7 Acknowledgements

This work was partially supported by ARO under award number W911NF-19-1-0317. We also thank the four anonymous reviewers and the area-chair for their detailed feedback that helped improve the presentation of our results.

References

Abernethy, J., Awasthi, P., Kleindessner, M., Morgenstern, J., and Zhang, J. (2020). Adaptive sampling to reduce disparate performance. *arXiv:2006.06879*.

Agarwal, A., Beygelzimer, A., Dudík, M., Langford, J., and Wallach, H. (2018). A reductions approach to fair classification. In *Proceedings of the 35th International Conference on Machine Learning*, pages 60–69.

Anahideh, H., Asudeh, A., and Thirumuruganathan, S. (2020). Fair active learning. *arXiv:2001.01796*.

Audibert, J.-Y. and Tsybakov, A. B. (2007). Fast learning rates for plug-in classifiers. *The Annals of statistics*, 35(2):608–633.

Auer, P., Cesa-Bianchi, N., and Fischer, P. (2002). Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47:235–256.

Barocas, S. and Selbst, A. (2016). Big data's disparate impact. *California Law Review*, 104.

Caton, S. and Haas, C. (2020). Fairness in machine learning: A survey. *arXiv:2010.04053*.

Celis, L. E., Keswani, V., and Vishnoi, N. (2020). Data preprocessing to mitigate bias: A maximum entropy based approach. In *International Conference on Machine Learning*, pages 1349–1359.

Chen, I., Johansson, F., and Sontag, D. (2018). Why is my classifier discriminatory? In *Advances in Neural Information Processing Systems*.

Diana, E., Gill, W., Kearns, M., Kenthapadi, K., and Roth, A. (2021). Minimax group fairness: Algorithms and experiments. *arXiv:2011.03108v2*.

Dua, D. and Graff, C. (2017). UCI machine learning repository.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. (2012). Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226.

Feldman, M., Friedler, S., Moeller, J., Scheidegger, C., and Venkatasubramanian, S. (2015). Certifying and removing disparate impact. In *proceedings of the 21th international conference on knowledge discovery and data mining*, pages 259–268.

Geoffrion, A. (1968). Proper efficiency and the theory of vector maximization. *Journal of mathematical analysis and applications*, 22(3):618–630.

Hardt, M., Price, E., and Srebro, N. (2016). Equality of opportunity in supervised learning. In *Advances in neural information processing systems*, pages 3315–3323.

Holstein, K., Vaughan, J. W., III, H. D., Dudík, M., and Wallach, H. M. (2018). Improving fairness in machine learning systems: What do industry practitioners need? *CoRR*, abs/1812.05239.

Hooker, S. (2021). Moving beyond "algorithmic bias is a data problem". *Patterns*, 2(4):100241.

Jo, E. S. and Gebru, T. (2019). Lessons from archives: Strategies for collecting sociocultural data in machine learning. *CoRR*, abs/1912.10389.

Kendall, M. G. (1948). *Rank correlation methods.* Griffin.

Krizhevsky, A. and Hinton, G. (2009). Learning multiple layers of features from tiny images. Technical Report 0, University of Toronto, Toronto, Ontario.

Lattimore, T. and Szepesvári, C. (2020). *Bandit algorithms*. Cambridge University Press.

Lipton, Z., Chouldechova, A., and McAuley, J. (2017). Does mitigating ML's impact disparity require treatment disparity? *arXiv:1711.07076*.

Mann, H. B. (1945). Nonparametric tests against trend. *Econometrica*, 13(3):245–259.

Martinez, N., Bertran, M., and Sapiro, G. (2020). Minimax pareto fairness: A multi objective perspective. In *International Conference on Machine Learning*, pages 6755–6764.

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., and Gebru, T. (2018). Model cards for model reporting. *CoRR*, abs/1810.03993.

Polyanskiy, Y. and Wu, Y. (2015). Lecture notes on information theory.

Srinivas, N., Krause, A., Kakade, S. M., and Seeger, M. W. (2009). Gaussian process bandits without regret: An experimental design approach. *CoRR*, abs/0912.3995.

Suresh, H. and Guttag, J. V. (2019). A framework for understanding unintended consequences of machine learning. *CoRR*, abs/1901.10002.

Ustun, B., Liu, Y., and Parkes, D. (2019). Fairness without harm: Decoupled classifiers with preference guarantees. In *Proceedings of the 36th International Conference on Machine Learning*, pages 6373–6382.

Woodworth, B., Gunasekar, S., Ohannessian, M., and Srebro, N. (2017). Learning non-discriminatory predictors. In *Proceedings of the 2017 Conference on Learning Theory*, pages 1920–1953.

Xiao, H., Rasul, K., and Vollgraf, R. (2017). Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *CoRR*, abs/1708.07747.

Zemel, R., Wu, Y., Swersky, K., Pitassi, T., and Dwork, C. (2013). Learning fair representations. In *International Conference on Machine Learning*, pages 325–333.

Zhang, Song, and Qi (2017). Age progression/regression by conditional adversarial autoencoder. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE.

A Proof of Proposition 1

Proof. First we note that there must exist at least one optimum mixture distribution π^* since by the Assumption 1, the objective function is continuous and the domain of optimization, Δ_m, is compact. We also note that any optimal π^* must lie in the interior of the simplex Δ_m as a consequence of Assumption 2 (this follows by contradiction).

Next, we show that any such π^* must equalize the $L(z, f_{\pi^*} \coloneqq \mathbb{E}_z\left[\ell\left(f_{\pi^*}, X, Y\right)\right]$ for all $z \in \mathcal{Z}$. Indeed, assume that this is not the case and enumerate the elements of $\mathcal{Z}$ as $z_1, \ldots, z_m$ in decreasing order of $L(z, f_{\pi^*})$ value. Also introduce ϱ to denote $\min_{z \neq z_1} L(z_1, f_{\pi^*}) - L(z, f_{\pi^*})$ and (for now) assume that $\varrho > 0$.

Now define π_ϵ as follows: $\pi_\epsilon(z) = \pi^*(z)$ for $z \in \mathcal{Z}' \coloneqq \mathcal{Z} \setminus \{z_1, z_m\}$, $\pi_\epsilon(z_1) = \pi^*(z_1) + \epsilon$ and $\pi_\epsilon(z_m) = \pi^*(z_m) - \epsilon$. Due to the continuity assumption on the mapping $\pi \mapsto L(z, f_\pi)$ for all $z \in \mathcal{Z}$ in Assumption 1, we note that there exists $\epsilon_0 < \pi_{\min} \coloneqq \min_{z \in \mathcal{Z}} \pi^*(z)$, such that for all $z \neq z_1$ we have $L(z, f_{\pi_\epsilon}) \leq L(z_1, f_{\pi^*}) - \varrho/2$ for all $\epsilon \leq \epsilon_0$. Note that $\pi_{\min}$ must be strictly greater than 0, since π^* lies in the interior of Δ_m. Finally, due to the monotonicity assumption in Assumption 1, there must exist $\varrho_1 > 0$ such that $L(z_1, f_{\pi_{\epsilon_0}}) = L(z_1, f_{\pi^*}) - \varrho_1 < L(z_1, f_{\pi^*})$. Together, these results imply that

$$\max_{z \in \mathcal{Z}} L(z, f_{\pi_{\epsilon_0}}) \leq \max\{L(z_1, f_{\pi^*}) - \varrho/2,\ L(z_1, f_{\pi^*}) - \varrho_1\} < L(z_1, f_{\pi^*}),$$

thus contradicting the optimality assumption on π^*. This completes the proof of the statement that any optimal π^* must ensure that $L(z, f_{\pi^*}) = L(z', f_{\pi^*})$ for all $z, z' \in \mathcal{Z}$.

Finally, the fact that $L(z, f_{\pi^*}) = L(z', f_{\pi^*})$ for any optimal π^* also ensures the uniqueness of the optimum mixture distribution. Again, assume that this is not the case and there exists another optimal $\tilde{\pi}^* \neq \pi^*$ achieving optimum value in (1). Then there must exist a z such that $\pi(z) \neq \pi^*(z)$, and hence $L(z, f_{\pi^*}) \neq L(z, f_{\tilde{\pi}^*})$. Hence, by Assumption 1, we have $\max_{z \in \mathcal{Z}} L(z, f_{\pi^*}) \neq \max_{z \in \mathcal{Z}} L(z, f_{\tilde{\pi}^*})$, which contradicts our hypothesis that both π^* and $\tilde{\pi}^*$ achieve the optimum value in (2). This concludes the proof of the uniqueness of π^*. □

B Details of Algorithm 2

The detailed pseudo-code of $\mathcal{A}_{\texttt{opt}}$ is given in Algorithm 2.

B.1 Proof of Theorem 1

Recall the UCB first defined in Equation 4:

$$U_t(z, \hat{f}_t) \coloneqq \frac{1}{|\mathcal{D}_z|} \sum_{(x,y) \in \mathcal{D}_z} \ell(\hat{f}_t, x, y) \ + \ e_z\left(N_{z,t}\right) + \frac{2C}{\pi_t(z)} \underbrace{\sum_{z' \in \mathcal{Z}} \pi_t(z') e_{z'}(N_{z',t})}_{\coloneqq \rho_t}. \tag{7}$$

With the parameter C as defined in Assumption 3 and sequences $e_z(N)$ satisfying Assumption 4.

Before proceeding further, we introduce the following notation:

- For any $z \in \mathcal{Z}$ and at any time $t \leq n$, we use $N_{z,t}$ to refer to the number of samples drawn corresponding to the feature z prior to time t, i.e., $N_{z,t} = \sum_{i=1}^{t-1} \mathbb{1}_{\{z_i = z\}}$.
- $\widehat{L}_t(z, f) = \frac{1}{N_{z,t}} \sum_{(x,y) \in \mathcal{D}_{(z)}} \ell(f, x, y)$, denotes the empirical risk of a classifier f on the samples corresponding to attribute z,
- $L(z, f) = \mathbb{E}_z\left[\ell(f, X, Y)\right]$ denotes the population risk of classifier f corresponding to attribute z.
- $\rho_t = \sum_{z \in \mathcal{Z}} \pi_t(z) e_{z, N_{z,t}}$ will be used to quantify the uniform deviation of the classifier $\hat{f}_t$ from the corresponding optimal classifier f_{π_t}. Recall that π_t is the mixture distribution of the $|\mathcal{Z}|$ attributes at time t constructed by the algorithm, i.e., $\pi_t(z) = \frac{N_{z,t}}{t}$ for $z \in \mathcal{Z}$.

Algorithm 2: Optimistic Sampling for Fair Classification ($\mathcal{A}_{\texttt{opt}}$)

Input: n (budget), $\mathcal{F}$ (function class), ℓ (loss function), $\xi \in (0,1)$ (forced exploration term)

Initialize: $\mathcal{D}_0 = \emptyset$; $e_{z,1} = +\infty$ and $\mathcal{D}_{(z)} = \emptyset$, for all $z \in \mathcal{Z}$;

/* Draw two independent samples from each $z \in \mathcal{Z}$ */

for $t = 1, \ldots, m$ **do**

  $z \leftarrow z_t \in \mathcal{Z}$;

  $\left(X_t^{(i)}, Y_t^{(i)}\right)_{i=1,2} \sim P_z$; $\quad \mathcal{D}_{(z)} = \{(X_t^{(1)}, Y_t^{(1)})\}$; $\quad \mathcal{D}_t \leftarrow \mathcal{D}_{t-1} \cup \{(X_t^{(2)}, Y_t^{(2)})\}$;

end

$\pi_t \leftarrow \left(\frac{1}{m}, \ldots, \frac{1}{m}\right)$; $\quad \hat{f}_t \in \arg\min_{f \in \mathcal{F}} \frac{1}{t} \sum_{(x,y) \in \mathcal{D}_t} \ell(f, x, y)$; $\quad N_{z,t} \leftarrow 1, \ \forall z \in \mathcal{Z}$;

for $t = m+1, \ldots, n$ **do**

  /* choose the next distribution $P_{XY|Z=z}$ to sample */

  if $\min_{z \in \mathcal{Z}} N_{z,t} < t^{\xi}$ **then**

    $z_t \in \arg\min_{z \in \mathcal{Z}} N_{z,t}$ // Forced exploration

  else

    $z_t = \arg\max_{z \in \mathcal{Z}} U_t(z, \hat{f}_t)$

  end

  /* Collect data */

  $\left(X_t^{(i)}, Y_t^{(i)}\right)_{i=1,2} \sim P_{z_t}$

  /* Perform the updates */

  Update $\left(e_z(N_{z,t})\right)_{z \in \mathcal{Z}}$,

  $\mathcal{D}_t \leftarrow \mathcal{D}_{t-1} \cup \{(X_t^{(1)}, Y_t^{(1)})\}$, $\quad \mathcal{D}_{(z_t)} \leftarrow \mathcal{D}_{(z_t)} \cup \{(X_t^{(2)}, Y_t^{(2)}\}$.

