Supplementary Material

Here we provide complete proofs of our hardness results stated in the main body of the paper. We use a technique from (Kearns and Li, 1993) called the *method of induced distributions*. The idea is to construct two distributions that are sufficiently different, yet can be made indistinguishable by the adversary. Then no learner can "guess" the underlying distribution with high probability and so any learner will incur high loss and/or exhibit high unfairness on at least one of the two distributions, regardless of the amount of available data. The proofs of the four results use the same technique and are structured in a similar way, with the key challenge being to design the corresponding constructions of the learning problem, that is, of the hypothesis space, the distributions and the adversaries. These constructions are in each case tailored to the fairness measure and the type of bound we want to show.

Appendix A. Pareto lower bounds proofs

Theorem 1 *Let $0 \leq \alpha < 0.5, 0 < P_0 \leq 0.5$. For any input set $\mathcal{X}$ with at least four distinct points, there exists a finite hypothesis space $\mathcal{H}$, such that for any learning algorithm $\mathcal{L} : \cup_{n\in\mathbb{N}}(\mathcal{X} \times A \times \mathcal{Y})^n \to \mathcal{H}$, there exists a distribution $\mathbb{P}$ for which $\mathbb{P}(A = 0) = P_0$, a malicious adversary $\mathcal{A}$ of power α and a hypothesis $h^* \in \mathcal{H}$, such that with probability at least* 0.5

$$L(\mathcal{L}(S^p), \mathbb{P}) - L(h^*, \mathbb{P}) \geq \min\left\{\frac{\alpha}{1-\alpha}, 2P_0(1-P_0)\right\}$$

and

$$\mathcal{D}^{par}(\mathcal{L}(S^p), \mathbb{P}) - \mathcal{D}^{par}(h^*, \mathbb{P}) \geq \min\left\{\frac{\alpha}{2P_0(1-P_0)(1-\alpha)}, 1\right\} \geq \min\left\{\frac{\alpha}{2P_0}, 1\right\}.$$

Proof Let $\eta = \frac{\alpha}{1-\alpha}$, so that $\alpha = \frac{\eta}{1+\eta}$.

Case 1 First assume that $\eta = \frac{\alpha}{1-\alpha} \leq 2P_0(1-P_0)$. Take 4 distinct points $\{x_1, x_2, x_3, x_4\} \in \mathcal{X}$. We consider 2 distributions $\mathbb{P}_0$ and $\mathbb{P}_1$, where each $\mathbb{P}_i$ is defined as

$$\mathbb{P}_i(x, a, y) = \begin{cases} 1 - P_0 - \eta/2 & \text{if } x = x_1, a = 1, y = 1 \\ P_0 - \eta/2 & \text{if } x = x_2, a = 0, y = 0 \\ \eta/2 & \text{if } x = x_3, a = i, y = \neg i \\ \eta/2 & \text{if } x = x_4, a = \neg i, y = i \\ 0 & \text{otherwise} \end{cases}$$

Here we use the $\neg$ notation to denote negation, so that $\neg i = 1$ if $i = 0$ and $\neg i = 0$ if $i = 1$. Note that these are valid distributions, since $\eta \leq 2P_0(1-P_0) \leq 2P_0 \leq 2(1-P_0)$ by assumption and also that $P_0 = \mathbb{P}_i(A = 0)$ for both $i \in \{0, 1\}$. Consider the hypothesis space $\mathcal{H} = \{h_0, h_1\}$, with

$$h_0(x_1) = 1 \quad h_0(x_2) = 0 \quad h_0(x_3) = 1 \quad h_0(x_4) = 0$$

and

$$h_1(x_1) = 1 \quad h_1(x_2) = 0 \quad h_1(x_3) = 0 \quad h_1(x_4) = 1.$$

Note that $L(h_i, \mathbb{P}_i) = 0$ for both $i = 0, 1$. Moreover,

$$\begin{aligned}
\mathcal{D}^{par}(h_0, \mathbb{P}_0) &= \left|\mathbb{P}_{(X,A,Y)\sim\mathbb{P}_0}(h_0(X) = 1|A = 0) - \mathbb{P}_{(X,A,Y)\sim\mathbb{P}_0}(h_0(X) = 1|A = 1)\right| \\
&= \left|\frac{\eta/2}{P_0 - \eta/2 + \eta/2} - \frac{1 - P_0 - \eta/2}{1 - P_0 - \eta/2 + \eta/2}\right| \\
&= \left|\frac{\eta}{2P_0} - \frac{2 - 2P_0 - \eta}{2(1 - P_0)}\right| \\
&= \left|\frac{\eta}{2P_0(1 - P_0)} - 1\right| \\
&= 1 - \frac{\eta}{2P_0(1 - P_0)},
\end{aligned}$$

since $\eta \leq 2P_0(1 - P_0)$ by assumption. Furthermore,

$$\begin{aligned}
\mathcal{D}^{par}(h_1, \mathbb{P}_0) &= \left|\mathbb{P}_{(X,A,Y)\sim\mathbb{P}_0}(h_1(X) = 1|A = 0) - \mathbb{P}_{(X,A,Y)\sim\mathbb{P}_0}(h_1(X) = 1|A = 1)\right| \\
&= |0 - 1| \\
&= 1
\end{aligned}$$

Therefore, $\mathcal{D}^{par}(h_1, \mathbb{P}_0) - \mathcal{D}^{par}(h_0, \mathbb{P}_0) = \frac{\eta}{2P_0(1-P_0)}$. Similarly,

$$\begin{aligned}
\mathcal{D}^{par}(h_1, \mathbb{P}_1) &= \left|\mathbb{P}_{(X,A,Y)\sim\mathbb{P}_1}(h_1(X) = 1|A = 0) - \mathbb{P}_{(X,A,Y)\sim\mathbb{P}_1}(h_1(X) = 1|A = 1)\right| \\
&= \left|\frac{\eta/2}{P_0 - \eta/2 + \eta/2} - \frac{1 - P_0 - \eta/2}{1 - P_0 - \eta/2 + \eta/2}\right| \\
&= 1 - \frac{\eta}{2P_0(1 - P_0)}
\end{aligned}$$

and

$$\begin{aligned}
\mathcal{D}^{par}(h_0, \mathbb{P}_1) &= \left|\mathbb{P}_{(X,A,Y)\sim\mathbb{P}_1}(h_0(X) = 1|A = 0) - \mathbb{P}_{(X,A,Y)\sim\mathbb{P}_1}(h_0(X) = 1|A = 1)\right| \\
&= |0 - 1| \\
&= 1,
\end{aligned}$$

so that $\mathcal{D}^{par}(h_0, \mathbb{P}_1) - \mathcal{D}^{par}(h_1, \mathbb{P}_1) = \frac{\eta}{2P_0(1-P_0)}$.

