A. Error and Fairness for Randomized Classifiers

Let D denote the distribution over triples (X, A, Y). The accuracy of a classifier $h \in \mathcal{H}$ is measured by 0-1 error, $\mathrm{err}(h) \coloneqq \mathbb{P}_D[h(X) \neq Y]$, which for a randomized classifier Q becomes

$$\mathrm{err}(Q) \coloneqq \mathop{\mathbb{P}}_{(X,A,Y)\sim D,\, h\sim Q}[h(X) \neq Y] = \sum_{h\in\mathcal{H}} Q(h)\,\mathrm{err}(h) \ .$$

The fairness constraints on a classifier h are $\mathbf{M}\boldsymbol{\mu}(h) \leq \mathbf{c}$. Recall that $\mu_j(h) \coloneqq \mathbb{E}_D[g_j(X, A, Y, h(X)) \mid \mathcal{E}_j]$. For a randomized classifier Q we define its moment μ_j as

$$\mu_j(Q) \coloneqq \mathop{\mathbb{E}}_{(X,A,Y)\sim D,\, h\sim Q}\Big[g_j(X, A, Y, h(X)) \,\Big|\, \mathcal{E}_j\Big] = \sum_{h\in\mathcal{H}} Q(h)\mu_j(h) \ ,$$

where the last equality follows because $\mathcal{E}_j$ is independent of the choice of h.

B. Proof of Theorem 1

The proof follows immediately from the analysis of Freund & Schapire (1996) applied to the Exponentiated Gradient (EG) algorithm (Kivinen & Warmuth, 1997), which in our specific case is also equivalent to Hedge (Freund & Schapire, 1997).

Let $\Lambda \coloneqq \{\boldsymbol{\lambda} \in \mathbb{R}_+^{|\mathcal{K}|} : \|\boldsymbol{\lambda}'\|_1 \leq B\}$ and $\Lambda' \coloneqq \{\boldsymbol{\lambda}' \in \mathbb{R}_+^{|\mathcal{K}|+1} : \|\boldsymbol{\lambda}'\|_1 = B\}$. We associate any $\boldsymbol{\lambda} \in \Lambda$ with the $\boldsymbol{\lambda}' \in \Lambda'$ that is equal to $\boldsymbol{\lambda}$ on coordinates 1 through $|\mathcal{K}|$ and puts the remaining mass on the coordinate $\lambda'_{|\mathcal{K}|+1}$.

Consider a run of Algorithm 1. For each $\boldsymbol{\lambda}_t$, let $\boldsymbol{\lambda}'_t \in \Lambda'$ be the associated element of Λ'. Let $\mathbf{r}_t \coloneqq \mathbf{M}\widehat{\boldsymbol{\mu}}(h_t) - \widehat{\mathbf{c}}$ and let $\mathbf{r}'_t \in \mathbb{R}^{|\mathcal{K}|+1}$ be equal to $\mathbf{r}_t$ on coordinates 1 through $|\mathcal{K}|$ and put zero on the coordinate $r'_{t,|\mathcal{K}|+1}$. Thus, for any $\boldsymbol{\lambda}$ and the associated $\boldsymbol{\lambda}'$, we have, for all t,

$$\boldsymbol{\lambda}^\top \mathbf{r}_t = (\boldsymbol{\lambda}')^\top \mathbf{r}'_t \ , \tag{7}$$

and, in particular,

$$\boldsymbol{\lambda}_t^\top \left(\mathbf{M}\widehat{\boldsymbol{\mu}}(h_t) - \widehat{\mathbf{c}}\right) = \boldsymbol{\lambda}_t^\top \mathbf{r}_t = (\boldsymbol{\lambda}'_t)^\top \mathbf{r}'_t \ . \tag{8}$$

We interpret $\mathbf{r}'_t$ as the reward vector for the $\boldsymbol{\lambda}$-player. The choices of $\boldsymbol{\lambda}'_t$ then correspond to those of the EG algorithm with the learning rate η. By the assumption of the theorem we have $\|\mathbf{r}'_t\|_\infty = \|\mathbf{r}_t\|_\infty \leq \rho$. The regret bound for EG, specifically, Corollary 2.14 of Shalev-Shwartz (2012), then states that for any $\boldsymbol{\lambda}' \in \Lambda'$,

$$\sum_{t=1}^T (\boldsymbol{\lambda}')^\top \mathbf{r}'_t \leq \sum_{t=1}^T (\boldsymbol{\lambda}'_t)^\top \mathbf{r}'_t + \underbrace{\frac{B\log(|\mathcal{K}|+1)}{\eta} + \eta\rho^2 BT}_{=:\zeta_T} \ .$$

