function $w \rightarrow \mathbb{1}\{w > 1 - \delta\}$ evaluated at w_i s. Since f is an increasing function, there exists a threshold t such that ρ_k can be calculated as an average of threshold function $o \rightarrow \mathbb{1}\{o > t\}$ evaluated at o_i s. Therefore, the problem of learning ρ_k s reduces to learning a single parameter t . The next theorem demonstrates that for every available budget B , by appropriately choosing the number of type 2 samples, we can estimate ρ_k s sufficiently accurately with minimal cost, such that our previous bounds on V_{unit} can be nearly realized with high probability.

Theorem D.1. *Given a set of welfare-observation pairs $\{w_i, o_i\}$, choose threshold \hat{t} as the largest o_i for which $w_i \leq 1 - \delta$. Estimate ρ_k at each unit as the average of $\mathbb{1}\{o > \hat{t}\}$. Consider ULA based on estimated ρ_k s, given a budget of $B \gg c$ for treating K units. Under Assumption 6.1 and 6.3, for any $0 < \alpha < 1/2$, by choosing an appropriate number of samples, the decrease of V_{unit} due to the extra step of sampling and*

learning is bounded with probability at least $1 - 2\alpha$ by

$$\Delta V_{\text{unit}} \geq -\sqrt{KN}\delta\left(\gamma/\underline{\gamma} + \ln(1/\alpha) + z_\alpha(KN)^{-1/4}(\gamma/\underline{\gamma})^{1/2}\right),$$

where z_α is the z -score at α significance level. This implies that, with probability at least $1 - 2\alpha$, the relative decrease of V_{unit} is bounded by

$$O\left(\sqrt{\frac{c}{B}}(\gamma/\underline{\gamma} + \ln(1/\alpha))\right).$$

See proof on page 48.

Theorem D.1 suggests learning ρ_k s under noise-free observations comes at the relative cost of $O(\sqrt{c/B})$ which we expect to be very small. In fact, this is not limited to ρ_k s. The next theorem shows that for a 1-Lipschitz u , a slightly weaker bound of $\tilde{O}(\sqrt{c/B})$ on the relative decrease of ULA's value holds in expectation.

Theorem D.2. *Consider ULA based on unit-level statistics defined as the average of a 1-Lipschitz function $u(\cdot)$ at each unit. Given a random set \mathcal{S} of welfare-observation pairs, we can reconstruct a quantized f^{-1} and use that to estimate unit-level statistics based on type 1 observations (the details come in the proof). Assume V_{unit} is N -Lipschitz in the unit-level statistics (for example, this is the case when unit-level statistics are average treatment effects). Given a budget of $B \gg c$ for treating K units, under Assumption 6.1, by an appropriately chosen number of samples, the expected decrease of V_{unit} due to the extra step of sampling and learning is bounded by*

$$\mathbb{E}_{\mathcal{S}}[\Delta V_{\text{unit}}] \geq -3\sqrt{KN} - \frac{\delta'}{2\underline{\gamma}}\sqrt{KN} \ln(KN),$$

where δ' is the maximum treatment effect an individual may experience. This implies that, the expected relative decrease of V_{unit} is bounded by

$$\tilde{O}\left(\sqrt{\frac{c}{B}}(1/\underline{\gamma} + 1/\delta')\right).$$

See proof on page 49.

D.4 Learning Unit-Level Statistics Under Noisy Observations $o = f(w, z)$

Estimation Under Known Welfare-To-Observation Likelihood. In the simplest model, suppose the function f and the distribution of noise z denoted by p_z are known. Also, suppose welfare distribution can be represented as a member of a parametric statistical model parameterized by a few-dimensional parameter θ . Marginalizing over z , we obtain a statistical model with parameter θ over the space of observation. Maximizing likelihood, under mild assumptions, we can obtain a consistent estimator $\hat{\theta}_k$ for unit k with true parameter θ_k . Assuming there are n_1 observations available from this unit, $\hat{\theta}_k$ will have a covariance matrix of $\frac{I^{-1}(\theta_k)}{n}$, where $I(\theta_k)$ is the Fisher information matrix. We can then analytically calculate expected T'_k or ρ_k for this unit. For instance, denoting the cumulative density function of welfare associated with θ as P_θ , the estimated $\hat{\rho}_k$ will be consistent with a variance of $\nabla_\theta P_\theta(1 - \delta)^\top \frac{I^{-1}(\theta_k)}{n} \nabla_\theta P_\theta(1 - \delta)$. For a regular statistical model, we can assume $I(\theta_k)$ is well-conditioned (its smallest eigenvalue is $\Omega(1)$) and P_θ is $O(1)$ -Lipschitz. These are sufficient conditions to guarantee $\|I(\theta_k)^{-1/2} \nabla_\theta P_\theta\|_2^2 = O(1)$. Then, $|\mathbb{E}_w[\hat{\rho}_k] - \mathbb{E}_w[\rho_k]| = O(\frac{1}{\sqrt{n_1}})$ with high probability. In fact, for each individual, there may already be a few type 1 samples almost freely available. Then, when $n_1 = \Omega(N)$, the $O(\frac{1}{\sqrt{n_1}})$ error in estimating $\mathbb{E}_w[\rho_k]$ will be smaller or on par with the unavoidable $O(\frac{1}{\sqrt{N}})$ finite sample error in $|\rho_k - \mathbb{E}_w[\rho_k]|$, making $\hat{\rho}_k$ optimal.

In summary, when the welfare-to-observation likelihood is known and the Fisher information of the specified statistical model is well-conditioned, no type 2 samples or learning is required to estimate unit-level statistics. It is worth noting that a known likelihood does not necessarily imply that observations can fully resolve uncertainty about individual welfare. In fact, for each individual, only a few observations may be

relevant at the time of intervention. For instance, a student might have only a few grades early in their school when interventions are most effective. Furthermore, individual welfare is subject to change, and previous years' grades may not be helpful due to potential changes in the student's life circumstances. However, these issues are less relevant when considering units as a whole.

Learning Under Limited Distortion of Welfare Distribution. The following proposition provides an example where noisy observations of welfare are available with an unknown noise distribution, but the noise does not strongly distort the observations. In this case, by paying a fixed cost, we can obtain an ϵ -accurate predictor for unit-level statistics.

Proposition D.3. *Consider noisy observation $o = f(w, z)$ where z is an independent randomness (noise). Assume there exists a scaled and shifted version of observation \tilde{o} such that $\text{EMD}(p_w, p_{\tilde{o}}) \leq \epsilon$, where EMD is earth mover's distance. Suppose $u(\cdot)$ is 1-Lipschitz and there are many type 1 observations available from each unit for free. By observing average treatment effects from $1/\epsilon^2$ units, we can obtain a predictor from observations to unit-level statistic that has an average absolute error of $O(\epsilon)$.*

Proof. For a 1-Lipschitz $u(\cdot)$, Kantorovich-Rubinstein duality implies $|\mathbb{E}_w[u(w)] - \mathbb{E}_{\tilde{o}}[u(\tilde{o})]| \leq \text{EMD}(p_w, p_{\tilde{o}}) \leq \epsilon$. Therefore, when the input to u is appropriately shifted and scaled, the estimated unit-level statistic $\hat{U}_k := \mathbb{E}_{\tilde{o}}[u(\tilde{o}) | \mathcal{U}_k]$ will remain in the ϵ -radius of the true statistic $U_k := \mathbb{E}_w[u(w) | \mathcal{U}_k]$. We can find the scale and shift by solving a linear regression: Assume U_k s are available from m units denoted by \mathcal{S} . Solve the regression to find the scale and shift that minimizes $\sum_{k \in \mathcal{S}} |\hat{U}_k - U_k|$. Such estimates will have an L_1 -norm risk of $O(\epsilon + \sqrt{1/m})$. Choosing $m = 1/\epsilon^2$ will complete the proof. \square

Learning Under a Nonparametric Model. Next, we show efficient learning of unit-level statistics is possible under more general nonparametric models. Assume that observation $o \in \mathcal{O}$ is an unknown noisy function of w : $o = f(w, z)$. Units differ in how welfare is distributed within each unit, but the function f and the distribution of noise z are common across units. We denote the density of welfare and observation at unit k by p_w^k and p_o^k , respectively. For demonstration purposes, we assume that \mathcal{O} is a closed interval of \mathbb{R} with $O(1)$ length.

We assume the distortion from the noise is limited in the sense that there exists $v : \mathcal{O} \rightarrow \mathbb{R}$ such that for every unit k we have $\mathbb{E}_{p_w^k}[u(w)] \approx \mathbb{E}_{p_o^k}[v(o)]$ while $\max_o |v(o)| = O(\max_w |u(w)|)$. Consider a Hilbert space $\mathbb{H} = \mathbb{R}^{\mathcal{O}}$ equipped with an inner product $\langle p, q \rangle_{\mathbb{H}} = \int_{\mathcal{O}} p(o)q(o)do$, for $p, q \in \mathbb{H}$. Define a kernel $\mathcal{K} : [M] \times [M] \rightarrow \mathbb{R}$ by its feature mapping $\phi : [M] \rightarrow \mathbb{H}$ where $\phi(k) = p_o^k$. Hence $\mathcal{K}(k, k') = \int_{\mathcal{O}} p_o^k(o)p_o^{k'}(o)do$.

Since only the ranking of statistics matters in ULA, we can normalize any u to be in $[0, 1]$. Our assumptions on the observation model ensure that there exists $v^* \in \mathbb{H}$ with $\|v^*\|_{\infty} = \Lambda = O(1)$ such that the unit-level statistic at unit k can be computed by $h^*(k) = \langle v^*, \phi(k) \rangle_{\mathbb{H}} = \mathbb{E}_{p_w^k}[u(w)]$. Note that h^* is linear in v^* . Consider hypothesis class

$$\mathcal{H} = \{h : k \rightarrow \langle v_h, \phi(k) \rangle_{\mathbb{H}}, \|v_h\|_{\mathbb{H}} \leq \Lambda\}. \quad (21)$$

The next assumption enables efficient learning of v_h such that $h(k)$ induces an accurate ranking of units, as presented in a follow-up lemma.

Assumption D.4. We limit the concentration of p_o^k by assuming $r := \sup_{k \in [M]} \mathcal{K}(k, k)$ is $O(1)$.

Lemma D.5 (Mohri et al. (2018), Corollary 10.2). *Consider n independent samples \mathcal{S} drawn from an arbitrary joint distribution \mathcal{D} over $[M] \times [M]$. For $\beta > 0$ and a sample (k, k', y) , where $y = \chi\{h^*(k) > h^*(k')\}$, define the β -margin loss $l_{\beta}(y(h(k) - h(k'))) := \min\{1, \max\{0, 1 - y(h(k) - h(k'))/\beta\}\}$. Also, define the zero-one risk and empirical β -margin risk by*

$$R(h) := \mathbb{E}_{(k, k', y) \sim \mathcal{D}} [l_{0+}(y(h(k) - h(k')))],$$

$$\hat{R}_{\beta}(h) := \frac{1}{n} \sum_{(k, k', y) \in \mathcal{S}} l_{\beta}(y(h(k) - h(k'))).$$

Under Assumption D.4, with probability at least $1 - \alpha$, for every h in the hypothesis class \mathcal{H} of Equation (21), we have

$$R(h) \leq \hat{R}_\beta(h) + \frac{4r}{\beta} \sqrt{\frac{1}{n}} + \sqrt{\frac{\log(1/\alpha)}{n}},$$

where $r = O(1)$ by assumption.

Direct application of this lemma is not possible since we only have finite-sample estimates of $h^*(k)$ and $h(k)$. Let z_α be the z-score of α significance level. If there are n_1 type 1 samples available for calculating $h(k)$, then $\hat{h}(k) = \frac{1}{n_1} \sum_{j, o^{(j)} \sim p_o^k} v_h(o^{(j)})$ will be in $z_\alpha \sqrt{\text{Var}(v_h)/n_1}$ radius of $h(k)$ with probability at least $1 - \alpha$. Observing n_2 type 2 samples from each unit, we can estimate $h^*(k)$ with $\tilde{h}^*(k) = \frac{1}{n_2} \sum_{j, w^{(j)} \sim p_w^k} u(w^{(j)})$ up to an error of $z_\alpha \sqrt{\text{Var}(u)/n_2} \leq z_\alpha \sqrt{1/(4n_2)}$ with probability at least $1 - \alpha$. Using these finite-sample estimates, we can present the following lemma.

Lemma D.6 (Extension of Lemma D.5). *Given β , n_2 , and α , consider n independent samples drawn from an arbitrary joint distribution \mathcal{D} over $([M] \times [M]) \setminus \{(k, k') : |h^*(k) - h^*(k')| < \max\{\beta, z_\alpha/\sqrt{n_2}\}\}$. Let h be the minimizer of empirical β -margin risk calculated with n_1 type 1 and n_2 type 2 samples. Under Assumption D.4, with probability at least $1 - 8\alpha$, we have*

$$R(h) \leq \frac{4z_\alpha}{\beta} \sqrt{\frac{\sqrt{r}\Lambda}{n_1}} + \frac{4r}{\beta} \sqrt{\frac{1}{n}} + \sqrt{\frac{\log(1/\alpha)}{n}}.$$

Proof. Consider a sample (k, k') from \mathcal{D} . Our finite-sample evaluation of this sample's label based on type 2 data is $\tilde{y} = \chi\{\tilde{h}^*(k) > \tilde{h}^*(k')\}$ where $\tilde{h}^*(k) = \frac{1}{n_2} \sum_{j, w^{(j)} \sim p_w^k} u(w^{(j)})$. Since \mathcal{D} only encompasses (k, k') such that $|h^*(k) - h^*(k')| > z_\alpha/\sqrt{n_2}$ and $|\tilde{h}^*(k) - h^*(k)| \geq z_\alpha \sqrt{1/4n_2}$ with probability at least $1 - \alpha$, we can conclude $\tilde{y} = y$ with probability at least $1 - 2\alpha$.