Consider a (randomized) malicious adversary $\mathcal{A}_i$ of power α, that given a clean distribution $\mathbb{P}_i$, changes every marked point to $(x_3, \neg i, i)$ with probability 0.5 and to $(x_4, i, \neg i)$ otherwise. Under a distribution $\mathbb{P}_i$ and an adversary $\mathcal{A}_i$, the probability of seeing a point $(x_3, i, \neg i)$ is $\frac{\eta}{2}(1 - \alpha) = \frac{\eta}{2}\frac{1}{1+\eta} = \alpha/2$, which is equal to the probability of seeing a point $(x_3, \neg i, i)$. Therefore, denoting the probability distribution of the corrupted dataset, under a

clean distribution $\mathbb{P}_i$ and an adversary $\mathcal{A}_i$, by $\mathbb{P}'_i$ (as a shorthand for $\mathbb{P}_i^{\mathcal{A}_i}$), we have

$$\mathbb{P}'_i(x,a,y) = \begin{cases} (1-\alpha)(1-P_0-\eta/2) & \text{if } x = x_1, a = 1, y = 1 \\ (1-\alpha)(P_0-\eta/2) & \text{if } x = x_2, a = 0, y = 0 \\ \alpha/2 & \text{if } x = x_3, a = i, y = \neg i \\ \alpha/2 & \text{if } x = x_3, a = \neg i, y = i \\ \alpha/2 & \text{if } x = x_4, a = \neg i, y = i \\ \alpha/2 & \text{if } x = x_4, a = i, y = \neg i \\ 0 & \text{otherwise} \end{cases}$$

In particular, $\mathbb{P}'_0 = \mathbb{P}'_1$, so the two initial distributions $\mathbb{P}_0$ and $\mathbb{P}_1$ become indistinguishable under the adversarial manipulation.

Fix an arbitrary learner $\mathcal{L} : \cup_{n\in\mathbb{N}}(\mathcal{X} \times A \times \mathcal{Y})^n \to \{h_0, h_1\}$. Note that, if the clean distribution is $\mathbb{P}_0$, the events (in the probability space defined by the sampling of the poisoned train data)

$$\begin{aligned}\{L(\mathcal{L}(S^p), \mathbb{P}_0) - L(h_0, \mathbb{P}_0) \geq \eta\} &= \{\mathcal{L}(S^p) = h_1\} \\ &= \left\{\mathcal{D}^{par}(\mathcal{L}(S^p), \mathbb{P}_0) - \mathcal{D}^{par}(h_0, \mathbb{P}_0) \geq \frac{\eta}{2P_0(1-P_0)}\right\}\end{aligned}$$

are all the same. Similarly, if the clean distribution is $\mathbb{P}_1$

$$\begin{aligned}\{L(\mathcal{L}(S^p), \mathbb{P}_1) - L(h_1, \mathbb{P}_1) \geq \eta\} &= \{\mathcal{L}(S^p) = h_0\} \\ &= \left\{\mathcal{D}^{par}(\mathcal{L}(S^p), \mathbb{P}_1) - \mathcal{D}^{par}(h_1, \mathbb{P}_1) \geq \frac{\eta}{2P_0(1-P_0)}\right\}.\end{aligned}$$

Therefore, depending on whether we choose $\mathbb{P}_0$ or $\mathbb{P}_1$ as a clean distribution, we have

$$\begin{aligned}&\mathbb{P}_{S^p\sim\mathbb{P}'_0}\left((L(\mathcal{L}(S^p), \mathbb{P}_0) - L(h_0, \mathbb{P}_0) \geq \eta) \wedge \left(\mathcal{D}^{par}(\mathcal{L}(S^p), \mathbb{P}_0) - \mathcal{D}^{par}(h_0, \mathbb{P}_0) \geq \frac{\eta}{2P_0(1-P_0)}\right)\right) \\ &\quad = \mathbb{P}_{S^p\sim\mathbb{P}'_0}\left(\mathcal{L}(S^p) = h_1\right)\end{aligned}$$

and

$$\begin{aligned}&\mathbb{P}_{S^p\sim\mathbb{P}'_1}\left((L(\mathcal{L}(S^p), \mathbb{P}_1) - L(h_1, \mathbb{P}_1) \geq \eta) \wedge \left(\mathcal{D}^{par}(\mathcal{L}(S^p), \mathbb{P}_1) - \mathcal{D}^{par}(h_1, \mathbb{P}_1) \geq \frac{\eta}{2P_0(1-P_0)}\right)\right) \\ &\quad = \mathbb{P}_{S^p\sim\mathbb{P}'_1}\left(\mathcal{L}(S^p) = h_0\right)\end{aligned}$$

Finally, note that $\mathbb{P}'_0 = \mathbb{P}'_1$, so that either $\mathbb{P}_{S^p\sim\mathbb{P}'_0}\left(\mathcal{L}(S^p) = h_1\right) \geq 1/2$ or $\mathbb{P}_{S^p\sim\mathbb{P}'_1}\left(\mathcal{L}(S^p) = h_0\right) \geq 1/2$. Therefore, for at least one of $i = 0, 1$, both

$$L(\mathcal{L}(S^p), \mathbb{P}_i) - L(h_i, \mathbb{P}_i) \geq \eta = \frac{\alpha}{1-\alpha}$$

and

$$\mathcal{D}^{par}(\mathcal{L}(S^p), \mathbb{P}_i) - \mathcal{D}^{par}(h_i, \mathbb{P}_i) \geq \frac{\eta}{2P_0(1-P_0)} = \frac{\alpha}{2P_0(1-P_0)(1-\alpha)}$$

both hold with probability at least $1/2$ when the choice of distribution and adversary is $\mathbb{P}_i$ and $\mathcal{A}_i$ respectively. This concludes the proof in the first case.