  Update $\pi_t \leftarrow \frac{t-1}{t}\pi_t + \frac{1}{t}\mathbb{1}_{\{z_t\}}$

  /* Update the classifier $\hat{f}_t$ */

  $\hat{f}_t \in \arg\min_{f \in \mathcal{F}} \frac{1}{t} \sum_{(x,y) \in \mathcal{D}_t} \ell(f, x, y)$

end

Output: π_n

Suppose at the end of n rounds, the resulting mixture distribution is π_n. Then introduce the set of *over-represented* attributes, $\mathcal{Z}_o \subset \mathcal{Z}$, defined as $\mathcal{Z}_o := \{z \in \mathcal{Z} : \pi_n(z) > \pi^*(z)\}$ and refer to the set $\mathcal{Z}_u := \mathcal{Z} \setminus \mathcal{Z}_o$ as the *under-represented* attributes. Note that $\mathcal{Z}_o$ is empty only if $\pi_n = \pi^*$. In this case, the sampling algorithm has learned the optimal mixing distribution resulting in zero excess risk. Hence, for the rest of the proof, we will assume that $\mathcal{Z}_o$ is non-empty. Note that for any $z_0 \in \mathcal{Z}_o$, Assumption 1 ensures that we must have $L(z_0, f_{\pi_n}) < L(z_0, f_{\pi^*})$ since $\pi_n(z_0) > \pi^*(z_0)$.

Lemma 1. *Suppose $t_0 \leq n$ denotes the last time that Algorithm 2 queried any feature belonging to the subset $\mathcal{Z}_o$, and denote the corresponding feature by z_0. Then, if n is large enough to ensure that $n^{1-\xi} \geq \left(\frac{1}{\pi_{\min}}\right)$ where $\pi_{\min} = \min_{z \in \mathcal{Z}} \pi^*(z)$ and that $2 \max_{z \in \mathcal{Z}} e_z\left((\pi_{\min} n)^{\xi}\right) \leq \epsilon_0$ (introduced in Assumption 3), we have the following with probability at least $1 - \delta$:*

$$L(z, f_{\pi_{t_0}}) \leq \underbrace{L(z, f_{\pi^*})}_{:=M^*} + \underbrace{2e_{z_0}(N_{z_0,t_0}) + \left(4C/\pi^*(z_0)\right)\rho_{t_0}}_{:=B_0}. \tag{8}$$

Recall that $\rho_t = \sum_z \pi_t(z) e_z(N_{z,t})$, C is the constant from Assumption 3, and π^* is the optimal mixture distribution defined in (2)*.*

Proof. Throughout this proof we assume that the $1 - \delta/2$ probability event introduced while defining the UCB term in (4) occurs for the two sequence of samples drawn by Algorithm 2: the first used in updating $\mathcal{D}_t$ for training the classifier $\hat{f}_t$, and the second used in updating $(\mathcal{D}_z)_{z \in \mathcal{Z}}$ for estimating the loss of $\hat{f}_t$. Thus, both of these events occur with probability at least $1 - \delta$.

First note the following chain of inequalities: $\pi_{t_0}(z_0) \geq \pi_n(z_0) > \pi^*(z_0) \geq \pi_{\min}$. Then at time t_0, it must be the case that $N_{z_0,t_0} \geq \pi_{\min} n$ as z_0 belongs to $\mathcal{Z}_o$. This, along with the fact that

$n \geq (1/\pi_{\min})^{1/(1-\xi)}$ means that $N_{z_0,t_0} \geq \pi_{\min} n \geq n^{-1+\xi} n = n^{\xi}$, and hence the query to z_0 must have been made due to the condition in Line 11 of Algorithm 2, and not due to the forced exploration step in Line 9.

Next, we have the following chain of inequalities for any $z \neq z_0$:

$$\begin{aligned}
L(z, f_{\pi_{t_0}}) &\overset{\text{(a)}}{\leq} L(z, \hat{f}_{t_0}) + |L(z, \hat{f}_{t_0}) - L(z, f_{\pi_{t_0}})| \\
&\overset{\text{(b)}}{\leq} L(z, \hat{f}_{t_0}) + \left(2C/\pi_{t_0(z)}\right) \rho_{t_0} \\
&\overset{\text{(c)}}{\leq} \widehat{L}_{t_0}(z, \hat{f}_{t_0}) + e_z(N_{z,t}) + \left(2C/\pi_{t_0(z)}\right) \rho_{t_0} \\
&\overset{\text{(d)}}{\leq} \widehat{L}_{t_0}(z_0, \hat{f}_{t_0}) + e_{z_0}(N_{z_0,t}) + \left(2C/\pi_{t_0}(z_0)\right) \rho_{t_0} \\
&\overset{\text{(e)}}{\leq} L(z_0, \hat{f}_{t_0}) + 2e_{z_0}(N_{z_0,t_0}) + \left(2C/\pi_{t_0}(z_0)\right) \rho_{t_0} \\
&\overset{\text{(f)}}{\leq} L(z_0, f_{\pi_{t_0}}) + 2e_{z_0}(N_{z_0,t_0}) + \left(4C/\pi_{t_0}(z_0)\right) \rho_{t_0} \\
&\overset{\text{(g)}}{\leq} L(z_0, f_{\pi^*}) + 2e_{z_0}(N_{z_0,t_0}) + \left(4C/\pi^*(z_0)\right) \rho_{t_0} \\
&\overset{\text{(h)}}{=} L(z, f_{\pi^*}) \ + \ 2e_{z_0}(N_{z_0,t_0}) + \left(4C/\pi^*(z_0)\right) \rho_{t_0} \\
&= M^* + B_0.
\end{aligned}$$

In the above display:
(a) follows from an application of triangle inequality,
(b) applies Assumption 3 and uses the fact that n is large enough to ensure that the suboptimality of $\hat{f}_{t_0}$ w.r.t. $f_{\pi_{t_0}}$ is no larger than ϵ_0.
(c) follows from the definition of event Ω_2,
(d) follows from the attribute selection rule in Line 11,
(e) again uses the uniform deviation event Ω_2,
(f) follows from another application of Assumption 3,
(g) uses the fact that $\pi_{t_0}(z_0) > \pi^*(z_0)$, the monotonicity condition from Assumption 1,
(h) uses the result of Proposition 1 to write $L(z_0, f_{\pi^*}) = L(z, f_{\pi^*}) \coloneqq M^*$. □

Next, we show that for a large enough value of n, due to the forced exploration step (Line 9 in Algorithm 2), the values of $\pi_{t_0}(z)$ for $z \in \mathcal{Z}_u$ are not too small.

Lemma 2. *For any given $A < \pi_{\min} = \min_{z \in \mathcal{Z}} \pi^*(z)$, there exists an $n_0 < \infty$, defined in* (12) *below, such that for all $n \geq n_0$, the following are true at time t_0 (recall that t_0 denotes the last time at which an attribute from $\mathcal{Z}_o$ was queried by the algorithm):*

$$\pi_{t_0}(z) \geq \frac{(m-1)\pi^*(z)}{m} + \frac{A}{m}, \qquad \textit{for all } \ z \in \mathcal{Z}_u, \tag{9}$$

$$\textit{and} \quad t_0 \geq n \left(\frac{\pi_{\min}}{2\pi_{\min} - A} \right). \tag{10}$$

Recall that in the above display $m = |\mathcal{Z}|$.

Proof. First note that for any $z \in \mathcal{Z}_u$, we have

$$M^* \leq L(z, f_{\pi_{t_0}}) \leq M^* + B_0,$$

where we have used the notation $M^* = L(z, f_{\pi^*})$ introduced in (8). The left inequality above is due to the monontonicity condition of Assumption 1, while the right inequality is from Lemma 1.

Introducing the notation $\pi_{\min} = \min_{z \in \mathcal{Z}} \pi^*(z)$, we note that by definition $t_0 \geq \pi_{\min} n$. Recall that the term B_0 is defined as $B_0 = 2e_{z_0}(N_{z_0,t_0}) + \left(4C/\pi^*(z_0)\right) \rho_{t_0}$. Now, since $N_{z_0,t_0} \geq \pi_{\min} n$ and due to the monotonicity of $e_z(N_{z,t})$, we can upper bound $e_{z_0}(N_{z_0,t_0})$ with $e_{z_0}(\pi_{\min} n)$. Next, recall that $\rho_{t_0} = \sum_{z \in \mathcal{Z}} \pi_{t_0}(z) e_z(N_{z,t_0}) \leq \max_{z \in \mathcal{Z}} e_z(N_{z,t_0})$. Since $t_0 \geq N_{z_0,t_0} \geq \pi_{\min} n$, and due to the fact that forced-exploration step, i.e., Line 9 of Algorithm 2, was not needed at time t_0, we must have $N_{z,t_0} \geq (\pi_{\min} n)^{\xi}$ for all $z \in \mathcal{Z}_u$. Thus the second term in the definition of B_0 can be simply

bounded with ${}^{4C}/_{\pi^*(z_0)} \max_{z\in\mathcal{Z}} e_z((\pi_{\min} n)^\xi) \le {}^{4C}/_{\pi_{\min}} \max_{z\in\mathcal{Z}} e_z((\pi_{\min} n)^\xi)$. Combining these two steps, we finally get that $B_0 \le ({}^{4C}/_{\pi_{\min}} + 2) \max_{z\in\mathcal{Z}} e_z((\pi_{\min} n)^\xi)$.

By the monotonicity of the terms $e_z(N_{z,t})$ and ρ_t, we note that $\lim_{n\to\infty} B_0 = 0$, since $\lim_{n\to\infty} \max_{z\in\mathcal{Z}} e_z((\pi_{\min} n)^\xi) = 0$. Thus as n goes to infinity, $L(z, f_{\pi_{t_0}})$ converges to the optimal value M^*, which by continuity of the mapping $\pi \mapsto L(z, f_\pi)$ for all $z \in \mathcal{Z}$ implies that $\pi_{t_0} \to \pi^*$. We can use this fact to define a sufficient number of samples, denoted by n_0, beyond which it can be ensured that π_{t_0} satisfies the statement in (9).

$$\pi_{\min} = \min_{z\in\mathcal{Z}} \pi^*(z); \qquad b = \inf\left\{ \max_{z\in\mathcal{Z}} L(z, f_\pi) - M^* : \|\pi^* - \pi\|_\infty > \frac{\pi_{\min} - A}{m} \right\}; \tag{11}$$

$$n_0 \coloneqq \max\left\{ n_0', \frac{1}{\pi_{\min}^2} \right\}; \qquad n_0' \coloneqq \min\left\{ n \ge 1 : \left(\frac{4C}{\pi_{\min}} + 2\right) \max_{z\in\mathcal{Z}} e_z((\pi_{\min} n)^\xi) \le b \right\}. \tag{12}$$

Thus the definition of the term b, combined with the upper bound on B_0 due to Lemma 1 and the forced-exploration rule ensure that for $n \ge n_0$, we must have $\pi_{t_0}(z) \ge \pi^*(z)/2$ for all $z \in \mathcal{Z}_u$ as required by (9).

Since we will use the above computation of n_0' several times, we formalize it in terms of the following definition.

Definition 2 (`SmallestBudget`)**.** Given constants $c > 0$, p, q and $r \in (0, 1]$, the function `SmallestBudget` returns the following

$$\texttt{SmallestBudget}(c, p, q, r) = \min\left\{ n \ge 1 : \max_z e_z(N_{pq}) \le \gamma/c \right\}, \quad \text{where}$$
$$\gamma = \inf\left\{ \max_{z\in\mathcal{Z}} L(z, f_\pi) - M^* : \|\pi^* - \pi\|_\infty > r \right\} \quad \text{and} \quad N_{pq} = (pn)^q.$$

Note that n_0' in (12) is equal to $\texttt{SmallestBudget}\,((4C/\pi_{\min} + 2),\, \pi_{\min},\, \xi,\, (\pi_{\min} - A)/m)$.

