*Table 3.* Association tests code names

| | |
|---|---|
| FLvINS | Flowers vs. insects (Greenwald et al., 1998) |
| INSTvWP | Instruments vs. weapons (Greenwald et al., 1998) |
| MNTvPHS | Mental vs. physical disease (Monteith & Pettit, 2011) |
| EAvAA | Europ-Amer vs Afr-Amer names (Caliskan et al., 2017) |
| EAvAA(Bertrand & Mullainathan, 2004) | Europ-Amer vs Afr-Amer names (Bertrand & Mullainathan, 2004) |
| MNvFN | Male vs. female names (Nosek et al., 2002a) |
| MTHvART | Math vs. arts (Nosek et al., 2002a) |
| SCvART(Nosek et al., 2002b) | Science vs. arts (Nosek et al., 2002b) |
| YNGvOLD | Young vs. old people's names (Nosek et al., 2002a) |
| PLvUPL | Pleasant vs. unpleasant (Greenwald et al., 1998) |
| TMPvPRM | Temporary vs. permanent (Monteith & Pettit, 2011) |
| PLvUPL(Nosek et al., 2002a) | Pleasant vs. unpleasant (Nosek et al., 2002a) |
| CARvFAM | Career vs. family (Nosek et al., 2002a) |
| MTvFT | Male vs. female terms (Nosek et al., 2002a) |
| MTvFT(Nosek et al., 2002b) | Male vs. female terms (Nosek et al., 2002b) |

## A. Relation between groupwise and pairwise comparison

In case of pairwise comparison, we have $|I_1| = \cdots = |I_G| = 2$. As mentioned in the Algorithm 1, we at first mean-center each group, which is assumed to nullify the variability along the directions of the relevant attributes. Lets consider $I_1 = \{\varphi_{1_1}, \varphi_{1_2}\}$. Then:

$$\begin{aligned} H\Phi_{I_1} &= \left(\varphi_{1_1} - \frac{\varphi_{1_1} + \varphi_{1_2}}{2}, \varphi_{1_2} - \frac{\varphi_{1_1} + \varphi_{1_2}}{2}\right)^\top \\ &= \left(\frac{\varphi_{1_1} - \varphi_{1_2}}{2}, \frac{\varphi_{1_2} - \varphi_{1_1}}{2}\right)^\top \end{aligned}$$

Hence the combined matrix can be written as:

$$M_{pairs} = \frac{1}{4|G|} \sum_{i=1}^{|G|} \left(\varphi_{i_1} - \varphi_{i_2}\right) \left(\varphi_{i_1} - \varphi_{i_2}\right)^\top$$

which is equivalent to consider the difference between the pairs of each individual groups (upto a constant). On the other hand, we have more than two observations in each group, the grand matrix following Algorithm 1 becomes:

$$M_{general} = \frac{1}{N} \sum_{i=1}^{G} \sum_{j=1}^{|I_G|} \left(\varphi_{i_j} - \bar{\varphi}_i\right) \left(\varphi_{i_j} - \bar{\varphi}_i\right)^\top$$

where $N = \sum_{i=1}^{G} |I_G|$, total number of observations. Hence, in case of $|I_G| = 2$, we essentially don't need to mean center as we are taking the difference between the observations of each pair. When $G$ is essentially fixed, i.e. $|I_G| \approx N$, the error in estimating $\mathsf{ran}A_*$ due to mean centering contributes a higher order term (See Theorem 3.2 for more details) which is essentially negligible. In case of pairwise comparison, although there is no error due to mean centering, we pay a constant as we are effectively loosing one observation in the each pair.

## B. Theoretical properties of EXPLORE

In this section, we investigate the theoretical properties of EXPLORE. We provide statistical guarantees corresponding to the estimation using the scaled logistic link (Section 2.2). To keep things simple, we tweak (2.5) so that it is strongly identifiable:

$$y_i \mid z_{i_1}, z_{i_2} \sim \mathsf{Ber}((2-\epsilon)\sigma(-\langle D_i, \Sigma_0 \rangle))$$

for some small $\epsilon > 0$. The log-likelihood of samples $(x_1, y_1), \ldots, (x_n, y_n)$ is

$$\ell_n(\Sigma) = \tfrac{1}{n} \textstyle\sum_{i=1}^{n} \left[ y_i \log F_*(x_i' \Sigma x_i) \right. \\ \left. + (1 - y_i) \log \left(1 - F_*(x_i' \Sigma x_i)\right) \right],$$

where $F_* = (2 - \epsilon)\sigma$.

**Proposition B.1.** *The population version of the likelihood function $\ell(\Sigma)$ is concave in $\Sigma$ and uniquely maximized at $\Sigma_0$.*

*Proof.* The population version of the likelihood function is:

$$\ell(\Sigma) = E\left[ Y \log F_*(X'\Sigma X) + (1 - Y) \log \left(1 - F_*(X'\Sigma X)\right) \right] = g(X'\Sigma X) \tag{B.1}$$

As the function $\Sigma \longrightarrow X'\Sigma X$ is affine in $\Sigma$, we only need to show that $g$ is concave. From equation B.1, the function $g(.)$ can be define as: $g(t) = y \log F_*(t) + (1 - y) \log \left(1 - F_*(t)\right)$ on $t \in \mathbb{R}^+$ for any fixed $y \in \{0, 1\}$. The function $F_*$ is double differentiable with the derivatives as below:

$$\begin{aligned} F_*(x) &= \frac{2 - \epsilon}{1 + e^t}, \;\; 1 - F_*(t) = \frac{e^t - 1 + \epsilon}{1 + e^t} \\ F_*'(t) &= -(2 - \epsilon) \frac{e^t}{(1 + e^t)^2} \\ F_*''(t) &= -(2 - \epsilon) \frac{e^t (1 - e^t)}{(1 + e^t)^3} \end{aligned}$$

