A Auditing Multiple Groups

Here we consider the case when there are more than two groups. Suppose we have $J+1$ groups $\{0, 1, \ldots, J\}$. In accordance with Definition 1, the null and alternative become

$$H_0 : \mathbb{E}_\rho[\varphi(X)|\xi_i] = \mathbb{E}_\rho[\varphi(X)|\xi_j], \quad \forall i, j \in \{0, \ldots, J\}, \tag{13}$$
$$H_1 : \exists i, j \in \{0, \ldots, J\} \text{ such that } \mathbb{E}_\rho[\varphi(X)|\xi_i] \neq \mathbb{E}_\rho[\varphi(X)|\xi_j]. \tag{14}$$

As before, let $\mu_i = \mathbb{E}_\rho[\varphi(X)|\xi_i]$, $i \in \{0, \ldots, J\}$. One could derive a sequential test by applying Algorithm 1 to each pair of means μ_i and μ_j. Game-theoretically, this can be interpreted as splitting your initial wealth among multiple games and playing each simultaneously. If you grow rich enough in any one game, you reject the null. Of course, one needs to adjust the significance level to account for the number of games being played, thus reducing the (nonasymptotic) power of the test.

Of course, it is not necessary to test each mean against all others. We need only test whether $\mu_b = \mu_{b+1}$ for all $b \in \{0, \ldots, J\}$. That is, we can play J games instead of $\Omega(J^2)$ games. In order to ensure this constitutes a level-α test, we reject when the wealth process of any game is at least $(\alpha J)^{-1}$. The union bound then ensures that the type-I error of this procedure is bounded by α. Moreover, the asymptotic power remains one since, if $\mu_i \neq \mu_j$ for some i, j then $\mu_b \neq \mu_{b+1}$ for some b. The guarantees we've provided on Algorithm 1 ensure that the wealth process for this particular game will eventually grow larger than $(\alpha J)^{-1}$, thus our test will reject. We summarize this discussion with the following proposition, which is the equivalent of Proposition 1 for auditing multiple groups.

Proposition 4. *Let $\alpha \in (0, 1)$. Consider running Algorithm 1 on groups $b, b+1$, for $b \in \{0, 1, \ldots, J\}$ in parallel with input parameter α/K. This constitutes a level-α sequential test for (13) with aymptotic power one against (14). If we receive an audit from each group at each timestep, then the expected stopping time τ of this procedure obeys*

$$\mathbb{E}[\tau] \lesssim \min_{b \in \{0, \ldots, J-1\}} \frac{1}{|\mu_b - \mu_{b+1}|^2} \log\left(\frac{J}{|\mu_b - \mu_{b+1}|^2 \alpha}\right). \tag{15}$$

The expected stopping time follows from Proposition 1 after correcting for the significance level and the difference between the means. We take the minimum over all b because the procedure rejects as soon as any of the wealth processes grow too large. Equivalent versions of Propositions 2 and 3 for multiple groups can be obtained similarly.

B Omitted Proofs

B.1 Proof of Proposition 1

We break the proof into three components.

Level-α sequential test. Combining the discussion at the beginning of Section 3 with Ville's inequality demonstrates why our procedure constitutes a level-α sequential test. However, let us prove it formally here for completeness. Let $P \in H_0$ and note that $\mathbb{E}_P[\widehat{Y}_t^0 - \widehat{Y}_t^1] = \mathbb{E}_P[\varphi(X_t^0) - \varphi(X_t^1)] = \mu_0 - \mu_1 = 0$. Therefore, using the fact that λ_t is predictable (i.e., $\mathcal{F}_{t-1}$-measurable)

$$\mathbb{E}_P[\mathcal{K}_t|\mathcal{F}_{t-1}] = \mathbb{E}_P\left[\prod_{j=1}^{t}(1 + \lambda_j(\widehat{Y}_j^0 - \widehat{Y}_j^1))\middle|\mathcal{F}_{t-1}\right] = \mathcal{K}_{t-1}(1 + \lambda_t \mathbb{E}_P[\widehat{Y}_t^0 - \widehat{Y}_t^1]) = \mathcal{K}_{t-1},$$

