Classification with abstention but without disparities
Supplementary material

Nicolas Schreuder[1] Evgenii Chzhen[2]

[1]CREST, ENSAE Paris, Palaiseau, France
[2]LMO, Université Paris-Saclay, CNRS, Inria, Orsay, France

STRUCTURE OF APPENDIX

Appendix 1 is devoted to the proof of Theorem 3.2. Appendix 2 reminds and proves auxiliary results that are used in the rest of the supplementary material. The proof of Proposition 5.1 is split across Appendix 3 for the control of the reject rate and Appendix 4 for the control of the demographic parity violation. Appendix 5 contains the proof of Proposition 5.2. Finally, Appendix 6 provides a constructive proof of Proposition 6.1.

1 DERIVATION OF THE OPTIMAL PREDICTION

Recall that we are interested in solving the following optimization problem

$$\begin{aligned}
&\min_{g:\mathbb{R}^d\times[K]\to\{0,1,r\}} \mathbb{P}(g(\boldsymbol{X},S)\neq Y \mid g(\boldsymbol{X},S))\neq r)\\
&\quad\text{s.t.}\begin{cases}\mathbb{P}(g(\boldsymbol{X},S))\neq r \mid S=s)=\alpha_s, \quad \forall s\in[K]\\ \mathbb{P}(g(\boldsymbol{X},S)=1 \mid S=s, g(\boldsymbol{X},S)\neq r)=\mathbb{P}(g(\boldsymbol{X},S)=1 \mid g(\boldsymbol{X},S))\neq r), \quad \forall s\in[K]\end{cases}.
\end{aligned} \tag{1}$$

Simplifications. First we simplify the quantities involved in the above problem. Set $\bar{\alpha}=\sum_{s=1}^K p_s\alpha_s$ and recall that we defined the random variable $\eta(\boldsymbol{X},S)=\mathbb{E}[Y \mid \boldsymbol{X},S]$. Observe that for any g such that $\mathbb{P}(g(\boldsymbol{X},S)\neq r \mid s=s)=\alpha_s$, we can write

$$\begin{aligned}
&\mathbb{P}(g(\boldsymbol{X},S)\neq Y \mid g(\boldsymbol{X},S)\neq r)=\sum_{s=1}^K \frac{p_s}{\bar{\alpha}}\mathbb{E}_{\boldsymbol{X}|S=s}\left[(1-\eta(\boldsymbol{X},S)\mathbb{1}_{g(\boldsymbol{X},S)=1}+\eta(\boldsymbol{X},S)\mathbb{1}_{g(\boldsymbol{X},S)=0}\right] \enspace,\\
&\mathbb{P}(g(\boldsymbol{X},S)\neq r|S=s)=\mathbb{E}_{\boldsymbol{X}|S=s}\left[\mathbb{1}_{g(\boldsymbol{X},S)=1}+\mathbb{1}_{g(\boldsymbol{X},S)=0}\right] \enspace,\\
&\mathbb{P}(g(\boldsymbol{X},S)=1 \mid g(\boldsymbol{X},S)\neq r)=\sum_{s=1}^K \frac{p_s}{\bar{\alpha}}\mathbb{E}_{\boldsymbol{X}|S=s}\left[\mathbb{1}_{g(\boldsymbol{X},S)=1}\right] \enspace,\\
&\mathbb{P}(g(\boldsymbol{X},S)=1|S=s, g(\boldsymbol{X},S)\neq r)=\frac{1}{\alpha_s}\mathbb{E}_{\boldsymbol{X}|S=s}\left[\mathbb{1}_{g(\boldsymbol{X},S)=1}\right] \enspace.
\end{aligned}$$

Lagrangian. We introduce the Lagrangian $\mathcal{L}$ of the constrained minimization problem as

$$\begin{aligned}
\mathcal{L}(g,\boldsymbol{\lambda},\boldsymbol{\gamma})=\mathbb{P}(g(\boldsymbol{X},S)\neq Y \mid g(\boldsymbol{X},S)\neq r)+\sum_{s=1}^K\lambda_s(\mathbb{P}(g(\boldsymbol{X},S)\neq r|S=s)-\alpha_s)\\
+\sum_{s=1}^K\gamma_s(\mathbb{P}(g(\boldsymbol{X},S)=1|S=s, g(\boldsymbol{X},S)\neq r)-\mathbb{P}(g(\boldsymbol{X},S)=1|g(\boldsymbol{X},S)\neq r)) \enspace.
\end{aligned}$$

Using the simpler expressions we derived in the previous paragraph and rather straight-forward algebraic manipulations, the Lagrangian can be expressed as

$$\mathcal{L}(g, \boldsymbol{\lambda}, \boldsymbol{\gamma}) = \sum_{s=1}^{K} \mathbb{E}_{\boldsymbol{X}|S=s}\left[H_{(\boldsymbol{X},s)}(g, \boldsymbol{\lambda}, \boldsymbol{\gamma})\right] - \sum_{s=1}^{K} \lambda_s \alpha_s \enspace ,$$

where, setting $\bar{\gamma} := \sum_{s=1}^{K} \gamma_s$, we defined the function

$$H_{(\boldsymbol{x},s)}(g, \boldsymbol{\lambda}, \boldsymbol{\gamma}) = \begin{cases} 0, & \text{if } g(\boldsymbol{x}, s) = r \\ \frac{p_s}{\bar{\alpha}}\eta(\boldsymbol{x}, s) + \lambda_s, & \text{if } g(\boldsymbol{x}, s) = 0 \\ \frac{p_s}{\bar{\alpha}}(1 - \eta(\boldsymbol{x}, s) - \bar{\gamma}) + \lambda_s + \frac{\gamma_s}{\alpha_s}, & \text{if } g(\boldsymbol{x}, s) = 1 \end{cases} \enspace . \tag{2}$$

Using this Lagrangian, our initial problem in Eq. (1) can be expressed as the following $\min\max$ problem

$$\min_{g} \max_{(\boldsymbol{\lambda},\boldsymbol{\gamma}) \in \mathbb{R}^K \times \mathbb{R}^K} \mathcal{L}(g, \boldsymbol{\lambda}, \boldsymbol{\gamma}) \enspace ,$$

where the minimum is take w.r.t. all classifiers with abstention. Weak duality then implies that

$$\min_{g} \max_{(\boldsymbol{\lambda},\boldsymbol{\gamma}) \in \mathbb{R}^K \times \mathbb{R}^K} \mathcal{L}(g, \boldsymbol{\lambda}, \boldsymbol{\gamma}) \geq \max_{(\boldsymbol{\lambda},\boldsymbol{\gamma}) \in \mathbb{R}^K \times \mathbb{R}^K} \min_{g} \mathcal{L}(g, \boldsymbol{\lambda}, \boldsymbol{\gamma}) \enspace .$$

Solving dual problem. In what follows we focus our attention on the dual $\max\min$ problem, which can be solved analytically. We first solve the inner minimization problem of the $\max\min$ formulation for any $(\boldsymbol{\lambda}, \boldsymbol{\gamma})$,

$$\min_{g} \mathcal{L}(g, \boldsymbol{\lambda}, \boldsymbol{\gamma}) \enspace , \tag{3}$$

and then show that strong duality holds under our assumptions. Recall that the Lagrangian is given by $\mathcal{L}(g, \boldsymbol{\lambda}, \boldsymbol{\gamma}) = \sum_{s=1}^{K} \mathbb{E}_{\boldsymbol{X}|S=s}\left[H_{(\boldsymbol{X},s)}(g, \boldsymbol{\lambda}, \boldsymbol{\gamma})\right] - \sum_{s=1}^{K} \lambda_s \alpha_s$ (with H defined in Eq. (2)), hence the problem in Eq. (3) can be solved point-wise (i.e., minimizing $H_{(x,s)}$ at every point (x, s)). Hence, it is sufficient to solve

$$\min_{z \in \{0,1,r\}} H_{(\boldsymbol{x},s)}(z, \boldsymbol{\lambda}, \boldsymbol{\gamma}) \enspace ,$$

for any $s \in [K]$ and any $\boldsymbol{x} \in \mathbb{R}^d$. One can check that, for any given couple $(\boldsymbol{x}, s) \in \mathbb{R}^d \times [K]$, the minimizer of the above expression is given by

$$\tilde{g}(\boldsymbol{x}, s) = \begin{cases} r, & \text{if } 0 \leq \min(\frac{p_s}{\bar{\alpha}}\eta(\boldsymbol{x}, s) + \lambda_s, \frac{p_s}{\bar{\alpha}}(1 - \eta(\boldsymbol{x}, s) - \bar{\gamma}) + \lambda_s + \frac{\gamma_s}{\alpha_s}) \\ \mathbb{1}(\frac{p_s}{\bar{\alpha}}(1 - 2\eta(\boldsymbol{x}, s) - \bar{\gamma}) + \frac{\gamma_s}{\alpha_s} < 0), & \text{otherwise} \end{cases} \enspace .$$

Using the fact that $2\min(a, b) = a + b - |a - b|$, the minimum in the above system simplifies to

$$\tilde{g}(\boldsymbol{x}, s) = \begin{cases} r, & \text{if } \left|\frac{p_s}{2\bar{\alpha}}(1 - 2\eta(\boldsymbol{x}, s) - \bar{\gamma}) + \frac{\gamma_s}{2\alpha_s}\right| \leq \lambda_s + \frac{p_s}{2\bar{\alpha}}(1 - \bar{\gamma}) + \frac{\gamma_s}{2\alpha_s} \\ \mathbb{1}(\frac{p_s}{\bar{\alpha}}(1 - 2\eta(\boldsymbol{x}, s) - \bar{\gamma}) + \frac{\gamma_s}{\alpha_s} < 0), & \text{otherwise} \end{cases} \enspace .$$

Note that $\tilde{g}$ actually depends on (λ, γ), but we omit this dependency for the sake of simplicity. Plugging back the expression for $\tilde{g}$ in the function H we get

$$H_{(\boldsymbol{x},s)}(\tilde{g}, \boldsymbol{\lambda}, \boldsymbol{\gamma}) = \left(\frac{p_s}{2\bar{\alpha}}(1 - \bar{\gamma}) + \lambda_s + \frac{\gamma_s}{2\alpha_s} - \left|\frac{p_s}{2\bar{\alpha}}(1 - 2\eta(\boldsymbol{x}, s) - \bar{\gamma}) + \frac{\gamma_s}{2\alpha_s}\right|\right)_{-} \enspace ,$$

where $(a)_{-} := \min(a, 0)$. Thus, for every fixed (λ, γ) we have $\min_g \mathcal{L}(g, \boldsymbol{\lambda}, \boldsymbol{\gamma}) = \mathcal{L}(\tilde{g}, \boldsymbol{\lambda}, \boldsymbol{\gamma})$. Maximizing the latter expression over the dual variables (λ, γ), we derive the dual optimization problem, which reads as

$$\max_{(\boldsymbol{\lambda},\boldsymbol{\gamma})} \left\{ \sum_{s=1}^{K} \mathbb{E}_{\boldsymbol{X}|S=s} \left(\frac{p_s}{2\bar{\alpha}}(1 - \bar{\gamma}) + \lambda_s + \frac{\gamma_s}{2\alpha_s} - \left|\frac{p_s}{2\bar{\alpha}}(1 - 2\eta(\boldsymbol{x}, s) - \bar{\gamma}) + \frac{\gamma_s}{2\alpha_s}\right|\right)_{-} - \sum_{s=1}^{K} \lambda_s \alpha_s \right\} \enspace .$$

Replacing the maximization problem above by its minimization analogue, we conclude that the optimal (for the dual) Lagrange multipliers $(\boldsymbol{\lambda}^*, \boldsymbol{\gamma}^*)$ are a solution of

$$\min_{(\boldsymbol{\lambda},\boldsymbol{\gamma})} \left\{ \sum_{s=1}^{K} \mathbb{E}_{\boldsymbol{X}|S=s} \left(\left|\frac{p_s}{2\bar{\alpha}}(1 - 2\eta(\boldsymbol{X}, S) - \langle\boldsymbol{\gamma}, \mathbf{1}\rangle) + \frac{\langle\boldsymbol{\gamma}, \boldsymbol{e}_s\rangle}{2\alpha_s}\right| - \frac{p_s}{2\bar{\alpha}}(1 - \langle\boldsymbol{\gamma}, \mathbf{1}\rangle) - \langle\boldsymbol{\lambda}, \boldsymbol{e}_s\rangle - \frac{\langle\boldsymbol{\gamma}, \boldsymbol{e}_s\rangle}{2\alpha_s}\right)_{+} + \langle\boldsymbol{\lambda}, \boldsymbol{\alpha}\rangle \right\} \enspace , \tag{4}$$

where for any real number y, $(y)_+ := \max(x, 0)$ and for any $s \in [K]$, $\boldsymbol{e}_s$ is the s-basis vector of $\mathbb{R}^K$.