For the proof of the statement in (10), we note that since $\pi_{t_0}(z) \ge \pi^*(z) - \frac{\pi_{\min} - A}{m}$ for all $z \ne z_0$, the time t_0 must satisfy the following:

$$\begin{aligned} t_0 &= N_{z_0,t_0} + \sum_{z\ne z_0} N_{z,t_0} \ge \pi^*(z_0) n + \sum_{z\ne z_0} \left(\pi^*(z) - \frac{\pi_{\min} - A}{m}\right) t_0 \\ &= \pi^*(z_0) n + \left(1 - \pi^*(z_0) + \frac{(m-1)(\pi_{\min} - A)}{m}\right) t_0. \end{aligned}$$

This, in turn, implies that

$$t_0 \ge n \left(\frac{\pi^*(z_0)}{\pi^*(z_0) + \frac{m-1}{m}(\pi_{\min} - A)}\right) \ge n \left(\frac{\pi_{\min}}{2\pi_{\min} - A}\right).$$

This completes the proof of (10). □

We now state a basic result about the behavior of π_t:

Lemma 3. *Suppose the empirical mixture distribution is π_r at some time r, and in the time interval $\{r+1, \ldots, t\}$ the agent only queries attributes z from $\mathcal{Z}' \subset \mathcal{Z}$ such that $\sum_{z\in\mathcal{Z}'} \pi_r(z) < 1$. Then there exists at least one $z \in \mathcal{Z}'$ such that $\pi_t(z) > \pi_r(z)$.*

Proof. The statement follows by contradiction. Assume that the conclusion stated above is not true, and $\pi_t(z) \le \pi_r(z)$ for all $z \in \mathcal{Z}'$. Introducing b_z to denote the number of times attribute $z \in \mathcal{Z}'$ is queried in the time interval $\{r+1, \ldots, t\}$, we then have

$$\pi_t(z) = \frac{r\pi_r(z) + b_z}{t} \;\Rightarrow\; (t-r)\pi_t(z) + r\underbrace{(\pi_t(z) - \pi_r(z))}_{\le 0} = b_z \;\Rightarrow\; (t-r)\pi_t(z) \ge b_z. \tag{13}$$

Note that $\sum_{z\in\mathcal{Z}'}\pi_t(z) = \frac{t-r(1-\sum_z \pi_r(z))}{t} < 1$ by assumption that $\sum_{z\in\mathcal{Z}'}\pi_r(z) < 1$. Combining this with (13), we get the required contradiction as follows:

$$(t-r) > (t-r)\sum_{z\in\mathcal{Z}'}\pi_t(z) \geq \sum_{z\in\mathcal{Z}'} b_z = (t-r).$$

□

Before proceeding, we introduce some notations. As stated earlier, due to the definition of the term t_0, we know that in the rounds $t \in \{t_0+1,\dots,n\}$, the algorithm only queries the attributes belonging to the set $\mathcal{Z}_u$. If the set $\mathcal{Z}_u$ is empty, that means t_0 must be equal to n and the algorithm stops there. Otherwise, the interval $\{t_0+1,\dots,n\}$ can be partitioned into $\{t_0+1,\dots,t_1\}$, $\{t_1+1,\dots,t_2\}$, $\dots$, $\{t_s+1,\dots,n\}$ for appropriately defined $t_1,\dots,t_s$ and $s \leq |\mathcal{Z}|-1$ as follows.

- First we introduce the term $\mathcal{Z}_t$ to denote the 'active set' of attributes at time t, i.e., the set of attributes that are queried at least once after time t. Note that we have $\mathcal{Z}_{t_0} = \mathcal{Z}_u$.
- Then (for $t \geq t_0$) we define a subset $\mathcal{Z}_o^{(1)}$ of $\mathcal{Z}_t$ as those attributes $z \in \mathcal{Z}_t$ for which we have $\pi_{t_0}(z) < \pi_n(z)$. By Lemma 3, we know that $\mathcal{Z}_o^{(1)}$ must be non-empty.
- Next, we define t_1 as the last time $t \leq n$ at which an attribute $z \in \mathcal{Z}_o^{(1)}$ is queried by the algorithm.
- If $t_1 = n$, then we stop and $s = 1$. Otherwise, we repeat the previous two steps with $\mathcal{Z}_{t_1} = \mathcal{Z}_{t_0} \setminus \mathcal{Z}_o^{(1)}$.

To clarify the above introduced notation, we present an example.

Example 2. *Consider a problem with set of attributes $\mathcal{Z} = \{\mathfrak{a}_1,\mathfrak{a}_2,\mathfrak{a}_3,\mathfrak{a}_4\}$, $n = 20$ and $\pi^* = (0.25, 0.25, 0.25, 0.25)$. Suppose an adaptive algorithm[3] selects the following sequence of attributes (given a budget of 20):*

$$\mathfrak{a}_1,\ \mathfrak{a}_2,\ \mathfrak{a}_3,\ \mathfrak{a}_4,\ \mathfrak{a}_1,\ \mathfrak{a}_1,\ \mathfrak{a}_1,\ \mathfrak{a}_1,\ \mathfrak{a}_1,\ \mathfrak{a}_1,\ \mathfrak{a}_4,\ \mathfrak{a}_1,\ \mathfrak{a}_1,\quad \mathfrak{a}_2,\ \mathfrak{a}_2,\ \mathfrak{a}_2,\ \mathfrak{a}_3,\ \mathfrak{a}_3,\ \mathfrak{a}_3,\quad \mathfrak{a}_4. \tag{14}$$

Then as we can see, the algorithm ends up with $\pi_n = (9/20, 1/5, 1/5, 3/20)$.

- *Comparing π_n with π^*, we observe that the set $\mathcal{Z}_o$ is $\{\mathfrak{a}_1\}$ and $\mathcal{Z}_u = \{\mathfrak{a}_2,\mathfrak{a}_3,\mathfrak{a}_4\}$.*
- *The last time an element of $\mathcal{Z}_o$ is queried, that is t_0, is equal to 13 and the corresponding attribute is $z_0 = \mathfrak{a}_1$. The mixture distribution at time t_0 is $\pi_{t_0} = (9/13, 1/13, 1/13, 2/13)$.*
- *Comparing π_{t_0} with π_n, we observe that the set $\mathcal{Z}_o^{(1)}$ is $\{\mathfrak{a}_2,\mathfrak{a}_3\}$ since their fractions increase in the period $[t_0+1, n]_{\mathbb{N}}$, and $\mathcal{Z}_u^{(1)} = \{\mathfrak{a}_4\}$. The last time an element of $\mathcal{Z}_o^{(1)}$ is queried is $t_1 = 19$, and the corresponding attribute is $z_1 = \mathfrak{a}_3$.*
- *Finally, since only one element remains, we have $\mathcal{Z}_o^{(2)} = \{\mathfrak{a}_4\}$, $\mathcal{Z}_u^{(2)} = \emptyset$ and thus $t_2 = n$ and $z_2 = \mathfrak{a}_4$. Also note that the total number of phases above is 3 and hence the term $s = 3-1 = 2$.*

Lemma 4. *Suppose, z_1 is the element of $\mathcal{Z}_o^{(1)}$ (introduced above) queried by the algorithm at time t_1. Then, the following is true at time t_1 for all $z \in \mathcal{Z}\setminus\{z_1\}$*

$$L(z, f_{\pi_{t_1}}) \leq M^* + B_0 + B_1, \qquad \textit{where} \qquad B_1 = 2e_{z_1}(N_{z_1,t_1}) + \left({}^{8C}/_{\pi^*(z_1)}\right)\rho_{t_1} \tag{15}$$

Furthermore, repeating this process till the budget is exhausted, we get for any $z \in \mathcal{Z}$ and an $s < m$

$$L(z, f_{\pi_n}) \leq M^* + \sum_{i=0}^{s} B_i, \qquad \textit{where} \qquad B_i = 2e_{z_i,N_{z_i,t_i}} + \left({}^{8C}/_{\pi_{\min}\pi^*(z_i)}\right)\rho_{t_i}, \quad \textit{for } i > 1. \tag{16}$$

Here s is the (random) number of 'phases' (see Example 2), and is always upper bounded by $m = |\mathcal{Z}|$.

[3] note that the sequence of attributes have been chosen only for illustrating the notation, and do not satisfy the forced exploration condition in Algorithm 2

Proof. To prove (15), we note that at time t_1 for any $z \neq z_1$, we have

$$\begin{aligned}
L(z, f_{\pi_{t_1}}) &\overset{\text{(a)}}{\leq} L(z, \hat{f}_{t_1}) + \frac{2C}{\pi_{t_1}(z)}\rho_{t_1} \;\leq\; \widehat{L}_{t_1}(z, \hat{f}_{t_1}) + e_{z,N_{z,t_1}} + \frac{2C}{\pi_{t_1}(z)}\rho_{t_1} \\
&\overset{\text{(b)}}{\leq} \widehat{L}_{t_1}(z_1, \hat{f}_{t_1}) + e_{z_1,N_{z_1,t_1}} + \frac{2C}{\pi_{t_1}(z_1)}\rho_{t_1} \\
&\overset{\text{(c)}}{\leq} L(z_1, f_{\pi_{t_1}}) + 2e_{z_1,N_{z_1,t_1}} + \frac{4C}{\pi_{t_1}(z_1)}\rho_{t_1} \\
&\overset{\text{(d)}}{\leq} L(z_1, f_{\pi_{t_0}}) + 2e_{z_1,N_{z_1,t_1}} + \frac{4C}{\pi_{t_1}(z_1)}\rho_{t_1} \\
&\overset{\text{(e)}}{\leq} L(z_1, f_{\pi_{t_0}}) + 2e_{z_1,N_{z_1,t_1}} + \frac{4C}{\frac{(m-1)\pi^*(z_1)}{m} + \frac{A}{m}}\rho_{t_1} \\
&\overset{\text{(f)}}{=} L(z_1, f_{\pi_{t_0}}) + B_1 \;\leq\; M^* + B_0 + B_1.
\end{aligned}$$

In the above display
(a) uses Assumption 3 and the event Ω_2 introduced while defining the UCB in (4),
(b) follows from the point selection rule in Line 11 of Algorithm 2,
(c) uses Assumption 3,
(d) uses that from the definition of z_1, we must have $\pi_{t_1}(z_1) \geq \pi_{t_0}(z_1)$, and due to monotonicity assumption (Assumption 1), $L(z_1, f_{\pi_{t_0}}) > L(z_1, f_{\pi_{t_1}})$, and
(e) uses the fact that $\pi_{t_1}(z_1) \geq \pi_{t_0}(z_1) \geq \pi^*(z_1)/2$ proved in (9), and
(f) uses Lemma 1 to bound $L(z_1, f_{\pi_{t_0}})$ with $M^* + B_0$ to get (15).

Now, assume that the budget n satisfies $n \geq n_1 \coloneqq \max\{n_0, n_1'\}$; where

$$n_1' = \texttt{SmallestBudget}\left(c_0 + c_1,\ \pi_{\min},\ \xi,\ \frac{2(\pi_{\min} - A)}{m}\right), \quad \text{where} \tag{17}$$

$$c_0 = \frac{4C}{A_0} + 2; \quad c_1 = \frac{4C}{A_1} + 2; \quad \text{and} \quad A_i \coloneqq \frac{i}{m}A + \frac{(m-i)\pi_{\min}}{m},\ \text{for } 0 \leq i \leq m. \tag{18}$$

Then by the definition of the `SmallestBudget` function (introduced in Definition 2 in Appendix B.1), we must have $\pi_{t_1}(z) \geq \frac{(m-2)\pi^*(z)}{m} + \frac{2A}{m}$ for all $z \in \mathcal{Z}$.

