Random Matrix Methods for Machine Learning

Romain Couillet
University Grenoble Alpes, France
romain.couillet@gipsa-lab.grenoble-inp.fr

Zhenyu Liao
Huazhong University of Science and Technology, China
zhenyu_liao@hust.edu.cn [1]

April 17, 2023

Erratum

Theorem 2.11 (Inspired by Mestre [2008]). *Under the setting of Theorem 6 with* $\mathbb{E}[|\mathbf{Z}_{ij}|^4] < \infty$ *and* $\max_{1\leq i\leq p} \operatorname{dist}(\lambda_i(\mathbf{C}), \operatorname{supp}(\nu)) \to 0$, *let* $f : \mathbb{C} \to \mathbb{C}$ *be a complex function analytic on the complement of* $\gamma(\mathbb{C} \setminus \operatorname{supp}(\mu))$ *in* $\mathbb{C}$ *with* γ *defined in (2.39). Then,*

$$\frac{1}{p}\sum_{i=1}^{p} f(\lambda_i(\mathbf{C})) - \frac{1}{2c\pi\imath}\oint_{\Gamma_\mu} f\left(\frac{-1}{m_{\frac{1}{n}\mathbf{X}^\intercal\mathbf{X}}(\omega)}\right)\omega m'_{\frac{1}{n}\mathbf{X}^\intercal\mathbf{X}}(\omega)d\omega \xrightarrow{a.s.} 0,$$

for some complex positively oriented contour $\Gamma_\mu \subset \mathbb{C}$ *surrounding* $\operatorname{supp}(\mu)\setminus\{0\}$. *In particular, if* $c < 1$, *the result holds for any* f *analytic on* $\{z \in \mathbb{C},\ \Re[z] > 0\}$ *with* Γ_μ *chosen as any such contour within* $\{z \in \mathbb{C},\ \Re[z] > 0\}$.

Section Equation (2.43).

$$\ell_a - \hat{\ell}_a \xrightarrow{a.s.} 0, \quad \hat{\ell}_a = -\frac{n}{p_a}\frac{1}{2\pi\imath}\oint_{\Gamma_\mu^{(a)}} \omega\frac{m'_{\frac{1}{n}\mathbf{X}^\intercal\mathbf{X}}(\omega)}{m_{\frac{1}{n}\mathbf{X}^\intercal\mathbf{X}}(\omega)}d\omega. \tag{2.43}$$

Section 3.1.1 "GLRT asymptotics" around Equation (3.2). As a consequence, in order to set a maximum false alarm rate (or false positive, or Type I error) of $r > 0$ in the limit of large n, p, one must choose a threshold $f(\alpha)$ for T_p such that

$$\mathbb{P}(T_p \geq f(\alpha)) = r,$$

that is, such that

$$\mu_{\mathrm{TW}_1}([A_p, +\infty)) = r, \quad A_p = (f(\alpha) - (1+\sqrt{c})^2)(1+\sqrt{c})^{-\frac{4}{3}}c^{\frac{1}{6}}n^{\frac{2}{3}} \tag{3.2}$$

with μ_{TW_1} the Tracy-Widom measure in Theorem 2.15.

Section 3.1.2 "Linear and Quadratic Discriminant Analysis" before Remark 3.1 Plugging this result into the expression of $T^{(\gamma)}_{\mathrm{LDA}}(\mathbf{x})$, we find that

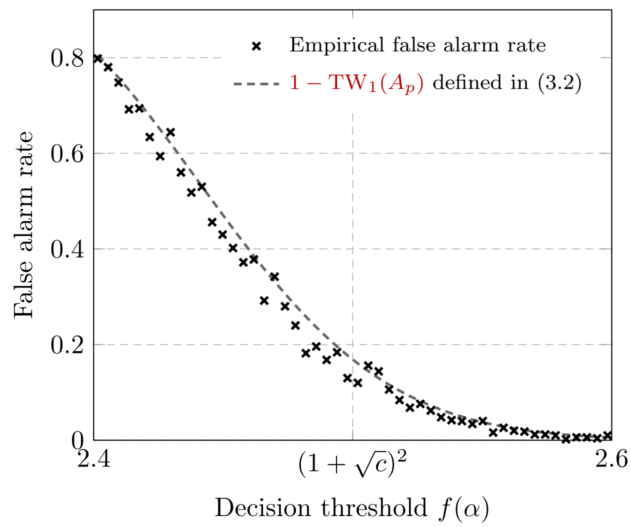

Figure 1: Comparison between empirical false alarm rates and $1 - \mathrm{TW}_1(A_p)$ for A_p of the form in (3.2), as a function of the threshold $f(\alpha) \in [(1+\sqrt{c})^2 - 5n^{-2/3}, (1+\sqrt{c})^2 + 5n^{-2/3}]$, for $p = 256$, $n = 1\,024$ and $\sigma = 1$. Results obtained from 500 runs. Link to code: Matlab and Python.