Dual is jointly convex. Let us check that the objective function of the problem in Eq. (4) of the above optimization problem is jointly convex in $(\boldsymbol{\lambda}, \boldsymbol{\gamma})$. First of all, the mappings

$$(\boldsymbol{\lambda}, \boldsymbol{\gamma}) \mapsto \frac{p_s}{2\bar{\alpha}}(1 - 2\eta(\boldsymbol{x}, s) - \langle \boldsymbol{\gamma}, \mathbf{1}\rangle) + \frac{\langle \boldsymbol{\gamma}, \boldsymbol{e}_s\rangle}{2\alpha_s} \quad \text{and} \quad (\boldsymbol{\lambda}, \boldsymbol{\gamma}) \mapsto -\frac{p_s}{2\bar{\alpha}}(1 - \langle \boldsymbol{\gamma}, \mathbf{1}\rangle) - \langle \boldsymbol{\lambda}, \boldsymbol{e}_s\rangle - \frac{\langle \boldsymbol{\gamma}, \boldsymbol{e}_s\rangle}{2\alpha_s} \ ,$$

are affine. Since taking the absolute value of an affine mapping gives a convex mapping (as a maximum between two affine, hence convex, functions), the sum of the absolute value of the first mapping with the second mapping is a convex function. Furthermore, the composition with the positive part function preserves convexity since this operation can be expressed as taking the maximum between two convex functions. Finally, by linearity of expectation, we notice that the objective is expressed as a finite sum of jointly convex functions and conclude that it is jointly convex in $(\boldsymbol{\lambda}, \boldsymbol{\gamma})$.

Optimality conditions for (λ^*, γ^*). The objective function of problem in Eq. (4) is not smooth everywhere due to the presence of absolute values and positive part functions. However, thanks to Assumption 3.1, the set of points at which the objective function is not differentiable has zero Lebesgue measure and can thus be ignored. The First-Order Optimality Conditions (FOOC) on the optimal Lagrange multipliers $(\boldsymbol{\lambda}^*, \boldsymbol{\gamma}^*)$ then read as

$$\begin{aligned}
\alpha_s &= \mathbb{P}_{\boldsymbol{X}|S=s}\left(\left|\frac{p_s}{2\bar{\alpha}}(1 - 2\eta(\boldsymbol{X}, s) - \langle \boldsymbol{\gamma}^*, \mathbf{1}\rangle) + \frac{\langle \boldsymbol{\gamma}^*, \boldsymbol{e}_s\rangle}{2\alpha_s}\right| \geq \frac{p_s}{2\bar{\alpha}}(1 - \langle \boldsymbol{\gamma}^*, \mathbf{1}\rangle) + \langle \boldsymbol{\lambda}^*, \boldsymbol{e}_s\rangle + \frac{\langle \boldsymbol{\gamma}^*, \boldsymbol{e}_s\rangle}{2\alpha_s}\right), \forall s \\
0 &= \sum_{s=1}^{K}\left(\frac{p_s}{\bar{\alpha}}\mathbf{1} - \frac{\boldsymbol{e}_s}{\alpha_s}\right)\mathbb{P}_{\boldsymbol{X}|S=s}\left(\min\left(2\eta(\boldsymbol{X}, S), \eta(\boldsymbol{X}, S) - \frac{\bar{\alpha}\lambda_s}{p_s}\right) \geq \frac{\bar{\alpha}\gamma_s}{p_s\alpha_s} + 1 - \bar{\gamma}\right) \ .
\end{aligned} \tag{FOOC}$$

Feasibility of $\tilde{g}$ for the primal problem Let us check that $\tilde{g}$ with (λ^*, γ^*) is actually feasible for the primal problem. Using the definition of $\tilde{g}$ and the first-order optimal condition on $\boldsymbol{\lambda}^*$ we obtain, for any $s \in [K]$,

$$\begin{aligned}
\mathbb{P}(\tilde{g}(\boldsymbol{X}, S) \neq r \mid S = s) &= \mathbb{P}_{\boldsymbol{X}|S=s}\left(\left|\frac{p_s}{2\bar{\alpha}}(1 - 2\eta(\boldsymbol{X}, s) - \langle \boldsymbol{\gamma}, \mathbf{1}\rangle) + \frac{\langle \boldsymbol{\gamma}, \boldsymbol{e}_s\rangle}{2\alpha_s}\right| \geq \frac{p_s}{2\bar{\alpha}}(1 - \langle \boldsymbol{\gamma}, \mathbf{1}\rangle) + \langle \boldsymbol{\lambda}, \boldsymbol{e}_s\rangle + \frac{\langle \boldsymbol{\gamma}, \boldsymbol{e}_s\rangle}{2\alpha_s}\right) \\
&= \alpha_s \ ,
\end{aligned}$$

which proves that $\tilde{g}$ satisfies the first set of constraints (i.e., controlled reject rate). For the demographic parity constraints, one obtains

$$\begin{aligned}
\mathbb{P}_{\boldsymbol{X}|S=s}(\tilde{g}(\boldsymbol{X}, S) = 1 \mid \tilde{g}(\boldsymbol{X}, S) \neq r) &= \frac{1}{\alpha_s}\mathbb{P}_{\boldsymbol{X}|S=s}(\tilde{g}(\boldsymbol{X}, S) = 1) \\
&= \frac{1}{\alpha_s}\mathbb{P}_{\boldsymbol{X}|S=s}\left(\min\left(2\eta(\boldsymbol{X}, S), \eta(\boldsymbol{X}, S) - \frac{\bar{\alpha}\lambda_s}{p_s}\right) \geq \frac{\alpha\gamma_s}{p_s\alpha_s} + 1 - \bar{\gamma}\right) \ , \\
\mathbb{P}_{(\boldsymbol{X},S)}(\tilde{g}(\boldsymbol{X}, S) = 1 \mid \tilde{g}(\boldsymbol{X}, S) \neq r) &= \sum_{s=1}^{K}\frac{p_s}{\bar{\alpha}}\mathbb{P}_{\boldsymbol{X}|S=s}\left(\min\left(2\eta(\boldsymbol{X}, S), \eta(\boldsymbol{X}, S) - \frac{\bar{\alpha}\lambda_s}{p_s}\right) \geq \frac{\bar{\alpha}\gamma_s}{p_s\alpha_s} + 1 - \bar{\gamma}\right) \ .
\end{aligned}$$

The second equation of the first-order optimality conditions (**FOOC**) guarantees that for, any $s \in [K]$,

$$\mathbb{P}_{\boldsymbol{X}|S=s}(\tilde{g}(\boldsymbol{X}, S) = 1 \mid \tilde{g}(\boldsymbol{X}, S) \neq r) = \mathbb{P}_{(\boldsymbol{X},S)}(\tilde{g}(\boldsymbol{X}, S) = 1 \mid \tilde{g}(\boldsymbol{X}, S) \neq r) \ ,$$

that is, it guarantees that the classifier $\tilde{g}$ satisfies the demographic parity constraint.

We conclude that the classifier $\tilde{g}$ is feasible for the primal problem and thus that the strong duality holds, implying the claimed expression for the optimal classifier with abstention.

2 AUXILIARY RESULTS

We will need a tight control on the sup-norm of the difference between CDF and empirical CDF. The next result is [Massart, 1990, Corollary 1].

Theorem 2.1. *Let $\boldsymbol{Z}, \boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n$ be $n + 1$ i.i.d. continuous random variable sampled from $\mathbb{P}$ on $\mathcal{Z}$, then for any $\delta > 0$, with probability at least $1 - \delta$,*

$$\sup_{z \in \mathbb{R}}\left|\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}\left(Z_i \leq z\right) - \mathbb{P}(Z \leq z)\right| \leq \sqrt{\frac{\log(2/\delta)}{2n}} \ .$$

3 CONTROL OF REJECT RATE

In this section we derive a finite sample control on the reject rate of the proposed procedure claimed in the main body. We start by recalling the result that we want to prove.

Proposition 3.1. *For all $\delta \in (0, 1)$, the proposed algorithm satisfies with probability at least $1 - \delta$ that*

$$|\mathbb{P}(\hat{g}(\boldsymbol{X}, S) \neq r \mid S = s) - \alpha_s| \leq \sqrt{\frac{2\log(2K/\delta)}{n_s}} + \frac{2}{n_s}, \quad \forall s \in [K] \enspace .$$

The rest of this section is devoted to the proof of this result. In what follows, all the derivations should be understood conditionally on $\hat{\eta}$. In simple words, the estimator $\hat{\eta}$ is treated as fixed and the only randomness comes from the unlabeled data. According to the definition of our estimator,

$$\mathbb{P}_{\boldsymbol{X}|S=s}\left(\hat{g}(\boldsymbol{X}, s) \neq r\right) = \mathbb{P}_{\boldsymbol{X}|S=s}\left(\hat{G}(\boldsymbol{X}, s, \hat{\boldsymbol{\lambda}}, \hat{\gamma}) > 0\right) \enspace ,$$

where $\hat{G}$ was defined in Section 4.

Using the triangle inequality we can upper bound $|\mathbb{P}_{\boldsymbol{X}|S=s}\left(\hat{G}(\boldsymbol{X}, s, \hat{\boldsymbol{\lambda}}, \hat{\gamma}) > 0\right) - \alpha_s|$ by two terms

$$\underbrace{\left|\mathbb{P}_{\boldsymbol{X}|S=s}\left(\hat{G}(\boldsymbol{X}, s, \hat{\boldsymbol{\lambda}}, \hat{\gamma}) > 0\right) - \hat{\mathbb{P}}_{\boldsymbol{X}|S=s}\left(\hat{G}(\boldsymbol{X}, s, \hat{\boldsymbol{\lambda}}, \hat{\gamma}) > 0\right)\right|}_{\mathtt{T}_1} + \underbrace{\left|\hat{\mathbb{P}}_{\boldsymbol{X}|S=s}\left(\hat{G}(\boldsymbol{X}, s, \hat{\boldsymbol{\lambda}}, \hat{\gamma}) > 0\right) - \alpha_s\right|}_{\mathtt{T}_2} \enspace , \tag{5}$$

which are treated separately.

Control of $\mathtt{T}_1$. The first term $\mathtt{T}_1$ can be controlled using tools from empirical process theory. One can directly observe by definition of $\hat{G}$ that

$$\begin{aligned}
\mathtt{T}_1 &\leq \sup_{(\boldsymbol{\lambda}, \boldsymbol{\gamma}) \in \mathbb{R}^K \times \mathbb{R}^K} \left|\mathbb{P}_{\boldsymbol{X}|S=s}\left(\hat{G}(\boldsymbol{X}, s, \boldsymbol{\lambda}, \gamma) > 0\right) - \hat{\mathbb{P}}_{\boldsymbol{X}|S=s}\left(\hat{G}(\boldsymbol{X}, s, \boldsymbol{\lambda}, \gamma) > 0\right)\right| \\
&\leq \sup_{(a,b) \in \mathbb{R} \times \mathbb{R}} \left|\mathbb{P}_{\boldsymbol{X}|S=s}\left(\left|\frac{p_s}{2\bar{\alpha}}\hat{\eta}(\boldsymbol{X}, S) - a\right| - a + b > 0\right) - \hat{\mathbb{P}}_{\boldsymbol{X}|S=s}\left(\left|\frac{p_s}{2\bar{\alpha}}\hat{\eta}(\boldsymbol{X}, S) - a\right| - a + b > 0\right)\right| \\
&\leq \sup_{(a,c) \in \mathbb{R} \times \mathbb{R}} \left|\mathbb{P}_{\boldsymbol{X}|S=s}\left(\left|\frac{p_s}{2\bar{\alpha}}\hat{\eta}(\boldsymbol{X}, S) - a\right| > c\right) - \hat{\mathbb{P}}_{\boldsymbol{X}|S=s}\left(\left|\frac{p_s}{2\bar{\alpha}}\hat{\eta}(\boldsymbol{X}, S) - a\right| > c\right)\right| \\
&\leq 2\sup_{a \in \mathbb{R}} \left|\mathbb{P}_{\boldsymbol{X}|S=s}(\hat{\eta}(\boldsymbol{X}, S) \leq a) - \hat{\mathbb{P}}_{\boldsymbol{X}|S=s}(\hat{\eta}(\boldsymbol{X}, S) \leq a)\right| \enspace ,
\end{aligned} \tag{6}$$

where we used the triangle inequality and the fact that $(\hat{\eta}(\boldsymbol{X}, S) \mid S = s)$ is a continuous random variable to obtain the last inequality.

