Uncertainty and the Social Planner's Problem: Why Sample Complexity Matters

Cyrus Cousins

August 8, 2022

Abstract

Welfare measures overall utility across a population, whereas malfare measures overall disutility, and the social planner's problem can be cast either as maximizing the former or minimizing the latter. We show novel bounds on the expectations and tail probabilities of estimators of welfare, malfare, and regret of per-group (dis)utility values, where estimates are made from a finite sample drawn from each group. In particular, we consider *estimating* these quantities for individual functions (e.g., allocations or classifiers) with standard probabilistic bounds, and optimizing and *bounding generalization error* over *hypothesis classes* (i.e., we quantify overfitting) using Rademacher averages. We then study algorithmic fairness through the lens of sample complexity, finding that because marginalized or minority groups are often understudied, and fewer data are therefore available, the social planner is more likely to overfit to these groups, thus even models that *seem fair in training* can be *systematically biased* against such groups. We argue that this effect can be mitigated by ensuring sufficient sample sizes for each group, and our sample complexity analysis characterizes these sample sizes. Motivated by these conclusions, we present *progressive sampling* algorithms that efficiently use data to optimize various fairness objectives.

Keywords

Algorithmic Fairness ♢ Minimax Fair Learning ♢ Multi-Group Agnostic PAC Learning ♢ Fair PAC Learning
Social Planner's Problem ♢ Welfare Estimation ♢ Malfare Estimation ♢ Sampling Methods

1 Introduction

Machine learning systems in settings like facial recognition [Buolamwini and Gebru, 2018, Cook et al., 2019, Cavazos et al., 2020] and medicine [Ashraf et al., 2018, Bærøe et al., 2022, Chen et al., 2018] exhibit *differential accuracy* across race, gender, and other protected groups. This can lead to discrimination: for example, facial recognition in policing yields disproportionate false-arrest rates [Garvie et al., 2016], and machine learning in medicine can lead to inequity of health outcomes [DeCamp and Lindvall, 2020], both of which exacerbate existing structural and societal inequalities impacting minority groups. In recent years, researchers have proposed welfare-centric fair learning models, which constrain or optimize welfare [Speicher et al., 2018, Heidari et al., 2018, Rolf et al., 2020, Hu and Chen, 2020, Siddique et al., 2020, Cousins et al., 2022, Do and Usunier, 2022] or malfare [Martinez et al., 2020, Abernethy et al., 2020, Lahoti et al., 2020, Diana et al., 2021, Cousins, 2021, Shekhar et al., 2021] to promote *fair learning* across *all groups*, as well as regret-based methods [Blum and Lykouris, 2020, Rothblum and Yona, 2021], which similarly promote fairness by minimizing the *maximum dissatisfaction* of any group, relative to their preferred outcome, i.e., how much excess risk is incurred, or utility is lost, to any group by *compromising* on a *shared solution*.

We study sampling and learning problems in the optimization of welfare, malfare, and regret objectives. In particular, our setting subsumes the *minimax fair learning* [Martinez et al., 2020, Abernethy et al., 2020, Diana et al., 2021, Lahoti et al., 2020, Shekhar et al., 2021] — also known as *Group Distributionally Robust Optimization* (Group DRO) [Hu et al., 2018, Oren et al., 2019, Sagawa et al., 2019] — and the *fair-PAC learning* [Cousins, 2021] settings, by considering arbitrary malfare or welfare functions, as well as the *multi-group agnostic PAC learning* [Blum and Lykouris, 2020, Rothblum and Yona, 2021] setting, by considering *arbitrary malfare functions* — rather than just the maximum — of per-group regret values. This extension naturally and smoothly interpolates between minimizing *utilitarian* (i.e., weighted average) and *egalitarian* (i.e., maximum) malfare of risk or regret. Crucially, this allows for fine-grained control over the desired fairness concept, and mitigates the minority rule issues of minimax methods, while remaining axiomatically grounded in *cardinal welfare theory*. We bound the generalization error of optimizing welfare, malfare, and regret objectives, and find that while the power-mean malfare is always easy to estimate, due to Lipschitz-continuity (as studied by Cousins [2021]), our learning algorithms work for any malfare, welfare, or regret objective that is *continuous* and *monotonic* in per-group (dis)utility values.

We then study algorithmic fairness through the lens of sample complexity, finding that because marginalized or minority groups are often understudied, and fewer data are therefore available, the social planner is more likely to *overfit* to these groups. Consequently, even models that *seem fair* in training can be *systematically biased* against such groups. Section 3 shows that this effect can be mitigated with sufficient per-group sample sizes, and section 4 presents *progressive sampling* methods, which dynamically sample until a near-optimal model (w.r.t. some fairness objective) is learned. Our analysis rigorously addresses issues raised by, e.g., Chen et al. [2018], who ask how sampling-error impacts fairness, and suggest using learning curves to draw sufficient per-group sample sizes so as to ensure small sampling-error, and Shekhar et al. [2021], who study the problem of optimally allocating sampling effort for the special case of egalitarian malfare.

Our bounds leverage the specific character of the objective at hand; for example, utilitarian welfare is sensitive to the *average* confidence radius across groups, whereas egalitarian welfare is more sensitive to the confidence radii of disadvantaged (i.e., low-utility or high-risk) groups. Furthermore, our progressive sampling methods are tailored to three realistic models of data generation: in the *joint sampling* model, each sample contains a piece of information for every group, in the *mixture sampling* model, samples are annotated with (sets of) group labels, and in the *conditional sampling* model, we are allowed to choose from which groups to sample. In section 4.3, we find that when one considers the specific fairness objective at hand, the optimal decisions as to where to invest sampling effort based on partial information are highly non-trivial, and of great import to fairness. While our settings and modelling assumptions are practically motivated, this is a highly theoretical paper, and all novel results are meticulously proven in appendix A.

Contributions We summarize the contributions of this work as follows.

1) We generalize the regret objective, and unify it with the welfare and malfare objectives for fair machine learning. All three are cardinal objective families, which may be *computed* or *estimated* to measure the fairness of a system, or *optimized* in machine learning or economics settings to create fair systems, i.e., they address the *social planners problem*.

2) We introduce various statistical estimators for regret, welfare, and malfare. We then bound the tails, expectations, and statistical biases of these estimators through a unified analysis in terms of *uniform convergence bounds*.
3) We introduce three practical models of sampling (data collection) for populations consisting of multiple groups, with philosophical connections to *stratified sampling*. We then show that the knowledge-gain in optimizing or estimating our fairness concepts depends intimately on the model class, per group distributions, and the fairness concept at hand.
4) We present two *progressive sampling algorithms*, tailored to our sampling models, to optimize our fairness objectives. We introduce novel technical conditions under which progressive samplers can estimate not only Lipschitz-continuous objectives, but also strictly-monotonic continuous objectives, which is of independent interest beyond the fairness sphere.

2 Learning Framework and Objectives

In this section, we introduce the functional form of the objects and random spaces that we operate over, and we define our learning objectives. In particular, section 2.1 presents the welfare, malfare, and regret objectives, which compile per-group sentiment values into a *cardinal objective value* that can be optimized and analyzed, then section 2.2 reifies this abstract mathematics with three realistic models of data-collection, each of which requires its own statistical treatment to efficiently learn from data, i.e., to optimize and bound objectives, while minimizing the cost of obtaining said data.

We henceforth assume a *supervised learning setting*, where $\mathcal{X}$ is the domain and $\mathcal{Y}$ is the codomain. We also assume either a *loss function*[1] $\ell(\cdot,\cdot)\colon \mathcal{Y}' \times \mathcal{Y} \to \mathbb{R}_{0+}$ or a *utility function* $\mathrm{u}(\cdot,\cdot)\colon \mathcal{Y}' \times \mathcal{Y} \to \mathbb{R}_{0+}$, which map predictions and labels onto *negatively connoted* loss or disutility, or *positively connoted* gain or utility, generically termed a *sentiment function* $\mathrm{s}(\cdot,\cdot)\colon \mathcal{Y}' \times \mathcal{Y} \to \mathbb{R}_{0+}$. In most supervised learning settings, a single probability distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$ suffices, but we assume a set $\mathcal{Z}$ of g groups, and we model the experiences and conditions of each group as its own distribution, i.e., we have $\mathcal{D}_1, \dots, \mathcal{D}_g$. For convenience, we often compose the sentiment function with a predictor or model $h(\cdot) : \mathcal{X} \to \mathcal{Y}'$, taking $(\mathrm{s} \circ h)(x, y) \doteq \mathrm{s}(h(x), y)$, thus we quantify model performance for group i as $\mathbb{E}_{\mathcal{D}_i}[\mathrm{s} \circ h]$.

2.1 Fair Learning with Malfare, Welfare, and Regret Objectives

Here we define the welfare, malfare, and regret objectives. While the details differ, each of these is a function of the expected utility or loss (generically sentiment) of some $h : \mathcal{X} \to \mathcal{Y}$ for each of the g groups, and we are interested in selecting the model or hypothesis h from some *hypothesis class* $\mathcal{H} \subseteq \mathcal{X} \to \mathcal{Y}'$ that optimizes the given objective.

Malfare and Welfare A welfare function $\mathrm{W}(\mathcal{S}; \boldsymbol{w})$ measures overall positive utility $\mathcal{S}$ across a population weighted by $\boldsymbol{w}$, whereas a malfare function $\mathrm{M}(\mathcal{S}; \boldsymbol{w})$ measures overall disutility $\mathcal{S}$, and generically, we say an aggregator function $\mathrm{M}(\mathcal{S}; \boldsymbol{w})$ measures overall sentiment $\mathcal{S}$. The prototypical example is the *utilitarian* (or *Benthamite*) aggregate, defined as $\mathrm{M}_1(\mathcal{S}; \boldsymbol{w}) \doteq \mathcal{S} \cdot \boldsymbol{w}$, which simply averages sentiment across the population (e.g., welfare as per-capita income, or malfare as per-capita medical expenditure), and the second-fiddle is the *egalitarian* (or *Rawlsian*) welfare $\mathrm{W}_{-\infty}(\mathcal{S}; \boldsymbol{w})$ (minimum) or malfare $\mathrm{M}_{\infty}(\mathcal{S}; \boldsymbol{w})$ (maximum), which summarizes a population's sentiment as that of its most disadvantaged member. We assume throughout that $\boldsymbol{w} \in (0, 1)^g$ is a *probability vector*, thus $\|\boldsymbol{w}\|_1 = 1$, and $\mathcal{S} \in \mathbb{R}_{0+}^g$ is *nonnegative*. *Ab initio*, our first objective is, via the social planner's problem, to maximize welfare [Rolf et al., 2020, Hu and Chen, 2020, Siddique et al., 2020, Do and Usunier, 2022], or by extension (e.g., in chores manna or harm allocation [Kulkarni et al., 2021, Heidari et al., 2021], or in machine learning [Abernethy et al., 2020, Cousins, 2021]) to minimize malfare, i.e., we seek to approximate

$$h^\star \doteq \operatorname*{argmin}_{h \in \mathcal{H}} \mathrm{M}\left(i \mapsto \mathop{\mathbb{E}}_{\mathcal{D}_i}[\ell \circ h]; \boldsymbol{w} \right) \quad , \quad \text{or} \quad h^\star \doteq \operatorname*{argmax}_{h \in \mathcal{H}} \mathrm{W}\left(i \mapsto \mathop{\mathbb{E}}_{\mathcal{D}_i}[\mathrm{u} \circ h]; \boldsymbol{w} \right) \quad . \tag{1}$$

Intuitively, the utilitarian case seeks to optimize overall or average sentiment, whereas the egalitarian case instead seeks to lift up the most disadvantaged group, and thus promote equality, perhaps at the expense of overall (total) utility.

[1] Often $\mathcal{Y}' = \mathcal{Y}$, such as in standard classification and regression settings, but this is not universally the case. For instance, in probabilistic classification or regression (i.e., conditional density estimation), $\mathcal{Y}'$ is a space of *distributions over* $\mathcal{Y}$, either parametric [Nelder and Wedderburn, 1972, Cousins and Riondato, 2019] or nonparametric [Rosenblatt, 1956, Parzen, 1962], and in *interval estimation*, $\mathcal{Y}'$ is a space of *contiguous sets over* $\mathcal{Y}$. Similarly, for *recommender systems*, $\mathcal{Y}'$ can be a set of *items*, from which one is to be recommended, and $\mathcal{Y}$ can be a *subset of these items* that a given user would like.

Notably, welfare maximization generalizes *utility maximization* to multiple groups, and malfare minimization likewise generalizes *risk minimization*, and the well-studied *minimax fair learning* framework arises as the special-case of *egalitarian malfare minimization*. In general, we assume only *monotonicity* and *continuity* of aggregator functions; however, there are a set of relatively standard axioms that, when taken together, restricts the class of interest to the *power-mean family* [Cousins, 2021]. This is convenient, as all power-mean malfare functions are Lipschitz-continuous, which in section 4 leads to stronger estimation guarantees and more efficient sampling algorithms than ε-δ limit-continuity.

Definition 2.1 (Axioms of Cardinal Welfare and Malfare)**.** Suppose an aggregator function $\mathrm{M}(\mathcal{S};\boldsymbol{w})$. For each item, assume (if necessary) that the axiom applies for all $\mathcal{S},\mathcal{S}' \in \mathbb{R}^g_{0+}$, scalars $\alpha,\beta \in \mathbb{R}_{0+}$, and probability vectors $\boldsymbol{w} \in (0,1)^g$.
1) *Strict Monotonicity*: $\mathcal{S}' \neq \mathbf{0} \implies \mathrm{M}(\mathcal{S};\boldsymbol{w}) < \mathrm{M}(\mathcal{S}+\mathcal{S}';\boldsymbol{w})$.
2) *Weighted Symmetry*: Suppose $g' \in \mathbb{Z}_+$, $\mathcal{S}' \in \mathbb{R}^{g'}_{0+}$, and probability vector $\boldsymbol{w}' \in (0,1)^{g'}$, such that for all $u \in \mathbb{R}_{0+}$, it holds that $\sum_{i \text{ s.t. } \mathcal{S}_i = u} \boldsymbol{w}_i = \sum_{i \text{ s.t. } \mathcal{S}'_i = u} \boldsymbol{w}'_i$. Then $\mathrm{M}(\mathcal{S};\boldsymbol{w}) = \mathrm{M}(\mathcal{S}';\boldsymbol{w}')$.
3) *Continuity*: $\mathrm{M}(\mathcal{S};\boldsymbol{w})$ is a continuous function (in the standard ε-δ limit-continuity sense) in both $\mathcal{S}$ and $\boldsymbol{w}$.
4) *Independence of Unconcerned Agents*: $\mathrm{M}(\langle \mathcal{S}_{1:g-1},\alpha\rangle;\boldsymbol{w}) \leq \mathrm{M}(\langle \mathcal{S}'_{1:g-1},\alpha\rangle;\boldsymbol{w}) \Longrightarrow \mathrm{M}(\langle \mathcal{S}_{1:g-1},\beta\rangle;\boldsymbol{w}) \leq \mathrm{M}(\langle \mathcal{S}'_{1:g-1},\beta\rangle;\boldsymbol{w})$.
5) *Multiplicative Linearity*: $\mathrm{M}(\alpha\mathcal{S};\boldsymbol{w}) = \alpha\mathrm{M}(\mathcal{S};\boldsymbol{w})$.
6) *Unit Scale*: $\mathrm{M}(\mathbf{1};\boldsymbol{w}) = \mathrm{M}(\langle 1,\ldots,1\rangle;\boldsymbol{w}) = 1$.
7) *Pigou-Dalton Transfer Principle*: Suppose $\mu = \boldsymbol{w}\cdot\mathcal{S} = \boldsymbol{w}\cdot\mathcal{S}'$, and for all $i \in \mathcal{Z}$: $|\mu - \mathcal{S}'_i| \leq |\mu - \mathcal{S}_i|$. Then for utility and welfare, $\mathrm{W}(\mathcal{S}';\boldsymbol{w}) \geq \mathrm{W}(\mathcal{S};\boldsymbol{w})$, and for disutility and malfare, $\mathrm{M}(\mathcal{S}';\boldsymbol{w}) \leq \mathrm{M}(\mathcal{S};\boldsymbol{w})$.

These are the axioms employed by Cousins [2021] in the construction of the *fair-PAC learning* framework, and we briefly argue they are quite natural, though henceforth we only *require* monotonicity (1) and at times continuity (3). Axioms 1–4 are essentially the standard *axioms of cardinal welfare* [Sen, 1977, Roberts, 1980], modified to include the weights $\boldsymbol{w}$, and omitting any of them leads to rather perverse aggregator functions. Axiom 5 (multiplicative linearity) strengthens the traditional *independence of common scale* axiom, and ensures that the *units* of welfare or malfare must match those of sentiment, and axiom 6 (unit scale) merely specifies a multiplicative constant. Taken together, axioms 1–6 strengthen the Debreu-Gorman theorem [Debreu, 1959, Gorman, 1968], to uniquely characterizes all aggregator functions as *weighted power-means*. Finally, axiom 7, the *Pigou-Dalton transfer principle* Pigou [1912], Dalton [1920], characterizes *fairness* in the sense of *equitable redistribution* of utility (welfare) or disutility (malfare).

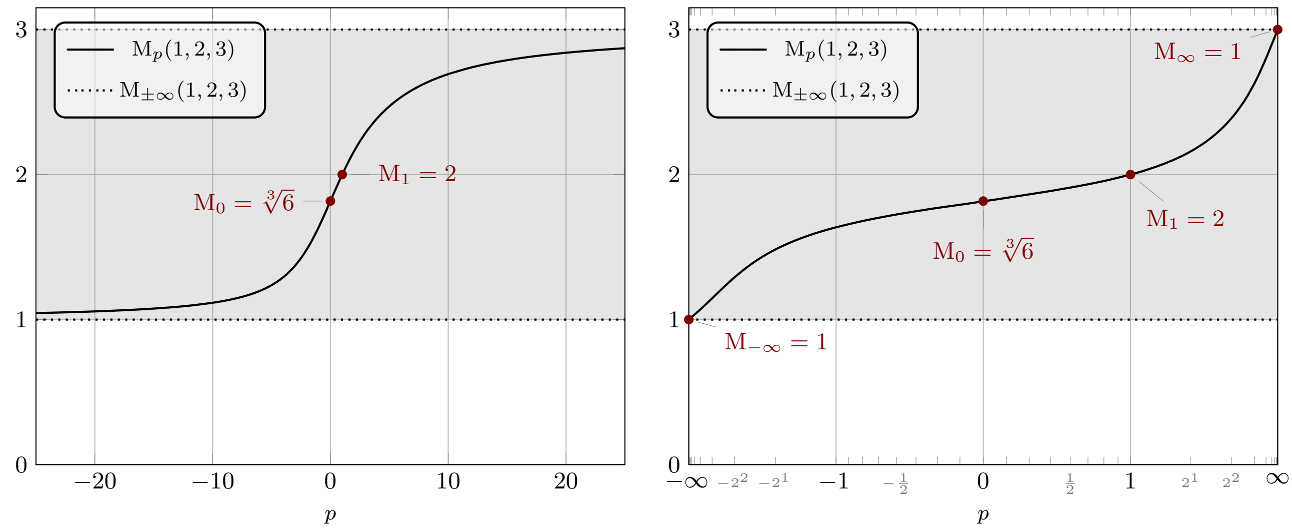
Theorem 2.2 (Aggregator Function Properties [Cousins, 2021, theorems 2.4 and 2.5])**.** Suppose aggregator function $\mathrm{M}(\mathcal{S};\boldsymbol{w})$, and assume arbitrary sentiment vector $\mathcal{S} \in \mathbb{R}^g_{0+}$ and probability vector $\boldsymbol{w} \in (0,1)^g$. The following then hold.
1) *Power-Mean Factorization*: Axioms 1–6 imply $\exists p \in \mathbb{R}$ s.t.

$$\mathrm{M}(\mathcal{S};\boldsymbol{w}) = \mathrm{M}_p(\mathcal{S};\boldsymbol{w}) \doteq f_p^{-1}\left(\sum_{i=1}^{g}\boldsymbol{w}_i f_p(\mathcal{S}_i)\right) =_{p\neq 0} \sqrt[p]{\sum_{i=1}^{g}\boldsymbol{w}_i\mathcal{S}_i^p} \ , \quad \text{with} \ \begin{cases} p = 0 & f_0(x) \doteq \ln(x) \\ p \neq 0 & f_p(x) \doteq \mathrm{sgn}(p)x^p \end{cases} .$$

2) *Fair Welfare and Malfare*: Axioms 1–7 imply $p \in (-\infty, 1]$ for welfare and $p \in [1,\infty)$ for malfare.
3) *Lipschitz-Continuity*: For all $p \geq 1$, it holds that $\left|\mathrm{M}_p(\mathcal{S};\boldsymbol{w}) - \mathrm{M}_p(\mathcal{S}';\boldsymbol{w})\right| \leq \mathrm{M}_p(|\mathcal{S}-\mathcal{S}'|;\boldsymbol{w}) \leq \max_{i\in\mathcal{Z}}|\mathcal{S}_i - \mathcal{S}'_i|$.

In closing, we note that utilitarian philosophy is often criticized for permitting great inequality by ignoring the needs of smaller or less visible groups, whereas egalitarian philosophy is criticized for ignoring the masses in favor of outliers and disadvantaged groups, and its inherent susceptibility to minority rule. Concretely, utilitarian aggregates only weakly satisfy the Pigou-Dalton principle, thus do not incentivize equitable redistribution (of wealth or suffering), and egalitarian aggregates satisfy only weak (i.e., not strict) monotonicity, thus only incentivize gains in the *most disadvantaged* group(s). Power-means provide a spectrum of intermediaries, so exactly how tradeoffs should be made may depend on the application, as well as the culturosocietal values of the social planner and their society. They are also *statistically convenient*, as many of our estimation guarantees hold in terms of generic Lipschitz-continuity assumptions, and thus apply to any power-mean malfare function. Figure 1 illustrates the behavior of the power mean function around the utilitarian $p = 1$, as well as the egalitarian limits of $p \to \pm\infty$.

The Malfare of Regret We now discuss *regret*, which is a property of the hypothesis class $\mathcal{H}$ and per-group distributions $\mathcal{D}_{1:g}$. Intuitively, regret measures the *relative dissatisfaction* of group i with some $h \in \mathcal{H}$, relative to their preferred outcome $\boldsymbol{h}^\star_i \in \mathcal{H}$. In particular, we define the (per-group) preferred outcome $\boldsymbol{h}^\star_i$ as the model that group i would select



(a) Natural axes plot of the power mean, with $p \in [-25, 25]$. Behavior around $p = 0$ is clear, and behavior as $p \to \pm\infty$ is suggested (but not confirmed) by observing extreme values of p. The utilitarian and geometric welfare are also visually depicted.

(b) Sigmoidally transformed plot of the power mean. The x-axis is transformed by the tangent function to emphasize behavior around $p \approx 0$, as well as for $p \to \pm\infty$. Minor ticks are *linearly spaced* on the $[-1, 1]$ region, and *powers of* 2 outside this region.

Figure 1: Plots of the unweighted power mean of (dis)utility values $\langle 1, 2, 3 \rangle$, as a function of the power p. Observe that the limits as $p \to \pm\infty$ are the minimum and maximum, and these and other significant values are marked on the plots. Figure 1a presents a natural axes plot, which necessarily is limited to a finite region of p, and figure 1b presents a sigmoidally transformed plot, which shows the entire spectrum of $p \in [-\infty, \infty]$.

for themselves, i.e.,

$$\boldsymbol{h}_i^\star \doteq \operatorname*{argmin}_{h \in \mathcal{H}} \mathop{\mathbb{E}}_{\mathcal{D}_i}[\ell \circ h] \quad \text{or} \quad \boldsymbol{h}_i^\star \doteq \operatorname*{argmax}_{h \in \mathcal{H}} \mathop{\mathbb{E}}_{\mathcal{D}_i}[\mathrm{u} \circ h] \enspace , \tag{2}$$

for loss or utility, respectively, and we let $\mathcal{S}_i^\star$ denote the *optimal expected sentiment* for group i, i.e., $\mathcal{S}_i^\star \doteq \mathbb{E}_{\mathcal{D}_i}[\mathrm{s} \circ \boldsymbol{h}_i^\star]$. We now formally define the *regret* of group i on some outcome or model $h \in \mathcal{H}$ as

$$\mathrm{Reg}_i(h) \doteq \mathop{\mathbb{E}}_{\mathcal{D}_i}[\ell \circ h] - \mathcal{S}_i^\star \enspace , \quad \mathrm{Reg}_i(h) \doteq \mathcal{S}_i^\star - \mathop{\mathbb{E}}_{\mathcal{D}_i}[\mathrm{u} \circ h] \enspace , \quad \text{or generically,} \quad \mathrm{Reg}_i(h) \doteq \left| \mathop{\mathbb{E}}_{\mathcal{D}_i}[\mathrm{s} \circ h] - \mathcal{S}_i^\star \right| \enspace . \tag{3}$$

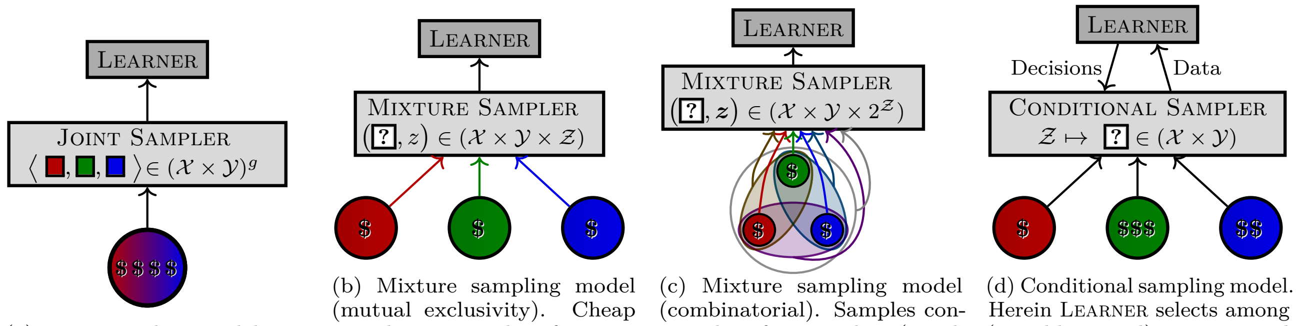
Intuitively $\mathrm{Reg}_i(h)$ is nonnegative by construction (hence the absolute value in the generic form), and it quantifies the amount by which group i prefers their optimal $\boldsymbol{h}_i^\star$ to h.