in the large n_0, n_1, p limit,

$$
\begin{aligned}
T_{\mathrm{LDA}}^{(\gamma)}(\mathbf{x}) &= \frac{(-1)^\ell}{2}(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)^\mathsf{T} \bar{\mathbf{Q}}^\circ (\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1) - \frac{g_0(-\gamma)}{2c_0} + \frac{g_1(-\gamma)}{2c_1} \\
&+ \mathbf{z}^\mathsf{T} \mathbf{C}_\ell^{\frac{1}{2}} \mathbf{Q}^\circ \mathbf{U} \begin{bmatrix} 1 \\ -1 \\ \frac{1}{\gamma \tilde{g}_0(-\gamma)} \\ -\frac{1}{\gamma \tilde{g}_1(-\gamma)} \end{bmatrix} + o(1)
\end{aligned}
$$

where we used in particular the fact that $\frac{1-\gamma\tilde{g}_0(-\gamma)}{\gamma\tilde{g}_0(-\gamma)} = g_0(-\gamma)$.

Theorem 2.11 (Optimal decision threshold)**.** *Since* $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_p)$*, it is clear that the expectation* $\mathbb{E}[T_{\mathrm{LDA}}^{(\gamma)}(\mathbf{x})]$ *is dominated by* $\pm\frac{1}{2}(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)^\mathsf{T} \bar{\mathbf{Q}}^\circ (\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)$ *which is positive when* $\ell = 0$ *and negative when* $\ell = 1$*, as expected. Yet, the term* $\frac{g_1(-\gamma)}{2c_1} - \frac{g_0(-\gamma)}{2c_0}$ *intervenes as a bias. If* $\mathbf{C}_0 = \mathbf{C}_1$ *(which is indeed the assumption of LDA)* and *the training set is "balanced" with* $c_0 = c_1$*, then* $g_0 = g_1$ *and this bias disappears; however, for* $\mathbf{C}_0, \mathbf{C}_1$ *distinct, this bias in general remains and must be accounted for in the decision threshold which, therefore,* should not be zero.

Section 5.1.1 "Regression with random neural network" after Equation (5.12). The fact that this denominator scales like $\|\gamma\bar{\mathbf{Q}}\|$ as $\gamma \to 0$ explains the major difference between the training and test error behavior in

Figure 5.5. Due to the γ^2 prefactor in $\bar{E}_{\rm train}$, the training error is guaranteed to be finite (even possibly to vanish) as $\gamma \to 0$. But for the test error, since $\gamma\bar{\mathbf{Q}} \to 0$ as N approaches n from each side, if the numerator term $\frac{1}{\hat{n}} \operatorname{tr} \bar{\mathbf{K}}_{\hat{\mathbf{X}}\hat{\mathbf{X}}} - \frac{1}{\hat{n}} \operatorname{tr}(\mathbf{I}_n + \gamma\bar{\mathbf{Q}})(\bar{\mathbf{K}}_{\mathbf{X}\hat{\mathbf{X}}}\bar{\mathbf{K}}_{\mathbf{X}\hat{\mathbf{X}}}^{\mathsf{T}}\bar{\mathbf{Q}})$ does not scale like $\gamma\bar{\mathbf{Q}}$, then $\bar{E}_{\rm test}$ diverges to infinity at $N = n$. A first counterexample is of course when $\hat{\mathbf{X}} = \mathbf{X}$, for which the numerator term of $\bar{E}_{\rm test}$ is now

$$\frac{1}{\hat{n}} \operatorname{tr} \bar{\mathbf{K}}_{\hat{\mathbf{X}}\hat{\mathbf{X}}} - \frac{1}{\hat{n}} \operatorname{tr}(\mathbf{I}_n + \gamma\bar{\mathbf{Q}})(\bar{\mathbf{K}}_{\mathbf{X}\hat{\mathbf{X}}}\bar{\mathbf{K}}_{\mathbf{X}\hat{\mathbf{X}}}^{\mathsf{T}}\bar{\mathbf{Q}}) = \frac{\gamma^2}{n} \operatorname{tr} \bar{\mathbf{Q}}\bar{\mathbf{K}}\bar{\mathbf{Q}}$$

Bibliography

Xavier Mestre. Improved Estimation of Eigenvalues and Eigenvectors of Covariance Matrices Using Their Sample Estimates. *IEEE Transactions on Information Theory*, 54(11):5113–5129, 2008. ISSN 0018-9448. doi: 10.1109/tit.2008.929938.