By our assumption (see Remark 3.3), the random variables $\hat{\eta}(\boldsymbol{X}_i, s), (\hat{\eta}(\boldsymbol{X}, S) \mid S = s)$ for $i \in \mathcal{I}_s$ are i.i.d. continuous conditionally on $\hat{\eta}$. Thus, applying Theorem 2.1 we conclude that with probability at least $1 - \delta$ it holds that

$$\mathtt{T}_1 \leq \sqrt{\frac{2\log(2/\delta)}{n_s}} \enspace . \tag{7}$$

Control of $\mathtt{T}_2$. The control of the second term $\mathtt{T}_2$ requires a more involved analysis. Since $\hat{\boldsymbol{\lambda}}$ is a minimizer of (3), the first order optimality condition for convex non-smooth minimization problems states that for any $s \in [K]$, there exists $\rho_s \in [0, 1]$ such that

$$\alpha_s = \hat{\mathbb{P}}_{\boldsymbol{X}|S=s}\left(\hat{G}(\boldsymbol{X}, s, \hat{\boldsymbol{\lambda}}, \hat{\gamma}) > 0\right) + \rho_s\hat{\mathbb{P}}_{\boldsymbol{X}|S=s}\left(\hat{G}(\boldsymbol{X}, s, \hat{\boldsymbol{\lambda}}, \hat{\gamma}) = 0\right) \enspace .$$

Thus, the second term of Eq. (5) can be bounded as

$$\mathtt{T}_2 = \left|\hat{\mathbb{P}}_{\boldsymbol{X}|S=s}\left(\hat{G}(\boldsymbol{X}, s, \hat{\boldsymbol{\lambda}}, \hat{\gamma}) > 0\right) - \alpha_s\right| \leq \hat{\mathbb{P}}_{\boldsymbol{X}|S=s}\left(\hat{G}(\boldsymbol{X}, s, \hat{\boldsymbol{\lambda}}, \hat{\gamma}) = 0\right) \enspace . \tag{8}$$

The control of $\hat{\mathbb{P}}_{\boldsymbol{X}|S=s}\left(\hat{G}(\boldsymbol{X}, s, \hat{\boldsymbol{\lambda}}, \hat{\gamma}) = 0\right)$ is provided by the following result.

Lemma 3.2. *Assume that $(\hat{\eta}(\boldsymbol{X}, S) \mid S = s, \hat{\eta})$ is almost surely continuous, then for any $s \in [K]$, for any $(\boldsymbol{\lambda}, \boldsymbol{\gamma})$,*

$$\hat{\mathbb{P}}_{\boldsymbol{X}|S=s}\left(\hat{G}(\boldsymbol{X}, s, \boldsymbol{\lambda}, \boldsymbol{\gamma}) = 0\right) \leq \frac{2}{n_s} \enspace , \qquad \textit{a.s.}$$

Proof. Fix some arbitrary $(\boldsymbol{\lambda}, \boldsymbol{\gamma}) \in \mathbb{R}^K \times \mathbb{R}^K$. We recall that by definition of the empirical measure $\hat{\mathbb{P}}_{\boldsymbol{X}|s}$ we have

$$\hat{\mathbb{P}}_{\boldsymbol{X}|S=s}\left(\hat{G}(\boldsymbol{X}, s, \boldsymbol{\lambda}, \boldsymbol{\gamma}) = 0\right) = \frac{1}{n_s}\sum_{i=1}^{n_s} \mathbb{1}(\hat{G}(\boldsymbol{X}_i, s, \boldsymbol{\lambda}, \boldsymbol{\gamma}) = 0) \enspace . \tag{9}$$

The proof goes by contradiction. Assume that the event

$$\Omega := \left\{\frac{1}{n_s}\sum_{i=1}^{n_s} \mathbb{1}(\hat{G}(\boldsymbol{X}_i, s, \boldsymbol{\lambda}, \boldsymbol{\gamma}) = 0) \geq \frac{3}{n_s}\right\} \enspace , \tag{10}$$

happens with positive probability. Then, on Ω, there exist three indexes i_1, i_2, i_3 such that

$$\hat{G}(\boldsymbol{X}_{i_j}, s, \boldsymbol{\lambda}, \boldsymbol{\gamma}) = 0 \enspace , \quad j = 1, 2, 3 \enspace .$$

Furthermore, $\hat{G}(\boldsymbol{X}, s, \boldsymbol{\lambda}, \boldsymbol{\gamma}) = 0$ implies that either

$$\frac{\hat{p}_s}{\bar{\alpha}}\hat{\eta}(\boldsymbol{X}, s) + \langle \boldsymbol{\lambda}, \boldsymbol{e}_s \rangle = 0 \quad \text{ or } \quad \frac{\hat{p}_s}{\bar{\alpha}}(\hat{\eta}(\boldsymbol{X}, s) + \langle \boldsymbol{\gamma}, 1 \rangle - 1) - \langle \boldsymbol{\lambda}, \boldsymbol{e}_s \rangle + \frac{\langle \boldsymbol{\gamma}, \boldsymbol{e}_s \rangle}{\alpha_s} = 0 \enspace .$$

By the pigeonhole principle, on Ω, there exist $i, j \in \{i_1, i_2, i_3\}, i \neq j$ such that

$$\hat{\eta}(\boldsymbol{X}_i, s) = \hat{\eta}(\boldsymbol{X}_j, s) \enspace ,$$

which contradicts our assumption that $(\hat{\eta}(\boldsymbol{X}, S) \mid S = s, \hat{\eta})$ is continuous almost surely, hence the event Ω has zero probability.

Remark 3.3. *Recall that the assumption of continuity of $(\hat{\eta}(\boldsymbol{X}, S) \mid S = s, \hat{\eta})$ can always be fulfilled with the help of additional randomization. More formally, one needs to replace $\hat{\eta}$ by its smoothed version using additional randomization such as in Algorithm 1. To keep things simple, we avoid this technicality in our proof and simply assume that $(\hat{\eta}(\boldsymbol{X}, S) \mid S = s, \hat{\eta})$ is indeed continuous. The statement of this result is straightforwardly adapted to the perturbed version of $\hat{\eta}$.*

□

Lemma 3.2 allows to control the second term in Eq. (5) yielding

$$\mathtt{T}_2 \leq \frac{2}{n_s} \enspace . \tag{11}$$

Putting together. Substituting Eqs. (7) and (11) into Eq. (6), we deduce that for all $s \in [K]$ we have, with probability at least $1 - \delta$,

$$\left|\mathbb{P}_{\boldsymbol{X}|S=s}\left(\hat{G}(\boldsymbol{X}, s, \hat{\boldsymbol{\lambda}}, \hat{\boldsymbol{\gamma}}) > 0\right) - \alpha_s\right| \leq \sqrt{\frac{2\log(2/\delta)}{n_s}} + \frac{2}{n_s} \enspace .$$

Finally, taking the union bound we deduce that, with probability at least $1 - \delta$, we have for all $s \in [K]$

$$\left|\mathbb{P}_{\boldsymbol{X}|S=s}\left(\hat{G}(\boldsymbol{X}, s, \hat{\boldsymbol{\lambda}}, \hat{\boldsymbol{\gamma}}) > 0\right) - \alpha_s\right| \leq \sqrt{\frac{2\log(2K/\delta)}{n_s}} + \frac{2}{n_s} \enspace .$$

The proof of Proposition 3.1 is concluded.

4 CONTROL OF DEMOGRAPHIC PARITY VIOLATION

In this section we derive a finite sample control on the demographic parity violation of the proposed procedure, claimed in the main body. We start by recalling the result that we want to prove.

Proposition 4.1. *For any $\delta \in (0,1)$, the proposed algorithm satisfies with probability at least $1-\delta$, for any $s \in [K]$,*

$$\left|\mathbb{P}_{\boldsymbol{X}|S=s}\left(\hat{g}(\boldsymbol{X},s)=1 \mid \hat{g}(\boldsymbol{X},s)\neq r\right)-\mathbb{P}_{(\boldsymbol{X},S)}\left(\hat{g}(\boldsymbol{X},S)=1 \mid \hat{g}(\boldsymbol{X},S)\neq r\right)\right| \leq \frac{1}{\alpha_s}v_{n_s}^{\delta,K}+\frac{1}{\bar{\alpha}}\sum_{s=1}^{K}p_s v_{n_s}^{\delta,K} \quad ,$$

where

$$v_n^{\delta,K} := \left(3\sqrt{\frac{\log(4K/\delta)}{n}}+\frac{4}{n}\right) \quad .$$

Remark 4.2. *One can observe from the proof that the event on which Proposition 4.1 holds is contained in the event on which Proposition 3.1 holds, thus both hold simultaneously with high probability.*

The rest of this section is devoted to the proof of this result.

Problem splitting. Similarly to the control of the reject rate, we start by splitting our problem in several parts. Let us set

$$(\mathrm{DP}^s) := \left|\mathbb{P}_{\boldsymbol{X}|S=s}\left(\hat{g}(\boldsymbol{X},s)=1 \mid \hat{g}(\boldsymbol{X},s)\neq r\right)-\mathbb{P}_{(\boldsymbol{X},S)}\left(\hat{g}(\boldsymbol{X},S)=1 \mid \hat{g}(\boldsymbol{X},S)\neq r\right)\right| \quad ,$$

for all $s \in [K]$. The triangle inequality yields that (DP^s) can be bounded by five terms as follows

$$\begin{aligned}(\mathrm{DP}^s) \leq\ & \left|\mathbb{P}_{\boldsymbol{X}|S=s}\left(\hat{g}(\boldsymbol{X},s)=1 \mid \hat{g}(\boldsymbol{X},s)\neq r\right)-\alpha_s^{-1}\mathbb{P}_{\boldsymbol{X}|S=s}\left(\hat{g}(\boldsymbol{X},s)=1\right)\right| \\ & + \left|\alpha_s^{-1}\mathbb{P}_{\boldsymbol{X}|S=s}\left(\hat{g}(\boldsymbol{X},s)=1\right)-\alpha_s^{-1}\hat{\mathbb{P}}_{\boldsymbol{X}|S=s}\left(\hat{g}(\boldsymbol{X},s)=1\right)\right| \\ & + \left|\alpha_s^{-1}\hat{\mathbb{P}}_{\boldsymbol{X}|S=s}\left(\hat{g}(\boldsymbol{X},s)=1\right)-\bar{\alpha}^{-1}\sum_{s\in[K]}p_s\hat{\mathbb{P}}_{\boldsymbol{X}|S=s}\left(\hat{g}(\boldsymbol{X},s)=1\right)\right| \\ & + \left|\bar{\alpha}^{-1}\sum_{s\in[K]}p_s\hat{\mathbb{P}}_{\boldsymbol{X}|S=s}\left(\hat{g}(\boldsymbol{X},s)=1\right)-\bar{\alpha}^{-1}\sum_{s\in[K]}p_s\mathbb{P}_{\boldsymbol{X}|S=s}\left(\hat{g}(\boldsymbol{X},s)=1\right)\right| \\ & + \left|\bar{\alpha}^{-1}\sum_{s\in[K]}p_s\mathbb{P}_{\boldsymbol{X}|S=s}\left(\hat{g}(\boldsymbol{X},s)=1\right)-\mathbb{P}_{(\boldsymbol{X},S)}\left(\hat{g}(\boldsymbol{X},S)=1 \mid \hat{g}(\boldsymbol{X},S)\neq r\right)\right| \\ =\ & (\mathrm{DP}_1^s)+(\mathrm{DP}_2^s)+(\mathrm{DP}_3^s)+(\mathrm{DP}_4)+(\mathrm{DP}_5) \quad .\end{aligned}$$

The second and the fourth terms will be controlled using empirical process theory. We can get a bound on the first and fifth terms through our control of the reject rate. The third term is controlled via the first-order optimality condition on $\hat{\boldsymbol{\gamma}}$, similarly to Lemma 3.2.