Let $\hat{h}^*(k) = \frac{1}{n_1} \sum_{j, o^{(j)} \sim p_o^k} v^*(o^{(j)})$ be the empirical evaluation of $h^*(k)$ based on type 1 observations. For (k, k') drawn from \mathcal{D} we know $|h^*(k) - h^*(k')| \geq \beta$. Since $|\hat{h}^*(k) - h^*(k)| \leq z_\alpha \sqrt{\text{Var}(v^*)/n_1}$ with probability at least $1 - \alpha$, we can argue $l_\beta(y(\hat{h}^*(k) - \hat{h}^*(k'))) \leq \frac{2z_\alpha}{\beta} \sqrt{\text{Var}(v^*)/n_1}$ with probability at least $1 - 2\alpha$.

Putting it together, the finite-sample evaluation of empirical margin risk of h^* will be bounded by $\frac{2z_\alpha}{\beta} \sqrt{\text{Var}(v^*)/n_1}$ with probability at least $1 - 4\alpha$. Let h be the minimizer of the finite-sample evaluation of empirical margin risk. A similar argument shows the true empirical margin risk of h , denoted by $\hat{R}_\beta(h)$, cannot differ from its finite-sample evaluation more than $\frac{2z_\alpha}{\beta} \sqrt{\text{Var}(v_h)/n_1}$ with probability at least $1 - 4\alpha$. Therefore, since h was the minimizer of the finite-sample evaluation of empirical margin risk, we have

$$\hat{R}_\beta(h) \leq \frac{2z_\alpha}{\beta} \sqrt{\frac{\text{Var}(v^*)}{n_1}} + \frac{2z_\alpha}{\beta} \sqrt{\frac{\text{Var}(v_h)}{n_1}}, \quad (22)$$

with probability at least $1 - 8\alpha$. By definition of the hypothesis class, $\|v_h^2\|_{\mathbb{H}} \leq \Lambda$. Then for any unit k , Assumption D.4 and Cauchy-Schwarz inequality imply

$$\text{Var}(v_h) \leq \max_k \mathbb{E}_{p_o^k}[v_h^2(o)] = \max_k \langle \phi(k), v_h^2 \rangle_{\mathbb{H}} \leq \max_k \|\phi(k)\|_{\mathbb{H}} \|v_h^2\|_{\mathbb{H}} \leq \sqrt{r}\Lambda.$$

Since $h^* \in \mathcal{H}$, the same bound holds on $\text{Var}(v^*)$. Using these variance bounds to further bound $\hat{R}_\beta(h)$ in Equation (22), and plugging this into Lemma D.5 complete the proof. \square

Thus far, we have shown that efficiently ranking units is possible for a general bounded $u(\cdot)$. The following assumption will enable us to connect the ranking risk with the ULA's value in subsequent analysis.

Assumption D.7. Denote the unit-level statistics at unit k by U_k . We assume U_k s are not very concentrated: There exists $\lambda = O(1)$ such that $\Pr(U_k \in [a, a + b]) \leq \lambda b$.

Since type 1 data is abundant and almost freely available, we neglect finite-sample issues with n_1 and assume $c_1 \approx 0$. The next theorem shows that by appropriately choosing the number of samples and margin β , the relative cost of learning an accurate ranking of units will be negligible for a sufficiently large budget.

Theorem D.8. *Suppose ULA treats K units and ULA's value is $(N\delta')$ -Lipschitz in unit-level statistics. By appropriate choice of the algorithm and the number of samples as detailed in the proof, the relative decrease in ULA's value due to the cost of learning a ranking of units is $O(K^{-8/9}M^{2/3}N^{-2/9})$ with high probability, where M is the total number of units and N is the number of individuals per unit. As a result, for $K = \Omega(M^{3/4}/N^{1/4})$, the effect of learning starts to become negligible asymptotically.*

Proof. From m random units, sample n_2 true welfare from each unit. Without loss of generality, assume ULA allocates resources to the units with the smallest statistics. We assume the ξ -quantile of the unit-level statistics is known and denoted by $U_{(\xi)}$. One can later verify that quantiles can effectively be estimated with our sampled data without incurring a significant additional cost. Define $\Delta := \max\{\beta, z_\alpha/\sqrt{n_2}\}$. Targeting K units, define \mathcal{D} as the product of two distributions \mathcal{D}_1 and \mathcal{D}_2 , where \mathcal{D}_1 is a uniform distribution over any unit with a statistic less than $U_{(K/M)}$ and \mathcal{D}_2 is a uniform distribution over any unit with a statistic more than $U_{(K/M)} + \Delta$. Since Assumption D.7 limits the concentration of U_k s, the m random samples from the units give $n = \Theta(\frac{K}{M}m^2)$ samples from \mathcal{D} .

We minimize the empirical β -margin risk of Lemma D.6 using finite-sample evaluations of welfare. Since by definition of \mathcal{D} this problem is realizable, assuming empirical risk minimization can be performed perfectly, we have $\hat{R}_\beta(h) = 0$. Next, we bound $R(h)$. For notational brevity, we normalize ULA's value by $N\delta'$, where δ' is the maximum treatment effect an individual may experience.

- For K' units ($K' \leq K$) in the support of \mathcal{D}_2 to substitute K' units in the support of \mathcal{D}_1 , at least K'^2 pairwise misranking is required. Since every pairwise misranking at least contributes $1/(KM)$ to the risk, we have $K' \leq \sqrt{KMR(h)}$, where Lemma D.6 requires $R(h) = O(\frac{1}{\beta} \sqrt{\frac{M}{Km^2}})$ with high probability. Therefore, the change in ULA's value due to misranking is $O(K^{1/4}M^{3/4}m^{-1/2}\beta^{-1/2})$.
- Assuming $c_2 \approx c$, the cost of sampling wastes a budget equivalent to treating mn_2 individuals. Hence, the change in ULA's (normalized) value due to the sampling cost is $O(mn_2/N)$.
- Finally, assuming ULA's (normalized) value is 1-Lipschitz in unit-level statistics, the change in ULA's value due to misranking of individuals with their statistic in $[U_{(K/M)}, U_{(K/M)} + \Delta]$, is $O(\Delta(M\lambda\Delta)) = O(M\Delta^2) = O(M\beta^2 + M/n_2)$.

Adding these all together, the decrease in ULA's (normalized) value is bounded by

$$O(K^{1/4}M^{3/4}m^{-1/2}\beta^{-1/2} + mn_2/N + M\beta^2 + M/n_2).$$

Choose $\beta = K^{1/10}M^{-1/10}m^{-1/5}$ and $n_2 = \sqrt{MN/m}$. Then, the above bound reduces to

$$O(K^{1/5}M^{4/5}m^{-2/5} + \sqrt{Mm/N}).$$

Finally, choose $m = K^{2/9}M^{1/3}N^{5/9}$. Then the above bound simplifies to

$$O(K^{1/9}M^{2/3}N^{-2/9}).$$

Since ULA's (normalized) value is $\Theta(K)$, the relative change in ULA's value is bounded by $O(K^{-8/9}M^{2/3}N^{-2/9})$. Hence, this effect becomes negligible when K exceeds $\Omega(M^{3/4}/N^{1/4})$. \square

In a comparison with naive sampling from all units (Appendix D.1), a direct calculation shows that Theorem D.8 provides more efficient learning if $N = O(M^3)$. In particular, for $M = \Theta(N)$, the effect of learning a ranking according to Theorem D.8 becomes negligible when $K = \Omega(\sqrt{M})$, while naive sampling from all units requires $K = \Omega(M^{2/3})$. It is also worth noting that the guarantee in Theorem D.8 holds with high probability, which is a stronger result than the one presented in Appendix D.1 that holds in expectation.

Theorem D.8 has a pessimistic view of where incorrect ranking can occur. Optimistically, if for any pair of units, the probability of mistake is $O(R(h))$, then a simple calculation shows sampling from $O(\log(M))$ units is enough to get a sufficiently accurate ranking for ULA.

E Additional Statements

Proposition E.1. *In ILA with an ϵ -accurate predictor, if there are at least $I = \lfloor \frac{B-p(\epsilon)}{c} \rfloor$ individuals with a welfare of $(1 - \delta - 2\epsilon)$ or less, the ILA's maximum value of*

$$\max \left\{ \frac{B - p(\epsilon)}{c}, 0 \right\} \delta$$

can be realized.

Proof. To see this, first, observe that for the I individuals with the lowest welfare, the estimated welfare will be less than or equal to $(1 - \delta - \epsilon)$. This ensures the I^{th} estimated welfare after sorting, $\hat{w}_{s(I)}$, will be less than or equal to $(1 - \delta - \epsilon)$. Therefore, the true welfare of $s(1), \dots, s(I)$ should be less than or equal to $(1 - \delta)$. For such a group of targeted individuals, the treatment effect is at its maximum of δ . \square

Theorem E.2 (Sufficient conditions for a dominant ULA in case of a sublinear budget). *Consider ULA with a (q, q') -within-unit allocation and a budget $B = o(C)$. Assume $q' \leq (1 - \bar{\rho})q$. Define normalized inequality $\hat{G}_\rho := G_\rho / (1 - \bar{\rho}) \in [0, 1]$, and consider inequality threshold*

$$\hat{G}_\rho^{(1)} := 1 - \frac{q}{\bar{\rho}} \frac{1 - \bar{\rho}}{1 - q}.$$

- *If $q > \bar{\rho}$, within-unit allocation is effective and ULA achieves $(1 - q'/(1 - q))$ of the maximal value of $B \frac{\delta}{c}$. This implies $(1 + O(q')) \cdot \text{ULA} \succeq \text{ILA}$.
If further $p(\epsilon)/B > q'/(1 - q)$, we have $\text{ULA} \succ \text{ILA}$.*
- *If $q \leq \bar{\rho}$ but a minimal inequality is present as $\hat{G}_\rho > \hat{G}_\rho^{(1)}$, within-unit allocation is still effective and similar results hold true as the previous case.*
- *If none of the above conditions hold, but $p(\epsilon)$ consumes a significant portion of the budget as $p(\epsilon) > B \frac{q_c - (q - q')}{1 - q}$, we have $\text{ULA} \succ \text{ILA}$.*

Proof. First of all, note that for a sublinear budget, the condition $K \leq M(1 - \bar{\rho})$ of Lemma 4.2 (and so Corollary 4.4) is met. In the first two cases, within-unit allocation is effective, and Corollary 4.4 shows ULA at max loses a value of $B \frac{\delta}{c} \left(\frac{q'}{1 - q} \right)$ compared to the optimal. On the other hand, Equation (6) shows ILA at least loses a value of $p(\epsilon) \frac{\delta}{c}$. Therefore, ULA's value will exceed ILA if $B \frac{q'}{1 - q} < p(\epsilon)$. For a least effective within-unit allocation, Corollary 4.4 shows that ILA would lose a maximum value of $B \frac{\delta}{c} \left(\frac{q_c - (q - q')}{1 - q} \right)$. Therefore, a similar argument shows that ULA will dominate ILA if $B \left(\frac{q_c - (q - q')}{1 - q} \right) < p(\epsilon)$. \square

Proposition E.3. *If 1) τ is strictly decreasing, 2) $\tau \in \mathbb{C}^1$, 3) τ is either convex or concave, and 4) $\frac{d\tau}{dw}$ is convex, then $r(w)$ of Property 1 will be increasing in w for $w_c = 0$ and every valid choice of a and b . Therefore, under these conditions, τ won't have Property 1.*

Proof. Since τ is strictly decreasing, for every choice of $b > 0$, $r(w)$ is well-defined. Denote the derivative of τ by τ' . Taking the derivative of r w.r.t. w and rearranging the terms, we obtain

$$\frac{1}{(\tau(w) - \tau(w + b))^2} \left\{ \tau'(w)(\tau(w + a) - \tau(w + b)) - \tau'(w + a)(\tau(w) - \tau(w + b)) + \tau'(w + b)(\tau(w) - \tau(w + a)) \right\}.$$

The sign of the derivative is determined by its numerator, denoted by $n(w)$. Define

$$\begin{aligned} c_1(w) &:= \tau(w + a) - \tau(w + b), \\ c_2(w) &:= \tau(w) - \tau(w + a). \end{aligned}$$

Then, the numerator of the derivative can be written as

$$n(w) = c_1(w)\tau'(w) + c_2(w)\tau'(w+b) - (c_1(w) + c_2(w))\tau'(w+a).$$

Note that a strictly decreasing τ and $b \geq a > 0$ implies c_1 and c_2 are positive. Then, for a convex τ' , Jensen's inequality implies

$$n(w) \geq (c_1(w) + c_2(w))\tau'\left(w + b\frac{c_2(w)}{c_1(w) + c_2(w)}\right) - (c_1(w) + c_2(w))\tau'(w+a). \quad (23)$$

- For a concave τ and $b \geq a > 0$, the Jensen's inequality gives $\tau(w+a) \geq (1-a/b)\tau(w) + (a/b)\tau(w+b)$. This allows us to write

$$\frac{c_2(w)}{c_1(w) + c_2(w)} = \frac{\tau(w) - \tau(w+a)}{\tau(w) - \tau(w+b)} \leq \frac{\tau(w) - (1-a/b)\tau(w) - (a/b)\tau(w+b)}{\tau(w) - \tau(w+b)} = \frac{a}{b}. \quad (24)$$

Plugging this into Equation (23) and using the fact that τ' is decreasing since τ is concave, we obtain

$$\tau'\left(w + b\frac{c_2(w)}{c_1(w) + c_2(w)}\right) \geq \tau'(w+a). \quad (25)$$

- For a convex τ and $b \geq a > 0$, the Jensen's inequality gives $\tau(w+a) \leq (1-a/b)\tau(w) + (a/b)\tau(w+b)$. So, Equation (24) holds in the reverse direction. Plugging this into Equation (23) and using the fact that τ' is increasing since τ is convex, we again obtain Equation (25).