Therefore, by equations (7) and (8), we also have for any $\boldsymbol{\lambda} \in \Lambda$,

$$\sum_{t=1}^T \boldsymbol{\lambda}^\top \mathbf{r}_t \leq \sum_{t=1}^T \boldsymbol{\lambda}_t^\top \mathbf{r}_t + \zeta_T \ . \tag{9}$$

This regret bound can be used to bound the suboptimality of $L(\widehat{Q}_T, \widehat{\boldsymbol{\lambda}}_T)$ in $\widehat{\boldsymbol{\lambda}}_T$ as follows:

$$\begin{aligned}
L(\widehat{Q}_T, \boldsymbol{\lambda}) &= \frac{1}{T}\sum_{t=1}^{T}\left(\widehat{\mathrm{err}}(h_t) + \boldsymbol{\lambda}^\top\left(\mathbf{M}\widehat{\boldsymbol{\mu}}(h_t) - \widehat{\mathbf{c}}\right)\right) \\
&= \frac{1}{T}\sum_{t=1}^{T}\left(\widehat{\mathrm{err}}(h_t) + \boldsymbol{\lambda}^\top \mathbf{r}_t\right) \\
&\leq \frac{1}{T}\sum_{t=1}^{T}\left(\widehat{\mathrm{err}}(h_t) + \boldsymbol{\lambda}_t^\top \mathbf{r}_t\right) + \frac{\zeta_T}{T} \qquad (10) \\
&= \frac{1}{T}\sum_{t=1}^{T} L(h_t, \boldsymbol{\lambda}_t) + \frac{\zeta_T}{T} \\
&\leq \frac{1}{T}\sum_{t=1}^{T} L(\widehat{Q}_T, \boldsymbol{\lambda}_t) + \frac{\zeta_T}{T} \qquad (11) \\
&= L\left(\widehat{Q}_T, \frac{1}{T}\sum_{t=1}^{T}\boldsymbol{\lambda}_t\right) + \frac{\zeta_T}{T} = L(\widehat{Q}_T, \widehat{\boldsymbol{\lambda}}_T) + \frac{\zeta_T}{T}\ . \qquad (12)
\end{aligned}$$

Equation (10) follows from the regret bound (9). Equation (11) follows because $L(h_t, \boldsymbol{\lambda}_t) \leq L(Q, \boldsymbol{\lambda}_t)$ for all Q by the choice of h_t as the best response of the Q-player. Finally, equation (12) follows by linearity of $L(Q, \boldsymbol{\lambda})$ in $\boldsymbol{\lambda}$. Thus, we have for all $\boldsymbol{\lambda} \in \Lambda$,

$$L(\widehat{Q}_T, \widehat{\boldsymbol{\lambda}}_T) \geq L(\widehat{Q}_T, \boldsymbol{\lambda}) - \frac{\zeta_T}{T}\ . \qquad (13)$$

Also, for any Q,

$$\begin{aligned}
L(Q, \widehat{\boldsymbol{\lambda}}_T) &= \frac{1}{T}\sum_{t=1}^{T} L(Q, \boldsymbol{\lambda}_t) \qquad (14) \\
&\geq \frac{1}{T}\sum_{t=1}^{T} L(h_t, \boldsymbol{\lambda}_t) \qquad (15) \\
&\geq \frac{1}{T}\sum_{t=1}^{T} L(h_t, \widehat{\boldsymbol{\lambda}}_T) - \frac{\zeta_T}{T} \qquad (16) \\
&= L(\widehat{Q}_T, \widehat{\boldsymbol{\lambda}}_T) - \frac{\zeta_T}{T}\ , \qquad (17)
\end{aligned}$$

where equation (14) follows by linearity of $L(Q, \boldsymbol{\lambda})$ in $\boldsymbol{\lambda}$, equation (15) follows by the optimality of h_t with respect to $\widehat{\boldsymbol{\lambda}}_t$, equation (16) from the regret bound (9), and equation (17) by linearity of $L(Q, \boldsymbol{\lambda})$ in Q. Thus, for all Q,

$$L(\widehat{Q}_T, \widehat{\boldsymbol{\lambda}}_T) \leq L(Q, \widehat{\boldsymbol{\lambda}}_T) + \frac{\zeta_T}{T}\ . \qquad (18)$$

Equations (13) and (18) immediately imply that for any $T \geq 1$,

$$\nu_T \leq \frac{\zeta_T}{T} = \frac{B\log(|\mathcal{K}|+1)}{\eta T} + \eta\rho^2 B\ ,$$

proving the first part of the theorem.

The second part of the theorem follows by plugging in $\eta = \frac{\nu}{2\rho^2 B}$ and verifying that if $T \geq \frac{4\rho^2 B^2 \log(|\mathcal{K}|+1)}{\nu^2}$ then

$$\nu_T \leq \frac{B\log(|\mathcal{K}|+1)}{\frac{\nu}{2\rho^2 B}\cdot\frac{4\rho^2 B^2\log(|\mathcal{K}|+1)}{\nu^2}} + \frac{\nu}{2\rho^2 B}\cdot\rho^2 B = \frac{\nu}{2} + \frac{\nu}{2}\ .$$

C. Proofs of Theorems 2 and 3

The bulk of this appendix proves the following theorem, which will immediately imply Theorems 2 and 3.