The proof of (16) essentially follows by repeating the above argument a further $s-1$ times. However, there is one minor difference. In proving (15), specifically in Step **(e)**, we used the fact that $\pi_{t_1}(z_1) \geq A_1 \coloneqq \frac{A}{m} + \frac{(m-1)\pi_{\min}}{m}$. For $i \geq 2$, we can similarly use the fact that $\pi_{t_i}(z_i) \geq A_i \coloneqq \frac{iA + (m-1)\pi_{\min}}{m}$. The corresponding requirement on n is that, with

$$\begin{aligned}
&n \geq n_i \coloneqq \max\{n_0, \ldots, n_{i-1}, n_i'\}, \qquad \text{where} \\
&n_i' = \texttt{SmallestBudget}\left(c_0 + \ldots c_i,\ \pi_{\min},\ \xi,\ \frac{(i+1)(\pi_{\min} - A)}{m}\right), \qquad \text{and} \\
&c_i = \frac{4C}{A_i} + 2, \qquad \text{for all} \quad 0 \leq i \leq m.
\end{aligned} \tag{19}$$

□

Lemma 5. *Finally, we obtain that* $\max_{z\in\mathcal{Z}} L(z, \hat{f}_{\pi_n}) - M^* = \mathcal{O}\left(\frac{|\mathcal{Z}|C}{\pi_{\min}} \max_{z\in\mathcal{Z}} e_z(N_A)\right)$*, where* $N = A\left(\frac{\pi_{\min}}{2\pi_{\min} - A}\right) n$.

Proof. First note that for all t_i, $i = 1, 2, \ldots, s$, we have $\rho_t \leq \max_{z\in\mathcal{Z}} e_z(N_{z,t})$. Now, we note the following: for any $i = 0, \ldots s$, we have $\min_{z\in\mathcal{Z}} \pi_{t_i}(z) \geq A_i \geq A$. Also for all $i = 0, \ldots s$, we also have trivially, using (10), $t_i \geq N_A \coloneqq \pi_{\min}/(2\pi_{\min} - A)n$. Together these two statements imply the following:

- $B_0 \leq c_0 \max_{z\in\mathcal{Z}} e_z(N_{z,t_0}) \leq c_0 \max_{z\in\mathcal{Z}} e_z(N_A) = \left(\frac{4C}{\pi_{\min}} + 2\right) \max_{z\in\mathcal{Z}} e_z(N_A)$.

- Similarly, for $i \geq 1$ we have $B_i \leq c_i \max_{z\in\mathcal{Z}} e_z(N_{z,t_i}) \leq c_i \max_{z\in\mathcal{Z}} e_z(N_A) \leq \left(\frac{4C}{A_i}+2\right)\max_{z\in\mathcal{Z}} e_z(N_A)$.

Combining the above two points, we get the following

$$\sum_{i=1}^{s} B_i \leq \sum_{i=1}^{s}\left(\frac{4C}{A_{i-1}}+2\right)\max_{z\in\mathcal{Z}} e_z(N_A)$$

which, if $A \geq \pi_{\min}/2$, implies

$$\max_{z\in\mathcal{Z}} L(z, f_{\pi_n}) - M^* = \mathcal{O}\left(\frac{|\mathcal{Z}|C}{\pi_{\min}}\max_{z\in\mathcal{Z}} e_z(N_A)\right)$$

as required. Note that the above result holds under the assumption that n is large enough to ensure that the conditions in (19) is satisfied for all $0 \leq i \leq s$. Since $s \leq m-1$, a sufficient condition for this is that the condition in (19) is satisfied for all $0 \leq i \leq m-1$. □

B.1.1 Knowledge of parameter C

In our analysis above, we did not impose any condition on the forced exploration parameter ξ; instead we used knowledge of the parameter C from Assumption 3. We now show that if for some $0 < \xi < 1$, it is known that $\lim_{N\to\infty}\max_{z\in\mathcal{Z}}\frac{e_z(N^\xi)}{N^{\xi-1}} = 0$, then we can remove the C dependent term from (4) and obtain the same guarantees for $\mathcal{A}_{\texttt{opt}}$ as in Theorem 1.

The proof will follow the same outline as in the previous section, and to avoid repetition, we obtain a result analogous to Lemma 1. In particular, with the same definition of t_0 and z_0 as in Lemma 1 and Lemma 2, the following is true at time t_0:

$$\begin{aligned}
L(z, f_{\pi_{t_0}}) &\leq \widehat{L}_{t_0}(z, f_{\pi_{t_0}}) + e_z(N_{z,t_0}) \\
&\leq \underbrace{\widehat{L}_{t_0}(z, \hat{f}_{t_0}) + e_z(N_{z,t_0})}_{=U_{t_0}(z,\hat{f}_{t_0})} + \frac{2C}{\pi_{t_0}(z)}\rho_{t_0} \\
&\overset{(a)}{\leq} U_{t_0}(z_0, \hat{f}_{t_0}) + \frac{2C}{\pi_{t_0}(z)}\rho_{t_0} \\
&\leq L(z_0, f_{\pi_{t_0}}) + 2e_z(N_{z_0,t_0}) + 2C\rho_{t_0}\left(\frac{1}{\pi_{t_0}(z_0)} + \frac{1}{\pi_{t_0}(z)}\right) \\
&\overset{(b)}{\leq} L(z_0, f_{\pi_{t_0}}) + \left(\frac{2C}{\pi_{\min}} + \frac{2C}{t_0^{\xi-1}}\right)\rho_{t_0} \\
&\overset{(c)}{\leq} L(z_0, f_{\pi_{t_0}}) + \underbrace{\left(\frac{2C}{\pi_{\min}} + \frac{2C}{t_0^{\xi-1}}\right)\max_{z\in\mathcal{Z}} e_z(t_0^\xi)}_{:=B_0'}
\end{aligned}$$

In the above display,
(a) uses the fact that at time t_0, the attribute z_0 was selected by maximizing $U_{t_0}(z, \hat{f}_{t_0})$,
(b) uses the fact that $\pi_{t_0}(z_0) > \pi_{\min}$ and $\pi_{t_0}(z) \geq t_0^{\xi-1}$ for all $z \neq z_0$ since the forced exploration step was not invoked at time t_0, and
(c) uses the fact that $\rho_{t_0} \leq \max_{z\in\mathcal{Z}} e_z(N_{z,t_0}) \leq \max_{z\in\mathcal{Z}} e_z(t_0^\xi)$ due to the monotonicity of $e_z(N)$ and the fact that $N_{z,t_0} \geq t_0^\xi$ for all $z \in \mathcal{Z}$ at time t_0.

Now, to continue with the rest of the proof as in Appendix B.1, we need that the term B_0' converges to zero as n (and hence, t_0) goes to infinity. The first term of B_0', i.e., $2C\max_z e_z(t_0^\xi)/\pi_{\min}$, converges to zero from the condition that the uniform confidence bound converges to zero as the number of samples, t_0^ξ, goes to infinity; while the second term, $\frac{2C}{t_0^{\xi-1}}\max_z e_z(t_0^\xi)$ converges to zero from the assumption on ξ made at the beginning of this section. Using this fact, we can then proceed as in Lemmas 2, 4, and 5 to get the final result.

B.2 Relation to Active Learning in Bandits

Our formulation of the minimax fair classification problem diverges from the Active learning in bandit (ALB) problem due to the fact that drawing samples from one attribute can reduce the performance of another. To make the discussion concrete, consider an ALB problem with two distributions $Q_1 \sim N(\mu_1, \sigma_1^2)$ and $Q_2 \sim N(\mu_2, \sigma_2^2)$ for $\mu_1, \mu_2 \in \mathbb{R}$ and $\sigma_1^2, \sigma_2^2 \in (0, \infty)$. Let $N_{i,t}$ denote the number of samples allocated by an agent to Q_i for $i = 1, 2$ and $t \geq 1$. Then, we can construct high probability confidence sequences around $\hat{\mu}_{i,t}$ (the empirical counterparts to μ_i at time t) of non-increasing (in t) lengths $e_{i,t}$ such that $|\hat{\mu}_{i,t} - \mu_i| \leq e_{i,t}$. If at time t, the agent decides to draw a sample from arm 1, it would not change the quality of its estimate of Q_2 at time $t + 1$ and beyond.

On the other hand, let us consider Instance 2 of $\texttt{SyntheticModel1}(\boldsymbol{\mu})$ (introduced in Definition 1) with $\mathcal{F}$ being the set of all linear classifiers passing through origin. This is illustrated in Figure 8, where the circles denote the one-square-deviation region around the mean values. The shaded circles correspond to the attribute $Z = v$, and the circles with dashed boundaries correspond to the label $Y = 0$. For example, $P_{X|YZ}(\cdot|Y = 0, Z = u)$ is denoted by the top left circle. Suppose at some time t_1, the current classifier $\hat{f}_{t_1}$ (represented by the solid gray line passing through origin in Figure 8) has low (resp. high) accuracy for the protected attribute $Z = v$ (resp. $Z = u$). To remedy this, the agent may draw more samples from the distribution of attribute $Z = v$ to skew the training dataset distribution towards $Z = v$. This would result in the updated classifier $\hat{f}_{t_2}$ (the dashed gray line in Figure 8) at some time $t_2 > t_1$ to achieve high accuracy for the attribute $Z = v$. But this increased accuracy for $Z = v$ comes at the cost of a reduction in the prediction accuracy for the attribute $Z = u$, as shown in Figure 8.

The above discussion highlights the key distinguishing feature of our problem from the prior work in ALB. Because of this distinction, the existing analyses of the ALB algorithms do not carry over directly to our case. As a result, to quantify the performance of Algorithm 2, we device a new 'multi-phase' approach which is described in detail in Appendix. B.1.

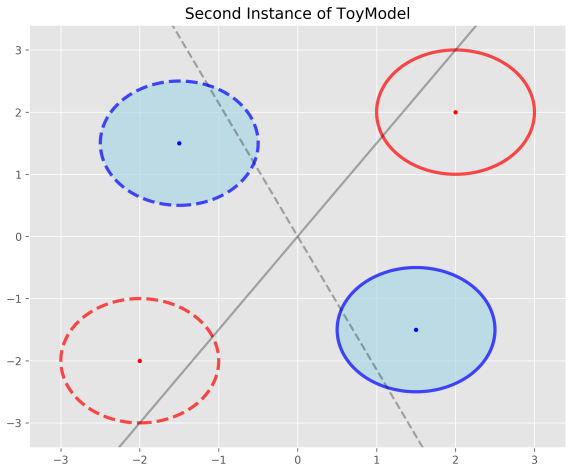

Figure 8: Figure demonstrates the second instance of $\texttt{SyntheticModel1}$ introduced in Section 2.2. Here the circles denote the one-standard-deviation regions of the distributions of features X conditioned on Y and Z. Shaded circles correspond to $Z = u$ and circles with dashed boundaries represent $Y = 0$. The solid gray line represents a possible linear classifier that may be learned if the dataset has fewer examples from $Z = v$. If an adaptive algorithm addresses this by populating the data-set with numerous examples from $Z = v$, a possible updated linear classifier is shown by the dashed gray line. This demonstrates the main distinguishing feature of our problem w.r.t. the active learning in bandits (ALB) problem: *allocating samples to improve the performance on one attribute can have an adverse effect on the performance of the resulting classifier on other attributes.*

C Details of ϵ-greedy strategy

The ϵ-greedy strategy originally proposed by Abernethy et al. (2020) proceeds as follows at time t: with probability $1-\epsilon$, draw a pair (X_t, Y_t) from the distribution P_{z_t} where z_t is the attribute with largest empirical loss; and with probability ϵ draw (Z_t, X_t, Y_t) from the population distribution P_{XYZ}. If $\pi_Z(\cdot)$ denotes the marginal of the population distribution over $\mathcal{Z}$, i.e., $\pi_Z(z) = \sum_{x,y} P_{XYZ}(x, y, z)$; then the ϵ-greedy strategy can be equivalently described as follows: For any $t = 1, 2, \ldots$, do

- Draw a random variable $Q_t \sim \texttt{Bernoulli}(\epsilon)$.
- If $Q_t = 1$, draw $Z_t \sim \pi_Z$. Else set $Z_t \in \arg\max_{z\in\mathcal{Z}} \hat{L}_t(z, \hat{f}_t)$.
- Draw a pair of samples from the distribution P_{Z_t}.

For simplicity of presentation, we assume that π_Z is the uniform distribution, i.e., $\pi_Z(z) = 1/m$ for all $z \in \mathcal{Z}$. However, our theoretical results can be easily generalized to the case of any π_Z which places non-zero mass on all $z \in \mathcal{Z}$.