Several authors [Blum and Lykouris, 2020, Rothblum and Yona, 2021] minimize the worst-case (over groups) *regret* of the selected $\hat{h}$, and the statistical and computational questions that arise are studied under the umbrella of "multi-group agnostic PAC learning." We generalize this notion, optimizing not just *worst-case* (i.e., egalitarian), but *arbitrary malfare functions*, of per-group regret values, which allows for greater flexibility and resistance to the usual issues of egalitarian malfare. In particular, we seek

$$h^\star \doteq \operatorname*{argmin}_{h \in \mathcal{H}} \mathrm{M}\left(i \mapsto \mathrm{Reg}_i(h); \boldsymbol{w}\right) = \operatorname*{argmin}_{h \in \mathcal{H}} \mathrm{M}\left(i \mapsto \left| \mathop{\mathbb{E}}_{\mathcal{D}_i}[\mathrm{s} \circ h] - \mathcal{S}_i^\star \right|; \boldsymbol{w}\right) \enspace . \tag{4}$$

Curiously, since we seek to measure overall regret, and regret is a nonnegative quantity with negative connotation, we always summarize it with a malfare function $\mathrm{M}(\cdot; \boldsymbol{w})$, even when we began with a *utility function*. Intuitively, this is because we can never hope to select a shared function $\hat{h}$ that group i prefers to $\boldsymbol{h}_i^\star$, thus *excess dissatisfaction* is always positive in both the loss and utility cases. In some sense, the malfare of regret thus measures the *price of sharing* in a society, as the shared model $\hat{h}$ is naturally compared [Dwork et al., 2018] to letting each group select their own model $\hat{h}_i$.

Why Consider the Malfare of Regret? Previous work summarizes regret across groups by taking the *largest regret* amongst them. This is analogous to game-theoretic regret (i.e., the maximum over agents of *utility differences* between *adjacent profiles*), but even there, *any malfare function* could reasonably aggregate per-group regret values. We argue



(a) Joint sampling model. Expensive samples each contain data for all groups.

(b) Mixture sampling model (mutual exclusivity). Cheap samples contain data for a single random (or adversarially selected) group.

(c) Mixture sampling model (combinatorial). Samples contain data for a random (or adversarially selected) *nonempty combination of groups.*

(d) Conditional sampling model. Herein LEARNER selects among (variably-priced) groups, and then SAMPLER draws samples from the selected group.

Figure 2: Visualization of sampling models for group-centric fair machine learning. The data collected, its topology, and which or how decisions are made are described for each sampling model. Note that the *mixture sampling model* is split into the simpler *mutually exclusive* case (figure 2b), wherein each sample pertains to a single group $z \in \mathcal{Z}$, as well as the *combinatorial* case (figure 2c), in which each sample pertains to some subset $\boldsymbol{z} \in 2^{\mathcal{Z}}$ of the set of all groups $\mathcal{Z}$.

that considering only *egalitarian regret* may act as an enforcer of the status quo, if one group is particularly happy with their $\boldsymbol{h}_i^\star$ and is thus aggrieved by any compromise — perhaps best summarized by the adage, "To those accustomed to privilege, equality feels like oppression." We mitigate this issue by summarizing regret with a power-mean malfare function $\text{M}_p(\cdot; \boldsymbol{w})$, instead of the egalitarian malfare, in order to lessen the impact of the most aggrieved group. In particular, this class smoothly and nonlinearly interpolates between the worst-case (egalitarian) $\text{M}_\infty(\cdot; \boldsymbol{w})$ regret and the utilitarian $\text{M}_1(\cdot; \boldsymbol{w})$ welfare or malfare.

Fascinatingly, we find that utilitarian regret minimization reduces to utilitarian malfare or welfare optimization, as all terms involving per-group optimal sentiment can be factored into an additive constant from these objectives; observe

$$\text{M}_1\big(i \mapsto \text{Reg}_i(h); \boldsymbol{w}\big) = \text{M}_1\left(i \mapsto \left|\mathop{\mathbb{E}}_{\mathcal{D}_i}[\text{s} \circ h] - \mathcal{S}_i^\star\right|; \boldsymbol{w}\right) = \left|\left(\sum_{i=1}^{g} \boldsymbol{w}_i \mathop{\mathbb{E}}_{\mathcal{D}_i}[\text{s} \circ h]\right) - \boldsymbol{w} \cdot \mathcal{S}^\star\right| = \begin{cases} \text{s} = \ell & \text{M}_1\big(\mathbb{E}_{\mathcal{D}_i}[\ell \circ h]; \boldsymbol{w}\big) - \boldsymbol{w} \cdot \mathcal{S}^\star \\ \text{s} = \text{u} & \boldsymbol{w} \cdot \mathcal{S}^\star - \text{W}_1\big(\mathbb{E}_{\mathcal{D}_i}[\text{u} \circ h]; \boldsymbol{w}\big) \end{cases}, \tag{5}$$

namely $\mathcal{S}^\star$ appears only in the additive constant $\boldsymbol{w} \cdot \mathcal{S}^\star$, which is independent of h. From this perspective, we conclude that while the utilitarian regret is not particularly interesting, the power-mean malfare of regret *interpolates between* minimizing largest regret, with its minority rule issues, and optimizing utilitarian welfare or malfare, where $\boldsymbol{w}$ parameterizes the utilitarian objective, and p precisely specifies how the objective trades off between the two extremes.

2.2 Three Sampling Models for Populations with Multiple Groups

In order to study efficient sampling, we must first *quantify* the *cost* of a sampling-based estimation routine, which requires a *sampling model.* Within a single-group population, methods like *i.i.d. sampling*, *importance sampling*, or *sampling without replacement* are near-ubiquitous, and all can measure cost as *sample size* $m \in \mathbb{Z}_+$, where $\mathbb{Z}_+$ denotes the *positive integers*; however, in group-sensitive settings, we must consider how samples from different groups are obtained, and what the cost of collecting these samples is. In the context of this work, we don't argue for a one-size-fits-all solution, but rather we discuss three sampling models, and show that they fit key applications in the computer science domain and beyond.

1) *Joint Sampling*: Each i.i.d. sample contains a piece of information for each of the g groups, with arbitrary dependencies *between groups.* For example, per-group representatives could be shown a shared $x \in \mathcal{X}$ and asked for their feedback, which would then be used to establish some $\mathcal{Y}_i$ for each group i. Thus each sample is in the space $\mathcal{X} \times \mathcal{Y}^g$ if the $\mathcal{X}$ components are shared between groups, or more generally in $(\mathcal{X} \times \mathcal{Y})^g$. This setting also arises in *multi-objective reinforcement learning* [Siddique et al., 2020, Cousins et al., 2022], as well as various bandit problems and empirical game theoretic analysis [Viqueira et al., 2020], where each query of an *action* or *strategy profile* yields a sample of the utility values of each *player*, *agent*, or *group.*

2) *Mixture Sampling*: For each sample, the data are only relevant to a nonempty *subset of groups* $\boldsymbol{z} \in 2^{\mathcal{Z}}$, thus samples

are elements of $\mathcal{X} \times \mathcal{Y} \times 2^{\mathcal{Z}}$. This generality is useful for studying concepts like *intersectionalism* and *multicalibration* [Rothblum and Yona, 2021], where individuals may belong to multiple groups, (e.g., at the interface of both gender and race), and is in some sense more data-efficient than associating each sample with a single group, but the case of *mutually exclusive* groups (i.e., each sample belongs to exactly one group, thus samples are in $\mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$) is also computationally and philosophically convenient [Dwork et al., 2018]. This model naturally represents a mixed population being sampled i.i.d., where the group identities of the sample are left up to chance (i.e., roughly proportional to group frequencies), and is thus the most appropriate model for learning from [Cousins, 2021] existing datasets with *group identity features* [Ding et al., 2021].

3) *Conditional Sampling*: Here we *actively choose* from which group to sample, in contrast to the mixture sampling model, where we simply cast our net and "get what we get." In particular, we sample i.i.d. $(\mathcal{X}, \mathcal{Y})$ pairs *conditioned on* some group $z \in \mathcal{Z}$, thus we may select sample sizes $\boldsymbol{m}_{1:g} \in \mathbb{Z}_+^g$ and draw a sample $(\boldsymbol{x}, \boldsymbol{y}) \in (\mathcal{X} \times \mathcal{Y})^{\boldsymbol{m}_1} \times \cdots \times (\mathcal{X} \times \mathcal{Y})^{\boldsymbol{m}_g}$. This is a natural model in *active sampling* [Abernethy et al., 2020] and *scientific inquiry* settings, where initial results guide further study and resource expenditure, and similar conditional sampling structure arises in *stratified sampling* settings.

Each of our three sampling models, as well as how they interact with learning or estimation routines, is illustrated in figure 2. In mixture sampling, we generally assume *unit cost* $C = 1$ per sample, and in joint sampling, we assume *constant cost* $C > 1$ per joint-sample, as it is more expensive to set up a properly controlled joint sampling distribution. On the other hand, in conditional sampling, some groups may be more difficult or costly to study than others, so we assume a *cost model* $\boldsymbol{C}_{1:g} \in \mathbb{R}_+^g$, where $\boldsymbol{C}_i$ is the per-sample cost for group i, thus the total cost of drawing a sample with per-group sizes $\boldsymbol{m}_{1:g}$ is $\boldsymbol{m} \cdot \boldsymbol{C}$. Note that the extra control of the conditional sampling model is extremely convenient and very powerful, however it is generally more expensive than mixture sampling. These costs are entirely application dependent, so we take no stance on which is preferable, and rather focus on developing efficient learning algorithms under each sampling model.

3 Statistical Analysis and Estimation Guarantees

In this section, we discuss the statistics of estimating malfare and welfare functions. In particular, we assume a set $\mathcal{Z}$ of g groups, and we want to estimate the malfare, welfare, or regret of per-group expected loss or utility of some $h : \mathcal{X} \to \mathcal{Y}'$, i.e.,

$$\hat{\mathrm{M}} \approx \mathrm{M}\big(i \mapsto \mathop{\mathbb{E}}_{\mathcal{D}_i}[\mathrm{s} \circ h]; \boldsymbol{w}\big) \enspace , \quad \text{or} \quad \hat{\mathrm{M}} \approx \mathrm{M}\big(i \mapsto \mathrm{Reg}_i(h); \boldsymbol{w}\big) \enspace ,$$

where $\mathcal{D}_{1:g}$ are distributions over $\mathcal{X} \times \mathcal{Y}$, $\mathrm{s}(\cdot, \cdot) : \mathcal{Y}' \times \mathcal{Y} \to \mathbb{R}_{0+}$, $(\mathrm{s} \circ h)(x, y) \doteq \mathrm{s}(h(x), y)$, and $\mathrm{M}(\cdot; \boldsymbol{w})$ generically represents some aggregator function. Estimating the expected loss or utility of one group is a well-studied sampling problem, but generalizing to the welfare, malfare, or regret of multiple groups introduces some subtleties. We start by noting that while the empirical mean is an unbiased estimator of expected utility or loss *of a single group*, in general there is no unbiased estimator of welfare or malfare (essentially due to their nonlinear nature, much like with the *standard deviation*). Thus rather than unbiased estimators, we seek *additive error* (AE) bounds of the form $\mathbb{P}\big(|\mathrm{M} - \hat{\mathrm{M}}| \leq \varepsilon\big) \geq 1 - \delta$, where ε is the *confidence radius* (a.k.a. the *margin of error*), and δ is the *failure probability* (or, by alternative convention, $1 - \delta$ is the *level of confidence*).

In machine learning, it does not suffice to estimate the welfare or malfare of a *single function* $h(\cdot) : \mathcal{X} \to \mathcal{Y}$, as we optimize over a *hypothesis class* $\mathcal{H} \subseteq \mathcal{X} \to \mathcal{Y}'$, thus we seek some sample-dependent $\hat{h} \in \mathcal{H}$ with true objective value within ε of that of the optimal $h^\star \in \mathcal{H}$. At times, we are also interested in related statistics, like the objective values of $\hat{h}$ and $h^\star$, and in general, tools to bound the deviations between the empirical and true objective values for any $h \in \mathcal{H}$ are sufficient to bound these quantities. The rest of this section pursues such bounds, assuming a fixed *failure probability* δ and *sample size* $\boldsymbol{m}_i$ for each group $i \in \mathcal{Z}$. In particular, section 3.1 reviews known results for uniformly estimating expectations across $\mathcal{H}$, section 3.2 builds upon these results to uniformly estimate malfare, welfare, and regret values, and section 3.3 then studies how varying per-group sentiment values and confidence radii impacts these bounds, and quantifies the incremental value of sampling from each group as a function of these quantities.

3.1 Uniform Convergence Bounds for Mean Estimation

In this work, the common functional form of our additive error (AE) bounds is *data dependent uniform convergence*, vectorized to operate over samples from multiple groups, rather than on a single-group sample. Occasionally, we are interested in the scalar form $\mathrm{AES}(m, \delta, \boldsymbol{x}, \boldsymbol{y}) : \mathbb{Z}_+ \times (0,1) \times \mathcal{X}^m \times \mathcal{Y}^m \to \mathbb{R}_{0+}$, which operates on a single group, but unless otherwise stated, we refer to the vector bound $\mathrm{AEV}(\boldsymbol{m}, \delta, \boldsymbol{x}, \boldsymbol{y}) : \mathbb{Z}_+^g \times (0,1) \times (\mathcal{X}^{\boldsymbol{m}_1} \times \cdots \times \mathcal{X}^{\boldsymbol{m}_g}) \times (\mathcal{Y}^{\boldsymbol{m}_1} \times \cdots \times \mathcal{Y}^{\boldsymbol{m}_g}) \to \mathbb{R}_{0+}$. In particular, given a sample $(\boldsymbol{x}, \boldsymbol{y}) \sim \mathcal{D}_1^{\boldsymbol{m}_1} \times \cdots \times \mathcal{D}_g^{\boldsymbol{m}_g}$, we require a *random function*[2] $\mathrm{AEV}(\dots)$ such that

$$\hat{\boldsymbol{\varepsilon}} \leftarrow \mathrm{AEV}(\boldsymbol{m}, \delta, \boldsymbol{x}, \boldsymbol{y}) \implies \mathop{\mathbb{P}}_{\boldsymbol{x},\boldsymbol{y},\hat{\boldsymbol{\varepsilon}}} \left(\max_{i \in \mathcal{Z}} \sup_{h \in \mathcal{H}} \left| \mathop{\hat{\mathbb{E}}}_{\boldsymbol{x}_{i,:},\boldsymbol{y}_{i,:}}[\mathrm{s} \circ h] - \mathop{\mathbb{E}}_{\mathcal{D}_i}[\mathrm{s} \circ h] \right| - \hat{\boldsymbol{\varepsilon}}_i > 0 \right) < \delta \ . \tag{6}$$

Section 3.2 explores how $\mathrm{AEV}(\dots)$ can be used to bound malfare, welfare, and regret, and the remainder of this subsection is dedicated to showing non-trivial bounds of this form for machine learning applications. All of our additive error bounds assume *bounded sentiment range* $r \doteq \sup_{y' \in \mathcal{Y}', y \in \mathcal{Y}} \mathrm{s}(y', y)$, but this can usually be relaxed if we instead assume a *moment condition*, e.g., each $\mathrm{s} \circ h$ is sub-exponential, sub-gamma, sub-Poisson, or sub-Gaussian [Boucheron et al., 2013]. We often assume for the sake of intuition that uniform convergence rates exhibit convergence in probability to $\mathrm{AES}(m, \delta, \dots) \in \boldsymbol{\Theta}\sqrt{\ln \frac{1}{\delta} / m}$, which agrees with approximate bounds from many central limit theorems, though section 4 requires less restrictive convergence rates to show correctness of our estimation routines.

Data-Dependent Bounds Data-dependent uniform convergence bounds, i.e., those of the form $\mathrm{AES}(m, \delta, \boldsymbol{x}, \boldsymbol{y})$, are invaluable for studying a population about which little is known. However, they can't be evaluated until data are available, thus we cannot determine *a priori* how much data will be required to meet a given confidence radius. This contrasts distribution-dependent bounds, which take the form $\mathrm{AES}_{\mathcal{D}}(m, \delta)$, and depend on the distribution $\mathcal{D}$ (but not the data $(\boldsymbol{x}, \boldsymbol{y})$), and therefore must make often-problematic or unrealistic assumptions about the data-distribution. Even further in this direction are distribution-free bounds, which depend on neither the distribution nor the data, but must thus yield worst-case dependence on the data distribution. These three classes of bounds may be related, in the sense that each RHS can be used to bound each LHS, as

$$\mathrm{AES}(m, \delta) \le \sup_{\mathcal{D} \text{ over } \mathcal{X} \times \mathcal{Y}} \mathrm{AES}_{\mathcal{D}}(m, \delta) \le \sup_{(\boldsymbol{x}, \boldsymbol{y}) \in (\mathcal{X} \times \mathcal{Y})^m} \mathrm{AES}(m, \delta, \boldsymbol{x}, \boldsymbol{y}) \ .$$

In section 4, when constructing schedules for progressive sampling, we often assume knowledge of $\mathrm{AES}(m, \delta)$, but this is usually possible via this worst-case RHS bound.

Learning and Uniform Convergence We first present simple bounds for *bounded finite hypothesis classes*, which depend on the *sentiment range* r, hypothesis class size $|\mathcal{H}|$, variances $\mathbb{V}[\cdot]$, and empirical variances $\hat{\mathbb{V}}[\cdot]$.

Theorem 3.1 (Uniform Convergence for Bounded Finite Hypothesis Classes)**.** We may bound the distribution-free $\mathrm{AES}(m, \delta)$, the distribution-dependent $\mathrm{AES}_{\mathcal{D}}(m, \delta)$, and the data-dependent $\mathrm{AES}(m, \delta, \boldsymbol{x}, \boldsymbol{y})$ scalar additive error as

1) $\boldsymbol{\varepsilon} \leftarrow \sqrt{\dfrac{2\frac{1}{4} r^2 \ln \frac{2|\mathcal{H}|}{\delta}}{m}}$ [Hoeffding, 1963] ;

2) $\boldsymbol{\varepsilon} \leftarrow \dfrac{r \ln \frac{2|\mathcal{H}|}{\delta}}{3m} + \sup_{h \in \mathcal{H}} \sqrt{\dfrac{2\, \mathbb{V}_{\mathcal{D}}[\mathrm{s} \circ h] \ln \frac{2|\mathcal{H}|}{\delta}}{m}}$ [Bennett, 1962] ; &

3) $\hat{\boldsymbol{\varepsilon}} \leftarrow \dfrac{7r \ln \frac{2|\mathcal{H}|+1}{\delta}}{3(m-1)} + \sup_{h \in \mathcal{H}} \sqrt{\dfrac{2\, \hat{\mathbb{V}}_{\boldsymbol{x},\boldsymbol{y}}[\mathrm{s} \circ h] \ln \frac{2|\mathcal{H}|+1}{\delta}}{(m-1)}}$ [Cousins and Riondato, 2020] ; respectively.

Note that supremum variances and empirical variances are properties of the distribution and sample, respectively. Dependence on variance is necessary (similar terms appear in mean-estimation lower-bounds [Devroye et al., 2016, Lugosi and Mendelson, 2019]), however the $\ln|\mathcal{H}|$ union bound terms are loose, and the bounds are vacuous for infinite

[2] Note that $\mathrm{AEV}(\dots)$ can be a randomized algorithm (e.g., involving Monte-Carlo Rademacher averages [Cousins and Riondato, 2020], bootstrapping, permutation testing, or other such methods), thus in general, its output $\hat{\boldsymbol{\varepsilon}}$ is a *random variable* that depends on the data $\boldsymbol{x}, \boldsymbol{y}$. Going forward, we present only scalar bounds, but it is to be understood that given additive error *scalar bound* $\mathrm{AES}(\dots)$ and a finite group count g, we may construct the additive error *vector bound* $\mathrm{AEV}(\boldsymbol{m}, \delta, \boldsymbol{x}, \boldsymbol{y}) \leftarrow \langle \mathrm{AES}(\boldsymbol{m}_1, \frac{\delta}{g}, \boldsymbol{x}_{1,:}, \boldsymbol{y}_{1,:}), \dots, \mathrm{AES}(\boldsymbol{m}_g, \frac{\delta}{g}, \boldsymbol{x}_{g,:}, \boldsymbol{y}_{g,:}) \rangle$ via the union bound.

(continuous) $\mathcal{H}$. We now state results using Rademacher averages [Bartlett and Mendelson, 2002, Shalev-Shwartz and Ben-David, 2014] that tolerate infinite $\mathcal{H}$, while preserving the variance-dependence of item 2.

Theorem 3.2 (Uniform Convergence with Rademacher Averages)**.** Suppose hypothesis class $\mathcal{H}$ and sentiment function $\mathrm{s}(\cdot,\cdot)$, take $(\boldsymbol{x},\boldsymbol{y}) \sim \mathcal{D}^m$ and $\boldsymbol{\sigma} \sim \mathcal{U}^m(\pm 1)$, i.e., $\boldsymbol{\sigma}$ is uniformly distributed on $(\pm 1)^m$, and define the *Rademacher average* $\mathfrak{R}_m(\mathrm{s}\circ\mathcal{H},\mathcal{D})$ and *Bousquet variance proxy* $\mathfrak{V}_m(\mathrm{s}\circ\mathcal{H},\mathcal{D})$ [see Bousquet, 2002] as

$$\mathfrak{R}_m(\mathrm{s}\circ\mathcal{H},\mathcal{D}) \doteq \mathop{\mathbb{E}}_{\boldsymbol{x},\boldsymbol{y},\boldsymbol{\sigma}}\left[\sup_{h\in\mathcal{H}}\left|\frac{1}{m}\sum_{i=1}^{m}\mathrm{s}\circ h(\boldsymbol{x}_i)\boldsymbol{\sigma}_i\right|\right] \quad , \quad \mathfrak{V}_m(\mathrm{s}\circ\mathcal{H},\mathcal{D}) \doteq \sup_{h\in\mathcal{H}}\mathop{\mathbb{V}}_{\mathcal{D}}[\mathrm{s}\circ h] + 4r\mathfrak{R}_m(\mathrm{s}\circ\mathcal{H},\mathcal{D}) \ . \tag{7}$$

We may then bound $\mathrm{AES}_{\mathcal{D}}(m,\delta)$ as $\boldsymbol{\varepsilon} \leftarrow 2\mathfrak{R}_m(\mathrm{s}\circ\mathcal{H},\mathcal{D}) + \frac{r\ln\frac{1}{\delta}}{3m} + \sqrt{\frac{2\mathfrak{V}_m(\mathrm{s}\circ\mathcal{H},\mathcal{D})\ln\frac{1}{\delta}}{m}}$.

Data-dependent analogues of theorem 3.2 are possible using *empirical Rademacher averages* and *variances* at no asymptotic cost [Cousins and Riondato, 2020]. In the worst case, theorem 3.2 performs comparably to union bounds, i.e., theorem 3.1 item 2, however it improves when *correlations* exist between elements of $\mathcal{H}$, because the *effective size* of $\mathcal{H}$ is smaller for the purposes of realizing the supremum in the Rademacher average, see (7). The abstract inequalities of theorem 3.2 are quite opaque, so we now provide concrete bounds on the Rademacher averages of practical infinite hypothesis classes. The below results hold *for any distribution* $\mathcal{D}$, and are thus distribution-free, although similar distribution-dependent or data-dependent bounds are possible.

Property 3.3 (Practical Bounds on Rademacher Averages)**.** 1) Suppose $\mathcal{H}$ has Vapnik-Chervonenkis (VC) dimension d [Vapnik and Chervonenkis, 1968, 1971], and $\ell(\hat{y},y) \doteq 1 - \mathbb{1}_y(\hat{y})$ is the 0-1 loss. Then for some absolute constant c, $\mathfrak{R}_m(\ell\circ\mathcal{H},\mathcal{D}) \leq \sqrt{\frac{cd}{m}}$, which implies bounds for linear classifiers, bounded-depth decision trees [Leboeuf et al., 2020], and many classes of neural network [Anthony and Bartlett, 2009].
2) Suppose $\mathcal{X} \doteq \{\vec{x}\in\mathbb{R}^\infty \mid \|\vec{x}\|_2 \leq R\}$ is the R-radius $\mathcal{L}_2$ ball in $\mathbb{R}^\infty$, $\mathcal{H} \doteq \{\vec{x}\mapsto\vec{w}\cdot\vec{x} \mid \|\vec{w}\|_2 \leq \gamma\}$ is a γ-regularized linear class, $\mathcal{Y} \doteq [-R\gamma, R\gamma]$, and $\ell(\cdot,\cdot)$ is a λ-Lipschitz loss function s.t. $\ell(y,y) = 0$. Then $r \leq 2\lambda R\gamma$ and $\mathfrak{R}_m(\ell\circ\mathcal{H},\mathcal{D}) \leq \frac{2\lambda R\gamma}{\sqrt{m}}$. This implies bounds for (kernelized) SVM, generalized linear models [Nelder and Wedderburn, 1972], and bounded linear regression.

3.2 From Mean Estimation to Welfare, Malfare, and Regret Bounds

We now adapt the additive error bounds of section 3.1 on expectations to bound malfare, welfare, and regret in terms of empirical estimates thereof. In particular, the strategy for each is to combine tail bounds for mean-estimation with the *monotonicity axiom* (definition 2.1 item 1) to bound the tails and expectations of our desiderata. We use the uniform convergence bounds of section 3.1 to bound the error of these estimates, thus we need only propagate this uncertainty through the appropriate aggregator functions. In general, aggregator functions are nonlinear, and optimizing over $\mathcal{H}$ results in estimation bias, thus the plug-in estimator is biased, however, we still obtain *tail bounds* on our objectives via $\mathrm{AEV}(\dots)$. Because the plug-in estimator is biased, we also consider various LCB-and-UCB-style estimates, which when optimized yield safer function choices and partially control for overfitting. Finally, in some cases, integrating over worst-case uncertainty from the tail bounds of $\mathrm{AEV}(\dots)$ yields convenient bounds on the expectation (and thus the bias) of the plug-in estimator.

Welfare and Malfare Due to the lack of an unbiased estimator for welfare and malfare, we study the simple plug-in estimator $\hat{\mathrm{M}}$, as employed by [Cousins and Riondato, 2020], and introduce a pair of lower and upper estimators $(\hat{\mathrm{M}}^{\downarrow},\hat{\mathrm{M}}^{\uparrow})$. In particular, we take

$$\hat{\mathrm{M}} \doteq \underbrace{\mathrm{M}\left(i\mapsto \mathop{\hat{\mathbb{E}}}_{\boldsymbol{x}_{i,:},\boldsymbol{y}_{i,:}}[\mathrm{s}\circ h];\boldsymbol{w}\right)}_{\text{Plug-In Estimate}} \ , \ \ \hat{\mathrm{M}}^{\downarrow} \doteq \underbrace{\mathrm{M}\left(i\mapsto 0 \vee \mathop{\hat{\mathbb{E}}}_{\boldsymbol{x}_{i,:},\boldsymbol{y}_{i,:}}[\mathrm{s}\circ h] - \hat{\boldsymbol{\varepsilon}}_i;\boldsymbol{w}\right)}_{\text{LCB Estimate}} \ , \ \ \& \ \hat{\mathrm{M}}^{\uparrow} \doteq \underbrace{\mathrm{M}\left(i\mapsto c \wedge \mathop{\hat{\mathbb{E}}}_{\boldsymbol{x}_{i,:},\boldsymbol{y}_{i,:}}[\mathrm{s}\circ h] + \hat{\boldsymbol{\varepsilon}}_i;\boldsymbol{w}\right)}_{\text{UCB Estimate}} \ , \tag{8}$$

where $\vee$ and $\wedge$ are the (minimum precedence) infix binary maximum and minimum operators, respectively. By monotonicity (axiom 1), it holds that $\hat{\mathrm{M}}^{\downarrow} \leq \hat{\mathrm{M}} \leq \hat{\mathrm{M}}^{\uparrow}$. The *lower* and *upper confidence bound* estimates are convenient, both to show high probability bounds, and to sandwich the plug-in estimator, which we use to bound its bias. We first show tail bounds for the estimation of welfare and malfare in terms of their plug-in, LCB, and UCB estimates, and we then bound the bias of $\hat{\mathrm{M}}$. We note that this approach may seem backwards, as often tail bounds are given in terms of expectations. In this setting, due to the bias of each estimator, we employ tail-bound integration methods to bound expectations, hence the primary position of tail bounds.

