STAT7017 Assignment 2

Rui Qiu

2018-08-20

Q1

Proof:

The Stieltjes transform G of ρ_m is,

$$
\begin{aligned}
G(z) &= \int \frac{\rho_M(t)}{z-t} dt \\
&= \int \frac{\frac{1}{n_1}\sum_{j=1}^{n_1} \delta_{\lambda_j}(t)}{z-t} dt \\
&= \frac{1}{n_1}\sum_{j=1}^{n_1} \int \frac{1}{z-t} \delta_{\lambda_j}(t) dt \\
&= \frac{1}{n_1}\sum_{j=1}^{n_1} \frac{1}{z-\lambda_j} \\
&= -\frac{1}{n_1}[\mathrm{tr}(M - z\mathbf{I}_{n_1})^{-1}]
\end{aligned}
$$

Note that

$$
\sum_{i=1}^{p} \lambda_i^k = tr(A^k), tr(A+B) = tr(A) + tr(B).
$$

According to the Stieltjes transformation of the M-P law,

$$
G(z) = \mathbb{E}G(z) = -\frac{1}{n_1}[\mathrm{tr}(M - z\mathbf{I}_{n_1})^{-1}]
$$

Q2

The following derivations are based on the assumption that $\sigma_X = \sigma_W = 1$.

Since

$$
\mathbb{E}[cX] = c\mathbb{E}[X]
$$

and σ_W, σ_X are constants, it is equal to show that

$$
\mathbb{E}[f_\alpha(Z)] = 0.
$$

$$
\mathbb{E}[f_\alpha(z)] = \mathbb{E}\left[\frac{[z]_+ + \alpha[-z]_+ - \frac{1+\alpha}{\sqrt{2\pi}}}{\sqrt{\frac{1}{2}(1+\alpha)^2 - \frac{1}{2\pi}(1+\alpha)^2}}\right]
$$

Since $z \sim N(0,1)$, $\mathbb{E}(z) = \mathbb{E}(-z)$. Meanwhile, the denominator is constant. What we are really interested here is

$$
\begin{aligned}
K &= \mathbb{E}\left([z]_+ + \alpha[-z]_+ - \frac{1+\alpha}{\sqrt{2\pi}}\right) \\
&= \mathbb{E}[z]_+ + \alpha\mathbb{E}[z]_+ - \frac{1+\alpha}{\sqrt{2\pi}} \\
\mathbb{E}[x]_+ &= \int_0^\infty x \cdot \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) dx \\
&= \frac{1}{\sqrt{2\pi}} \int_0^\infty x \exp\left(-\frac{x^2}{2}\right) dx \\
&= \frac{1}{\sqrt{2\pi}} \left(-\exp(-\frac{x^2}{2})\right)\Bigg|_0^\infty \\
&= \frac{1}{\sqrt{2\pi}} \\
K &= (1+\alpha)\frac{1}{\sqrt{2\pi}} - \frac{1+\alpha}{\sqrt{2\pi}} = 0
\end{aligned}
$$

Therefore, $\mathbb{E}[f(\sigma_W \sigma_X z)] = \frac{0}{\text{a non-zero constant}} = 0$.

Next, set $x = \sigma_W \sigma_X z$, then

$$
\begin{aligned}
f'(x) &= \frac{\frac{1}{2}(1-\alpha) \cdot \sqrt{\frac{1}{2}(1+\alpha^2) - \frac{1}{2\pi}(1+\alpha)^2}}{\frac{1}{2}(1+\alpha^2) - \frac{1}{2\pi}(1+\alpha)^2} \\
\mathbb{E}[f'(x)] = f'(x) &= \frac{\frac{1}{2}(1-\alpha)}{\sqrt{\frac{1}{2}(1+\alpha^2) - \frac{1}{2\pi}(1+\alpha)^2}}
\end{aligned}
$$

$$
\mathbb{E}[f'(x)]^2 = \frac{(1-\alpha)^2}{2(1+\alpha^2) - \frac{2}{\pi}(1+\alpha)^2}.
$$