Theorem 4. *Let $(\widehat{Q}, \widehat{\boldsymbol{\lambda}})$ be any ν-approximate saddle point of L with*

$$\widehat{c}_k = c_k + \varepsilon_k \quad \textit{and} \quad \varepsilon_k \geq \sum_{j \in \mathcal{J}} |M_{k,j}| \left(2R_{n_j}(\mathcal{H}) + \frac{2}{\sqrt{n_j}} + \sqrt{\frac{\ln(2/\delta)}{2n_j}} \right) \ .$$

Let $Q^\star$ minimize $\mathrm{err}(Q)$ subject to $\mathbf{M}\boldsymbol{\mu}(Q) \leq \mathbf{c}$. Then with probability at least $1 - (|\mathcal{J}|+1)\delta$, the distribution $\widehat{Q}$ satisfies

$$\mathrm{err}(\widehat{Q}) \leq \mathrm{err}(Q^\star) + 2\nu + 4R_n(\mathcal{H}) + \frac{4}{\sqrt{n}} + \sqrt{\frac{2\ln(2/\delta)}{n}} \ ,$$

$$\textit{and for all } k, \quad \gamma_k(\widehat{Q}) \leq c_k + \frac{1+2\nu}{B} + 2\varepsilon_k \ .$$

Let $\Lambda \coloneqq \{\boldsymbol{\lambda} \in \mathbb{R}_+^{|\mathcal{K}|} : \ \|\boldsymbol{\lambda}'\|_1 \leq B\}$ denote the domain of $\boldsymbol{\lambda}$. In the remainder of the section, we assume that we are given a pair $(\widehat{Q}, \widehat{\boldsymbol{\lambda}})$ which is a ν-approximate saddle point of L, i.e.,

$$\begin{aligned} L(\widehat{Q}, \widehat{\boldsymbol{\lambda}}) &\leq L(Q, \widehat{\boldsymbol{\lambda}}) + \nu \quad \text{for all } Q \in \Delta, \\ \text{and} \quad L(\widehat{Q}, \widehat{\boldsymbol{\lambda}}) &\geq L(\widehat{Q}, \boldsymbol{\lambda}) - \nu \quad \text{for all } \boldsymbol{\lambda} \in \Lambda. \end{aligned} \tag{19}$$

We first establish that the pair $(\widehat{Q}, \widehat{\boldsymbol{\lambda}})$ satisfies an approximate version of complementary slackness. For the statement and proof of the following lemma, recall that $\widehat{\boldsymbol{\gamma}}(Q) = \mathbf{M}\widehat{\boldsymbol{\mu}}(Q)$, so the empirical fairness constraints can be written as $\widehat{\boldsymbol{\gamma}}(Q) \leq \widehat{\mathbf{c}}$ and the Lagrangian L can be written as

$$L(Q, \boldsymbol{\lambda}) = \widehat{\mathrm{err}}(Q) + \sum_{k \in \mathcal{K}} \lambda_k (\widehat{\gamma}_k(Q) - \widehat{c}_k) \ . \tag{20}$$

Lemma 1 (Approximate complementary slackness)**.** *The pair $(\widehat{Q}, \widehat{\boldsymbol{\lambda}})$ satisfies*

$$\sum_{k \in \mathcal{K}} \widehat{\lambda}_k (\widehat{\gamma}_k(\widehat{Q}) - \widehat{c}_k) \geq B \max_{k \in \mathcal{K}} \big(\widehat{\gamma}_k(\widehat{Q}) - \widehat{c}_k\big)_+ - \nu \ ,$$

where we abbreviate $x_+ = \max\{x, 0\}$ for any real number x.

Proof. We show that the lemma follows from the optimality conditions (19). We consider a dual variable $\boldsymbol{\lambda}$ defined as

$$\boldsymbol{\lambda} = \begin{cases} \mathbf{0} & \text{if } \widehat{\boldsymbol{\gamma}}(\widehat{Q}) \leq \widehat{\mathbf{c}}, \\ B\mathbf{e}_{k^\star} & \text{otherwise, where } k^\star = \arg\max_k [\widehat{\gamma}_k(\widehat{Q}) - \widehat{c}_k], \end{cases}$$

where $\mathbf{e}_k$ denotes the kth vector of the standard basis. Then we have by equations (19) and (20) that