Since in both cases Equation (25) holds true, Equation (23) implies $n(w) \geq 0$, and the proof is complete. \square

Theorem E.4. Consider ULA with a (q, q') -within-unit allocation and a budget B that is no more than the cost of treating $M\bar{T}'/\delta'$ units. Under Assumption 6.3, suppose τ has Property 1 for w_c , $a = (1-q-q')/\gamma$, and $b = (1-2q')/\gamma$. Define $T'_c := \gamma\Gamma_b(w_c) + q'\delta'$ and two inequality thresholds

$$\tilde{G}_T^{(1)} := 1 - \frac{\bar{T}'}{T'_c} \leq \tilde{G}_T^{(2)} := 1 - \frac{1}{4}\frac{\bar{T}'}{T'_c} - \frac{3}{4}\frac{\bar{T}'}{\delta'}. \quad (26)$$

Then, if any of the following conditions holds, for every treated unit k , we will have $T'_k \geq T'_c$:

- If $G_T > \tilde{G}_T^{(2)}$ and the budget is not excessively large (not surpassing the cost of treating $M\bar{T}'/2\delta'$ units).
- If $G_T > \tilde{G}_T^{(1)}$ and the budget does not exceed

$$\frac{B}{C} < (1-q)\frac{\bar{T}' - T'_c + \max\{\bar{T}' - T'_c, G_T T'_c\}}{2(\delta' - T'_c)}. \quad (27)$$

Then, defining $T'_{\text{eff}} := \bar{T}'/(1-G_T)$ and neglecting $O(B^2/C^2)$, if any of the above conditions holds, we have

$$V_{\text{unit}} \geq KQ_\tau(T'_{\text{eff}} - q'\delta'; q, q', \gamma).$$

Proof. Denote the profile of all T'_k 's by \mathbf{T}' . Let $s(\cdot)$ be the ascending sort of T'_k 's. For notational simplicity, assume $s(k) = k$ without loss of generality. We also drop the dependency of $Q_\tau(\cdot; q, q', \gamma)$ on q, q' , and γ as these are the constants of the problem. Lemma 6.4 guarantees that unit k if treated, provides a value of at least $Q_\tau(T'_k - q'\delta')$. Then, treating K units ensures a value of

$$V_{\text{unit}} \geq \tilde{V}(\mathbf{T}') := \sum_{k=M-K+1}^M Q_\tau(T'_k - q'\delta').$$

Next, we look for \mathbf{T}' minimizing $\tilde{V}(\mathbf{T}')$ while satisfying a mean of \bar{T}' and a Gini index of G_T . The proof has two parts: First, we find sufficient conditions under which all the treated units have a T' larger than or equal to

$$T'_c := \gamma \Gamma_{(1-2q')/\gamma}(w_c) + q' \delta'.$$

Then, for such units, we further lower bound $\tilde{V}(\mathbf{T}')$.

Ensuring for All Treated Units, $T'_k \geq T'_c$. As the first step, we relax the problem, allowing for unit profiles with a Gini index of G_T or more. This requires

$$2 \sum_{k \in [M]} k T'_k \geq M(M+1) \bar{T}' + M^2 \bar{T}' G_T. \quad (28)$$

Consider two units l and u ($u > l$). For a fixed mean, the Gini index or equivalently the left-hand side of Equation (28) can be increased if we increase T'_u and decrease T'_l by the same amount. Now assume there are K_+ units with $T'_k \geq T'_c$. Neglecting two units at max, for a fixed K_+ , the Gini index will be maximized when units are concentrated on the boundaries T'_{\min} , T'_c , and T'_{\max} . We denote the number of units on these boundaries by K_0 , K_1 , and K_2 . These numbers should be integer but by neglecting at max three units, we can assume they can take any nonnegative real value. We further relax the problem and set $T'_{\min} = 0$ and $T'_{\max} = \delta'$. We do a proof by contradiction. If any of the K treated units has $T'_k < T'_c$, we should have $K > K_+$. Since $K_+ \geq K_2$, necessarily $K > K_2$. The mean constraint requires $K_1 T'_c + K_2 \delta' = M \bar{T}'$. Using this and $K_0 + K_1 + K_2 = M$, we can solve for K_0 and K_1 in terms of K_2 :

$$K_1 = \frac{M \bar{T}' - K_2 \delta'}{T'_c}, \quad (29)$$

$$K_0 = \frac{M(T'_c - \bar{T}') + K_2(\delta' - T'_c)}{T'_c}. \quad (30)$$

Then, the nonnegativity of K_0 and K_1 requires

$$\frac{\bar{T}' - T'_c}{\delta' - T'_c} \leq \frac{K_2}{M} \leq \frac{\bar{T}'}{\delta'}. \quad (31)$$

The Gini index constraint also imposes a lower bound on K_2 . For a unit profile concentrated on the boundaries, Equation (28) can be written as

$$(K_0 + 1 + M - K_2) K_1 T'_c + (M - K_2 + 1 + M) K_2 \delta' \geq M(M+1) \bar{T}' + M^2 \bar{T}' G_T.$$

Plugging K_0 and K_1 from Equation (29) into this and doing a direct calculation, we obtain

$$-\left(\frac{K_2}{M}\right)^2 \frac{\delta'}{\bar{T}'} + 2\left(\frac{K_2}{M}\right) + \frac{T'_c(1 - G_T) - \bar{T}'}{\delta' - T'_c} \geq 0. \quad (32)$$

If $B = o(c)$ and consequently $K/M = o(1)$, asymptotically this condition requires $G_T \leq \tilde{G}_T^{(1)} := 1 - \bar{T}/T'_c$. Therefore, in the case of a sublinear budget, $G_T > \tilde{G}_T^{(1)}$ results in contradiction and ensures for no treated unit $T'_k < T'_c$. In general, since $K_2/M \leq \bar{T}'/\delta'$ (Equation (31)), the quadratic function of Equation (32) is increasing in the valid range of K_2 . Under the assumption $K/M \leq \bar{T}'/\delta'$, then $K > K_2$ requires Equation (32) to be holding for $K_2 = K$ as well. Now consider two cases:

- If the budget is not excessively large and $K/M \leq \bar{T}'/(2\delta')$, Equation (32) for $K_2 = K = \bar{T}'/(2\delta')$ requires

$$G_T \leq \tilde{G}_T^{(2)} := 1 - \frac{1}{4} \frac{\bar{T}'}{T'_c} - \frac{3}{4} \frac{\bar{T}'}{\delta'}.$$

Therefore, $G_T > \tilde{G}_T^{(2)}$ results in a contradiction.

- Neglecting the squared term in Equation (32), K/M should be at least $(\bar{T}' - T'_c(1 - G_T))/(2\delta' - 2T'_c)$, which is binding only if $G_T > \tilde{G}_T^{(1)} := 1 - \bar{T}'/T'_c$. Furthermore, Equation (31) and $K \geq K_2$ requires K/M to be also no less than $(\bar{T}' - T'_c)/(\delta' - T'_c)$. Putting these together,

$$\frac{K}{M} \geq \frac{1}{2} \frac{\bar{T}' - T'_c + \max\{\bar{T}' - T'_c, G_T T'_c\}}{\delta' - T'_c}.$$

Hence, a K less than the above amount results in a contradiction.

This completes the first part of the proof.

Lower Bounding $\tilde{V}(\mathbf{T}')$. We next lower bound $\tilde{V}(\mathbf{T}')$, under the conditions that $T'_k \geq T'_c$ for $k > M - K$, subject to the mean and Gini index constraints. As the first step, we further relax the Gini constraint of Equation (28) and require

$$2 \sum_{k=1}^{M-K} kT'_k + 2M \sum_{k=M-K+1}^M T'_k \geq M(M+1)\bar{T}' + M^2\bar{T}'G_T. \quad (33)$$

Let \mathbf{T}'^* be the optimal solution of the relaxed problem. We now construct a solution \mathbf{T}' from \mathbf{T}'^* that does not increase \tilde{V} while maintaining a similar mean and meeting the relaxed Gini constraint of Equation (33) with a lower or equal margin. As the first step, construct a new solution as

$$T'_k = \begin{cases} T_k'^*, & k \leq M - K, \\ \frac{1}{K} \sum_{k>M-K} T_k'^* & \text{o.w.} \end{cases}$$

This solution maintains the same mean and meets the relaxed Gini constraint with the same margin. We argue $\tilde{V}(\mathbf{T}') \leq \tilde{V}(\mathbf{T}'^*)$: First, we show $Q_\tau(t)$ is a convex function for $t \geq T'_c - q'\delta'$ or equivalently $T' \geq T'_c$. To see this, note that

$$\frac{dQ_\tau}{dt} = (\Gamma_b^{-1})'(t/\gamma) \Gamma'_a(\Gamma_b^{-1}(t/\gamma)) = \frac{\Gamma'_a(w)}{\Gamma'_b(w)} = \frac{\tau(w) - \tau(w+a)}{\tau(w) - \tau(w+b)}, \quad (34)$$

where $w = \Gamma_b^{-1}(t/\gamma)$. Then Property 1 implies $\frac{dQ_\tau}{dt}$ is decreasing in w or equivalently increasing in t for $w \leq w_c$ or equivalently $t = T' - q'\delta' \geq \gamma\Gamma_b(w_c) = T'_c - q'\delta'$. Since we have already shown $T'_k \geq T'_c$ for all of the top K units, Jensen's inequality implies $\tilde{V}(\mathbf{T}') \leq \tilde{V}(\mathbf{T}'^*)$. Therefore, if \mathbf{T}'^* was optimal, the constructed \mathbf{T}' should be optimal as well.

In the second step, choose any two units l, u from $[M-K]$ such that $T'_l < T'_u$. Decrease T'_l and increase T'_u for the same amount while the order of units is preserved. Such a change will increase the margin of the relaxed Gini index constraint but does not change the mean of the units and the objective. By repetitively applying this operation, we obtain an optimal solution where all units except for one are concentrated around either 0 or $T'_u := \frac{1}{K} \sum_{k>M-K} T'_k$. Hence, we can represent this solution by two numbers K_0 and K_1 which are the number the number of units with T'_k of 0 and T'_u . These numbers should be integers however neglecting the value of two units in maximum, we can assume they are nonnegative real numbers satisfying the sum constraint $K_0 + K_1 = M$. The mean constraint also requires $K_1 T'_u = M\bar{T}'$. Solving for K_0 and K_1 in terms of T'_u we obtain $K_1/M = \bar{T}'/T'_u$ and $K_0/M = 1 - \bar{T}'/T'_u$. The nonnegativity of K_0 and $K_1 \geq K$ imposes

$$\max\{\bar{T}', T'_c\} \leq T'_u \leq \frac{M}{K}\bar{T}'.$$

The relaxed Gini constraint also imposes another lower bound on T'_u . From Equation (33) we have

$$(K_0 + 1 + M - K)(K_1 - K)T'_u + 2MKT'_u \geq M(M+1)\bar{T}' + M^2\bar{T}'G_T.$$

Substituting K_0 and K_1 in terms of T'_u and simplifying equation, we obtain

$$\frac{K^2 - K}{M^2} \left(\frac{T'_u}{\bar{T}'} \right)^2 + (1 - G_T) \left(\frac{T'_u}{\bar{T}'} \right) - 1 \geq 0.$$

Then a direct calculation shows T'_u should be at least as large as

$$T'_u \geq \frac{2\bar{T}'}{1 - G_T + \sqrt{(1 - G_T)^2 + 2(K^2 - K)/M^2}}. \quad (35)$$

Using this lower bound in $\tilde{V}(\mathbf{T}') = KQ_\tau(T'_u - q'\delta')$ then completes the proof. □

F Missing Proofs

Theorem 6.7 (Sufficient conditions for a dominant ULA in case of a heterogeneous effect). *Consider ULA with a (q, q') -within-unit allocation and a budget B that is no more than the cost of treating $M\bar{T}'/\delta'$ units. Under Assumption 6.3 and conditions of Theorem 6.2, suppose an elbow-shaped effect (Figure 6) that satisfies Property 1 for $w_c = 1 - \delta - a$, $a = (1 - q - q')/\gamma$, and $b = (1 - 2q')/\gamma$. Then $T'_c \leq \delta' - q^2/2$. Consider inequality thresholds*

$$G_T^{(1)} := 1 - \frac{\bar{T}'}{\delta' - q^2/(2\gamma)} \leq G_T^{(2)} := 1 - \frac{1}{4} \frac{\bar{T}'}{\delta' - q^2/(2\gamma)} - \frac{3}{4} \frac{\bar{T}'}{\delta'}.$$

If either $G_T > G_T^{(2)}$, or $G_T > G_T^{(1)}$ and the budget meets Equation (36), then we have $T'_k \geq \delta$ for every treated unit k . Furthermore,

- Defining $g := (G_T - G_T^{(1)})/(1 - G_T)$ and

$$\alpha := 1 - \frac{1 - q - q'}{4\gamma} - \frac{1}{2(q - q')}g - \frac{\gamma}{4} \frac{1 - 2q}{(q - q')^3} g^2,$$

and neglecting $O(l^2)$, ULA achieves $(1 - q'/(1 - q))(1 - \alpha l/\delta)$ of the optimal value.