Now consider the case that $\alpha = 1$,

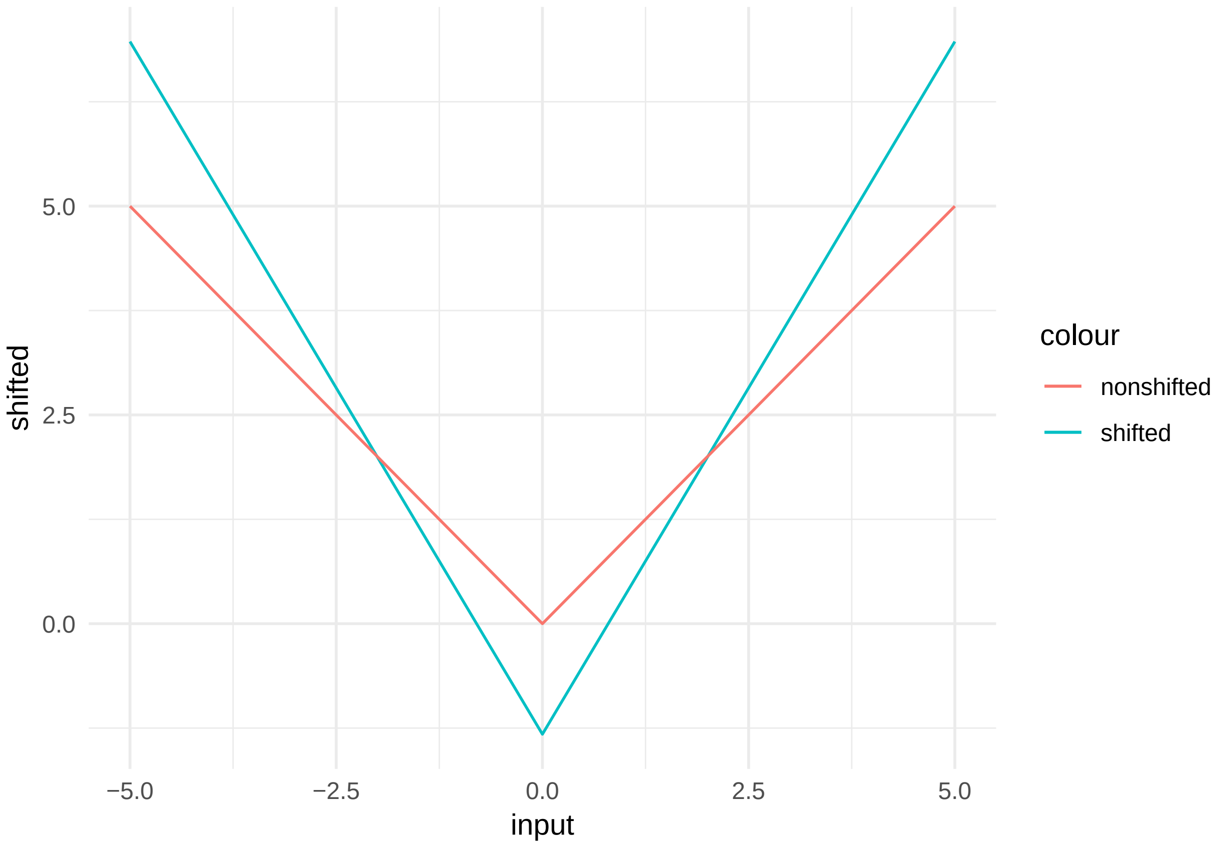
$$
f_\alpha(x) = \frac{[x]_+ + [-x]_+ - \frac{2}{\sqrt{2\pi}}}{\sqrt{1-\frac{2}{\pi}}} = \frac{|x| - \frac{2}{\sqrt{2\pi}}}{\sqrt{1-\frac{2}{\pi}}}.
$$

It looks like a "shifted and streched aboslution value function". For example,

```r
library(ggplot2)
f <- function(x) {
    return((abs(x)-2/sqrt(2*pi))/(sqrt(1-2/pi)))
}
x <- seq(-5,5,0.1)
dat <- as.data.frame(cbind(x,f(x),abs(x)))
colnames(dat) <- c("input", "shifted","nonshifted")
head(dat)
```

```
##   input  shifted nonshifted
## 1  -5.0 6.970876        5.0
## 2  -4.9 6.804986        4.9
## 3  -4.8 6.639096        4.8
## 4  -4.7 6.473207        4.7
## 5  -4.6 6.307317        4.6
## 6  -4.5 6.141427        4.5
```

```r
ggplot(dat,aes(x=input)) +
    geom_line(aes(y=shifted, color="shifted")) +
    geom_line(aes(y=nonshifted, color="nonshifted")) +
    theme_minimal()
```



To ensure $\zeta = 0$, we can simply plug in $\alpha = 1$ back to $\xi = \frac{(1-\alpha)^2}{2(1+\alpha^2)-\frac{2}{\pi}(1+\alpha)^2} = 0$.

