Homework 2

Due by Monday 27 August 2018 10:00

$\sim$ *Version 3* $\sim$

In this homework, we are going to consider the recent paper by researchers at Google titled *Nonlinear Random Matrix Theory for Deep Learning* (NIPS 2017). When people talk about AI they usually mean 'Deep Learning' which is the idea that you can obtain amazing results by using a cascade of neural networks. Unfortunately, although the results are good, there is little mathematical theory to explain why it works so well. This paper suggests that maybe we can use Random Matrix Theory to achieve this goal.

Layers in neural networks roughly have the form $Y = f(WX)$ where X is a data matrix, $W = [w_{ij}]$ is a matrix of weights, and $f : \mathbf{R} \to \mathbf{R}$ is a nonlinear activation function (e.g., $f(x) := max(0, x)$) that applies elementwise on the matrix WX; see Figure 2 in [1]. Fitting a (one-layer) Deep Learning model involves finding the weight matrix W given some training data (Y, X) (e.g., labels Y for inputs X). A multi-layer model is a cascade that takes the form of (15) in [2].

In [2], they consider the following stylised setup. $X = [X_{i\mu}]$ is a $n_0 \times m$ random data matrix with i.i.d. elements $X_{i\mu} \sim N(0, \sigma_x^2)$ and $W = [W_{ij}]$ is a random weight matrix with i.i.d. elements $W_{ij} \sim N(0, \sigma_w^2/n_0)$. They set $\phi = n_0/m$ and $\psi = n_0/n_1$ and consider the regime where $n_0, n_1, m \to \infty$ while ϕ and ψ remain constant. The function $f : \mathbf{R} \to \mathbf{R}$ must satisfy

$$\mathbf{E}[f(\sigma_w \sigma_x Z)] = 0, \qquad \mathbf{E}[|f(\sigma_w \sigma_x Z)|^k] < \infty \text{ for } k > 1,$$

for random variable $Z \sim N(0, 1)$. Let $Y = f(WX)$ where f is applied elementwise to the matrix WX, they then proceed to study the $n_1 \times n_1$ matrix $M := \frac{1}{m} YY'$.

Question 1 [5 Points]

The empirical spectral density of M is given by

$$\rho_M(t) = \frac{1}{n_1} \sum_{j=1}^{n_1} \delta_{\lambda_j}(t)$$

where $\lambda_1, \ldots, \lambda_{n_1}$ are the n_1 eigenvalues of M. Show that the Stieltjes transform G of ρ_M is given by

$$G(z) = -\frac{1}{n_1} \text{tr}(M - zI_{n_1})^{-1},$$

where I_{n_1} is a $n_1 \times n_1$ identity matrix and tr is the trace operator.

Question 2 [5 Points]

Let the nonlinear activation function $f : \mathbf{R} \to \mathbf{R}$ be given by

$$f_\alpha(x) = \frac{[x]_+ + \alpha[-x]_+ - \frac{1+\alpha}{\sqrt{2\pi}}}{\sqrt{\frac{1}{2}(1+\alpha^2) - \frac{1}{2\pi}(1+\alpha)^2}}$$

where $[x]_+$ is the positive part of x. Show that

$$\mathbf{E}[f_\alpha(Z)] = 0 \quad \text{and} \quad \zeta := \mathbf{E}[f'_\alpha(Z)]^2 = \frac{(1-\alpha)^2}{2(1+\alpha)^2 - \frac{2}{\pi}(1+\alpha)^2}$$

for $Z \sim N(0,1)$. In particular, show that when $\alpha = 1$ we can write f_α as a simpler (well known) function and that we also get $\zeta = 0$. *Note that we are taking $\sigma_x = \sigma_w = 1$ which seems to be the assumption that they are making in the paper.*

Question 3 [5 Points]

Consider the case $\zeta = 0$ (ie., see Section 3.2.2). Show that the Stieltjes transform G_{MP} of the standard (i.e., $\sigma^2 = 1$) Marchenko-Pastur distribution with shape $y = \phi/\psi$, satisfies the equation

$$zG^2 + \Big((1-z)\frac{\psi}{\phi} - 1\Big)G + \frac{\psi}{\phi} = 0.$$

Note: this is different to (14) in [2], and this is using their definition of Stieltjes transform given by (5).

Question 4 [5 Points]

Reproduce the numerical experiment of Figure 1 in [2] but only for the case that $\alpha = 1$ but with $L = 1$, $L = 5$, and $L = 10$.

Note 1: If you computer is not fast enough, you can consider $L = 1$, $L = 3$, and $L = 5$ with $n_0 = \{100, 250, 500, 750, 1000, 1500\}$. Note 2: There is another typo in the paper, you should consider the eigenvalues of $\frac{1}{m}Y^L(Y^L)^T$ where the L is the last layer (i.e., notice the division by m). Note 3: If `E` contains the eigenvalues, `y` is the parameter of the MP distribution, and `dmp` is the density of the MP, then I considered a numerical approximation of the $L^2(a,b)$ difference by:

```
a <- (1-sqrt(y))^2
b <- (1+sqrt(y))^2
h <- hist(E, breaks=length(E), xlim=c(0,1.5*b), freq=FALSE)
domain <- h$mids > a & h$mids < b
x <- h$mids[domain]
d <- h$density[domain]
plot(x, dmp(x, y), type='s', col=3)
lines(x, d, type='s')
sqrt(sum((dmp(x, y) - d)^2)) # metric
```

Question 5 [5 Points]

Now redo the numerical experiment for the nonlinear activation function f_α for $\alpha \approx 1$ (but less than 1). What is your conclusion about this numerical experiment? And your opinion of the paper overall?

References

[1] LeCun, Bengio, and Hinton (2014). *Deep learning*. Nature.
[2] Pennington and Worah (2017). *Nonlinear random matrix theory for deep learning*. NIPS.

This homework is to be submitted through Wattle in digital form only as per ANU policy. The R code must be supplied. If you use any references (note: this will never count against you), please clearly indicate which ones.