Assessment 4

Due by Wednesday 4 October 2023 9:00am

Question 1 [4 marks]

Suppose that A is a 2×2 Wishart distributed random matrix (i.e., $A \sim W_p(n, \Sigma)$ with $p = 2$ and $n \geq 2$). Prove using elementary methods that $\det(A)$ is distributed like $\det(\Sigma)$ times two independent chi-squared random variables with degrees of freedom n and $n-1$.

Question 2 [4 marks]

Consider a sequence of two-dimensional random vectors $\mathbb{x}_1, \mathbb{x}_2, \ldots, \mathbb{x}_n$ where the coordinates of $\mathbb{x}_n := (x_{n1}, x_{n2})'$. Suppose that

$$\sqrt{n}\,\mathbb{x}_n \to \mathbf{N}_2\Big(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}3 & 1\\1 & 2\end{bmatrix}\Big),$$

then what is the asymptotic distribution of $\xi := \sqrt{n}\,(x_{n1} + x_{n2}^2)$ as $n \to \infty$?

Question 3 [12 marks]

Algorithms and tools built using neural networks have gain a lot of promenance over the last 10 years. However, this uptake is probably more due to increases in processing power of modern computers as well as in the availability of large datasets rather than in the development of new mathematics. It is interesting to understand how random matrix theory (RMT) can play a role in understanding how neural networks function. Given p-dimensional data observations $\mathbb{x}_1, \mathbb{x}_2, \ldots, \mathbb{x}_n$, we stack them into a data matrix $\mathbb{X} = [\mathbb{x}_1, \ldots, \mathbb{x}_n]$ of size $p \times n$. A classic neural network model is given by $f(W\mathbb{X})$ where W is a $r \times p$ matrix of weights and $f : \mathbb{R} \to \mathbb{R}$ is a function (applied component-wise). This model and its connection to RMT has been recently studied in **[A]** and **[B]**. Setting $Z := f(W\mathbb{X})$, the authors in both papers study the Gram matrix

$$\frac{1}{n} ZZ^T,$$

where W is a random matrix with univariate normal distributed $N(0, \sigma_w^2/p)$ entries and $\mathbb{X}$ has entries with distribution $N(0, \sigma_x^2)$. This is sometimes called a *random feature network*.

[2] (a) In **[A]**, they consider the case where f is a variant of the *rectified linear unit* (ReLU) activation function given by f_α; see equation (17) in paper. Implement this function f_α and plot it for various values of α and compare it to the (classic) ReLU function.

[2] (b) The paper defines (see Theorem 1) that

$$\eta := \mathbb{E}[f(\sigma_w \sigma_x \xi)^2], \quad \zeta := \mathbb{E}[\sigma_w \sigma_x f'(\sigma_w \sigma_x \xi)]^2,$$

where $\xi \sim N(0, 1)$. The paper claims that when $f = f_\alpha$, it is straightforward to check that (see near equation (18) in paper),

$$\eta = 1, \quad \zeta = \frac{(1-\alpha)^2}{2(1+\alpha)^2 - \frac{2}{\pi}(1+\alpha)^2}.$$

Proceed to check this is true. Note that $[x]_+ = \max(0, x) = x\mathbf{1}_{\{x>0\}}$.

[2] (c) Consider the empirical spectral distribution (ESD) of $\frac{1}{n}Z_\alpha Z_\alpha^T$ for various choices of α where $Z_\alpha := f_\alpha(W\mathbb{X})$ and take $\sigma_w = \sigma_x = 1$. Compare the histogram of the ESD against the Marchenko-Pastur density. Justify and discuss your choices of parameters (e.g., see Section 3.2.2 in **[A]**). What is the value of α that best matches the Marchenko-Pastur distribution?

[2] (d) In Section 4.1 of **[A]**, they consider a *deep feedforward neural network* with ℓth-layer post-activation matrix given by,

$$Y^\ell = f(W^\ell Y^{\ell-1}), \quad Y^0 = \mathbb{X}.$$

Implement this model in R.

[2] (e) We would like to understand the distance between a limit spectral distribution (LSD) $\bar{\rho}_1$ and the empirical spectral distribution (ESD) of $Y^\ell(Y^\ell)^T$. The paper gives a distance metric

$$d(\bar{\rho}_1, \rho_1) := \int |\bar{\rho}_1(x) - \rho_\ell(x)|\, dx$$

Implement this function and apply it to the case $\ell = 1$ and taking $\bar{\rho}_1$ to be the Marchenko-Pastur density.

[2] (f) Use the model implemented in (d) to perform the experiment given in Figure 1 of **[A]** in the case where $f = f_\alpha$ and we are close to the Marchenko-Pastur distribution for the first-layer limiting distribution. Reproduce the figure in the cases $\ell = 1$ and $ell = 3$ (Note that $\ell = L$ in Figure 1).

References

[A] Pennington, Worah (2017). Nonlinear random matrix theory for deep learning. NeuroIPS 2017.

[B] Louart, Liao, Couillet (2018). A random matrix approach to neural networks. Annals of Applied Probability, Vol 28., No. 2.