We show below that $g''(t) \leq 0$ for all $t$ which proves the concavity of $\ell(\Sigma)$:

$$\begin{aligned} & g(t) = y \log F_*(t) + (1 - y) \log \left(1 - F_*(t)\right) \\ \Rightarrow & g'(t) = y \frac{F_*'(t)}{F_*(t)} - (1 - y) \frac{F_*'(t)}{1 - F_*(t)} \\ \Rightarrow & g''(t) = y \frac{F_*(t) F_*''(t) - (F_*'(t))^2}{F_*^2(t)} - (1 - y) \frac{(1 - F_*(t)) F_*''(t) + (F_*'(t))^2}{(1 - F_*(t))^2} \end{aligned} \tag{B.2}$$

For the first summand in the double derivative we have:

$$\begin{aligned} \frac{F_*(t) F_*''(t) - (F_*'(t))^2}{F_*^2(t)} &= \frac{-(2 - \epsilon)^2 \frac{e^t (1 - e^t)}{(1 + e^t)^4} - (2 - \epsilon)^2 \frac{e^{2t}}{(1 + e^t)^4}}{\frac{(2 - \epsilon)^2}{(1 + e^t)^4}} \\ &= -\frac{e^t (1 - e^t) + e^{2t}}{(1 + e^t)^2} = -\frac{e^t}{(1 + e^t)^2} < 0 \; \forall \, t \in \mathbb{R}^+ \end{aligned} \tag{B.3}$$

For the second summand:

$$\begin{aligned} \frac{(1 - F_*(t)) F_*''(t) + (F_*'(t))^2}{(1 - F_*(t))^2} &= \frac{-(2 - \epsilon) \frac{(e^t - 1 + \epsilon) e^t (1 - e^t)}{(1 + e^t)^4} + (2 - \epsilon)^2 \frac{e^{2t}}{(1 + e^t)^4}}{\frac{(e^t - 1 + \epsilon)^2}{(1 + e^t)^2}} \\ &= \frac{(2 - \epsilon) \left[ (2 - \epsilon) e^{2t} - (e^t - 1 + \epsilon) e^t (1 - e^t) \right]}{(e^t - 1 + \epsilon)^2 (1 + e^t)^2} \\ &= \frac{(2 - \epsilon) \left[ (2 - \epsilon) e^{2t} + (e^t - 1 + \epsilon) e^t (e^t - 1) \right]}{(e^t - 1 + \epsilon)^2 (1 + e^t)^2} \geq 0 \; \forall \, t \in \mathbb{R}^+ \end{aligned} \tag{B.4}$$

Combining equations B.2, B.3 and B.4 we get:

$$g''(t) = -y \frac{e^t}{(1 + e^t)^2} - (1 - y) \frac{(2 - \epsilon) \left[ (2 - \epsilon) e^{2t} + (e^t - 1 + \epsilon) e^t (e^t - 1) \right]}{(e^t - 1 + \epsilon)^2 (1 + e^t)^2} < 0 \; \forall \, t \in \mathbb{R}^+$$

This proves the strict concavity. To prove that $\Sigma_0$ is the unique maximizer, observe that:

$$\begin{aligned}\ell(\Sigma) &= E\left[Y \log F_*(X'\Sigma X) + (1-Y)\log\left(1 - F_*(X'\Sigma X)\right)\right] \\ &= E\left[F_*(X'\Sigma_0 X)\log F_*(X'\Sigma X) + (1 - F_*(X'\Sigma_0 X))\log\left(1 - F_*(X'\Sigma X)\right)\right] \\ &= \ell(\Sigma_0) - E\left(KL(Bern(F_*(X'\Sigma_0 X)) \,||\, Bern(F_*(X'\Sigma X)))\right)\end{aligned}$$

Hence $\ell(\Sigma_0) \geq \ell(\Sigma)$ for all $\Sigma \in \Theta$ as KL divergence is always non-negative. Next, let $\Sigma_1$ be any other maximizer. Then,

$$\begin{aligned}& E\left(KL(Bern(F_*(X'\Sigma_0 X)) \,||\, Bern(F_*(X'\Sigma X)))\right) = 0 \\ \Rightarrow & KL(Bern(F_*(X'\Sigma_0 X)) \,||\, Bern(F_*(X'\Sigma X))) = 0 \text{ a.s. in } X \\ \Rightarrow & F_*(X'\Sigma_0 X) = F_*(X'\Sigma X) \text{a.s. in } X \\ \Rightarrow & X'(\Sigma - \Sigma_0)X = 0 \text{ a.s. in } X \\ \Rightarrow & \Sigma = \Sigma_0\end{aligned}$$

as the interior of the support of $X$ is non null. This proves the uniqueness of the maximizer. □

The maximum likelihood estimator (MLE) $\widehat{\Sigma}$ is

$$\widehat{\Sigma} = \arg\max_{\Sigma} \ell_n(\Sigma)$$

The asymptotic properties of $\widehat{\Sigma}$ (consistency and asymptotic normality) are well-established in the statistical literature (e.g. see van der Vaart (1998)). Here we study the non-asymptotic convergence rate of the MLE. We start by stating our assumptions.