Q3

According to equation (11) in the paper.

$$P = \frac{G - \frac{1-\psi}{z}}{\psi/z} = \frac{zG - 1 + \psi}{\psi} = \frac{z}{\psi}G - \frac{1}{\psi} + 1$$

With $\zeta = 0$, also WLOG set $\eta = 1$, according to equation (12) & (13) (plug in $t = \frac{1}{z\psi}$):

$$P = 1 + \frac{1}{z\psi}[1 + (P-1)\phi][1 + (P-1)\psi]$$

Let $Q = P - 1$:

$$Q = P - 1 = \frac{z}{\psi}G - \frac{1}{\psi}$$

Rearrange the previous equation:

$$Q \cdot z\psi = (1 + Q\phi)(1 + Q\psi) = 1 + Q(\phi + \psi) + Q^2\phi\psi$$

$$\phi\psi Q^2 + (\phi + \psi - z\psi)Q + 1 = 0$$

Now plug in Q with z, ψ, G:

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

$$\left(\frac{\phi z^2}{\psi}\right)G^2 + \left(\frac{z\phi}{\psi} + z - z^2 - \frac{2z}{\psi^2}\right) \cdot \phi\psi G + \frac{\phi}{\psi} - \frac{\phi}{\psi} - 1 + 1 + z = 0$$

$$\frac{\phi z^2}{\psi}G^2 + \left(z[1 - \frac{\phi}{\psi} - z]\right)G + z = 0 \quad (\star)$$

$(\star)$ times $\frac{\psi}{\phi z}$ to generate our target equation:

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

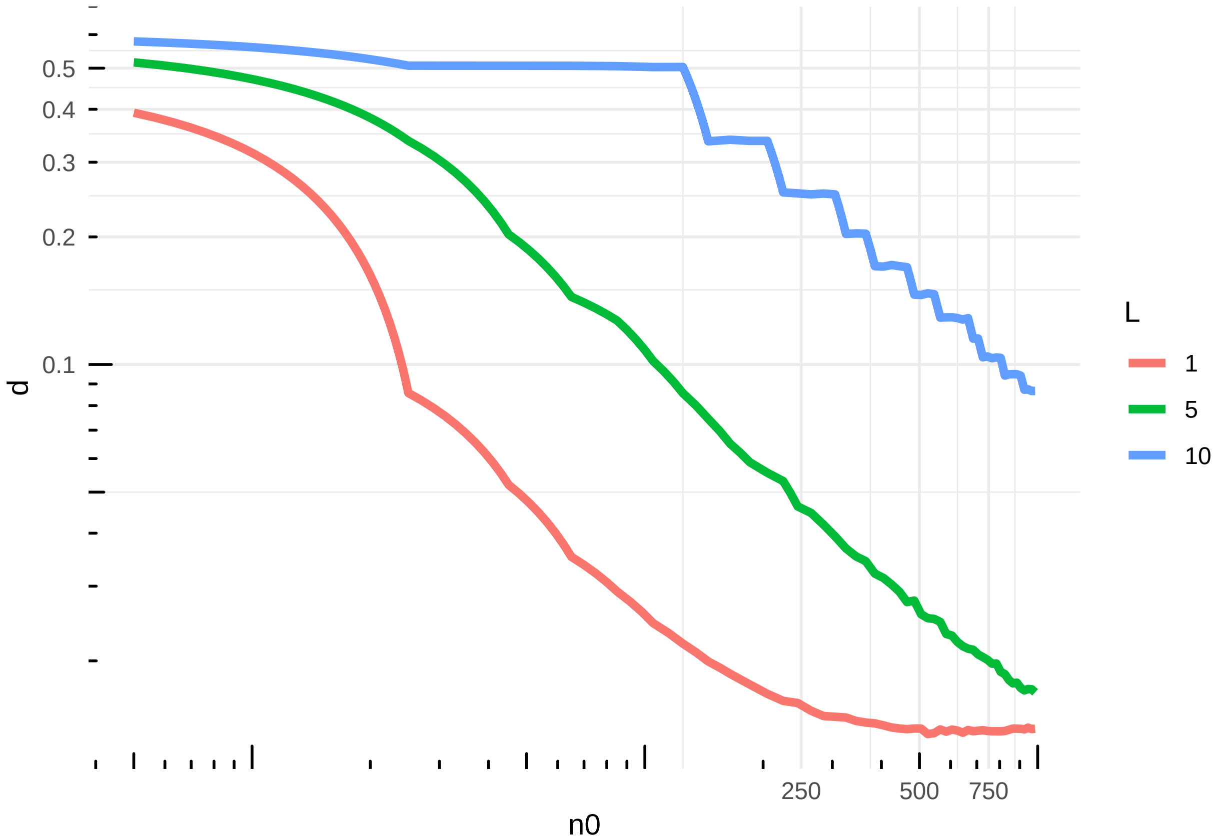
$$zG^2 + \left((1 - z)\frac{\psi}{\phi} - 1\right)G + \frac{\psi}{\phi} = 0.$$