Case 2 Now suppose that $\eta = \frac{\alpha}{1-\alpha} > 2P_0(1-P_0)$. Let $\alpha_1 \in (0, 0.5)$ be such that $\frac{\alpha_1}{1-\alpha_1} = 2P_0(1-P_0)$. Note that since $f(x) = \frac{x}{1-x}$ is monotonically increasing in (0,1), α_1 is unique and $\alpha_1 < \alpha$.

Now repeat the same construction as in Case 1, but with $\eta_1 = \frac{\alpha_1}{1-\alpha_1} = 2P_0(1-P_0)$. For every marked point, the adversary does the same as in Case 1 with probability α_1/α and does not change the point otherwise. Then the same argument as in Case 1 shows that for one $i \in \{0,1\}$, both

$$L(\mathcal{L}(S^p), \mathbb{P}_i) - L(h_i, \mathbb{P}_i) \geq \eta_1 = \frac{\alpha_1}{1-\alpha_1} = 2P_0(1-P_0)$$

and

$$\mathcal{D}^{par}(\mathcal{L}(S^p), \mathbb{P}_i) - \mathcal{D}^{par}(h_i, \mathbb{P}_i) \geq \frac{\eta_1}{2P_0(1-P_0)} = 1$$

both hold with probability at least 1/2. This concludes the proof of Theorem 1. ■

Theorem 2 *Let $0 \leq \alpha < 0.5, P_{10} \leq P_{11} < 1$ be such that $P_{10} + P_{11} < 1$. For any input set $\mathcal{X}$ with at least five distinct points, there exists a finite hypothesis space $\mathcal{H}$, such that for any learning algorithm $\mathcal{L} : \cup_{n\in\mathbb{N}}(\mathcal{X} \times A \times \mathcal{Y})^n \to \mathcal{H}$, there exists a distribution $\mathbb{P}$ for which $\mathbb{P}(A = a, Y = 1) = P_{1a}$ for $a \in \{0,1\}$, a malicious adversary $\mathcal{A}$ of power α and a hypothesis $h^* \in \mathcal{H}$, such that with probability at least* 0.5

$$L(\mathcal{L}(S^p), \mathbb{P}) - L(h^*, \mathbb{P}) > \min\left\{\frac{\alpha}{1-\alpha}, 2P_{10}, 2(1-P_{10}-P_{11})\right\}$$

and

$$\mathcal{D}^{opp}(\mathcal{L}(S^p), \mathbb{P}) - \mathcal{D}^{opp}(h^*, \mathbb{P}) \geq \min\left\{\frac{\alpha}{2(1-\alpha)P_{10}}, 1, \frac{1-P_{10}-P_{11}}{P_{10}}\right\}.$$

Proof Let $\eta = \frac{\alpha}{1-\alpha}$, so that $\alpha = \frac{\eta}{1+\eta}$.

Case 1 Assume that $\eta = \frac{\alpha}{1-\alpha} \leq 2\min\{P_{10}, 1-P_{10}-P_{11}\}$. Take 5 distinct points $\{x_1, x_2, x_3, x_4, x_5\} \in \mathcal{X}$. We consider 2 distributions $\mathbb{P}_0$ and $\mathbb{P}_1$, where each $\mathbb{P}_i$ is defined as

$$\mathbb{P}_i(x,a,y) = \begin{cases} P_{11} & \text{if } x = x_1, a = 1, y = 1 \\ P_{10} - \eta/2 & \text{if } x = x_2, a = 0, y = 1 \\ \eta/2 & \text{if } x = x_3, a = i, y = \neg i \\ \eta/2 & \text{if } x = x_4, a = \neg i, y = i \\ 1 - P_{10} - P_{11} - \eta/2 & \text{if } x = x_5, a = 0, y = 0 \\ 0 & \text{otherwise} \end{cases}$$

Note that these are valid distributions, since $\eta \leq 2P_{10}, \eta \leq 2(1-P_{10}-P_{11})$ by assumption, and that $P_{1a} = \mathbb{P}_i(A = a, Y = 1)$ for both $a \in \{0,1\}, i \in \{0,1\}$. Consider the hypothesis space $\mathcal{H} = \{h_0, h_1\}$, with

$$h_0(x_1) = 1 \quad h_0(x_2) = 1 \quad h_0(x_3) = 1 \quad h_0(x_4) = 0 \quad h_0(x_5) = 0$$

and

$$h_1(x_1) = 1 \quad h_1(x_2) = 1 \quad h_1(x_3) = 0 \quad h_1(x_4) = 1 \quad h_1(x_5) = 0$$

Note that $L(h_i, \mathbb{P}_i) = 0$ and $\mathcal{D}^{opp}(h_i, \mathbb{P}_i) = 0$ for both $i = 0, 1$. Note also that $L(h_1, \mathbb{P}_0) = L(h_0, \mathbb{P}_1) = \eta$. Moreover,

$$\begin{aligned}\mathcal{D}^{opp}(h_1, \mathbb{P}_0) &= \left|\mathbb{P}_{(X,A,Y)\sim\mathbb{P}_0}(h_1(X) = 1|A = 0, Y = 1) - \mathbb{P}_{(X,A,Y)\sim\mathbb{P}_0}(h_1(X) = 1|A = 1, Y = 1)\right| \\ &= \left|\frac{P_{10} - \eta/2}{P_{10} - \eta/2 + \eta/2} - 1\right| \\ &= \frac{\eta}{2P_{10}}\end{aligned}$$

and similarly $\mathcal{D}^{opp}(h_0, \mathbb{P}_1) = \frac{\eta}{2P_{10}}$.