**Assumption B.2.** *The feature space $\mathcal{X}$ is a bounded subset of $\mathbb{R}^d$, i.e. there exits $R < \infty$ such that $\|X\| = \|\varphi_1 - \varphi_2\| \leq U$ for all $X \in \mathcal{X}$.*

**Assumption B.3.** *The parameter space $\Theta$ is a subset of $\mathbb{S}^d_{++}$ and $\sup\{\lambda_{max}(\Sigma) : \Sigma \in \Theta\} \leq C_+ < \infty$.*

Under these assumptions, we establish a finite sample concentration result for our estimator $\widehat{\Sigma}$:

**Theorem B.4.** *Under assumptions B.2 and B.3 we have the following*

$$\sqrt{n}\|\widehat{\Sigma} - \Sigma_0\|_{op} \leq t$$

*with probability atleast $1 - e^{-bt^2}$ for some constant $b > 0$.*

*Proof.* We break the proof of the theorem into a few small lemmas. Consider the collection $\mathcal{G} = \{g_\Sigma : \Sigma \in \Theta\}$, where

$$g_\Sigma(X, Y) = [Y \log F_*(X'\Sigma X) + (1-Y)\log\left(1 - F(X'\Sigma X)\right]$$

The problem of estimating $\Sigma_0$ using MLE can be viewed as a risk minimization problem over the collection of functions $\mathcal{G}$, which we are going to exploit later this section. Lemma B.5 below provides a lower bound on the deviation of $l(\Sigma)$ from $l(\Sigma_0)$ in terms of $\|\Sigma - \Sigma_0\|_{op}$:

**Lemma B.5.** *Under assumptions B.2 and B.3, we have a quadratic lower bound on the excess risk:*

*Proof.* From the definition of our model in ExPLORE, $F_*(t) = (2-\epsilon)/(1+e^t)$ which implies $F'_*(t) = -(2-\epsilon)e^t/(1+e^t)^2$. As $X$ is bounded (Assumption B.2),

$$\langle XX^T, \Sigma\rangle \leq \lambda_{max}(\Sigma)\|X\|_2^2 \leq C_+ U^2$$

for all $X \in \mathcal{X}, \Sigma \in \Theta$, where the constants $C_+$ and $U$ are as defined in Assumptions B.3 and B.2 respectively. Hence, there exists $\tilde{K} > 0$ such that $|F'_*(X'(\alpha\Sigma + (1-\alpha)\Sigma_0)X)| \geq \tilde{K}$ for all $X, \Sigma$. For notational simplicity define $D = XX^T$. From the definition of $l(\Sigma)$ we have:

$$\begin{aligned}l(\Sigma) &= E\left(F_*(\langle D, \Sigma_0\rangle)\log F_*(\langle D, \Sigma\rangle) + (1 - F_*(\langle D, \Sigma_0\rangle))(1 - \log F_*(\langle D, \Sigma\rangle))\right) \\ &= l(\Sigma_0) - E\left[KL\left(Bern(F_*(\langle D, \Sigma^0\rangle)) \,||\, Bern(F_*(\langle D, \Sigma\rangle))\right)\right]\end{aligned}$$

$$\leq l(\Sigma^0) - 2E\left[\left(F_*(\langle D, \Sigma_0\rangle) - F_*(\langle D, \Sigma\rangle)\right)^2\right] \tag{B.5}$$

where the last inequality follows from Pinsker's inequality. Using equation B.5 we can conclude:

$$\begin{aligned}
l(\Sigma^0) - l(\Sigma) &\geq 2E\left[\left(F_*(\langle D, \Sigma_0\rangle) - F_*(\langle D, \Sigma\rangle)\right)^2\right] \\
&\geq 2\tilde{K}^2 E\left[\left(\langle D, \Sigma - \Sigma^0\rangle\right)^2\right] \\
&\geq 2\tilde{K}^2\|\Sigma - \Sigma_0\|_{op}^2 E\left[\left(\langle D, \frac{\Sigma - \Sigma_0}{\|\Sigma - \Sigma_0\|_{op}}\rangle\right)^2\right] \\
&\geq 2\tilde{K}^2\|\Sigma - \Sigma_0\|_{op}^2 E\left[\left(X^T\frac{\Sigma - \Sigma_0}{\|\Sigma - \Sigma_0\|_{op}}X\right)^2\right] \\
&\geq 2c\tilde{K}^2\|\Sigma - \Sigma_0\|_{op}^2
\end{aligned}$$

Here we have used the fact that

$$\inf_{T\in S_d^{++}:\|T\|_{op}=1} E\left[\left(X^TTX\right)^2\right] = c > 0$$

To prove the fact, assume on the contrary that the infimum is 0. The set of all matrices $T$ with $\|T\|_{op} = 1$ is compact subset of $\mathbb{R}^{d\times d}$. Now consider the function:

$$f : T \longrightarrow E\left[\left(X^TTX\right)^2\right]$$

By DCT, $f$ is a continuous function. Hence the infimum will be attained, which means that we can find a matrix $M$ such that $M \in S_d^{++}$ and $\|M\|_{op} = 1$ such that $E\left[\left(X^TMX\right)^2\right] = 0$. Hence $X^TMX = 0$ almost surely. As the support of $A$ contains an open set, we can conclude $M = 0$ which contradicts $\|M\|_{op} = 1$.

□

Next we establish an upper bound on the variability of the centered function $g_\Sigma - g_{\Sigma_0}$ in terms of the distance function, which is stated in the following lemma:

**Lemma B.6.** *Under the aforementioned assumptions,*

$$Var\left(g_\Sigma - g_{\Sigma_0}\right) \lesssim d^2(\Sigma, \Sigma_0)$$

*where* $d(\Sigma, \Sigma_0) = \|\Sigma - \Sigma_0\|_{op}$.