Q4

$\alpha = 1, L = 1, 5, 10, \phi \equiv \frac{n_0}{m} = 1, \psi \equiv \frac{n_0}{n_1} = \frac{3}{2}$. The reproduced experiement is plotted below.

```
start_time <- Sys.time()
set.seed(7017)

library(ggplot2)

alpha <- 1
n0 <- seq(5,1000,20)
phi <- 1
m <- n0 / phi
psi <- 1.5
L <- c(1,5,10)
sigmax <- 1
sigmaw <- 1

f <- function(x,alpha=1) {
    numerator <- max(x,0)+alpha*max(-x,0)-(1+alpha)/sqrt(2*pi)
    denominator <- sqrt(0.5*(1+alpha^2)-0.5/pi*(1+alpha)^2)
```

```r
    return(numerator/denominator)
}

dat <- data.frame()

for (l in L) {
    for (n in n0) {
        X <- matrix(rnorm(n*n,0,sigmax),nrow=n)

        i <- 0
        while (i <= l) {
            if (i==0) {
                Yl <- X
            } else {
                Wl <- matrix(rnorm(ceiling(nrow(Yl)/psi)*nrow(Yl),
                                   0,sigmaw/sqrt(nrow(Yl))),ncol=nrow(Yl))
                # reshape Wl
                Yl <- Wl%*%Yl
                # as f contains max() inside, which ruins our
                # element-wise calculation (in matrix)
                Yl <- matrix(sapply(Yl,f),ncol=ncol(Yl),byrow=T)
            }
            i <- i + 1
        }
        covmat <- Yl%*%t(Yl)
        M <- covmat/nrow(Yl)

        # transform these eigenvalues to density then we are done
        rho_l <- eigen(covmat)$values
        rho_l <- density(rho_l,0.05)$y

        # also tranform this
        rho_1_bar <- -1/pi*Im(0.5-0.5*sqrt(as.complex(1-4/eigen(M)$values)))
        rho_1_bar <- density(rho_1_bar,0.05)$y

        # dist <- mean(abs(rho_1_bar-rho_l))
        # this is an approximation
        # according to derivations in lecture on 27 August
        dist <- mean(abs(1/ceiling(nrow(Yl)/psi)-rho_l))
        dat <- rbind(dat, c(l, n, dist))
    }
}
colnames(dat) <- c("L","n0","d")
dat$L <- factor(dat$L)
ggplot(dat, aes(x=n0,y=d,color=L)) +
    geom_line(size=1.5) +
    coord_trans(x="log10",y="log10") +
    annotation_logticks(scaled=F) +
    theme_minimal()
```




```r
end_time <- Sys.time()

end_time - start_time
```