High-probability event. Let us describe in details the high-probability event on which we will place ourselves for controlling all the terms appearing in the bound for (DP^s), uniformly over the groups $s \in [K]$.

Proposition 3.1 states that there exists an event Ω^r that holds with probability at least $1-K\delta$ and on which, for any $\delta \in (0, 1/K)$, the proposed algorithm satisfies with probability at least $1-K\delta$ that

$$|\mathbb{P}(\hat{g}(\boldsymbol{X},S)\neq r \mid S=s)-\alpha_s| \leq u_{n_s}^{\delta}, \quad \forall s\in[K], \qquad \text{where} \qquad u_n^{\delta} := \sqrt{\frac{2\log(2/\delta)}{n}}+\frac{2}{n} \quad , \quad \forall n \geq 1 \quad .$$

Furthermore, for any $s \in [K]$, using the fact that the random variable $(\eta(\boldsymbol{X},S) \mid S=s)$ is continuous, the event

$$\Omega_s^M := \left\{\sup_{a\in\mathbb{R}}\left|\mathbb{P}_{\boldsymbol{X}|S=s}\left(\eta(\boldsymbol{X},s)>a\right)-\hat{\mathbb{P}}_{\boldsymbol{X}|S=s}\left(\eta(\boldsymbol{X},s)>a\right)\right| \leq \sqrt{\frac{\log(2/\delta)}{2n_s}}\right\} \quad ,$$

holds with probability at least $1-\delta$ (by Massart's bound, recalled in Theorem 2.1). By a simple union bound argument, the intersection of those events, denoted by $\Omega^M := \cap_{s\in[K]}\Omega^M_s$, then holds with probability at least $1-2K\delta$.

In what follows we place ourselves on the event

$$\Omega_0 := \Omega^r \cap \Omega^M \ \ ,$$

which holds with probability at least $1-2K\delta$.

First-order optimality condition for $\hat{\boldsymbol{\gamma}}$. Recall that $(\hat{\boldsymbol{\lambda}}, \hat{\boldsymbol{\gamma}})$ is a solution of

$$\min_{(\boldsymbol{\lambda},\boldsymbol{\gamma})} \left\{ \langle \boldsymbol{\lambda}, \boldsymbol{\alpha} \rangle + \hat{\mathbb{E}}_{\boldsymbol{X}|S=s}(\hat{G}(\boldsymbol{X}, s, \boldsymbol{\lambda}, \boldsymbol{\gamma}))_+ \right\} \ \ ,$$

where the function $\hat{G}$ is defined as

$$\hat{G}(\boldsymbol{x}, s, \boldsymbol{\lambda}, \boldsymbol{\gamma}) = \left| \frac{\hat{p}_s}{2\bar{\alpha}}(1 - 2\hat{\eta}(\boldsymbol{x}, s) - \langle \boldsymbol{\gamma}, \mathbf{1} \rangle) + \frac{\langle \boldsymbol{\gamma}, \boldsymbol{e}_s \rangle}{2\alpha_s} \right| - \frac{\hat{p}_s}{2\bar{\alpha}}(1 - \langle \boldsymbol{\gamma}, \mathbf{1} \rangle) - \langle \boldsymbol{\lambda}, \boldsymbol{e}_s \rangle - \frac{\langle \boldsymbol{\gamma}, \boldsymbol{e}_s \rangle}{2\alpha_s} \ \ .$$

The positive part of $\hat{G}$ can be expressed as

$$(\hat{G}(\boldsymbol{x}, s, \boldsymbol{\lambda}, \boldsymbol{\gamma}))_+ = \max(0, m_+(\boldsymbol{x}, s, \boldsymbol{\lambda}, \boldsymbol{\gamma}), m_-(\boldsymbol{x}, s, \boldsymbol{\lambda}, \boldsymbol{\gamma})) \ \ ,$$

where we used the following short-hand notation

$$m_+(\boldsymbol{x}, s, \boldsymbol{\lambda}, \boldsymbol{\gamma}) = -\frac{p_s}{\bar{\alpha}}\hat{\eta}(\boldsymbol{x}, s) - \lambda_s \quad \text{and} \quad m_-(\boldsymbol{x}, s, \boldsymbol{\lambda}, \boldsymbol{\gamma}) = \frac{p_s}{\bar{\alpha}}(\hat{\eta}(\boldsymbol{x}, s) + \langle \boldsymbol{\gamma}, 1 \rangle - 1) - \frac{\langle \boldsymbol{\gamma}, \boldsymbol{e}_s \rangle}{\alpha_s} - \lambda_s \ \ .$$

Observe that since $\{m_-(\boldsymbol{x}, s, \boldsymbol{\lambda}, \boldsymbol{\gamma}) > \max(0, m_+(\boldsymbol{x}, s, \boldsymbol{\lambda}, \boldsymbol{\gamma}))\} \ \Leftrightarrow \ \{\hat{g}(\boldsymbol{X}, s) = 1\}$, then the first-order optimality condition on $\hat{\gamma}$ written in vector form reads as

$$\exists (\rho_s)_{s=1}^K \in [0,1]^K \quad \text{s.t.} \quad \sum_{s=1}^K \left(\frac{p_s}{\bar{\alpha}}\mathbf{1} - \frac{1}{\alpha_s}\boldsymbol{e}_s \right) \left(\hat{\mathbb{P}}_{\boldsymbol{X}|S=s}\left(\hat{g}(\boldsymbol{X}, s) = 1\right) + \rho_s \hat{\mathbb{P}}_{\boldsymbol{X}|S=s}\left(\Delta_s(\hat{\boldsymbol{\lambda}}, \hat{\gamma})\right) \right) = 0 \ \ ,$$

where we define the event $\Delta_s(\boldsymbol{\lambda}, \boldsymbol{\gamma}) := \{m_-(\boldsymbol{X}, s, \boldsymbol{\lambda}, \boldsymbol{\gamma}) = \max(0, m_+(\boldsymbol{X}, s, \boldsymbol{\lambda}, \boldsymbol{\gamma}))\}$. In scalar form the previous condition can be expressed as: for any $s \in [K]$, there exists $\rho_s \in [0,1]$ such that

$$\sum_{s=1}^K \frac{p_s}{\bar{\alpha}} \left(\hat{\mathbb{P}}_{\boldsymbol{X}|S=s}\left(\hat{g}(\boldsymbol{X}, s) = 1\right) + \rho_s \hat{\mathbb{P}}_{\boldsymbol{X}|S=s}\left(\Delta_s(\hat{\boldsymbol{\lambda}}, \hat{\gamma})\right) \right) = \frac{1}{\alpha_s} \left(\hat{\mathbb{P}}_{\boldsymbol{X}|S=s}\left(\hat{g}(\boldsymbol{X}, s) = 1\right) + \rho_s \hat{\mathbb{P}}_{\boldsymbol{X}|S=s}\left(\Delta_s(\hat{\boldsymbol{\lambda}}, \hat{\gamma})\right) \right) \ \ .$$

Control of the first term ($\mathbf{DP}_1^s$). Re-arranging terms and using the fact that $\mathbb{P}_{\boldsymbol{X}|S=s}\left(\hat{g}(\boldsymbol{X}, S) = 1\right) \le \mathbb{P}_{\boldsymbol{X}|S=s}\left(\hat{g}(\boldsymbol{X}, S) \neq r\right) = \mathbb{P}_{\boldsymbol{X}|S=s}\left(\hat{g}(\boldsymbol{X}, S) = 1\right) + \mathbb{P}_{\boldsymbol{X}|S=s}\left(\hat{g}(\boldsymbol{X}, S) = 0\right)$, we obtain

$$\begin{aligned}
(\mathrm{DP}_1^s) &:= \left| \mathbb{P}_{\boldsymbol{X}|S=s}\left(\hat{g}(\boldsymbol{X}, s) = 1 \mid \hat{g}(\boldsymbol{X}, s) \neq r\right) - \alpha_s^{-1}\mathbb{P}_{\boldsymbol{X}|S=s}\left(\hat{g}(\boldsymbol{X}, S) = 1\right) \right| \\
&= \left| \frac{1}{\alpha_s} - \frac{1}{\mathbb{P}_{\boldsymbol{X}|S=s}\left(\hat{g}(\boldsymbol{X}, s) \neq r\right)} \right| \mathbb{P}_{\boldsymbol{X}|S=s}\left(\hat{g}(\boldsymbol{X}, S) = 1\right) \\
&\le \left| \frac{1}{\alpha_s} - \frac{1}{\mathbb{P}_{\boldsymbol{X}|S=s}\left(\hat{g}(\boldsymbol{X}, s) \neq r\right)} \right| \mathbb{P}_{\boldsymbol{X}|S=s}\left(\hat{g}(\boldsymbol{X}, S) \neq r\right) = \frac{1}{\alpha_s} \left| \mathbb{P}_{\boldsymbol{X}|S=s}\left(\hat{g}(\boldsymbol{X}, S) \neq r\right) - \alpha_s \right| \ \ .
\end{aligned}$$

We can conclude that, on the event Ω_0, for all $s \in [K]$

$$(\mathrm{DP}_1^s) \le \frac{1}{\alpha_s} \left| \mathbb{P}_{\boldsymbol{X}|S=s}\left(\hat{g}(\boldsymbol{X}, S) \neq r\right) - \alpha_s \right| \le \frac{u_{n_s}^{\delta}}{\alpha_s} \ \ .$$

Control of the second term ($\mathbf{DP}_2^s$). The second term is given by the empirical process

$$(\mathrm{DP}_2^s) := \alpha_s^{-1}\left|\mathbb{P}_{\boldsymbol{X}|S=s}\left(\hat{g}(\boldsymbol{X},S)=1\right) - \hat{\mathbb{P}}_{\boldsymbol{X}|S=s}\left(\hat{g}(\boldsymbol{X},S)=1\right)\right| \enspace .$$

The event $\{\hat{g}(\boldsymbol{X},s)=1\}$ is the same as the event

$$S(\boldsymbol{\lambda},\boldsymbol{\gamma}) := \left\{\left|\frac{p_s}{2\bar{\alpha}}(1-2\hat{\eta}(\boldsymbol{X},s)-\langle\boldsymbol{1},\hat{\boldsymbol{\gamma}}\rangle)+\frac{\hat{\boldsymbol{\gamma}}_s}{2\alpha_s}\right| > \frac{p_s}{2\bar{\alpha}}(1-\langle\boldsymbol{1},\boldsymbol{\gamma}\rangle)+\hat{\boldsymbol{\lambda}}_s+\frac{\hat{\boldsymbol{\gamma}}_s}{2\alpha_s}, \quad 2\hat{\eta}(\boldsymbol{X},s)\geq 1+\frac{\bar{\alpha}\hat{\boldsymbol{\gamma}}_s}{\alpha_s p_s}-\langle\hat{\boldsymbol{\gamma}},\boldsymbol{1}\rangle\right\} \enspace ,$$

which can be compressed to

$$\left\{\hat{\eta}(\boldsymbol{X},s) > \max\left(\frac{1}{2}+\frac{\bar{\alpha}\hat{\boldsymbol{\gamma}}_s}{2\alpha_s p_s}-\frac{1}{2}\left\langle\hat{\boldsymbol{\gamma}},\boldsymbol{1}\right\rangle, \frac{\bar{\alpha}}{p_s}\left(\hat{\boldsymbol{\lambda}}_s+\frac{\hat{\boldsymbol{\gamma}}_s}{\alpha_s}\right)+1-\langle\boldsymbol{1},\boldsymbol{\gamma}\rangle\right)\right\} \enspace .$$

Following this observation, the second term (DP_2^s) of interest is expressed as

$$(\mathrm{DP}_2^s) = \alpha_s^{-1}\sup_{(\boldsymbol{\lambda},\boldsymbol{\gamma})}\left|\mathbb{P}_{\boldsymbol{X}|S=s}\left(S(\boldsymbol{\lambda},\boldsymbol{\gamma})\right)-\hat{\mathbb{P}}_{\boldsymbol{X}|S=s}\left(S(\boldsymbol{\lambda},\boldsymbol{\gamma})\right)\right| \leq \alpha_s^{-1}\sup_{a\in\mathbb{R}}\left|\mathbb{P}_{\boldsymbol{X}|S=s}\left(\hat{\eta}(\boldsymbol{X},s)>a\right)-\hat{\mathbb{P}}_{\boldsymbol{X}|S=s}\left(\hat{\eta}(\boldsymbol{X},s)>a\right)\right| \enspace .$$