Theorem 3.4 (Welfare and Malfare Tail Bounds)**.** Suppose sentiment function $\mathrm{s}(\cdot,\cdot) : \mathcal{Y}' \times \mathcal{Y} \to \mathbb{R}_{0+}$, per-group probability distributions $\mathcal{D}_{1:g}$, sample size vector $\boldsymbol{m} \in \mathbb{Z}_+^g$, samples $(\boldsymbol{x}, \boldsymbol{y}) \sim \mathcal{D}_1^{\boldsymbol{m}_1} \times \cdots \times \mathcal{D}_g^{\boldsymbol{m}_g}$, failure probability $\delta \in (0,1)$, and additive error bound $\mathrm{AEV}(\dots)$, and let $\hat{\boldsymbol{\varepsilon}} \leftarrow \mathrm{AEV}(\boldsymbol{m}, \delta, \boldsymbol{x}, \boldsymbol{y})$. Then for all $h \in \mathcal{H}$ and all monotonic aggregator functions $\mathrm{M}(\cdot; \boldsymbol{w})$, it holds with probability at least $1-\delta$ over $\boldsymbol{x}$, $\boldsymbol{y}$, and $\hat{\boldsymbol{\varepsilon}}$ that

$$\underbrace{\mathrm{M}\left(i \mapsto 0 \vee \mathop{\mathbb{E}}_{\mathcal{D}_i}[\mathrm{s} \circ h] - \hat{\boldsymbol{\varepsilon}}_i; \boldsymbol{w}\right)}_{\textsc{True LB}} \leq \underbrace{\mathrm{M}\left(i \mapsto \mathop{\hat{\mathbb{E}}}_{\boldsymbol{x}_{i,:}, \boldsymbol{y}_{i,:}}[\mathrm{s} \circ h]; \boldsymbol{w}\right)}_{\textsc{Plug-In Estimate } \hat{\mathrm{M}}} \leq \underbrace{\mathrm{M}\left(i \mapsto c \wedge \mathop{\mathbb{E}}_{\mathcal{D}_i}[\mathrm{s} \circ h] + \hat{\boldsymbol{\varepsilon}}_i; \boldsymbol{w}\right)}_{\textsc{True UB}} \ , \quad \& \tag{9}$$

$$\underbrace{\mathrm{M}\left(i \mapsto 0 \vee \mathop{\hat{\mathbb{E}}}_{\boldsymbol{x}_{i,:}, \boldsymbol{y}_{i,:}}[\mathrm{s} \circ h] - \hat{\boldsymbol{\varepsilon}}_i; \boldsymbol{w}\right)}_{\textsc{LCB Estimate } \hat{\mathrm{M}}^{\downarrow}} \leq \underbrace{\mathrm{M}\left(i \mapsto \mathop{\mathbb{E}}_{\mathcal{D}_i}[\mathrm{s} \circ h]; \boldsymbol{w}\right)}_{\textsc{True Aggregate M}} \leq \underbrace{\mathrm{M}\left(i \mapsto c \wedge \mathop{\hat{\mathbb{E}}}_{\boldsymbol{x}_{i,:}, \boldsymbol{y}_{i,:}}[\mathrm{s} \circ h] + \hat{\boldsymbol{\varepsilon}}_i; \boldsymbol{w}\right)}_{\textsc{UCB Estimate } \hat{\mathrm{M}}^{\uparrow}} \ , \tag{10}$$

thus if $\mathrm{M}(\cdot; \boldsymbol{w})$ is λ-Lipschitz-continuous[3] w.r.t. some norm $\|\cdot\|_{\mathrm{M}}$, we have

$$\left| \underbrace{\mathrm{M}\left(i \mapsto \mathop{\hat{\mathbb{E}}}_{\boldsymbol{x}_{i,:}, \boldsymbol{y}_{i,:}}[\mathrm{s} \circ h]; \boldsymbol{w}\right)}_{\textsc{Plug-In Estimate}} - \underbrace{\mathrm{M}\left(i \mapsto \mathop{\mathbb{E}}_{\mathcal{D}_i}[\mathrm{s} \circ h]; \boldsymbol{w}\right)}_{\textsc{True Aggregate}} \right| \leq \lambda \|\hat{\boldsymbol{\varepsilon}}\|_{\mathrm{M}} \ . \tag{11}$$

From (10), we see that minimizing $\hat{\mathrm{M}}^{\uparrow}$ (or maximizing $\hat{\mathrm{M}}^{\downarrow}$) is in some sense a *safe choice*, as w.h.p. we can bound the true aggregate value in terms of the UCB or LCB. This idea is reminiscent of the *sample variance penalization* algorithm of Maurer and Pontil [2009], wherein ERM is supplanted by minimizing an *upper-bound* on risk; in that case with variance-dependent bounds, but here the bound depends on the structure of the malfare or welfare objective at hand. It should also be noted that while the final Lipschitz form (11) is concise and convenient for all Lipschitz-continuous aggregator functions (e.g., all $p \geq 1$ power-mean malfare functions, see theorem 2.2 item 3), it can be quite loose. For example, under $\pm$ uncertainty intervals, the egalitarian welfare $\mathrm{W}_{-\infty}(\langle 4\pm1, 9\pm8\rangle; \boldsymbol{w}) = \min(4\pm1, 9\pm8)$ must be on the interval 3 ± 2, despite (11) giving a confidence radius of 8. Thus while (11) is convenient for intuition and analysis, when possible (9) or (10) should be favored.

From Tail Bounds to Expectations While theorem 3.4 gives high-probability bounds on the gap between empirical and true welfare or malfare, it does not actually bound the expectation (and thus the *statistical bias*) of the plug-in estimator. Unlike many simple large deviation bounds, the expectation of the plug-in estimator $\hat{\mathrm{M}}$ does not even appear in the theorem. Nevertheless, we now bound the integral over a worst-case distribution of possibilities for both the lower and upper confidence bound estimators. This bounds the bias of the plug-in estimator, and then corollary 3.6 derives a particularly convenient form for these bounds using Bernstein-type variance-sensitive bounds.

Theorem 3.5 (Welfare and Malfare Expectation Bounds)**.** Suppose as in theorem 3.4, and assume also that $\mathrm{AEV}(\boldsymbol{m}, \delta, \boldsymbol{x}, \boldsymbol{y}) = \mathrm{AEV}(\boldsymbol{m}, \delta)$ is a deterministic *distribution-free* or *distribution-dependent* (but *not data-dependent*) bound. Then

$$\left| \mathrm{M} - \mathbb{E}[\hat{\mathrm{M}}] \right| \leq \mathbb{E}\left[|\mathrm{M} - \hat{\mathrm{M}}| \right] \leq \lambda \int_0^1 \left\| \mathrm{AEV}(\boldsymbol{m}, \delta) \right\|_{\mathrm{M}} \mathrm{d}\delta \ .$$

The above theorems give general recipes for bounding tails and expectations, so for demonstrative purposes, we instantiate them with theorem 3.1 for malfare estimation. Similar bounds can be derived for learning with theorem 3.2.

Corollary 3.6 (Bernstein-Type Malfare Bounds)**.** Suppose as in theorem 3.1, and also per-group sample size m (i.e., $\boldsymbol{m} = \langle m, \dots, m\rangle$) and $p \geq 1$ power-mean malfare function $\mathrm{M}_p(\cdot; \boldsymbol{w})$. Now, let variance proxy v be defined in three cases as $v \doteq \mathrm{M}_{1/2}(\boldsymbol{v}; \boldsymbol{w}) = \left(\sum_{i=1}^g \boldsymbol{w}_i \sqrt{\boldsymbol{v}_i}\right)^2$ for $p = 1$, $v \doteq \boldsymbol{w} \cdot \boldsymbol{v}$ for $p \in (1, 2]$, or $v \doteq \|\boldsymbol{v}\|_\infty$ for $p > 2$. Then for all $\delta \in (0,1)$, we have

[3] In other words, if for all $\mathcal{S}, \mathcal{S}'$, it holds that $\left|\mathrm{M}(\mathcal{S}; \boldsymbol{w}) - \mathrm{M}(\mathcal{S}'; \boldsymbol{w})\right| \leq \lambda \left\|\mathcal{S} - \mathcal{S}'\right\|_{\mathrm{M}}$.

1) $\mathbb{P}\left(\left|\mathrm{M} - \hat{\mathrm{M}}\right| \geq \frac{r \ln \frac{2g}{\delta}}{3m} + \sqrt{\frac{2v \ln \frac{2g}{\delta}}{m}}\right) \leq \delta$;

2) $\mathbb{E}\left[\left|\mathrm{M} - \hat{\mathrm{M}}\right|\right] \leq \frac{r \ln(2eg)}{3m} + \sqrt{\frac{2v \ln(2eg)}{m}}$; &

3) $\mathrm{M} \leq \mathbb{E}[\hat{\mathrm{M}}] \leq \mathrm{M} + \frac{r \ln(eg)}{3m} + \sqrt{\frac{2v \ln(eg)}{m}}$.

Estimating the Malfare of Regret Regret is difficult to bound, as it depends both on the expected sentiment of the selected $\hat{h}$, and also on $\mathcal{H}$ through the (unknown) per-group optimal sentiments $\mathcal{S}^\star_{1:g}$. We thus introduce the estimators

$$\hat{\mathcal{S}}_i \doteq \inf_{h \in \mathcal{H}} \hat{\mathbb{E}}_{\boldsymbol{x}_{i,:}, \boldsymbol{y}_{i,:}}[\ell \circ h] \ , \quad \text{or} \quad \hat{\mathcal{S}}_i \doteq \sup_{h \in \mathcal{H}} \hat{\mathbb{E}}_{\boldsymbol{x}_{i,:}, \boldsymbol{y}_{i,:}}[\mathrm{u} \circ h] \ , \tag{12}$$

for loss or utility (note that these are downward and upward biased estimates), respectively, cf. (2). By analogy with (3), the plug-in estimator for the regret malfare minimizer is then

$$\hat{h} \doteq \operatorname*{argmin}_{h \in \mathcal{H}} \mathrm{M}\left(i \mapsto \left|\hat{\mathbb{E}}_{\boldsymbol{x}_{i,:}, \boldsymbol{y}_{i,:}}[\mathrm{s} \circ h] - \hat{\mathcal{S}}_i\right|; \boldsymbol{w}\right) \ . \tag{13}$$

The following theorem bounds the difference between the true and empirical malfare of regret.

Theorem 3.7 (Regret Estimation Bounds)**.** Suppose sentiment function $\mathrm{s}(\cdot,\cdot) \colon \mathcal{Y}' \times \mathcal{Y} \to \mathbb{R}_{0+}$, per-group probability distributions $\mathcal{D}_{1:g}$, sample size vector $\boldsymbol{m} \in \mathbb{Z}_+^g$, samples $(\boldsymbol{x}, \boldsymbol{y}) \sim \mathcal{D}_1^{\boldsymbol{m}_1} \times \cdots \times \mathcal{D}_g^{\boldsymbol{m}_g}$, failure probability $\delta \in (0, 1)$, and additive error bound $\mathrm{AEV}(\dots)$, and let $\hat{\boldsymbol{\varepsilon}} \leftarrow \mathrm{AEV}(\boldsymbol{m}, \delta, \boldsymbol{x}, \boldsymbol{y})$. Then for all $h \in \mathcal{H}$ and all monotonic malfare functions $\mathrm{M}(\cdot; \boldsymbol{w})$, it holds with probability at least $1 - \delta$ over $\boldsymbol{x}$, $\boldsymbol{y}$, and $\hat{\boldsymbol{\varepsilon}}$ that

$$\underbrace{\mathrm{M}\left(i \mapsto 0 \vee \left|\mathbb{E}_{\mathcal{D}_i}[\mathrm{s} \circ h] - \mathcal{S}^\star_i\right| - 2\hat{\boldsymbol{\varepsilon}}_i; \boldsymbol{w}\right)}_{\text{True Regret Malfare LB}} \leq \underbrace{\mathrm{M}\left(i \mapsto \left|\hat{\mathbb{E}}_{\boldsymbol{x}_{i,:}, \boldsymbol{y}_{i,:}}[\mathrm{s} \circ h] - \hat{\mathcal{S}}_i\right|; \boldsymbol{w}\right)}_{\text{Plug-In Regret Malfare}} \leq \underbrace{\mathrm{M}\left(i \mapsto c \wedge \left|\mathbb{E}_{\mathcal{D}_i}[\mathrm{s} \circ h] - \mathcal{S}^\star_i\right| + 2\hat{\boldsymbol{\varepsilon}}_i; \boldsymbol{w}\right)}_{\text{True Regret Malfare UB}} \ , \ \& \tag{14}$$

$$\underbrace{\mathrm{M}\left(i \mapsto 0 \vee \left|\hat{\mathbb{E}}_{\boldsymbol{x}_{i,:}, \boldsymbol{y}_{i,:}}[\mathrm{s} \circ h] - \hat{\mathcal{S}}_i\right| - 2\hat{\boldsymbol{\varepsilon}}_i; \boldsymbol{w}\right)}_{\text{LCB Estimate}} \leq \underbrace{\mathrm{M}\left(i \mapsto \left|\mathbb{E}_{\mathcal{D}_i}[\mathrm{s} \circ h] - \mathcal{S}^\star_i\right|; \boldsymbol{w}\right)}_{\text{True Regret Malfare}} \leq \underbrace{\mathrm{M}\left(i \mapsto c \wedge \left|\hat{\mathbb{E}}_{\boldsymbol{x}_{i,:}, \boldsymbol{y}_{i,:}}[\mathrm{s} \circ h] - \hat{\mathcal{S}}_i\right| + 2\hat{\boldsymbol{\varepsilon}}_i; \boldsymbol{w}\right)}_{\text{UCB Estimate}} \ , \tag{15}$$

thus if $\mathrm{M}(\cdot; \boldsymbol{w})$ is λ-Lipschitz-continuous w.r.t. some norm $\|\cdot\|_{\mathrm{M}}$, we have

$$\left|\underbrace{\mathrm{M}\left(i \mapsto \left|\hat{\mathbb{E}}_{\boldsymbol{x}_{i,:}, \boldsymbol{y}_{i,:}}[\mathrm{s} \circ h] - \hat{\mathcal{S}}_i\right|; \boldsymbol{w}\right)}_{\text{Plug-In Regret Malfare}} - \underbrace{\mathrm{M}\left(i \mapsto \left|\mathbb{E}_{\mathcal{D}_i}[\mathrm{s} \circ h] - \mathcal{S}^\star_i\right|; \boldsymbol{w}\right)}_{\text{True Regret Malfare}}\right| \leq 2\lambda \|\hat{\boldsymbol{\varepsilon}}\|_{\mathrm{M}} \ . \tag{16}$$

Note that similar bounds on the *expectation* of the regret plug-in estimator can be shown along the lines of corollary 3.6, *mutatis mutandis* for regret. Note also that theorem 3.7 matches theorem 3.4 up to a 2-factor attached to the confidence radius, thus in some sense regret is "about twice as difficult" to estimate as malfare or welfare.

3.3 Information Asymmetry and Where Best to Sample

An intuitive notion of fairness would suggest that we should draw equally-sized samples for each group, or perhaps samples proportional to population frequencies. If the goal is to optimize or bound welfare, malfare, or regret, such intuitive notions should be rejected, as they are critically flawed. We now discuss the ways in which samples drawn from one group or another may be more or less valuable to for the purposes of estimating or optimizing these objectives.

As a brief thought experiment, suppose we want to estimate the egalitarian welfare of a population consisting of two groups. Suppose also that their utilities are similarly difficult to estimate, and their expected utilities are $\langle 1, 10 \rangle$. In such a setting, nearly all sampling effort should be invested in estimating the utility of group one, as once group two is estimated to within ± 9 additive error, there is *no further benefit* to improving their estimate. Thus the optimal (as measured by sufficient sample size) sampling strategy depends on the true expected utilities, the difficulties of estimating utilities for each group, and the objective in question, and *in no way* resembles the naïve uniform or proportional so-called "fair sampling strategies" described above. We argue that such naïve strategies are dangerous, as they introduce subtle biases and fairness issues, but the rationale for alternative sampling strategies is only apparent through the lens of sample complexity.

We now ask the questions, "Given a sample, what do we need to obtain sharper bounds?" and "How much will bounds improve with a larger sample?" We begin with a soft discussion as to why samples from different groups may contribute more or less information to an estimate, which we measure as the improvement to tail bounds that additional samples may yield. In particular, for malfare, we discuss the improvement to upper bounds, but the entire discussion can be directly translated to welfare and lower bounds in the usual manner. We then quantify these factors mathematically, and we develop these ideas further in section 4.3, where they are used to adaptively choose from which group to sample. Answering these questions tells us what fundamentally is needed to solve an estimation or optimization task, which allows us to intelligently guide scientific inquiry by directing limited resources towards studying the most informative populations, where informativeness is measured by the *amount of improvement* to tail bounds.

Philosophical Discussion We now discuss the three main factors driving heterogeneity in sampling impact.

1) *Variable estimation difficulty or overfit potential*: Often it is inherently more difficult to give bounds on the expected sentiment for some groups than for others. This can be due to differences in variances (see theorem 3.1) or in uniform convergence bounds (see theorem 3.2), and in general, occurs when $\hat{\boldsymbol{\varepsilon}} \leftarrow \mathrm{AEV}(\dots)$ has $\hat{\boldsymbol{\varepsilon}}_i \ll \hat{\boldsymbol{\varepsilon}}_j$ for some group $i \neq j$, even while $\boldsymbol{m}_i \approx \boldsymbol{m}_j$. There are many possible causes of such asymmetries, but we now give two examples in the machine learning sphere:
   A) The class $\mathcal{H}$ can be less complicated when projected onto group i than on group j, e.g., if $\mathcal{H}$ is linear classifiers, and group i exists in a lower-dimensional subspace than group j (see item 1).
   B) Group i may itself be more self-similar (less diverse) than group j; for instance with tabular data, there may simply be no members of group i that attain certain feature values, which would for instance, impact bounds for decision trees. It is entirely possible for minority groups to exhibit either more or less intra-group diversity than a majority group, thus this effect can work in either direction.
2) *Variable task difficulty*: Some groups may be inherently easier or harder to satisfy than others; e.g., regression and classification problems are generally easier for groups with labels that are more homogeneous, and regret varies with the optimal expected sentiment $\mathcal{S}_i^\star$. Similarly, for recommender systems, if a group is generally satisfied by a larger number of options, they will generally have lower risk. This is crucial, because most malfare and welfare functions are more sensitive[4] to high-risk or low-utility groups, thus the ease of satisfying a group effects their impact on malfare and welfare values.
3) *Aggregator function interactions*: Complicated interactions also occur through the malfare or welfare function. When learning over $\mathcal{H}$, the set of near-optimal functions is more relevant than those that are clearly bad choices overall, and groups that tend to be mutually satisfied (i.e., are correlated) are less impactful to the overall objective. Weight values in malfare or welfare functions may also differ between groups, and higher-weighted groups are usually more impactful.

Quantifying the Incremental Value of Sampling We measure the impact of sampling by asking the question, "What is the incremental value of a single sample drawn for some group?" In particular, we quantify the value of the sample as the *reduction in uncertainty*, as measured by the infimum UCB (over $\mathcal{H}$), and although this is inherently a discrete question, we approximate the answer for the power-mean malfare with tools from the calculus of infinitesimals. We consider the power-mean malfare family for its simplicity and convenient differentiability properties, but similar analysis is possible for welfare or regret bounds. For the sake of intuition, we lead by presenting a parametric Gaussian example in figure 3.

[4] In particular, this holds for all $p \neq 1$ power means, and is axiomatically justified by the *Pigou-Dalton transfer principle* (definition 2.1 item 7).

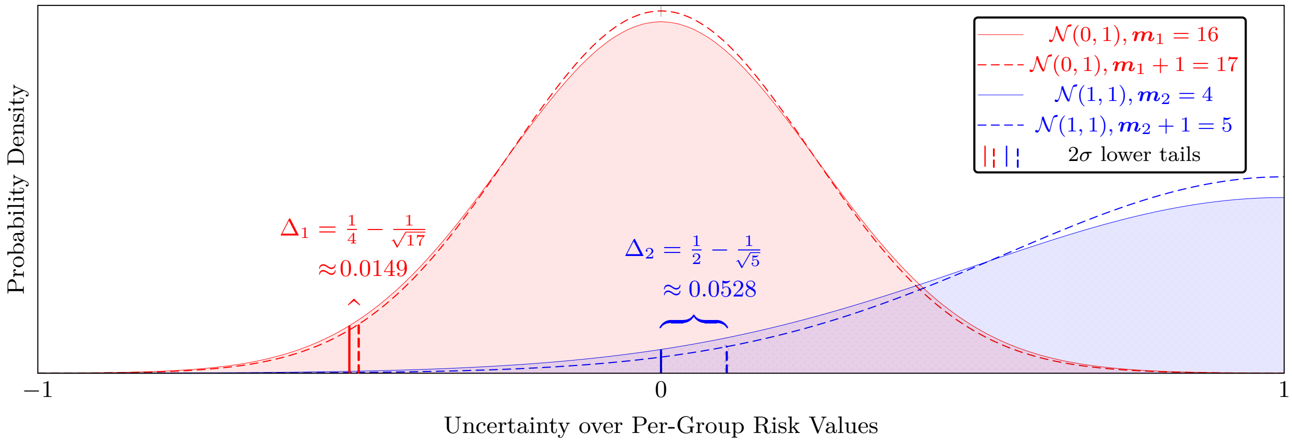


Figure 3: In this example, rather than using the general tail bounds of section 3.1, we use a simple frequentist[◇] Gaussian analysis, where loss values for each $\mathcal{D}_i$ are distributed $\mathcal{N}(\boldsymbol{\mu}_i, 1)$, for unknown means $\boldsymbol{\mu} \doteq \langle 1, 2\rangle$. Here we obtain $\approx 95\%$ confidence intervals with 2σ tails, and plot the lower tails of groups 1 and 2, of sizes $\boldsymbol{m} \doteq \langle 16, 4\rangle$. Letting Δ_i denote the improvement to bound radius for group i, and assuming equal sampling cost $\boldsymbol{C}_1 = \boldsymbol{C}_2$, we see that the reduction in uncertainty for the *unweighted utilitarian malfare* objective is optimized by sampling group 2, with a reduction of $\frac{\Delta_1}{2}$, and the reduction for the *egalitarian* objective is optimized by sampling group 1, with a reduction of Δ_2.

[◇] We can of course relax the assumption of known variance via confidence estimation through the Student's t-distribution. Note also that similar analysis is possible in a Bayesian framework, e.g., if we can assume a Gaussian conjugate prior [DeGroot, 1970, Gelman et al., 2004].

Note that all such analysis is necessarily heuristic, as we fundamentally cannot answer this question without more information: it is precisely because we are trying to *estimate unknown means* that we can't know how the samples we draw will impact the *empirical means*. For now, we heuristically assume that our estimated expectations are reasonably accurate, and consider what will happen as tail bounds sharpen with additional samples. The strategy we thus employ is to make a reasonable guess as to how sampling might impact the UCB by assuming that the empirical mean will not be strongly affected, and all confidence intervals over m samples will contract at a $\boldsymbol{\Theta}\sqrt{\frac{1}{m}}$ rate.

Property 3.8 (Incremental Gain of Sampling)**.** Suppose $\boldsymbol{w}$-weighted power-mean malfare $\mathrm{M}_p(\cdot; \boldsymbol{w})$, sample $(\boldsymbol{x}, \boldsymbol{y})$ with group sample sizes $\boldsymbol{m}_{1:g}$, and let $\boldsymbol{x}', \boldsymbol{y}'$ extend $\boldsymbol{x}, \boldsymbol{y}$ to sample sizes $\boldsymbol{m}'$, where $\boldsymbol{m}' = \boldsymbol{m} + \mathbb{1}_i$, i.e., group i has one additional sample. Now, let $\hat{\boldsymbol{\varepsilon}} \leftarrow \mathrm{AEV}(\boldsymbol{m}, \delta, \boldsymbol{x}, \boldsymbol{y})$ and $\tilde{\boldsymbol{\varepsilon}} \leftarrow \mathrm{AEV}(\boldsymbol{m}', \delta, \boldsymbol{x}', \boldsymbol{y}')$, and take $\hat{h} \doteq \operatorname{argmin}_{h \in \mathcal{H}} \mathrm{M}_p\Big(i \mapsto \hat{\mathbb{E}}_{\boldsymbol{x}_{i,:}, \boldsymbol{y}_{i,:}}[\ell \circ h] + \hat{\boldsymbol{\varepsilon}}_i; \boldsymbol{w}\Big)$, $\hat{\mathrm{M}} \doteq \mathrm{M}_p\Big(i \mapsto \hat{\mathbb{E}}_{\boldsymbol{x}_{i,:}, \boldsymbol{y}_{i,:}}[\ell \circ \hat{h}]; \boldsymbol{w}\Big)$, $\hat{\mathrm{M}}^{\uparrow} \doteq \mathrm{M}_p\Big(i \mapsto \hat{\mathbb{E}}_{\boldsymbol{x}_{i,:}, \boldsymbol{y}_{i,:}}[\ell \circ \hat{h}] + \hat{\boldsymbol{\varepsilon}}_i; \boldsymbol{w}\Big)$, and $\tilde{\mathrm{M}}^{\uparrow} \doteq \inf_{h \in \mathcal{H}} \mathrm{M}_p\Big(i \mapsto \hat{\mathbb{E}}_{\boldsymbol{x}'_{i,:}, \boldsymbol{y}'_{i,:}}[\ell \circ h] + \tilde{\boldsymbol{\varepsilon}}_i; \boldsymbol{w}\Big)$.