- Neglecting $O(q'^2)$, if

$$\frac{p(\epsilon)}{B} > \frac{q'}{1 - q} + \alpha \frac{l}{\delta} - \frac{l}{\delta} \frac{1}{8\gamma} \left(\frac{B}{C} + \underline{\gamma}\epsilon/2 \right),$$

we will further have ULA \succ ILA.

Proof. Since τ is l -Lipschitz when $w < 1 - \delta$, for $w_c = 1 - \delta - a$, we have

$$\begin{aligned} T'_c &= \gamma\Gamma_b(w_c) + q'\delta' = \gamma\Gamma_a(w_c) + \gamma(\delta - (b - a)/2)(b - a) + q'\delta' \\ &\leq \gamma\delta'a + \gamma(\delta' - (b - a)/2)(b - a) + q'\delta' = \gamma\delta'b - \gamma(b - a)^2/2 + q'\delta' \\ &\leq \delta' - q^2/(2\gamma). \end{aligned}$$

Plugging this into Equation (26) give inequality thresholds $G_T^{(1)}$ and $G_T^{(2)}$.

Examining Equation (27), if $\bar{T}' - T'_c \geq G_T T'_c$, then the upper bound on the budget decreases with increasing T'_c . On the other hand, if $\bar{T}' - T'_c < G_T T'_c$, utilizing the property of the Gini index that G_T cannot exceed $1 - \bar{T}'/\delta'$, one can verify that the upper bound is again decreasing in T'_c . Therefore, to satisfy Equation (27), it suffices to have

$$\begin{aligned} \frac{B}{C} &< (1 - q) \frac{\bar{T}' - \delta' + q^2/(2\gamma) + \max\{\bar{T}' - \delta' + q^2/(2\gamma), (\delta' - q^2/(2\gamma))G_T\}}{2(\delta' - \delta' + q^2/(2\gamma))} \\ &= \frac{1}{2}(1 - q)(2\gamma\delta'/q^2 - 1)(-G_T^{(1)} + \max\{-G_T^{(1)}, G_T\}). \end{aligned} \quad (36)$$

In the proof of Theorem E.4 (in particular, Equation (34)) we showed

$$\frac{dQ_\tau}{dt} = \frac{\tau(w) - \tau(w + a)}{\tau(w) - \tau(w + b)},$$

for $w = \Gamma_b^{-1}(t/\gamma)$. Then we argued Property 1 implies $Q_\tau(t)$ is a convex function for $t \geq T'_c - q'\delta'$. Now we

further show that Q_τ is strongly convex in the regime of interest where $w + b \geq w_c$ and $w + a \leq w_c$:

$$\begin{aligned}
\frac{d^2 Q_\tau}{dt^2} &= \frac{dw}{dt} \frac{d}{dw} \left(\frac{dQ_\tau}{dt} \right) \\
&= \frac{1 - \tau'(w)(\tau(w+a) - \tau(w+b)) + \tau'(w+a)(\tau(w) - \tau(w+b)) - \tau'(w+b)(\tau(w) - \tau(w+a))}{\gamma (\tau(w) - \tau(w+b))^3} \\
&\geq \frac{1}{\gamma} \frac{\frac{l}{2}(\tau(w+a) - \tau(w+b)) - l(\tau(w) - \tau(w+b)) + (\tau(w) - \tau(w+a))}{(\tau(w) - \tau(w+b))^3} \\
&= \frac{1}{\gamma} \frac{(\tau(w) - \tau(w+a)) - \frac{l}{2}(\tau(w) - \tau(w+b)) + O(l^2)}{(\tau(w) - \tau(w+b))^3} \\
&\geq \frac{l}{2\gamma} \frac{2a-b}{(b-a)^3} + O(l^2). \tag{37}
\end{aligned}$$

Here we denoted the derivative of τ by τ' and used $l \geq |\tau'(w)| \geq l/2$ for $w \leq w_c + a$ and $\tau'(w+b) = -1$. Neglecting $O(l^2)$, this second derivative is positive for $a > b/2$ or equivalently $q < 1/2$. We denote the lower bound of Equation (37) by β_2 . Using a similar technique, we can lower bound the first derivative as well:

$$\frac{dQ_\tau}{dt} \geq \frac{l}{2} \frac{a}{b-a} + O(l^2).$$

Neglecting $O(l^2)$, we denote this lower bound by β_1 . At $w = w_c$, we have

$$Q_\tau(T'_c - q'\delta') = \gamma \Gamma_a(w_c) \geq \gamma a(\delta + la/4).$$

Also note that for a nonnegative $\tilde{G}_T^{(1)}$, we have

$$T'_{\text{eff}} - T'_c = \frac{T'_c}{1 - G_T} \left(\frac{\bar{T}'}{T'_c} - 1 + G_T \right) \geq \bar{T}' \frac{G_T - \tilde{G}_T^{(1)}}{1 - G_T} \geq \bar{T}' \frac{G_T - G_T^{(1)}}{1 - G_T}.$$

Now we have all the pieces to lower bound $Q_\tau(T'_{\text{eff}} - q'\delta')$:

$$\begin{aligned}
Q_\tau(T'_{\text{eff}} - q'\delta') &\geq Q_\tau(T'_c - q'\delta') + \beta_1(T'_{\text{eff}} - T'_c) + \frac{1}{2}\beta_2(T'_{\text{eff}} - T'_c)^2 \\
&\geq \gamma a \left(\delta + \frac{la}{4} + \frac{\beta_1 \bar{T}' G_T - G_T^{(1)}}{\gamma \bar{T}' (1 - G_T)} + \frac{\beta_2}{2\gamma} \left(\bar{T}' \frac{G_T - G_T^{(1)}}{1 - G_T} \right)^2 \right).
\end{aligned}$$

Plugging β_1 , β_2 , a , and b into the above equation and using $K \approx \frac{B}{Nc(1-q)}$ in Theorem E.4 will complete the proof of the near-optimality of ULA.

We next drop $O(q'^2)$ from our calculation and obtain sufficient conditions for the dominance of ULA. If ULA is not dominant, necessarily,

$$(B - p(\epsilon)) \frac{\delta'}{c} > B \frac{\delta'}{c} (1 - O(l, q')).$$

This requires $p(\epsilon)/B = O(l, q')$. For such a small price of prediction, using Theorem 6.2, it is sufficient to have

$$B\delta' \left(1 - \frac{q'}{1-q} - \alpha \frac{\delta}{\delta'} \right) > (B - p(\epsilon)) \left(\delta' - \frac{l}{8\gamma} (B/C + \underline{\gamma}\epsilon/2) \right) \approx B\delta' - p(\epsilon)\delta - B \frac{l}{8\gamma} (B/C + \underline{\gamma}\epsilon/2).$$

Rearranging the terms and solving for $p(\epsilon)/B$ will prove the stated sufficiency condition. \square

Proof of Lemma 4.1. Recall that within an intervened unit, resources will be allocated to the lower $(1 - q)$ fraction of individuals, potentially including a q' fraction from the top q cut. Depending on the relative size of q to ρ_k , two possibilities arise:

- $\rho_k \leq q$: In the worst case, the top ρ_k portion may erroneously receive the resources in place of the population in need. However, if $q' \leq \rho_k$, only a proportion of size q' may be misclassified. The treatment on these individuals may have zero effect. Hence, $T_k/\delta \geq 1 - q - \min\{\rho_k, q'\}$.
- $\rho_k > q$: Since at least $(q - q')$ fraction of top individuals rightfully will not receive the resources, at most $\rho_k - (q - q')$ fraction of top individuals will be among treated. These individuals might experience a zero effect. Therefore, we have $T_k/\delta \geq 1 - q - (\rho_k - (q - q'))$.

In both cases, we can write the lower bound as $T_k/\delta \geq 1 - q - \text{loss}(\rho_k)$, where loss is defined in Equation (7). \square

Proof of Lemma 4.2. Given a unit profile $\boldsymbol{\rho}$, let $s(\cdot)$ be the ascending sorting of ρ_k s. For notational simplicity, assume $s(k) = k$ without loss of generality. The largest possible loss of any unit profile $\boldsymbol{\rho}$ with a mean of $\bar{\rho}$ and Gini index of G_ρ is the solution of the optimization problem

$$\max_{\boldsymbol{\rho}} \sum_{k \in [K]} \text{loss}(\rho_k) \quad (38)$$

$$\text{s.t.} \quad \sum_{k \in [M]} \rho_k = M\bar{\rho}, \quad (\text{Mean constraint})$$

$$\sum_{k \in [M]} k\rho_k - \frac{M(M+1)}{2}\bar{\rho} - \frac{M^2}{2}\bar{\rho}G_\rho = 0, \quad (\text{Gini index constraint}) \quad (39)$$

$$1 \geq \rho_k \geq 0, \quad (\text{Valid range constraint})$$

$$\rho_{k+1} \geq \rho_k, \text{ for } k \in [M-1]. \quad (\text{Valid sorting constraint})$$

The second constraint (Equation (39)) is obtained by rearranging terms in the Gini index formula. Since $\text{loss}(\cdot)$ is a piecewise linear function and the constraints are all linear functions of $\boldsymbol{\rho}$, the above optimization problem is a linear program. We upper bound the solution to this problem in the following.

As the first step, we relax the Gini index constraint (Equation (39)) to include every unit profile with a Gini index of G_ρ or more:

$$\sum_{k \in [M]} k\rho_k - \frac{M(M+1)}{2}\bar{\rho} - \frac{M^2}{2}\bar{\rho}G_\rho \geq 0. \quad (40)$$

The solution of the linear program with this relaxed constraint will serve as an upper bound for $\text{loss}_K(\boldsymbol{\rho})$.

Depending on the magnitude of ρ_K (corresponding to the last treated unit) compared to q and q' , three regimes can be defined:

1. $\rho_K < q'$
2. $q' \leq \rho_K < q$
3. $q \leq \rho_K$

For a fixed K , we upper bound the solution of the linear program in each regime and take the maximum of these three upper bounds as a bound for the original problem.

Second Regime. Let ρ^* be an optimal solution in the second regime where $q' \leq \rho_K^* < q$. We show a new solution ρ can be constructed from ρ^* with a similar mean, larger or equal Gini index, and larger or equal loss. Hence, this solution should also be optimal. Let $l = \min \{k : \rho_k^* > q'\}$ and $u = \max \{k : \rho_k^* < 1\}$. Define $\Delta = \min \{\rho_l^* - q', 1 - \rho_u^*\}$. This is the maximum amount that we can move ρ_l^* back and ρ_u^* forward while keeping them in the $[q', 1]$ range. Define the new values $\rho_l = \rho_l^* - \Delta$, $\rho_u = \rho_u^* + \Delta$, and $\rho_k = \rho_k^*$ for $k \notin \{l, u\}$. The new solution ρ admits the same mean as ρ^* , preserves the ordering, and increases the margin of the Gini index constraint (Equation (40)) by $(u - l)\Delta \geq 0$. Further, since $\rho_K^* < q$, we already know for $l \leq k \leq K$, $\text{loss}(\rho_k)$ is the minimum loss in range $[q', 1]$ and this change can only increase $\text{loss}(\rho)$. Therefore, ρ should also be an optimal solution. By repetitively applying this operation, we can obtain an optimal solution ρ such that no more than one unit lies in $(q', 1)$. During this process, ρ_K may enter the third regime, which is fine, as it shows the loss of the solution in the third regime is larger. In the rest of the proof, we continue in the second regime where $\rho_K < q$.

Let us call the constructed unit profile, which we know is optimal, ρ^* . We build on the same idea and further manipulate ρ^* to obtain another optimal unit profile ρ . Let $l = \min \{k : \rho_k^* > 0\}$ and $u = \max \{k : \rho_k^* < q'\}$. Define $\Delta = \min \{\rho_l^*, q' - \rho_u^*\}$. This is the maximum amount that we can move ρ_l^* back and ρ_u^* forward while keeping them in the $[0, q']$ range. Define the new values $\rho_l = \rho_l^* - \Delta$, $\rho_u = \rho_u^* + \Delta$, and $\rho_k = \rho_k^*$ for $k \notin \{l, u\}$. The new solution ρ again has the same mean as ρ^* , preserves the ordering, and increases the margin of the Gini index constraint by $(u - l)\Delta \geq 0$. Moreover, since $\text{loss}(\cdot)$ is linear in the $[0, q']$ range and $u \leq K$, this operation does not change the objective function. Therefore, if ρ^* was optimal, ρ also remains optimal. By repetitively applying this operation, we can obtain an optimal solution ρ such that no more than one unit admits $0 < \rho_k < q'$.