Consider a (randomized) malicious adversary $\mathcal{A}_i$ of power α, that given a clean distribution $\mathbb{P}_i$, changes every marked point to $(x_3, \neg i, i)$ with probability 0.5 and to $(x_4, i, \neg i)$ otherwise. Under a distribution $\mathbb{P}_i$ and an adversary $\mathcal{A}_i$, the probability of seeing a point $(x_3, i, \neg i)$ is $\frac{\eta}{2}(1 - \alpha) = \frac{\eta}{2}\frac{1}{1+\eta} = \alpha/2$, which is equal to the probability of seeing a point $(x_3, \neg i, i)$. Therefore, denoting the probability distribution of the corrupted dataset, under a clean distribution $\mathbb{P}_i$ and an adversary $\mathcal{A}_i$, by $\mathbb{P}'_i$, we have

$$\mathbb{P}'_i(x, a, y) = \begin{cases} (1 - \alpha)P_{11} & \text{if } x = x_1, a = 1, y = 1 \\ (1 - \alpha)(P_{10} - \eta/2) & \text{if } x = x_2, a = 0, y = 1 \\ \alpha/2 & \text{if } x = x_3, a = i, y = \neg i \\ \alpha/2 & \text{if } x = x_3, a = \neg i, y = i \\ \alpha/2 & \text{if } x = x_4, a = \neg i, y = i \\ \alpha/2 & \text{if } x = x_4, a = i, y = \neg i \\ (1 - \alpha)(1 - P_{10} - P_{11} - \eta/2) & \text{if } x = x_5, a = 0, y = 0 \\ 0 & \text{otherwise} \end{cases}$$

In particular, $\mathbb{P}'_0 = \mathbb{P}'_1$, so the two initial distributions $\mathbb{P}_0$ and $\mathbb{P}_1$ become indistinguishable under the adversarial manipulation.

Fix an arbitrary learner $\mathcal{L} : \cup_{n\in\mathbb{N}}(\mathcal{X} \times A \times \mathcal{Y})^n \to \{h_0, h_1\}$. Note that, if the clean distribution is $\mathbb{P}_0$, the events (in the probability space defined by the sampling of the poisoned train data)

$$\{L(\mathcal{L}(S^p), \mathbb{P}_0) - L(h_0, \mathbb{P}_0) \geq \eta\} = \{\mathcal{L}(S^p) = h_1\} = \left\{\mathcal{D}^{opp}(\mathcal{L}(S^p), \mathbb{P}_0) - \mathcal{D}^{opp}(h_0, \mathbb{P}_0) \geq \frac{\eta}{2P_{10}}\right\}$$

are all the same. Similarly, if the clean distribution is $\mathbb{P}_1$

$$\{L(\mathcal{L}(S^p), \mathbb{P}_1) - L(h_1, \mathbb{P}_1) \geq \eta\} = \{\mathcal{L}(S^p) = h_0\} = \left\{\mathcal{D}^{opp}(\mathcal{L}(S^p), \mathbb{P}_1) - \mathcal{D}^{opp}(h_1, \mathbb{P}_1) \geq \frac{\eta}{2P_{10}}\right\}.$$

Therefore, depending on whether we choose $\mathbb{P}_0$ or $\mathbb{P}_1$ as a clean distribution, we have

$$\mathbb{P}_{S^p\sim\mathbb{P}'_0}\left(L(\mathcal{L}(S^p), \mathbb{P}_0) - L(h_0, \mathbb{P}_0) \geq \eta \wedge \mathcal{D}^{opp}(\mathcal{L}(S^p), \mathbb{P}_0) - \mathcal{D}^{opp}(h_0, \mathbb{P}_0) \geq \frac{\eta}{2P_{10}}\right)$$

$$= \mathbb{P}_{S^p \sim \mathbb{P}_0'} \left(\mathcal{L}(S^p) = h_1 \right)$$

and

$$\mathbb{P}_{S^p \sim \mathbb{P}_1'} \left(L(\mathcal{L}(S^p), \mathbb{P}_1) - L(h_1, \mathbb{P}_1) \geq \eta \wedge \mathcal{D}^{opp}(\mathcal{L}(S^p), \mathbb{P}_1) - \mathcal{D}^{opp}(h_1, \mathbb{P}_1) \geq \frac{\eta}{2P_{10}} \right)$$
$$= \mathbb{P}_{S^p \sim \mathbb{P}_1'} \left(\mathcal{L}(S^p) = h_0 \right)$$

Finally, note that $\mathbb{P}_0' = \mathbb{P}_1'$, so that either $\mathbb{P}_{S^p \sim \mathbb{P}_0'} \left(\mathcal{L}(S^p) = h_1 \right) \geq 1/2$ or $\mathbb{P}_{S^p \sim \mathbb{P}_1'} \left(\mathcal{L}(S^p) = h_0 \right) \geq 1/2$. Therefore, for at least one of $i = 0, 1$, both

$$L(\mathcal{L}(S^p), \mathbb{P}_i) - L(h_i, \mathbb{P}_i) \geq \eta = \frac{\alpha}{1-\alpha}$$

and

$$\mathcal{D}^{opp}(\mathcal{L}(S^p), \mathbb{P}_i) - \mathcal{D}^{opp}(h_i, \mathbb{P}_i) \geq \frac{\eta}{2P_{10}} = \frac{\alpha}{2P_{10}(1-\alpha)}$$

both hold with probability at least $1/2$. This concludes the proof of the first case.

Case 2 Now assume that $\frac{\alpha}{1-\alpha} > 2 \min \{P_{10}, 1 - P_{10} - P_{11}\}$. We distinguish two cases:

Case 2.1 Suppose that $P_{10} \leq 1 - P_{10} - P_{11}$. We have that $\frac{\alpha}{1-\alpha} > 2P_{10}$. Then, denote by α_1 the unique number between $(0, 0.5)$, such that $\frac{\alpha_1}{1-\alpha_1} = 2P_{10} = 2 \min \{P_{10}, 1 - P_{10} - P_{11}\}$, and note that $\alpha_1 < \alpha$. Then repeat the same construction as in Case 1, but with $\eta_1 = \frac{\alpha_1}{1-\alpha_1}$ and an adversary that with probability α_1/α does the same as in Case 1 and leaves a marked point untouched otherwise.