On the event Ω_0, which is contained in the event Ω_s^M, we have for all $s\in[K]$

$$(\mathrm{DP}_2^s) \leq \frac{1}{\alpha_s}\sqrt{\frac{\log(2/\delta)}{2n_s}} \enspace .$$

Control of the third term ($\mathbf{DP}_3^s$). The third term can be controlled with the first-order optimality condition on $\hat{\boldsymbol{\gamma}}$ similarly to Lemma 3.2 and multiple triangle inequalities as

$$\begin{aligned}(\mathrm{DP}_3^s) &:= \left|\alpha_s^{-1}\hat{\mathbb{P}}_{\boldsymbol{X}|S=s}\left(\hat{g}(\boldsymbol{X},S)=1\right)-\bar{\alpha}^{-1}\sum_{s\in[K]}p_s\hat{\mathbb{P}}_{\boldsymbol{X}|S=s}\left(\hat{g}(\boldsymbol{X},S)=1\right)\right| \\ &= \left|\frac{\rho_s}{\alpha_s}\hat{\mathbb{P}}_{\boldsymbol{X}|S=s}\left(\Delta_s(\hat{\boldsymbol{\lambda}},\hat{\boldsymbol{\gamma}})\right)-\sum_{s=1}^{K}\frac{p_s}{\bar{\alpha}}\rho_s\hat{\mathbb{P}}_{\boldsymbol{X}|S=s}\left(\Delta_s(\hat{\boldsymbol{\lambda}},\hat{\boldsymbol{\gamma}})\right)\right| \\ &\leq \frac{1}{\alpha_s}\hat{\mathbb{P}}_{\boldsymbol{X}|S=s}\left(\Delta_s(\hat{\boldsymbol{\lambda}},\hat{\boldsymbol{\gamma}})\right)+\sum_{s=1}^{K}\frac{p_s}{\bar{\alpha}}\hat{\mathbb{P}}_{\boldsymbol{X}|S=s}\left(\Delta_s(\hat{\boldsymbol{\lambda}},\hat{\boldsymbol{\gamma}})\right) \enspace .\end{aligned}$$

The following lemma gives an almost sure upper bound on $\hat{\mathbb{P}}_{\boldsymbol{X}|S=s}\left(\Delta_s(\hat{\boldsymbol{\lambda}},\hat{\boldsymbol{\gamma}})\right)$ for any $s\in[K]$.

Lemma 4.3. *Assume that $(\hat{\eta}(\boldsymbol{X},S)\mid S=s,\hat{\eta})$ is almost surely continuous, then for any $s\in[K]$, for any $(\boldsymbol{\lambda},\boldsymbol{\gamma})$,*

$$\hat{\mathbb{P}}_{\boldsymbol{X}|S=s}\left(\Delta_s(\hat{\boldsymbol{\lambda}},\hat{\boldsymbol{\gamma}})\right) \leq \frac{2}{n_s}, \qquad \textit{a.s.}$$

Proof. This proof is similar to proof of Lemma 3.2. Assume by contradiction that the stated bound is not true. Then, it happens with positive probability that

$$\frac{1}{n_s}\sum_{i=1}^{n_s}\mathbb{1}\left\{m_-(\boldsymbol{X}_i,s,\boldsymbol{\lambda},\boldsymbol{\gamma})=\max(0,m_+(\boldsymbol{X}_i,s,\boldsymbol{\lambda},\boldsymbol{\gamma}))\right\} \geq \frac{3}{n_s} \enspace ,$$

which implies that there exist a triplet i_1,i_2,i_3 such that

$$m_-(\boldsymbol{X}_{i_j},s,\boldsymbol{\lambda},\boldsymbol{\gamma})=\max(0,m_+(\boldsymbol{X}_{i_j},s,\boldsymbol{\lambda},\boldsymbol{\gamma})), \quad \text{for } j=1,2,3 \enspace .$$

By the pigeonhole principle, there must exist a couple $(i,j), i\neq j$ among this triplet such that either

$$m_-(\boldsymbol{X}_i,s,\boldsymbol{\lambda},\boldsymbol{\gamma})=m_-(\boldsymbol{X}_j,s,\boldsymbol{\lambda},\boldsymbol{\gamma}) \quad \text{or} \quad m_-(\boldsymbol{X}_i,s,\boldsymbol{\lambda},\boldsymbol{\gamma})-m_+(\boldsymbol{X}_i,s,\boldsymbol{\lambda},\boldsymbol{\gamma})=m_-(\boldsymbol{X}_j,s,\boldsymbol{\lambda},\boldsymbol{\gamma})-m_+(\boldsymbol{X}_j,s,\boldsymbol{\lambda},\boldsymbol{\gamma}) \enspace .$$

In both cases one must have $\hat{\eta}(\boldsymbol{X}_i,s)=\hat{\eta}(\boldsymbol{X}_j,s)$ which happens with probability 0 by the continuity assumption and leads to a contradiction. The proof of lemma is concluded □

Plugging in the bounds from Lemma 4.3 yields for all $s\in[K]$

$$(\mathrm{DP}_3^s) \leq \frac{2}{n_s\alpha_s}+\frac{2}{\bar{\alpha}}\sum_{s=1}^{K}\frac{p_s}{n_s} \enspace .$$

Control of the fourth term $(\mathbf{DP}_4)$. The fourth term can be seen as a sum of empirical processes:

$$\begin{aligned}(\mathrm{DP}_4) &:= \bar{\alpha}^{-1}\left|\sum_{s\in[K]} p_s\hat{\mathbb{P}}_{\boldsymbol{X}|S=s}\left(\hat{g}(\boldsymbol{X},S)=1\right) - \sum_{s\in[K]} p_s\mathbb{P}_{\boldsymbol{X}|S=s}\left(\hat{g}(\boldsymbol{X},S)=1\right)\right| \\ &\le \bar{\alpha}^{-1}\sum_{s=1}^{K} p_s\left|\hat{\mathbb{P}}_{\boldsymbol{X}|S=s}\left(\hat{g}(\boldsymbol{X},S)=1\right) - \mathbb{P}_{\boldsymbol{X}|S=s}\left(\hat{g}(\boldsymbol{X},S)=1\right)\right| \ .\end{aligned}$$

On the event Ω_0, we can control the fourth term from the bound we have on the second term (which holds uniformly over the classes s) as

$$(\mathrm{DP}_4) \le \frac{1}{\bar{\alpha}}\sum_{s\in K} p_s\sqrt{\frac{\log(2/\delta)}{2n_s}} \ .$$

Control of the fifth term $(\mathbf{DP}_5)$. Finally, the fifth term can be bounded using the same trick as for the first term. On the event Ω_0, we have

$$\begin{aligned}(\mathrm{DP}_5) &:= \left|\bar{\alpha}^{-1}\sum_{s\in[K]} p_s\mathbb{P}_{\boldsymbol{X}|S=s}\left(\hat{g}(\boldsymbol{X},S)=1\right) - \mathbb{P}_{(\boldsymbol{X},S)}\left(\hat{g}(\boldsymbol{X},s)=1 \mid \hat{g}(\boldsymbol{X},s)\neq r\right)\right| \\ &= \left|\frac{1}{\bar{\alpha}} - \frac{1}{\sum_{s=1}^{K} p_s\mathbb{P}_{\boldsymbol{X}|S=s}\left(\hat{g}(\boldsymbol{X},s)\neq r\right)}\right|\sum_{s=1}^{K} p_s\mathbb{P}_{\boldsymbol{X}|S=s}\left(\hat{g}(\boldsymbol{X},S)=1\right) \\ &\le \left|\frac{1}{\bar{\alpha}} - \frac{1}{\sum_{s=1}^{K} p_s\mathbb{P}_{\boldsymbol{X}|S=s}\left(\hat{g}(\boldsymbol{X},s)\neq r\right)}\right|\sum_{s=1}^{K} p_s\mathbb{P}_{\boldsymbol{X}|S=s}\left(\hat{g}(\boldsymbol{X},S)\neq r\right) \\ &= \frac{1}{\bar{\alpha}}\left|\sum_{s=1}^{K} p_s(\mathbb{P}_{\boldsymbol{X}|S=s}\left(\hat{g}(\boldsymbol{X},s)\neq r\right) - \alpha_s)\right| \le \frac{1}{\bar{\alpha}}\sum_{s=1}^{K} p_s u^{\delta}_{n_s} \ .\end{aligned}$$

Conclusion. Putting everything together, we have shown that, on the event $\mathbf{A}$ which holds with probability at least $1-2K\delta$, we have, for any $s\in[K]$,

$$(\mathrm{DP}^s) \le \frac{1}{\alpha_s}\left(3\sqrt{\frac{\log(2/\delta)}{2n_s}} + \frac{4}{n_s}\right) + \frac{2}{\bar{\alpha}}\sum_{s=1}^{K} p_s\left(3\sqrt{\frac{\log(2/\delta)}{2n_s}} + \frac{4}{n_s}\right) \ .$$

5 CONTROL OF THE EXCESS RISK

Define the sequence

$$u_n^{\delta,K} := \sqrt{\frac{2\log(4K/\delta)}{n}} + \frac{2}{n} \ , \quad \forall n\ge 1 \ .$$

We state and prove a slightly more precise bound then the one presented in the main body.

Proposition 5.1. *Assume that $u_{n_s}^{\delta,K} < \alpha_s < 1-\frac{2}{n_s}$ for any $s\in[K]$ and that Assumption 3.1 holds. Then, for any $\delta\in(0,1)$, the excess risk of the post-processing classifier with abstention $\hat{g}$ defined in Eq (2) satisfies, with probability at least $1-\delta$,*

$$\mathcal{E}(\hat{g}) \le \left(\frac{1}{\bar{\alpha}} + \frac{1}{\bar{\alpha}-\sum_s p_s u_{n_s}^{\delta,K}}\right)\|\eta-\hat{\eta}\|_1 + 6\sum_{s=1}^{K}\left(\frac{p_s}{\bar{\alpha}} + \frac{1}{\alpha_s}\right)u_{n_s}^{\delta,K} \ .$$

A quick inspection of the proof shows that the high-probability event on which the stated bound holds is the same as the event on which Proposition 4.1 holds, which is contained in the event on which Proposition 3.1 holds. Thus, we can control the excess risk and the violation of the constraints on the same high-probability event.

Proof. Since, using Assumption 3.1 we have established strong duality, the following equality holds

$$R(g^*) = \max_{(\boldsymbol{\lambda},\boldsymbol{\gamma})} \left\{ \sum_{s=1}^{K} \mathbb{E}_{\boldsymbol{X}|S=s} \left(\frac{p_s}{2\bar{\alpha}}(1-\bar{\gamma}) + \lambda_s + \frac{\gamma_s}{2\alpha_s} - \left| \frac{p_s}{2\bar{\alpha}}(1-2\eta(\boldsymbol{x},s)-\bar{\gamma}) + \frac{\gamma_s}{2\alpha_s} \right| \right)_{-} - \sum_{s=1}^{K} \lambda_s \alpha_s \right\} \quad . \tag{12}$$

Besides, we can bound the risk of any classifier g as

$$\begin{aligned} R(g) &= \sum_{s=1}^{K} \frac{p_s}{\mathbb{P}(g(\boldsymbol{X},S) \neq r)} \mathbb{E}_{\boldsymbol{X}|S=s} \left[(1-\eta(\boldsymbol{X},s)\mathbb{1}_{g(\boldsymbol{X},s)=1} + \eta(\boldsymbol{X},s)\mathbb{1}_{g(\boldsymbol{X},s)=0} \right] \\ &\leq \sum_{s=1}^{K} \frac{p_s}{\mathbb{P}(g(\boldsymbol{X},S) \neq r)} \mathbb{E}_{\boldsymbol{X}|S=s} \left[(1-\hat{\eta}(\boldsymbol{X},s)\mathbb{1}_{g(\boldsymbol{X},s)=1} + \hat{\eta}(\boldsymbol{X},s)\mathbb{1}_{g(\boldsymbol{X},s)=0} \right] + \frac{\|\eta - \hat{\eta}\|_1}{\mathbb{P}(g(\boldsymbol{X},S) \neq r)} \quad . \end{aligned} \tag{13}$$