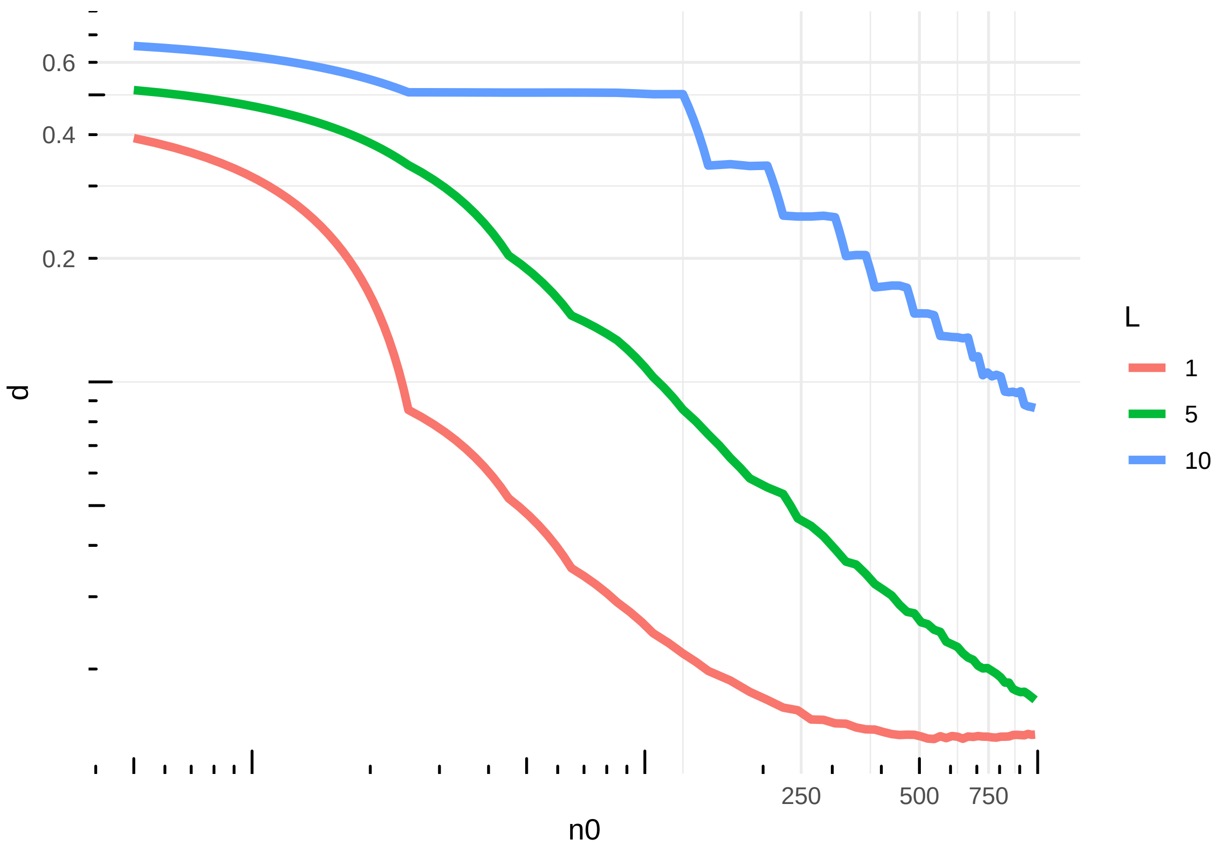
```
## Time difference of 4.863343 mins
```

The whole processing time takes about 3.5 minutes.

As we can see, when the network size (n_0) increases, the distance between the l-th layer empirical eigenvalue distribution and the theoretical first layer limiting distribution is converging to a rather ideal amount.

Q5

Set $\alpha = 0.99 \approx 1$. Another reproduced experiement is plotted below.




```
## Time difference of 4.076794 mins
```

Recall that although we have made some changes in α by setting it slightly below 1, there is no obvious difference from our previous plot.

So generally we can conclude that

- when n_0 increases, ρ_l approaches $\bar{\rho}_1$;
- the value of α does not affect the final convergence, but it might affect "how fast" it converges in some cases;
- the convergence depends on a single scalar statistic of the non-linearity.

Overall, this paper might be a little bit challenging, some part of the derivation is way too abstract. To be honest, there might be quite an amount of typos/errors in this paper, which cause some confusion in the assignment. However, if we look at it from a bright side, at least it does indeed provide an insight in explaining deep learning with random matrix theory.

References

1. LeCun, Bengio, and Hinton, "Deep Learning", *Nature*, 2014.
2. Pennington and Worah, "Nonlinear Random Matrix Theory for Deep Learning", *NIPS*, 2017.
3. C. Louart, Z. Liao, and R. Couillet, "A Random Matrix Approach to Neural Networks", 2017. [Online]. Available: https://arxiv.org/abs/1702.05419. [Accessed: 26-Aug-2018].
4. S. O'Rourke, "A Note on the Marchenko-Pastur Law for a Class of Random Matrices with Dependent Entries", 2012. [Online]. Available: https://arxiv.org/abs/1201.3554. [Accessed: 26-Aug-2018].

5. Z. Liao, R. Couillet, “The Dynamics of Learning: A Random Matrix Approach”, 2018. [Online]. Available: https://arxiv.org/abs/1805.11917. [Accessed: 26-Aug-2018].