Neglecting a maximum of two units, in the optimal solution ρ constructed so far, ρ_k only takes a value from $\{0, q', 1\}$. Hence, we can represent this solution by three numbers K_0 , K_1 , and K_2 which are the number of units with ρ_k of 0, q' , and 1, respectively. These numbers should be integers; however, neglecting the value of three units in maximum, we can assume they are nonnegative real numbers satisfying the sum constraint $K_0 + K_1 + K_2 = M$. The mean constraint for this solution is

$$K_1 q' + K_2 = M \bar{\rho}.$$

Solving for K_2 , we have $K_2 = M \bar{\rho} - K_1 q'$. The nonnegativity of K_2 requires $K_1 \leq M \bar{\rho} / q'$. Plugging K_2 into the sum constraint and solving for K_0 , we obtain $K_0 = M(1 - \bar{\rho}) - K_1(1 - q')$. The nonnegativity of K_0 requires $K_1 \leq M(1 - \bar{\rho}) / (1 - q')$. The Gini index constraint for ρ is

$$\sum_{k \in [M]} k \rho_k = \frac{1}{2}(K_0 + 1 + K_0 + K_1)K_1 q' + \frac{1}{2}(M - K_2 + 1 + M)K_2 \geq \frac{1}{2}M(M + 1)\bar{\rho} + \frac{1}{2}M^2 \bar{\rho} G_\rho.$$

Plugging K_0 and K_2 into this and simplifying equations, we obtain

$$\frac{K_1}{M} \leq \sqrt{\frac{\bar{\rho}(1 - \bar{\rho} - G_\rho)}{q'(1 - q')}}.$$

Note that for values bounded between 0 and 1 with a mean of $\bar{\rho}$, the Gini index cannot exceed $(1 - \bar{\rho})$. Putting these together,

$$0 \leq \frac{K_1}{M} \leq \min \left\{ \frac{\bar{\rho}}{q'}, \frac{1 - \bar{\rho}}{1 - q'}, \sqrt{\frac{\bar{\rho}(1 - \bar{\rho} - G_\rho)}{q'(1 - q')}} \right\}.$$

This imposes the following bounds on K_0 :

$$1 - \bar{\rho} \geq \frac{K_0}{M} \geq \max \left\{ 0, 1 - \bar{\rho} - \bar{\rho} \frac{1 - q'}{q'}, 1 - \bar{\rho} - \sqrt{\frac{1 - q'}{q'}} \bar{\rho}(1 - \bar{\rho} - G_\rho) \right\}. \quad (41)$$

In regime 2, we have $q' \leq \rho_K < q$. This requires $K_0 \leq K \leq K_0 + K_1$. Substituting K_1 in terms of K_0 and performing a direct calculation, we obtain the following upper bounds on K_0 :

$$\frac{K_0}{M} \leq \frac{K}{M}, \quad \frac{K_0}{M} \leq \frac{1 - \bar{\rho}}{q'} - \frac{K}{M} \frac{1 - q'}{q'}. \quad (42)$$

For a solution to exist in regime 2, the upper bounds of Equation (42) should be larger than the lower bound of Equation (41). One can verify that for $K \leq 1 - \bar{\rho}$, the upper bound in the right-hand side of Equation (42) is larger than the lower bound. Hence, there remains one condition for the solution of regime 2 to exist:

$$\frac{K}{M} \geq 1 - \bar{\rho} - \min \left\{ \bar{\rho} \frac{1 - q'}{q'}, \sqrt{\frac{1 - q'}{q'} \bar{\rho} (1 - \bar{\rho} - G_\rho)} \right\}.$$

In regime 2, we can write the objective as

$$\text{loss}_K(\boldsymbol{\rho}) = (K - K_0)q'.$$

Maximizing the objective then corresponds to minimizing K_0 subject to Equation (41). It can be verified that when $q' < q_c = \bar{\rho} \frac{1 - \bar{\rho} - G_\rho}{1 - \bar{\rho} - \bar{\rho} G_\rho}$, the lower bound in Equation (41) is zero, resulting in $\text{loss}_K(\boldsymbol{\rho}) = Kq'$. When $q' \geq q_c$, the lower bound is at least $K_0 \geq 1 - \bar{\rho} - \sqrt{\bar{\rho} (1 - \bar{\rho} - G_\rho) (1 - q') / q'}$. For K_0 equal to this lower bound, $\text{loss}_K(\boldsymbol{\rho}) \leq (K - K_0)q'$.

Third Regime. Let $\boldsymbol{\rho}^*$ be an optimal solution in the third regime where $q \leq \rho_K^*$. We follow a similar approach and show that a new optimal solution $\widetilde{\boldsymbol{\rho}}$ can be constructed from $\boldsymbol{\rho}^*$ that admits a specific structure. First of all, Let us upper bound $\text{loss}(\rho_k)$ with $\widetilde{\text{loss}}(\rho_k) = \max \{q', \rho_k - (q - q')\}$. We will consider maximizing $\widetilde{\text{loss}}_K(\boldsymbol{\rho}) := \sum_{k \in [K]} \widetilde{\text{loss}}(\rho_k)$ instead of $\text{loss}_K(\boldsymbol{\rho})$ in the linear program of Equation (38). Fix ρ_K^* . Let $l = \min \{k : k > 0\}$ and $u = \max \{k : \rho_k^* < \rho_K^*\}$. Define $\Delta = \min \{\rho_l^*, \rho_K^* - \rho_u^*\}$. This is the maximum amount we can move ρ_l^* back and ρ_u^* forward while keeping them in the $[0, \rho_K^*]$ range. Define the new values $\rho_l = \rho_l^* - \Delta$, $\rho_u = \rho_u^* + \Delta$, and $\rho_k = \rho_k^*$ for $k \notin \{l, u\}$. The new solution $\boldsymbol{\rho}$ admits the same mean as $\boldsymbol{\rho}^*$, preserves the ordering, and increases the margin of the Gini index constraint by $(u - l)\Delta \geq 0$. Further, since $\widetilde{\text{loss}}(\cdot)$ is convex in $[0, \rho_K^*]$, this change can only increase the loss. Therefore, $\boldsymbol{\rho}$ should be optimal as well. By repetitively applying this operation, we can obtain an optimal solution $\boldsymbol{\rho}$ such that no more than one unit falls in $0 < \rho_k < \rho_K^* = \rho_K$.

Let us call the constructed unit profile $\boldsymbol{\rho}^*$. Again, fix ρ_K^* . Let $l = \min \{k : k > \rho_K^*\}$ and $u = \max \{k : \rho_k^* < 1\}$. Define $\Delta = \min \{\rho_l^* - \rho_K^*, 1 - \rho_u^*\}$. This is the maximum amount we can move ρ_l^* back and ρ_u^* forward while keeping them in the $[\rho_K^*, 1]$ range. Define the new values $\rho_l = \rho_l^* - \Delta$, $\rho_u = \rho_u^* + \Delta$, and $\rho_k = \rho_k^*$ for $k \notin \{l, u\}$. This new solution has the same mean as $\boldsymbol{\rho}^*$ with a higher Gini index. It also leaves the objective function unaffected. Hence, the constructed solution is also optimal. By repetitively applying this operation, we can obtain an optimal solution $\boldsymbol{\rho}$ such that for no more than one unit $\rho_K^* < \rho_k < 1$.

For notational simplicity, denote ρ_K^* by ρ , a free parameter in the constructed solution. Neglecting a maximum of two units, ρ_k in the constructed optimal solution can only take a value from $\{0, \rho, 1\}$. So, we can represent this solution with ρ and three integers K_0, K_1 , and K_2 which are the number of units with ρ_k equal to 0, ρ , and 1, respectively. Although these numbers are integers, we can treat them as nonnegative real numbers, neglecting the effect of three units in maximum. We require $K_0 + K_1 + K_2 = M$. The mean constraint also requires

$$K_1 \rho + K_2 = M \bar{\rho}.$$

Solving these two equations for K_1 and K_2 in terms of K_0 , we obtain $K_1 = (M(1 - \bar{\rho}) - K_0)/(1 - \rho)$ and $K_2 = (M(\bar{\rho} - \rho) + K_0 \rho)/(1 - \rho)$. The nonnegativity of K_1 and K_2 imposes a bound on K_0 : $1 - \bar{\rho}/\rho \leq K_0/M \leq 1 - \bar{\rho}$. The Gini index constraint requires

$$\sum_{k \in [M]} k \rho_k = \frac{1}{2}(K_0 + 1 + K_0 + K_1)K_1 \rho + \frac{1}{2}(M - K_2 + 1 + M)K_2 \geq \frac{1}{2}M(M + 1)\bar{\rho} + \frac{1}{2}M^2 \bar{\rho} G_\rho.$$

Plugging K_1 and K_2 as a function of K_0 into this constraint and simplifying equations, we obtain

$$\frac{K_0}{M} \geq 1 - \bar{\rho} - \sqrt{\frac{1 - \rho}{\rho}} \sqrt{\bar{\rho} (1 - \bar{\rho} - G_\rho)}.$$

Since we are in the third regime, $K_0 \leq K$, and $K \leq K_0 + K_1$. The latter imposes another upper bound on K_0 : $K_0 \leq M(1 - \bar{\rho})/\rho - K(1 - \rho)/\rho$. Altogether, the following bounds are in place on K_0 :

$$\min \left\{ \frac{K}{M}, 1 - \bar{\rho}, \frac{1 - \bar{\rho}}{\rho} - \frac{K}{M} \frac{1 - \rho}{\rho} \right\} \geq \frac{K_0}{M} \geq \max \left\{ 0, 1 - \frac{\bar{\rho}}{\rho}, 1 - \bar{\rho} - \sqrt{\frac{1 - \rho}{\rho}} \sqrt{\bar{\rho}(1 - \bar{\rho} - G_\rho)} \right\}. \quad (43)$$

One can verify that for $K \leq 1 - \bar{\rho}$, the second and third term of the upper bound in the left-hand side of Equation (43) are always larger than or equal to the lower bound. Hence, for the solution of the third region to exist, it is required $K/M \geq \max \{1 - \bar{\rho}/\rho, 1 - \bar{\rho} - \sqrt{\bar{\rho}(1 - \bar{\rho} - G_\rho)(1 - \rho)/\rho}\}$. After a direct calculation, this imposes two upper bounds on ρ :

$$\rho \leq \frac{\bar{\rho}}{1 - \frac{K}{M}}, \quad \rho \leq \bar{\rho} \frac{1 - \bar{\rho} - G_\rho}{1 - \bar{\rho} - \bar{\rho}G_\rho - 2\frac{K}{M}(1 - \bar{\rho} - \frac{K}{2M})}. \quad (44)$$

Define

$$q_u(K/M) := \min \left\{ \frac{\bar{\rho}}{1 - \frac{K}{M}}, \bar{\rho} \frac{1 - \bar{\rho} - G_\rho}{1 - \bar{\rho} - \bar{\rho}G_\rho - 2\frac{K}{M}(1 - \bar{\rho} - \frac{K}{2M})} \right\}. \quad (45)$$

Then the upper bounds of Equation (44) can be summarized as $\rho \leq q_u(K/M)$. Since we are in regime 3 and $\rho \geq q$, a necessary condition in this regime is $q \leq q_u(K/M)$.

The upper bounded loss for the constructed $\boldsymbol{\rho}$ in regime 3 is

$$\widetilde{\text{loss}}_K(\boldsymbol{\rho}) = K_0 q' + (K - K_0)(\rho - q + q') = (K - K_0)(\rho - q) + K q'. \quad (46)$$

For a fixed ρ , to maximize $\widetilde{\text{loss}}_K(\boldsymbol{\rho})$, we have to minimize K_0 . From Equation (43) we can see

$$\frac{K_0}{M} \geq \begin{cases} 1 - \bar{\rho} - \sqrt{\frac{1 - \rho}{\rho}} \sqrt{\bar{\rho}(1 - \bar{\rho} - G_\rho)}, & \rho > q_c, \\ 0, & \text{o.w.}, \end{cases} \quad (47)$$

where

$$q_c := \bar{\rho} \frac{1 - \bar{\rho} - G_\rho}{1 - \bar{\rho} - \bar{\rho}G_\rho}.$$

Note the tradeoff in the choice of ρ : Equation (46) indicates that a larger ρ leads to a larger loss. However, Equation (47) shows that a larger ρ also results in a larger K_0 , reducing the loss. Consider two possibilities:

- $\rho < q_c$: Since $\rho \geq q$ in regime 3, we should have $q < q_c$. The lower bound of Equation (47) is not binding in this case, and K_0 can be as low as 0. Since $\rho \leq q_c$, a direct result of Equation (46) is

$$\widetilde{\text{loss}}_K(\boldsymbol{\rho}) < K(q_c - q + q').$$

- $\rho \geq q_c$: If $B = o(C)$, we have $K/M = o(1)$ and neglecting $o(1)$, the second upper bound of Equation (44) reduces to q_c . Plugging $K_0 = 0$ and $\rho = q_c + o(\bar{\rho})$ into Equation (46), we obtain

$$\widetilde{\text{loss}}_K(\boldsymbol{\rho}) \leq K(q_c - q + q').$$

If $B = \Theta(C)$ and consequently $K/M = \Theta(1)$, we show if $q \leq q_c/2$, the maximum loss occurs at $\rho = q_c$. To show this, substitute K_0 with its lower bound of Equation (47) in $\widetilde{\text{loss}}_K(\boldsymbol{\rho})$:

$$\widetilde{\text{loss}}_K(\boldsymbol{\rho}) = \left(K - M(1 - \bar{\rho}) + M \sqrt{\frac{1 - \rho}{\rho}} \sqrt{\bar{\rho}(1 - \bar{\rho} - G_\rho)} \right) (\rho - q) + K q'.$$