C.1 Analysis of the ϵ-greedy Sampling Strategy

The ϵ-greedy strategy proposed in Abernethy et al. (2020) proceeds as follows: for any $t \geq 1$, it maintains a candidate classifier $\hat{f}_t$ along with its empirical loss on m validation sets corresponding to the attributes $z \in \mathcal{Z}$. At each round t, the algorithm selects an attribute z_t as follows: with probability ϵ, z_t is drawn from a fixed distribution over $\mathcal{Z}$, and with probability $1-\epsilon$, it is set to the distribution with the largest empirical loss. Having chosen z_t, the algorithm draws a pair of samples from P_{z_t} and updates the training and validation sets, as well as the classifier $\hat{f}_t$. This process is continued until the sampling budget is exhausted.

Abernethy et al. (2020) present two theoretical results on the performance of their ϵ-greedy strategy. The first one (Abernethy et al., 2020, Theorem 1), is an asymptotic consistency result derived under somewhat restrictive assumptions: $\mathcal{X} = \mathbb{R}$, $\mathcal{F}$ is the class of threshold classifiers, and the learner has access to an oracle which returns the exact loss, $L(z, f_{\pi_t})$, for every π_t. In their second result (Abernethy et al., 2020, Theorem 2), they analyze the greedy version (i.e., $\epsilon = 0$) of the algorithm with $|\mathcal{Z}| = 2$, and show that at time n, either the excess risk is of $\mathcal{O}\big(\max_{z\in\mathcal{Z}} \sqrt{2d_{VC}(\ell\circ\mathcal{F})\log(2/\delta)/N_{z,n}}\big)$, or the algorithm draws a sample from the attribute with the largest loss. However, due to the greedy nature of the algorithm analyzed, there are no guarantees that $N_{z,n} = \Omega(n)$, and thus, in the worst case the above excess risk bound is $\mathcal{O}(1)$.

We now show how the techniques we developed for the analysis of Algorithm 2 can be suitably employed to study the ϵ-greedy strategy under the same assumptions as in Theorem 1. In our results, we assume that the distribution according to which ϵ-greedy selects an attribute with probability ϵ at each round t is uniform. This is just for simplicity and all our results hold for any distribution which places non-zero mass on all $z \in \mathcal{Z}$.

We first present an intuitive result that says if the ϵ-greedy strategy is too *exploratory* (ϵ is large), the excess risk will not converge to zero. We present an illustration of this result using Instance I of the SyntheticModel1 introduced in Section 2.2 in Figure 9 in Appendix C.

Proposition 2. *If the ϵ-greedy strategy is implemented with $\epsilon > m\pi_{\min} := |\mathcal{Z}| \min_{z\in\mathcal{Z}} \pi^*(z)$, then its excess risk is $\Omega(1)$ with probability at least $1-\delta$ for n large enough (see Equation 21 in Appendix C.2 for the precise condition on n).*

Proof Outline. The result follows from two observations: **(i)** For all t larger than a term $\tau_0(\delta, \epsilon, \pi_{\min})$, defined in (21), with probability at least $1-\delta$, for any $z \in \mathcal{Z}$, we must have $\pi_t(z) \geq \frac{\pi_{\min}+\epsilon/m}{2}$, and **(ii)** Suppose $\pi_{\min}$ is achieved by some attribute $z_{\min}$, i.e., $\pi^*(z_{\min}) = \pi_{\min}$. Then, the first statement implies that the excess risk of the ϵ-greedy algorithm is at least $\min_{\pi:\pi(z_{\min})\geq(\pi_{\min}+\epsilon/m)/2} L(z_{\min}, f_\pi) - M^*$, which is strictly greater than zero by Assumption 1. The detailed proof is reported in Appendix C.2. □

According to Proposition 2, for the excess risk to converge to zero, the ϵ-greedy strategy must be implemented with $\epsilon \leq \pi_{\min}/m$. We next derive an upper-bound on the excess risk of ϵ-greedy, similar to the one in Theorem 1 for Algorithm 2.

Theorem 2. *Let Assumptions 1-3 hold and ϵ-greedy implemented with $0 < \epsilon < m\pi_{\min}$. If the query budget n is sufficiently large (see Equations 23 and 24 in Appendix C.2 for the exact requirements), then for any $0 < \beta < \frac{m\pi_{\min}}{\epsilon} - 1$, with probability at least $1 - 2\delta$, we have*

$$\mathcal{R}_n(\epsilon\text{-greedy}) = \mathcal{O}\Big(\frac{|\mathcal{Z}|C}{q} \max_{z \in \mathcal{Z}} e_z(N_q)\Big), \tag{20}$$

where $q = \epsilon/m$, $N_q = qn\big(\pi_{\min} - q(1+\beta)\big)(1-\beta)$, and C is the parameter introduced in Assumption 3.

Remark 4. *The assumption on ϵ in Theorem 2 implies that $q = \epsilon/m < \pi_{\min}$, and as a result $qn\pi_{\min} < \pi_{\min}^2 n$. Given this and the monotonicity of the size of the confidence interval $e_z(N)$ w.r.t. N, we may conclude that the bound on the excess risk of Alg. 2 (Equation 5) is always tighter than the one for ϵ-greedy strategy (Equation 20). Note that for the classifiers with finite VC-dimension, the bound in* (20) *is of the same order in n as the one in* (5)*, but with a larger leading constant.*

C.2 Proof of Proposition 2

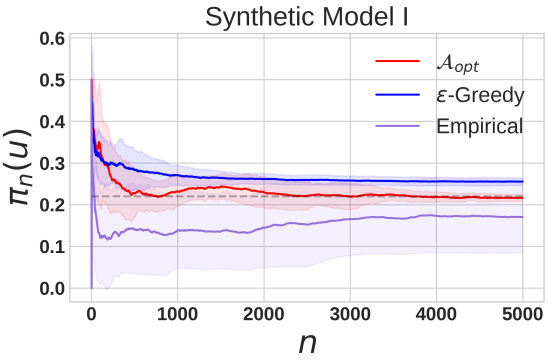

Figure 9: This figure shows an instance when the ϵ-greedy strategy is over-exploratory which results in the empirical mixture distribution π_t being strictly away from π^*, and thus resulting in $\Omega(1)$ excess risk even as n goes to infinity. For the ϵ-greedy strategy in this particular figure, we used $\epsilon = .5$ with $m = 2$ and $\pi_{\min} = \min_z \pi^*(z) \approx 0.23$. This provides a numerical demonstration of the statement of Proposition 2. Note that the $\mathcal{A}_{\texttt{opt}}$ strategy is more resilient to the wrong choice of the exploration parameter c_0: in this figure we used $c_0 = 1.0$ (much larger than the 0.1 value used in experiments) and the corresponding $\pi_n(u)$ for $\mathcal{A}_{\texttt{opt}}$ still eventually converges towards $\pi^*(u)$.

Fix any $z \in \mathcal{Z}$, and decompose $N_{z,t}$ into $N_{z,t}^{(0)} + N_{z,t}^{(1)}$ where, $N_{z,t}^{(1)} = \sum_{i=1}^t \mathbb{1}_{\{z_i = z\}} Q_i$ is the number of times up to round t the attribute z was queried due to the exploration step (i.e., $Q_t = 1$) of the ϵ-greedy algorithm.

Now, define $M_{z,0} = 0$ and $M_{z,t} = M_{z,t-1} + Q_t \mathbb{1}_{\{\tilde{z}_t = z\}} - \epsilon/m$. Due to the fact that Q_t and $\tilde{z}_t$ are independent, it is easy to check that $(M_{z,t})_{t \geq 0}$ forms a martingale sequence. For some given $\delta > 0$, introduce the notation $\delta_t = \frac{6\delta}{m\pi^2 t^2}$ and $b_t = \sqrt{\frac{t \log(2/\delta_t)}{2}}$. We then have the following:

$$\begin{aligned}
\sum_{t=1}^n P\left(M_{z,t} > b_t\right) &\leq \sum_{t=1}^n \left(\prod_{i=1}^t \mathbb{E}\left[\exp\left(\lambda Q_i \mathbb{1}_{\{\tilde{z}_i = z\}} - \epsilon/m\right) \middle| (Q_j)_{j=1}^{i-1}\right] \right) e^{-\lambda b_t} \\
&\leq \sum_{t=1}^n e^{t\lambda^2/8 - \lambda b_t} \overset{(a)}{=} \sum_{t=1}^n e^{-2b_t^2/t} \leq \sum_{t=1}^\infty \frac{\delta_t}{2} = \frac{\delta}{2m}.
\end{aligned}$$

In the above display, **(a)** follows by setting $\lambda = 4b_t/t$. By repeating the same argument with the martingale sequence $\{-M_{z,t} : t \geq 0\}$, we get that $P\left(|M_{z,t}| \leq b_t, \forall t \geq 1,\ \forall z \in \mathcal{Z}\right) \geq 1 - \delta$. In other words, it implies that, with probability at least $1 - \delta$, we have the following:

$$N_{z,t}^{(1)} \geq \frac{t}{\epsilon} - b_t, \quad \forall\, t \geq 1,\ \forall z \in \mathcal{Z}.$$

This implies that $N_{z,t} \geq \epsilon/m - b_t$ for all z, t. In particular, if $z_{\min}$ is the attribute such that $\pi^*(z_{\min}) = \min_{z \in \mathcal{Z}} \pi^*(z) = \pi_{\min}$, then this implies that $\hat{\pi}_t(z_{\min}) \geq \epsilon/m - b_t/t$ for all t, z. Using the fact that $\lim_{t\to\infty} b_t/t = 0$, we define $\tau_0 = \tau_0(\delta, \epsilon, \pi_{\min})$, as follows:

$$\tau_0(\delta, \epsilon, \pi_{\min}) \coloneqq \min \left\{ t \geq 1 : \sqrt{\frac{\log\left(m\pi^2 t^2/3\delta\right)}{2t}} \leq \frac{\epsilon - m\pi_{\min}}{2m} \right\}. \tag{21}$$

This implies that with probability at least $1 - \delta$, $\pi_t(z_{\min}) \geq (\pi_{\min} + \epsilon/m)/2$ for all $t \geq \tau_0$. Hence the excess risk of the ϵ-greedy algorithm under these conditions is at least $\min_{\pi:\pi(z_{\min}) \geq (\pi_{\min}+\epsilon/m)/2} \mathbb{E}_{z_{\min}} [\ell(f_\pi, X, Y)] - M^*$, which is an $\Omega(1)$ term due to Assumption 1.

C.3 Proof of Theorem 2

First we partition the interval $[1, n]_{\mathbb{N}}$ into T_0 and T_1, where $\mathrm{T}_j = \{t : Q_t = j\}$ for $j = 0, 1$. In other words, T_0 denotes the times at which the ϵ-greedy strategy is *greedy* while T_1 denotes the times at which $Q_t = 1$ and the ϵ-greedy strategy is exploratory.

Then we introduce the following terms:

- Define $\mathcal{Z}_o$ and $\mathcal{Z}_u$ as in the proof of Theorem 1. That is, $\mathcal{Z}_o = \{z \in \mathcal{Z} : \pi_n(z) > \pi^*(z)\}$ and $\mathcal{Z}_u = \mathcal{Z} \setminus \mathcal{Z}_o$. We assume, as before, that $\mathcal{Z}_o \neq \emptyset$. Then, we define $t_0 \coloneqq \max\{t \in \mathrm{T}_0 : z_t \in \mathcal{Z}_o\}$, and denote the corresponding attribute (queried at time t_0) with z_0.
- Then, by appealing to Lemma 3, we note that there must exist a nonempty $\mathcal{Z}_o^{(1)} \subset \mathcal{Z}_u$ such that $\pi_n(z) > \pi_{t_0}(z)$ for all $z \in \mathcal{Z}_o^{(1)}$. Using this define $t_1 = \max\{t \in \mathrm{T}_0 : z_t \in \mathcal{Z}_o^{(1)}\}$ and use z_1 to denote the corresponding attribute queried at time t_1. We can proceed in this way to define $\{(t_j, z_j) : 2 \leq j \leq s\}$ for an appropriate $s \leq m - 1$ (such that $t_s = n$).