Then the *incremental impact* of sampling from group i on the UCB is approximately

$$\hat{\mathrm{M}}^{\uparrow} - \tilde{\mathrm{M}}^{\uparrow} \approx \frac{\hat{\boldsymbol{\varepsilon}}_i \boldsymbol{w}_i}{2\boldsymbol{m}_i + \frac{3}{2}} \left(\frac{\hat{\mathbb{E}}_{\boldsymbol{x}_{i,:}, \boldsymbol{y}_{i,:}}[\ell \circ \hat{h}] + \hat{\boldsymbol{\varepsilon}}_i}{\hat{\mathrm{M}}^{\uparrow}} \right)^{p-1} \approx \frac{\hat{\boldsymbol{\varepsilon}}_i \boldsymbol{w}_i}{2\boldsymbol{m}_i} \left(\frac{\hat{\mathbb{E}}_{\boldsymbol{x}_{i,:}, \boldsymbol{y}_{i,:}}[\ell \circ \hat{h}]}{\hat{\mathrm{M}}} \right)^{p-1} . \tag{17}$$

Observe that (17) characterizes the knowledge gain of sampling from group i. This gain is *proportional* to the current bound radius $\hat{\boldsymbol{\varepsilon}}_i$, the group weight $\boldsymbol{w}_i$, and the $(p-1)$th power of the ratio of the UCB risk of group i to the UCB malfare, i.e., $(\hat{\mathbb{E}}_{\boldsymbol{x}_{i,:}, \boldsymbol{y}_{i,:}}[\ell \circ \hat{h}] + \hat{\boldsymbol{\varepsilon}}_i / \hat{\mathrm{M}}^{\uparrow})^{p-1}$, and *inversely proportional* to the amount of effort $\boldsymbol{m}_i$ already put forth into studying group i. These terms line up with the soft arguments at the top of section 3.3 as to where sampling should occur, but it is only via precisely studying sample complexity and estimation error that we gain quantifiable mathematical insight. In particular, the weight term $\boldsymbol{w}_i$ appears directly, and $\frac{\hat{\boldsymbol{\varepsilon}}_i}{\boldsymbol{m}_i}$ captures both the difficulty of estimating this group, and also the diminishing incremental improvement produced by further sampling. The ratio between the risk of group i and the malfare then captures how important group i is relative to the other groups, and this term being raised to the $(p-1)$th power nonlinearly adjusts its impact; higher p saturate high-risk groups, tending towards egalitarianism, wherein the most disadvantaged group becomes the most important, whereas in the $p = 1$ (utilitarian) case, this term is 1. Finally, for optimization problems, the dependence on $\hat{h}$ captures other dependencies; namely the behavior of $\mathrm{M}(\cdot; \boldsymbol{w})$ *near the optimal* $h \in \mathcal{H}$ is what matters.

This analysis parallels concerns in the *stratified sampling* regime, wherein subpopulations are sampled individually, generally to produce an improved mean estimator. In particular, we suggest a form of *disproportionate allocation*, i.e., per-group sample sizes are not necessarily proportional to their population frequencies. Rather than simply considering *variances* to estimate *means*, we holistically consider the objective and uncertainty over various quantities, thus our sample-size selection-strategy is a variant of the *minimax sampling ratio* [Shahrokh Esfahani and Dougherty, 2014] method. Chen et al. [2018] also suggest disproportionate allocation in fair machine learning, albeit only for bounding absolute differences of per-group fairness statistics. Similar concerns also arise in *optimizing* minimax-fair models, wherein Abernethy et al. [2020] present an algorithm that takes gradient steps to improve a model for the highest-risk group, though it is unclear whether such methods generalize beyond the egalitarian case.

4 Progressive and Active Sampling Algorithms

Section 3 considers *fixed sample sizes* $\boldsymbol{m}_{1:g}$ and *failure probabilities* δ, and bounds the *confidence radius* ε. In this section, we want a fixed ε-δ additive error guarantee, but we are willing to let an algorithm select the sample size m (or per-group sample sizes $\boldsymbol{m}_{1:g}$). In particular, due to the cost of sampling and processing data, we want our algorithm to minimize m (or cost measured as some function of $\boldsymbol{m}$), while constraining ε and δ to user-supplied levels. Some cases are simpler than others; the *joint sampling model* yields a standard progressive sampling method with a fixed sampling schedule, and the method under *mixture sampling* is similar, except a subtle conditioning argument allows us to use variably-sized per-group sample sizes based on the order groups are sampled in. For the *conditional sampling model*, we develop an *active sampling* approach, which makes cost-sensitive decisions as to which group to sample at each iteration. More details on sampling schedules and other aspects of our progressive sampling algorithms are given in appendix B.

4.1 Convergence and Sampling Schedules in Progressive Sampling

We can't simply draw samples one-by-one, compute bounds using $\hat{\boldsymbol{\varepsilon}} \leftarrow \mathrm{AEV}(\boldsymbol{m}, \delta, \boldsymbol{x}, \boldsymbol{y})$ after each sample, and terminate when a sufficiently sharp bound is available, because the possibility of *early termination* leads to the *multiple comparisons problem*, wherein by chance the desired confidence radius is met at some timestep. Progressive sampling algorithms correct for this by establishing, usually *a priori*, a *sampling schedule* $\boldsymbol{s}$ and *failure probability schedule* $\boldsymbol{\delta}$, which usually dictate that, at timestep t, we take a tail-bound with $\delta = \boldsymbol{\delta}_t$ and sample size $\boldsymbol{s}_t$, while ensuring that all bounds hold simultaneously (by union bound) with probability at least $1 - \delta$. Due to this union bound, it is *statistically inefficient* to take bounds after drawing every sample.[5] Furthermore, for technical reasons, we henceforth assume a few mild regularity conditions:

1) The *sampling schedule* $\boldsymbol{s} \in \mathbb{Z}_+^\infty$ is a strictly monotonically increasing sequence, i.e., for all $t \in \mathbb{Z}_+$, $\boldsymbol{s}_t \leq \boldsymbol{s}_{t+1}$;
2) The *failure probability schedule* $\boldsymbol{\delta} \in [0,1)^\infty$ is a sequence that sums to some $\delta \in (0,1)$, i.e., $\sum_{i=1}^{\infty} \boldsymbol{\delta}_t = \|\boldsymbol{\delta}\|_1 = \delta$; &
3) The *distribution-free bound*[6] $\sup_{\boldsymbol{x},\boldsymbol{y}} \|\mathrm{AEV}(\boldsymbol{m}, \delta, \boldsymbol{x}, \boldsymbol{y})\|$ is monotonically decreasing in $\boldsymbol{m}_{1:g}$ and δ for any norm $\|\cdot\|$.

In order to prove that a progressive sampling algorithm produces a (probabilistically) correct answer, it is crucial to show that it does not loop indefinitely. We now introduce ε-convergent schedules, which require all sentiment values to eventually be ε-δ estimated w.r.t. some norm $\|\cdot\|_{\mathrm{M}}$, which we then translate into guarantees on welfare, malfare, or regret, as per theorems 3.4 and 3.7.

Definition 4.1 (ε-Uniformly-Convergent Schedule)**.** For any $\varepsilon \geq 0$, a sampling schedule $\boldsymbol{s}$ and failure probability schedule $\boldsymbol{\delta}$ are ε-uniformly-convergent w.r.t. $\mathrm{AEV}(\dots)$ and some norm $\|\cdot\|_{\mathrm{M}}$ if

$$\inf_{t \in \mathbb{Z}_+} \sup_{(\boldsymbol{x},\boldsymbol{y}) \in (\mathcal{X} \times \mathcal{Y})^{\boldsymbol{s}_t \times g}} \left\|\mathrm{AEV}(\langle \boldsymbol{s}_t, \dots, \boldsymbol{s}_t \rangle, \boldsymbol{\delta}_t, \boldsymbol{x}, \boldsymbol{y})\right\|_{\mathrm{M}} \leq \varepsilon \ . \tag{18}$$

[5] Indeed, in some of the earliest progressive sampling work, John and Langley [1996] did employ *arithmetic schedules*, i.e., those with constant differences between successive sample sizes. We discuss only the more efficient *geometric schedules* introduced by Provost et al. [1999]. Early work in progressive sampling had no failure probability schedules $\boldsymbol{\delta}$, considering statistical error and multiple comparisons only through heuristic convergence estimation techniques, thus the statistical cost of performing this correction was not yet fully appreciated.

[6] Note that in the case where $\hat{\boldsymbol{\varepsilon}}$ is a *random variable*, as in (6) and the related discussion in the prologue of section 3.1, we assume this property holds *with certainty*, which can be achieved by combining the radomized data-dependent bound with a worst-case distribution-free bound. Alternatively, if it does not hold *with certainly*, but rather *almost certainly*, or even *in probability*, the subsequent analysis largely applies, albeit with these inherited probabilistic guarantees.

Intuitively, definition 4.1 captures the idea that no matter how unlucky we are with the sampled $\boldsymbol{x}, \boldsymbol{y}$, if AEV(...) bounds tails once for each timestep t of the schedule, with per-group samples of size at least $\boldsymbol{s}_t$ and failure probability $\boldsymbol{\delta}_t$, then at some point an ε-estimate of the objective will be produced. Note that neither data-dependent AEV(...) bounds on sentiment values, nor sufficient per-group error radii to estimate the objective, are known *a priori*, thus it is not always possible to select a *sufficient* static sample size, however, definition 4.1 is *more flexible*, as it requires *only the existence* of a (possibly unknown) sufficient sample size. Even when a sufficient sample size is known, unless it is also *necessary*, progressive sampling is usually more sample-efficient, often terminating closer to the *necessary sample size*.

With this definition in hand, we now construct finite ε-, and infinite 0-, uniformly-convergent schedules. In the context of this work (see theorems 4.5 and 4.6), the finite schedule can be employed with a Lipschitz-continuous objective and an *a priori* known distribution-free bound on AEV(...), and when the objective is continuous, but not Lipschitz-continuous (e.g., the *geometric welfare* $\mathrm{W}_0(\cdot; \boldsymbol{w})$, see theorem 2.2 item 1), or the class $\mathcal{H}$ is uniformly-convergent, but at an unknown rate (e.g., sparse linear classifiers in an unknown bounded dimension, see property 3.3 item 1), the infinite schedule can still be used. Both are based on geometrically-increasing sample sizes, which are efficient because they never "overshoot" any sample size by more than a constant factor, yet they cover an exponentially large range of sample sizes in a linear number of timesteps.

Definition 4.2 (Geometric-Uniform Schedule)**.** Suppose *optimistic size* $\alpha \geq 1$, *common ratio* $\beta > 1$, and *schedule length* $T \in \mathbb{Z}_+$. The geometric-uniform schedule then takes (geometric) $\boldsymbol{s}_t \doteq \lceil \alpha \beta^t \rceil$ and (uniform) $\boldsymbol{\delta}_t \doteq \frac{\delta}{T} \mathbb{1}_{1,\dots,T}(t)$.

Definition 4.3 (Double-Geometric Schedule)**.** Suppose *optimistic size* $\alpha > 0$ and *common ratio* $\beta > 1$. The double-geometric schedule then takes (geometric) $\boldsymbol{s}_t \doteq \lceil \alpha \beta^t \rceil$ and (geometric) $\boldsymbol{\delta}_t \doteq \frac{\delta(\beta-1)}{\beta^t}$.

Lemma 4.4 (Sufficient Conditions for Uniformly-Convergent Geometric Schedules)**.** Suppose as in definition 4.2, and assume also that

$$\sup_{(\boldsymbol{x},\boldsymbol{y}) \in (\mathcal{X} \times \mathcal{Y})^{\boldsymbol{s}_T \times g}} \left\| \mathrm{AEV}\left(\langle \boldsymbol{s}_T, \dots, \boldsymbol{s}_T \rangle, \tfrac{\delta}{T}, \boldsymbol{x}, \boldsymbol{y}\right) \right\|_{\mathrm{M}} \leq \varepsilon \ . \tag{19}$$

Then the geometric-uniform schedule $(\boldsymbol{s}, \boldsymbol{\delta})$ is ε-uniformly-convergent.

Furthermore, suppose as in definition 4.3, $\alpha \geq \frac{1}{\delta}$, and assume that

$$\lim_{m \to \infty} \sup_{(\boldsymbol{x},\boldsymbol{y}) \in (\mathcal{X} \times \mathcal{Y})^{m \times g}} \left\| \mathrm{AEV}\left(\langle m, \dots, m \rangle, \tfrac{\beta-1}{\beta(m+1)}, \boldsymbol{x}, \boldsymbol{y}\right) \right\|_{\mathrm{M}} = 0 \ . \tag{20}$$

Then the double-geometric schedule $(\boldsymbol{s}, \boldsymbol{\delta})$ is 0-uniformly-convergent.

The initial and final sample sizes of the geometric-uniform schedule are $\boldsymbol{s}_1 = \lceil \alpha\beta \rceil$ and $\boldsymbol{s}_T = \lceil \alpha \beta^T \rceil$, and often one can set ${}^{\boldsymbol{s}_1}/_{\beta}$ and $\boldsymbol{s}_T$ to *minimal sufficient* and *maximal necessary* sample sizes (as a function of T, the objective, and other parameters). To maximize *statistical efficiency* while controlling the value of β, we may select the minimal T such that $\lceil \log_\beta \frac{\boldsymbol{s}_T}{\boldsymbol{s}_1} \rceil = T$.[7] In particular, assuming a λ-Lipschitz objective, the Hoeffding (item 1) and empirical Bernstein (item 3) bounds of theorem 3.1 imply ε-uniformly convergent schedules via (19) of length $T \in \mathbf{\Theta}(\log \frac{\lambda r}{\varepsilon})$. For the double-geometric schedule, we may similarly set ${}^{\boldsymbol{s}_1}/_{\beta}$ to a minimal sufficient sample size, and here there is no T parameter (the schedule is infinite), thus we may simply select β as desired. This yields 0-uniformly convergent schedules, since each of the bounds of theorem 3.1 satisfy (20), as do those of theorem 3.2, so long as $\lim_{m \to \infty} \max_{i \in \mathcal{Z}} \mathfrak{R}_m(\mathrm{s} \circ \mathcal{H}, \mathcal{D}_i) = 0$.

Both of the above schedule types are *efficient*, in the sense that for the smallest (per-group) static sample size $m^\star$ at which we obtain the bound $\boldsymbol{\varepsilon}^\star$, some $\hat{m} \leq \lceil \beta m^\star \rceil$ is contained in the schedule at some timestep t, and the bound $\hat{\boldsymbol{\varepsilon}} \leftarrow \mathrm{AEV}(\langle \hat{m}, \dots, \hat{m} \rangle, \boldsymbol{\delta}_t, \dots)$ exceeds $\boldsymbol{\varepsilon}^\star$ only because it uses a smaller δ value, i.e., because $\boldsymbol{\delta}_t < \delta$. In particular, assuming all bounds are asymptotically $\mathbf{\Theta}\sqrt{u}$ for $u \doteq \ln \frac{g}{\delta}$, we have for each group i that

$$\frac{\boldsymbol{\varepsilon}^\star_i}{\hat{\boldsymbol{\varepsilon}}_i} \in \mathbf{\Theta}\sqrt{\frac{u}{\log(T) + u}} \ , \quad \& \quad \frac{\boldsymbol{\varepsilon}^\star_i}{\hat{\boldsymbol{\varepsilon}}_i} \in \mathbf{\Theta}\sqrt{\frac{u}{\log(m^\star) + u}} \ . \tag{21}$$

for the geometric-uniform and double-geometric schedules, respectively. Note also that $\log(T) \in \mathbf{\Theta}(\log\log \frac{r\lambda}{\varepsilon})$, whereas $\log(m^\star) \in \mathbf{O}(\log \frac{r\lambda u}{\varepsilon})$, thus (21) shows us that the geometric-uniform schedule is preferable to the double-uniform schedule, unless $m^\star$ is *exponentially smaller* than the above bound, e.g., if $\lambda = \infty$, or if a nonlinear objective is more stable to perturbations of each $\mathcal{S}_i$ about its optimum than the Lipschitz constant λ would indicate.

[7]The base-β logarithm arises intuitively, as the *number of times* the sample size must increase by a factor β to reach size $\boldsymbol{s}_T$ from $\boldsymbol{s}_1$.

Algorithm 1 Fair Learning with Linear Progressive Sampling under the Joint and Mixture Sampling Models

1: **procedure** LinearPSLoss$(\mathcal{H}, \ell(\cdot,\cdot), \mathcal{D}, \mathrm{AEV}(\dots), \boldsymbol{s}, \boldsymbol{\delta}, \varepsilon, \mathrm{M}(\cdot; \boldsymbol{w}), \textsc{Reg}) \to (\hat{h}, \hat{\mu}, \hat{\varepsilon}, \mathrm{M}^{\star\downarrow})$
2: **input**: Hypothesis class $\mathcal{H} \subseteq \mathcal{X} \to \mathcal{Y}'$, loss function $\ell(\cdot,\cdot) : \mathcal{Y}' \times \mathcal{Y} \to [0, c]$, joint or mixture distribution $\mathcal{D}$, additive error vector bound $\mathrm{AEV}(\boldsymbol{m}, \delta, \boldsymbol{x}, \boldsymbol{y})$, schedule $\boldsymbol{s} \in \mathbb{Z}_+^\infty$ and $\boldsymbol{\delta} \in [0,1)^\infty$, confidence radius ε, weighted malfare $\mathrm{M}(\cdot; \boldsymbol{w})$, and Boolean Reg
3: **output**: Empirically UCB-optimal $\hat{h}$, empirical malfare estimate $\hat{\mu}$, confidence radius $\hat{\varepsilon}$, and lower bound on minimal malfare $\mathrm{M}^{\star\downarrow}$
4: $\boldsymbol{m}_{1:g} \leftarrow \boldsymbol{0}$; $\boldsymbol{x}_{1:g} \leftarrow \langle\langle\rangle, \dots, \langle\rangle\rangle$; $\boldsymbol{y}_{1:g} \leftarrow \langle\langle\rangle, \dots, \langle\rangle\rangle$ ▷ Initialize per-group sample counts, empty per-group sample lists
5: **for** $t \in 1, 2, \dots$ **do** ▷ Progressive sampling timesteps
6: **if** $\mathcal{D}$ is joint sampler **then**
7: $\big(\boldsymbol{x}_{1:g, \boldsymbol{s}_{t-1}+1:\boldsymbol{s}_t}, \boldsymbol{y}_{1:g, \boldsymbol{s}_{t-1}+1:\boldsymbol{s}_t}\big) \sim \mathcal{D}^{\boldsymbol{s}_t - \boldsymbol{s}_{t-1}}$; $\forall i \in \mathcal{Z}\colon\ \boldsymbol{m}_i \leftarrow \boldsymbol{s}_t$ ▷ Sample from joint distribution (assume $\boldsymbol{s}_0 = 0$)
8: **else if** $\mathcal{D}$ is mixture sampler **then**
9: **while** $\min_i \boldsymbol{m}_i < \boldsymbol{s}_t$ **do**
10: $(x, y, \boldsymbol{z}) \sim \mathcal{D}$ ▷ Draw $\mathcal{X} \times \mathcal{Y} \times 2^{\mathcal{Z}}$ triplet (domain, codomain, groups)
11: $\forall i \in \boldsymbol{z}\colon\ \boldsymbol{m}_i \leftarrow \boldsymbol{m}_i + 1$; $(\boldsymbol{x}_{i,\boldsymbol{m}_i}, \boldsymbol{y}_{i,\boldsymbol{m}_i}) \leftarrow (x, y)$ ▷ Increment counts and store samples for each group i associated with (x, y)
12: **end while**
13: **end if**
14: $\hat{\boldsymbol{\varepsilon}}_{1:g} \leftarrow (1 + \mathbb{1}_{\textsc{Reg}})\,\mathrm{AEV}(\boldsymbol{m}, \boldsymbol{\delta}_t, \boldsymbol{x}, \boldsymbol{y})$ ▷ Bound additive error of per-group supremum deviations (w.h.p.)
15: $\forall i \in \mathcal{Z}\colon\ \hat{\mathcal{S}}_i \leftarrow \big(\inf_{h \in \mathcal{H}} \hat{\mathbb{E}}_{\boldsymbol{x}_{i,:}, \boldsymbol{y}_{i,:}}[\ell \circ h]\big)$ **if** Reg **else** 0 ▷ Set regret baseline of per-group minimal empirical risks (or 0 if $\neg$Reg)
16: $\hat{h} \leftarrow \operatorname*{argmin}_{h \in \mathcal{H}} \mathrm{M}\big(i \mapsto c \wedge \hat{\mathbb{E}}_{\boldsymbol{x}_{i,:}, \boldsymbol{y}_{i,:}}[\ell \circ h] - \hat{\mathcal{S}}_i + \hat{\boldsymbol{\varepsilon}}_i; \boldsymbol{w}\big)$ ▷ Compute UCB-optimal $\hat{h}$
17: $\mathrm{M}^{\star\downarrow} \leftarrow \inf_{h \in \mathcal{H}} \mathrm{M}\big(i \mapsto 0 \vee \hat{\mathbb{E}}_{\boldsymbol{x}_{i,:}, \boldsymbol{y}_{i,:}}[\ell \circ h] - \hat{\mathcal{S}}_i - \hat{\boldsymbol{\varepsilon}}_i; \boldsymbol{w}\big)$ ▷ Lower-bound optimal $\mathrm{M}^\star$
18: $(\hat{\mathrm{M}}^\downarrow, \hat{\mathrm{M}}^\uparrow) \leftarrow \Big(\mathrm{M}\big(i \mapsto 0 \vee \hat{\mathbb{E}}_{\boldsymbol{x}_{i,:}, \boldsymbol{y}_{i,:}}[\ell \circ \hat{h}] - \hat{\mathcal{S}}_i - \hat{\boldsymbol{\varepsilon}}_i; \boldsymbol{w}\big), \mathrm{M}\big(i \mapsto c \wedge \hat{\mathbb{E}}_{\boldsymbol{x}_{i,:}, \boldsymbol{y}_{i,:}}[\ell \circ \hat{h}] - \hat{\mathcal{S}}_i + \hat{\boldsymbol{\varepsilon}}_i; \boldsymbol{w}\big)\Big)$ ▷ LCB and UCB on $\hat{h}$ (regret) malfare
19: **if** $\hat{\mathrm{M}}^\uparrow \le \mathrm{M}^{\star\downarrow} + 2\varepsilon$ **then** ▷ Check if desired error guarantee is met (termination condition)
20: $(\hat{\mu}, \hat{\varepsilon}) \leftarrow \Big(\frac{1}{2}(\hat{\mathrm{M}}^\downarrow + \hat{\mathrm{M}}^\uparrow), \frac{1}{2}(\hat{\mathrm{M}}^\uparrow - \hat{\mathrm{M}}^\downarrow)\Big)$ ▷ Symmetric estimate $\hat{\mu}$ and confidence radius $\hat{\varepsilon}$ of (regret) malfare of $\hat{h}$
21: **return** $\big(\hat{h}, \hat{\mu}, \hat{\varepsilon}, \mathrm{M}^{\star\downarrow}\big)$ ▷ Return UCB-optimal $\hat{h}$, $\hat{\varepsilon}$-estimate of $\mathrm{M}(\cdot; \boldsymbol{w})$, and lower-bound on optimal malfare $\mathrm{M}^{\star\downarrow}$
22: **end if**
23: **end for**
24: **end procedure**
25: **procedure** LinearPSUtility$(\mathcal{H}, \mathrm{u}(\cdot,\cdot), \mathcal{D}, \mathrm{AEV}(\dots), \boldsymbol{s}, \boldsymbol{\delta}, \varepsilon, \mathrm{M}(\cdot; \boldsymbol{w}), \textsc{Reg}) \to (\hat{h}, \hat{\mu}, \hat{\varepsilon}, \mathrm{M}^{\star\uparrow})$
26: **input**: Utility function $\mathrm{u}(\cdot,\cdot) : \mathcal{Y}' \times \mathcal{Y} \to [0, c]$, weighted aggregator function $\mathrm{M}(\cdot; \boldsymbol{w})$ (malfare if Reg, otherwise welfare), see line 2
27: **output**: Empirically LCB-optimal $\hat{h}$, empirical welfare $\hat{\mu}$, confidence radius $\hat{\varepsilon}$, and UB on maximal welfare $\mathrm{M}^{\star\uparrow}$ (*or similar for regret*)
28: $(\hat{h}, \hat{\mu}, \hat{\varepsilon}, \mathrm{M}^{\star\uparrow}) \leftarrow$ LinearPSLoss$(\mathcal{H}, c - \mathrm{u}(\cdot,\cdot), \mathcal{D}, \mathrm{AEV}(\dots), \boldsymbol{s}, \boldsymbol{\delta}, \varepsilon, (2\mathbb{1}_{\textsc{Reg}} - 1)\mathrm{M}(\mathcal{S}_i \mapsto c - \mathcal{S}_i; \boldsymbol{w}), \textsc{Reg})$ ▷ Negate to flip inf and sup
29: **return** $\big(\hat{h}, r - \hat{\mu}, \hat{\varepsilon}, (2\mathbb{1}_{\textsc{Reg}} - 1)\mathrm{M}^{\star\uparrow}\big)$
30: **end procedure**
31: **procedure** LinearPSEstimate$(h, \mathrm{s}(\cdot,\cdot), \mathcal{D}, \mathrm{AEV}(\dots), \boldsymbol{s}, \boldsymbol{\delta}, \varepsilon, \mathrm{M}(\cdot; \boldsymbol{w})) \to (\hat{\mu}, \hat{\varepsilon})$
32: **input**: Hypothesis $h(\cdot) : \mathcal{X} \to \mathcal{Y}'$, sentiment function $\mathrm{s} : (\mathcal{Y}' \times \mathcal{Y}) \to [0, c]$, weighted aggregator function $\mathrm{M}(\cdot; \boldsymbol{w})$, see line 2
33: **output**: Empirical aggregator function estimate $\hat{\mu}$ and confidence radius $\hat{\varepsilon}$
34: $(\_, \hat{\mu}, \hat{\varepsilon}, \_) \leftarrow$ LinearPSLoss$(\{h\}, \mathrm{s}(\cdot,\cdot), \mathcal{D}, \mathrm{AEV}(\dots), \boldsymbol{s}, \boldsymbol{\delta}, \varepsilon, \mathrm{M}(\cdot; \boldsymbol{w}), \textsc{False})$ ▷ Estimate the malfare or welfare of $\mathrm{s} \circ h$
35: **return** $(\hat{\mu}, \hat{\varepsilon})$
36: **end procedure**

4.2 The Linear Progressive Sampling Algorithm

The core of *linear progressive sampling* (algorithm 1) is quite simple. At timestep $t = 1$, we *guess* that a sample of size $\boldsymbol{s}_1$ for all groups will be sufficient to ε-δ optimize the objective, we draw at least such a sample (line 7 for joint sampling, or lines 9–12 for mixture sampling), compute tail bounds (line 14), then determine the UCB-optimal $\hat{h}$ (line 16). If our bounds indicate that $\hat{h}$ is provably near-optimal, algorithm 1 terminates, otherwise, our guess was incorrect, so we increment t, draw at least $\boldsymbol{s}_t$ samples (per-group), and repeat. The basic principle is quite flexible, so algorithm 1 can maximize welfare or minimize malfare of risk or regret via the LinearPSLoss(...) and LinearPSUtility(...) routines. Furthermore, in addition to *learning* and *optimization* tasks over some class $\mathcal{H}$, algorithm 1 can be applied to *estimation tasks*: given a single function h, it can estimate the malfare or welfare of $\mathrm{s} \circ h$ via the LinearPSEstimate(...) routine.

Theorem 4.5 shows that algorithm 1 learns an optimal $h \in \mathcal{H}$ to within user-specified ε-δ additive error. We require only monotonicity (axiom 1) and continuity (axiom 3) of $\mathrm{M}(\cdot; \boldsymbol{w})$, though the power-mean malfare family is convenient, as Lipschitz-continuity (theorem 2.2 item 3) permits efficient ε-uniformly-convergent schedules (definition 4.2). NB this result generalizes to welfare objectives, *mutatis mutandis* (flipping infima and suprema), via the negation reduction of LinearPSUtility(...), i.e., lines 25–30, and to function estimation via LinearPSEstimate(...).