Take a derivative of $\widetilde{\text{loss}}_K$ with respect to ρ and simplify the equations:

$$\frac{\partial}{\partial \rho} \widetilde{\text{loss}}_K = K - M(1 - \bar{\rho}) + M \sqrt{\frac{1 - \rho}{\rho}} \sqrt{\bar{\rho}(1 - \bar{\rho} - G_\rho)} \left(1 - \frac{1}{2} \frac{1 - q/\rho}{1 - \rho} \right). \quad (48)$$

First of all, observe that $\sqrt{(1-\rho)/\rho}$ and the last term inside the parentheses are decreasing in ρ , hence $\widetilde{\text{loss}}_K$ is concave in the feasible range of ρ . Define

$$q_l(K/M) := q_c u(K/M). \quad (49)$$

We next show if $q \leq q_l(K/M)$, then $\frac{\partial \widetilde{\text{loss}}_K}{\partial \rho}$ is nonpositive at the left boundary where $\rho = q_c$. To observe this, plug $\rho = q_c$ into Equation (48) and simplify the equations:

$$\frac{\partial}{\partial \rho} \widetilde{\text{loss}}_K|_{\rho=q_c} = K - \frac{M}{2} \left(1 - \frac{q}{q_c}\right) \frac{1 - \bar{\rho} - \bar{\rho} G_\rho}{1 - \bar{\rho}}.$$

Then, one can verify that the above derivative is nonpositive for $q \leq q_l(K/M)$. Combining this with our observation of concavity shows that the maximum of $\widetilde{\text{loss}}_K$ occurs at $\rho = q_c$. For this choice of ρ , we have

$$\widetilde{\text{loss}}_K(\boldsymbol{\rho}) < K(q_c - q + q').$$

First Regime. Following a similar argument as the third regime, we can show there exists an optimal solution $\boldsymbol{\rho}$ such that neglecting two units, every other unit has ρ_k from $\{0, \rho, 1\}$, where $\rho \leq q'$. We can again represent this solution with K_0, K_1 , and K_2 corresponding to the number of units with $\rho_k = 0, \rho$, and 1 , respectively. One can show exactly similar upper bounds on ρ (Equation (44)) and a similar lower bound for K_0 (Equation (47)). In the first regime, the objective at $\boldsymbol{\rho}$ can be written as

$$\text{loss}_K(\boldsymbol{\rho}) = (K - K_0)\rho.$$

If $\rho < q_c$, the zero lower bound for K_0 can be achieved. In this case,

$$\text{loss}_K(\boldsymbol{\rho}) \leq K \min\{q_c, q'\}.$$

If $\rho \geq q_c$, it is necessary to have $q' \geq q_c$. In this case, we can upper-bound the objective by

$$\text{loss}_K(\boldsymbol{\rho}) \leq Kq'.$$

Worst of All Regimes. Finally, we consider the largest objective among all regimes as an upper bound on the optimal solution of the problem. For a sublinear budget $B = o(C)$,

- If $q \geq q' \geq q_c$, the largest loss comes from regime 1 where $\text{loss}_K(\boldsymbol{\rho}) \leq Kq'$.
- If $q \geq q_c > q'$, the largest loss comes from both regimes 1 and 2 where $\text{loss}_K(\boldsymbol{\rho}) \leq Kq'$.
- If $q_c > q \geq q'$, the largest loss comes from regime 3 where neglecting $o(1)$ we have $\text{loss}_K(\boldsymbol{\rho}) \leq K(q_c - q + q')$.

For a linear budget $B = \Theta(C)$, using the definitions of $q_l(\cdot)$ and $q_u(\cdot)$ from Equations (45) and (49)

- If $q > q_u(K/M)$ and $q' \geq q_c$, the largest loss comes from regime 1 where $\text{loss}_K(\boldsymbol{\rho}) \leq Kq'$.
- If $q > q_u(K/M) \geq q_c > q'$, the largest loss comes from both regimes 1 and 2 where $\text{loss}_K(\boldsymbol{\rho}) \leq Kq'$.
- If $q_l(K/M) \geq q \geq q'$, the largest loss comes from regime 3 where $\text{loss}_K(\boldsymbol{\rho}) \leq K(q_c - q + q')$.

This completes the proof. Note that in presenting the results in the main text, we used $q_c u(K/M) \geq q_u(K/M)$. □

Proof of Theorem 4.5. Recall from Lemma 4.2 that $\text{loss}_K(\rho) \leq Kq'$ if $q > q_u(K/M)$, where $q_u(K/M)$ is defined in Equation (45) as

$$q_u(K/M) := \min \left\{ \frac{\bar{\rho}}{1 - \frac{K}{M}}, \bar{\rho} \frac{1 - \bar{\rho} - G_\rho}{1 - \bar{\rho} - \bar{\rho}G_\rho - 2\frac{K}{M}(1 - \bar{\rho} - \frac{K}{2M})} \right\}. \quad (50)$$

Observe that $q_u(\cdot)$ is an increasing function as long as $K \leq M(1 - \bar{\rho})$. So, for $K \leq M(1 - \bar{\rho})/2$, in order to have $q > q_u(K/M)$, it suffices to have $q > q_u((1 - \bar{\rho})/2)$. In particular, it suffices to have q larger than the second term in the minimum in Equation (50):

$$q > \bar{\rho} \frac{1 - \bar{\rho} - G_\rho}{1 - \bar{\rho} - \bar{\rho}G_\rho - \frac{3}{4}(1 - \bar{\rho})^2}.$$

Rearranging the terms, we find this condition can be met by a sufficiently high inequality: $G_\rho/(1 - \bar{\rho}) > 1 - \frac{1}{4} \frac{q}{\bar{\rho}} \frac{1 - \bar{\rho}}{1 - q}$.

Even when inequality is not this high, a not-too-large budget can ensure $q > q_u(K/M)$. Solving for K/M in $q > q_u(K/M)$, we obtain

$$\frac{K}{M} < \max \left\{ 1 - \frac{\bar{\rho}}{q}, 1 - \bar{\rho} - \sqrt{(1 - \bar{\rho})^2 - (1 - \frac{q_c}{q})(1 - \bar{\rho} - \bar{\rho}G_\rho)} \right\}.$$

The first term in the maximum is positive only when $q > \bar{\rho}$. Similarly, the second term is also positive only when $q > q_c$. Since $q_c \leq \bar{\rho}$, a positive bound on K/M requires $q > q_c$ which is equivalent to a sufficiently high inequality: $G_\rho/(1 - \bar{\rho}) > \frac{1 - q/\bar{\rho}}{1 - q} = 1 - \frac{q}{\bar{\rho}} \frac{1 - \bar{\rho}}{1 - q}$. In this case, the second term of the maximum will be decreasing in G_ρ . Therefore, it suffices to set G_ρ to its maximum of $(1 - \bar{\rho})$. Then, plugging $K/M \approx \frac{B}{(1 - q)C}$ into this condition gives Equation (13).

In the third case, where inequality does not satisfy $G_\rho/(1 - \bar{\rho}) > \frac{1 - q/\bar{\rho}}{1 - q}$, within-unit allocation is least effective. In this case, Lemma 4.2 suggests that for $q \leq q_l(K/M) := q_c u(K/M)$, the loss will be bounded by $K(q_c - q + q')$ which gives a similar lower bound on V_{unit} as Equation (10). Then a similar argument as the third case of Theorem E.2 shows that ULA will dominate ILA if $p(\epsilon)/B > \frac{q_c - (q - q')}{1 - q}$. There remains to show $q \leq q_l(K/M)$. For instance, if the budget is not larger than the amount for treating $M(1 - \bar{\rho})/4$ units and $q \leq q_c/2$, one can verify that regardless of the level of inequality, we will have $q \leq q_l(K/M)$. In the worst case when such conditions don't hold, within-unit allocation may not be effective at all, and we can resort to a blind within-unit allocation corresponding to $q = q' = 0$. This gives the sufficient condition of $p(\epsilon)/B > q_c$ for ULA dominance. \square

Proof of Lemma 5.1. Consider the following unit profile:

$$\rho_k^* = \begin{cases} \rho_l, & k \leq K, \\ \rho_u, & k > K. \end{cases}$$

Here, ρ_k^* s are concentrated around a lower (ρ_l) and an upper value (ρ_u). For a unit profile with mean $\bar{\rho}$, we should have $K\rho_l + (M - K)\rho_u = M\bar{\rho}$. Solving for ρ_u , we obtain $\rho_u = (M\bar{\rho} - K\rho_l)/(M - K)$. A valid ρ_u lies in the range $1 \geq \rho_u \geq \rho_l$. One can verify the upper bound of 1 is already satisfied if $K \leq M(1 - \bar{\rho})$ and $\rho_l \geq 0$. A simple calculation also shows the lower bound is satisfied for $\rho_l \leq \bar{\rho}$. Therefore, for a valid choice of ρ_l and ρ_u , we just need $\bar{\rho} \geq \rho_l \geq 0$.

Our choice of ρ_l and ρ_u should also result in a Gini index of G_ρ or less:

$$\begin{aligned} & \sum_{k \in [M]} k\rho_k^* - \frac{M(M+1)}{2}\bar{\rho} - \frac{M^2}{2}\bar{\rho}G_\rho \leq 0 \\ \iff & K(K+1)\rho_l + (M+K+1)(M-K)\rho_u - M(M+1)\bar{\rho} - M^2\bar{\rho}G_\rho \leq 0. \end{aligned}$$

Plugging ρ_u in terms of ρ_l into this and simplifying equations, we obtain

$$\rho_l \geq \bar{\rho} - \frac{M}{K} G \bar{\rho}.$$

Since $\text{loss}_K(\boldsymbol{\rho}^*) = K \text{loss}(\rho_l)$, we are interested in the smallest ρ_l possible, which is either the above bound or 0. Note that this choice satisfies $\bar{\rho} \geq \rho_l \geq 0$, and results in Equation (14). \square

Proof of Theorem 5.4. As Lemma 5.1 suggests, if $G_\rho \geq K/M$, there exists $\boldsymbol{\rho}^* \in \mathcal{P}$ for which ULA's loss is zero. Assuming $K \leq M(1 - \bar{\rho})/2$, then it suffices to have $G_\rho/(1 - \bar{\rho}) \geq 1/2$ for an optimal ULA.

If inequality is not this high, we can still hope for a small loss if $\rho_{\text{eff}} := \bar{\rho}(1 - (M/K)G_\rho)$ is less than or equal to q . In this case, $\text{loss}(\rho_{\text{eff}}) \leq q$ and $\text{loss}_K(\boldsymbol{\rho}^*) \leq Kq'$. Rearranging terms, $\rho_{\text{eff}} \leq q$ can be interpreted as a condition on G_ρ : $G_\rho \geq (K/M)(1 - q/\bar{\rho})$. Using $K \leq M(1 - \bar{\rho})/2$, it suffices to have $G_\rho/(1 - \bar{\rho}) \geq (1 - q/\bar{\rho})/2$. Then for a loss of Kq' , the value of ULA will be at least as the lower bound in Equation (9). Therefore, ULA achieves $(1 - q'/(1 - q))$ of the maximal value of $B \frac{\delta}{c}$, and we get $\text{ILA} \not\leq (1 + O(q')) \cdot \text{ULA}$.

We can get rid of the $O(q')$ term in the previous case if $p(\epsilon)/B$ is sufficiently large. For $\rho_{\text{eff}} \leq q$, we can expand $\text{loss}(\rho_{\text{eff}})$ using Equation (7) and obtain

$$\text{loss}_K(\boldsymbol{\rho}^*) \leq K \min \{\rho_{\text{eff}}, q'\} = \min \{\bar{\rho}(K - MG_\rho), Kq'\}.$$

This corresponds to a reduction of

$$\min \left\{ B \frac{\bar{\rho}}{1 - q} - \bar{\rho} G_\rho C, B \frac{q'}{1 - q} \right\} \frac{\delta}{c}$$

from ULA's value. Since ILA would lose at least a value of $p(\epsilon) \frac{\delta}{c}$, if $p(\epsilon)/B > q'/(1 - q)$ or $p(\epsilon) > \bar{\rho}(B/(1 - q) - G_\rho C)$ holds, ULA will have a larger value than ILA for $\boldsymbol{\rho}^*$, and we have $\text{ILA} \not\leq \text{ULA}$.

Finally, if $\rho_{\text{eff}} > q$, within-unit allocation is not effective and we can use a blind within-unit allocation instead. For $q = q' = 0$, Equation (7) simply gives $\text{loss}(\rho_{\text{eff}}) = \rho_{\text{eff}}$. Then a similar argument as the previous case shows ULA would beat ILA for $\boldsymbol{\rho}^*$ if $p(\epsilon) > \bar{\rho}(B - G_\rho C)$. Since by assumption $K \leq M(1 - \bar{\rho})$, we have $C \geq B/(1 - \bar{\rho})$. Therefore, it is sufficient to have $p(\epsilon)/B > \bar{\rho}(1 - G_\rho/(1 - \bar{\rho}))$. \square

Proof of Theorem 6.2. The proof has two parts. First, for a fixed budget, we characterize the welfare density that maximizes V_{ind} subject to Assumption 6.1. Then for such a class of densities, we upper bound V_{ind} as a function of the budget. Throughout the proof, we denote the remaining budget by $\tilde{B} := B - p(\epsilon)$ and define $\tilde{\gamma} := \gamma - \underline{\gamma}$.