First, we define the $1 - \delta$ probability event Ω_3 as follows (see proof of Proposition 2 in Appendix C.2 for derivation):

$$\Omega_3 = \left\{ \left| N_{z,t}^{(1)} - \frac{\epsilon}{m} t \right| \leq \underbrace{\sqrt{\frac{t \log(m\pi^2 t^2/3\delta)}{2}}}_{\coloneqq b_t} \right\}. \tag{22}$$

For the rest of this proof, we will assume that the $(1 - 2\delta)$ probability event $\cap_{i=1}^3 \Omega_i$ holds, where Ω_1 and Ω_2 are the same uniform deviation results that were used in defining the UCB in (4).

Then we have the following:

Lemma 6. *At time t_0, we have $L(z, f_{\pi_{t_0}}) \leq M^* + \tilde{B}_0$ where $\tilde{B}_0 \coloneqq 4\left(1 + C/q - \beta_{t_0}/t_0\right) \max_{z \in \mathcal{Z}} e_{z, N_{z,t_0}}$ and $q = \epsilon/m$.*

Proof.

$$\begin{aligned}
L(z, f_{\pi_{t_0}}) &\leq \widehat{L}_{t_0}(z, f_{\pi_{t_0}}) + 2e_z(N_{z,t_0}) \\
&\leq \widehat{L}_{t_0}(z, \hat{f}_{t_0}) + 2e_z(N_{z,t_0}) + \frac{2C}{\pi_{t_0}(z)} \rho_{t_0} \\
&\leq \widehat{L}_{t_0}(z_0, \hat{f}_{t_0}) + 2e_z(N_{z,t_0}) + \frac{2C}{\pi_{t_0}(z)} \rho_{t_0} \\
&\leq L(z_0, f_{\pi_{t_0}}) + 2\left(e_z(N_{z,t_0}) + e_{z_0}(N_{z_0,t_0})\right) + 2C\rho_{t_0} \left(\frac{1}{\pi_{t_0}(z_0)} + \frac{1}{\pi_{t_0}(z)} \right) \\
&\leq L(z_0, f_{\pi_{t_0}}) + 4\left(1 + \frac{C}{\epsilon/m - b_{t_0}/t_0}\right) \max_{z \in \mathcal{Z}} e_z(N_{z,t_0}) \\
&= L(z_0, f_{\pi_{t_0}}) + \tilde{B}_0.
\end{aligned}$$

□

Lemma 7. *Suppose n is large enough to satisfy the conditions in* (24) *for some fixed* $0 < \beta < \pi_{\min}/q - 1$. *Then we have* $\tilde{B}_0 \leq 4\,(1 + C/q(1-\beta))\max_{z\in\mathcal{Z}} e_z(N_q)$, *where* $N_q := nq\,(\pi_{\min} - q(1+\beta))\,(1-\beta)$.

Due to the definitions of event Ω_3 and the time t_0, it must be the case that $N_{z_0,t_0} \geq N^{(0)}_{z_0,n} \geq \pi^*(z_0)n - N^{(1)}_{z_0,n} \geq (\pi^*(z_0) - \epsilon/m = -b_n/n)\,n$. Next, we assume that n is large enough, such that the following are satisfied for some $0 < \beta$:

- First, we assume that n is large enough to ensure that $b_n/n \leq \beta\epsilon/m$ for some $0 < \beta < m\pi_{\min}/\epsilon - 1$, i.e.,

$$n \geq \tilde{n_0}(\beta) := \min\left\{t \geq 1 : \frac{\log t}{t} \leq \frac{\beta^2\epsilon^2}{m^2} - \frac{1}{2}\log\left(m\pi^2/3\delta\right)\right\}. \tag{23}$$

  This implies that $t_0 \geq n\,(\pi^*(z_0) - \epsilon/m(1+\beta)) \geq n\left(\pi_{\min} - \frac{\epsilon(1+\beta)}{m}\right)$.

- Next, we assume that $b_{t_0}/t_0 \leq \beta\epsilon/m$. A sufficient condition for this is

$$n\left(\pi_{\min} - \frac{\epsilon(1+\beta)}{m}\right) \geq \tilde{n}_0(\beta) \quad \Rightarrow \quad n \geq \frac{\tilde{n}_0(\beta)}{\pi_{\min} - \epsilon(1+\beta)/m}. \tag{24}$$

Then, with the notation $N_q = n\,(\pi_{\min} - q(1+\beta))\,(q(1-\beta))$, where $q = \epsilon/m$, we have the following:

$$\tilde{B}_0 = 4\left(1 + \frac{C}{q - b_{t_0}t_0}\right)\max_{z\in\mathcal{Z}} e_z(N_{z,t_0}) \leq 4\left(1 + \frac{C}{q(1-\beta)}\right)\max_{z\in\mathcal{Z}} e_z(N_q).$$

Now, proceeding in the same way as in Lemma 4 and Lemma 5, we can define the terms $\tilde{B}_j$ for $j \geq 1$ (analogous to the terms B_j introduced in Lemma 4) to show that with probability at least $1 - 2\delta$, the excess risk resulting from the ϵ-`greedy` strategy satisfies:

$$\max_{z\in\mathcal{Z}} L(z, f_{\pi_n}) - M^* \leq \sum_{j=0}^{s-1}\tilde{B}_j = \mathcal{O}\left(\frac{|\mathcal{Z}|C}{q(1-\beta)}\right)\max_{z\in\mathcal{Z}} e_z(N_q),$$

as required.

C.4 Comparison with $\mathcal{A}_{\texttt{opt}}$

The ϵ-`greedy` strategy differs from $\mathcal{A}_{\texttt{opt}}$ in two major ways:

1. The excess risk bound derived in Theorem 1 for $\mathcal{A}_{\texttt{opt}}$ is always tighter than the corresponding bound for ϵ-`greedy` strategy in Theorem 2. In particular, for the family of classifiers with finite VC dimension, both the algorithms achieve same convergence rates w.r.t. n, but the ϵ-`greedy` strategy has a larger leading constant.
2. From a practical point of view, the ϵ-`greedy` strategy is less robust to the choice of parameter ϵ as compared to the $\mathcal{A}_{\texttt{opt}}$ strategy. For instance, as shown in Figure 9, choosing a large value of ϵ may result in the mixture distribution (π_t) not converging to π^*, whereas even with much larger values of c_0, the π_t from $\mathcal{A}_{\texttt{opt}}$ algorithm still eventually converges to π^*.

D Proof of Lower Bound

In this section, we first formally state the lower bound result and then present its proof. We denote by $\mathcal{M}$ the class of `SyntheticModel1`s introduced in Definition 1. We also define the following class of problems:

Definition 3. Let $\mathcal{Q}$ denote the class of problems defined by the triplets $(\boldsymbol{\mu}, \mathcal{F}, \ell_{01})$, where $\boldsymbol{\mu} \in \mathcal{M}$ is an instance of the `SyntheticModel1`, $\mathcal{F}$ is the class of linear classifiers in two dimensions, and ℓ_{01} is the $0-1$ loss.

For the function class $\mathcal{F}$, we know that $e_z(N) = \mathcal{O}(\sqrt{\log(n/\delta)/N})$ for both $z \in \mathcal{Z} = \{u, v\}$, which implies that the expected excess risk achieved by both $\mathcal{A}_{\texttt{opt}}$ and ϵ-`greedy` strategies is of $\mathcal{O}\left(\sqrt{\log(n)/n}\right)$. We now prove that this convergence rate (in terms of n) for this class of problems.

Proposition 3. *Suppose $\mathcal{A}$ is any adaptive sampling scheme which is applied with a budget n to a problem $Q \in \mathcal{Q}$ introduced in Definition 3. Then, we have*

$$\max_{Q \in \mathcal{Q}} \mathbb{E}_Q \left[\mathcal{R}_n\left(\mathcal{A}\right)\right] = \Omega\left(1/\sqrt{n}\right). \tag{25}$$

To prove this proposition, we consider two problem instances in the class of problems, $\mathcal{Q}$, used in the statement of Proposition 3, denoted by Q_μ and Q_γ. The instance Q_μ has the synthetic model with mean vectors $\mu_{0u} = (-r, r), \mu_{1u} = (r, -r), \mu_{0v} = (-r', -r')$ and $\mu_{1v} = (r', r')$ for some $r' > r > 0$, and Q_γ has the mean vectors $\mu_{0u} = (-r', r'), \mu_{1u} = (r', -r'), \mu_{0v} = (-r, -r)$ and $\mu_{1v} = (r, r)$. For both these problem instances, implementing an adaptive sampling algorithm $\mathcal{A}$ with a budget n induces a probability measure on the space $(\mathcal{X} \times \mathcal{Y} \times \mathcal{Z})^n$. We use $\mathbb{P}_\mu$ and $\mathbb{P}_\gamma$ (resp. $\mathbb{E}_\mu$ and $\mathbb{E}_\gamma$) to denote the probability measures (resp. expectations) for the problem instances Q_μ and Q_γ respectively.

Then we have the following KL-divergence decomposition result.

Lemma 8. *For any event E, we have the following:*

$$\frac{n(r'-r)^2}{2} = D_{KL}\left(\mathbb{P}_\mu, \mathbb{P}_\gamma\right) \geq d_{KL}\left(\mathbb{P}_\mu(E),\, \mathbb{P}_\gamma(E)\right) \geq 2\left(\mathbb{P}_\mu(E) - \mathbb{P}_\gamma(E)\right)^2.$$

In the above display, $d_{KL}(a_1, a_2)$ for $a_1, a_2 \in [0, 1]$ denotes the KL-divergence between two Bernoulli random variables with parameters a_1 and a_2 respectively.

As a consequence of the above result, we have $|\mathbb{P}_\mu(E) - \mathbb{P}_\gamma(E)| \leq (r'-r)\sqrt{n/2}$.

Proof. The first inequality is a consequence of the data-processing inequality for KL-divergence (Polyanskiy and Wu, 2015, Corollary 2.2), while the second inequality follows from an application Pinsker's inequality (Polyanskiy and Wu, 2015, Theorem 6.5).

We now show the derivation of the first equality in the statement. Let $\mathcal{H}_t$ denote the history at the beginning of round t, i.e., $\mathcal{H}_t = (Z_1, X_1, Y_1, \ldots, Z_{t-1}, X_{t-1}, Y_{t-1})$. Then for any sequence of $(Z_1, X_1, Y_1, \ldots, Z_n, X_n, Y_n)$, we have

$$\mathbb{P}_\mu\left(Z_1, \ldots, Y_n\right) = \prod_{t=1}^{n} \mathcal{A}\left(Z_t | \mathcal{H}_{t-1}\right) Q_\mu\left(Y_t | Z_t\right) Q_\mu\left(X_t | Y_t, Z_t\right).$$

We can write a similar expression for $\mathbb{P}_\gamma$ as well. Now, proceeding as in (Lattimore and Szepesvári, 2020, Lemma 15.1), we get the following divergence decomposition result

$$\begin{aligned} D_{KL}\left(\mathbb{P}_\mu, \mathbb{P}_\gamma\right) &= \mathbb{E}_\mu\left[N_{u,n}\right] D_{KL}\left(Q_\mu(\cdot, \cdot | Z = u), Q_\gamma(\cdot, \cdot | Z = u)\right) + \mathbb{E}_\mu\left[N_{v,n}\right] D_{KL}\left(Q_\mu(\cdot, \cdot | Z = v), Q_\gamma(\cdot, \cdot | Z = v)\right) \\ &= \left(\mathbb{E}_\mu\left[N_{u,n}\right] + \mathbb{E}_\mu\left[N_{v,n}\right]\right) \frac{(r'-r)^2}{2} = \frac{n(r'-r)^2}{2}, \end{aligned}$$

where the last equality uses the expression for KL-divergence between two multi-variate Gaussian distributions. □

Every linear classifier $f \in \mathcal{F}$ can be equivalently parametrized by a normal vector $w \in \mathbb{R}^2$ such that $f(x) = \text{sign}(\langle w, x \rangle)$. We know that if $\pi(u) = 1$, then the optimal linear classifier is the one with normal vector $w_1 = (1, -1)$. That is $f^*(x) = \text{sign}(\langle w, x \rangle)$. Similarly, when $\pi(u) = 0$, the optimal linear classifier is the one with $w_0 = (1, 1)$, and it varies continuously from w_0 to w_1 as $\pi(u)$ increases from 0 to 1. Now, let a be the value of $\pi(u)$ at which the optimal linear classifier $f_\pi(x) = \text{sign}(\langle w, x \rangle)$ with $w = (1, 0)$. In the next lemma, we consider the event $E = \{\pi_n(u) > a\}$.