Setting $\mathbb{A}_s(g) := \frac{p_s}{\bar{\alpha}} \mathbb{E}_{\boldsymbol{X}|S=s} \left[(1-\hat{\eta}(\boldsymbol{X},s))\mathbb{1}_{g(\boldsymbol{X},s)=1} + \hat{\eta}(\boldsymbol{X},s)\mathbb{1}_{g(\boldsymbol{X},s)=0} \right]$, we have for any classifier g,

$$R(g) \leq \sum_{s=1}^{K} \mathbb{A}_s(g) + \frac{\|\eta - \hat{\eta}\|_1}{\mathbb{P}(g(\boldsymbol{X},S) \neq r)} + \frac{1}{\bar{\alpha}} \left| \mathbb{P}(g(\boldsymbol{X},S) \neq r) - \bar{\alpha} \right| \quad .$$

In what follows we bound $\mathrm{r}_1(g) := \sum_{s=1}^{K} \mathbb{A}_s(g)$. Re-arranging terms we deduce that

$$\begin{aligned} \mathrm{r}_1(g) = &\sum_{s=1}^{K} \frac{p_s}{\bar{\alpha}} \mathbb{E}_{\boldsymbol{X}|S=s} \left[(1-\hat{\eta}(\boldsymbol{X},S))\mathbb{1}_{g(\boldsymbol{X},S)=1} + \hat{\eta}(\boldsymbol{X},S)\mathbb{1}_{g(\boldsymbol{X},S)=0} \right] \pm \sum_{s=1}^{K} \hat{\lambda}_s \left\{ \mathbb{E}_{\boldsymbol{X}|S=s} \left[\mathbb{1}_{g(\boldsymbol{X},S)=1} + \mathbb{1}_{g(\boldsymbol{X},S)=0} \right] - \alpha_s \right\} \\ &\pm \sum_{s=1}^{K} \frac{\hat{\gamma}_s}{\alpha_s} \mathbb{E}_{\boldsymbol{X}|S=s} \left[\mathbb{1}_{g(\boldsymbol{X},S)=1} \right] - \left(\sum_{s'=1}^{K} \hat{\gamma}_{s'} \right) \left(\sum_{s=1}^{K} \frac{p_s}{\bar{\alpha}} \mathbb{E}_{\boldsymbol{X}|S=s} \left[\mathbb{1}_{g(\boldsymbol{X},S)=1} \right] \right) \\ &= \sum_{s=1}^{K} \mathbb{E}_{\boldsymbol{X}|S=s} \left[\hat{H}_{(\boldsymbol{X},s)}(g, \hat{\boldsymbol{\lambda}}, \hat{\boldsymbol{\gamma}}) \right] - \sum_{s=1}^{K} \hat{\lambda}_s \alpha_s - \sum_{s=1}^{K} \hat{\lambda}_s \left\{ \mathbb{E}_{\boldsymbol{X}|S=s} \left[\mathbb{1}_{g(\boldsymbol{X},S)=1} + \mathbb{1}_{g(\boldsymbol{X},S)=0} \right] - \alpha_s \right\} \\ &- \sum_{s=1}^{K} \frac{\hat{\gamma}_s}{\alpha_s} \mathbb{E}_{\boldsymbol{X}|S=s} \left[\mathbb{1}_{g(\boldsymbol{X},S)=1} \right] - \left(\sum_{s'=1}^{K} \hat{\gamma}_{s'} \right) \left(\sum_{s=1}^{K} \frac{p_s}{\bar{\alpha}} \mathbb{E}_{\boldsymbol{X}|S=s} \left[\mathbb{1}_{g(\boldsymbol{X},S)=1} \right] \right) \quad , \end{aligned}$$

where

$$\hat{H}_{(\boldsymbol{x},s)}(g, \boldsymbol{\lambda}, \boldsymbol{\gamma}) = \begin{cases} 0, & \text{if } g(\boldsymbol{x},s) = r \\ \frac{p_s}{\bar{\alpha}} \hat{\eta}(\boldsymbol{x},s) + \lambda_s, & \text{if } g(\boldsymbol{x},s) = 0 \\ \frac{p_s}{\bar{\alpha}}(1-\hat{\eta}(\boldsymbol{x},s)-\bar{\gamma}) + \lambda_s + \frac{\gamma_s}{\alpha_s}, & \text{if } g(\boldsymbol{x},s) = 1 \end{cases} \quad ,$$

with $\bar{\gamma} = \sum_{s=1}^{K} \gamma_s$. Note that, by the definition of $\hat{g}$, it holds that

$$\sum_{s=1}^{K} \mathbb{E}_{\boldsymbol{X}|S=s} \left[\hat{H}_{(\boldsymbol{X},s)}(\hat{g}, \hat{\boldsymbol{\lambda}}, \hat{\boldsymbol{\gamma}}) \right] = \mathbb{E}(-\hat{G}(\boldsymbol{X}, s, \hat{\boldsymbol{\lambda}}, \hat{\boldsymbol{\gamma}}))_{-} \quad .$$

Thus, the following holds

$$\begin{aligned} \mathrm{r}_1(\hat{g}) = &\sum_{s=1}^{K} \mathbb{E}_{\boldsymbol{X}|S=s} \left(\frac{p_s}{2\bar{\alpha}}(1-\bar{\hat{\gamma}}) + \hat{\lambda}_s + \frac{\hat{\gamma}_s}{2\alpha_s} - \left| \frac{p_s}{2\bar{\alpha}}(1-2\hat{\eta}(\boldsymbol{X},s)-\bar{\hat{\gamma}}) + \frac{\hat{\gamma}_s}{2\alpha_s} \right| \right)_{-} - \sum_{s=1}^{K} \hat{\lambda}_s \alpha_s \\ &- \sum_{s=1}^{K} \hat{\lambda}_s \left(\mathbb{P}(\hat{g}(\boldsymbol{X},S) \neq r \mid S=s) - \alpha_s \right) \\ &- \sum_{s=1}^{K} \hat{\gamma}_s \left(\frac{\mathbb{P}(\hat{g}(\boldsymbol{X},S) = 1 \mid S=s)}{\alpha_s} - \sum_{s'=1}^{K} \frac{p_{s'}}{\bar{\alpha}} \mathbb{P}(\hat{g}(\boldsymbol{X},S) = 1 \mid S=s') \right) \quad . \end{aligned} \tag{14}$$

Substituting Eq. (14) into Eq. (13) we obtain the following upper bound on $R(\hat{g})$

$$
\begin{aligned}
R(\hat{g}) \leq \sum_{s=1}^{K} \mathbb{E}_{\boldsymbol{X}|S=s} \left(\frac{p_s}{2\bar{\alpha}}(1 - \hat{\bar{\gamma}}) + \hat{\lambda}_s + \frac{\hat{\gamma}_s}{2\alpha_s} - \left| \frac{p_s}{2\bar{\alpha}}(1 - 2\hat{\eta}(\boldsymbol{X}, s) - \hat{\bar{\gamma}}) + \frac{\hat{\gamma}_s}{2\alpha_s} \right| \right)_- - \sum_{s=1}^{K} \hat{\lambda}_s \alpha_s \\
- \sum_{s=1}^{K} \hat{\lambda}_s \left(\mathbb{P}(\hat{g}(\boldsymbol{X}, S) \neq r \mid S = s) - \alpha_s \right) - \sum_{s=1}^{K} \hat{\gamma}_s \left(\frac{\mathbb{P}(\hat{g}(\boldsymbol{X}, S) = 1 \mid S = s)}{\alpha_s} - \sum_{s'=1}^{K} \frac{p_{s'}}{\bar{\alpha}} \mathbb{P}(\hat{g}(\boldsymbol{X}, S) = 1 \mid S = s') \right) \\
+ \frac{\|\eta - \hat{\eta}\|_1}{\mathbb{P}(\hat{g}(\boldsymbol{X}, S) \neq r)} + \frac{1}{\bar{\alpha}} \left| \mathbb{P}(\hat{g}(\boldsymbol{X}, S) \neq r) - \bar{\alpha} \right| \quad ,
\end{aligned}
$$

which holds almost surely.

Define the excess risk $\mathcal{E}(\hat{g}) := R(\hat{g}) - R(g^*)$. Note that, using the fact that mapping $x \mapsto (x)_-$ is 1-Lipschitz followed by the triangle inequality, the difference

$$
\left| \left(\frac{p_s}{2\bar{\alpha}}(1 - \bar{\gamma}) + \lambda_s + \frac{\gamma_s}{2\alpha_s} - \left| \frac{p_s}{2\bar{\alpha}}(1 - 2\hat{\eta}(\boldsymbol{x}, s) - \bar{\gamma}) + \frac{\gamma_s}{2\alpha_s} \right| \right)_- - \left(\frac{p_s}{2\bar{\alpha}}(1 - \bar{\gamma}) + \lambda_s + \frac{\gamma_s}{2\alpha_s} - \left| \frac{p_s}{2\bar{\alpha}}(1 - 2\eta(\boldsymbol{x}, s) - \bar{\gamma}) + \frac{\gamma_s}{2\alpha_s} \right| \right)_- \right| \quad ,
$$

can be upper bounded by $\frac{p_s}{\bar{\alpha}} |\hat{\eta}(\boldsymbol{x}, s) - \eta(\boldsymbol{x}, s)|$, for any $(\boldsymbol{x}, s, \boldsymbol{\lambda}, \boldsymbol{\gamma})$. Thus, replacing $(\boldsymbol{\lambda}^*, \boldsymbol{\gamma}^*)$ by $(\hat{\boldsymbol{\lambda}}, \hat{\boldsymbol{\gamma}})$ (recall that $(\boldsymbol{\lambda}^*, \boldsymbol{\gamma}^*)$ is optimal) in the expression for $R(g^*)$ in Eq. (12) we obtain

$$
\begin{aligned}
\mathcal{E}(\hat{g}) \leq & \frac{\|\eta - \hat{\eta}\|_1}{\bar{\alpha}} + \frac{1}{\bar{\alpha}} \left| \mathbb{P}(\hat{g}(\boldsymbol{X}, S) \neq r) - \bar{\alpha} \right| + \frac{\|\eta - \hat{\eta}\|_1}{\mathbb{P}(g(\boldsymbol{X}, S) \neq r)} + \sum_{s=1}^{K} |\hat{\lambda}_s| \left| \mathbb{P}(\hat{g}(\boldsymbol{X}, S) \neq r \mid S = s) - \alpha_s \right| \\
& + \sum_{s=1}^{K} |\hat{\gamma}_s| \left| \frac{\mathbb{P}(\hat{g}(\boldsymbol{X}, S) = 1 \mid S = s)}{\alpha_s} - \sum_{s'=1}^{K} \frac{p_{s'}}{\bar{\alpha}} \mathbb{P}(\hat{g}(\boldsymbol{X}, S) = 1 \mid S = s') \right| \quad .
\end{aligned}
\tag{15}
$$

In what follows we provide a control of all the terms appearing in the above derived bound.

Using the fact that on the event of Proposition 4.1 we have, with probability at least $1 - 2K\delta$,

$$
\left| \mathbb{P}(\hat{g}(\boldsymbol{X}, S) \neq r \mid S = s) - \alpha_s \right| \leq u^{\delta}_{n_s}, \quad \forall s \in [K], \quad \text{with} \quad u^{\delta}_{n} := \sqrt{\frac{2\log(2/\delta)}{n}} + \frac{2}{n} \quad , \quad \forall n \geq 1 \ ,
$$

we deduce that with probability at least $1 - 2K\delta$ the following three inequalities hold

$$
\begin{aligned}
\frac{1}{\bar{\alpha}} \left| \mathbb{P}(g(\boldsymbol{X}, S) \neq r) - \bar{\alpha} \right| &\leq \frac{1}{\bar{\alpha}} \sum_{s=1}^{K} p_s u^{\delta}_{n_s} \quad , \\
\frac{\|\eta - \hat{\eta}\|_1}{\mathbb{P}(\hat{g}(\boldsymbol{X}, S) \neq r)} &\leq \frac{\|\eta - \hat{\eta}\|_1}{\bar{\alpha} - \sum_s p_s u^{\delta}_{n_s}} \quad , \\
\sum_{s=1}^{K} |\hat{\lambda}_s| \left| \mathbb{P}(\hat{g}(\boldsymbol{X}, S) \neq r \mid S = s) - \alpha_s \right| &\leq \sum_{s=1}^{K} |\hat{\lambda}_s| u^{\delta}_{n_s} \quad .
\end{aligned}
\tag{16}
$$

Note that by the assumption of the proposition, the term $\bar{\alpha} - \sum_s p_s u^{\delta}_{n_s} > 0$.