Given a welfare density p_w consistent with Assumption 6.1, define a new density $q_w(w) := p_w(w) - u \delta(w - w_1) + u \delta(w - w_2)$, where $\delta(\cdot)$ is Dirac delta function and $u \geq 0$ is chosen such that Assumption 6.1 is not violated. For a fixed w_t , increasing u from 0 changes the budget by

$$\frac{1}{C} \frac{\partial \tilde{B}}{\partial u} = -\sigma(w_t - w_1) + \sigma(w_t - w_2).$$

Without loss of generality, suppose $w_2 > w_1$. Then a nondecreasing σ implies $\frac{\partial \tilde{B}}{\partial u} \leq 0$. To spend the same budget, we need to increase w_t :

$$\frac{1}{C} \frac{\partial \tilde{B}}{\partial w_t} = \mathbb{E}[\sigma'(w_t - w)] \geq 0.$$

Since we calculate the derivative at $u = 0$, the expectation is w.r.t. $w \sim p_w$. Putting these together, a fixed spending requires

$$\frac{dw_t}{du} = \frac{\sigma(w_t - w_1) - \sigma(w_t - w_2)}{\mathbb{E}[\sigma'(w_t - w)]}.$$

When we increase w_t in this way, the (expected) value of ILA changes by

$$\begin{aligned} \frac{1}{MN} \frac{dV_{\text{ind}}}{du} &= -\tau(w_1)\sigma(w_t - w_1) + \tau(w_2)\sigma(w_t - w_2) + \frac{dw_t}{du} \mathbb{E}[\tau(w)\sigma'(w_t - w)] \\ &= -\tau(w_1)\sigma(w_t - w_1) + \tau(w_2)\sigma(w_t - w_2) + (\sigma(w_t - w_1) - \sigma(w_t - w_2)) \frac{\mathbb{E}[\tau(w)\sigma'(w_t - w)]}{\mathbb{E}[\sigma'(w_t - w)]}. \end{aligned} \quad (51)$$

Consider the special case of $w_2 = 1$. At this point $\tau(w_2) = 0$. We assume \tilde{B} is more than $\underline{\gamma}\epsilon$ and less than $\gamma(1 - \epsilon)$. This ensures for a tight choice of w_t , $1 - \epsilon \geq w_t \geq \epsilon$ which implies $\sigma(w_t) = 1$ and $\sigma(w_t - 1) = 0$. We also relax Assumption 6.1 at the right-most boundary and assume density can be concentrated at $w = 1$. Then Equation (51) shows a mass can move from w_1 to 1 while increasing the value if and only if $\tau(w_1) < \mathbb{E}[\tau(w)\sigma'(w_t - w)]/\mathbb{E}[\sigma'(w_t - w)]$. Therefore, we can characterize the value-maximizing p_w for $w < 1$ by $p_w(w) = \gamma \mathbb{1}\{w \leq a\} + \underline{\gamma} \mathbb{1}\{w > a\}$, where

$$\tau(a) = \frac{\mathbb{E}[\tau(w)\sigma'(w_t - w)]}{\mathbb{E}[\sigma'(w_t - w)]}.$$

This is of course a recursive formula and not the full characterization. However, it allows us to bound a in terms of w_t :

- Using the concavity of τ and the symmetry of σ' , a simple argument in terms of the worst-case σ' shows

$$\tau(a) \geq \frac{\underline{\gamma} \int_{w_t - \epsilon}^{w_t + \epsilon} \frac{1}{2\epsilon} \tau(w) dw + \bar{\gamma} \int_{w_t - \epsilon/2}^a \frac{1}{\epsilon} \tau(w) dw}{\underline{\gamma} + \bar{\gamma}(a - w_t - \epsilon/2)/\epsilon}. \quad (52)$$

Since $l \geq |\frac{d\tau}{dw}| \geq l/2$, we can further lower bound the integrals in the numerator:

$$\begin{aligned} \frac{1}{2\epsilon} \int_{w_t - \epsilon}^{w_t + \epsilon} \tau(w) dw &\geq \frac{1}{2} (\tau(w_t) - l\epsilon/2 + \tau(t) + l\epsilon/4) = \tau(w) - l\epsilon/8, \\ \frac{1}{\epsilon} \int_{w_t - \epsilon/2}^a \tau(w) dw &\geq (\tau(a) + l(a - w_t + \epsilon/2)/4)(a - w_t + \epsilon/2)/\epsilon. \end{aligned}$$

Plugging these into Equation (52) and simplifying equations, we obtain

$$\tau(a) - \tau(w_t) \geq -\frac{l\epsilon}{8} + \frac{\bar{\gamma} l}{\underline{\gamma} 4} \frac{(a - w_t + \epsilon/2)^2}{\epsilon}.$$

Using $\tau(a) - \tau(w_t) \leq l(w_t - a)$, we can further simplify the above equation and obtain

$$a \leq w_t - \alpha\epsilon,$$

where

$$\alpha := \frac{1}{2} + 2(\underline{\gamma}/\bar{\gamma}) - \sqrt{4(\underline{\gamma}/\bar{\gamma})^2 + (5/2)(\underline{\gamma}/\bar{\gamma})}.$$

Note that for $\gamma \geq 3\underline{\gamma}$, we have $\alpha \in [0, 1/2]$.

- Using Jensen's inequality for a concave τ and the symmetry of σ' , one can show

$$\tau(a) \leq \frac{\underline{\gamma}\tau(w_t) + \bar{\gamma} \int_{w_t - \epsilon}^a \frac{1}{2\epsilon} \tau(w) dw}{\underline{\gamma} + \bar{\gamma}(a - w_t - \epsilon)/(2\epsilon)}. \quad (53)$$

Since $l \geq |\frac{d\tau}{dw}| \geq l/2$, we can further upper bound the integral in the numerator:

$$\frac{1}{2\epsilon} \int_{w_t - \epsilon}^a \tau(w) dw \leq (\tau(a) + l(a - w_t + \epsilon)/2)(a - t + \epsilon)/(2\epsilon).$$

Plugging this into Equation (53) and simplifying equations, we obtain

$$\tau(a) - \tau(w_t) \leq \frac{\bar{\gamma} l (a - w_t + \epsilon)^2}{\underline{\gamma} 4 \epsilon}.$$

Using $\tau(a) - \tau(w_t) \geq l(w_t - a)/2$, we can further simplify the above equation and obtain

$$a \geq w_t - \beta \epsilon,$$

where

$$\beta := 1 + (\underline{\gamma}/\bar{\gamma}) - \sqrt{(\underline{\gamma}/\bar{\gamma})^2 + 2(\underline{\gamma}/\bar{\gamma})}.$$

Next, we upper bound V_{ind} for the partially characterized optimal p_w . Using $|\frac{d\tau}{dw}| \geq l/2$, for a nondecreasing σ , we have

$$\frac{1}{MN} V_{\text{ind}} \leq \tau(0) \mathbb{E}[\sigma(w_t - w)] - \gamma \frac{la^2}{4} \sigma(w_t - a) - \underline{\gamma} (w_t - a) \frac{la}{2} \sigma(0) - \underline{\gamma} \frac{lw_t}{2} \int_{w_t}^1 \sigma(w_t - w) dw.$$

When all of the budget \tilde{B} is spent, the first term is equal to $\delta' \tilde{B}/C$. Using $\sigma'(\Delta w) \geq 1/\epsilon$ and $w_t \leq 1 - \epsilon$, we can lower bound the integral by $\epsilon/8$. Since $a \leq w_t$, we also have $\sigma(w_t - a) \geq \sigma(0) = 1/2$. Plugging these into the above equation, we obtain

$$\frac{1}{MN} V_{\text{ind}} \leq \tilde{V}(w_t, a) := \delta' \tilde{B}/C - \gamma \frac{la^2}{8} - \underline{\gamma} (w_t - a) \frac{la}{4} - \underline{\gamma} \frac{lw_t \epsilon}{16}.$$

When the budget constraint is met tightly, using the symmetry of σ and $1 - \epsilon \geq w_t \geq \epsilon$, we have

$$\tilde{B}/C = \mathbb{E}[\sigma(w_t - w)] \leq \bar{\gamma} a + \underline{\gamma} w_t.$$

We call this the budget constraint. Maximizing $\tilde{V}(w_t, a)$ subject to this budget constraint and two additional constraints $a \leq w_t - \alpha \epsilon$ and $a \geq w_t - \beta \epsilon$ shows that when $\bar{\gamma} \geq 4\underline{\gamma}\beta$, the budget constraint is tight and $a = w_t - \alpha \epsilon$. In this case,

$$\tilde{V}(w_t, a) = \delta' \tilde{B}/C - \frac{l}{8\gamma} \left((\tilde{B}/C - \underline{\gamma}\alpha\epsilon)^2 + 2\underline{\gamma}\alpha\epsilon(\tilde{B}/C - \underline{\gamma}\alpha\epsilon) + \underline{\gamma}\epsilon(\tilde{B}/C + \bar{\gamma}\alpha\epsilon)/2 \right).$$

Then a straightforward calculation shows that for $\bar{\gamma} \geq 2\underline{\gamma}\alpha$,

$$\frac{1}{MN} V_{\text{ind}} \leq \tilde{V}(w_t, a) \leq (\tilde{B}/C) \left(\delta' - \frac{l}{8\gamma} (\tilde{B}/C + \underline{\gamma}\epsilon/2) \right).$$

□

Proof of Lemma 6.4. Our goal is to find a welfare density that minimizes treatment effect T_k while adhering to T'_k and Assumption 6.3. Define $w_{(\alpha)}^k$ as the α^{th} quantile of welfare at unit k . For notational brevity, we drop the superscript k in the following. Let p_w^* be the optimal welfare density with the α^{th} quantile denoted by $w_{(\alpha)}^*$. We show a new valid welfare density p_w can be constructed from p_w^* that gives a T_k lower or equal to what p_w^* could give. Fix $w_{(q')}$, $w_{(1-q)}^*$, and $w_{(1-q')}$. Define four regimes for welfare:

$$\mathcal{R}_1 = [0, w_{(q')}^*), \mathcal{R}_2 = [w_{(q')}^*, w_{(1-q)}^*), \mathcal{R}_3 = [w_{(1-q)}^*, w_{(1-q')}^*), \mathcal{R}_4 = [w_{(1-q')}^*, 1].$$

Note that during within-unit allocation, in the worst case, individuals from the top q' fraction will replace those in the bottom q' fraction and receive treatment. In this case, T_k will only come from the individuals with welfare in \mathcal{R}_2 and \mathcal{R}_4 . Since $\tau(\cdot)$ is nonincreasing, for fixed boundaries of the regimes, our aim is to move the mass within these \mathcal{R}_2 and \mathcal{R}_4 towards larger values while adhering to Assumption 6.3. As the first step, we relax the problem and assume Assumption 6.3 only needs to be held in \mathcal{R}_2 and \mathcal{R}_3 . The minimum of T_k under this relaxation serves as a lower bound for the original problem. If there exist $w^+ \in \mathcal{R}_2 \cup \mathcal{R}_4$, $w^- \in \mathcal{R}_1 \cup \mathcal{R}_3$, $\Delta^+ \geq 0$, and $\Delta^- \geq 0$ such that

1. Either Δ^+ or Δ^- is not zero ,
2. $w^+ + \Delta^+$ and $w^- - \Delta^-$ lie in the same regimes as w^+ and w^- ,
3. $p_w^*(w^+) > 0$ and $p_w^*(w^-) > 0$,
4. If $w^+ \in \mathcal{R}_2$, $p_w^*(w^+ + \Delta^+) < \gamma$,
5. If $w^- \in \mathcal{R}_3$, $p_w^*(w^- - \Delta^-) < \gamma$,
6. $\tau(w^+ + \Delta^+) + \tau(w^- - \Delta^-) = \tau(w^+) + \tau(w^-)$,

then construct a new solution

$$p_w(w) = p_w^*(w) + d \mathbb{1}\{w = w^+ + \Delta^+\} - d \mathbb{1}\{w = w^+\} + d \mathbb{1}\{w = w^- - \Delta^-\} - d \mathbb{1}\{w = w^-\} , \quad (54)$$

where $d > 0$ is the maximum value that keeps p_w a valid probability density function. For instance, if $w^+ \in \mathcal{R}_2$ and $w^- \in \mathcal{R}_3$,

$$d = \min \{p^*(w^+) - 0, p^*(w^-) - 0, \gamma - p^*(w^+ + \Delta^+), \gamma - p^*(w^- - \Delta^-)\} .$$

It is straightforward to verify that the constructed p_w preserves T'_k and meets Assumption 6.3 at \mathcal{R}_2 and \mathcal{R}_3 , but has a T_k lower or equal to what p_w^* could give. By repetitively applying this operation for all w^+ , w^- , Δ^+ , and Δ^- that meet the conditions, we obtain a density function that follows one of the following structures:

1. All the mass in \mathcal{R}_2 and \mathcal{R}_4 has moved towards the right-most boundaries, and all the mass in \mathcal{R}_1 and \mathcal{R}_3 has moved towards the left-most boundaries (Figure 9(c)):

$$p_w(w) = \gamma \mathbb{1}\left\{w \in \left[w_{(1-q)}^* - \frac{1-q-q'}{\gamma}, w_{(1-q)}^* + \frac{q-q'}{\gamma}\right]\right\} + q' \delta(w) + q' \delta(w-1) . \quad (55)$$

Here, $\delta(\cdot)$ is the Dirac delta function.