Theorem 4.5 (Linear PS Guarantees)**.** Suppose $(\hat{h}, \hat{\mu}, \hat{\varepsilon}, \mathrm{M}^{\star\downarrow}) \leftarrow$ LinearPSLoss$(\mathcal{H}, \ell(\cdot,\cdot), \mathcal{D}, \mathrm{AEV}(\dots), \boldsymbol{s}, \boldsymbol{\delta}, \varepsilon, \mathrm{M}(\cdot; \boldsymbol{w}), \textsc{Reg})$,

Algorithm 2 Fair Learning with Braided Progressive Sampling under the Conditional Sampling Model

1: **procedure** $\textsc{BraidedPSLoss}(\mathcal{H}, \ell(\cdot,\cdot), \mathcal{D}_{1:g}, \boldsymbol{C}_{1:g}, \mathrm{AES}(\dots), \boldsymbol{s}, \boldsymbol{\delta}, \varepsilon, \mathrm{M}(\cdot; \boldsymbol{w}), \textsc{Reg}) \to (\hat{h}, \hat{\mu}, \hat{\varepsilon}, \mathrm{M}^{\star\downarrow})$
2: **input**: Hypothesis class $\mathcal{H}$, loss function $\ell(\cdot,\cdot)$, per-group distributions $\mathcal{D}_{1:g}$, cost model $\boldsymbol{C}_{1:g} \in \mathbb{R}_+^g$, additive error scalar bound $\mathrm{AES}(m, \delta, \boldsymbol{x}, \boldsymbol{y})$, schedule $\boldsymbol{s} \in \mathbb{Z}_+^\infty$ and $\boldsymbol{\delta} \in [0,1)^\infty$, confidence radius ε, weighted malfare $\mathrm{M}(\cdot; \boldsymbol{w})$, and Boolean $\textsc{Reg}$
3: **output**: Empirically UCB-optimal $\hat{h}$, empirical malfare estimate $\hat{\mu}$, confidence radius $\hat{\varepsilon}$, and lower bound on minimal malfare $\mathrm{M}^{\star\downarrow}$
4: $\boldsymbol{t}_{1:g} \leftarrow \boldsymbol{1}$ ▷ Initialize per-group timestep indices
5: $\forall i \in \mathcal{Z}\colon\ (\boldsymbol{x}_{i,1:\boldsymbol{s}_1}, \boldsymbol{y}_{i,1:\boldsymbol{s}_1}) \sim \mathcal{D}_i^{\boldsymbol{s}_1};\ \hat{\boldsymbol{\varepsilon}}_i \leftarrow (1 + \mathbb{1}_{\textsc{Reg}})\,\mathrm{AES}(\boldsymbol{s}_1, \frac{\boldsymbol{\delta}_1}{g}, \boldsymbol{x}_{i,:}, \boldsymbol{y}_{i,:})$ ▷ Draw initial sample for all groups & bound error
6: **loop** ▷ Loop over braided algorithm iterations
7: $\forall i, j \in \mathcal{Z}, t \in \mathbb{Z}_+ :\ \tilde{\boldsymbol{\varepsilon}}_{j,t}^{(i)} \leftarrow \left(\hat{\boldsymbol{\varepsilon}}_j \ \textbf{ if } \ i \neq j \ \textbf{ else } \ \hat{\boldsymbol{\varepsilon}}_j \sqrt{\frac{\boldsymbol{s}_{\boldsymbol{t}_j} \ln \frac{g}{\boldsymbol{\delta}_{\boldsymbol{t}_j}}}{\boldsymbol{s}_{t+\boldsymbol{t}_j} \ln \frac{g}{\boldsymbol{\delta}_{t+\boldsymbol{t}_j}}}} \right)$ ▷ Estimate of $\hat{\boldsymbol{\varepsilon}}_j$ after sampling group i for t more iterations
8: $\forall i \in \mathcal{Z}\colon\ \hat{\mathcal{S}}_i \leftarrow \left(\inf_{h \in \mathcal{H}} \hat{\mathbb{E}}_{\boldsymbol{x}_{i,:}, \boldsymbol{y}_{i,:}} [\ell \circ h] \right) \ \textbf{if} \ \textsc{Reg} \ \textbf{else} \ 0$ ▷ Set regret baseline of per-group minimal empirical risks (or 0 if $\neg\textsc{Reg}$)
9: $\hat{h} \leftarrow \operatorname{argmin}_{h \in \mathcal{H}} \mathrm{M}\big(i \mapsto c \wedge \hat{\mathbb{E}}_{\boldsymbol{x}_{i,:}, \boldsymbol{y}_{i,:}} [\ell \circ h] - \hat{\mathcal{S}}_i + \hat{\boldsymbol{\varepsilon}}_i; \boldsymbol{w}\big)$ ▷ Compute UCB-optimal $\hat{h}$
10: $\mathrm{M}^{\star\downarrow} \leftarrow \inf_{h \in \mathcal{H}} \mathrm{M}\big(i \mapsto 0 \vee \hat{\mathbb{E}}_{\boldsymbol{x}_{i,:}, \boldsymbol{y}_{i,:}} [\ell \circ h] - \hat{\mathcal{S}}_i - \hat{\boldsymbol{\varepsilon}}_i; \boldsymbol{w}\big)$ ▷ Lower-bound optimal $\mathrm{M}^\star$
11: $(\hat{\mathrm{M}}^\downarrow, \hat{\mathrm{M}}^\uparrow) \leftarrow \Big(\mathrm{M}\big(i \mapsto 0 \vee \hat{\mathbb{E}}_{\boldsymbol{x}_{i,:}, \boldsymbol{y}_{i,:}} [\ell \circ \hat{h}] - \hat{\mathcal{S}}_i - \hat{\boldsymbol{\varepsilon}}_i; \boldsymbol{w}\big), \mathrm{M}\big(i \mapsto c \wedge \hat{\mathbb{E}}_{\boldsymbol{x}_{i,:}, \boldsymbol{y}_{i,:}} [\ell \circ \hat{h}] - \hat{\mathcal{S}}_i + \hat{\boldsymbol{\varepsilon}}_i; \boldsymbol{w}\big) \Big)$ ▷ LCB and UCB on $\hat{h}$ (regret) malfare
12: **if** $\hat{\mathrm{M}}^\uparrow \leq \mathrm{M}^{\star\downarrow} + 2\varepsilon$ **then** ▷ Check if desired error guarantee is met (termination condition)
13: $(\hat{\mu}, \hat{\varepsilon}) \leftarrow \Big(\frac{1}{2}(\hat{\mathrm{M}}^\downarrow + \hat{\mathrm{M}}^\uparrow), \frac{1}{2}(\hat{\mathrm{M}}^\uparrow - \hat{\mathrm{M}}^\downarrow) \Big)$ ▷ Symmetric estimate of $\hat{\mu}$ of malfare or regret of $\hat{h}$
14: **return** $(\hat{h}, \hat{\mu}, \hat{\varepsilon}, \mathrm{M}^{\star\downarrow})$
15: **end if**
16: $i \leftarrow \operatorname{argmax}_{i \in \mathcal{Z}} \sup_{t \in \mathbb{Z}_+} \underbrace{\frac{1}{\boldsymbol{C}_i(\boldsymbol{s}_{t+\boldsymbol{t}_i} - \boldsymbol{s}_{\boldsymbol{t}_i})}}_{\textsc{Reciprocal Cost}} \underbrace{\Big(\hat{\mathrm{M}}^\uparrow - \mathrm{M}\Big(j \mapsto c \wedge \hat{\mathbb{E}}_{\boldsymbol{x}_{i,:}, \boldsymbol{y}_{i,:}} [\ell \circ h] - \hat{\mathcal{S}}_i + \tilde{\boldsymbol{\varepsilon}}_{j,t}^{(i)}; \boldsymbol{w} \Big) \Big)}_{\textsc{Estimated (Regret) Malfare Improvement}}$ ▷ Maximize improvement:cost ratio
17: $(\boldsymbol{x}_{i,1+\boldsymbol{s}_{\boldsymbol{t}_i}:\boldsymbol{s}_{1+\boldsymbol{t}_i}}, \boldsymbol{y}_{i,1+\boldsymbol{s}_{\boldsymbol{t}_i}:\boldsymbol{s}_{1+\boldsymbol{t}_i}}) \sim \mathcal{D}_i^{\boldsymbol{s}_{1+\boldsymbol{t}_i} - \boldsymbol{s}_{\boldsymbol{t}_i}};\ \boldsymbol{t}_i \leftarrow \boldsymbol{t}_i + 1;\ \hat{\boldsymbol{\varepsilon}}_i \leftarrow (1 + \mathbb{1}_{\textsc{Reg}})\,\mathrm{AES}(\boldsymbol{s}_{\boldsymbol{t}_i}, \frac{\boldsymbol{\delta}_{\boldsymbol{t}_i}}{g}, \boldsymbol{x}_{i,:}, \boldsymbol{y}_{i,:})$ ▷ Sample group i & bound error
18: **end loop**
19: **end procedure**

$\mathrm{M}(\mathcal{S}; \boldsymbol{w})$ is continuous and monotonic in $\mathcal{S}$ with (possibly infinite) Lipschitz constant λ_{M} w.r.t. $\|\cdot\|_{\mathrm{M}}$, and the schedules $(\boldsymbol{s}, \boldsymbol{\delta})$ are $\frac{\varepsilon}{\lambda_{\mathrm{M}}(1+\mathbb{1}_{\textsc{Reg}})}$-uniformly-convergent w.r.t. $\mathrm{AEV}(\dots)$ and $\|\cdot\|_{\mathrm{M}}$. Now take μ to be the true objective value of $\hat{h}$ and $\mu^\star$ to be the true objective value of the optimal $h^\star$, i.e., if $\textsc{Reg} = \textsc{False}$, take $\mu \doteq \mathrm{M}(i \mapsto \mathbb{E}_{\mathcal{D}_i}[\ell \circ \hat{h}]; \boldsymbol{w})$ and $\mu^\star \doteq \inf_{h \in \mathcal{H}} \mathrm{M}(i \mapsto \mathbb{E}_{\mathcal{D}_i}[\ell \circ h]; \boldsymbol{w})$, or if $\textsc{Reg} = \textsc{True}$, take (see equation 3) $\mu \doteq \mathrm{M}(i \mapsto \mathrm{Reg}_i(\hat{h}); \boldsymbol{w})$ and $\mu^\star \doteq \inf_{h \in \mathcal{H}} \mathrm{M}\left(i \mapsto \mathrm{Reg}_i(h); \boldsymbol{w}\right)$. Then, with probability at least $1 - \delta$, the output $(\hat{h}, \hat{\mu}, \hat{\varepsilon}, \mathrm{M}^{\star\downarrow})$ obeys
1) $|\hat{\mu} - \mu| \leq \hat{\varepsilon} \leq \varepsilon$; & 2) $\mathrm{M}^{\star\downarrow} \leq \mu^\star \leq \mu \leq \hat{\mu} + \hat{\varepsilon} \leq \mathrm{M}^{\star\downarrow} + 2\varepsilon$.

4.3 The Braided Progressive Sampling Algorithm

Under the joint and mixture sampling models (algorithm 1), progress is *linear* (i.e., *sequential*, as no decisions are made except when to terminate); we begin with (at least) $\boldsymbol{s}_1$ samples per group, and advance until we reach a sufficient sample size to terminate with the desired guarantee. For the conditional sampling model, we present *braided progressive sampling* (algorithm 2), which is *actively making decisions*, thus linear analysis is not applicable. At each *iteration* (line 6) of algorithm 2, a group index $i \in \mathcal{Z}$ is chosen (line 16) to optimize an estimate of knowledge-gain via logic similar to that of section 3.3 (due to space limitations, the details are deferred to appendix B.2), and group i is sampled for one additional *timestep* (line 17), i.e., the sample associated with group i is extended from size $\boldsymbol{s}_{\boldsymbol{t}_i}$ to $\boldsymbol{s}_{1+\boldsymbol{t}_i}$, where $\boldsymbol{t}_i$ denotes the current timestep for group i. The remainder of algorithm 2 is essentially the same as algorithm 1; after sampling, we optimize (line 9) a UCB-optimal $\hat{h}$, bound the objective (lines 10–11), and terminate if the user supplied guarantee is met, otherwise we continue.

There is thus a *lattice* of possible sample size vectors $\boldsymbol{m}$, i.e., the possibilities are the Cartesian product $\boldsymbol{s} \times \cdots \times \boldsymbol{s}$. To avoid a union bound over this (exponentially large) lattice, we analyze the method as a *braid*, in that g progressive sampling sequences are *concurrently active*, and at each iteration we select some group i, and advance the schedule by one timestep for only group i (thus we have g independent strands, advancing and intertwining in some random order). Consequently, we must use (line 17) the additive error *scalar bound* $\hat{\boldsymbol{\varepsilon}}_i \leftarrow \mathrm{AES}\big(\boldsymbol{m}_i, \frac{\boldsymbol{\delta}_{\boldsymbol{t}_i}}{g}, \boldsymbol{x}_{i,:}, \boldsymbol{y}_{i,:}\big)$, i.e., we operate on *one group at a time*, rather than over all groups as in the linear algorithm (algorithm 1 line 14). Similar analysis is employed for *multi-armed bandits*, where a union bound is taken over all timesteps and each arm being sampled. With algorithm 2 explained, we now show a correctness result, analogous to theorem 4.6 for the linear algorithm. Note that, as with algorithm 1, we can generalize algorithm 2 and its guarantees to utility and welfare functions, using the reduction of $\textsc{LinearPSUtility}(\dots)$, and similarly we can provide guarantees for function estimation via the logic of $\textsc{LinearPSEstimate}(\dots)$.

Theorem 4.6 (Braided PS Guarantees)**.** Suppose $(\hat{h}, \hat{\mu}, \hat{\varepsilon}, \mathrm{M}^{\star\downarrow}) \leftarrow \textsc{BraidedPSLoss}(\mathcal{H}, \ell(\cdot,\cdot), \mathcal{D}, \mathrm{AES}(\dots), \boldsymbol{s}, \boldsymbol{\delta}, \varepsilon, \mathrm{M}(\cdot; \boldsymbol{w}), \textsc{Reg})$, $\mathrm{M}(\mathcal{S}; \boldsymbol{w})$ is continuous and strictly monotonic in $\mathcal{S}$ with (possibly infinite) Lipschitz constant λ_{M} w.r.t. $\|\cdot\|_{\mathrm{M}}$, and the schedules $(\boldsymbol{s}, \boldsymbol{\delta})$ are $\frac{\varepsilon}{\lambda_{\mathrm{M}}(1+\mathbb{1}_{\textsc{Reg}})}$-uniformly-convergent w.r.t. $\|\cdot\|_{\mathrm{M}}$ and the *additive error vector bound* $\mathrm{AEV}(\boldsymbol{m}, \delta, \boldsymbol{x}, \boldsymbol{y}) \leftarrow \left\langle \mathrm{AES}(\boldsymbol{m}_1, \frac{\delta}{g}, \boldsymbol{x}_1, \boldsymbol{y}_1), \dots, \mathrm{AES}(\boldsymbol{m}_g, \frac{\delta}{g}, \boldsymbol{x}_g, \boldsymbol{y}_g) \right\rangle$. Now take μ to be the true objective value of $\hat{h}$ and $\mu^\star$ to be the true objective value of the optimal $h^\star$ (see theorem 4.5). Then, with probability at least $1-\delta$, we have
1) $|\hat{\mu} - \mu| \leq \hat{\varepsilon} \leq \varepsilon$; & 2) $\mathrm{M}^{\star\downarrow} \leq \mu^\star \leq \mu \leq \hat{\mu} + \hat{\varepsilon} \leq \mathrm{M}^{\star\downarrow} + 2\varepsilon$.

5 Conclusion

This work generalizes existing theories of fair machine learning, with welfare, malfare, and regret objectives, thus subsuming the *minimax fair learning* [Martinez et al., 2020, Abernethy et al., 2020, Diana et al., 2021, Lahoti et al., 2020, Shekhar et al., 2021], *multi-group agnostic PAC learning* [Blum and Lykouris, 2020, Rothblum and Yona, 2021], and *fair-PAC learning* [Cousins, 2021] settings, while enjoying rigorous statistical learning guarantees and the axiomatization of cardinal welfare theory. In particular, we bound the *generalization error* and *sample complexity* of UCB-optimal models, either given a fixed sample, or to meet a user-supplied ε-δ optimality guarantee via progressive sampling. Our bounds leverage the specific character of the objective at hand, and our progressive sampling methods are tailored to three realistic models of data generation. We stress that while training UCB-optimal models is analytically convenient, there is also an important fairness impact to this decision, as fair malfare functions (e.g., egalitarian) place strong emphasis on the most disadvantaged groups, which are often understudied minority groups. Cousins [2021] notes that optimizing *empirical malfare* $\hat{\mathrm{M}}$ overfits to small numbers of sampled minorities, however we argue that training UCB-optimal models (i.e., optimizing $\hat{\mathrm{M}}^{\uparrow}$) factors *uncertainty* into training, so that the needs of understudied groups (i.e., those with large $\hat{\boldsymbol{\varepsilon}}_i$ values) are better addressed.

Our active learning setting under the conditional sampling model is philosophically intriguing, as we find that optimally investing sampling effort under uncertainty is challenging, depends on the objective at hand, and has important fairness impact. In section 3.3, we see that a host of factors involving the objective, function class $\mathcal{H}$, and per-group distributions $\mathcal{D}_{1:g}$ all interact to determine the sharpness of welfare, malfare, and regret bounds, and property 3.8 quantifies the incremental UCB improvement of sampling each group. This analysis answers questions raised by Chen et al. [2018] as to how sampling-error impacts fairness, and generalizes the analysis of Shekhar et al. [2021] from the egalitarian special-case to arbitrary power-mean malfare functions. Algorithm 2 then incorporates these ideas into an *active sampling algorithm*, which dynamically select groups to sample based on projected UCB improvement. Notably, algorithm 1 does use uniform sample sizes under the joint sampling model, and uses whatever is available under the mixture sampling model, as these are natural choices under these sampling models. In contrast, under the conditional sampling model, algorithm 2 is able to make more intelligent decisions as to where to allocate sampling effort.

We thus conclude that (welfare-centric) fairness, statistical uncertainty, and sample complexity analysis are tightly intertwined, and must all be considered to best allocate resources in service of the social planner. We are hopeful that this analysis and algorithmic study will lead to a greater emphasis on sample complexity and finite sample analysis for the social planner's problem, which is traditionally analyzed in terms of the asymptotic Bayesian methods of classical economics. In particular, we are hopeful that this analysis emphasizes and mathematically supports the call for greater visibility of minority groups and the importance of incorporating diverse data into (fair) machine learning systems.

Acknowledgments

This work was supported in part by NSF award RI-1813444 and DARPA/AFRL grant FA8750.

References

Jacob Abernethy, Pranjal Awasthi, Matthäus Kleindessner, Jamie Morgenstern, Chris Russell, and Jie Zhang. Active sampling for min-max fairness. *arXiv preprint arXiv:2006.06879*, 2020.

Martin Anthony and Peter L Bartlett. *Neural network learning: Theoretical foundations*. Cambridge University Press, 2009.

Ahmed Ashraf, Shehroz Khan, Nikhil Bhagwat, Mallar Chakravarty, and Babak Taati. Learning to unlearn: Building immunity to dataset bias in medical imaging studies. *Machine Learning for Health Workshop at Advances at Neural Information Processing Systems*, 2018.

Kristine Bærøe, Torbjørn Gundersen, Edmund Henden, and Kjetil Rommetveit. Can medical algorithms be fair? Three ethical quandaries and one dilemma. *BMJ Health & Care Informatics*, 29(1), 2022.

Peter L Bartlett and Shahar Mendelson. Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002.

George Bennett. Probability inequalities for the sum of independent random variables. *Journal of the American Statistical Association*, 57(297):33–45, 1962.

Avrim Blum and Thodoris Lykouris. Advancing subgroup fairness via sleeping experts. In *Innovations in Theoretical Computer Science Conference*, volume 11, 2020.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

Olivier Bousquet. A Bennett concentration inequality and its application to suprema of empirical processes. *Comptes Rendus Mathematique*, 334(6):495–500, 2002.

Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91. PMLR, 2018.

Jacqueline G Cavazos, P Jonathon Phillips, Carlos D Castillo, and Alice J O'Toole. Accuracy comparison across face recognition algorithms: Where are we on measuring race bias? *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 2020.

Irene Chen, Fredrik D Johansson, and David Sontag. Why is my classifier discriminatory? *Advances in Neural Information Processing Systems*, 31, 2018.

Cynthia M Cook, John J Howard, Yevgeniy B Sirotin, Jerry L Tipton, and Arun R Vemury. Demographic effects in facial recognition and their dependence on image acquisition: An evaluation of eleven commercial systems. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 1(1):32–41, 2019.

Cyrus Cousins. An axiomatic theory of provably-fair welfare-centric machine learning. In *Advances in Neural Information Processing Systems*, 2021.

Cyrus Cousins and Matteo Riondato. CaDET: Interpretable parametric conditional density estimation with decision trees and forests. *Machine Learning*, 108(8):1613–1634, 2019.

Cyrus Cousins and Matteo Riondato. Sharp uniform convergence bounds through empirical centralization. *Advances in Neural Information Processing Systems*, 33, 2020.

Cyrus Cousins, Kavosh Asadi, and Michael L. Littman. Fair e3: Efficient welfare-centric fair reinforcement learning. In *5th Multidisciplinary Conference on Reinforcement Learning and Decision Making*. RLDM, 2022.

Hugh Dalton. The measurement of the inequality of incomes. *The Economic Journal*, 30(119):348–361, 1920.

Gerard Debreu. Topological methods in cardinal utility theory. 1959.

Matthew DeCamp and Charlotta Lindvall. Latent bias and the implementation of artificial intelligence in medicine. *Journal of the American Medical Informatics Association*, 27(12):2020–2023, 2020.

M DeGroot. *Optimal statistical decisions*. McGraw-Hill, New York, 1970.

Luc Devroye, Matthieu Lerasle, Gabor Lugosi, Roberto I Oliveira, et al. Sub-Gaussian mean estimators. *The Annals of Statistics*, 44(6):2695–2725, 2016.

Emily Diana, Wesley Gill, Michael Kearns, Krishnaram Kenthapadi, and Aaron Roth. Minimax group fairness: Algorithms and experiments. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 66–76, 2021.

Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. Retiring adult: New datasets for fair machine learning. *Advances in Neural Information Processing Systems*, 34, 2021.

Virginie Do and Nicolas Usunier. Optimizing generalized Gini indices for fairness in rankings. In *45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2022.

Cynthia Dwork, Nicole Immorlica, Adam Tauman Kalai, and Mark D. M. Leiserson. Decoupled classifiers for group-fair and efficient machine learning. In *Conference on Fairness, Accountability and Transparency, FAT 2018, 23–24 February 2018, New York, NY, USA*, volume 81 of *Proceedings of Machine Learning Research*, pages 119–133. PMLR, 2018.

Clare Garvie, Alvaro Bedoya, and Jonathan Frankle. The perpetual line-up. Unregulated police face recognition in America, 2016.

URL https://www.perpetuallineup.org.

Andrew Gelman, John B. Carlin, Hal S. Stern, and Donald B. Rubin. *Bayesian Data Analysis*. Chapman and Hall, second edition, 2004.

William M Gorman. The structure of utility functions. *The Review of Economic Studies*, 35(4):367–390, 1968.

Hoda Heidari, Claudio Ferrari, Krishna Gummadi, and Andreas Krause. Fairness behind a veil of ignorance: A welfare analysis for automated decision making. In *Advances in Neural Information Processing Systems*, pages 1265–1276, 2018.

Hoda Heidari, Solon Barocas, Jon M Kleinberg, and Karen Levy. On modeling human perceptions of allocation policies with uncertain outcomes. In *EC '21: The 22nd ACM Conference on Economics and Computation, Budapest, Hungary, July 18–23, 2021*, pages 589–609. ACM, 2021.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

Lily Hu and Yiling Chen. Fair classification and social welfare. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 535–545, 2020.

Weihua Hu, Gang Niu, Issei Sato, and Masashi Sugiyama. Does distributionally robust supervised learning give robust classifiers? In *International Conference on Machine Learning*, pages 2029–2037. PMLR, 2018.

George H John and Pat Langley. Static versus dynamic sampling for data mining. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, pages 367–370, 1996.

Rucha Kulkarni, Ruta Mehta, and Setareh Taki. Indivisible mixed manna: On the computability of MMS + PO allocations. In *Proceedings of the 22nd ACM Conference on Economics and Computation*, pages 683–684, 2021.

Preethi Lahoti, Alex Beutel, Jilin Chen, Kang Lee, Flavien Prost, Nithum Thain, Xuezhi Wang, and Ed Chi. Fairness without demographics through adversarially reweighted learning. *Advances in neural information processing systems*, 33:728–740, 2020.

Jean-Samuel Leboeuf, Frédéric LeBlanc, and Mario Marchand. Decision trees as partitioning machines to characterize their generalization properties. *Advances in Neural Information Processing Systems*, 33, 2020.

Gábor Lugosi and Shahar Mendelson. Mean estimation and regression under heavy-tailed distributions: A survey. *Foundations of Computational Mathematics*, 19(5):1145–1190, 2019.

Natalia Martinez, Martin Bertran, and Guillermo Sapiro. Minimax Pareto fairness: A multi objective perspective. In *International Conference on Machine Learning*, pages 6755–6764. PMLR, 2020.

Andreas Maurer and Massimiliano Pontil. Empirical Bernstein bounds and sample-variance penalization. In *International Conference on Computational Learning Theory*. Springer, 2009.

John Ashworth Nelder and Robert WM Wedderburn. Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, 135(3):370–384, 1972.

Yonatan Oren, Shiori Sagawa, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust language modeling. *arXiv preprint arXiv:1909.02060*, 2019.

Emanuel Parzen. On estimation of a probability density function and mode. *The annals of mathematical statistics*, 33(3):1065–1076, 1962.

Arthur Cecil Pigou. *Wealth and welfare*. Macmillan and Company, limited, 1912.

Foster Provost, David Jensen, and Tim Oates. Efficient progressive sampling. In *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 23–32, 1999.

Kevin WS Roberts. Interpersonal comparability and social choice theory. *The Review of Economic Studies*, pages 421–439, 1980.

Esther Rolf, Max Simchowitz, Sarah Dean, Lydia T Liu, Daniel Björkegren, Moritz Hardt, and Joshua Blumenstock. Balancing competing objectives with noisy data: Score-based classifiers for welfare-aware machine learning. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13–18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 8158–8168. PMLR, 2020.

Murray Rosenblatt. Remarks on some nonparametric estimates of a density function. *The Annals of Mathematical Statistics*, pages 832–837, 1956.

Guy N Rothblum and Gal Yona. Multi-group agnostic PAC learnability. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18–24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 9107–9115. PMLR, 2021.

Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks. In *International Conference on Learning Representations*, 2019.

Amartya Sen. On weights and measures: Informational constraints in social welfare analysis. *Econometrica: Journal of the Econometric Society*, pages 1539–1572, 1977.

Mohammad Shahrokh Esfahani and Edward R Dougherty. Effect of separate sampling on classification accuracy. *Bioinformatics*, 30 (2):242–250, 2014.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge University Press, 2014.

Shubhanshu Shekhar, Greg Fields, Mohammad Ghavamzadeh, and Tara Javidi. Adaptive sampling for minimax fair classification. *Advances in Neural Information Processing Systems*, 34, 2021.