$$\begin{aligned} \widehat{\mathrm{err}}(\widehat{Q}) + \sum_{k \in \mathcal{K}} \widehat{\lambda}_k (\widehat{\gamma}_k(\widehat{Q}) - \widehat{c}_k) &= L(\widehat{Q}, \widehat{\boldsymbol{\lambda}}) \\ &\geq L(\widehat{Q}, \boldsymbol{\lambda}) - \nu = \widehat{\mathrm{err}}(\widehat{Q}) + \sum_{k \in \mathcal{K}} \lambda_k (\widehat{\gamma}_k(\widehat{Q}) - \widehat{c}_k) - \nu \ , \end{aligned}$$

and the lemma follows by our choice of $\boldsymbol{\lambda}$. □

Next two lemmas bound the empirical error of $\widehat{Q}$ and also bound the amount by which $\widehat{Q}$ violates the empirical fairness constraints.

Lemma 2 (Empirical error bound)**.** *The distribution $\widehat{Q}$ satisfies $\widehat{\mathrm{err}}(\widehat{Q}) \leq \widehat{\mathrm{err}}(Q) + 2\nu$ for any Q satisfying the empirical fairness constraints, i.e., any Q such that $\widehat{\boldsymbol{\gamma}}(Q) \leq \widehat{\mathbf{c}}$.*

Proof. Assume that Q satisfies $\widehat{\gamma}(Q) \leq \widehat{\mathbf{c}}$. Since $\widehat{\boldsymbol{\lambda}} \geq \mathbf{0}$, we have

$$L(Q, \widehat{\boldsymbol{\lambda}}) = \widehat{\mathrm{err}}(Q) + \widehat{\boldsymbol{\lambda}}^\top \big(\widehat{\gamma}(Q) - \widehat{\mathbf{c}}\big) \leq \widehat{\mathrm{err}}(Q) \ .$$

The optimality conditions (19) imply that

$$L(\widehat{Q}, \widehat{\boldsymbol{\lambda}}) \leq L(Q, \widehat{\boldsymbol{\lambda}}) + \nu \ .$$

Putting these together, we obtain

$$L(\widehat{Q}, \widehat{\boldsymbol{\lambda}}) \leq \widehat{\mathrm{err}}(Q) + \nu \ .$$

We next invoke Lemma 1 to lower bound $L(\widehat{Q}, \widehat{\boldsymbol{\lambda}})$ as

$$\begin{aligned} L(\widehat{Q}, \widehat{\boldsymbol{\lambda}}) = \widehat{\mathrm{err}}(\widehat{Q}) + \sum_{k \in \mathcal{K}} \widehat{\lambda}_k (\widehat{\gamma}_k(\widehat{Q}) - \widehat{c}_k) &\geq \widehat{\mathrm{err}}(\widehat{Q}) + B \max_{k \in \mathcal{K}} \big(\widehat{\gamma}_k(\widehat{Q}) - \widehat{c}_k\big)_+ - \nu \\ &\geq \widehat{\mathrm{err}}(\widehat{Q}) - \nu \ . \end{aligned}$$

Combining the upper and lower bounds on $L(\widehat{Q}, \widehat{\boldsymbol{\lambda}})$ completes the proof. □

Lemma 3 (Empirical fairness violation)**.** *Assume that the empirical fairness constraints $\widehat{\gamma}(Q) \leq \widehat{\mathbf{c}}$ are feasible. Then the distribution $\widehat{Q}$ approximately satisfies all empirical fairness constraints:*

$$\max_{k \in \mathcal{K}} \Big(\widehat{\gamma}_k(\widehat{Q}) - \widehat{c}_k\Big) \leq \frac{1 + 2\nu}{B} \ .$$

Proof. Let Q satisfy $\widehat{\gamma}(Q) \leq \widehat{\mathbf{c}}$. Applying the same upper and lower bound on $L(\widehat{Q}, \widehat{\boldsymbol{\lambda}})$ as in the proof of Lemma 2, we obtain

$$\widehat{\mathrm{err}}(\widehat{Q}) + B \max_{k \in \mathcal{K}} \big(\widehat{\gamma}_k(\widehat{Q}) - \widehat{c}_k\big)_+ - \nu \ \leq \ L(\widehat{Q}, \widehat{\boldsymbol{\lambda}}) \ \leq \ \widehat{\mathrm{err}}(Q) + \nu \ .$$

We can further upper bound $\widehat{\mathrm{err}}(Q) - \widehat{\mathrm{err}}(\widehat{Q})$ by 1 and use $x \leq x_+$ for any real number x to complete the proof. □

It remains to lift the bounds on empirical classification error and constraint violation into the corresponding bounds on true classification error and the violation of true constraints. We will use the standard machinery of uniform convergence bounds via the (worst-case) Rademacher complexity.