Furthermore, on the same event, using the notations of the proof of Proposition 4.1, we have for any $s \in [K]$

$$
\left| \frac{\mathbb{P}_{\boldsymbol{X}|S=s}\left(\hat{g}(\boldsymbol{X}, S) = 1\right)}{\alpha_s} - \sum_{s'=1}^{K} \frac{p_{s'}}{\bar{\alpha}} \mathbb{P}_{\boldsymbol{X}|S=s'}\left(\hat{g}(\boldsymbol{X}, S) = 1\right) \right| \leq (\mathrm{DP}^s_2) + (\mathrm{DP}^s_3) + (\mathrm{DP}_4) \leq \frac{1}{\alpha_s} v^{\delta}_{n_s} + \frac{2}{\bar{\alpha}} \sum_{s=1}^{K} p_s v^{\delta}_{n_s} \quad .
\tag{17}
$$

where $v^{\delta}_{n} = \sqrt{\frac{\log(2/\delta)}{2n}} + \frac{2}{n}$. Hence, substituting Eqs. (16) and (17) into Eq. (15) we deduce that with probability at least

$1 - 2K\delta$

$$\begin{aligned}
\mathcal{E}(\hat{g}) &\leq \left(\frac{1}{\bar{\alpha}} + \frac{1}{\bar{\alpha} - \sum_s p_s u^\delta_{n_s}}\right) \|\eta - \hat{\eta}\|_1 + \sum_{s=1}^K \left(\frac{p_s}{\bar{\alpha}} + |\hat{\lambda}_s|\right) u^\delta_{n_s} + \sum_{s=1}^K \left(\frac{|\hat{\gamma}_s|}{\alpha_s} + \frac{2p_s}{\bar{\alpha}}(\sum_{s'}|\hat{\gamma}_{s'}|)\right) v^\delta_{n_s}\\
&= \left(\frac{1}{\bar{\alpha}} + \frac{1}{\bar{\alpha} - \sum_s p_s u^\delta_{n_s}}\right) \|\eta - \hat{\eta}\|_1 + \sum_{s=1}^K \left(\frac{2p_s}{\bar{\alpha}} + 2|\hat{\lambda}_s| + \frac{|\hat{\gamma}_s|}{\alpha_s} + \frac{2p_s}{\bar{\alpha}}(\sum_{s'}|\hat{\gamma}_s|)\right) \sqrt{\frac{\log(1/\delta)}{2n_s}}\\
&+ \sum_{s=1}^K \left(\frac{p_s}{\bar{\alpha}} + |\hat{\lambda}_s| + \frac{|\hat{\gamma}_s|}{\alpha_s} + \frac{2p_s}{\bar{\alpha}}(\sum_{s'}|\hat{\gamma}_{s'}|)\right) \frac{2}{n_s} \quad .
\end{aligned}$$

In order to finish the proof it remains to provide a bound on $|\hat{\lambda}_s|$ and $|\hat{\gamma}_s|$. Proposition 5.2, proven below, establishes this bound and yields

$$\begin{aligned}
\mathcal{E}(\hat{g}) &\leq \left(\frac{1}{\bar{\alpha}} + \frac{1}{\bar{\alpha} - \sum_s p_s u^\delta_{n_s}}\right) \|\eta - \hat{\eta}\|_1 + \sum_{s=1}^K \left[\left(\frac{4p_s}{\bar{\alpha}} + \frac{3}{\alpha_s}\right)\sqrt{\frac{2\log(2/\delta)}{n_s}} + \left(\frac{6}{\alpha_s} + \frac{6p_s}{\bar{\alpha}}\right)\frac{2}{n_s}\right]\\
&\leq \left(\frac{1}{\bar{\alpha}} + \frac{1}{\bar{\alpha} - \sum_s p_s u^\delta_{n_s}}\right) \|\eta - \hat{\eta}\|_1 + 6\sum_{s=1}^K \left(\frac{p_s}{\bar{\alpha}} + \frac{1}{\alpha_s}\right) u^\delta_{n_s} \quad .
\end{aligned}$$

the proof is concluded after the observation that thanks to our assumption we have $\bar{\alpha} - \sum_s p_s u^\delta_{n_s} \geq \bar{\alpha}/2$. □

Boundedness of optimal parameters

Proposition 5.2. *The minimization problem in Eq.* (4) *admits a global minimizer* $(\boldsymbol{\lambda}^*, \boldsymbol{\gamma}^*)$ *which satisfies*

$$\|\boldsymbol{\gamma}^*\|_1 \leq 2 \quad \textit{and} \quad |\lambda^*_s| \leq \frac{p_s}{\bar{\alpha}} \vee \frac{|\boldsymbol{\gamma}^*_s|}{\alpha_s} \quad .$$

Furthermore, if for any s*,* $n_s > \frac{2}{\alpha_s \wedge (1-\alpha_s)}$ *and* $\hat{\eta}(\cdot, s) \in [0,1]$*, the same holds for Eq.* (3)*, that is,*

$$\|\hat{\boldsymbol{\gamma}}\|_1 \leq 2 \quad \textit{and} \quad |\hat{\lambda}_s| \leq \frac{p_s}{\bar{\alpha}} \vee \frac{|\hat{\boldsymbol{\gamma}}_s|}{\alpha_s} \quad .$$

Proof. Overloading the notation, we denote the conditional expectation of Y given $S = s$ by $\eta(s)$. Denote by $H(\boldsymbol{\lambda}, \boldsymbol{\gamma})$ the objective function of the minimization problem in Eq. (4).

Existence of global minizer. Fix arbitrary $(\boldsymbol{\lambda}, \boldsymbol{\gamma}) \in \mathbb{R}^K \times \mathbb{R}^K$ such that $\sum_{s=1}^K \gamma_s = 0$. Since the function $x \mapsto (|x| - b)_+$ is convex for any $b \in \mathbb{R}$ we can lower bound $H(\boldsymbol{\lambda}, \boldsymbol{\gamma})$ using Jensen's inequality as

$$\begin{aligned}
H(\boldsymbol{\lambda}, \boldsymbol{\gamma}) &= \frac{1}{2}\sum_{s=1}^K \frac{1}{\alpha_s}\mathbb{E}_{\boldsymbol{X}|S=s}\left(\left|\frac{\alpha_s p_s}{\bar{\alpha}}(1 - 2\eta(\boldsymbol{X}, s)) + \gamma_s\right| - \frac{p_s\alpha_s}{\bar{\alpha}} - 2\alpha_s\lambda_s - \gamma_s\right)_+ + \sum_{s=1}^K \lambda_s\alpha_s\\
&\geq \frac{1}{2}\sum_{s=1}^K \frac{1}{\alpha_s}\left(\left|\frac{\alpha_s p_s}{\bar{\alpha}}(1 - 2\eta(s)) + \gamma_s\right| - \frac{p_s\alpha_s}{\bar{\alpha}} - 2\alpha_s\lambda_s - \gamma_s\right)_+ + \sum_{s=1}^K \lambda_s\alpha_s \quad .
\end{aligned}$$

Furthermore, since $\alpha_s \leq 1$ for any s and by assumption, $\bar{\gamma} = 0$, we can further lower bound $H(\boldsymbol{\lambda}, \boldsymbol{\gamma})$ as

$$\begin{aligned}
H(\boldsymbol{\lambda}, \boldsymbol{\gamma}) &\geq \frac{1}{2}\sum_{s=1}^K \left(\left|\frac{\alpha_s p_s}{\bar{\alpha}}(1 - 2\eta(s)) + \gamma_s\right| - \frac{p_s\alpha_s}{\bar{\alpha}} - 2\alpha_s\lambda_s - \gamma_s\right)_+ + \sum_{s=1}^K \lambda_s\alpha_s\\
&\geq \frac{1}{2}\left(\|\boldsymbol{\gamma}\|_1 - \sum_{s=1}^K \frac{\alpha_s p_s}{\bar{\alpha}}\,|(1 - 2\eta(s))| - 1 - 2\sum_{s=1}^K \lambda_s\alpha_s\right)_+ + \sum_{s=1}^K \lambda_s\alpha_s\\
&\geq \frac{\|\boldsymbol{\gamma}\|_1}{2} - 1 \ , \qquad (18)
\end{aligned}$$

where we used the triangle inequality for the second inequality and we lower bounded the positive part by the number itself and upper bounded $|1 - 2\eta(s)|$ by one.

Besides, notice that

$$\begin{aligned} H(\boldsymbol{\lambda}, \boldsymbol{\gamma}) &= \frac{1}{2}\sum_{s=1}^{K} \frac{1}{\alpha_s}\mathbb{E}_{\boldsymbol{X}|S=s}\left(\left|\frac{\alpha_s p_s}{\bar{\alpha}}(1 - 2\eta(\boldsymbol{X}, s)) + \gamma_s\right| - \frac{p_s \alpha_s}{\bar{\alpha}} - 2\alpha_s\lambda_s - \gamma_s\right)_+ + \sum_{s=1}^{K}\lambda_s\alpha_s \\ &\geq \sum_{s=1}^{K}\frac{1}{\alpha_s}\mathbb{E}_{\boldsymbol{X}|S=s}\left(-\frac{\alpha_s p_s}{\bar{\alpha}}\eta(\boldsymbol{X}, s) - \alpha_s\lambda_s\right)_+ + \sum_{s=1}^{K}\lambda_s\alpha_s \\ &\geq \sum_{s=1}^{K}\left\{\left(-\frac{p_s}{\bar{\alpha}}\eta(s) - \lambda_s\right)_+ + \lambda_s\alpha_s\right\} \enspace . \end{aligned}$$

The last expression in the above derived inequality can be further lower bounded as

$$\sum_{s=1}^{K}\left\{\left(-\frac{p_s}{\bar{\alpha}}\eta(s) - \lambda_s\right)_+ + \lambda_s\alpha_s\right\} \geq \sum_{s=1}^{K}\{\alpha_s \wedge (1 - \alpha_s)\}|\lambda_s| - \sum_{s=1}^{K}\frac{p_s}{\bar{\alpha}}\eta(s)\{(2\alpha_s) \vee 1\} \enspace . \tag{19}$$

At this moment we have shows that for any $(\boldsymbol{\lambda}, \boldsymbol{\gamma}) \in \mathbb{R}^K \times \mathbb{R}^K$ such that $\sum_{s=1}^{K}\gamma_s = 0$ we have

$$H(\boldsymbol{\lambda}, \boldsymbol{\gamma}) \geq \frac{\|\boldsymbol{\gamma}\|_1}{2} - 1 \quad \text{and} \quad H(\boldsymbol{\lambda}, \boldsymbol{\gamma}) \geq \sum_{s=1}^{K}\{\alpha_s \wedge (1 - \alpha_s)\}|\lambda_s| - \sum_{s=1}^{K}\frac{p_s}{\bar{\alpha}}\eta(s)\{(2\alpha_s) \vee 1\} \enspace .$$

To exploit the above result, we observe that for any $(\boldsymbol{\lambda}, \boldsymbol{\gamma}) \in \mathbb{R}^K \times \mathbb{R}^K$ and for any $c \in \mathbb{R}$ the transformation

$$\gamma_s \mapsto \gamma_s + \frac{p_s\alpha_s}{\bar{\alpha}}c \quad \text{and} \quad \lambda_s \mapsto \lambda_s \quad s \in [K] \enspace ,$$

does not change the value of the objective function. Take any minimizing sequence $(\boldsymbol{\lambda}^k, \boldsymbol{\gamma}^k)$ of H. Due to the above, translation invariance observation, we transform $(\boldsymbol{\lambda}^k, \boldsymbol{\gamma}^k)$ to another minimizing sequence with the property

$$\sum_{s=1}^{K}\gamma_s^k = 0, \qquad \forall k \in \mathbb{N} \enspace . \tag{20}$$

With a slight abuse of notation we denote this (potentially new) minimizing sequence by $(\boldsymbol{\lambda}^k, \boldsymbol{\gamma}^k)$. By definition of $(\boldsymbol{\lambda}^k, \boldsymbol{\gamma}^k)$ as the minimizing sequence, for any $\epsilon > 0$ there exists $N \in \mathbb{N}$ such that

$$H(\boldsymbol{\lambda}^k, \boldsymbol{\gamma}^k) \leq H(\mathbf{0}, \mathbf{0}) + \epsilon, \qquad \forall k \geq N \enspace .$$

Since $H(\mathbf{0}, \mathbf{0}) = \sum_{s=1}^{K}\frac{p_s}{2\bar{\alpha}}\left(|1 - 2\eta(\boldsymbol{X}, s)| - 1\right)_+ = 0$, then it holds for all $k \geq N$ that

$$H(\boldsymbol{\lambda}^k, \boldsymbol{\gamma}^k) \leq \epsilon, \qquad \forall k \geq N \enspace .$$

Furthermore, since for all $k \in \mathbb{N}$ the property in Eq. (20) holds, then using Eqs. (18) and (19) we obtain

$$\|\boldsymbol{\gamma}^k\|_1 \leq 2(1 + \epsilon) \quad \text{and} \quad \sum_{s=1}^{K}\{\alpha_s \wedge (1 - \alpha_s)\}|\lambda_s^k| \leq \epsilon + \sum_{s=1}^{K}\frac{p_s}{\bar{\alpha}}\eta(s)\{(2\alpha_s) \vee 1\} \enspace .$$

Thus for all $k \geq N$ the minimizing sequence $(\boldsymbol{\lambda}^k, \boldsymbol{\gamma}^k)$ is bounded. Extracting convergent sub-sequence and using the fact that H is continuous we conclude that the global minimizer exists. So far we have shown that the minimizer is attained and that at the optimum we have $\|\gamma^*\|_1 \leq 2$ (same holds for $\|\hat{\gamma}\|_1$). Note that the above argument also give a bound on $|\lambda_2^*|$ as well as $|\hat{\lambda}_s|$, yet this bound is not satisfactory and we can derive a better one.