so $(\mathcal{K}_t)_{t \geq 1}$ is a P-martingale, with initial value 1. Moreover, it is nonegative since $|\lambda_t| \leq 1/2$ for all t by definition of ONS. Thus, Ville's inequality implies $P(\exists t \geq 1 : \mathcal{K}_t \geq 1/\alpha) \leq \alpha$, meaning that rejecting at $1/\alpha$ yields a level-α sequential test. Finally, as discussed in the main paper, the last lines of Algorithm 1 are justified by the randomized Ville's inequality of Ramdas and Manole [43], which states that, for all stopping times n,

$$P(\exists t \leq n : \mathcal{K}_t \geq 1/\alpha \text{ or } \mathcal{K}_n > U/\alpha) \leq \alpha,$$

where $U \sim \text{Unif}(0, 1)$ is independent of everything else.

Asymptotic power. Next, let us demonstrate that Algorithm 1 has asymptotic power one. That is, for $P \in H_1$, $P(\tau < \infty) = 1$. It suffices to show that $P(\tau = \infty) = 0$. To see this, define

$$g_t := \widehat{Y}_t^0 - \widehat{Y}_t^1, \quad S_t := \sum_{i=1}^t g_i, \quad V_t := \sum_{i=1}^t g_i^2.$$

We have the following guarantee on the wealth process, which can be translated from results concerning ONS from Cutkosky and Orabona [44, Theorem 1]:

$$\mathcal{K}_t \geq \frac{1}{V_t} \exp\left\{\frac{S_t^2}{4(V_t + |S_t|)}\right\} \geq \frac{1}{t} \exp\left\{\frac{S_t^2}{8t}\right\}, \quad \forall t \geq 1. \tag{16}$$

Since $\{\tau = \infty\} \subset \{\tau \geq t\}$ for all $t \geq 1$, we have $P(\tau = \infty) \leq \liminf_{t\to\infty} P(\tau > t) \leq \liminf_{t\to\infty} P(\mathcal{K}_t < 1/\alpha)$, where the final inequality is by definition of the algorithm. Using the second inequality of (16),

$$P(\mathcal{K}_t < 1/\alpha) \leq P\bigg(\exp\left\{\frac{S_t^2}{8t}\right\} < t/\alpha\bigg) \leq P\bigg(-\sqrt{\frac{8\log(t/\alpha)}{t}} < \frac{S_t}{t} < \sqrt{\frac{8\log(t/\alpha)}{t}}\bigg).$$

By the SLLN, S_t/t converges to $\mu_0 - \mu_1 \neq 0$ almost surely. On the other hand, $8\log(t/\alpha)/t \to 0$. Thus, if we let A_t be the event that $\exp(S_t^2/8t) < t/\alpha$, we see that $\mathbf{1}(A_t) \to 0$ almost surely. Hence, by the dominated convergence theorem,

$$P(\tau = \infty) \leq \liminf_{t\to\infty} P(A_t) = \liminf_{t\to\infty} \int \mathbf{1}(A_t) dP = \int \liminf_{t\to\infty} \mathbf{1}(A_t) dP = 0.$$

This completes the argument.