Let $\mathcal{F}$ be a class of functions $f : \mathcal{Z} \to [0, 1]$ over some space $\mathcal{Z}$. Then the (worst-case) *Rademacher complexity* of $\mathcal{F}$ is defined as

$$R_n(\mathcal{F}) \coloneqq \sup_{z_1, \ldots, z_n \in \mathcal{Z}} \mathbb{E}\left[\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n \sigma_i f(z_i) \right| \right] \ ,$$

where the expectation is over the i.i.d. random variables $\sigma_1, \ldots, \sigma_n$ with $\mathbb{P}[\sigma_i = 1] = \mathbb{P}[\sigma_i = -1] = 1/2$.

We first prove concentration of generic moments derived from classifiers $h \in \mathcal{H}$ and then move to bounding the deviations from true classification error and true fairness constraints.

Lemma 4 (Concentration of moments)**.** *Let $g : \mathcal{X} \times \mathcal{A} \times \{0, 1\} \times \{0, 1\} \to [0, 1]$ be any function and let D be a distribution over (X, A, Y). Then with probability at least $1 - \delta$, for all $h \in \mathcal{H}$,*

$$\Big|\widehat{\mathbb{E}}\big[g(X, A, Y, h(X))\big] - \mathbb{E}\big[g(X, A, Y, h(X))\big]\Big| \leq 2R_n(\mathcal{H}) + \frac{2}{\sqrt{n}} + \sqrt{\frac{\ln(2/\delta)}{2n}} \ ,$$

where the expectation is with respect to D and the empirical expectation is based on n i.i.d. draws from D.

Proof. Let $\mathcal{F} \coloneqq \{f_h\}_{h \in \mathcal{H}}$ be the class of functions $f_h : (x, y, a) \mapsto g\big(x, y, a, h(x)\big)$. By Theorem 3.2 of Boucheron et al. (2005), we then have with probability at least $1 - \delta$, for all h,

$$\Big|\widehat{\mathbb{E}}\big[g(X, A, Y, h(X))\big] - \mathbb{E}\big[g(X, A, Y, h(X))\big]\Big| = \Big|\widehat{\mathbb{E}}[f_h] - \mathbb{E}[f_h]\Big| \leq 2R_n(\mathcal{F}) + \sqrt{\frac{\ln(2/\delta)}{2n}} \ . \tag{21}$$

We will next bound $R_n(\mathcal{F})$ in terms of $R_n(\mathcal{H})$. Since $h(x) \in \{0,1\}$, we can write

$$f_h(x,y,a) = h(x)g(x,a,y,1) + \Big(1-h(x)\Big)g(x,a,y,0) = g(x,a,y,0) + h(x)\Big(g(x,a,y,1) - g(x,a,y,0)\Big) \ .$$

Since $\big|g(x,a,y,0)\big| \leq 1$ and $\big|g(x,a,y,1) - g(x,a,y,0)\big| \leq 1$, we can invoke Theorem 12(5) of Bartlett & Mendelson (2002) for bounding function classes shifted by an offset, in our case $g(x,a,y,0)$, and Theorem 4.4 of Ledoux & Talagrand (1991) for bounding function classes under contraction, in our case $g(x,a,y,1) - g(x,a,y,0)$, yielding

$$R_n(\mathcal{F}) \leq \frac{1}{\sqrt{n}} + R_n(\mathcal{H}) \ .$$

Together with the bound (21), this proves the lemma. □

Lemma 5 (Concentration of loss)**.** *With probability at least* $1-\delta$*, for all* $Q \in \Delta$*,*

$$|\widehat{\mathrm{err}}(Q) - \mathrm{err}(Q)| \leq 2R_n(\mathcal{H}) + \frac{2}{\sqrt{n}} + \sqrt{\frac{\ln(2/\delta)}{2n}} \ .$$

Proof. We first use Lemma 4 with $g : (x,a,y,\hat{y}) \mapsto \mathbf{1}\{y \neq \hat{y}\}$ to obtain, with probability $1-\delta$, for all h,

$$\Big|\widehat{\mathrm{err}}(h) - \mathrm{err}(h)\Big| = \Big|\widehat{\mathbb{E}}[f_h] - \mathbb{E}[f_h]\Big| \leq 2R_n(\mathcal{H}) + \frac{2}{\sqrt{n}} + \sqrt{\frac{\ln(2/\delta)}{2n}} \ .$$

The lemma now follows for any Q by taking a convex combination of the corresponding bounds on $h \in \mathcal{H}$.[7] □

Finally, we show a result for the concentration of the empirical constraint violations to their population counterparts. We will actually show the concentration of the individual moments $\widehat{\mu}_j(Q)$ to $\mu_j(Q)$ uniformly for all $Q \in \Delta$. Since $\mathbf{M}$ is a fixed matrix not dependent on the data, this also directly implies concentration of the constraints $\widehat{\boldsymbol{\gamma}}(Q) = \mathbf{M}\widehat{\boldsymbol{\mu}}(Q)$ to $\boldsymbol{\gamma}(Q) = \mathbf{M}\boldsymbol{\mu}(Q)$. For this result, recall that $n_j = |\{i \in [n] : \ (X_i,A_i,Y_i) \in \mathcal{E}_j\}|$ and $p^\star_j = \mathbb{P}[\mathcal{E}_j]$.