Lemma 9. *Suppose π_n denotes the mixture distribution returned by the algorithm $\mathcal{A}$ after n rounds when applied to a problem $Q \in \{Q_\mu, Q_\gamma\}$. Introduce the event $E = \{\pi_n(u) > a\}$ and assume that $r' < 2r$. Recall that $a \in (0, 1)$ is value of $\pi(u)$ such that the corresponding optimal linear classifier has the normal vector $(1, 0)$; that is, $f_\pi(x) = \text{sign}(\langle x, (1, 0) \rangle)$. Then we have the following:*

$$\mathbb{E}_\mu\left[\mathcal{R}_n(\mathcal{A})\right] \geq \frac{c_r}{2\sqrt{2}}(r'-r)\mathbb{P}_\mu(E^c) \quad \textit{and} \quad \mathbb{E}_\gamma\left[\mathcal{R}_n(\mathcal{A})\right] \geq \frac{c_r}{2\sqrt{2}}(r'-r)\mathbb{P}_\gamma(E),$$

where $c_r \coloneqq \min_{x\in[r,2r]} \left|\Phi'\left(\frac{x}{\sqrt{2}}\right)\right|$ *and* $\Phi(\cdot)$ *denotes the cumulative distribution function (cdf) of a standard Normal random variable.*

Proof. We present the details only for the first inequality as the second inequality follows in an entirely analogous manner by replacing E^c with E.

$$\begin{aligned}
\mathbb{E}_\mu\left[\mathcal{R}_n(\mathcal{A})\right] &= \mathbb{E}_\mu\left[\mathcal{R}_n(\mathcal{A})\mathbb{1}_{\{E\}}\right] + \mathbb{E}_\mu\left[\mathcal{R}_n(\mathcal{A})\mathbb{1}_{\{E^c\}}\right] \\
&\geq \mathbb{E}_\mu\left[\mathcal{R}_n(\mathcal{A})\mathbb{1}_{\{E^c\}}\right] \\
&\overset{(a)}{\geq} \left(\frac{\Phi(r/\sqrt{2}) - \Phi(r'/\sqrt{2})}{2}\right)\mathbb{P}_\mu(E^c) \\
&\geq \min_{x\in[r,2r]}\left|\Phi'(x/\sqrt{2})\right|\left(\frac{r'-r}{2\sqrt{2}}\right)\mathbb{P}_\mu\left(E^c\right) \\
&= \frac{c_r}{2\sqrt{2}}(r'-r)\mathbb{P}_\mu\left(E^c\right).
\end{aligned}$$

The key observation in the proof which relies on the choice of $\mathcal{F}$ as the set of all linear classifiers, is in step **(a)** above. This step uses the fact that under the event E^c, when $\pi_n(u) \leq a$, the minimax loss must be at least greater than $\frac{\Phi(r/\sqrt{2})-\Phi(r'/\sqrt{2})}{2}$ (here $\Phi(\cdot)$ denotes the cdf of the standard normal random variable). This follows from the fact that under model Q_μ, the classifier w^*_μ which corresponds to the optimal mixture distribution π^*_μ must satisfy the condition that $\langle w^*_\mu, (1,0)\rangle \geq 0$; whereas under the event E^c the optimal classifier w^*_a satisfies $\langle w^*_a, (1,0)\rangle \leq 0$ (by definition of event E). □

The final result now follows by combining the results of Lemma 8 and Lemma 9. In particular, we have the following:

$$\begin{aligned}
\max_{Q\in\mathcal{Q}}\mathbb{E}_Q\left[\mathcal{R}_n\left(\mathcal{A}\right)\right] \geq \max_{Q\in\{Q_\mu,Q_\gamma\}}\mathbb{E}_Q[\mathcal{R}_n\left(\mathcal{A}\right)] &\overset{(a)}{\geq} \frac{1}{2}\left(\mathbb{E}_\mu\left[\mathcal{R}_n\left(\mathcal{A}\right)\right] + \mathbb{E}_\gamma\left[\mathcal{R}_n\left(\mathcal{A}\right)\right]\right) \\
&\overset{(b)}{\geq} \frac{c_r}{4\sqrt{2}}(r'-r)\left(\mathbb{P}_\mu\left(E^c\right) + \mathbb{P}_\gamma\left(E\right)\right) \\
&\overset{(c)}{\geq} \frac{c_r}{4\sqrt{2}}(r'-r)\left(1 - \left|\mathbb{P}_\mu\left(E\right) - \mathbb{P}_\gamma\left(E\right)\right|\right) \\
&\overset{(d)}{\geq} \frac{c_r}{4\sqrt{2}}(r'-r)\left(1 - (r'-r)\sqrt{\frac{n}{2}}\right) \\
&\overset{(e)}{\geq} \frac{c_r}{16\sqrt{n}}.
\end{aligned}$$

In the above display:
(a) uses the fact that average is smaller than maximum,
(b) lower bounds $\mathbb{E}_\mu\left[\mathcal{R}_n\left(\mathcal{A}\right)\right]$ and $\mathbb{E}_\gamma\left[\mathcal{R}_n\left(\mathcal{A}\right)\right]$ using the result of Lemma 9,
(c) uses the fact that $\mathbb{P}_\mu(E^c) + \mathbb{P}_\gamma(E) = 1 - \mathbb{P}_\mu(E^c) + \mathbb{P}_\gamma(E) \geq 1 - |\mathbb{P}_\mu(E^c) - \mathbb{P}_\gamma(E)|$,
(d) follows by using the bound $|\mathbb{P}_\mu(E^c) - \mathbb{P}_\gamma(E)| \leq (r'-r)\sqrt{n/2}$ derived in Lemma 8, and finally,
(e) follows by setting $r' = r + 1/\sqrt{2n}$.

D.1 Extending to $|\mathcal{Z}| > 2$

The same idea employed for the case of $m = |\mathcal{Z}| = 2$ can be generalized for larger even values of m. In particular, let $\mathcal{X} = \mathbb{R}^2$, $\mathcal{Y} = \{0,1\}$ as before, and for some even $m > 2$ let $\mathcal{Z} = \{1, 2, \ldots, m\}$. Furthermore, let $\mathcal{Z}_o = \{1, 3, \ldots, m-1\}$ and $\mathcal{Z}_e = \{2, \ldots, m\}$ denote the odd and even numbered elements of $\mathcal{Z}$ (in fact any partition of $\mathcal{Z}$ into two sets of size $m/2$ each works). For any $i \in \mathcal{Z}$, we will use $\mathcal{Z}_i$ to denote $\mathcal{Z}_o$ if i is odd, and $\mathcal{Z}_e$ if i is even.

Now, consider $m+1$ problem instances for the above $(\mathcal{X}, \mathcal{Y}, \mathcal{Z})$, denoted by $Q_0, \ldots, Q_m$, and defined as follows:

- For Q_0, we have $P_{Y|Z}(Y=1|Z=z)=1/2$ for all $z\in\mathcal{Z}$ and $P_{X|YZ}(\cdot|Y=y,Z=z)\sim N(\mu_{yz},I_2)$ where

$$\mu_{yz}=\begin{cases}(r,r), & \text{for } y=0, z\in\mathcal{Z}_o\\ (-r,-r), & \text{for } y=1, z\in\mathcal{Z}_o\\ (r,-r), & \text{for } y=0, z\in\mathcal{Z}_e\\ (-r,r), & \text{for } y=1, z\in\mathcal{Z}_e\end{cases} \tag{26}$$

- For all other $z\geq 1$, the problem Q_z is exactly the same as Q_0 except for $P_{X|YZ}(\cdot|Y=y,Z=z)$ for $y\in\{0,1\}$. In particular, $\mu_{oz}=(r',r')$, $\mu_{1z}=(-r',-r')$ if $z\in\mathcal{Z}_o$ and $\mu_{0z}=(r',-r')$ and $\mu_{1z}=(-r',r')$ if $z\in\mathcal{Z}_e$.

Now, as in the $|\mathcal{Z}|=2$ case, for any $i\in\mathcal{Z}$, define $a=\sum_{z\in\mathcal{Z}_i}\pi(z):w^*_\pi=(1,0)$, and note that since $r'>r$, we must have $a>1/2$. Introduce the notation E_i to denote the event $\{\sum_{z\in\mathcal{Z}_i}\pi_n(z)>a\}$, where π_n denotes the empirical mixture distribution constructed by a given algorithm $\mathcal{A}$ after n samples. Recall that for any $i\in\mathcal{Z}$, the set $\mathcal{Z}_i$ is equal to $\mathcal{Z}_o$ or $\mathcal{Z}_e$ depending on whether i is even or odd.

Then, proceeding as before, we have

$$\begin{aligned}
\max_{0\leq i\leq m}\mathbb{E}_{Q_i}\left[\mathcal{R}_n(\mathcal{A})\right]&\geq\frac{1}{m}\sum_{i=1}^m\mathbb{E}_{Q_i}\left[\mathcal{R}_n\left(\mathcal{A}\right)\right]\overset{(i)}{\geq}\frac{c_r(r'-r)}{2\sqrt{2}}\frac{1}{m}\sum_{i=1}^m\mathbb{P}_{Q_i}\left(E_i^c\right)\\
&\geq\frac{c_r(r'-r)}{2\sqrt{2}}\frac{1}{m}\sum_{i=1}^m\left(1-\mathbb{P}_{Q_i}\left(E_i\right)\right)\\
&\overset{(ii)}{\geq}\frac{c_r(r'-r)}{2\sqrt{2}}\frac{1}{m}\sum_{i=1}^m\left(1-\mathbb{P}_{Q_0}\left(E_i\right)-d_{TV}\left(\mathbb{P}_{Q_i},\mathbb{P}_{Q_0}\right)\right)\\
&\overset{(iii)}{\geq}\frac{c_r(r'-r)}{2\sqrt{2}}\left(1-\frac{1}{m}\left(\frac{m}{2}+\sum_{i=1}^m(r'-r)\sqrt{\frac{\mathbb{E}_{Q_0}[N_{i,n}]}{2}}\right)\right)\\
&\overset{(iv)}{\geq}\frac{c_r(r'-r)}{2\sqrt{2}}\left(\frac{1}{2}-\frac{1}{m}\left((r'-r)\sqrt{\frac{nm}{2}}\right)\right)\\
&=\frac{c_r(r'-r)}{2\sqrt{2}}\left(\frac{1}{2}-(r'-r)\sqrt{\frac{n}{2m}}\right)\\
&\overset{(v)}{=}\Omega\left(\sqrt{\frac{m}{n}}\right).
\end{aligned}$$

In the above display,
(i) follows from the definition of event E_i,
(ii) uses the fact that $|\mathbb{P}_{Q_0}(E_i)-\mathbb{P}_{Q_i}(E_i)|\leq d_{TV}(\mathbb{P}_{Q_0},\mathbb{P}_{Q_i})$,
(iii) uses Pinsker's inequality to bound the total-variation distance,
(iv) uses Cauchy-Schwarz inequality along with the fact that $\sum_{i=1}^m N_{i,n}=n$,
(v) follows by setting $r'=r+\sqrt{m/2n}$.

E Details of Experiments

E.1 Synthetic Datasets

Data. We used the two synthetic models introduced in Section 2.2 for generating the training set. Here we provide a formal definition for `SyntheticModel2`, an instance of which is illustrated in Figure 7.

Definition 4 (`SyntheticModel2`)**.** Set $\mathcal{X}=\mathbb{R}^2$, $\mathcal{Y}=\{0,1\}$, and $\mathcal{Z}=\{u,v\}$ and fix an angle θ and a constant $a\in\mathbb{R}$. Then $P_{X|Y=y,Z=u}=\mathcal{U}(c_y,c_y+a)$. That is, within attribute $Z=u$ and for either

label $Y = y$, X is uniformly distributed along one dimension. To get the conditional distributions for $Z = v$ we draw from mixtures of three Gaussians: $\widetilde{P}_{X=x|Y=y,Z=v} = \sum_{i=1}^{3} \rho_{i,y} P_{i,y}(x)$ where each $P_{i,y} = \mathcal{N}(\mu_{i,y}, \Sigma_{i,y})$ and $\rho_{i,y}$ are non-negative and sum to one over i. The resulting sample is then uniformly rotated counterclockwise by θ in the plane.