Then the same argument as in Case 1 gives that for some $i \in \{0, 1\}$, with probability at least 0.5, both of the following hold

$$L(\mathcal{L}(S^p), \mathbb{P}_i) - L(h_i, \mathbb{P}_i) \geq \frac{\alpha_1}{1-\alpha_1} = 2P_{10}$$

and

$$\mathcal{D}^{opp}(\mathcal{L}(S^p), \mathbb{P}_i) - \mathcal{D}^{opp}(h_i, \mathbb{P}_i) \geq \frac{\eta_1}{2P_{10}} = 1.$$

Case 2.2 In the case when $1 - P_{10} - P_{11} < P_{10}$ we have that $\frac{\alpha}{1-\alpha} > 2(1 - P_{10} - P_{11})$. Then, denote by α_2 the unique number between $(0, 0.5)$, such that $\frac{\alpha_2}{1-\alpha_2} = 2(1 - P_{10} - P_{11}) = 2 \min \{P_{10}, 1 - P_{10} - P_{11}\}$, and note that $\alpha_2 < \alpha$. Then repeat the same construction as in Case 1, but with $\eta_2 = \frac{\alpha_2}{1-\alpha_2}$ and an adversary that with probability α_2/α does the same as in Case 1 and leaves a marked point untouched otherwise.

Then the same argument as in Case 1 gives that for some $i \in \{0, 1\}$, with probability at least 0.5, both of the following hold

$$L(\mathcal{L}(S^p), \mathbb{P}_i) - L(h_i, \mathbb{P}_i) \geq \frac{\alpha_2}{1-\alpha_2} = 2(1 - P_{10} - P_{11})$$

and

$$\mathcal{D}^{opp}(\mathcal{L}(S^p), \mathbb{P}_i) - \mathcal{D}^{opp}(h_i, \mathbb{P}_i) \geq \frac{\eta_2}{2P_{10}} = \frac{1 - P_{10} - P_{11}}{P_{10}}.$$

This concludes the proof of Theorem 2. ■

Appendix B. Hurting fairness without affecting accuracy - proofs

Theorem 3 *Let $0 \leq \alpha < 0.5, 0 < P_0 \leq 0.5$. For any input set $\mathcal{X}$ with at least four distinct points, there exists a finite hypothesis space $\mathcal{H}$, such that for any learning algorithm $\mathcal{L} : \cup_{n\in\mathbb{N}}(\mathcal{X} \times A \times \mathcal{Y})^n \to \mathcal{H}$, there exists a distribution $\mathbb{P}$ for which $\mathbb{P}(A = 0) = P_0$, a malicious adversary $\mathcal{A}$ of power α and a hypothesis $h^* \in \mathcal{H}$, such that with probability at least* 0.5

$$L(\mathcal{L}(S^p), \mathbb{P}) = L(h^*, \mathbb{P}) = \min_{h\in\mathcal{H}} L(h, \mathbb{P})$$

and

$$\mathcal{D}^{par}(\mathcal{L}(S^p), \mathbb{P}) - \mathcal{D}^{par}(h^*, \mathbb{P}) \geq \min\left\{\frac{\alpha}{2P_0(1-P_0)(1-\alpha)}, 1\right\} \geq \min\left\{\frac{\alpha}{2P_0}, 1\right\}.$$

Proof Let $\eta = \frac{\alpha}{1-\alpha}$, so that $\alpha = \frac{\eta}{1+\eta}$.

Case 1 First assume that $\eta = \frac{\alpha}{1-\alpha} \leq 2P_0(1-P_0)$. Take 4 distinct points $\{x_1, x_2, x_3, x_4\} \in \mathcal{X}$. We consider 2 distributions $\mathbb{P}_0$ and $\mathbb{P}_1$, where each $\mathbb{P}_i$ is defined as

$$\mathbb{P}_i(x, a, y) = \begin{cases} 1 - P_0 - \eta/2 & \text{if } x = x_1, a = 1, y = 1 \\ P_0 - \eta/2 & \text{if } x = x_2, a = 0, y = 0 \\ \eta/2 & \text{if } x = x_3, a = i, y = 1 \\ \eta/2 & \text{if } x = x_4, a = \neg i, y = 1 \\ 0 & \text{otherwise} \end{cases}$$

Note that these are valid distributions, since $\eta \leq 2P_0(1-P_0) \leq 2P_0 \leq 2(1-P_0)$ by assumption and also that $P_0 = \mathbb{P}_i(A = 0)$ for both $i \in \{0, 1\}$. Consider the hypothesis space $\mathcal{H} = \{h_0, h_1\}$, with

$$h_0(x_1) = 1 \quad h_0(x_2) = 0 \quad h_0(x_3) = 1 \quad h_0(x_4) = 0$$

and

$$h_1(x_1) = 1 \quad h_1(x_2) = 0 \quad h_1(x_3) = 0 \quad h_1(x_4) = 1.$$

Note that $L(h_i, \mathbb{P}_i) = L(h_{\neg i}, \mathbb{P}_i) = \eta/2$ for both $i = 0, 1$. Moreover,

$$\begin{aligned} \mathcal{D}^{par}(h_0, \mathbb{P}_0) &= \left|\mathbb{P}_{(X,A,Y)\sim\mathbb{P}_0}(h_0(X) = 1|A = 0) - \mathbb{P}_{(X,A,Y)\sim\mathbb{P}_0}(h_0(X) = 1|A = 1)\right| \\ &= \left|\frac{\eta/2}{P_0 - \eta/2 + \eta/2} - \frac{1 - P_0 - \eta/2}{1 - P_0 - \eta/2 + \eta/2}\right| \\ &= \left|\frac{\eta}{2P_0} - \frac{2 - 2P_0 - \eta}{2(1-P_0)}\right| \\ &= \left|\frac{\eta}{2P_0(1-P_0)} - 1\right| \\ &= 1 - \frac{\eta}{2P_0(1-P_0)}, \end{aligned}$$

since $\eta \leq 2P_0(1-P_0)$ by assumption. Furthermore, $\mathcal{D}^{par}(h_1, \mathbb{P}_0) = 1$, so that $\mathcal{D}^{par}(h_1, \mathbb{P}_0) - \mathcal{D}^{par}(h_0, \mathbb{P}_0) = \frac{\eta}{2P_0(1-P_0)}$. Similarly,

$$\begin{aligned}
\mathcal{D}^{par}(h_1, \mathbb{P}_1) &= \left|\mathbb{P}_{(X,A,Y)\sim\mathbb{P}_1}(h_1(X) = 1|A = 0) - \mathbb{P}_{(X,A,Y)\sim\mathbb{P}_1}(h_1(X) = 1|A = 1)\right| \\
&= \left|\frac{\eta/2}{P_0 - \eta/2 + \eta/2} - \frac{1 - P_0 - \eta/2}{1 - P_0 - \eta/2 + \eta/2}\right| \\
&= 1 - \frac{\eta}{2P_0(1-P_0)}
\end{aligned}$$

and $\mathcal{D}^{par}(h_0, \mathbb{P}_1) = 1$.