Lemma 6 (Concentration of conditional moments)**.** *For any* $j \in \mathcal{J}$*, with probability at least* $1-\delta$*, for all* Q*,*

$$\big|\widehat{\mu}_j(Q) - \mu_j(Q)\big| \leq 2R_{n_j}(\mathcal{H}) + \frac{2}{\sqrt{n_j}} + \sqrt{\frac{\ln(2/\delta)}{2n_j}} \ .$$

If $np^\star_j \geq 8\log(2/\delta)$*, then with probability at least* $1-\delta$*, for all* Q*,*

$$\big|\widehat{\mu}_j(Q) - \mu_j(Q)\big| \leq 2R_{np^\star_j/2}(\mathcal{H}) + 2\sqrt{\frac{2}{np^\star_j}} + \sqrt{\frac{\ln(4/\delta)}{np^\star_j}} \ .$$

Proof. Our proof largely follows the proof of Lemma 2 of Woodworth et al. (2017), with appropriate modifications for our more general constraint definition. Let $S_j \coloneqq \{i \in [n] : \ (X_i,A_i,Y_i) \in \mathcal{E}_j\}$ be the set of indices such that the corresponding examples fall in the event $\mathcal{E}_j$. Note that we have defined $n_j = |S_j|$. Let $D(\cdot)$ denote the joint distribution of (X,A,Y). Then, conditioned on $i \in S_j$, the random variables $g_j(X_i,A_i,Y_i,h(X_i))$ are i.i.d. draws from the distribution $D(\cdot \mid \mathcal{E}_j)$, with mean $\mu_j(h)$. Applying Lemma 4 with g_j and the distribution $D(\cdot \mid \mathcal{E}_j)$ therefore yields, with probability $1-\delta$, for all h,

$$\big|\widehat{\mu}_j(h) - \mu_j(h)\big| \leq 2R_{n_j}(\mathcal{H}) + \frac{2}{\sqrt{n_j}} + \sqrt{\frac{\ln(2/\delta)}{2n_j}} \ ,$$

The lemma now follows by taking a convex combination over h. □

[7] The same reasoning applies for general error, $\mathrm{err}(h) = \mathbb{E}[g_{\mathrm{err}}(X,A,Y,h(X))]$, by using $g = g_{\mathrm{err}}$ in Lemma 4.

Proof of Theorem 4. We now use the lemmas derives so far to prove Theorem 4. We first use Lemma 6 to bound the gap between the empirical and population fairness constraints. The lemma implies that with probability at least $1 - |\mathcal{J}|\delta$, for all $k \in \mathcal{K}$ and all $Q \in \Delta$,

$$\begin{aligned}
\left|\widehat{\gamma}_k(Q) - \gamma_k(Q)\right| &= \left|\mathbf{M}_k\left(\widehat{\boldsymbol{\mu}}(Q) - \boldsymbol{\mu}(Q)\right)\right| \\
&\leq \sum_{j \in \mathcal{J}} |M_{k,j}| \left|\widehat{\mu}_j(Q) - \mu_j(Q)\right| \\
&\leq \sum_{j \in \mathcal{J}} |M_{k,j}| \left(2R_{n_j}(\mathcal{H}) + \frac{2}{\sqrt{n_j}} + \sqrt{\frac{\ln(2/\delta)}{2n_j}}\right) \\
&\leq \varepsilon_k \ . \qquad (22)
\end{aligned}$$

Note that our choice of $\widehat{\mathbf{c}}$ along with equation (22) ensure that $\widehat{\gamma}_k(Q^\star) \leq \widehat{c}_k$ for all $k \in \mathcal{K}$. Using Lemma 2 allows us to conclude that

$$\widehat{\mathrm{err}}(\widehat{Q}) \leq \widehat{\mathrm{err}}(Q^\star) + 2\nu \ .$$

We now invoke Lemma 5 twice, once for $\widehat{\mathrm{err}}(\widehat{Q})$ and once for $\widehat{\mathrm{err}}(Q^\star)$, proving the first statement of the theorem.