Umer Siddique, Paul Weng, and Matthieu Zimmer. Learning fair policies in multi-objective (deep) reinforcement learning with average and discounted rewards. In *International Conference on Machine Learning*, pages 8905–8915. PMLR, 2020.

Till Speicher, Hoda Heidari, Nina Grgić-Hlača, Krishna P Gummadi, Adish Singla, Adrian Weller, and Muhammad Bilal Zafar. A unified approach to quantifying algorithmic unfairness: Measuring individual & group unfairness via inequality indices. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2239–2248, 2018.

Vladimir Vapnik and Aleksei Chervonenkis. The uniform convergence of frequencies of the appearance of events to their probabilities. In *Doklady Akademii Nauk*, volume 181, pages 781–783. Russian Academy of Sciences, 1968.

Vladimir Vapnik and Aleksei Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16(2):264–281, 1971.

Enrique Areyan Viqueira, Cyrus Cousins, and Amy Greenwald. Improved algorithms for learning equilibria in simulation-based games. In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems*, pages 79–87, 2020.

A Proof Appendix

We now present proofs of all mathematical claims in the paper body. We first show in appendix A.1 the estimation guarantees, sample complexity, and incremental-sampling knowledge-gain bounds of section 3. We then derive in appendix A.2 the results of section 3, i.e., those related to sampling schedules and progressive sampling.

A.1 Statistical Estimation Guarantees for Malfare, Welfare, and Regret

In this section, we show all bounds related to uniform convergence of sentiment values, and estimating welfare, malfare, and regret. We begin with theorems 3.1 and 3.2 and property 3.3, which bound single-group uniform convergence rates in the form of additive error scalar bounds, i.e., AES(...), under various conditions.

Theorem 3.1 (Uniform Convergence for Bounded Finite Hypothesis Classes)**.** We may bound the distribution-free $\mathrm{AES}(m, \delta)$, the distribution-dependent $\mathrm{AES}_{\mathcal{D}}(m, \delta)$, and the data-dependent $\mathrm{AES}(m, \delta, \boldsymbol{x}, \boldsymbol{y})$ scalar additive error as

1) $\boldsymbol{\varepsilon} \leftarrow \sqrt{\frac{2\frac{1}{4}r^2 \ln \frac{2|\mathcal{H}|}{\delta}}{m}}$ [Hoeffding, 1963] ;

2) $\boldsymbol{\varepsilon} \leftarrow \frac{r \ln \frac{2|\mathcal{H}|}{\delta}}{3m} + \sup_{h \in \mathcal{H}} \sqrt{\frac{2\,\mathbb{V}_{\mathcal{D}}[\mathrm{s} \circ h] \ln \frac{2|\mathcal{H}|}{\delta}}{m}}$ [Bennett, 1962] ; &

3) $\hat{\boldsymbol{\varepsilon}} \leftarrow \frac{7r \ln \frac{2|\mathcal{H}|+1}{\delta}}{3(m-1)} + \sup_{h \in \mathcal{H}} \sqrt{\frac{2\,\hat{\mathbb{V}}_{\boldsymbol{x},\boldsymbol{y}}[\mathrm{s} \circ h] \ln \frac{2|\mathcal{H}|+1}{\delta}}{(m-1)}}$ [Cousins and Riondato, 2020] ; respectively.

Proof. First note that items 1 and 2 hold via union bound over the lower and upper tails of the sentiment of each $h \in \mathcal{H}$, hence the $\ln \frac{2|\mathcal{H}|}{\delta}$ term, and item 3 has a union bound over one additional tail, which is used to bound the supremum variance. In particular, item 1 holds via Hoeffding's [Hoeffding, 1963] inequality, and item 2 via Bennett's inequality [Bennett, 1962] (technically the sub-gamma form of Bennett's sub-Poisson bound, a.k.a. Bernstein's inequality), and item 3 holds via the supremum variance bound of Cousins and Riondato [2020, theorem 2], followed by Bennett's inequality. Note that here $\hat{\mathbb{V}}[\cdot]$ denotes the unbiased (Bessel-corrected) empirical variance, though the empirical supremum variance is still an upward-biased estimate of the supremum variance. □

Theorem 3.2 (Uniform Convergence with Rademacher Averages)**.** Suppose hypothesis class $\mathcal{H}$ and sentiment function $\mathrm{s}(\cdot,\cdot)$, take $(\boldsymbol{x}, \boldsymbol{y}) \sim \mathcal{D}^m$ and $\boldsymbol{\sigma} \sim \mathcal{U}^m(\pm 1)$, i.e., $\boldsymbol{\sigma}$ is uniformly distributed on $(\pm 1)^m$, and define the *Rademacher average* $\mathfrak{R}_m(\mathrm{s} \circ \mathcal{H}, \mathcal{D})$ and *Bousquet variance proxy* $\mathfrak{V}_m(\mathrm{s} \circ \mathcal{H}, \mathcal{D})$ [see Bousquet, 2002] as

$$\mathfrak{R}_m(\mathrm{s} \circ \mathcal{H}, \mathcal{D}) \doteq \mathop{\mathbb{E}}_{\boldsymbol{x},\boldsymbol{y},\boldsymbol{\sigma}} \left[\sup_{h \in \mathcal{H}} \left| \frac{1}{m} \sum_{i=1}^{m} \mathrm{s} \circ h(\boldsymbol{x}_i) \boldsymbol{\sigma}_i \right| \right] \quad , \quad \mathfrak{V}_m(\mathrm{s} \circ \mathcal{H}, \mathcal{D}) \doteq \sup_{h \in \mathcal{H}} \mathop{\mathbb{V}}_{\mathcal{D}}[\mathrm{s} \circ h] + 4r\mathfrak{R}_m(\mathrm{s} \circ \mathcal{H}, \mathcal{D}) \ . \tag{7}$$

We may then bound $\mathrm{AES}_{\mathcal{D}}(m, \delta)$ as $\boldsymbol{\varepsilon} \leftarrow 2\mathfrak{R}_m(\mathrm{s} \circ \mathcal{H}, \mathcal{D}) + \frac{r \ln \frac{1}{\delta}}{3m} + \sqrt{\frac{2\mathfrak{V}_m(\mathrm{s} \circ \mathcal{H}, \mathcal{D}) \ln \frac{1}{\delta}}{m}}$.

Proof. This result follows via the Rademacher symmetrization inequality [Boucheron et al., 2013, lemma 11.4], which upper-bounds the expected supremum deviation as twice the Rademacher average, followed by Bousquet's bound [Bousquet, 2002] on the supremum deviation, i.e., on the quantity $\mathrm{AES}_{\mathcal{D}}(\dots)$. □

Property 3.3 (Practical Bounds on Rademacher Averages)**.** 1) Suppose $\mathcal{H}$ has Vapnik-Chervonenkis (VC) dimension d [Vapnik and Chervonenkis, 1968, 1971], and $\ell(\hat{y}, y) \doteq 1 - \mathbb{1}_y(\hat{y})$ is the 0-1 loss. Then for some absolute constant c, $\mathfrak{R}_m(\ell \circ \mathcal{H}, \mathcal{D}) \leq \sqrt{\frac{cd}{m}}$, which implies bounds for linear classifiers, bounded-depth decision trees [Leboeuf et al., 2020], and many classes of neural network [Anthony and Bartlett, 2009].

2) Suppose $\mathcal{X} \doteq \{\vec{x} \in \mathbb{R}^\infty \mid \|\vec{x}\|_2 \leq R\}$ is the R-radius $\mathcal{L}_2$ ball in $\mathbb{R}^\infty$, $\mathcal{H} \doteq \{\vec{x} \mapsto \vec{w} \cdot \vec{x} \mid \|\vec{w}\|_2 \leq \gamma\}$ is a γ-regularized linear class, $\mathcal{Y} \doteq [-R\gamma, R\gamma]$, and $\ell(\cdot,\cdot)$ is a λ-Lipschitz loss function s.t. $\ell(y, y) = 0$. Then $r \leq 2\lambda R\gamma$ and $\mathfrak{R}_m(\ell \circ \mathcal{H}, \mathcal{D}) \leq \frac{2\lambda R\gamma}{\sqrt{m}}$. This implies bounds for (kernelized) SVM, generalized linear models [Nelder and Wedderburn, 1972], and bounded linear regression.

Proof. This result is a collection of standard results in statistical learning theory, most of which are fully cited in the property statement. We now carry out our bibliographic duties with regards to the remaining results. These bounds are often stated for the *empirical Rademacher average*, $\hat{\mathfrak{R}}_m(\ell \circ \mathcal{H}, (\boldsymbol{x}, \boldsymbol{y}))$ which conditions on the sample $(\boldsymbol{x}, \boldsymbol{y})$, but we state bounds that hold over *any possible distribution* by considering *worst-case* realizations of these samples.

We first show item 1. The bound $\mathfrak{R}_m(\mathcal{H}, \mathcal{D}) \leq \sqrt{\frac{cd}{m}}$ follows via Dudley's chaining or entropy integral arguments, along with Haussler's bound on the $\mathcal{L}_2$ covering number of Vapnik-Chervonenkis classes, see, e.g., Boucheron et al. [2013, theorem 13.7]. The specific classes mentioned in the theorem statement are of bounded VC dimension, hence their inclusion.

We now show item 2. First note that each of the cited model classes is Lipschitz-continuous, so their bounds follow from the general statement. Now, note that the range of $\mathcal{H}$ is $[-R\gamma, R\gamma]$, which follows directly from the Cauchy-Schwarz inequality. Composition with $\ell(\cdot, y)$, for $y \in [-R\gamma, R\gamma]$, then maps this range to a subset of $[0, 2\lambda R\gamma]$, by Lipschitz continuity and nonnegativity of $\ell(\cdot, \cdot)$, and the fact that $\ell(y, y) = 0$, hence we conclude $r \leq 2\lambda R\gamma$.

We now bound the Rademacher average. In particular, Shalev-Shwartz and Ben-David [2014, lemma 26.10] show that $\mathfrak{R}_m(\mathcal{H}, \mathcal{D}_{|\mathcal{X}}) \leq \frac{R\gamma}{\sqrt{m}}$, where $\mathcal{D}_{|\mathcal{X}}$ denotes the marginalization of the label space $\mathcal{Y}$ from the instance distribution $\mathcal{D}$, i.e., they bound the Rademacher average of the linear hypothesis class on the unlabeled distribution over $\mathcal{X}$. Technically, their definition of Rademacher average contains no absolute value inside the supremum, but this is immaterial, as Cousins and Riondato [2020, lemma 5] show that the same bound holds with or without the absolute value. We then apply the Ledoux-Talagrand contraction principle[8] [Boucheron et al., 2013, lemma 11.6] to compose the hypothesis class with a λ-Lipschitz loss function, which yields $\mathfrak{R}_m(\ell \circ \mathcal{H}, \mathcal{D}) \leq 2\lambda \mathfrak{R}_m(\mathcal{H}, \mathcal{D}_{|\mathcal{X}}) \leq \frac{2\lambda R\gamma}{\sqrt{m}}$. □

We now show theorems 3.4 and 3.5 and corollary 3.6, which bound tails and expectations of malfare and welfare values in terms of additive error vector bounds, i.e., in terms of $\mathrm{AEV}(\dots)$.

Theorem 3.4 (Welfare and Malfare Tail Bounds)**.** Suppose sentiment function $\mathrm{s}(\cdot, \cdot) : \mathcal{Y}' \times \mathcal{Y} \to \mathbb{R}_{0+}$, per-group probability distributions $\mathcal{D}_{1:g}$, sample size vector $\boldsymbol{m} \in \mathbb{Z}_+^g$, samples $(\boldsymbol{x}, \boldsymbol{y}) \sim \mathcal{D}_1^{\boldsymbol{m}_1} \times \cdots \times \mathcal{D}_g^{\boldsymbol{m}_g}$, failure probability $\delta \in (0, 1)$, and additive error bound $\mathrm{AEV}(\dots)$, and let $\hat{\boldsymbol{\varepsilon}} \leftarrow \mathrm{AEV}(\boldsymbol{m}, \delta, \boldsymbol{x}, \boldsymbol{y})$. Then for all $h \in \mathcal{H}$ and all monotonic aggregator functions $\mathrm{M}(\cdot; \boldsymbol{w})$, it holds with probability at least $1 - \delta$ over $\boldsymbol{x}$, $\boldsymbol{y}$, and $\hat{\boldsymbol{\varepsilon}}$ that

$$\underbrace{\mathrm{M}\left(i \mapsto 0 \vee \mathop{\mathbb{E}}_{\mathcal{D}_i}[\mathrm{s} \circ h] - \hat{\boldsymbol{\varepsilon}}_i; \boldsymbol{w}\right)}_{\text{True LB}} \leq \underbrace{\mathrm{M}\left(i \mapsto \mathop{\hat{\mathbb{E}}}_{\boldsymbol{x}_{i,:}, \boldsymbol{y}_{i,:}}[\mathrm{s} \circ h]; \boldsymbol{w}\right)}_{\text{Plug-In Estimate } \hat{\mathrm{M}}} \leq \underbrace{\mathrm{M}\left(i \mapsto c \wedge \mathop{\mathbb{E}}_{\mathcal{D}_i}[\mathrm{s} \circ h] + \hat{\boldsymbol{\varepsilon}}_i; \boldsymbol{w}\right)}_{\text{True UB}} \ , \quad \& \tag{9}$$

$$\underbrace{\mathrm{M}\left(i \mapsto 0 \vee \mathop{\hat{\mathbb{E}}}_{\boldsymbol{x}_{i,:}, \boldsymbol{y}_{i,:}}[\mathrm{s} \circ h] - \hat{\boldsymbol{\varepsilon}}_i; \boldsymbol{w}\right)}_{\text{LCB Estimate } \hat{\mathrm{M}}^{\downarrow}} \leq \underbrace{\mathrm{M}\left(i \mapsto \mathop{\mathbb{E}}_{\mathcal{D}_i}[\mathrm{s} \circ h]; \boldsymbol{w}\right)}_{\text{True Aggregate M}} \leq \underbrace{\mathrm{M}\left(i \mapsto c \wedge \mathop{\hat{\mathbb{E}}}_{\boldsymbol{x}_{i,:}, \boldsymbol{y}_{i,:}}[\mathrm{s} \circ h] + \hat{\boldsymbol{\varepsilon}}_i; \boldsymbol{w}\right)}_{\text{UCB Estimate } \hat{\mathrm{M}}^{\uparrow}} \ , \tag{10}$$

thus if $\mathrm{M}(\cdot; \boldsymbol{w})$ is λ-Lipschitz-continuous w.r.t. some norm $\|\cdot\|_{\mathrm{M}}$, we have

$$\left|\underbrace{\mathrm{M}\left(i \mapsto \mathop{\hat{\mathbb{E}}}_{\boldsymbol{x}_{i,:}, \boldsymbol{y}_{i,:}}[\mathrm{s} \circ h]; \boldsymbol{w}\right)}_{\text{Plug-In Estimate}} - \underbrace{\mathrm{M}\left(i \mapsto \mathop{\mathbb{E}}_{\mathcal{D}_i}[\mathrm{s} \circ h]; \boldsymbol{w}\right)}_{\text{True Aggregate}}\right| \leq \lambda \|\hat{\boldsymbol{\varepsilon}}\|_{\mathrm{M}} \ . \tag{11}$$

Proof. We first note that, by the definition of $\mathrm{AEV}(\dots)$, with probability at least $1 - \delta$, for all $h \in \mathcal{H}$ and groups $i \in \mathcal{Z}$, it holds that $\left|\mathbb{E}_{\mathcal{D}_i}[\mathrm{s} \circ h] - \mathcal{S}_i^{\star}\right| \leq \hat{\boldsymbol{\varepsilon}}_i$. Given this, note that (9) and (10) follow directly from the *monotonicity assumption* on $\mathrm{M}(\cdot; \boldsymbol{w})$. Similarly, (11) follows from the monotonicity assumption and the definition of Lipschitz-continuity. □

Theorem 3.5 (Welfare and Malfare Expectation Bounds)**.** Suppose as in theorem 3.4, and assume also that $\mathrm{AEV}(\boldsymbol{m}, \delta, \boldsymbol{x}, \boldsymbol{y}) = \mathrm{AEV}(\boldsymbol{m}, \delta)$ is a deterministic *distribution-free* or *distribution-dependent* (but *not data-dependent*) bound. Then

$$\left|\mathrm{M} - \mathbb{E}[\hat{\mathrm{M}}]\right| \leq \mathbb{E}\left[|\mathrm{M} - \hat{\mathrm{M}}|\right] \leq \lambda \int_0^1 \left\|\mathrm{AEV}(\boldsymbol{m}, \delta)\right\|_{\mathrm{M}} \mathrm{d}\delta \ .$$

[8]It is of course possible to remove the leading 2 factor, if we are willing to work with the Rademacher average definition without the absolute value.

Proof. First note that by Jensen's inequality and convexity of $|\cdot|$, we have

$$\left|\mathrm{M} - \mathbb{E}[\hat{\mathrm{M}}]\right| \leq \mathbb{E}\left[\left|\mathrm{M} - \hat{\mathrm{M}}\right|\right] \enspace ,$$

thus all that remains to be shown is

$$\mathbb{E}\left[\left|\mathrm{M} - \hat{\mathrm{M}}\right|\right] \leq \lambda \int_0^1 \left\|\mathrm{AEV}(\boldsymbol{m}, \delta)\right\|_{\mathrm{M}} \, \mathrm{d}\delta \enspace .$$

By the definition of Lipschitz-continuity, note that for any $\mathcal{S}, \mathcal{S}'$, it holds that

$$\left|\mathrm{M}(\mathcal{S}; \boldsymbol{w}) - \mathrm{M}(\mathcal{S}'; \boldsymbol{w})\right| \leq \lambda \left\|\mathcal{S} - \mathcal{S}'\right\|_{\mathrm{M}} \enspace .$$

Now, we conclude the desideratum as

$$\begin{aligned}
\mathbb{E}\left[\left|\mathrm{M} - \hat{\mathrm{M}}\right|\right] &= \int_0^\infty \mathbb{P}\left(\left|\mathrm{M} - \hat{\mathrm{M}}\right| > \varepsilon\right) \mathrm{d}\varepsilon && \text{Properties of Expectation (Hazard Formula)} \\
&\leq \int_0^\infty \mathbb{P}\left(\lambda\|\mathcal{S} - \hat{\mathcal{S}}\|_{\mathrm{M}} > \varepsilon\right) \mathrm{d}\varepsilon && |\mathrm{M} - \hat{\mathrm{M}}| \leq \lambda\|\mathcal{S} - \hat{\mathcal{S}}\|_{\mathrm{M}} \\
&= \int_0^1 \inf\left(\epsilon \geq 0 \,:\, \mathbb{P}\left(\lambda\|\mathcal{S} - \hat{\mathcal{S}}\|_{\mathrm{M}} > \varepsilon\right) \leq \delta\right) \mathrm{d}\delta && \text{Integral of Inverse Formula} \\
&\leq \lambda \int_0^1 \left\|\mathrm{AEV}(\boldsymbol{m}, \delta)\right\|_{\mathrm{M}} \mathrm{d}\delta \enspace . && \text{Definition of AEV}(\dots)\text{, Lipschitz-Continuity}
\end{aligned}$$

□

Corollary 3.6 (Bernstein-Type Malfare Bounds)**.** Suppose as in theorem 3.1, and also per-group sample size m (i.e., $\boldsymbol{m} = \langle m, \dots, m\rangle$) and $p \geq 1$ power-mean malfare function $\mathrm{M}_p(\cdot; \boldsymbol{w})$. Now, let variance proxy v be defined in three cases as $v \doteq \mathrm{M}_{1/2}(\boldsymbol{v}; \boldsymbol{w}) = \left(\sum_{i=1}^g \boldsymbol{w}_i\sqrt{\boldsymbol{v}_i}\right)^2$ for $p = 1$, $v \doteq \boldsymbol{w} \cdot \boldsymbol{v}$ for $p \in (1, 2]$, or $v \doteq \|\boldsymbol{v}\|_\infty$ for $p > 2$. Then for all $\delta \in (0, 1)$, we have

1) $\mathbb{P}\left(\left|\mathrm{M} - \hat{\mathrm{M}}\right| \geq \frac{r \ln \frac{2g}{\delta}}{3m} + \sqrt{\frac{2v \ln \frac{2g}{\delta}}{m}}\right) \leq \delta$;
2) $\mathbb{E}\left[\left|\mathrm{M} - \hat{\mathrm{M}}\right|\right] \leq \frac{r \ln(2eg)}{3m} + \sqrt{\frac{2v \ln(2eg)}{m}}$; &
3) $\mathrm{M} \leq \mathbb{E}[\hat{\mathrm{M}}] \leq \mathrm{M} + \frac{r \ln(eg)}{3m} + \sqrt{\frac{2v \ln(eg)}{m}}$.

Proof. We first show the tail bounds (item 1), which we then use to show the expectation bounds (items 2 and 3). In particular, we apply theorem 3.1 item 2 to the singleton function family $\mathcal{H} \doteq \{h\}$ (thus $|\mathcal{H}| = 1$), with a union bound over all g groups to bound per-group confidence radii, i.e.,

$$\boldsymbol{\varepsilon}_i \leftarrow \frac{r \ln \frac{2g}{\delta}}{3m} + \sup_{h \in \mathcal{H}} \sqrt{\frac{2\,\mathbb{V}_{\mathcal{D}_i}[\mathrm{s} \circ h] \ln \frac{2g}{\delta}}{m}} = \frac{r \ln \frac{2g}{\delta}}{3m} + \sqrt{\frac{2\boldsymbol{v}_i \ln \frac{2g}{\delta}}{m}} \enspace .$$

We then apply (11) of theorem 3.4 to bound malfare tails in terms of these confidence radii, via the Lipschitz property of theorem 2.2 item 3. Subsequently, theorem 3.5 converts these tail bounds into expectation bounds for item 2, and similar logic for 1-tailed bounds yields item 3. The resulting bounds have rather convenient forms for $p \in \{1, 2, \infty\}$, so we relate all other cases to these three via monotonicity.

We first show the case of $p = 1$. Observe that

$$\begin{aligned}
\mathrm{M}_1(\boldsymbol{\varepsilon}; \boldsymbol{w}) &= \sum_{i=1}^g \boldsymbol{w}_i \left(\frac{r \ln \frac{2g}{\delta}}{3m} + \sqrt{\frac{2\boldsymbol{v}_i \ln \frac{2g}{\delta}}{m}}\right) && \text{Definition of } \mathrm{M}_1(\cdot; \boldsymbol{w}),\ \boldsymbol{\varepsilon} \\
&= \frac{r \ln \frac{2g}{\delta}}{3m} + \left(\sum_{i=1}^g \boldsymbol{w}_i\sqrt{\boldsymbol{v}_i}\right)\sqrt{\frac{2 \ln \frac{2g}{\delta}}{m}} = \frac{r \ln \frac{2g}{\delta}}{3m} + \sqrt{\frac{2\left(\sum_{i=1}^g \boldsymbol{w}_i\sqrt{\boldsymbol{v}_i}\right)^2 \ln \frac{2g}{\delta}}{m}} \enspace . && \text{Algebra}
\end{aligned}$$

Now suppose $p \in [1, 2]$. Then

$$\begin{aligned}
\mathrm{M}_p(\boldsymbol{\varepsilon}; \boldsymbol{w}) &\leq \mathrm{M}_2(\boldsymbol{\varepsilon}; \boldsymbol{w}) && \textsc{Power-Mean Inequality} \\
&= \sqrt{\sum_{i=1}^{g} \boldsymbol{w}_i \left(\frac{r \ln \frac{2g}{\delta}}{3m} + \sqrt{\frac{2\boldsymbol{v}_i \ln \frac{2g}{\delta}}{m}} \right)^2} && \textsc{Definition of } \mathrm{M}_2(\cdot; \boldsymbol{w}),\ \boldsymbol{\varepsilon} \\
&= \sqrt{\left(\frac{r \ln \frac{2g}{\delta}}{3m} \right)^2 + 2 \frac{r \ln \frac{2g}{\delta}}{m} \cdot \frac{\sum_{i=1}^{g} \boldsymbol{w}_i \sqrt{2\boldsymbol{v}_i \ln \frac{2g}{\delta}}}{\sqrt{m}} + \frac{2\boldsymbol{w} \cdot \boldsymbol{v} \ln \frac{2g}{\delta}}{m}} && \textsc{Algebra} \\
&\leq \sqrt{\left(\frac{r \ln \frac{2g}{\delta}}{3m} \right)^2 + 2 \frac{r \ln \frac{2g}{\delta}}{m} \sqrt{\frac{2 \sum_{i=1}^{g} \boldsymbol{w}_i \boldsymbol{v}_i \ln \frac{2g}{\delta}}{m}} + \frac{2\boldsymbol{w} \cdot \boldsymbol{v} \ln \frac{2g}{\delta}}{m}} && \textsc{Jensen's Inequality} \\
&= \sqrt{\left(\frac{r \ln \frac{2g}{\delta}}{3m} + \sqrt{\frac{2\boldsymbol{w} \cdot \boldsymbol{v} \ln \frac{2g}{\delta}}{m}} \right)^2} = \frac{r \ln \frac{2g}{\delta}}{3m} + \sqrt{\frac{2\boldsymbol{w} \cdot \boldsymbol{v} \ln \frac{2g}{\delta}}{m}} \ . && \textsc{Factoring}
\end{aligned}$$

Finally, for $p \in [1, \infty]$, we have

$$\mathrm{M}_p(\boldsymbol{\varepsilon}; \boldsymbol{w}) \leq \mathrm{M}_\infty(\boldsymbol{\varepsilon}; \boldsymbol{w}) = \max_{i \in \mathcal{Z}} \frac{r \ln \frac{2g}{\delta}}{3m} + \sqrt{\frac{2\boldsymbol{v}_i \ln \frac{2g}{\delta}}{m}} = \frac{r \ln \frac{2g}{\delta}}{3m} + \sqrt{\frac{2\|\boldsymbol{v}\|_\infty \ln \frac{2g}{\delta}}{m}} \ .$$

We now show the *expectation bounds* of items 2 and 3. Via theorem 3.5, we seek to *bound integrals* of the form

$$\mathbb{E}\left[\!\left|\mathrm{M} - \hat{\mathrm{M}}\right|\!\right] \leq \int_0^1 \mathrm{M}_p\big(\mathrm{AEV}(\langle m, \dots, m \rangle, \delta); \boldsymbol{w}\big)\,\mathrm{d}\delta \leq \int_0^1 \sqrt{\frac{2v \ln \frac{g}{\delta}}{m}} + \frac{r \ln \frac{g}{\delta}}{3m}\,\mathrm{d}\delta \ ,$$

for the various v as defined above. Note that we will encounter either $\ln \frac{2g}{\delta}$ or $\ln \frac{g}{\delta}$ terms, depending on whether we employ 2-tailed or 1-tailed bounds, in items 2 or 3, respectively. The difficult part of these integrals is

$$\begin{aligned}
\int_0^1 \sqrt{\ln \frac{g}{\delta}}\,\mathrm{d}\delta &= \tfrac{1}{2}\sqrt{\pi} g \operatorname{erfc}\big(\sqrt{\ln g}\big) + \sqrt{\ln g} && \textsc{Integration} \\
&\leq \sqrt{\ln(eg)} \ . && \forall u \geq 1:\ \sqrt{\ln u} + u \tfrac{1}{2}\sqrt{\pi} \operatorname{erfc}(\sqrt{\ln u}) \leq \sqrt{\ln(eu)}
\end{aligned}$$

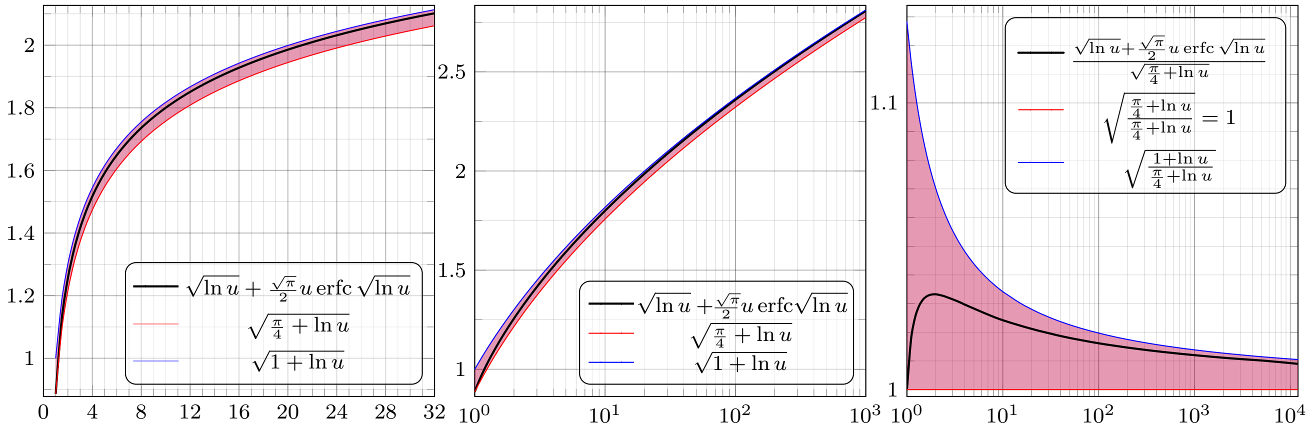
Note that the integration is exact, and the looseness of the erfc$(\cdot)$ (complementary error function) bound is quite minuscule, both multiplicatively and additively, as $\forall u \geq 1:\ 0 \leq \sqrt{\ln(eu)} - \Big(\sqrt{\ln u} + u\frac{1}{2}\sqrt{\pi}\operatorname{erfc}(\sqrt{\ln u})\Big) \leq 1 - \frac{\sqrt{\pi}}{2}$, see figure A1 for visualization. Observe also that the logarithmic terms in the fast-decaying summands integrate (improperly) as

$$\int_0^1 \ln \frac{g}{\delta}\,\mathrm{d}\delta = \ln(g) - \int_0^1 \ln \delta \,\mathrm{d}\delta = \ln(g) + 1 = \ln(eg) \ .$$

Substituting the values of v derived above into these indefinite integrals then yields the desiderata. In particular, item 2 substitutes into theorem 3.5 directly to bound $\mathbb{E}[\![\mathrm{M} - \hat{\mathrm{M}}|]$, and item 3 substitutes into a directed variant of theorem 3.5 (i.e., one considering single-sided error, rather than absolute error) to upper-bound $\mathbb{E}[\hat{\mathrm{M}}]$, paired with the observation that, by Jensen's inequality (as $p \geq 1$ power means are convex), $\mathrm{M} \leq \mathbb{E}[\hat{\mathrm{M}}]$. □

We now show theorem 3.7, which provides regret bounds analogous to the aggregator function bounds of theorem 3.4.