2. All the mass in \mathcal{R}_2 and \mathcal{R}_4 has moved towards the right-most boundaries. Choose $w^+ \in \mathcal{R}_3$, $w^- \in \mathcal{R}_3$, $\Delta^+ \geq 0$, and $\Delta^- \geq 0$ such that all the previous conditions are met and additionally $w^+ > w^-$. Then applying a similar operation as Equation (54) multiple times, move the mass in \mathcal{R}_3 towards its boundaries. Similarly, if we choose w^+ and w^- from \mathcal{R}_1 and apply the operations multiple times, the mass in \mathcal{R}_1 will be concentrated on its boundaries. These operations maintain T'_k . They also do not change T_k as T_k only depends on p_w in \mathcal{R}_2 and \mathcal{R}_4 . These operations make p_w look like Figure 9(a). Now consider increasing $w_{(1-q)}$ by moving the block around it forward while decreasing $w_{(1-q')}$ such that the resulting distribution preserves T'_k . Then at some point two blocks of mass merge as Figure 9(b). Since the mass under \mathcal{R}_2 is only moving forward, this operation can only decrease T_k . Keep moving the merged block forward while decreasing $w_{(q')}$ to preserve T'_k . This operation also can only decrease T_k . As $w_{(q')}$ approaches zero, we obtain a density like Figure 9(c), which has a similar structure as the first case.
3. All the mass in \mathcal{R}_1 and \mathcal{R}_3 has moved towards the left-most boundaries. We follow similar steps as the previous case. Choose $w^+ \in \mathcal{R}_2$, $w^- \in \mathcal{R}_2$, $\Delta^+ \geq 0$, and $\Delta^- \geq 0$ such that all the required conditions are met and additionally $w^+ > w^-$. Then applying a similar operation as Equation (54) multiple times, move the mass in \mathcal{R}_2 towards its boundaries. Similarly, if we choose w^+ and w^- from \mathcal{R}_4 and apply the operations multiple times, the mass in \mathcal{R}_4 will be concentrated on its boundaries. These operations maintain T'_k . Define $T_k^c := T'_k - T_k$. Since T_k^c only depends on p_w in \mathcal{R}_1 and \mathcal{R}_3 , these operations do not change T_k^c , consequently leaving T_k intact. Now consider decreasing $w_{(1-q)}$ by moving the block around it forward while decreasing $w_{(q')}$ such that the resulting distribution preserves T'_k . Then at some point two blocks of mass merge. This operation can only increase T_k^c and therefore decreases T_k . Keep moving the merged block forward while increasing $w_{(1-q')}$ to preserve T'_k . This operation also can only decrease T_k . As $w_{(1-q')}$ approaches one, we obtain a density like Figure 9(c), which has a similar structure as the first case.

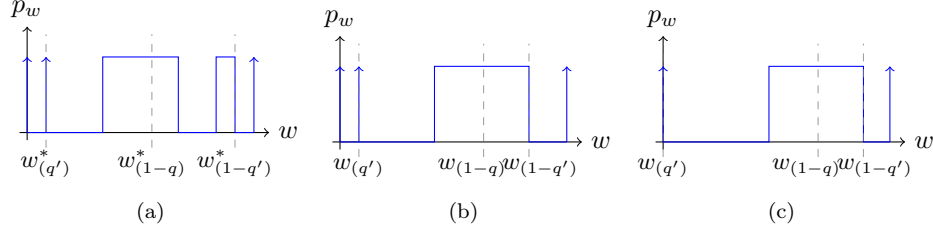


Figure 9: Demonstrating the proof of Lemma 6.4.

The above analysis shows there exists an optimal p_w with the structure following Equation (55) (Figure 9(c)). This solution should give an average treatment effect of T'_k when everyone gets treated. We can write T'_k and T_k resulting from p_w of Equation (55) as

$$\begin{aligned} T'_k &= q'\delta' + \gamma \Gamma_{(1-2q')/\gamma}(w_{(1-q)} - (1-q-q')/\gamma), \\ T_k &= \gamma \Gamma_{(1-q-q')/\gamma}(w_{(1-q)} - (1-q-q')/\gamma). \end{aligned}$$

Then solving for $w_{(1-q)}$ from the first equation and plugging it into the second equation completes the proof. \square

Proof of Theorem D.1. Choose a set \mathcal{S} of n_2 individuals uniformly at random and collect their true welfare data (type 2 data). Let $\hat{t} = \max\{o_i : i \in \mathcal{S}, w_i \leq 1 - \delta\}$ be the estimated threshold. Note that $\hat{t} \leq t$. At every unit k , estimate ρ_k from observations by $\hat{\rho}_k = \frac{1}{N} \sum_{i \in \mathcal{U}_k} \mathbb{1}\{o_i > \hat{t}\}$. The error of estimating ρ_k is

$$\hat{\rho}_k - \rho_k = \frac{1}{N} \sum_{i \in \mathcal{U}_k} \mathbb{1}\{o_i \in (\hat{t}, t]\} \geq 0.$$

Sort the ρ_k values in ascending order and, without loss of generality, let ρ_k be the k^{th} smallest value. Similarly, sort the $\hat{\rho}_k$ values in ascending order and let $\hat{s}(k)$ denote the unit with the k^{th} smallest value. In ULA, we intervene on first K units according to $\hat{s}(\cdot)$. For $k \in [K]$, if k is among $\hat{s}(1), \dots, \hat{s}(K)$, then our previous bound on loss from allocation to this unit will be valid. Otherwise, a unit not among true bottom K units has substituted unit k in allocation. In this case, for any $k' \in [K]$, we have $\hat{\rho}_{\hat{s}(k')} \leq \hat{\rho}_k$ and hence

$$\rho_{\hat{s}(k')} \leq \hat{\rho}_{\hat{s}(k')} \leq \hat{\rho}_k \leq \rho_k + \frac{1}{N} \sum_{i \in \mathcal{U}_k} \mathbb{1}\{o_i \in (\hat{t}, t]\}.$$

So, in the worst case, for $k \in [K]$, units with effective unit-level statistic $\tilde{\rho}_k = \rho_k + \frac{1}{N} \sum_{i \in \mathcal{U}_k} \mathbb{1}\{o_i \in (\hat{t}, t]\}$ will be treated. Recall that $\text{loss}(\cdot)$ in Equation (7) is 1-Lipschitz. This suggests that for a fixed \hat{t} , the increase of loss can be bounded by

$$\Delta \text{loss} \leq \frac{1}{N} \sum_{k \in [K]} \sum_{i \in \mathcal{U}_k} \mathbb{1}\{o_i \in (\hat{t}, t]\}.$$

Define $D := f^{-1}(t) - f^{-1}(\hat{t})$. The fact that f is an increasing function and $\hat{t} \leq t$ ensure $D \geq 0$. Assumption 6.3 limits the concentration of welfare at each unit by γ . Therefore, in the worst case, we can think of $\mathbb{1}\{o_i \in (\hat{t}, t]\}$ as $\text{Ber}(\gamma D)$. Denoting the z-score at the α significance level by z_α and using the central limit theorem to approximate the sum of Bernoulli random variables with a Gaussian distribution, we have

$$\Delta \text{loss}_K \leq \frac{1}{N} (KN\gamma D + z_\alpha \sqrt{KN} \sqrt{\gamma D(1-\gamma D)}) \leq K\gamma D + z_\alpha \sqrt{K/N} \sqrt{\gamma D},$$

with probability at least $1 - \alpha$. Observing $D > d$ requires none of n_2 welfare samples to lie in an interval of length d . Since the overall welfare density is lower bounded by $\underline{\gamma}$ (Assumption 6.1), we have $\Pr(D > d) \leq$

$(1 - \underline{\gamma}d)^{n_2}$. Then a direct calculation shows

$$D \leq (1 - \alpha^{1/n_2})/\underline{\gamma},$$

with probability at least $1 - \alpha$. Putting these together, the increase in loss due to inaccurate estimation of unit-level statistics is bounded by

$$\Delta \text{loss}_K \leq K \frac{\underline{\gamma}}{\underline{\gamma}} (1 - \alpha^{1/n_2}) + z_\alpha \sqrt{KN} \sqrt{\frac{\underline{\gamma}}{\underline{\gamma}}} \sqrt{1 - \alpha^{1/n_2}},$$

with probability at least $1 - 2\alpha$. Assuming $c_2 \approx c$, the cost n_2 type 2 samples at most reduces ULA's value by $n_2\delta$. A unit change in loss_K would also decrease ULA's value by $N\delta$. Therefore, with probability at least $1 - 2\alpha$, the decrease of ULA's value can be bounded by

$$\Delta V_{\text{unit}}/\delta \geq -KN \frac{\underline{\gamma}}{\underline{\gamma}} (1 - \alpha^{1/n_2}) - z_\alpha \sqrt{KN} \sqrt{\frac{\underline{\gamma}}{\underline{\gamma}}} \sqrt{1 - \alpha^{1/n_2}} - n_2.$$

Choose $n_2 = \sqrt{KN} \ln(1/\alpha)$. Then $\alpha^{1/n_2} = \exp(\ln(\alpha)/n_2) = \exp(-1/\sqrt{KN}) \approx 1 - 1/\sqrt{KN}$, and we have

$$\Delta V_{\text{unit}}/\delta \geq -\sqrt{KN} \left(\frac{\underline{\gamma}}{\underline{\gamma}}\right) - z_\alpha (KN)^{1/4} \sqrt{\frac{\underline{\gamma}}{\underline{\gamma}}} - \sqrt{KN} \ln(1/\alpha).$$

Using $V_{\text{unit}} = \Theta(KN\delta)$ and $K = \Theta(B/(Nc))$ complete the proof. \square

Proof of Theorem D.2. Our strategy is to reconstruct f by quantizing welfare into L levels. Choose a set \mathcal{S} of n_2 individuals uniformly at random and get their true welfare (type 2 data). Define

$$\hat{t}_l = \begin{cases} \max \{o_i : i \in \mathcal{S}, w_i \leq l/L\}, & l \in [L-1], \\ 1, & l = L. \end{cases}$$

Here, if a set was empty for l^{th} bin, we can assume $\hat{t}_l = \hat{t}_{l+1}$. Using the estimated thresholds, we can approximate f^{-1} and estimate the welfare of an observation o by

$$\hat{w} = \hat{f}^{-1}(o) = \frac{1}{L} \min \{l \in [L] : \hat{t}_l \geq o\}.$$

Note that by this definition, we have $\hat{w} \geq w$. Consider a welfare value w that lies in $[(l-1)/L, l/L]$ for an $l \in [L-1]$. If there exists $i \in \mathcal{S}$ such that $w_i \in [l/L, (l+1)/L]$, then $\hat{w} - w \leq 2/L$. Since \mathcal{S} are random draws and welfare density is at least $\underline{\gamma}$ (Assumption 6.1),

$$\Pr_{\mathcal{S}}(\hat{w} - w > 2/L) \leq \beta := (1 - \underline{\gamma}/L)^{n_2}.$$

For $l = L$, since $\hat{t}_L = 1$, we always have $\hat{w} - w \leq 1/L$, and the above argument holds more strongly. Denote the unit-level statistics at unit k by U_k : $U_k = \frac{1}{N} \sum_{i \in \mathcal{U}_k} u(w_i)$. Our estimation from U_k is $\hat{U}_k = \frac{1}{N} \sum_{i \in \mathcal{U}_k} u(\hat{w}_i)$, where $\hat{w}_i = \hat{f}^{-1}(o_i)$. For a 1-Lipschitz u , we have

$$(1 - \beta)(u(w_i) - 2/L) \leq \mathbb{E}_{\mathcal{S}}[u(\hat{w}_i)] \leq (1 - \beta)(u(w_i) + 2/L) + \beta.$$

Therefore, $|\mathbb{E}_{\mathcal{S}}[u(\hat{w}_i)] - u(w_i)| \leq \beta + 2/L$. This gives

$$|\mathbb{E}_{\mathcal{S}}[\hat{U}_k] - U_k| \leq \beta + 2/L.$$

Now if V_{unit} is N -Lipschitz in U_k for any $k \in [K]$, the expected decrease of V_{unit} can be bounded by

$$\mathbb{E}_{\mathcal{S}}[\Delta V_{\text{unit}}] \geq -KN(\beta + 2/L) - n_2\delta' = -KN((1 - \underline{\gamma}/L)^{n_2} + 2/L) - n_2\delta',$$

where δ' is the maximum treatment effect an individual can experience. Using the well-known property $\exp(x) \geq (1 + x/n)^n$ for $n \geq |x|$, by choosing $n_2 = (L/\underline{\gamma}) \cdot \ln(L)$ and $L = \sqrt{KN}$, we obtain

$$\mathbb{E}_{\mathcal{S}}[\Delta V_{\text{unit}}] \geq -KN(1/L + 2/L) - (L/\underline{\gamma}) \ln(L) \delta' = -3\sqrt{KN} - \frac{\delta'}{2\underline{\gamma}} \sqrt{KN} \ln(KN).$$

Using $V_{\text{unit}} = \Theta(KN\delta')$ and $K = \Theta(B/(Nc))$ complete the proof. \square