Consider a (randomized) malicious adversary $\mathcal{A}_i$ of power α, that given a clean distribution $\mathbb{P}_i$, changes every marked point to $(x_3, \neg i, 1)$ with probability 0.5 and to $(x_4, i, 1)$ otherwise. Under a distribution $\mathbb{P}_i$ and an adversary $\mathcal{A}_i$, the probability of seeing a point $(x_3, i, 1)$ is $\frac{\eta}{2}(1-\alpha) = \frac{\eta}{2}\frac{1}{1+\eta} = \alpha/2$, which is equal to the probability of seeing a point $(x_3, \neg i, 1)$. Therefore, denoting the probability distribution of the corrupted dataset, under a clean distribution $\mathbb{P}_i$ and an adversary $\mathcal{A}_i$, by $\mathbb{P}'_i$, we have

$$\mathbb{P}'_i(x, a, y) = \begin{cases}
(1-\alpha)(1 - P_0 - \eta/2) & \text{if } x = x_1, a = 1, y = 1 \\
(1-\alpha)(P_0 - \eta/2) & \text{if } x = x_2, a = 0, y = 0 \\
\alpha/2 & \text{if } x = x_3, a = i, y = 1 \\
\alpha/2 & \text{if } x = x_3, a = \neg i, y = 1 \\
\alpha/2 & \text{if } x = x_4, a = \neg i, y = 1 \\
\alpha/2 & \text{if } x = x_4, a = i, y = 1 \\
0 & \text{otherwise}
\end{cases}$$

In particular, $\mathbb{P}'_0 = \mathbb{P}'_1$, so the two initial distributions $\mathbb{P}_0$ and $\mathbb{P}_1$ become indistinguishable under the adversarial manipulation.

Fix an arbitrary learner $\mathcal{L} : \cup_{n\in\mathbb{N}}(\mathcal{X} \times A \times \mathcal{Y})^n \to \{h_0, h_1\}$. Note that, if the clean distribution is $\mathbb{P}_0$, the events (in the probability space defined by the sampling of the poisoned train data)

$$\{\mathcal{L}(S^p) = h_1\} = \left\{\mathcal{D}^{par}(\mathcal{L}(S^p), \mathbb{P}_0) - \mathcal{D}^{par}(h_0, \mathbb{P}_0) \geq \frac{\eta}{2P_0(1-P_0)}\right\}$$

are all the same. Similarly, if the clean distribution is $\mathbb{P}_1$

$$\{\mathcal{L}(S^p) = h_0\} = \left\{\mathcal{D}^{par}(\mathcal{L}(S^p), \mathbb{P}_1) - \mathcal{D}^{par}(h_1, \mathbb{P}_1) \geq \frac{\eta}{2P_0(1-P_0)}\right\}.$$

Therefore, depending on whether we choose $\mathbb{P}_0$ or $\mathbb{P}_1$ as a clean distribution, we have

$$\mathbb{P}_{S^p\sim\mathbb{P}'_0}\left(\mathcal{D}^{par}(\mathcal{L}(S^p), \mathbb{P}_0) - \mathcal{D}^{par}(h_0, \mathbb{P}_0) \geq \frac{\eta}{2P_0(1-P_0)}\right) = \mathbb{P}_{S^p\sim\mathbb{P}'_0}\left(\mathcal{L}(S^p) = h_1\right)$$

and

$$\mathbb{P}_{S^p\sim\mathbb{P}'_1}\left(\mathcal{D}^{par}(\mathcal{L}(S^p), \mathbb{P}_1) - \mathcal{D}^{par}(h_1, \mathbb{P}_1) \geq \frac{\eta}{2P_0(1-P_0)}\right) = \mathbb{P}_{S^p\sim\mathbb{P}'_1}\left(\mathcal{L}(S^p) = h_0\right)$$

Finally, note that $\mathbb{P}'_0 = \mathbb{P}'_1$, so that either $\mathbb{P}_{S^p \sim \mathbb{P}'_0}(\mathcal{L}(S^p) = h_1) \geq 1/2$ or $\mathbb{P}_{S^p \sim \mathbb{P}'_1}(\mathcal{L}(S^p) = h_0) \geq 1/2$. Furthermore, $L(\mathcal{L}(S^p), \mathbb{P}_i) = \eta/2$ holds for both $i \in \{0, 1\}$, for any realization of the randomness. Therefore, for at least one of $i = 0, 1$, both

$$L(\mathcal{L}(S^p), \mathbb{P}_i) = L(h_i, \mathbb{P}_i) = \frac{\eta}{2}$$

and

$$\mathcal{D}^{par}(\mathcal{L}(S^p), \mathbb{P}_i) - \mathcal{D}^{par}(h_i, \mathbb{P}_i) \geq \frac{\eta}{2P_0(1-P_0)} = \frac{\alpha}{2P_0(1-P_0)(1-\alpha)}$$

both hold with probability at least $1/2$. This concludes the proof in the first case.

Case 2 Now suppose that $\eta = \frac{\alpha}{1-\alpha} > 2P_0(1-P_0)$. Let $\alpha_1 \in (0, 0.5)$ be such that $\frac{\alpha_1}{1-\alpha_1} = 2P_0(1-P_0)$. Note that since $f(x) = \frac{x}{1-x}$ is monotonically increasing in (0,1), α_1 is unique and $\alpha_1 < \alpha$.