*Proof.* We start with the observation

$$g_{\Sigma_0}(X, Y) - g_\Sigma(X, Y) = Y\log\frac{F_*(X'\Sigma_0X)}{F_*(X'\Sigma X)} + (1-Y)\log\frac{1 - F_*(X'\Sigma_0X)}{1 - F_*(X'\Sigma X)}$$

From our assumption on the parameter space, we know there exists $p > 0$ such that $p \leq F_*(X'\Sigma X) \leq 1 - p$ for all $\Sigma \in \Theta$ and for all $X$ almost surely. Hence,

$$\begin{aligned}
|g_{\Sigma_0}(X, Y) - g_\Sigma(X, Y)| &\leq \left|\log\frac{F_*(X'\Sigma_0X)}{F_*(X'\Sigma X)}\right| + \left|\log\frac{1 - F_*(X'\Sigma_0X)}{1 - F_*(X'\Sigma X)}\right| \\
&\leq 2K|F_*(X'\Sigma X) - F_*(X'\Sigma_0X)| \quad [K \text{ is the upper bound on the derivative of log}] \\
&\leq K|X'(\Sigma - \Sigma_0)X| \quad [\text{As } F_*' \leq 1/2] \\
&\leq KU\|\Sigma - \Sigma_0\|_{op}
\end{aligned}$$

This concludes the lemma. □

The following lemma establishes an upper bound on the modulus of continuity of the centered empirical process:

**Lemma B.7.** *Under the aforementioned assumptions, we have for any $\delta > 0$:*

$$E\left(\sup_{d(\Sigma,\Sigma_0)\leq\delta}\left|\mathbb{P}_n\left(g_{\Sigma_0}-g_\Sigma\right)-P\left(g_{\Sigma_0}-g_\Sigma\right)\right|\right)\lesssim\delta$$

*Proof.* Fix $\delta > 0$. Define $\mathcal{H}_\delta = \{h_\Sigma = g_\Sigma - g_{\Sigma_0} : \|\Sigma - \Sigma_0\|_{op} \leq \delta\}$. We can write $g_\Sigma - g_{\Sigma_0} = h^1_\Sigma + h^2_\Sigma$ where

$$h^{(1)}_\Sigma = Y\log\frac{F_*(X'\Sigma_0X)}{F_*(X'\Sigma X)},\ h^{(2)}_\Sigma = (1-Y)\log\frac{1-F_*(X'\Sigma_0X)}{1-F_*(X'\Sigma X)}$$

Hence $H_\delta \subset H^{(1)}_\delta + H^{(2)}_\delta$ where $H^{(i)}_\delta = \{h^{(1)}_\Sigma :\ \|\Sigma - \Sigma_0\|_{op} \leq \delta\}$for $i \in \{1, 2\}$. Next, we argue that $H^{(1)}_\delta$ has finite VC dimension. To see this, consider the function $(X, Y) \to \langle XX', \Sigma\rangle$. As this is linear function, it has finite VC dimension. Now the function $\log \circ F_*$ is monotone. As composition of monotone functions keeps VC dimension finite, we see that $(X, Y) \to \log F_*(\langle XX', \Sigma\rangle)$ is also VC class. It is also easy to see that projection map $(X, Y) \to Y$ is VC class, which implies the functions $(X, Y) \to Y \log F_*(\langle XX', \Sigma\rangle)$ form a VC class. As $\Sigma_0$ is fixed, then we can easily conclude the class of functions $(X, Y) \to Y \log \frac{F_*(X'\Sigma_0X)}{F_*(X'\Sigma X)}$ has finite VC dimension. By similar argument we can establish $H^{(2)}_\delta$ also has finite VC dimension. Let's say $V_i$ be the VC dimension of $H^{(i)}_\delta$. Define $h_\delta$ to be envelope function of $H_\delta$. Then we have,

$$\begin{aligned}
|h_\delta(X,Y)| &= \left|\sup_{\|\Sigma-\Sigma_0\|_{op}\leq\delta} h_\Sigma(X,Y)\right| \\
&\leq \sup_{\|\Sigma-\Sigma_0\|_{op}\leq\delta}|h_\Sigma(X,Y)| \\
&\leq \sup_{\|\Sigma-\Sigma_0\|_{op}]\leq\delta}\left[\left|\log F_*(X'\Sigma_0X)-\log F_*(X'\Sigma X)\right|+\left|\log\left(1-F_*(X'\Sigma_0X)\right)-\log\left(1-F_*(X'\Sigma X)\right)\right|\right] \\
&\leq 2K_1\sup_{\|\Sigma-\Sigma_0\|_{op}\leq\delta}|X'(\Sigma-\Sigma_0)X| \leq 2K_1U\delta
\end{aligned}$$

Note that, $h_\delta$ can also serve as an envelope for both $H^{(1)}_\delta$ and $H^{(2)}_\delta$. Using the maximal inequality from classical empirical process theory (e.g. see Theorem 2.14.1 in (van der Vaart & Wellner, 1996)) we get:

$$E\left(\sup_{d(\Sigma,\Sigma_0)\leq\delta}\left|P_n\left(g_{\Sigma_0}-g_\Sigma\right)-P\left(g_{\Sigma_0}-g_\Sigma\right)\right|\right)\leq J(1,\mathcal{H}_\delta)\sqrt{Ph^2_\delta}\leq J(1,\mathcal{H}_\delta)2K_1U\delta \tag{B.6}$$

for all $\delta > 0$, where

$$\begin{aligned}
J(1,\mathcal{H}_\delta) &= \sup_Q\int_0^1\sqrt{1+\log N(\epsilon\|h_\delta\|_{Q,2},\mathcal{H}_\delta,L_2(Q))}\,d\epsilon \\
&\leq \sup_Q\int_0^1\sqrt{1+\log N\left(\epsilon\|h_\delta\|_{Q,2},\mathcal{H}^{(1)}_\delta+\mathcal{H}^{(2)}_\delta,L_2(Q)\right)}\,d\epsilon \\
&\leq \sup_Q\int_0^1\sqrt{1+\sum_{i=1}^2\log N\left(\epsilon\|h_\delta\|_{Q,2},\mathcal{H}^{(i)}_\delta,L_2(Q)\right)}\,d\epsilon \\
&\leq \sup_Q\int_0^1\sqrt{1+\sum_{i=1}^2\left[\log K+\log V_i+V_i\log 16e+2(V_i-1)\log\frac{1}{\epsilon}\right]}\,d\epsilon
\end{aligned}$$

which is finite. This completes the proof. □

The last ingradient of the proof is a result due of Massart and Nedelec (Massart et al., 2006), which, applied to our setting, yields an exponential tail bound. For the convenience of the reader, we present below a tailor-made version of their result which we apply to our problem:

**Theorem B.8** (Application of Talagarand's inequality)**.** *Let $\{Z_i = (X_i, Y_i)\}_{i=1}^n$ be i.i.d. observations taking values in the sample space $\mathcal{Z} : \mathcal{X} \times \mathcal{Y}$ and let $\mathcal{F}$ be a class of real-valued functions defined on $\mathcal{X}$. Let $\gamma$ be a bounded loss function on $\mathcal{F} \times \mathcal{Z}$ and suppose that $f^* \in \mathcal{F}$ uniquely minimizes the expected loss function $P(\gamma(f, .))$ over $\mathcal{F}$. Define the empirical risk as $\gamma_n(f) = (1/n)\sum_{i=1}^n \gamma(f, Z_i)$, and $\bar{\gamma}_n(f) = \gamma_n(f) - P(\gamma(f, .))$. Let $l(f^*, f) = P(\gamma(f, .)) - P(\gamma(f^*, .))$ be the excess risk. Assume that:*

1. *We have a pseudo-distance $d$ on $\mathcal{F} \times \mathcal{F}$ satisfying $Var_P[\gamma(f, .) - \gamma(f^*, .)] \leq d^2(f, f^*)$.*
2. *There exists $F \subseteq \mathcal{F}$ and a countable subset $F' \subseteq F$, such that for each $f \in F$, there is a sequence $\{f_k\}$ of elements of $F'$ satisfying $\gamma(f_k, z) \to \gamma(f, z)$ as $k \to \infty$, for every $z \in \mathcal{Z}$.*
3. *$l(f, f^*) \geq d^2(f^*, f) \; \forall \; f \in \mathcal{F}$*
4. *$\sqrt{n}E\left[\sup_{f \in F' : d(f, f^*) \leq \sigma} [\bar{\gamma}_n(f) - \bar{\gamma}_n(f^*)]\right] \leq \phi(\sigma)$ for every $\sigma > 0$ such that $\phi(\sigma) \leq \sqrt{n}\sigma$.*

*Let $\epsilon_*$ be such that $\sqrt{n}\epsilon_*^2 \geq \phi(\epsilon_*)$. Let $\hat{f}$ be the (empirical) minimizer of $\gamma_n$ over $F$ and $l(f^*, F) = \inf_{f \in F} l(f^*, f)$.Then, there exists an absolute constant $K$ such that for all $y \geq 1$, the following inequality holds:*

$$P\left(l(f^*, \hat{f}) > 2l(f^*, F) + Ky\epsilon_*^2\right) \leq e^{-y}$$

The collection of function is $\mathcal{G} = \{g_\Sigma : \|\Sigma - \Sigma_0\|_{op}\}$. The corresponding pseudo-distance is $d(g_\Sigma, g_{\Sigma_0}) = \|\Sigma - \Sigma_0\|_{op}$. Condition 2 is easily satisfied as our parameter space has countable dense set and our loss function is continuous with respect to the parameter. Condition 1 and 3 follows form Lemma B.6 and Lemma B.5 respectively. Condition 4 is satisfied via Lemma B.7 with $\phi(\sigma) = \sigma$. Hence, in our case, we can take $\epsilon_n = \sqrt{n}$ and conclude that, there exists a constant K such that, for all $t \geq 1$,

$$P\left(n(l(\Sigma_0) - l(\widehat{\Sigma})) \geq Kt\right) \leq e^{-t}$$

From Lemma B.5 we have $\|\widehat{\Sigma} - \Sigma_0\|_{op}^2 \lesssim l(\Sigma_0) - l(\widehat{\Sigma})$ which implies

$$P\left(\sqrt{n}\|\widehat{\Sigma} - \Sigma_0\|_{op}^2 \geq K_1 t\right) \leq e^{-t^2}$$

which completes the proof of the theorem. □

We can combine Theorem B.4 with Proposition 3.1 and Proposition 3.2 of (Yurochkin et al., 2020) to show that EXPLORE in conjunction with SENSR trains individually fair ML models. For simplicity, we keep the notations same as in (Yurochkin et al., 2020). Define $\mathcal{L} = \{\ell(\cdot, \theta) \; : \; \theta \in \Theta\}$ as the loss class. We assume that:

1. We assume the embeded feature space of $\varphi$ is bounded $R \triangleq \max\{\mathsf{diam}(\varphi), \mathsf{diam}_*(\varphi)\} < \infty$, where $\mathsf{diam}_*$ is the diameter of $\varphi$ in the (unknown) exact fair metric

   $$d_x^*(x_1, x_2) = \langle(\varphi_1 - \varphi_2), \Sigma_0(\varphi_1 - \varphi_2)\rangle^{1/2},$$

   and $\mathsf{diam}$ is the diameter in the learned fair metric

   $$\hat{d}_x(x_1, x_2) = \langle(\varphi_1 - \varphi_2), \widehat{\Sigma}(\varphi_1 - \varphi_2)\rangle^{1/2}.$$