Expected stopping time. Last, let us show the desired bound on the expected stopping time. Fix $P \in H_1$. Let τ be the stopping time of the test. Since it is nonnegative, we have

$$\mathbb{E}[\tau] = \sum_{t=1}^\infty P(\tau > t) = \sum_{t=1}^\infty P(\log \mathcal{K}_t < \log(1/\alpha)) = \sum_{t=1}^\infty P(E_t),$$

for $E_t = \{\log \mathcal{K}_t < \log(1/\alpha)\}$. Note that the second equality is by definition of the algorithm. Employing the first inequality of (16) yields

$$\begin{aligned} E_t &\subset \{S_t^2 < 4(V_t + |S_t|)(\log(1/\alpha) - \log(1/V_t)\} \\ &\subset \bigg\{S_t^2 < 4\bigg(V_t + \sum_{i\leq t} |g_i|\bigg)(\log(1/\alpha) - \log(1/V_t)\bigg\}. \end{aligned}$$

To analyze the probability of this event, we first develop upper bounds on $W_t := \sum_{i\leq t} |g_i|$ and V_t. We begin with W_t. Since W_t is the sum of independent random variables in $[0, 1]$, we apply the multiplicative Chernoff bound (e.g., [57]) to obtain

$$P(W_t > (1+\delta)\mathbb{E}[W_t]) \leq \exp(-\delta^2 \mathbb{E}[W_t]/3).$$

Setting the right hand side equal to $1/t^2$ and solving for δ gives $\delta = \sqrt{6\log t/\mathbb{E}W_t}$. Thus, with probability $1 - 1/t^2$, we have

$$W_t \leq \mathbb{E}W_t + \sqrt{6\mathbb{E}[W_t]\log t} = t + \sqrt{6t\log t} \leq 2t \quad \forall t \geq 17, \tag{17}$$

where we've used that $\mathbb{E}[W_t] = \sum_{i\leq t} \mathbb{E}[|g_i|] \leq t$ since $|g_i| \leq 1$. Following a nearly identical process for V_t, we have that with probability $1 - 1/t^2$,

$$V_t \leq \mathbb{E}[V_t] + \sqrt{6\mathbb{E}[V_t]\log t} \leq t + \sqrt{6t\log t} \leq 2t, \quad \forall t \geq 17, \tag{18}$$

where again we use that $|g_i|^2 \leq |g_i| \leq 1$. Let $H_t = \{V_t \leq 2t\} \cap \{W_t \leq 2t\}$. Then,

$$E_t \cap H_t \subset \{S_t^2 < 16t(\log(1/\alpha) + \log(2t)\} \subset \{|S_t| < \underbrace{4\sqrt{t\log(2t/\alpha)}}_{:=D}\}.$$

We now argue that $|S_t|$ is unlikely to be so small. Indeed, since S_t is the sum of independent random variables in $[-1, 1]$, applying a Chernoff bound for the third time gives $P(|S_t - \mathbb{E}S_t| \geq u) \leq 2\exp(-u^2/t)$. So, with probability $1 - 1/t^2$, by the reverse triangle inequality,

$$||S_t| - |\mathbb{E}S_t|| \leq |S_t - \mathbb{E}S_t| \leq \sqrt{t \log 2t^2},$$

implying that,

$$|S_t| \geq |\mathbb{E}S_t| - \sqrt{t \log 2t^2} \geq t\Delta - \sqrt{2t \log 2t}.$$

This final quantity is at least D for all $t \geq \frac{81}{\Delta^2} \log(\frac{162}{\Delta^2 \alpha})$. Now, combining what we've done thus far, by the law of total probability,

$$P(E_t) = P(E_t \cap H_t) + P(E_t|H_t^c)P(H_t^c) \leq P(|S_t| < D) + P(H_t^c) \leq 3/t^2,$$

and so, for t large enough such that (17), (18), and $S_t > D$ all hold, that is

$$T = \frac{81}{\Delta^2} \log\left(\frac{162}{\Delta^2 \alpha}\right),$$

we have

$$\mathbb{E}[\tau] = \sum_{t \geq 1} P(E_t) \leq T + \sum_{t \geq T} \frac{3}{t^2} \leq T + \frac{\pi^2}{2}.$$

This completes the proof.

B.2 Proof of Proposition 2

The proof is similar to that of Proposition 1, so we highlight only the differences.