The above shows that $Q^\star$ satisfies the empirical fairness constraints, so we can use Lemma 3, which together with equation (22) yields

$$\gamma_k(\widehat{Q}) \leq \widehat{\gamma}_k(\widehat{Q}) + \varepsilon_k \leq \widehat{c}_k + \frac{1+2\nu}{B} + \varepsilon_k = c_k + \frac{1+2\nu}{B} + 2\varepsilon_k \ ,$$

proving the second statement of the theorem. □

We are now ready to prove Theorems 2 and 3

Proof of Theorem 2. The first part of the theorem follows immediately from Assumption 1 and Theorem 4 (with $\delta/2$ instead of δ). The statement in fact holds with probability at least $1 - (|\mathcal{J}| + 1)\delta/2$. For the second part, we use the multiplicative Chernoff bound for binomial random variables. Note that $\mathbb{E}[n_j] = np_j^\star$, and we assume that $np_j^\star \geq 8\ln(2/\delta)$, so the multiplicative Chernoff bound implies that $n_j \leq np_j^\star/2$ with probability at most $\delta/2$. Taking the union bound across all j and combining with the first part of the theorem then proves the second part. □

Proof of Theorem 3. This follows immediately from Theorem 1 and the first part of Theorem 2. □

D. Additional Experimental Results

In this appendix we present more complete experimental results. We present experimental results for both the training and test data. We evaluate the exponentiated-gradient as well as the grid-search variants of our reductions. And, finally, we consider extensions of reweighting and relabeling beyond the specific tradeoffs proposed by Kamiran & Calders (2012). Specifically, we introduce a scaling parameter that interpolates between the prescribed tradeoff (specific importance weights or the number of examples to relabel) and the unconstrained classifier (uniform weights or zero examples to relabel). The training data results are shown in Figure 2. The test set results are shown in Figure 3.

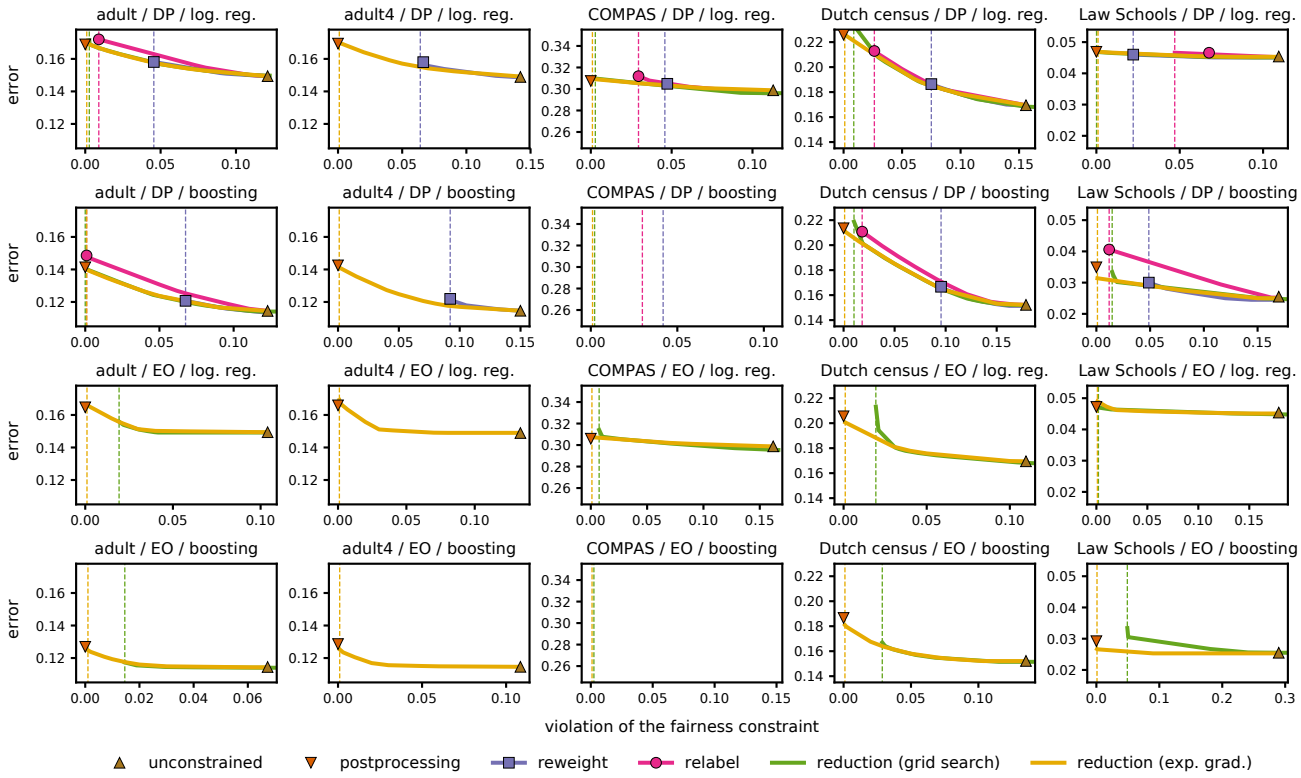

Figure 2. Training classification error versus constraint violation, with respect to DP (top two rows) and EO (bottom two rows). Markers correspond to the baselines. For our two reductions and the interpolants between reweighting (or relabeling) and the unconstrained classifier, we varied their tradeoff parameters and plot the Pareto frontiers of the sets of classifiers obtained for each method. Because the curves of the different methods often overlap, we use vertical dashed lines to indicate the lowest constraint violations. All data sets have binary protected attributes except for *adult4*, which has four protected attribute values, so relabeling is not applicable and grid search is not feasible for this data set. The exponentiated-gradient reduction dominates or matches other approaches as expected since it solves exactly for the points on the Pareto frontier of the set of all classifiers in each considered class.