Theorem 3.7 (Regret Estimation Bounds)**.** Suppose sentiment function $\mathrm{s}(\cdot, \cdot) \colon \mathcal{Y}' \times \mathcal{Y} \to \mathbb{R}_{0+}$, per-group probability distributions $\mathcal{D}_{1:g}$, sample size vector $\boldsymbol{m} \in \mathbb{Z}_+^g$, samples $(\boldsymbol{x}, \boldsymbol{y}) \sim \mathcal{D}_1^{\boldsymbol{m}_1} \times \dots \times \mathcal{D}_g^{\boldsymbol{m}_g}$, failure probability $\delta \in (0, 1)$, and additive error bound AEV$(\dots)$, and let $\hat{\boldsymbol{\varepsilon}} \leftarrow \mathrm{AEV}(\boldsymbol{m}, \delta, \boldsymbol{x}, \boldsymbol{y})$. Then for all $h \in \mathcal{H}$ and all monotonic malfare functions



(a) Natural axes plot. Note that the shaded region between the lower and upper bounds is difficult to see without scaling, and its height is maximized at $u = 1$.

(b) Semilog plot to emphasize asymptotic behavior. Note that even with a logarithmic x-axis, the absolute gap between the lower and upper bounds remains quite small.

(c) Semilog plot, wherein each function is divided by the lower bound $\sqrt{\frac{\pi}{4} + \ln u}$ to emphasize the decaying relative gap between lower and upper bounds.

Figure A1: Plots of various scalings of $\sqrt{\ln u} + \frac{\sqrt{\pi}}{2} u \operatorname{erfc} \sqrt{\ln u}$ and lower and upper bounds thereof (y-axis), i.e., $\sqrt{\frac{\pi}{4} + \ln u} \leq \sqrt{\ln u} + \frac{\sqrt{\pi}}{2} u \operatorname{erfc} \sqrt{\ln u} \leq \sqrt{1 + \ln u}$, see proof of corollary 3.6, against $u \geq 1$ (x-axis). The region sandwiched by the lower and upper bounds is shaded in purple, is quite small for all $u \geq 1$, and converges to 0 both additively and multiplicatively.

$\mathrm{M}(\cdot; \boldsymbol{w})$, it holds with probability at least $1 - \delta$ over $\boldsymbol{x}$, $\boldsymbol{y}$, and $\hat{\boldsymbol{\varepsilon}}$ that

$$\underbrace{\mathrm{M}\left(i \mapsto 0 \vee \left|\mathop{\mathbb{E}}_{\mathcal{D}_i}[\mathrm{s} \circ h] - \mathcal{S}_i^\star\right| - 2\hat{\boldsymbol{\varepsilon}}_i; \boldsymbol{w}\right)}_{\text{True Regret Malfare LB}} \leq \underbrace{\mathrm{M}\left(i \mapsto \left|\mathop{\hat{\mathbb{E}}}_{\boldsymbol{x}_{i,:}, \boldsymbol{y}_{i,:}}[\mathrm{s} \circ h] - \hat{\mathcal{S}}_i\right|; \boldsymbol{w}\right)}_{\text{Plug-In Regret Malfare}} \leq \underbrace{\mathrm{M}\left(i \mapsto c \wedge \left|\mathop{\mathbb{E}}_{\mathcal{D}_i}[\mathrm{s} \circ h] - \mathcal{S}_i^\star\right| + 2\hat{\boldsymbol{\varepsilon}}_i; \boldsymbol{w}\right)}_{\text{True Regret Malfare UB}} \ , \quad \& \tag{14}$$

$$\underbrace{\mathrm{M}\left(i \mapsto 0 \vee \left|\mathop{\hat{\mathbb{E}}}_{\boldsymbol{x}_{i,:}, \boldsymbol{y}_{i,:}}[\mathrm{s} \circ h] - \hat{\mathcal{S}}_i\right| - 2\hat{\boldsymbol{\varepsilon}}_i; \boldsymbol{w}\right)}_{\text{LCB Estimate}} \leq \underbrace{\mathrm{M}\left(i \mapsto \left|\mathop{\mathbb{E}}_{\mathcal{D}_i}[\mathrm{s} \circ h] - \mathcal{S}_i^\star\right|; \boldsymbol{w}\right)}_{\text{True Regret Malfare}} \leq \underbrace{\mathrm{M}\left(i \mapsto c \wedge \left|\mathop{\hat{\mathbb{E}}}_{\boldsymbol{x}_{i,:}, \boldsymbol{y}_{i,:}}[\mathrm{s} \circ h] - \hat{\mathcal{S}}_i\right| + 2\hat{\boldsymbol{\varepsilon}}_i; \boldsymbol{w}\right)}_{\text{UCB Estimate}} \ , \tag{15}$$

thus if $\mathrm{M}(\cdot; \boldsymbol{w})$ is λ-Lipschitz-continuous w.r.t. some norm $\|\cdot\|_{\mathrm{M}}$, we have

$$\left|\underbrace{\mathrm{M}\left(i \mapsto \left|\mathop{\hat{\mathbb{E}}}_{\boldsymbol{x}_{i,:}, \boldsymbol{y}_{i,:}}[\mathrm{s} \circ h] - \hat{\mathcal{S}}_i\right|; \boldsymbol{w}\right)}_{\text{Plug-In Regret Malfare}} - \underbrace{\mathrm{M}\left(i \mapsto \left|\mathop{\mathbb{E}}_{\mathcal{D}_i}[\mathrm{s} \circ h] - \mathcal{S}_i^\star\right|; \boldsymbol{w}\right)}_{\text{True Regret Malfare}}\right| \leq 2\lambda \|\hat{\boldsymbol{\varepsilon}}\|_{\mathrm{M}} \ . \tag{16}$$

Proof. This result follows essentially the same reasoning as theorem 3.4. The salient difference here is that we now bound both the estimation error of h and each $\boldsymbol{h}_i^\star$, which when summed yield the novel 2-factors. In particular, we have that with probability at least $1 - \delta$, it holds for each $i \in \mathcal{Z}$ and $h \in \mathcal{H}$ (simultaneously) that

$$\begin{aligned}
\left|\mathop{\mathbb{E}}_{\mathcal{D}_i}[\mathrm{s} \circ h] - \mathcal{S}_i^\star\right| &= \left|\left(\mathop{\hat{\mathbb{E}}}_{\boldsymbol{x}_{i,:}, \boldsymbol{y}_{i,:}}[\mathrm{s} \circ h] - \hat{\mathcal{S}}_i\right) + \left(\mathop{\mathbb{E}}_{\mathcal{D}_i}[\mathrm{s} \circ h] - \mathcal{S}_i^\star\right) - \left(\mathop{\hat{\mathbb{E}}}_{\boldsymbol{x}_{i,:}, \boldsymbol{y}_{i,:}}[\mathrm{s} \circ h] - \hat{\mathcal{S}}_i\right)\right| && \text{Algebra} \\
&\leq \left|\mathop{\hat{\mathbb{E}}}_{\boldsymbol{x}_{i,:}, \boldsymbol{y}_{i,:}}[\mathrm{s} \circ h] - \hat{\mathcal{S}}_i\right| + \left|\mathop{\mathbb{E}}_{\mathcal{D}_i}[\mathrm{s} \circ h] - \mathop{\hat{\mathbb{E}}}_{\boldsymbol{x}_{i,:}, \boldsymbol{y}_{i,:}}[\mathrm{s} \circ h]\right| + \left|\mathcal{S}_i^\star - \hat{\mathcal{S}}_i\right| && \text{Triangle Inequality} \\
&\leq \left|\mathop{\hat{\mathbb{E}}}_{\boldsymbol{x}_{i,:}, \boldsymbol{y}_{i,:}}[\mathrm{s} \circ h] - \hat{\mathcal{S}}_i\right| + \hat{\boldsymbol{\varepsilon}}_i + \hat{\boldsymbol{\varepsilon}}_i = \left|\mathop{\hat{\mathbb{E}}}_{\boldsymbol{x}_{i,:}, \boldsymbol{y}_{i,:}}[\mathrm{s} \circ h] - \hat{\mathcal{S}}_i\right| + 2\hat{\boldsymbol{\varepsilon}}_i \ . && \text{Definition of } \hat{\boldsymbol{\varepsilon}}
\end{aligned}$$

This result, paired with the assumed monotonicity of $\mathrm{M}(\cdot; \boldsymbol{w})$, yields (14) and (15), and then applying the definition of Lipschitz continuity yields (16). ∎

We now show property 3.8, which approximates the reduction to the UCB resultant from drawing a single additional sample for some group i.

Property 3.8 (Incremental Gain of Sampling)**.** Suppose $\boldsymbol{w}$-weighted power-mean malfare $\mathrm{M}_p(\cdot;\boldsymbol{w})$, sample $(\boldsymbol{x},\boldsymbol{y})$ with group sample sizes $\boldsymbol{m}_{1:g}$, and let $\boldsymbol{x}',\boldsymbol{y}'$ extend $\boldsymbol{x},\boldsymbol{y}$ to sample sizes $\boldsymbol{m}'$, where $\boldsymbol{m}'=\boldsymbol{m}+\mathbb{1}_i$, i.e., group i has one additional sample. Now, let $\hat{\boldsymbol{\varepsilon}}\leftarrow\mathrm{AEV}(\boldsymbol{m},\delta,\boldsymbol{x},\boldsymbol{y})$ and $\tilde{\boldsymbol{\varepsilon}}\leftarrow\mathrm{AEV}(\boldsymbol{m}',\delta,\boldsymbol{x}',\boldsymbol{y}')$, and take $\hat{h}\doteq\operatorname{argmin}_{h\in\mathcal{H}}\mathrm{M}_p\Big(i\mapsto\hat{\mathbb{E}}_{\boldsymbol{x}_{i,:},\boldsymbol{y}_{i,:}}[\ell\circ h]+\hat{\boldsymbol{\varepsilon}}_i;\boldsymbol{w}\Big)$, $\hat{\mathrm{M}}\doteq\mathrm{M}_p\Big(i\mapsto\hat{\mathbb{E}}_{\boldsymbol{x}_{i,:},\boldsymbol{y}_{i,:}}[\ell\circ\hat{h}];\boldsymbol{w}\Big)$, $\hat{\mathrm{M}}^{\uparrow}\doteq\mathrm{M}_p\Big(i\mapsto\hat{\mathbb{E}}_{\boldsymbol{x}_{i,:},\boldsymbol{y}_{i,:}}[\ell\circ\hat{h}]+\hat{\boldsymbol{\varepsilon}}_i;\boldsymbol{w}\Big)$, and $\tilde{\mathrm{M}}^{\uparrow}\doteq\inf_{h\in\mathcal{H}}\mathrm{M}_p\Big(i\mapsto\hat{\mathbb{E}}_{\boldsymbol{x}'_{i,:},\boldsymbol{y}'_{i,:}}[\ell\circ h]+\tilde{\boldsymbol{\varepsilon}}_i;\boldsymbol{w}\Big)$.

Then the *incremental impact* of sampling from group i on the UCB is approximately

$$\hat{\mathrm{M}}^{\uparrow}-\tilde{\mathrm{M}}^{\uparrow}\approx\frac{\hat{\boldsymbol{\varepsilon}}_i\boldsymbol{w}_i}{2\boldsymbol{m}_i+\frac{3}{2}}\left(\frac{\hat{\mathbb{E}}_{\boldsymbol{x}_{i,:},\boldsymbol{y}_{i,:}}[\ell\circ\hat{h}]+\hat{\boldsymbol{\varepsilon}}_i}{\hat{\mathrm{M}}^{\uparrow}}\right)^{p-1}\approx\frac{\hat{\boldsymbol{\varepsilon}}_i\boldsymbol{w}_i}{2\boldsymbol{m}_i}\left(\frac{\hat{\mathbb{E}}_{\boldsymbol{x}_{i,:},\boldsymbol{y}_{i,:}}[\ell\circ\hat{h}]}{\hat{\mathrm{M}}}\right)^{p-1}\quad. \tag{17}$$

Proof. We first assume that for each $h\in\mathcal{H}$, $i\in\mathcal{Z}$, it holds that $\hat{\mathbb{E}}_{\boldsymbol{x}_{i,:},\boldsymbol{y}_{i,:}}[\ell\circ h]\approx\hat{\mathbb{E}}_{\boldsymbol{x}'_{i,:},\boldsymbol{y}'_{i,:}}[\ell\circ h]$, and we use this approximation throughout. The result then follows directly from three observations.

First, note that $\frac{\partial}{\partial\mathcal{S}_i}\mathrm{M}_p(\mathcal{S};\boldsymbol{w})=\frac{\boldsymbol{w}_i\mathcal{S}_i^{p-1}}{\mathrm{M}_p^{p-1}(\mathcal{S};\boldsymbol{w})}$, thus for any UCB-optimal $\hat{h}$, we have the *subderivative*[9]

$$\frac{\partial_{\mathrm{sub}}}{\partial_{\mathrm{sub}}\hat{\boldsymbol{\varepsilon}}_i}\inf_{h\in\mathcal{H}}\mathrm{M}_p\left(i\mapsto\mathop{\hat{\mathbb{E}}}_{\boldsymbol{x}_{i,:},\boldsymbol{y}_{i,:}}[\ell\circ\hat{h}]+\hat{\boldsymbol{\varepsilon}}_i;\boldsymbol{w}\right)\ni\boldsymbol{w}_i\left(\frac{\hat{\mathbb{E}}_{\boldsymbol{x}_{i,:},\boldsymbol{y}_{i,:}}[\ell\circ\hat{h}]+\hat{\boldsymbol{\varepsilon}}_i}{\hat{\mathrm{M}}^{\uparrow}}\right)^{p-1}\quad.$$

Now, assuming $\boldsymbol{\Theta}\sqrt{\frac{1}{\boldsymbol{m}_i}}$ asymptotic behavior of $\hat{\boldsymbol{\varepsilon}}_i$, observe that

$$\tilde{\boldsymbol{\varepsilon}}_i\approx\hat{\boldsymbol{\varepsilon}}_i\sqrt{\frac{\boldsymbol{m}_i}{\boldsymbol{m}_i+1}}\implies(\hat{\boldsymbol{\varepsilon}}_i-\tilde{\boldsymbol{\varepsilon}}_i)\approx\hat{\boldsymbol{\varepsilon}}_i\left(1-\sqrt{\frac{\boldsymbol{m}_i}{\boldsymbol{m}_i+1}}\right)\approx\frac{\hat{\boldsymbol{\varepsilon}}_i}{2\boldsymbol{m}_i+\frac{3}{2}}\approx\frac{\hat{\boldsymbol{\varepsilon}}_i}{2\boldsymbol{m}_i}\quad,$$

and note that this approximation is quite sharp (see figure A2), as

$$\forall u\geq 1:\ \frac{1}{2u+\frac{3}{2}}\leq 1-\sqrt{\frac{u}{u+1}}\leq\frac{1}{2u+\sqrt{2}}\quad. \tag{22}$$

Finally, using the subderivative approximation, the impact of the additional sample on the malfare is approximately

$$\begin{aligned}
\hat{\mathrm{M}}^{\uparrow}-\tilde{\mathrm{M}}^{\uparrow}&\approx(\hat{\boldsymbol{\varepsilon}}_i-\tilde{\boldsymbol{\varepsilon}}_i)\frac{\partial_{\mathrm{sub}}}{\partial_{\mathrm{sub}}\hat{\boldsymbol{\varepsilon}}_i}\inf_{h\in\mathcal{H}}\mathrm{M}_p\left(i\mapsto\mathop{\hat{\mathbb{E}}}_{\boldsymbol{x}_{i,:},\boldsymbol{y}_{i,:}}[\ell\circ\hat{h}]+\hat{\boldsymbol{\varepsilon}}_i;\boldsymbol{w}\right)\Bigg|_{\hat{\boldsymbol{\varepsilon}}_i} && \textsc{Finite Difference Approximation}\ \mathop{\hat{\mathbb{E}}}_{\boldsymbol{x}_{i,:},\boldsymbol{y}_{i,:}}[\ell\circ h]\approx\mathop{\hat{\mathbb{E}}}_{\boldsymbol{x}'_{i,:},\boldsymbol{y}'_{i,:}}[\ell\circ h]\\
&\approx\hat{\boldsymbol{\varepsilon}}_i\left(1-\sqrt{\frac{\boldsymbol{m}_i}{\boldsymbol{m}_i+1}}\right)\boldsymbol{w}_i\left(\frac{\hat{\mathbb{E}}_{\boldsymbol{x}_{i,:},\boldsymbol{y}_{i,:}}[\ell\circ\hat{h}]+\hat{\boldsymbol{\varepsilon}}_i}{\hat{\mathrm{M}}^{\uparrow}}\right)^{p-1} && \textsc{See Above}\\
&\approx\frac{\hat{\boldsymbol{\varepsilon}}_i}{2\boldsymbol{m}_i+\frac{3}{2}}\,\boldsymbol{w}_i\left(\frac{\hat{\mathbb{E}}_{\boldsymbol{x}_{i,:},\boldsymbol{y}_{i,:}}[\ell\circ\hat{h}]+\hat{\boldsymbol{\varepsilon}}_i}{\hat{\mathrm{M}}^{\uparrow}}\right)^{p-1} && \textsc{See Above}\\
&\approx\frac{\hat{\boldsymbol{\varepsilon}}_i\boldsymbol{w}_i}{2\boldsymbol{m}_i}\left(\frac{\hat{\mathbb{E}}_{\boldsymbol{x}_{i,:},\boldsymbol{y}_{i,:}}[\ell\circ\hat{h}]}{\hat{\mathrm{M}}}\right)^{p-1}\quad. && \frac{1}{\boldsymbol{m}_i+\boldsymbol{\Theta}(1)}\approx\frac{1}{\boldsymbol{m}_i}\ \ \frac{\mathcal{S}_i+\boldsymbol{\varepsilon}_i}{\mathrm{M}_p(\mathcal{S}+\boldsymbol{\varepsilon};\boldsymbol{w})}\approx\frac{\mathcal{S}_i}{\mathrm{M}_p(\mathcal{S};\boldsymbol{w})}
\end{aligned}$$

□

A.2 Uniformly-Convergent Sampling Schedules and Progressive Sampling Guarantees

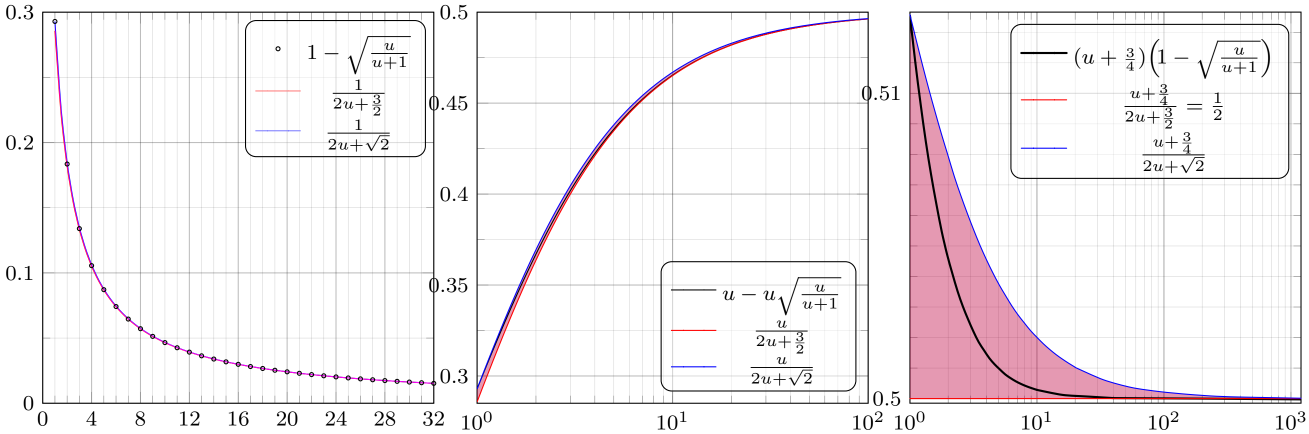
In this subsection, we show all results relating to sampling schedules and progressive sampling. Before showing the correctness of algorithms 1 and 2, we begin with the uniformly-convergent schedule analysis of lemma 4.4.

Lemma 4.4 (Sufficient Conditions for Uniformly-Convergent Geometric Schedules)**.** Suppose as in definition 4.2, and assume also that

$$\sup_{(\boldsymbol{x},\boldsymbol{y})\in(\mathcal{X}\times\mathcal{Y})^{\boldsymbol{s}_T\times g}}\left\|\mathrm{AEV}\left(\langle\boldsymbol{s}_T,\ldots,\boldsymbol{s}_T\rangle,\tfrac{\delta}{T},\boldsymbol{x},\boldsymbol{y}\right)\right\|_{\mathrm{M}}\leq\varepsilon\quad. \tag{19}$$

Then the geometric-uniform schedule $(\boldsymbol{s},\,\boldsymbol{\delta})$ is ε-uniformly-convergent.

[9] Technically, $\inf_{h\in\mathcal{H}}\mathrm{M}_p\left(i\mapsto\hat{\mathbb{E}}_{\boldsymbol{x}_{i,:},\boldsymbol{y}_{i,:}}[\ell\circ\hat{h}]+\hat{\boldsymbol{\varepsilon}}_i;\boldsymbol{w}\right)$ is not always concave in $\hat{\boldsymbol{\varepsilon}}$, thus this quantity is not always a subderivative, but it is at least a reasonable linear approximation. Furthermore, a sufficient condition for convexity is for $\mathrm{M}_p\left(i\mapsto\hat{\mathbb{E}}_{\boldsymbol{x}_{i,:},\boldsymbol{y}_{i,:}}[\ell\circ\hat{h}]+\hat{\boldsymbol{\varepsilon}}_i;\boldsymbol{w}\right)$, which is always convex in $\hat{\boldsymbol{\varepsilon}}$, to also be convex in the space of $\hat{h}\in\mathcal{H}$ (as is quite common in machine learning and optimization contexts).



(a) Natural axes plot. Note that the shaded region between the lower and upper bounds is difficult to see without scaling.

(b) Semilog plot, wherein each function is scaled by u to emphasize their asymptotic $\mathbf{\Theta}(\frac{1}{u})$ behavior.

(c) Semilog plot, wherein each function is scaled by $(u+\frac{3}{4})$ to better illustrate the gap between lower and upper bounds.

Figure A2: Plots of various scalings of $1-\sqrt{\frac{u}{u+1}}$ and lower and upper bounds thereof (y-axis), i.e., $\frac{1}{2u+\frac{3}{2}} \leq 1-\sqrt{\frac{u}{u+1}} \leq \frac{1}{2u+\sqrt{2}}$, see (22), against $u \geq 1$ (x-axis). The region sandwiched by the lower and upper bounds is shaded in purple, and we note that even when scaled by u, this region rapidly converges to 0 — in particular, the gap is bounded as $\frac{1}{2u+\sqrt{2}} - \frac{1}{2u+\frac{3}{2}} \leq \frac{\frac{3}{2}-\sqrt{2}}{4u^2}$.

Furthermore, suppose as in definition 4.3, $\alpha \geq \frac{1}{\delta}$, and assume that

$$\lim_{m\to\infty} \sup_{(\boldsymbol{x},\boldsymbol{y})\in(\mathcal{X}\times\mathcal{Y})^{m\times g}} \left\|\mathrm{AEV}\left(\langle m,\ldots,m\rangle, \tfrac{\beta-1}{\beta(m+1)}, \boldsymbol{x},\boldsymbol{y}\right)\right\|_{\mathrm{M}} = 0 \ . \tag{20}$$

Then the double-geometric schedule $(\boldsymbol{s}, \boldsymbol{\delta})$ is 0-uniformly-convergent.