The wealth process remains a martingale due to the IPW weights (7). Indeed, since λ_t and L_t are $\mathcal{F}_{t-1}$ measurable, under the null we have

$$\begin{aligned}\mathbb{E}[\mathcal{K}_t|\mathcal{F}_{t-1}] &= \mathbb{E}\left[\prod_{j=1}^{t}(1 + \lambda_j L_j(\widehat{Y}_j^0 \omega_j^0 - \widehat{Y}_j^1 \omega_j^1))\middle|\mathcal{F}_{t-1}\right] \\ &= \mathcal{K}_{t-1}(1 + \lambda_t L_t \mathbb{E}[\widehat{Y}_t^0 \omega_t^0 - \widehat{Y}_t^1 \omega_t^1]) = \mathcal{K}_{t-1}(1 + \lambda_t L_t(\mu_0 - \mu_1)) = \mathcal{K}_{t-1}.\end{aligned}$$

Moreover, as described in the text, multiplication by L_t ensures that $\mathcal{K}_t$ is nonnegative, since

$$\begin{aligned}|L_t||\widehat{Y}_t^0 \omega_t^0(X_t^0) - \widehat{Y}_t^1 \omega_t^1(X_t^1)| &\leq L_t|\widehat{Y}_t^0 \omega_t^0(X_t^0)| + L_t|\widehat{Y}_t^1 \omega_t^1(X_t^1)| \\ &\leq L_t \omega_t^0(X_t^0) + L_t \omega_t^1(X_t^1) \leq 1,\end{aligned}$$

since

$$L_t \leq \frac{1}{2\omega_t^b(X_t^b)},$$

for each b by definition. Therefore, $(\mathcal{K}_t)_{t \geq 1}$ is a nonnegative martingale and, as before, Ville's inequality implies that rejecting at $1/\alpha$ gives a level-α sequential test.

The asymptotic power follows by replacing g_t in Appendix B.1 with

$$h_t = L_t(\widehat{Y}_t^0 \omega_t^0 - \widehat{Y}_t^1 \omega_t^1).$$

Under the alternative, h_t has non-zero expected value, so identical arguments apply.

Regarding, the expected stopping time, we again argue about h_t instead of g_t. Since $|h_t| \leq 1$ (see above), the bounds on V_t and W_t remain as they are in the proof of Proposition 1. The bound on $|\mathbb{E}[S_t]|$ is where the proof departs that in Appendix B.1. In this case we have

$$\mathbb{E}[S_t|\mathcal{F}_{t-1}] = S_{t-1} + L_t \mathbb{E}[\widehat{Y}_t^0 \omega_t^0 - \widehat{Y}_t^1 \omega_t^1|\mathcal{F}_{t-1}] = S_{t-1} + L_t(\mu_0 - \mu_1).$$

Therefore,

$$\mathbb{E}[S_t] = \mathbb{E}[\mathbb{E}[S_t|\mathcal{F}_{t-1}]] = \mathbb{E}[S_{t-1} + L_t(\mu_0 - \mu_1)] = \mathbb{E}[S_{t-1}] + (\mu_0 - \mu_1)\mathbb{E}[L_t].$$

Induction thus yields

$$|\mathbb{E}[S_t]| = \left|(\mu_0 - \mu_1) \sum_{i \leq t} \mathbb{E}[L_i]\right| = \Delta \left|\sum_{i \leq t} \mathbb{E}[L_i]\right| \geq \Delta t L_{\inf}.$$

From here, we may replace Δ in the proof in Appendix B.1 with $\Delta L_{\inf}$ and the arithmetic remains the same. This yields the desired result.

B.3 Proof of Proposition 3

Again, the proof mirrors that of Proposition 1 so we highlight only the differences.