Refined bound on $\boldsymbol{\lambda}$. Recall that the first-order optimality condition on $\boldsymbol{\lambda}^*$ (see (**FOOC**)) is given by

$$\alpha_s = \mathbb{P}_{\boldsymbol{X}|S=s}\left(\left|\frac{p_s}{2\bar{\alpha}}(1 - 2\eta(\boldsymbol{X}, s) - \langle \boldsymbol{\gamma}^*, \mathbf{1}\rangle) + \frac{\langle \boldsymbol{\gamma}^*, \boldsymbol{e}_s\rangle}{2\alpha_s}\right| \geq \frac{p_s}{2\bar{\alpha}}(1 - \langle \boldsymbol{\gamma}^*, \mathbf{1}\rangle) + \langle \boldsymbol{\lambda}^*, \boldsymbol{e}_s\rangle + \frac{\langle \boldsymbol{\gamma}^*, \boldsymbol{e}_s\rangle}{2\alpha_s}\right), \qquad \forall s \in [K] \enspace .$$

Since $\eta(x, s) \in [0, 1]$, then for any $\boldsymbol{x} \in \mathbb{R}^d$ it holds that

$$-\frac{p_s}{\bar{\alpha}} - \frac{(\boldsymbol{\gamma}^*_s)_-}{\alpha_s} \leq \left|\frac{p_s}{2\bar{\alpha}}(1 - 2\eta(\boldsymbol{x}, s)) + \frac{\boldsymbol{\gamma}^*_s}{2\alpha_s}\right| - \frac{p_s}{2\bar{\alpha}} - \frac{\boldsymbol{\gamma}^*_s}{2\alpha_s} \leq -\frac{(\boldsymbol{\gamma}^*_s)_-}{\alpha_s} \enspace .$$

Therefore, if α_s is not in $\{0, 1\}$, we must have that

$$-\frac{p_s}{\bar{\alpha}} \leq \boldsymbol{\lambda}^*_s + \frac{(\boldsymbol{\gamma}^*_s)_-}{\alpha_s} \leq 0 \enspace ,$$

otherwise the considered probability is either equal to 0 or to 1. In particular, it implies that

$$|\boldsymbol{\lambda}^*_s| \leq \frac{p_s}{\bar{\alpha}} \vee \frac{|\boldsymbol{\gamma}^*_s|}{\alpha_s} \enspace .$$

Note that the same can be shown for $\hat{\boldsymbol{\lambda}}$ since Eq. (8) and Lemma 3.2 imply

$$\left|\hat{\mathbb{P}}_{\boldsymbol{X}|S=s}\left(\left|\frac{p_s}{2\bar{\alpha}}(1 - 2\hat{\eta}(\boldsymbol{X}, s) - \hat{\boldsymbol{\gamma}}_s) + \frac{\hat{\boldsymbol{\gamma}}_s}{2\alpha_s}\right| \geq \frac{p_s}{2\bar{\alpha}}(1 - \hat{\boldsymbol{\gamma}}_s) + \hat{\boldsymbol{\lambda}}_s + \frac{\boldsymbol{\gamma}_s}{2\alpha_s}\right) - \alpha_s\right| \leq \frac{2}{n_s}, \forall s \in [K] \enspace ,$$

and the assumption on n_s guarantee that the empirical probability is strictly between 0 and 1. The proof is concluded.

□

6 REDUCTION TO LINEAR PROGRAMMING

In this section we show that the minimization problem in Eq. (3) can be reduced to a problem of linear programming. Recall that our goal is to solve

$$\min_{(\boldsymbol{\lambda}, \boldsymbol{\gamma})} \left\{\langle \boldsymbol{\lambda}, \boldsymbol{\alpha}\rangle + \sum_{s=1}^{K} \hat{\mathbb{E}}_{\boldsymbol{X}|S=s}(\hat{G}(\boldsymbol{X}, s, \boldsymbol{\lambda}, \boldsymbol{\gamma}))_+\right\} \enspace , \tag{21}$$

where

$$\hat{G}(\boldsymbol{x}, s, \boldsymbol{\lambda}, \boldsymbol{\gamma}) = \left|\frac{p_s}{2\bar{\alpha}}(1 - 2\hat{\eta}(\boldsymbol{x}, s) - \langle \boldsymbol{\gamma}, \mathbf{1}\rangle) + \frac{\langle \boldsymbol{\gamma}, \boldsymbol{e}_s\rangle}{2\alpha_s}\right| - \frac{p_s}{2\bar{\alpha}}(1 - \langle \boldsymbol{\gamma}, \mathbf{1}\rangle) - \langle \boldsymbol{\lambda}, \boldsymbol{e}_s\rangle - \frac{\langle \boldsymbol{\gamma}, \boldsymbol{e}_s\rangle}{2\alpha_s} \enspace .$$

Similarly to the support vector machines, the reduction is achieved via the slack variables ζ_i, $i = 1, \ldots, n$. With these slack variables the above problem can be expressed as

$$\begin{aligned}
&\min_{(\boldsymbol{\lambda}, \boldsymbol{\gamma}, \boldsymbol{\zeta})} \langle \boldsymbol{\lambda}, \boldsymbol{\alpha}\rangle + \sum_{s=1}^{K} \sum_{i \in \mathcal{I}_s} \frac{\zeta_i}{n_s} \\
&\text{s.t.} \begin{cases}
\zeta_i \geq 0 & \forall i \in [n] \\
0 \leq \zeta_i + \langle \boldsymbol{\lambda}, \boldsymbol{e}_s\rangle + \frac{p_s}{\bar{\alpha}}\hat{\eta}(\boldsymbol{x}_i, s) & \forall i \in \mathcal{I}_s \forall s \in [K] \\
0 \leq \zeta_i + \left\langle \boldsymbol{\gamma}, \frac{1}{\alpha_s}\boldsymbol{e}_s - \frac{p_s}{\bar{\alpha}}\mathbf{1}\right\rangle + \langle \boldsymbol{\lambda}, \boldsymbol{e}_s\rangle + \frac{p_s}{\bar{\alpha}}\left(1 - \hat{\eta}(\boldsymbol{x}_i, s)\right) & \forall i \in \mathcal{I}_s \forall s \in [K]
\end{cases}
\end{aligned} \tag{\textbf{LP-Primal}}$$

To prove this result it is sufficient to observe that for all $x \in \mathbb{R}$ it holds that $(x)_+ = \min_{\zeta \geq x, \zeta \geq 0} \zeta$.

Introduce the following matrix notation

$$\boldsymbol{c} = \Bigg(\underbrace{{}^1\!/\!{n_1}, \dots, {}^1\!/\!{n_1}}_{\mathcal{I}_1}, \dots \underbrace{{}^1\!/\!{n_s}, \dots, {}^1\!/\!{n_s}}_{\mathcal{I}_s}, \dots, \underbrace{{}^1\!/\!{n_K}, \dots, {}^1\!/\!{n_K}}_{\mathcal{I}_K}, \alpha_1, \dots, \alpha_K, 0 \dots, 0\Bigg)$$

$$\boldsymbol{y} = (\boldsymbol{\zeta}^\top, \boldsymbol{\lambda}^\top, \boldsymbol{\gamma}^\top)$$

$$\boldsymbol{b} = \left(\left(\frac{p_1}{\bar{\alpha}}\hat{\eta}(\boldsymbol{x}_i, s)\right)_{i \in \mathcal{I}_1}, \dots, \left(\frac{p_K}{\bar{\alpha}}\hat{\eta}(\boldsymbol{x}_i, s)\right)_{i \in \mathcal{I}_K}, \left(\frac{p_1}{\bar{\alpha}}(1 - \hat{\eta}(\boldsymbol{x}_i, s))\right)_{i \in \mathcal{I}_1}, \dots, \left(\frac{p_K}{\bar{\alpha}}(1 - \hat{\eta}(\boldsymbol{x}_i, s))\right)_{i \in \mathcal{I}_K}\right)$$

$$\mathbf{A} = \left[\begin{array}{cccc|c|c}
-\mathbf{I}_{n_1 \times n_1} & \mathbf{0}_{n_1 \times n_2} & \dots & \mathbf{0}_{n_1 \times n_K} & -\mathbf{E}^1_{n_1 \times K} & \mathbf{0}_{n_2 \times K} \\
\mathbf{0}_{n_2 \times n_1} & -\mathbf{I}_{n_2 \times n_2} & \dots & \mathbf{0}_{n_2 \times n_K} & -\mathbf{E}^2_{n_2 \times K} & \mathbf{0}_{n_1 \times K} \\
\vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\
\mathbf{0}_{n_K \times n_1} & \mathbf{0}_{n_K \times n_2} & \dots & -\mathbf{I}_{n_K \times n_K} & -\mathbf{E}^K_{n_K \times K} & \mathbf{0}_{n_K \times K} \\
\hline
-\mathbf{I}_{n_1 \times n_1} & \mathbf{0}_{n_1 \times n_2} & \dots & \mathbf{0}_{n_1 \times n_K} & -\mathbf{E}^1_{n_1 \times K} & \frac{p_1}{\bar{\alpha}}\mathbf{1}_{n_1 \times K} - \frac{1}{\alpha_1}\mathbf{E}^1_{n_1 \times K} \\
\mathbf{0}_{n_2 \times n_1} & -\mathbf{I}_{n_2 \times n_2} & \dots & \mathbf{0}_{n_2 \times n_K} & -\mathbf{E}^2_{n_2 \times K} & \frac{p_2}{\bar{\alpha}}\mathbf{1}_{n_2 \times K} - \frac{1}{\alpha_2}\mathbf{E}^2_{n_2 \times K} \\
\vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\
\mathbf{0}_{n_K \times n_1} & \mathbf{0}_{n_K \times n_2} & \dots & -\mathbf{I}_{n_K \times n_K} & -\mathbf{E}^K_{n_K \times K} & \frac{p_K}{\bar{\alpha}}\mathbf{1}_{n_K \times K} - \frac{1}{\alpha_K}\mathbf{E}^K_{n_K \times K}
\end{array}\right]$$

where $\mathbf{E}^s_{n \times m}$ is a $n \times m$ matrix composed of zeros and ones, whose s^{th} column is equal to $\mathbf{1}$ and all other elements are zero, $\mathbf{1}_{n \times m}$ is a matrix of ones of size $n \times m$. Using the above notation, the problem in (**LP-Primal**) can be written as

$$\begin{aligned}
&\min_{\boldsymbol{y} \in \mathbb{R}^{n+2K}} \langle \boldsymbol{c}, \boldsymbol{y} \rangle \\
&\text{s.t.} \begin{cases} \mathbf{A}\boldsymbol{y} \leq \boldsymbol{b} \\ y_i \geq 0 \qquad i \in [n] \end{cases}
\end{aligned} \qquad \textbf{(LP-Primal-compacted)}$$

While the dimension of matrix $\mathbf{A}$ is $2n \times (n + 2K)$, this matrix has at most $4n + nK$ non-zero elements. This fact can be exploited if $n \gg K$, that it, the amount of *unlabeled* data is large compared to the amount of groups.

References

Pascal Massart. The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality. *The annals of Probability*, pages 1269–1283, 1990.