Proof. Recall that, via (18) of definition 4.1, the goal is to show

$$\inf_{t\in\mathbb{Z}_+} \sup_{(\boldsymbol{x},\boldsymbol{y})\in(\mathcal{X}\times\mathcal{Y})^{\boldsymbol{s}_t\times g}} \left\|\mathrm{AEV}(\langle \boldsymbol{s}_t,\ldots,\boldsymbol{s}_t\rangle, \boldsymbol{\delta}_t, \boldsymbol{x},\boldsymbol{y})\right\|_{\mathrm{M}} \leq \varepsilon \ .$$

We now show the results for the geometric-uniform and double-geometric schedules in separate parts.

The result for the geometric-uniform schedule holds by the regularity conditions assumed on AEV(…), where the infimum occurs, possibly nonuniquely, at $t = T$. In particular, this is implied by condition 3, thus we have

$$\begin{aligned}\inf_{t\in\mathbb{Z}_+}\sup_{(\boldsymbol{x},\boldsymbol{y})\in(\mathcal{X}\times\mathcal{Y})^{\boldsymbol{s}_t\times g}} \left\|\mathrm{AEV}(\langle \boldsymbol{s}_t,\ldots,\boldsymbol{s}_t\rangle, \boldsymbol{\delta}_t, \boldsymbol{x},\boldsymbol{y})\right\|_{\mathrm{M}} &= \sup_{(\boldsymbol{x},\boldsymbol{y})\in(\mathcal{X}\times\mathcal{Y})^{\boldsymbol{s}_T\times g}} \left\|\mathrm{AEV}(\langle \boldsymbol{s}_t,\ldots,\boldsymbol{s}_T\rangle, \tfrac{\delta}{T}, \boldsymbol{x},\boldsymbol{y})\right\|_{\mathrm{M}} && \text{Condition 3}\\ &\leq \varepsilon \ . && \text{By Assumption}\end{aligned}$$

We may thus conclude that the uniform-geometric schedule is ε-uniformly-convergent.

The result for the double-geometric schedule uses the regularity conditions on our bound AEV(…), and also derives a bound on $\boldsymbol{\delta}_t$ in terms of $\boldsymbol{s}_t$, which allows us to reason over only sample sizes, rather than the schedule itself. Let $t(m)$ denote the smallest timestep t for which a sample of size m is guaranteed to be drawn. In particular, ignoring the ceiling operator, and the fact that not every possible sample size $m \in \mathbb{Z}_+$ is in the schedule $\boldsymbol{s}$, the inverse of $m = \alpha\beta^t$ gives us $t(m) \geq \log_\beta \frac{m}{\alpha}$. However, the sample size m may be up to a factor β larger due to the schedule discretization, and rounded up, thus we have the matching upper-bound $t(m) \leq \log_\beta \frac{\beta(m+1)}{\alpha}$. With this result in hand, we now substitute and simplify to get

$$
\begin{aligned}
\boldsymbol{\delta}_{t(m)} &\geq \frac{\delta(\beta-1)}{\beta^{\log_\beta \frac{\beta(m+1)}{\alpha}}} && \text{See Above Definition 4.3} \\
&= \frac{\alpha\delta(\beta-1)}{\beta(m+1)} && \text{Algebra} \\
&\geq \frac{\beta-1}{\beta(m+1)} \ . && \alpha \geq \frac{1}{\delta}
\end{aligned}
$$

Now, observe that $\left\|\mathrm{AEV}(\dots)\right\|_{\mathrm{M}}$ is monotonic in $\boldsymbol{m}$ and δ by regularity condition 3, and thus, by the specified limit condition, may be taken arbitrarily close to 0 for sufficient values of m at timestep $t(m)$. We thus conclude that the double-geometric schedule is 0-uniformly-convergent. ☐

We now show the correctness results for algorithms 1 and 2, namely theorems 4.5 and 4.6.

Theorem 4.5 (Linear PS Guarantees)**.** Suppose $(\hat{h}, \hat{\mu}, \hat{\varepsilon}, \mathrm{M}^{\star\downarrow}) \leftarrow \textsc{LinearPSLoss}(\mathcal{H}, \ell(\cdot,\cdot), \mathcal{D}, \mathrm{AEV}(\dots), \boldsymbol{s}, \boldsymbol{\delta}, \varepsilon, \mathrm{M}(\cdot;\boldsymbol{w}), \textsc{Reg})$, $\mathrm{M}(\mathcal{S};\boldsymbol{w})$ is continuous and monotonic in $\mathcal{S}$ with (possibly infinite) Lipschitz constant λ_{M} w.r.t. $\|\cdot\|_{\mathrm{M}}$, and the schedules $(\boldsymbol{s}, \boldsymbol{\delta})$ are $\frac{\varepsilon}{\lambda_{\mathrm{M}}(1+\mathbb{1}_{\textsc{Reg}})}$-uniformly-convergent w.r.t. $\mathrm{AEV}(\dots)$ and $\|\cdot\|_{\mathrm{M}}$. Now take μ to be the true objective value of $\hat{h}$ and $\mu^\star$ to be the true objective value of the optimal $h^\star$, i.e., if $\textsc{Reg} = \textsc{False}$, take $\mu \doteq \mathrm{M}(i \mapsto \mathbb{E}_{\mathcal{D}_i}[\ell \circ \hat{h}]; \boldsymbol{w})$ and $\mu^\star \doteq \inf_{h \in \mathcal{H}} \mathrm{M}(i \mapsto \mathbb{E}_{\mathcal{D}_i}[\ell \circ h]; \boldsymbol{w})$, or if $\textsc{Reg} = \textsc{True}$, take (see equation 3) $\mu \doteq \mathrm{M}(i \mapsto \mathrm{Reg}_i(\hat{h}); \boldsymbol{w})$ and $\mu^\star \doteq \inf_{h \in \mathcal{H}} \mathrm{M}\left(i \mapsto \mathrm{Reg}_i(h); \boldsymbol{w}\right)$. Then, with probability at least $1-\delta$, the output $(\hat{h}, \hat{\mu}, \hat{\varepsilon}, \mathrm{M}^{\star\downarrow})$ obeys
1) $|\hat{\mu} - \mu| \leq \hat{\varepsilon} \leq \varepsilon$; & 2) $\mathrm{M}^{\star\downarrow} \leq \mu^\star \leq \mu \leq \hat{\mu} + \hat{\varepsilon} \leq \mathrm{M}^{\star\downarrow} + 2\varepsilon$.

Proof. We first show that algorithm 1 is guaranteed to eventually terminate[10] after reaching some timestep $\hat{t}$. Note that at timestep t, under the joint sampling model, $\mathrm{AEV}(\dots)$ is evaluated with sample sizes $\boldsymbol{m} = \langle \boldsymbol{s}_t, \dots, \boldsymbol{s}_t \rangle$, and under the mixture sampling model, sample sizes are at least that large, i.e., $\boldsymbol{m} \succeq \langle \boldsymbol{s}_t, \dots, \boldsymbol{s}_t \rangle$, which by the monotonicity regularity condition (item 3) on $\mathrm{AEV}(\dots)$, implies that the bounds with larger sample sizes are at least as sharp (in the worst case; for most reasonable bounds, increasing sample sizes yields improvement).

By the definition of $\frac{\varepsilon}{\lambda_{\mathrm{M}}(1+\mathbb{1}_{\textsc{Reg}})}$-$\|\cdot\|_{\mathrm{M}}$-uniform-convergence (definition 4.1), for finite λ, we eventually obtain a sufficiently accurate estimate (i.e., $\hat{\boldsymbol{\varepsilon}}$ is sufficiently close to $\mathbf{0}$) such that by the Lipschitz guarantees of theorem 3.4 or theorem 3.7, the algorithm will terminate (line 19). Furthermore, for infinite λ, $\lim_{t \to \infty} \hat{\boldsymbol{\varepsilon}} = \mathbf{0}$, which again implies eventual termination, now under only the continuity assumption. In particular, both $\mathrm{M}^{\star\downarrow}$ and $\hat{\mathrm{M}}^{\uparrow}$ are eventually ε-estimated, for a total error of $\leq 2\varepsilon$, which yields termination.

To see the correctness of this result, we first observe that by union bound over $t \in \mathbb{Z}_+$, all tail bounds of $\mathrm{AEV}(\dots)$ hold simultaneously with probability at least $1 - \sum_{i=1}^{\infty} \boldsymbol{\delta}_i = 1 - \delta$ (recall that δ is defined as such by regularity condition 2). Under the joint sampling model, this union bound is a simple bound over $\mathrm{AEV}(\dots)$ evaluated at sample size vectors $\langle \boldsymbol{s}_1, \dots, \boldsymbol{s}_1 \rangle, \langle \boldsymbol{s}_2, \dots, \boldsymbol{s}_2 \rangle, \dots$, and failure probability values $\boldsymbol{\delta}_1, \boldsymbol{\delta}_2, \dots$. At first glance, this strategy seems invalid for the mixture sampling model, as the sample sizes are actually not known *a priori*, since they depend on the groups sampled, however a subtle conditioning argument circumvents this objection. The simple trick we employ is to condition the algorithm on the infinite sequence of $\boldsymbol{z} \in 2^{\mathcal{Z}}$ samples drawn on line 10, and establish the schedule *after performing this conditioning operation*, thus the sample sizes are fixed from the perspective of the schedule, and the samples themselves remain *conditionally i.i.d.* given this order.This technique is entirely valid, because the sample sizes depend only on these $\boldsymbol{z}$ values, and are *conditionally independent* from the actual samples $\boldsymbol{x}$ and $\boldsymbol{y}$.

We henceforth assume that we are in the probability at least $1-\delta$ case where all $\hat{\boldsymbol{\varepsilon}}$ bounds taken over the course of the schedule hold. With this established, items 1 and 2 follow via theorem 3.4 if $\textsc{Reg} = \textsc{True}$, and via theorem 3.7 otherwise. In particular, item 1 holds, as $\hat{h} \in \mathcal{H}$, and $\hat{\varepsilon}$ was computed with $\mathrm{AEV}(\dots)$, we get $|\hat{\mu} - \mu| \leq \hat{\varepsilon}$, and by the termination condition, $\hat{\varepsilon} \leq \varepsilon$. To see item 2, observe that it is implied by item 1, coupled with the bound $\mathrm{M}^{\star\downarrow} \leq \mu^\star$, which holds for similar reasons. ☐

[10]Technically, under the mixture sampling model it is possible that the loop (lines 9–12) runs infinitely, but so long as each group is sampled with nonzero probability, this is a zero probability event.

Theorem 4.6 (Braided PS Guarantees)**.** Suppose $(\hat{h}, \hat{\mu}, \hat{\varepsilon}, \mathrm{M}^{\star\downarrow}) \leftarrow \textsc{BraidedPSLoss}(\mathcal{H}, \ell(\cdot,\cdot), \mathcal{D}, \mathrm{AES}(\dots), \boldsymbol{s}, \boldsymbol{\delta}, \varepsilon, \mathrm{M}(\cdot;\boldsymbol{w}), \textsc{Reg})$, $\mathrm{M}(\mathcal{S};\boldsymbol{w})$ is continuous and strictly monotonic in $\mathcal{S}$ with (possibly infinite) Lipschitz constant λ_{M} w.r.t. $\|\cdot\|_{\mathrm{M}}$, and the schedules $(\boldsymbol{s}, \boldsymbol{\delta})$ are $\frac{\varepsilon}{\lambda_{\mathrm{M}}(1+\mathbb{1}_{\textsc{Reg}})}$-uniformly-convergent w.r.t. $\|\cdot\|_{\mathrm{M}}$ and the *additive error vector bound* $\mathrm{AEV}(\boldsymbol{m}, \delta, \boldsymbol{x}, \boldsymbol{y}) \leftarrow \langle \mathrm{AES}(\boldsymbol{m}_1, \frac{\delta}{g}, \boldsymbol{x}_1, \boldsymbol{y}_1), \dots, \mathrm{AES}(\boldsymbol{m}_g, \frac{\delta}{g}, \boldsymbol{x}_g, \boldsymbol{y}_g)\rangle$. Now take μ to be the true objective value of $\hat{h}$ and $\mu^\star$ to be the true objective value of the optimal $h^\star$ (see theorem 4.5). Then, with probability at least $1-\delta$, we have

1) $|\hat{\mu} - \mu| \leq \hat{\varepsilon} \leq \varepsilon$; & 2) $\mathrm{M}^{\star\downarrow} \leq \mu^\star \leq \mu \leq \hat{\mu} + \hat{\varepsilon} \leq \mathrm{M}^{\star\downarrow} + 2\varepsilon$.

Proof. Proof of theorem 4.6 is quite similar to that of theorem 4.5. The primary difference is that the union bound is now over g individual schedules and all timesteps, i.e., we now have the total failure probability of all possible tail bounds (taken on line 17) is no greater than

$$\sum_{i=1}^{g}\sum_{t=1}^{\infty}\frac{\boldsymbol{\delta}_t}{g} = \frac{g}{g}\sum_{t=1}^{\infty}\boldsymbol{\delta}_t = \delta \ .$$

Note that while the *order* these bounds are taken is random, the braided structure takes no more than one bound for each $(i, \boldsymbol{t}_i)$ pair consisting of a *group* $i \in \mathcal{Z}$ and a *time index* $\boldsymbol{t}_i \in \mathbb{Z}_+$, and said bounds *do not depend* on the order in which they are taken. As the bounds themselves are always taken at the same sample sizes, we need only sum over failure probabilities of all possible tail bounds that may be taken by the algorithm to union-bound the total failure probability. As in the proof of the linear algorithm, we conclude that all tail bounds that could possibly be taken by the algorithm (now over all groups $i \in \mathcal{Z}$ and all timesteps $\boldsymbol{t}_i \in \mathbb{Z}_+$) hold simultaneously with probability at least $1-\delta$.

We now condition on the event that all bounds taken are correct (which holds with probability at least $1-\delta$). In this case, as in the linear algorithm, so long as the braided algorithm *terminates*, it produces a correct answer. The reasoning here is identical to that of theorem 4.5, as both algorithms share a termination condition.

Despite their similarity, it is not so straightforward as in the linear algorithm to show that the braided algorithm is *guaranteed to terminate* (i.e., it does not loop indefinitely). In particular, consider that while in algorithm 1, the appropriate uniformly-convergent schedule is sufficient to guarantee termination eventually, this would not be so in algorithm 2 *if it were to select groups arbitrarily* at each iteration, e.g., sampling from the same group at every iteration could loop indefinitely. However, this does not occur, due to the *group selection logic* on line 16. We consider first the 0-uniformly convergent schedule case, and then the general case, concluding eventual termination in both.

We first show that given a 0-uniformly convergent schedule, the algorithm will not loop indefinitely without sampling each group an infinite number of times, thus if the algorithm does not first terminate, any per-group time index vector $\boldsymbol{t}$ will eventually be exceeded (componentwise). To conclude this, it is sufficient to show that for each group i, the algorithm will either iterate until it either terminates, or group i is selected. This is because group i will eventually have the optimal improvement:cost ratio, and thus be selected to sample on line 16, which we now show. Consider that, by the *strict monotonicity* assumption, there is always nonzero (positive) projected improvement gain to improving the error bound $\hat{\boldsymbol{\varepsilon}}_i$, and thus to selecting any group i to sample. However, selecting other groups *ad nauseam* will take their cost terms to ∞, and thus their improvement:cost ratios to 0 (in the limit). It thus follows that, if the termination condition were never met, each group i would be sampled from an unbounded number of times, during which their error bound would converge as $\hat{\boldsymbol{\varepsilon}}_i \rightsquigarrow 0$, and thus $\hat{\boldsymbol{\varepsilon}} \rightsquigarrow \boldsymbol{0}$ by the *continuity* assumption. Consequently, it holds that $\hat{\mathrm{M}}^{\uparrow} \rightsquigarrow \mathrm{M}(h^\star)$ and $\mathrm{M}^{\star\downarrow} \rightsquigarrow \mathrm{M}(h^\star)$, which implies termination on line 12.

The case of an $\frac{\varepsilon}{\lambda_{\mathrm{M}}(1+\mathbb{1}_{\textsc{Reg}})}$-$\|\cdot\|_{\mathrm{M}}$-uniformly-convergent schedule follows similarly, except now groups may cease sampling if their error bounds are nonzero, but sufficiently small so as to ε-estimate the objective. If this holds for all groups (which again happens eventually if termination does not occur first), we again meet the termination condition of line 12, hence we conclude guaranteed termination in this case.

With termination shown in all cases, we now conclude that algorithm 2, produces a correct answer with the stated probability, by the same reasoning as in the analysis of algorithm 1, i.e., theorem 4.5. □

B A Traveller's Handbook to Progressive Sampling

In this appendix, we provide deeper intuition for our progressive sampling algorithms, with an emphasis on how they differ from standard progressive sampling approaches. We also describe how decisions are made as to which group to sample in the braided algorithm (algorithm 2), i.e., under the conditional sampling model. Note that this material is provided purely to supplement understanding of the methods; all proofs related to these algorithms are detailed in appendix A.2.

B.1 Tricks of the Trade: the Magician's Secrets, Revealed

Understanding progressive sampling algorithms essentially comes down to understanding *how union bounds are taken*, and *which tail bounds* may be taken and *when*, since correcting for the multiple comparisons problem is the central technical issue in such methods (e.g., an algorithm that computes bounds after every sample while using a union bound is extremely inefficient, because of the excessive union bound cost). For the most part, algorithms 1 and 2 are standard progressive sampling methods, however there are some subtle details that obscure the simple reasoning at their core. In particular, under the mixture sampling model, sample sizes are not always known *a priori*, and in the conditional sampling model, decisions are actually made dynamically that determine the order in which tail bounds are taken. Both of these decisions have ramifications that impact the core logic of a static alternating series of *sampling* and *bounding* steps, but as always, there is no real magic here: merely a few logical flourishes that permit these slight modifications to the standard flow of progressive sampling algorithms.

Under the joint sampling model, the linear progressive sampling algorithm is quite straightforward, as the total number of samples drawn b each timestep is completely known *a priori* (i.e., $\boldsymbol{s}_t$ from each group at timestep t), so a simple union bound over a deterministic schedule suffices. Under the mixture sampling model, there is some subtlety to why the algorithm works as it does, since the sample sizes at which tests are run are in fact a random variable dependent on the order in which groups are sampled. A naïve approach would be to consider a union bound over possible orders of sampling, or otherwise correct for the fact that multiple outcomes are possible, however this is ultimately not necessary, and such approaches would induce harmful corrective terms to probabilistic error-bounds and sample complexities. The simple trick we employ is to condition the algorithm on the infinite sequence of $\mathcal{Z}$ samples drawn on line 10, and establish the schedule *after performing this conditioning operation*, thus samples drawn under the mixture sampling model are in fact samples from individual groups in a *known order*. Of course, we don't "really" know this order, nor do we perform this conditioning anywhere in the algorithm; this is an analytical technique, and the simple fact of its existence suffices to show correctness. This analysis does not actually require a randomized order of $\mathcal{Z}$ samples, and in fact the algorithm works even if group identities are *adversarially selected*, as nowhere in the analysis do we actually assume sampled group identities (line 10) are *random*. However, it must be true under adversarial $\mathcal{Z}$ selection that each $(\mathcal{X}, \mathcal{Y})$ pair drawn is *conditionally independently* given each group in the set $\mathcal{Z}$, thus the adversarial mixture sampling analysis is most useful in the context of mutually exclusive groups (i.e., singleton $\boldsymbol{z}$ samples).

Because of these design decisions, all three algorithms are able to use simple progressive sampling parameters, such as a single sampling schedule $\boldsymbol{s}$ and failure probability schedule $\boldsymbol{\delta}$, albeit in slightly different ways. They also enjoy the characteristic sharpness of standard progressive sampling algorithms, with only an extra $\frac{1}{g}$ factor attached to failure probabilities (δ values) for the conditional sampling model, to accommodate the union bound over g simultaneous (braided) linear progressive sampling instances.

B.2 Selecting Where to Sample in the Braided Algorithm

We now discuss how the decision as to which group to sample from is made. Our reasoning here largely parallels that of property 3.8: we want to maximize the improvement made to the UCB-optimal $\hat{h}$. However, rather than apply the linear subderivative approximation, which is accurate for smallchanges (e.g., adding a single sample), we consider the impact of advancing the sampling schedule of each group by one timestep; for geometric schedules, this is a *multiplicative* — rather than an additive — change to the sample size. We must also include the cost of sampling when selecting where to sample, so this too enters the equation through the linear cost model $\boldsymbol{C}_{1:g}$. Intuitively, the idea is essentially to select the group i for which the *ratio of projected improvement* to *cost* is maximized.

Unfortunately, there is a slight wrinkle in the algorithm: at each iteration, we can not simply greedily maximize the improvement:cost ratio from a single timestep of sampling group i, because this leads into a failure mode, wherein if a group's risk is near c, i.e., near maximal, it will never be sampled. There is no simple workaround, as the malfare function may not even be defined for inputs larger than c (for instance if $\mathrm{M}(\mathcal{S}; \boldsymbol{w})$ is actually the function $c - \mathrm{W}_0(c - \mathcal{S}; \boldsymbol{w})$, i.e., geometric welfare through the reduction of the LinearPSUtility(...) procedure of algorithm 1). Our solution is to maximize not just the improvement:cost ratio of sampling for a *single* additional timestep, but over *any number* t of timesteps. For reasonable malfare functions, when the empirical risk is unchanged after taking the minimum with c (i.e., the expression $c \wedge \hat{\mathbb{E}}_{\boldsymbol{x}_{i,:},\boldsymbol{y}_{i,:}}[\ell \circ h]$ takes the value $\hat{\mathbb{E}}_{\boldsymbol{x}_{i,:},\boldsymbol{y}_{i,:}}[\ell \circ h]$ on line 16 of algorithm 1), we expect diminishing returns (in improvement:cost ratio) to further sampling, thus the $t = 1$ step greedy optimal choice should usually be selected.

In particular, given per-group uniform convergence bounds $\hat{\boldsymbol{\varepsilon}}_{1:g}$ and assuming each group is on timestep $\boldsymbol{t}_{1:g}$ of their sampling schedule, we estimate the UCB improvement[11] for sampling from group i for t timesteps as as

$$\tilde{\boldsymbol{\varepsilon}}^{(i)}_{j,t} \leftarrow \hat{\boldsymbol{\varepsilon}}_j \quad \text{for } i \neq j \ , \quad \& \quad \tilde{\boldsymbol{\varepsilon}}^{(i)}_{j,t} \leftarrow \hat{\boldsymbol{\varepsilon}}_j \sqrt{\frac{\boldsymbol{s}_{\boldsymbol{t}_j} \ln \frac{g}{\boldsymbol{\delta}_{\boldsymbol{t}_j}}}{t + \boldsymbol{s}_{\boldsymbol{t}_j} \ln \frac{g}{\boldsymbol{\delta}_{t+\boldsymbol{t}_j}}}} \quad \text{for } i = j \tag{23}$$

(see line 7). We then estimate the malfare improvement as $\Delta_{i,t} \leftarrow \hat{\mathrm{M}}^{\uparrow} - \mathrm{M}\left(j \mapsto c \wedge \hat{\mathbb{E}}_{\boldsymbol{x}_{i,:},\boldsymbol{y}_{i,:}}[\ell \circ h] + \tilde{\boldsymbol{\varepsilon}}^{(i)}_{j,t}; \boldsymbol{w}\right)$, and compute the cost of sampling from group i for t timesteps as $\boldsymbol{C}_i(\boldsymbol{s}_{t+\boldsymbol{t}_i} - \boldsymbol{s}_{\boldsymbol{t}_i})$. We then select the group with maximal projected improvement to sampling cost ratio, i.e., we select $i \leftarrow \operatorname{argmax}_{i \in \mathcal{Z}} \sup_{t \in \mathbb{Z}_+} \dfrac{\Delta_{i,t}}{\boldsymbol{C}_i(\boldsymbol{s}_{t+\boldsymbol{t}_i} - \boldsymbol{s}_{\boldsymbol{t}_i})}$ (see line 16).

Note that even if the estimated improvement is not accurate, we still gain information; either the upper bound decreases more than expected, which decreases the relative value of further sampling from this group (in which case we are less likely to sample from this group in the future), or it decreased less than expected (likely due to selection bias), in which case we may decide to sample it again, but now more work is required to get the same reduction in confidence radius (increased cost), so sampling from another group may now be optimal.

Further Notes on Long-Term Planning Optimizing over timestep count t, as well as group i, does seem like it may create some issues for the algorithm, however the impact is philosophically, computationally, and practically very small. In general, as long as some improvement to sampling a group is projected to happen *eventually*, evaluating the supremum over t exhaustively will eventually reach some t such that the cost is so large that no larger t can be optimal. For example, if the current (regret) malfare UCB is $\hat{\mathrm{M}}^{\uparrow}$, and improvement $\Delta_{i,t} > 0$ is projected after some timestep count t, then for cutoff timestep $t^{\uparrow}$, defined as the smallest integer $t^{\uparrow} > t$ such that

$$\frac{\hat{\mathrm{M}}^{\uparrow}}{\boldsymbol{C}_i(\boldsymbol{s}_{t^{\uparrow}+\boldsymbol{t}_i} - \boldsymbol{s}_{\boldsymbol{t}_i})} \leq \frac{\Delta_{i,t}}{\boldsymbol{C}_i(\boldsymbol{s}_{t+\boldsymbol{t}_i} - \boldsymbol{s}_{\boldsymbol{t}_i})} \ , \quad \text{or equivalently,} \quad \boldsymbol{s}_{t^{\uparrow}+\boldsymbol{t}_i} \geq \boldsymbol{s}_{\boldsymbol{t}_i} + (\boldsymbol{s}_{t+\boldsymbol{t}_i} - \boldsymbol{s}_{\boldsymbol{t}_i}) \frac{\hat{\mathrm{M}}^{\uparrow}}{\Delta_{i,t}} \ , \tag{24}$$

the improvement:cost ratio is never maximized after more than $t^{\uparrow}$ additional timesteps. We thus conclude evaluating the supremum over t in group selection (line 16) is not computationally intractable.

Note also that even if sampling group i for t timesteps yields an optimal improvement:cost ratio at this iteration, the algorithm may change course on the next iteration. For example, if more improvement than projected occurs, then sampling from another group j, which now has a greater impact on the malfare (e.g., group i is no longer maximal, and thus inconsequential to egalitarian malfare) now optimizes the improvement:cost ratio, or if less improvement occurs than projected, then projected improvement of group i at the next timestep may decrease, after which another group may have the optimal projected improvement:cost ratio. In other words, despite some element of *long-term planning* with this supremum over t in group selection, the algorithm does not commit to sampling from a group for more than a single iteration, which is important, because long-term projections are likely to be inaccurate, as they are made with less information than the greedy-optimal group selection at each iteration.

[11] Note that we can't simply assume $\sqrt{\frac{1}{m}}$ rates, as δ may also be changing. Even incorporating the $\ln \frac{1}{\delta}$ terms characteristic of exponential tail bounds is not always entirely accurate; and when the schedule and bound are fully specified, in some cases it may be possible to produce a better estimate of the bound improvement. In particular, if AES(...) is not data-dependent, we can simply take $\tilde{\boldsymbol{\varepsilon}}^{(i)}_{i,t} \leftarrow \mathrm{AES}(\boldsymbol{s}_{t+\boldsymbol{t}_i}, \frac{\boldsymbol{\delta}_{t+\boldsymbol{t}_i}}{g})$.