2. The loss functions in $\mathcal{L}$ is uniformly bounded, i.e. $0 \leq \ell(z, \theta) \leq M$ for all $z \in \mathcal{Z}$ and $\theta \in \Theta$ where $z = (x, y)$.
3. The loss functions in $\mathcal{L}$ is $L$-Lipschitz with respect to $d_x$, i.e.:

   $$\sup_{\theta \in \Theta}\left\{\sup_{(x_1, y), (x_2, y) \in \mathcal{Z}} |\ell((x_1, y), \theta) - \ell((x_2, y), \theta)|\right\} \leq L d_x(x_1, x_2);$$

Define $\delta^*$ to be bias term:

$$\min_{\theta\in\Theta}\sup_{P:W_*(P,P_*)\leq\epsilon}\left[\mathbb{E}_P\left(\ell(Z,\theta)\right)\right]=\delta^*$$

where $W_*$ is the Wasserstein distance with respect to the true matrix $\Sigma_0$ and $W$ is Wasserstein distance with respect to $\hat{\Sigma}$. Now for $x_1, x_2 \in \mathcal{X}$ we have:

$$\begin{aligned}\left|\hat{d}_x^2(x_1,x_2)-(d_x^*(x_1,x_2))^2\right| &= \left|(\varphi_1-\varphi_2)^\top\left(\widehat{\Sigma}-\Sigma^*\right)(\varphi_1-\varphi_2)\right| \\ &\leq \|\widehat{\Sigma}-\Sigma^*\|_{op}\|\varphi_1-\varphi_2\|_2^2 \\ &\leq R^2\|\widehat{\Sigma}-\Sigma^*\|_{op} \\ &\leq R^2K_1\frac{t}{\sqrt{n}}\end{aligned}$$

where the last inequality is valid with probability greater than or equal to $1-e^{-bt^2}$ from Theorem B.4. Hence we have with high probability:

$$\sup_{x_1,x_2\in\mathcal{X}}\left|\hat{d}_x^2(x_1,x_2)-(d_x^*(x_1,x_2))^2\right|\leq R^2K_1\frac{t}{\sqrt{n}}$$

Hence we can take $\delta_c = K_1 t/\sqrt{n}$ in Proposition 3.2 of (Yurochkin et al., 2020) to conclude that:

**Corollary B.9.** *If we assume he loss function $\ell \in \mathcal{L}$ and define the estimator $\hat{\theta}$ as:*

$$\widehat{\boldsymbol{\theta}} \in \arg\min_{\theta\in\Theta}\sup_{P:W(P,P_n)\leq\epsilon}\mathbb{E}_P\left[\ell(Z,h)\right],$$

*then the estimator $\hat{\theta}$ satisfies with probability greater than or equal to $1-t-e^{-t^2}$:*

$$\sup_{P:W_*(P,P_*)\leq\epsilon}\mathbb{E}_P\left[\ell(Z,\widehat{\boldsymbol{\theta}})\right]-\mathbb{E}_{P_*}\left[\ell(Z,\widehat{\boldsymbol{\theta}})\right]\leq\delta^*+2\delta_n, \tag{B.7}$$

*where $W$ and $W_*$ are the learned and exact fair Wasserstein distances induced by the learned and exact fair metrics (see Section 2.1 in Yurochkin et al. (2020)) and*

$$\delta_n\leq\frac{48\mathfrak{C}(\mathcal{L})}{\sqrt{n}}+\frac{48LR^2}{\sqrt{n\epsilon}}+\frac{LK_1tR^2}{\sqrt{n\epsilon}}+M\left(\frac{\log\frac{2}{t}}{2n}\right)^{\frac{1}{2}}.$$

*where $\mathfrak{C}(\mathcal{L})=\int_0^\infty\sqrt{\log\left(\mathcal{N}_\infty\left(\mathcal{L},r\right)\right)}\,dr$, with $\mathcal{N}_\infty\left(\mathcal{L},r\right)$ being the covering number of the loss class $\mathcal{L}$ with respect to the uniform metric.*

## C. Proofs of Theorems of Section 3

### C.1. Proof of Theorem 3.1

*Proof.* One key ingredient for the proof is a version of Davis-Kahane's $\sin\Theta$ theorem (Davis & Kahan, 1970), which we state here for convenience:

**Theorem C.1.** *Suppose $A, E \in \mathbb{R}^{d\times d}$. Define $\hat{A} = A + E$. Suppose $U$ (respectively $\hat{U}$) denote the top-k eigenvectors of $A$ (respectively $\hat{A}$). Define $\gamma = \lambda_k(A) - \lambda_{(k+1)}(A)$. Then if $\|E\|_{op} < \gamma$, we have:*

$$\|\hat{U}\hat{U}^T-UU^T\|_{op}\leq\frac{\|E\|_{op}}{\gamma-\|E\|_{op}}$$

In our context, let's define $U_k$ and $\hat{U}_k$ denote the eigenspace corresponding to top - $k$ eigenvectors of $\Sigma$ and $S_n$ respectively. Let $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d$ be the eigenvalues of $\Sigma$ and $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \geq \hat{\lambda}_d$ be eigenvalues of $S_n$. Applying the above theorem we obtain the following bound:

$$\|U_kU_k^*-\hat{U}_k\hat{U}_k^*\|_{op}\leq\frac{\|\Sigma-S_n\|_{op}}{\eta-\|\Sigma-S_n\|_{op}} \tag{C.1}$$

where $\eta = \lambda_k(\Sigma) - \lambda_{k+1}(\Sigma)$. To provide a high probability bound on $\|S_n - \Sigma\|_{op}$ we resort to Remark 5.40 ((Vershynin, 2011)), which implies that with probability $\geq 1 - 2e^{-ct^2}$:

$$\|\Sigma - S_n\|_{op} \leq \delta \vee \delta^2 \tag{C.2}$$

where $\delta = \frac{C\sqrt{d}+t}{\sqrt{n}}$. For $t < (\sqrt{n}\tilde{\gamma} - C\sqrt{d}) \wedge (\sqrt{n\tilde{\gamma}} - C\sqrt{d})$, $\eta > \delta \vee \delta^2$. Hence combining the bounds from equation C.1 and equation C.2 we have:

$$\|U_k U_k^* - \hat{U}_k \hat{U}_k^*\|_{op} \leq \frac{\delta \vee \delta^2}{\eta - (\delta \vee \delta^2)} \tag{C.3}$$