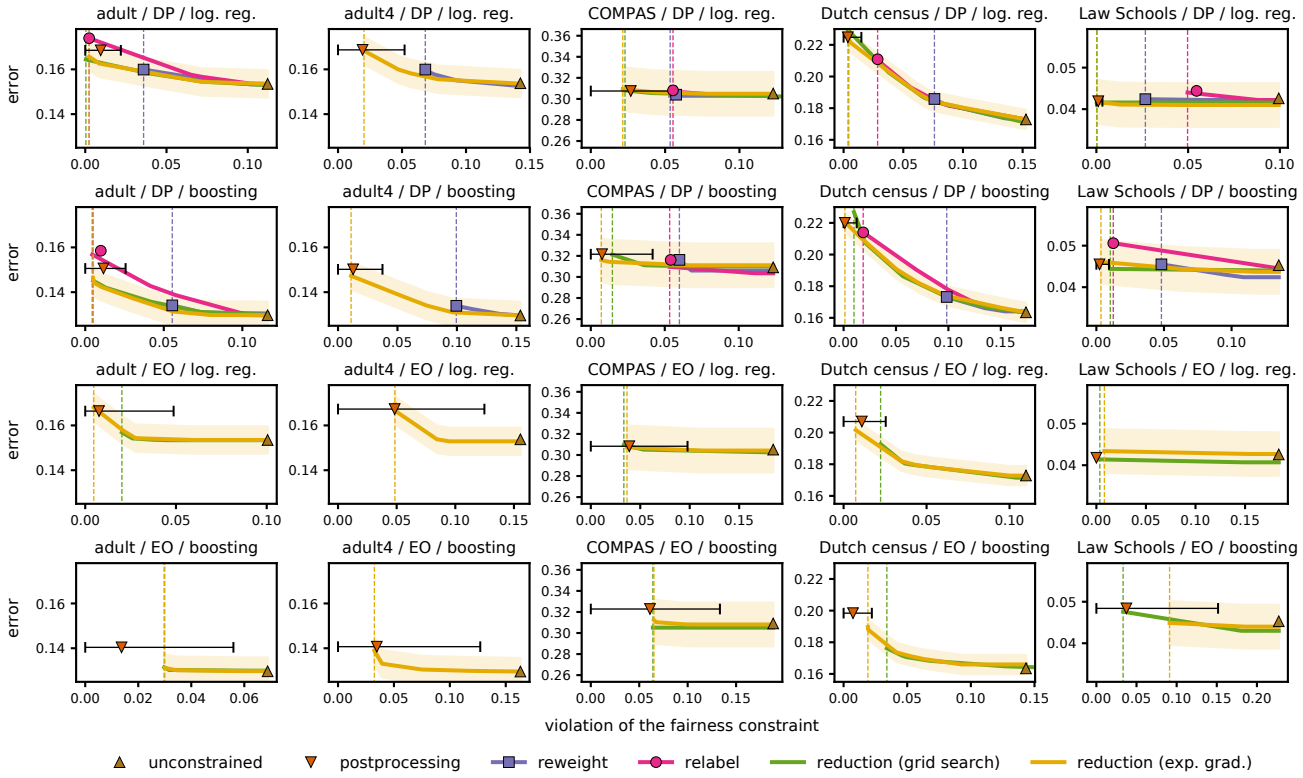

Figure 3. Test classification error versus constraint violation, with respect to DP (top two rows) and EO (bottom two rows). Markers correspond to the baselines. For our two reductions and the interpolants between reweighting (or relabeling) and the unconstrained classifier, we show convex envelopes of the classifiers taken from the *training* Pareto frontier of each method (i.e., the same classifiers as shown in Figure 2). Because the curves of the different methods often overlap, we use vertical dashed lines to indicate the lowest constraint violations. All data sets have binary protected attributes except for *adult4*, which has four protected attribute values, so relabeling is not applicable and grid search is not feasible for this data set. We show 95% confidence bands for the classification error of the exponentiated-gradient reduction and 95% confidence intervals for the constraint violation of post-processing. The exponentiated-gradient reduction dominates or matches performance of all other methods up to statistical uncertainty.