Now repeat the same construction as in Case 1, but with $\eta_1 = \frac{\alpha_1}{1-\alpha_1} = 2P_0(1-P_0)$. For every marked point, the adversary does the same as in Case 1 with probability α_1/α and does not change the point otherwise. Then the same argument as in Case 1 shows that for one $i \in \{0, 1\}$, both

$$L(\mathcal{L}(S^p), \mathbb{P}_i) = L(h_i, \mathbb{P}_i) = \frac{\eta_1}{2} = P_0(1-P_0)$$

and

$$\mathcal{D}^{par}(\mathcal{L}(S^p), \mathbb{P}_i) - \mathcal{D}^{par}(h_i, \mathbb{P}_i) \geq \frac{\eta_1}{2P_0(1-P_0)} = 1$$

both hold with probability at least $1/2$. This concludes the proof of Theorem 3. ■

Theorem 4 *Let $0 \leq \alpha < 0.5, P_{10} \leq P_{11} < 1$ be such that $P_{10} + P_{11} < 1$. For any input set $\mathcal{X}$ with at least five distinct points, there exists a finite hypothesis space $\mathcal{H}$, such that for any learning algorithm $\mathcal{L} : \cup_{n \in \mathbb{N}} (\mathcal{X} \times A \times \mathcal{Y})^n \to \mathcal{H}$, there exists a distribution $\mathbb{P}$ for which $\mathbb{P}(A = a, Y = 1) = P_{1a}$ for $a \in \{0, 1\}$, a malicious adversary $\mathcal{A}$ of power α and a hypothesis $h^* \in \mathcal{H}$, such that with probability at least* 0.5

$$L(\mathcal{L}(S^p), \mathbb{P}) = L(h^*, \mathbb{P}) = \min_{h \in \mathcal{H}} L(h, \mathbb{P})$$

and

$$\mathcal{D}^{opp}(\mathcal{L}(S^p), \mathbb{P}) - \mathcal{D}^{opp}(h^*, \mathbb{P}) \geq \min\left\{ \frac{\alpha}{2(1-\alpha)P_{10}} \left(1 - \frac{P_{10}}{P_{11}}\right), 1 - \frac{P_{10}}{P_{11}} \right\}.$$

Proof Let $\eta = \frac{\alpha}{1-\alpha}$, so that $\alpha = \frac{\eta}{1+\eta}$.

Case 1 First assume that $\eta \leq 2P_{10}$. Take 5 distinct points $\{x_1, x_2, x_3, x_4, x_5\} \in \mathcal{X}$. We consider 2 distributions $\mathbb{P}_0$ and $\mathbb{P}_1$, where each $\mathbb{P}_i$ is defined as

$$\mathbb{P}_i(x, a, y) = \begin{cases} P_{11} - \eta/2 & \text{if } x = x_1, a = 1, y = 1 \\ P_{10} - \eta/2 & \text{if } x = x_2, a = 0, y = 1 \\ \eta/2 & \text{if } x = x_3, a = i, y = 1 \\ \eta/2 & \text{if } x = x_4, a = \neg i, y = 1 \\ 1 - P_{10} - P_{11} & \text{if } x = x_5, a = 0, y = 0 \\ 0 & \text{otherwise} \end{cases}$$

Note that these are valid distributions, since $\eta \leq 2P_{10} \leq 2P_{11}$ by assumption, and that $P_{1a} = \mathbb{P}_i(A = a, Y = 1)$ for both $a \in \{0, 1\}, i \in \{0, 1\}$. Consider the hypothesis space $\mathcal{H} = \{h_0, h_1\}$, with

$$h_0(x_1) = 1 \quad h_0(x_2) = 1 \quad h_0(x_3) = 1 \quad h_0(x_4) = 0 \quad h_0(x_5) = 0$$

and

$$h_1(x_1) = 1 \quad h_1(x_2) = 1 \quad h_1(x_3) = 0 \quad h_1(x_4) = 1 \quad h_1(x_5) = 0$$

Note that $L(h_i, \mathbb{P}_i) = L(h_{\neg i}, \mathbb{P}_i) = \eta/2$. Moreover,

$$\begin{aligned}
\mathcal{D}^{opp}(h_0, \mathbb{P}_0) &= \left|\mathbb{P}_{(X,A,Y)\sim\mathbb{P}_0}(h_1(X) = 1 | A = 0, Y = 1) - \mathbb{P}_{(X,A,Y)\sim\mathbb{P}_0}(h_1(X) = 1 | A = 1, Y = 1)\right| \\
&= \left|1 - \frac{P_{11} - \eta/2}{P_{11} - \eta/2 + \eta/2}\right| \\
&= \frac{\eta}{2P_{11}}
\end{aligned}$$

and similarly $\mathcal{D}^{opp}(h_1, \mathbb{P}_0) = \frac{\eta}{2P_{10}}$. Since $P_{10} \leq P_{11}$, $\mathcal{D}^{opp}(h_0, \mathbb{P}_0) \leq \mathcal{D}^{opp}(h_1, \mathbb{P}_0)$ and

$$\mathcal{D}^{opp}(h_1, \mathbb{P}_0) - \mathcal{D}^{opp}(h_0, \mathbb{P}_0) = \frac{\eta}{2P_{10}}\left(1 - \frac{P_{10}}{P_{11}}\right).$$

Similarly $\mathcal{D}^{opp}(h_0, \mathbb{P}_1) = \frac{\eta}{2P_{10}}$ and $\mathcal{D}^{opp}(h_1, \mathbb{P}_1) = \frac{\eta}{2P_{11}}$, so that $\mathcal{D}^{opp}(h_1, \mathbb{P}_1) \leq \mathcal{D}^{opp}(h_0, \mathbb{P}_1)$ and

$$\mathcal{D}^{opp}(h_0, \mathbb{P}_1) - \mathcal{D}^{opp}(h_1, \mathbb{P}_1) = \frac{\eta}{2P_{10}}\left(1 - \frac{P_{10}}{P_{11}}\right).$$