Here the constant $C, c$ depends only on $\|x_i\|_{\psi_2}$. To conclude the proof, we need a bound on the bias term $\|U_k U_k^T - \tilde{A}_* \tilde{A}_*^T\|_{op}$, which is obtained from another application of Theorem C.1. From the representation of $\Sigma$ we have:

$$\Sigma = A_* A_*^T + B_* B_*^T + \sigma^2 I_d = \tilde{A}_* \Lambda \tilde{A}_*^T + B_* B_*^T + \sigma^2 I_d$$

where $\tilde{A}_*$ is the set of eigenvectors of $A_*$ and $\Lambda$ is the diagonal matrix of the eigenvalues. We can apply Theorem C.1 on $\Sigma$ taking $A = \tilde{A}_* \Lambda \tilde{A}_*^T$, $E = B_* B_*^T + \sigma^2 I_d$ and $\Sigma = \hat{A}$. Here $\lambda_k(A) = \lambda_{min}(A_* A_*^T)$ and $\lambda_{k+1}(A) = 0$. Hence $\gamma = \lambda_{min}(A_* A_*^T)$. As by our assumption $\|B_* B_*^T + \sigma^2 I_d\|_{op} < \gamma = \lambda_{min}(A_* A_*^T)$, we obtain :

$$\|U_k U_k^T - \tilde{A}_* \tilde{A}_*^T\|_{op} \leq \frac{\|B_* B_*^T + \sigma^2 I_d\|_{op}}{\lambda_{min}(A_* A_*^T) - \|B_* B_*^T + \sigma^2 I_d\|_{op}} = b \tag{C.4}$$

To conclude the theorem, we provide a bound on $\eta = \lambda_k(\Sigma) - \lambda_{k+1}(\Sigma)$. To upper bound $\lambda_{k+1}(\Sigma)$ we use Courant-Fisher theorem:

$$\begin{aligned}\lambda_{k+1}(\Sigma) = \inf_{S \subseteq \mathbb{R}^d : \dim(S) = d-k} \sup_{x \in S^{d-1} \cap S} x^T \Sigma x &\leq \sup_{x \in S^{d-1} \cap \tilde{A}_*^{\perp}} x^T \Sigma x \\ &= \sup_{x \in S^{d-1} \cap \tilde{A}_*^{\perp}} x^T B_* B_*^T x + \sigma^2 \leq \|B_* B_*^T\|_{op} + \sigma^2\end{aligned}$$

The lower bound on $\lambda_k(\Sigma)$ can be obtained easily as follows: For any $x \in S^{d-1}$:

$$x^T \Sigma x = x^T A_* A_*^T x + x^T B_* B_*^T x + \sigma^2 \geq \lambda_{min}(A_* A_*^T) + \sigma^2$$

This automatically implies $\lambda_k(\Sigma) \geq \lambda_{min}(A_* A_*^T) + \sigma^2$. Hence combining the bound on $\lambda_k(\Sigma)$ and $\lambda_{k+1}(\Sigma)$ we get:

$$\eta = \lambda_k(\Sigma) - \lambda_{k+1}(\Sigma) \geq \lambda_{min}(A_* A_*^T) - \|B_* B_*^T\|_{op} = \tilde{\gamma} \tag{C.5}$$

Combining equation C.2, C.4 and C.5 and using the fact that:

$$\left\| \hat{U} \hat{U}^\top - \tilde{A}_* \tilde{A}_*^\top \right\|_{op} = \left\| \hat{\Sigma} - \Sigma_0 \right\|_{op}$$

we conclude the theorem. □

### C.2. Proof of Theorem 3.2

*Proof.* The variance covariance matrix of $\varphi_i$ can be represented as following:

$$\Sigma_\varphi = A_* A_*^T + B_* B_*^T + \sigma^2 I_d$$

As in the proof of the previous theorem, define $\lambda_1 \geq \cdots \geq \lambda_d$ as the eigenvalues of $\Sigma_\varphi$ and $\hat{\lambda}_1 \geq \cdots \geq \hat{\lambda}_d$ as the eigenvalues of $S_n$. Also define by $U_k$ (respectively $\hat{U}_k$) to be the matrix containing top-k eigenvectors of $\Sigma$ (respectively $S_n$) and $\eta = \lambda_k - \lambda_{k+1}$. Using Davis-Kahan's $\sin\Theta$ theorem (see Theorem C.1), we conclude that:

$$\|\hat{U}_k \hat{U}_k^T - U_k U_K^T\|_{op} \leq \frac{\|S_n - \Sigma_\varphi\|_{op}}{\gamma - \|S_n - \Sigma_\varphi\|_{op}} \tag{C.6}$$

provided that $\eta > \|S_n - \Sigma_\varphi\|_{op}$. Using matrix concentration inequality (see remark 5.40 of ((Vershynin, 2011))) we get that with probability $> 1 - 2e^{-ct^2}$:

$$\|S_n - \Sigma\|_{op} \leq \delta \vee \delta^2 + \frac{t}{n} \tag{C.7}$$

where $\delta = (C\sqrt{d} + t)/\sqrt{n}$, for all $t \geq 0$. The difference between this and equation C.2 in Theorem 3.1 is the extra term $t/n$, which appears due to mean centering the samples. The constants $c, C$ only depends on the $\psi_2$ norm of $\varphi_i$. Combining equation C.6 and C.7 we conclude that, with high probability we have

$$\|\hat{U}_k\hat{U}_k^T - U_kU_k^T\|_{op} \leq \frac{\delta \vee \delta^2 + t/n}{\eta - (\delta \vee \delta^2) - t/n}$$

when $t/n + \delta \vee \delta^2 < \eta$. As before, we apply Theorem C.1 to control the bias. Towards that end, define $A = A_*A_*^T = \tilde{A}_*\Lambda\tilde{A}_*^T$, where $\tilde{A}_*$ is the matrix of eigenvectors of $A_*$ and $\Lambda$ is diagonal matrix with the eigenvalues of $A_*A_*^T$. Also define $E = B_*B_*^T + \sigma^2 I_d$ and $\hat{A} = \Sigma_\varphi$. Now, as before, $\lambda_k(A) = \lambda_{min}(A_*A_*^T)$ and $\lambda_{k+1}(A) = 0$. Hence $\gamma = \lambda_k(A) - \lambda_{k+1}(A) = \lambda_{min}(A_*A_*^T)$. Applying Theorem C.1 we conclude:

$$\|U_kU_k^T - \tilde{A}_*\tilde{A}_*^T\|_{op} \leq \frac{\|B_*B_*^T + \sigma^2 I_d\|_{op}}{\lambda_{min}(A_*A_*^T) - \|B_*B_*^T + \sigma^2 I_d\|_{op}} \tag{C.8}$$

Finally, we use Courant-Fischer Min-max theorem to provide an upper bound on $\eta = \lambda_k(\Sigma_\varphi) - \lambda_{k+1}(\Sigma_\varphi)$. As in the previous proof we have:

$$\begin{aligned}\lambda_{k+1}(\Sigma_\varphi) = \inf_{S\subseteq\mathbb{R}^d:\dim(S)=k+1} \sup_{x\in S^{d-1}\cap S} x^T\Sigma_\varphi x &\leq \sup_{x\in S^{d-1}\cap\tilde{A}_*^\perp} x^T\Sigma_\varphi x \\ &= \sup_{x\in S^{d-1}\cap\tilde{A}_*^\perp} x^TB_*B_*^Tx + \sigma^2 \leq \|B_*B_*^T\|_{op} + \sigma^2\end{aligned}$$

$$\begin{aligned}\lambda_{k+1}(\Sigma_\varphi) = \sup_{S\subseteq\mathbb{R}^d:\dim(S)=d-k} \sup_{x\in S^{d-1}\cap S} x^T\Sigma_\varphi x &\leq \sup_{x\in S^{d-1}\cap\tilde{A}_*^\perp} x^T\Sigma_\varphi x \\ &= \sup_{x\in S^{d-1}\cap\tilde{A}_*^\perp} x^TB_*B_*^Tx + \sigma^2 \leq \|B_*B_*^T\|_{op} + \sigma^2\end{aligned}$$

To get a lower bound on $\lambda_k(\Sigma_\varphi)$, we use the the other version of Courant-Fischer Minmax theorem:

$$\lambda_k(\Sigma_\varphi) = \max_{S:dim(S)=d-k+1} \min_{x\in S^{d-1}\cap S} x^T\Sigma x$$

Using this we conclude:

$$\lambda_k(\Sigma_\varphi) \geq \lambda_{min}(A_*A_*^T) + \sigma^2$$

Hence combining the bound on $\lambda_k(\Sigma_\varphi)$ and $\lambda_{k+1}(\Sigma_\varphi)$ we get:

$$\eta = \lambda_k(\Sigma_\varphi) - \lambda_{k+1}(\Sigma_\varphi) \geq \lambda_{min}(A_*A_*^T) - \|B_*B_*^T\|_{op} = \tilde{\gamma} \tag{C.9}$$

Combining equation C.7, C.8 and C.9 and using the fact that:

$$\left\|\hat{U}\hat{U}^\top - \tilde{A}_*\tilde{A}_*^\top\right\|_{op} = \left\|\hat{\Sigma} - \Sigma_0\right\|_{op}$$

we conclude the theorem. □

### C.3. Proof of Theorem 3.3

*Proof.* The proof of Theorem 3.3 essentially follows form Proposition 3.2 and Proposition 3.1 of (Yurochkin et al., 2020). Note that from Theorem 3.1, for any $x_1, x_2 \in \mathcal{X}$:

$$\left|\hat{d}_x^2(x_1, x_2) - (d_x^*(x_1, x_2))^2\right| = \left|(\varphi_1 - \varphi_2)^\top\left(\widehat{\Sigma} - \Sigma^*\right)(\varphi_1 - \varphi_2)\right|$$

$$
\begin{aligned}
&\leq \|\widehat{\Sigma} - \Sigma^*\|_{op} \|\varphi_1 - \varphi_2\|_2^2 \\
&\leq R^2 \|\widehat{\Sigma} - \Sigma^*\|_{op} \\
&\leq R^2 \left[ b + \frac{\delta \vee \delta^2}{\tilde{\gamma} - (\delta \vee \delta^2)} \right]
\end{aligned}
$$

where the last inequality is true with probability greater than or equal to $1 - 2e^{-ct^2}$ from Theorem 3.1. This justifies taking $\delta_c \geq \left[ b + \frac{\delta \vee \delta^2}{\tilde{\gamma} - (\delta \vee \delta^2)} \right]$ which along with Proposition 3.1 and 3.2 of (Yurochkin et al., 2020) completes the proof. □