First let us ensure that Algorithm 1 yields a level-α sequential test. As before, it suffices to demonstrate that the wealth process is a nonnegative martingale. The time-varying means do not change this fact from before:

$$\mathbb{E}[\mathcal{K}_t|\mathcal{F}_{t-1}] = \mathbb{E}\left[\prod_{j=1}^{t}(1+\lambda_j(\widehat{Y}_j^0 - \widehat{Y}_j^1))\middle|\mathcal{F}_{t-1}\right] = \mathcal{K}_{t-1}(1+\lambda_t\mathbb{E}[\widehat{Y}_t^0 - \widehat{Y}_t^1|\mathcal{F}_{t-1}]) = \mathcal{K}_{t-1},$$

since, under the null, $\mathbb{E}[Y_t^0|\mathcal{F}_{t-1}] = \mathbb{E}[\varphi(X)|\xi_0, \mathcal{F}_{t-1}] = \mu_0 = \mu_1 = \mathbb{E}[\varphi(X)|\xi_1, \mathcal{F}_{t-1}] = \mathbb{E}[Y_t^1|\mathcal{F}_{t-1}]$. Nonnegativity once again follows from the ONS strategy.

Asymptotic power follows an identical argument as in Appendix B.1, so we focus on expected stopping time. The event E_t remains the same as in Appendix B.1. We again apply a Chernoff bound to W_t (the values remain independent, even though they are not necessarily identically distributed), and obtain

$$W_t \leq \mathbb{E}W_t + \sqrt{6\mathbb{E}[W_t]\log t} = 2t,$$

for $t \geq 17$ with probability $1 - 1/t^2$, since again, $|g_i| \leq 1$ for each i. Similarly, $\mathbb{E}V_t \leq 2t$ with probability $1 - 1/t^2$ for $t \geq 17$. Let the shift begin at time n, and place $\Delta = \inf_{t \geq n} |\mu_0(t) - \mu_1(t)|$. Then $|\mathbb{E}S_t| \geq (t-n)\Delta$. As above, we want to find t such that

$$|S_t| \geq |\mathbb{E}S_t| - \sqrt{t\log 2t^2} \geq (t-n)\Delta - \sqrt{t\log 2t^2} \geq D.$$

Rearranging and simplifying this final inequality, we see that it suffices for t to satisfy

$$t - n \geq \frac{6}{\Delta}\sqrt{t\log(2t/\alpha)}. \tag{19}$$

We claim this holds for all

$$t \geq n + \max\left\{n, \frac{108}{\Delta^2}\log\left(\frac{108\cdot 4}{\Delta^2\alpha}\right)\right\}.$$

To see this, suppose first that $n \geq \beta$ where

$$\beta = \frac{108}{\Delta^2}\log\left(\frac{108\cdot 4}{\Delta^2\alpha}\right).$$

Then, at $t = 2n$, the right hand side of (19) is

$$\frac{6}{\Delta}\sqrt{2n\log(2n/\alpha)} \leq n,$$

where the final inequality holds for all $n \geq \beta$, which was assumed. Now suppose that $n < \beta$, so that (19) should hold for $t \geq n + \beta$. Since the left hand side of (19) grows faster than the right hand side, it suffices to show that it holds at $t = n + \beta$. To this end, write

$$\begin{aligned}
\frac{6}{\Delta}\sqrt{t\log(2t/\alpha)}\bigg|_{t=n+\beta} &\leq \frac{6}{\Delta}\sqrt{(n+\beta)\log(2n/\alpha + 2\beta/\alpha)} \\
&\leq \frac{6}{\Delta}\sqrt{2\beta\log(4\beta/\alpha)} \\
&= \frac{72}{\Delta^2}\sqrt{\log\left(\frac{108\cdot 4}{\Delta^2\alpha}\right)\log\left(\frac{108\cdot 4}{\Delta^2\alpha}\log\left(\frac{108\cdot 4}{\Delta^2\alpha}\right)\right)} \\
&= \frac{72}{\Delta^2}\sqrt{\log\left(\frac{108\cdot 4}{\Delta^2\alpha}\right)\log\left(\frac{108\cdot 4}{\Delta^2\alpha}\right) + \log\log\left(\frac{108\cdot 4}{\Delta^2\alpha}\right)} \\
&\leq \frac{108}{\Delta^2}\log\left(\frac{108\cdot 4}{\Delta^2\alpha}\right) = \beta,
\end{aligned}$$

where the final inequality uses the (loose) bound $\log\log(x) \leq \log^2(x)$.