To elaborate on the intuition provided for this model in the main text, our goal here is to design a problem where one attribute, here $Z = u$, has a low maximum expected accuracy, for a linear classifier, that can be attained with relatively few samples. Here, any linear classifier will misclassify at least half of all points in the overlap of the intervals $(c_0, c_0 + a)$ and $(c_1, c_1 + a)$. And any linear classifier which intercepts this overlap will attain the maximal possible expected accuracy on $Z = u$. Then for $Z = v$ we choose the $\rho_{i,y}$ and covariances such that most mass is contained in easily separable clusters, so that a classifier can attain high accuracy on $Z = v$ with few samples. But, the sparser clusters are chosen to have some overlap between $Y = 1$ and $Y = 0$. The overlap and sparseness require a learner draw more samples from $Z = v$ to accurately learn the optimal boundary.

Thus $\mathcal{A}_{\texttt{opt}}$ and `Greedy` will continually sample from $Z = u$, as any linear classifier will have relatively low accuracy on it, despite the fact that additional samples do not improve performance on this attribute. In contrast, $\mathcal{A}_{\texttt{opt}}+$ identifies this stagnation in performance and samples from $Z = v$, over time drawing sufficient samples from the sparse clusters to maximize accuracy on both attributes.

Classifier. We set $\mathcal{F}$ to the set of logistic regression classifiers. More specifically we used the `LogisticRegression` implementation in the `scikit-learn` package.

Details of experiments. For the experiments on `SyntheticModel1`, pictured in Figure 2, for each trial we set $n = 1000$ and tracked the $\pi_t(u)$ values for the three algorithms. We repeated the experiment for 100 trials and plot the resulting mean $\pi_t(u)$ values in the curves, along with the one-standard-deviation regions.

For experiments on `SyntheticModel2`, pictured in Figure 6a and Figure 6b, we used the heuristic algorithm, $\mathcal{A}_{\texttt{opt}}+$, with $c_1 = 0.1$ and the Mann-Kendall statistic tracking the accuracy after the previous 20 sample draws for each attribute. We ran 100 trials for each scheme and report the mean accuracy over both attributes, along with one-standard-deviation regions.

E.2 Real Datasets

We used the following datasets:

- *UTKFace* dataset. This dataset consists of large number of face images with annotations of age, gender and ethnicity. We used $Y = \{$`Male, Female`$\}$ and set $\mathcal{Z} = \{$`White, Black, Asian, Indian, Other`$\}$.
- *FashionMNIST* dataset. This dataset consists of 10 different classes, which were paired off to get five different binary classification tasks: {`(Tshirt, Shirt), (Trousers, Dress), (Pullover, Coat), (Sandals, Bag), (Sneakers, AnkleBoots)`}
- *Cifar10* dataset. This dataset also consists of 10 different classes, which were paired off as follows: { `(airplane, ship), (automobile, truck), (bird, cat), (deer, dog), (frog, horse)`}.
- *Adult* dataset. This is a low dimensional, binary classification problem where the inputs are a variety of demographic data about an individual and the task is to predict whether the individual makes more or less than $\$50,000$/year.
- *German* dataset. This is another low dimensional, binary classification problem where the inputs are demographic and financial information about an individual and the task is to categorize them as low or high risk for defaulting on a loan.

Data Transforms and Augmentation. We used the following pre-processing operations for the different datasets:

- *UTKFace.* For this dataset, we first resized the dataset to size $3 \times 50 \times 50$ from the original $3 \times 200 \times 200$. Then we also applied random horizontal flip with probability 0.5. For training we used the images from the file `UTKFace.tar.gz`, while for testing we used the file `crop_part1.tar.gz`. (Both of these files can be found at this link shared by the owner of the dataset).

- *FashionMNIST.* We only employed the normalization transform in this case, and used the default training and test splits.
- *Cifar10.* We employed a random crop transform (to size $3 \times 28 \times 28$ with padding 1), a random horizontal flip transform (with probability 0.5) and a normalization transform, and used the default training and test split.
- *Adult.* For the categorical inputs we used a one-hot encoding and normalized all numerical inputs to $[0, 1]$. We used the provided training/test split.
- *German* We used the same pre-processing as for the *Adult* dataset here. As described in the main paper, this dataset only provides 1000 examples and does not have a canonical train/test split. Moreover, results were very sensitive to the choice of split. So for every experiment we generated a new, random 70/30 training/test split and report results averaged over 500 such trials.

Details of Classifiers. We implemented the CNN using Pytorch. The CNN used for both *Cifar10* and *UTKFace* had the same architecture consisting of:

- First `Conv2d` layer with 32 output channels, kernel size 3, padding = 1, followed by `ReLU` followed by a `MaxPool2d` layer with kernel size 2 and stride 2.
- Second `Conv2d` layer with 64 output channels, kernel size 3, followed by `ReLU` followed by `MaxPool2d` with kernel size 2.
- The two `Conv2d` layers were followed by 3 fully connected layers with output sizes 600, 120 and 2 respectively.
- For training the neural network we used the Adam optimizer with `lr` $= 0.001$.

The CNN used for *FashionMNIST* was identical, except with the second convolutional layer omitted, due to the simpler nature of the dataset.

For both *Adult* and *German* datasets we again used the `LogisticRegression` implementation in the `scikit-learn` package.

Computing Infrastructure used. The image dataset experiments were run on Google Colab using the free GPU instances and on a shared computing cluster which provides a variety of different GPUs. So we cannot provide the exact details of the GPUs we used, but all experiments in the paper can comfortably run in less than day on a GTX 1080 TI.

E.3 Additional Experimental Results

Here we present results for experiments on additional datasets, these are all analogous to results presented in the main paper on other datasets.

First we have the *German* dataset, which is from the UCI Repository (Dua and Graff, 2017) and qualitatively similar to the *Adult* dataset. We again use a LR classifier and the exact implementation of each algorithm. We set $\mathcal{Z}$ to be male or female. The results, in Figure 10a, show, on average, an advantage for $\mathcal{A}_{\text{opt}}$ over both `Uniform` and `Greedy` schemes. But this dataset contains only 1000 examples in total. We generate a 70/30 training/test split, but find that our results strongly depend on

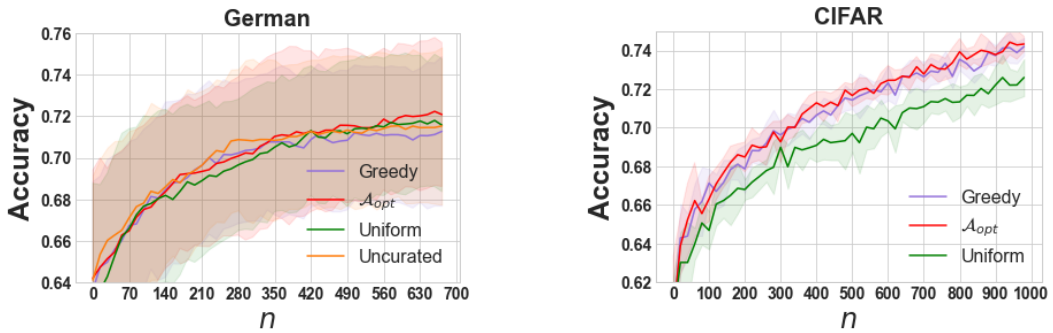

Figure 10: Left: Minimum test accuracy over all attributes for *German* dataset as a function of the sampling budget n, averaged over 500 trials. Right: Minimax error as a function of training round for *Cifar10*.

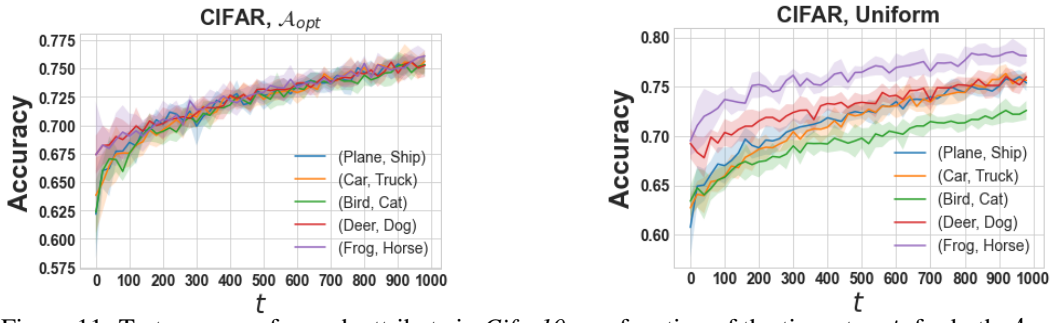

Figure 11: Test accuracy for each attribute in *Cifar10* as a function of the time step, t, for both $\mathcal{A}_{\texttt{opt}}$ and `Uniform` sampling schemes, averaged over 10 trials.

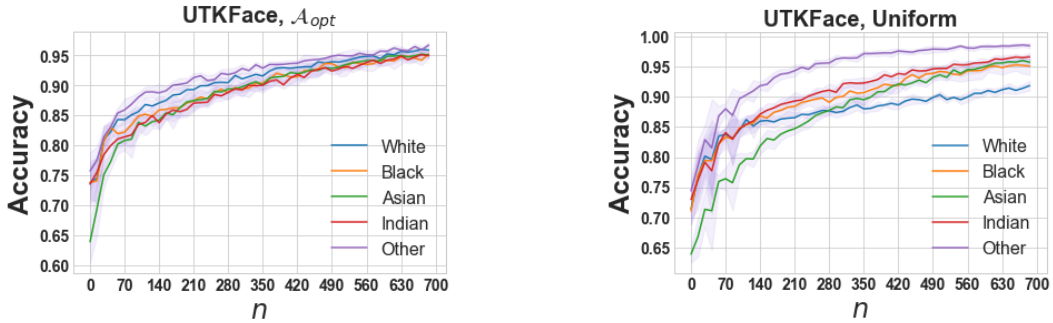

Figure 12: Test accuracy for each attribute in *UTKFace* as a function of the time step, t, for both $\mathcal{A}_{\texttt{opt}}$ and `Uniform` sampling schemes, averaged over 10 trials.

this split. To mitigate this, we run 500 trials and generate a new random split for each, but still find very high variance in our results, indicated by the shaded region in the figure.

Figure 10b shows the minimax error on *Cifar10*. This again demonstrates a significant improvement in worst case accuracy for the adaptive schemes over the `Uniform` scheme, which is equivalent to `Uncurated` for this dataset. Figures 11 and 12 show accuracy across all attributes as a function of training round for *Cifar10* and *UTKFace* for both $\mathcal{A}_{\texttt{opt}}$ and `Uniform` schemes, as in the analogous results for *FashionMNIST* this demonstrates the utility of $\mathcal{A}_{\texttt{opt}}$ for improving fairness between groups and lends credence to our assumption of the existence of an equalizing sampling distribution for real world.

Finally Table 1 summarizes the final test set accuracy, with standard deviation ranges, for each dataset and sampling scheme at the end of their respective training periods. As in other results, the adaptive schemes show an advantage over `Uniform` in all cases, and the efficacy of the `Uncurated` scheme depends on the nature of the dataset and chosen attributes.

Dataset	$\mathcal{A}_{opt}$	`Greedy`	`Uniform`	`Uncurated`
UTKFace	0.946 ± 0.003	0.946 ± 0.012	0.919 ± 0.008	0.933 ± 0.007
FashionMNIST	0.936 ± 0.004	0.924 ± 0.010	0.893 ± 0.002	0.893 ± 0.002
Cifar10	0.743 ± 0.004	0.742 ± 0.005	0.726 ± 0.010	0.726 ± 0.010
Adult	0.801 ± 0.001	0.800 ± 0.002	0.798 ± 0.002	0.797 ± 0.003
German	0.721 ± 0.035	0.713 ± 0.036	0.716 ± 0.032	0.715 ± 0.038

Table 1: Minimum test accuracy over different attributes achieved by the classifiers returned by the 4 sampling schemes.