Consider a (randomized) malicious adversary $\mathcal{A}_i$ of power α, that given a clean distribution $\mathbb{P}_i$, changes every marked point to $(x_3, \neg i, 1)$ with probability 0.5 and to $(x_4, i, 1)$ otherwise. Under a distribution $\mathbb{P}_i$ and an adversary $\mathcal{A}_i$, the probability of seeing a point $(x_3, i, 1)$ is $\frac{\eta}{2}(1 - \alpha) = \frac{\eta}{2}\frac{1}{1+\eta} = \alpha/2$, which is equal to the probability of seeing a point $(x_3, \neg i, 1)$. Therefore, denoting the probability distribution of the corrupted dataset, under a clean distribution $\mathbb{P}_i$ and an adversary $\mathcal{A}_i$, by $\mathbb{P}'_i$, we have

$$\mathbb{P}'_i(x, a, y) = \begin{cases}
(1 - \alpha)(P_{11} - \eta/2) & \text{if } x = x_1, a = 1, y = 1 \\
(1 - \alpha)(P_{10} - \eta/2) & \text{if } x = x_2, a = 0, y = 1 \\
\alpha/2 & \text{if } x = x_3, a = i, y = 1 \\
\alpha/2 & \text{if } x = x_3, a = \neg i, y = 1 \\
\alpha/2 & \text{if } x = x_4, a = \neg i, y = 1 \\
\alpha/2 & \text{if } x = x_4, a = i, y = 1 \\
(1 - \alpha)(1 - P_{10} - P_{11}) & \text{if } x = x_5, a = 0, y = 0 \\
0 & \text{otherwise}
\end{cases}$$

In particular, $\mathbb{P}'_0 = \mathbb{P}'_1$, so the two initial distributions $\mathbb{P}_0$ and $\mathbb{P}_1$ become indistinguishable under the adversarial manipulation.

Fix an arbitrary learner $\mathcal{L} : \cup_{n\in\mathbb{N}}(\mathcal{X} \times A \times \mathcal{Y})^n \to \{h_0, h_1\}$. Note that, if the clean distribution is $\mathbb{P}_0$, the events (in the probability space defined by the sampling of the poisoned train data)

$$\{\mathcal{L}(S^p) = h_1\} = \left\{\mathcal{D}^{opp}(\mathcal{L}(S^p), \mathbb{P}_0) - \mathcal{D}^{opp}(h_0, \mathbb{P}_0) \geq \frac{\eta}{2P_{10}}\left(1 - \frac{P_{10}}{P_{11}}\right)\right\}$$

are all the same. Similarly, if the clean distribution is $\mathbb{P}_1$

$$\{\mathcal{L}(S^p) = h_0\} = \left\{\mathcal{D}^{opp}(\mathcal{L}(S^p), \mathbb{P}_1) - \mathcal{D}^{opp}(h_1, \mathbb{P}_1) \geq \frac{\eta}{2P_{10}}\left(1 - \frac{P_{10}}{P_{11}}\right)\right\}.$$

Therefore, depending on whether we choose $\mathbb{P}_0$ or $\mathbb{P}_1$ as a clean distribution, we have

$$\mathbb{P}_{S^p\sim\mathbb{P}'_0}\left(\mathcal{D}^{opp}(\mathcal{L}(S^p), \mathbb{P}_0) - \mathcal{D}^{opp}(h_0, \mathbb{P}_0) \geq \frac{\eta}{2P_{10}}\left(1 - \frac{P_{10}}{P_{11}}\right)\right) = \mathbb{P}_{S^p\sim\mathbb{P}'_0}\left(\mathcal{L}(S^p) = h_1\right)$$

and

$$\mathbb{P}_{S^p\sim\mathbb{P}'_1}\left(\mathcal{D}^{opp}(\mathcal{L}(S^p), \mathbb{P}_1) - \mathcal{D}^{opp}(h_1, \mathbb{P}_1) \geq \frac{\eta}{2P_{10}}\left(1 - \frac{P_{10}}{P_{11}}\right)\right) = \mathbb{P}_{S^p\sim\mathbb{P}'_1}\left(\mathcal{L}(S^p) = h_0\right)$$

Finally, note that $\mathbb{P}'_0 = \mathbb{P}'_1$, so that either $\mathbb{P}_{S^p\sim\mathbb{P}'_0}\left(\mathcal{L}(S^p) = h_1\right) \geq 1/2$ or $\mathbb{P}_{S^p\sim\mathbb{P}'_1}\left(\mathcal{L}(S^p) = h_0\right) \geq 1/2$. Moreover, $L(\mathcal{L}(S^p), \mathbb{P}_i) = L(h_i, \mathbb{P}_i) = \eta/2$ holds for both $i \in \{0, 1\}$, for any realization of the randomness. Therefore, for at least one of $i = 0, 1$, both

$$L(\mathcal{L}(S^p), \mathbb{P}_i) = L(h_i, \mathbb{P}_i) = \frac{\eta}{2}$$

and

$$\mathcal{D}^{opp}(\mathcal{L}(S^p), \mathbb{P}_i) - \mathcal{D}^{opp}(h_i, \mathbb{P}_i) \geq \frac{\eta}{2P_{10}}\left(1 - \frac{P_{10}}{P_{11}}\right) = \frac{\alpha}{2P_{10}(1-\alpha)}\left(1 - \frac{P_{10}}{P_{11}}\right)$$

both hold with probability at least $1/2$. This concludes the proof in the first case.

Case 2 Now assume that $\frac{\alpha}{1-\alpha} > 2P_{10}$. Then denote by α_1 the unique number between $(0, 0.5)$, such that $\frac{\alpha_1}{1-\alpha_1} = 2P_{10}$, and note that $\alpha_1 < \alpha$. Then repeat the same construction as in Case 1, but with $\eta_1 = \frac{\alpha_1}{1-\alpha_1}$ and an adversary that with probability α_1/α does the same as in Case 1 and leaves a marked point untouched otherwise.

Then the same argument as in Case 1 gives that for some $i \in \{0, 1\}$, with probability at least 0.5, both of the following hold

$$L(\mathcal{L}(S^p), \mathbb{P}_i) = L(h_i, \mathbb{P}_i) = \frac{\eta_1}{2} = P_{10}$$

and

$$\mathcal{D}^{opp}(\mathcal{L}(S^p), \mathbb{P}_i) - \mathcal{D}^{opp}(h_i, \mathbb{P}_i) \geq \frac{\eta_1}{2P_{10}} = \frac{\eta_1}{2P_{10}}\left(1 - \frac{P_{10}}{P_{11}}\right) = 1 - \frac{P_{10}}{P_{11}}.$$

This concludes the proof of Theorem 4. ■