C Simulation Details

Code to recreate all plots and run the simulations is attached. Here we provide more extensive details on each figure.

Figure 1. Given Δ, we generate the two means μ_0 and μ_1 as $\mu_0 = 0.5 + \Delta/2$ and $\mu_1 = 0.5 - \Delta/2$. We take $\varphi(X)|\xi_b$ to be Ber(μ_b). (Thus, this simulates a scenario for which we witness the classifcation decisions, not e.g., a risk score.) We set $\alpha = 0.01$, so we reject when the wealth process is at least 100. We receive a pair of observations each timestep. Each experiment was run 100 times to generate the plotted standard deviation around the mean of each wealth process.

Figure 2. As above, we take the distribution of model observations $\varphi_t(X)|\xi_b$ to be Ber$(\mu_b(t))$. For the left hand side of Figure 2 we take $\mu_0(t) = \mu_1(t) = 0.3$ for $t = 1, \ldots, 99$. At $t = 100$, we add a logistic curve to μ_1. In particular, we let

$$\mu_1(t) = 0.3 + \frac{0.5}{1 + \exp((250 - t)/25)}, \quad t \geq 100.$$

We keep μ_0 at 0.3. For the right hand side of Figure 2, we let both μ_1 and μ_0 be noisy sine functions with different wavelengths. We take

$$\mu_0(t) = \frac{\sin(t/40)}{10} + 0.4 + \epsilon_t^0,$$

for all t, where $\epsilon_t^0 \sim N(0, 0.01)$. Meanwhile,

$$\mu_1(t) = \frac{\sin(t/20)}{10} + 0.4 + \frac{t}{1000} + \epsilon_t^1,$$

where, again, $\epsilon_t^1 \sim N(0, 0.01)$. The mean $\mu_1(t)$ thus has a constant upward drift over time. As before, we assume we receive a pair of observations at each timestep and we take $\alpha = 0.01$. We generate the means one, but run the sequential test 100 times in order to plot the standard deviation around the mean.

Figures 3 and 4. For a given sequential test and a given value of α, we run (i) the test under the null hypothesis, and (ii) the test under the alternative. Repeating 300 times and taking the average gives the FPR and average rejection time for this value of α. This procedure is how the leftmost two columns of Figure 3 were constructed. The final column then simply plots the FPR versus the value of α.

We used a random forest for both the credit default dataset and the US census data. For the credit default dataset, the model does not satisfy equality of opportunity [39] when Y indicates whether an individual has defaulted on their loan, and A indicates whether or not they have any university-level education. One can imagine loans being given or withheld on the basis of whether they are predicted to be returned; we might wish that this prediction does not hinge on educational attainment. For the census data, the model does not satisfy equality of opportunity when A indicates whether an individual has an optical issues, and Y indicates whether they are covered by public insurance. Admittedly, this example is somewhat less normative than the other. It is unclear whether we should expect perfect equality of opportunity in this context. However, we emphasize that our experiments are for illustrative purposes only. They are not meant as comments on the actual fairness or unfairness of these datasets. We interface with the census by means the folktables package [51].

For the credit default dataset and random forest classifier, we have $\Delta = |\mu_0 - \mu_1| = 0.034$, and $\Delta = 0.09$ for the census data. To construct the fair model (in order to test the null), we add Δ to the model predictions of the group with the lower mean. Thus, the distributions of predictions are different but the means are identical.

Figure 4 follows similar experimental logic, but we begin with a fair model (i.e., group predictions with the same mean), and then switch to the unfair random forest classifier at time $t = 400$.